Automatic Recognition of Artificial Reverberation Settings in Speech Recordings

Krzysztof KACHNIARZ*
Warsaw University of Technology, Faculty of Electronics and Information Technology, Institute of Radioelectronics and Multimedia Technology, Nowowiejska 15/19, 00-665 Warsaw, kkachni1@mion.elka.pw.edu.pl
**Promity Sp. z o.o., Wiejska 14/25, 00-490 Warsaw*

Marcin LEWANDOWSKI
Warsaw University of Technology, Faculty of Electronics and Information Technology, Institute of Radioelectronics and Multimedia Technology, Nowowiejska 15/19, 00-665 Warsaw, marcin.lewandowski@ire.pw.edu.pl

Abstract

The aim of this study is to create the method for automatic recognition of artificial reverberation settings extracted from a reference speech recordings. The proposed method employs machine-learning techniques to support the sound engineer in finding the ideal settings for artificial reverberation plugin available at a given Digital Audio Workstation (DAW), i.e. Gaussian Mixture Model (GMM) approach and deep Convolutional Neural Network (CNN) VGG13, which is a novel approach. Training set and data set are 1885 speech signals selected from a EMIME Bilingual Database which were processed with 66 artificial reverberation presets selected from Semantic Audio Labs's SAFE Reverb plugin database. Performance of the proposed automatic recognition method was evaluated using similarity measures between features of reference and analysed speech recordings. Evaluation procedure showed that a classical GMM approach gives 43.8% of recognition accuracy while proposed method with VGG13 deep CNN gives 99.94% of accuracy.

Keywords: artificial reverberation, machine learning, digital audio signal processing

1. Introduction

Artificial reverberation is one of the most common digital audio effect used in sound, music or video production. There are three different ways that reverberation can be created and added to a signal: convolution-based, delay-networks and physical modelling. Artificial reverberation algorithms have been under development beginning with [1]. Review of this more than 50 year development process can be found in [2] and during this time, uncountable artificial reverberation algorithms have been implemented. These algorithms running as a software plugins are equipped with numerous presets, which are the combination of various reverberation plugin settings. To efficiently create a desired room impression, the sound engineer must be familiar with all of these settings, which are different for each available plugin. Thus finding the best set of reverberation plugin parameters that identifies desired room acoustic features is time-consuming and non-trivial task. Over the years several techniques to simplify workflow with artificial reverberation plugins have been proposed [3-6]. Recently Reiss et al. [7] proposed a design of an adaptive digital audio effect for

artificial reverberation, controlled directly by desired reverberation characteristics, that allows it to learn from the user in a supervised way.

The aim of this study is to create the method for automatic recognition of artificial reverberation plugin settings (preset) extracted from a reference speech recordings. Based on how precisely the system recognizes proper plugin preset (with known low-level reverberation settings), the user can then fine tune individual parameters of the plugin to create desired acoustic impression. This approach is similar to [6], but instead of using GMM-based (Gaussian Mixture Model) system historically used in speaker recognition [8] the proposed method employs deep Convolutional Neural Network (CNN) VGG13-based technique, which was proposed in [9, 10].

2. Methodology

The proposed method for automatic recognition of artificial reverberation settings uses VGG13 neural network proposed in [10] and technique for signal preprocessing and feature extraction proposed in [9]. Prior to training phase, a long sequences of silence were removed from input signals by segmenting each audio file into frames and thresholding RMS energy of the frames. The silence threshold was chosen to be -56 dBFS. After silence-removal each file was transformed into log-scaled mel-frequency spectrograms with STFT window size of 1024 samples, hop size of 512 and 64 mel bands. Following feature extraction, each feature vector was split into chunks of a fixed size. The next step was data augmentation called mixup [11] as described in [9]. The VGG13 neural network structure was the same as proposed initially in [10].

The reverberation recognition efficiency of the VGG13 neural network approach was compared with efficiency of the GMM-UBM (Gaussian Mixture Model - Universal Background Model) method. This method has been used successfully for speaker recognition systems over the years and in the work [6] it was used for the artificial reverberation recommendation system. In this work the GMM-UBM method from [12] Matlab Toolbox was used. MFCC features were extracted from input signals instead of log-scaled mel-frequency spectrograms in neural network approach. From each input audio file, 20 MFCC cepstral coefficients were extracted. The window size was set to 25 ms and frame intervals were 10 ms. UBM training parameters were set according to [12].

3. Data Set and Reverb Presets

Speech recordings used to train and test VGG13 and GMM-UBM models were obtained from EMIME Bilingual Database [13] i.e. 145 sentences recorded in semi-anechoic chamber by 7 females and 6 males in English language, which makes a total of 1885 several-seconds audio files. The files were sampled at 22 kHz and 16 bit.

Training of VGG13 and GMM-UBM models were conducted with audio files processed with artificial reverberation VST plugin SAFE Reverb [14]. It's an open source software with an open API so there is a lot of various presets made by users. Many of them have parameters that are very similar to each other so based on low-level reverb preset settings cosine similarity was calculated between all of them and 66 presets were selected. To put all 66 reverb presets in context to each other, a classic

multidimensional scaling procedure was performed on correlation matrix between presets and the first two dimensions are presented in Fig. 1.

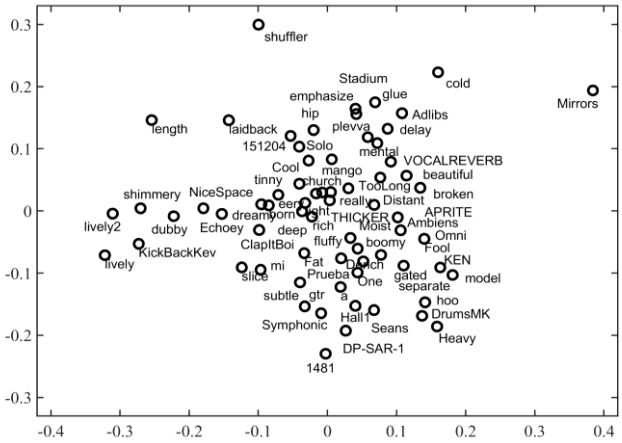

Figure 1. Similarity of the reverb presets displayed after multidimensional scaling

The whole data set used for training and testing comprises of 124410 audio files. Before applying each of the method for artificial reverberation recognition, the whole data set was split up into a three training sets and test sets, in a 90:10 proportion with a hold-out method:

- Standard set - randomly chosen 90% of audio files for a training set and 10% of audio files for a test set,
- Sentence independent set (to check if the model performance is independent of the sentence) - 15 specific sentences were selected for a test set and remaining 130 for a training set,
- Speaker independent set (to check if the model performance is independent of the speaker) - one specific male and one specific female voices were selected for a test set.

4. Results

To measure the performance of the VGG13 and GMM-UBM models, two metrics were chosen: accuracy and confusion matrix. Accuracy defines how often the model correctly recognizes reverberation settings in the reference audio files (taken from test sets). A confusion matrix summarizes the classification performance of a model with respect to test set. It is a two-dimensional matrix, indexed in one dimension by the true (actual) class of an object, e.g. the reverb setting and in the other by the reverb setting that the model assigns (predicts) from the test set audio files. Reverb settings are labelled as preset names from the SAFE Reverb plugin database and confusion matrix is normalized by class support size (number of elements in each class i.e. each reverb preset).

4.1. *Testing with Standard Data Set*

First of all, the training of the two models was performed with 90% of the standard data set. Testing was performed with remaining 10% of the standard data set. Model GMM-UBM achieved an accuracy score of 43.8% and VGG13 achieved an accuracy score of 99.94%. Confusion matrix for two models are shown in Fig. 2 and Fig. 3.

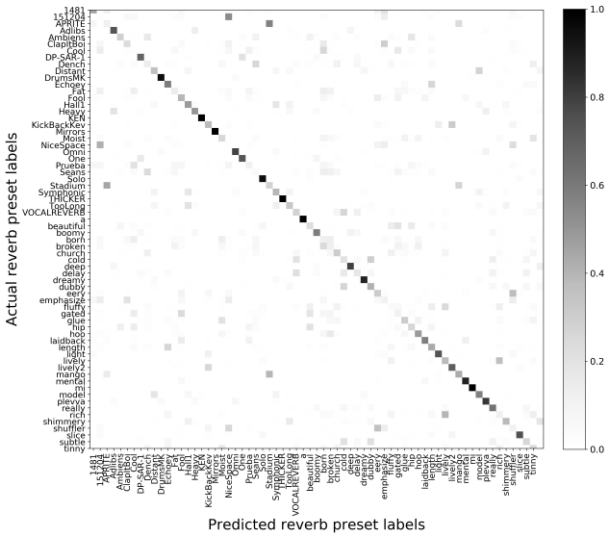

Figure 2. Confusion matrix of GMM-UBM model test with standard data set

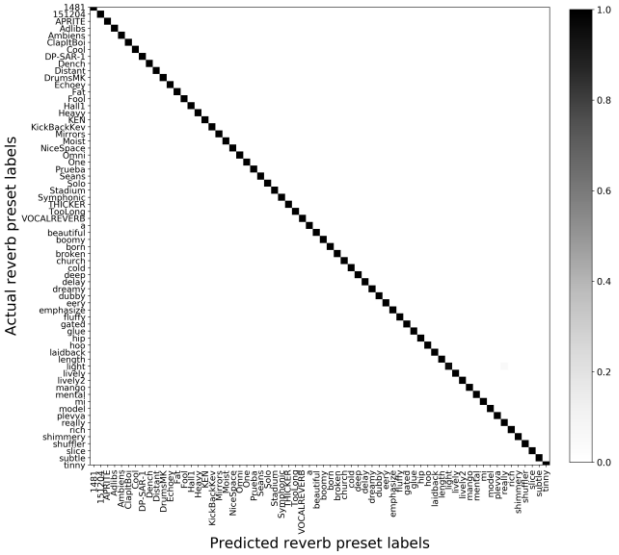

Figure 3. Confusion matrix of VGG13 model test with standard data set

4.2. *Testing with Sentence Independent Data Set*

Training of the two models was performed with 90% of the sentence independent data set. Testing was performed with remaining 10% of the data set. Model GMM-UBM achieved an accuracy score of 65.05% and VGG13 achieved an accuracy score of 99.88%. Confusion matrix for two models are shown in Fig. 4 and Fig. 5.

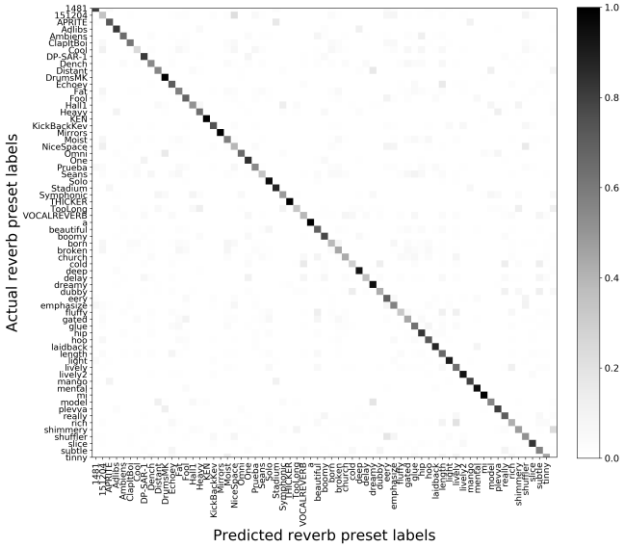

Figure 4. Confusion matrix of GMM-UBM model with sentence independent data set

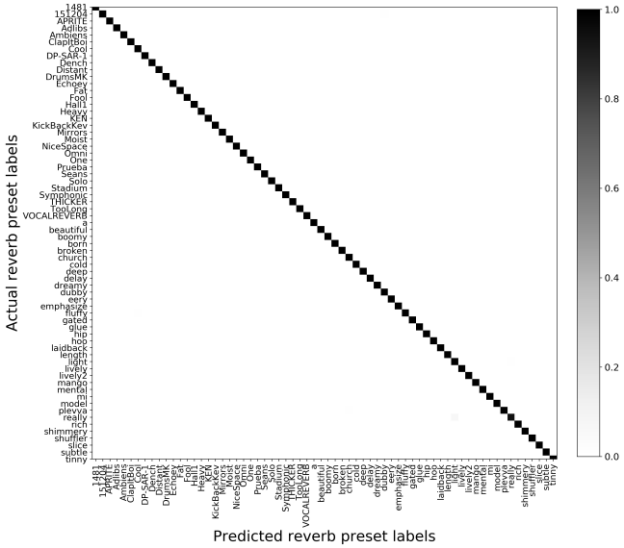

Figure 5. Confusion matrix of VGG13 model test with sentence independent data set

4.3. *Testing with Speaker Independent Data Set*

Training of the two models was performed with 90% of the speaker independent data set. Testing was performed with remaining 10% of the data set. Model GMM-UBM achieved an accuracy score of 47.88% and VGG13 achieved an accuracy score of 99.36%. Confusion matrix for two models are shown in Fig. 6 and Fig. 7.

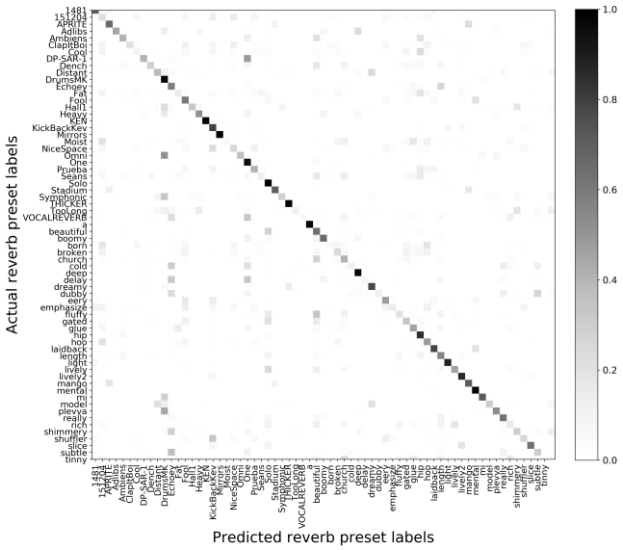

Figure 6. Confusion matrix of GMM-UBM model with speaker independent data set

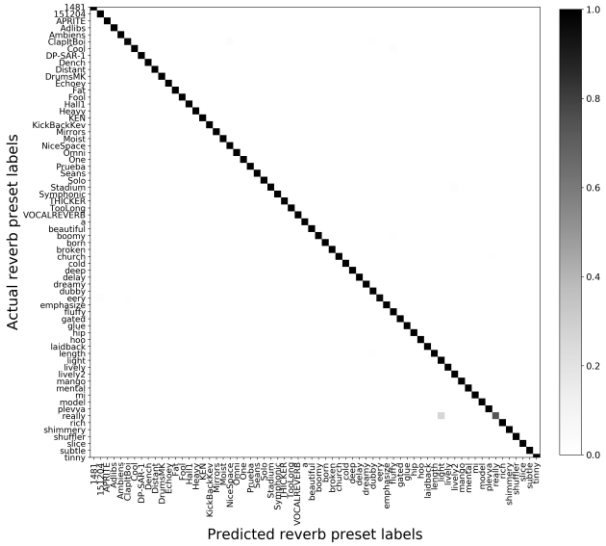

Figure 7. Confusion matrix of VGG13 test model with speaker independent data set

5. Conclusions

The new method for automatic recognition of artificial reverberation settings extracted from a features of the reference speech recordings was presented. This method is based on convolutional neural network trained in a supervised way. For comparison, the previously proposed recommendation system based on Gaussian Mixture Models was tested.

Application of that system could be numerous. The main goal of this system is to support the user in finding a reverberation preset that best matches desired room impression. Therefore, that system could support workflow of audio engineers in audio-video postproduction studios, consumer and professional sound studios or even be one of the features in Digital Audio Workstations. Another application of the proposed system could be support for dereverberation algorithms, which are important especially for speech intelligibility improvement. Algorithms of this kind require information about room acoustic's parameters in which recording has been made. The proposed system, based on recognized reverberation preset, could decode low-level reverb information and then help to suppress reverb level in the recording.

Evaluation tests have been conducted for three different data sets. In each case the accuracy obtained with model of convolutional neural network was much higher than accuracy obtained with model based on Gaussian Mixture Models. The recommendations of our model show that the system is almost always able to suggest similar reverb preset.

Future work will focus on exploring accuracy of the system in case of larger amount of presets and plugins. Also, it could be useful to integrate this model into some dereverberation algorithm and test it.

Acknowledgments

This work was supported by the statutory grant 504/04064/1034/40.00 from the Warsaw University of Technology.

References

1. M. R. Schroeder, B. F. Logan, *Colorless artificial reverberation*, IRE Transactions on Audio, **6** (1961) 209 – 214.
2. V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, J. S. Abel, *Fifty years of artificial reverberation*, IEEE Transactions on Audio, Speech, and Language Processing, **20**(5) (2012) 1421 – 1448.
3. J. Jullien, E. Kahle, M. Marin, O. Warusfel, *Spatializer: a perceptual approach*, 94th Convention of the Audio Engineering Society, Preprint, **3465** (1993).
4. M. F. Zbyszynski, A. Freed, *Control of VST plug-ins using OSC*, Proc. of the International Computer Music Conference, Spain 2005, 263 – 266.
5. Z. Rafii, B. Pardo, *Learning to control a reverberator using subjective perceptual descriptors*, 10th International Society for Music Information Retrieval (2009).

6. N. Peters, J. Choi, H. Lei, *Matching Artificial Reverb Settings to Unknown Room Recordings: a Recommendation System for Reverb Plugins*, 133rd Audio Engineering Society Convention, USA 2012.
7. E. T. Chourdakis, J. D. Reiss, *A machine-learning approach to application of intelligent artificial reverberation*, Journal of the Audio Engineering Society, 2017.
8. D. Reynolds, T. Quatieri, R. Dunn, *Speaker verification using adapted gaussian mixture models*, Digital Signal Processing, **10**(1-3) (2000) 19 – 41.
9. T. Iqbal, Q. Kong, M. Plumbley, W. Wang, *Stacked Convolutional Neural Networks For General-Purpose Audio Tagging*, Centre for Vision, Speech and Signal Processing, University of Surrey 2018.
10. K. Simonyan, A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 3rd ICLR 2015.
11. H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, *mixup: Beyond empirical risk minimization*, 6th ICLR 2015.
12. S. O. Sadjadi, M. Slaney, L. Heck, *MSR Identity Toolbox*, Microsoft Research 2013.
13. M. Wester, *The EMIME Bilingual Database*, The University of Edynburg: Centre for Speech Technology Research 2012.
14. Semantic Audio Labs, [Online], http://www.semanticaudio.co.uk/, [Access: 10.07.2019].