THE COMBINED AUGSBURG / PASSAU / TUM / ICL SYSTEM FOR DCASE 2017

Shahin Amiriparian[1,2,3], *Nicholas Cummins*[1,2], *Micheal Freitag*[1],
Kun Qian[1,2,3], *Zhao Ren*[1,2], *Vedhas Pandit*[1,2], *Björn Schuller*[1,2,4]

[1] Chair of Embedded Intelligence for Health Care & Wellbeing, Universität Augsburg, Germany
[2] Chair of Complex & Intelligent Systems, Universität Passau, Germany
[3] Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[4] GLAM – Group on Language, Audio & Music, Imperial College London, UK
shahin.amiriparian@informatik.uni-augsburg.de

ABSTRACT

This technical report covers the fusion of two approaches towards the Acoustic Scene Classification sub-task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2017). The first system uses a novel recurrent sequence to sequence autoencoder approach for unsupervised representation learning. The second system is based on the late fusion of support vector machines trained on either wavelet features or an archetypal acoustic feature set. A weighted late-fusion combination of these two systems achieved an accuracy of 90.1 % on the official development set and an accuracy of 59.1 % on the test set of the challenge.

1. OUR DCASE 2017 FUSION SYSTEM

Our combined approach for the DCASE 2017 acoustic scene classification challenge [1] is the decision fusion of predictions from two main systems.

Our first main system utilises a recurrent sequence to sequence autoencoder for the unsupervised feature representation from audio data [2]. This system comprises five main steps: (i) the extraction of mel-spectrograms from the raw audio files; (ii) the training of a recurrent sequence to sequence autoencoder on these spectra; (iii) the use of the activations from a fully connected layer between the decoder and encoder units – the learnt representations of the spectra – as the feature vectors; (iv) repetition of the step three for the different spectral representations made possible by the stereo recordings provided in the challenge dataset; and, (v) training of a multilayer perceptron on a greedily fused combination of the learnt feature vectors to predict the class labels. This approach achieved an accuracy of 88.0 % on the challenge development data.

Our second main system is comprised of support vector machines (SVM) models trained with either wavelet packet transform energy features, wavelet energy features or the 6k features *openSMILE* ComParE feature set [3]. A late fusion of these SVM systems achieved a development set accuracy of 83.2 %.

We perform decision-level fusion by computing the weighted sum of the label probabilities for each instance. The label with the highest probability is then chosen as the prediction for the instance. We optimised the weights in $[0; 1]$ in steps of 0.01 on the predefined cross-validation setup. A weight of 0.46 for the SVM based system and a weight of 0.54 for the recurrent sequence to sequence autoencoder resulted in the highest performance, with a classification accuracy of 90.1 %. This weight was then used to fuse the predictions on the evaluation set. The confusion matrix of our fused system on the development set is given in Figure 1.

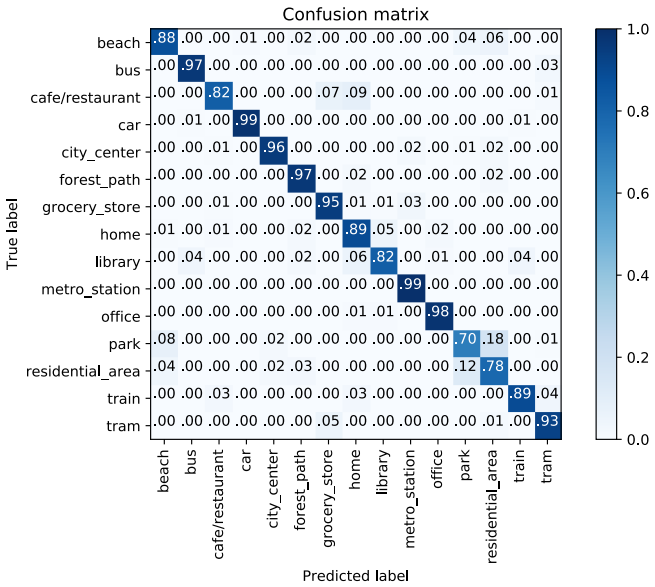

Figure 1: Confusion Matrix of our combined system on the DCASE 2017 development partition which achieved a classification accuracy of 90.1 %.

2. ACKNOWLEDGEMENTS

We thank our colleagues Zijiang Yang, and Zixing Zhang for their efforts and help. This research has received funding from the EU's FP7 under grant agreement No. 338164 (ERC StG iHEARu).

3. REFERENCES

[1] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System," in *Proc. of the DCASE Workshop*, Nov 2017, 5 pages.

[2] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio," in *Proc. of the DCASE 2017 Workshop*, Nov 2017, 5 pages.

[3] K. Qian, Z. Ren, V. Pandit, Z. Yang, Z. Zhang, and B. Schuller, "Wavelets Revisited for the Classification of Acoustic Scenes," in *Proc. of the DCASE 2017 Workshop*, Nov 2017, 5 pages.