# A SYSTEM FOR 2017 DCASE CHALLENGE USING DEEP SEQUENTIAL IMAGE AND WAVELET FEATURES

*Zhao Ren*[1,2], *Kun Qian*[1,2,3], *Vedhas Pandit*[1,2], *Zijiang Yang*[1,2], *Zixing Zhang*[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] MISP group, MMK, Technische Universität München, Germany
[4] Group on Language, Audio & Music, Imperial College London, UK

Zhao.Ren@informatik.uni-augsburg.de, schuller@ieee.org

## ABSTRACT

For the Acoustic Scene Classification task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2017), we propose a novel method to classify 15 different acoustic scenes using deep sequential learning for the audio scenes. First, deep representations extracted from the spectrogram and two types of scalograms using Convolutional Neural Networks, the ComparE features and two types of wavelet features are fed into the Gated Recurrent Neural Networks for classification separately. Predictions from the six models are then combined by a margin sampling value strategy. On the official development set of the challenge, the best accuracy on a four-fold cross-validation setup is 83.3%, which increases 8.5% compared with the baseline ($p < .001$ by one-tailed z-test).

## 1. THE PROPOSED SYSTEM

### 1.1. Setup

The proposed system is combined by the predictions in [1] and [2]. First, we achieved the best performance in [1] from the spectrogram (Short-Time Fourier Transform) and two types of scalograms (*bump* and *morse* wavelet) by Convolutional Neural Networks (CNNs) and Gated Recurrent Neural Networks (GRNNs). Second, the predictions in [2] are obtained from ComParE features and two types of wavelet features (wavelet packet transform energy and wavelet energy features) by GRNNs. Finally, we combine both results by margin sampling value (MSV) strategy, which is described in [1].

### 1.2. Results

The experimental results on development dataset are shown in Table 1. We observe that the performances in both [1] and [2] are significant improvement compared with the baseline 74.8% [3]. Further improvement is observed when fusing that two results. Table 2 shows the confusion matrix for the best performance 83.3%, combining the six models. The performance of our approach on evaluation dataset is 63.8%, outperforming the baseline 61.0%.

## 2. ACKNOWLEDGEMENTS

This work was partially supported by the European Union's Seventh Framework under grant agreements No. 338164 (ERC StG iHEARu), and the China Scholarship Council (CSC).

Table 1: Performance of the two models and the combination results by decision fusion.

| accuracy [%] | **Fold1** | **Fold2** | **Fold3** | **Fold4** | **Mean** |
|---|---|---|---|---|---|
| Image | 82.6 | 80.7 | 78.7 | 81.5 | 80.9 |
| ComParE + Wavelet | 82.6 | 81.8 | 81.0 | 85.0 | 82.6 |
| Image+ComparE+Wavelet | 84.8 | 82.6 | 82.2 | 83.6 | **83.3** |

Table 2: Confusion matrix of the development set for the proposed system, in which the values are averaged by the 4-fold cross validation.

| | | *Prediction* | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *beach* | *bus* | *cafe/rest.* | *car* | *city cent.* | *forest path* | *groc. store* | *home* | *library* | *metro st.* | *office* | *park* | *resid. area* | *train* | *tram* |
| *Actual* | beach | **67** | 0 | 0 | 0 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 1 | 4 | 0 | 1 |
| | bus | 0 | **75** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| | cafe/rest. | 0 | 0 | **62** | 0 | 1 | 0 | 5 | 5 | 1 | 1 | 2 | 0 | 0 | 1 | 1 |
| | car | 0 | 1 | 0 | **75** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| | city cent. | 0 | 0 | 0 | 0 | **71** | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 0 | 0 |
| | forest path | 1 | 0 | 1 | 0 | 2 | **73** | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| | groc. store | 1 | 0 | 4 | 0 | 0 | 0 | **65** | 1 | 1 | 6 | 0 | 0 | 0 | 0 | 0 |
| | home | 1 | 1 | 1 | 0 | 0 | 1 | 1 | **59** | 5 | 0 | 11 | 0 | 0 | 0 | 0 |
| | library | 1 | 0 | 1 | 0 | 0 | 2 | 2 | 5 | **59** | 3 | 3 | 0 | 1 | 2 | 0 |
| | metro st. | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 5 | **71** | 1 | 0 | 0 | 0 | 0 |
| | office | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | **74** | 0 | 0 | 0 | 0 |
| | park | 3 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 1 | 0 | 0 | **55** | 15 | 0 | 0 |
| | resid. area | 2 | 0 | 0 | 0 | 6 | 3 | 0 | 1 | 0 | 0 | 0 | 11 | **55** | 0 | 0 |
| | train | 0 | 9 | 4 | 1 | 6 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | **49** | 8 |
| | tram | 0 | 0 | 1 | 2 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | **68** |

## 3. REFERENCES

[1] Z. Ren, V. Pandit, K. Qian, Z. Yang, Z. Zhang, and B. Schuller, "Deep sequential image features on acoustic scene classification," in *Proc. DCASE Workshop*, 2017, in press.

[2] K. Qian, Z. Ren, V. Pandit, Z. Yang, Z. Zhang, and B. Schuller, "Wavelets revisited for the classification of acoustic scenes," in *Proc. DCASE Workshop*, 2017, in press.

[3] 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. DCASE Workshop*, 2017, in press.