Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)

Tampereen teknillinen yliopisto - Tampere University of Technology

Tuomas Virtanen, Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Emmanuel Vincent, Emmanouil Benetos & Benjamin Martinez Elizalde (eds.)

Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)

TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Tampereen teknillinen yliopisto - Tampere University of Technology

Tuomas Virtanen, Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Emmanuel Vincent, Emmanouil Benetos & Benjamin Martinez Elizalde (eds.)

Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)

Tampere University of Technology. Laboratory of Signal Processing
Tampere 2017

ISBN 978-952-15-4042-4

DEEP SEQUENTIAL IMAGE FEATURES FOR ACOUSTIC SCENE CLASSIFICATION

Zhao Ren[1,2], *Vedhas Pandit*[1,2], *Kun Qian*[1,2,3], *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]GLAM – 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, based on features extracted from Short-Time Fourier Transform and scalogram of the audio scenes using Convolutional Neural Networks. It is the first time to investigate the performance of *bump* and *morse* scalograms for acoustic scene classification in an according context. First, segmented audio waves are transformed into a spectrogram and two types of scalograms; then, 'deep features' are extracted from these using the pre-trained VGG16 model by probing at the fully connected layer. These representations are then fed into Gated Recurrent Neural Networks for classification separately. Predictions from the three systems are finally 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 80.9%, which increases by 6.1% when compared with the official baseline ($p < .001$ by one-tailed z-test).

Index Terms— Audio Scene Classification, Deep Sequential Learning, Scalogram, Convolutional Neural Networks, Gated Recurrent Neural Networks

1. INTRODUCTION

As a sub-field of computational auditory scene analysis (CASA) [1], acoustic scene classification attempts to identify the acoustic environment. It has been used in several applications such as context-aware computing [2], mobile robots [3], serious games [4] and many more. This year's scene classification task of the IEEE AASP Challenge – DCASE2017 [5] – provides a unique opportunity to present models and audio feature representations customised for this task. The challenge requires the participants to classify the audio data into fifteen classes based on the acoustic scene they represent. The corpus has been divided into a non-public evaluation set and four-folds, each featuring training set and development set.

Different from the representations extracted from 1D audio samples directly, such as energy, frequency, and voice-based features [6], features extracted from 2D spectrograms recently show significant improvement in music [7], snore sound [8], and acoustic scene classification [9]. Those methods mostly extract Short-Time Fourier Transformation (STFT) spectrograms from audio waves. In contrast, wavelet transformation incorporates multiple scales and localisation as an extension of Fourier transform to reach the optimum of the time-frequency resolution trade-off. Wavelet features have been shown to be efficient in snore sound classification [10–12] and acoustic scene classification [13] recently. Motivated by this success, we additionally investigate and present the capability of the deep feature representations of two types of scalograms in this study for the first time to our best knowledge in combination with pre-trained deep nets for image classification.

In the recent years, Convolutional Neural Networks (CNNs) became popular in deep learning for visual recognition tasks [14] thanks to their capability of highly nonlinear mapping of input images to output labels. Several CNN structures have been presented in succession, such as AlexNet [15], VGG [16], GoogLeNet [17], and ResNet [18]. It is also worth to design CNNs for processing spectrograms in acoustic tasks [7, 9, 19]. But as CNNs trained on a large scale data set are more robust and stable than those neural networks trained on relatively smaller number of (audio) samples, it might be worthwhile to reuse such nets to extract features from the spectrogram or scalogram for acoustic or other acoustic tasks through transfer learning [20].

In the transfer learning context, feeding powerful representations from CNNs into a classifier, such as a support vector machine (SVM), could achieve good prediction results [8]. However, as acoustic scene audio waves are relatively longer than speech signals which happen in a short time, there is a limitation in learning sequence by sequence using SVM. To break this bottleneck, several models for sequential learning have been proposed, including recurrent neural networks (RNNs) [21], long short-term memory (LSTM) RNNs [22], or Gated Recurrent Neural Networks (GRNNs) [23]. LSTM RNNs and CNNs are combined in [24] to improve the classification performance. Hence, sequential learning methods will be helpful to achieve a higher performance.

The contribution of our approach for acoustic scene classification is described as follows. First, we propose to extract the sequential features from a spectrogram (STFT) and two types of scalograms, namely (*bump* wavelet [25] and *morse* wavelets [26]), by the VGG16 model [16]. Second, we connect CNNs with a sequential learning method by feeding the features into GRNNs. Finally, the predictions from the three models are fused by the margin sampling value method. To our best knowledge, very little research has been undertaken exploring deep CNN feature representations of scalograms on sequential learning in audio analysis, let alone for acoustic scene classification.

In the following, our work aims at proposing a novel approach that makes use of CNNs and GRNNs by transfer learning, as well as presenting the experimental results obtained on DCASE2017.

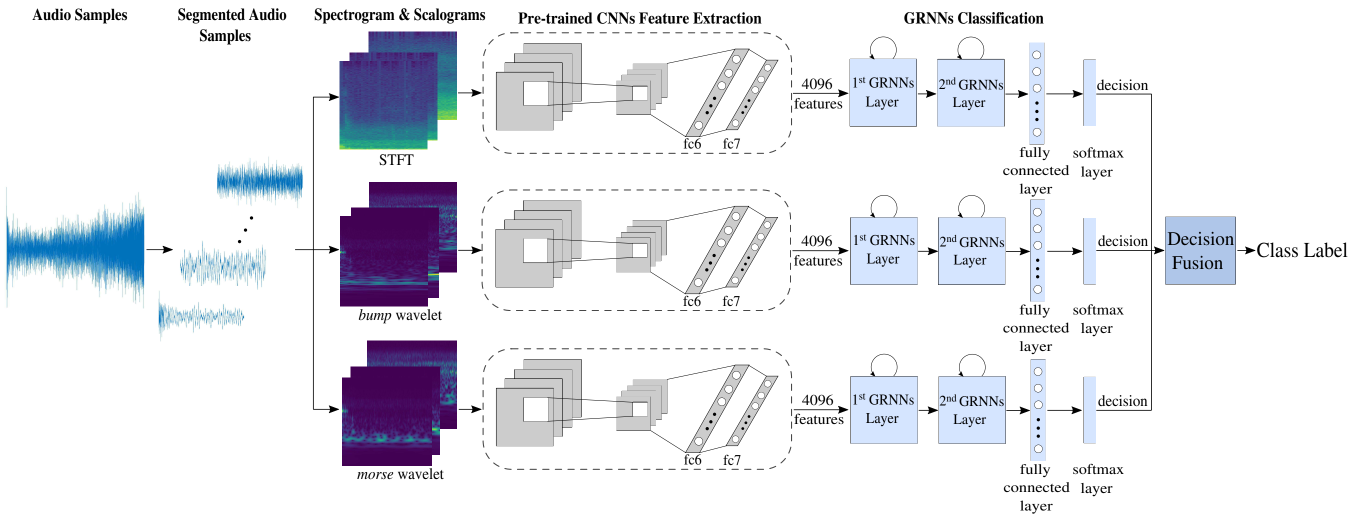


Figure 1: Framework of our proposed system. One type of spectrogram (STFT) and two types of scalograms (*bump* and *morse* wavelet) are generated from the segment audio files. Next, we use pre-trained CNNs to extract features from these images at the first fully connected layer *fc6*. After that, the features are fed into GRNNs to be trained and the final predictions are combined by a decision fusion strategy.

2. METHODOLOGY

An overview of our system can be seen in Figure 1. It mainly includes three components: deep sequential feature extraction, GRNN classification, and decision fusion, which will be introduced in this section after presenting the task.

2.1. Database

We evaluate our proposed system on the dataset of the acoustic scene classification task in the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events [5]. Each recording is split into several independent 10 s segments. The dataset contains 15 classes, including *beach*, *bus*, *cafe/restaurant*, *car*, *city center*, *forest path*, *grocery store*, *home*, *library*, *metro station*, *office*, *park*, *residential area*, *train*, and *tram*. The database is divided into an unlabelled evaluation set and four folds, each of which contains a training set and a development set. For each class, the development set contains 312 segments of 10 s from 52 minutes of audio recordings.

2.2. Deep Sequential Image Feature Extraction

2.2.1. *Spectrograms*

As written, we generate a STFT spectrogram and two types of scalograms, which are now described in more detail as follows.

a) STFT spectrogram. We use the STFT algorithm [27] with a Hamming window, a frame time of 40 ms and overlap of 20 ms, to compute the power spectral density by the dB power scale. At the time t, for a signal $x(t)$ with window function $\omega(t)$ and time index τ, the STFT is defined by

$$X(\tau,\omega) = \int_{-\infty}^{\infty} x(t)\omega(t-\tau)e^{-j\omega t}. \qquad (1)$$

b) Bump scalogram. As a special type of continuous wavelet transform (CWT), the *bump* wavelet [28] with the scale s and the window ω is defined in the Fourier domain as

$$\Psi(s\omega) = e^{\left(1-\frac{1}{1-(s\omega-\mu)^2/\sigma^2}\right)} 1_{[(\mu-\sigma)/s,(\mu+\sigma)/s]}, \qquad (2)$$

where μ and σ are two constant parameters.

c) Morse scalogram. The *morse* wavelet generation is proposed in [26], in which it is defined by

$$\Psi_{P,\gamma}(\omega) = u(\omega)\alpha_{P,\gamma}\omega^{\frac{P^2}{\gamma}} e^{-\omega^{\gamma}}, \qquad (3)$$

where $u(\omega)$ means the unit step, ω is the window, $\alpha_{P,\gamma}$ stands for a normalising constant, P and γ are the time-bandwidth product and the symmetry.

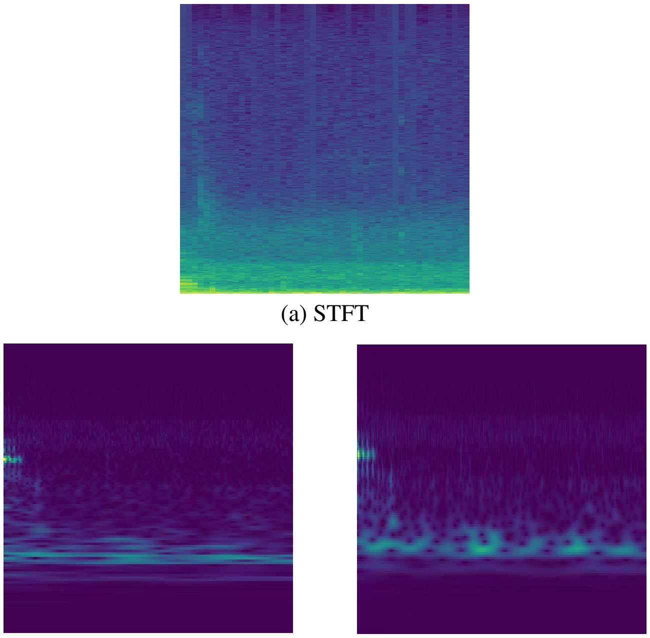
The spectrogram and scalograms are plotted by MATLAB using the *viridis* colour map, which was shown to be better suited than other colour maps or a gray image in [8]. It is a uniform colour map varying from blue to green to yellow. Moreover, the plots are (obviously) made to have no axes or margins, and are scaled to squared images with 224×224 pixels for VGG16-based feature extraction, as shown in Figure 2.

2.2.2. *Convolutional Neural Networks*

With the spectrograms and scalograms for acoustic scenes, the pre-trained CNNs are employed to extract our deep spectrum features. We use the VGG16 model provided by MatConvNet [29] as it worked successfully in the ImageNet Challenge 2014 [1]. VGG16 consists of 16 layers, including 13 convolutional layers, and 3 fully connected layers. The convolutional layers are split into five stacks with maxpooling layers, which use the same kernel size 3×3 and different numbers of channels [64, 128, 256, 512, 512]. The final fully connected layer is followed by a softmax function to generate a 1000-label classification on ImageNet data set. A framework of the VGG16 architecture is described in Table 1.

For our deep sequential feature extraction, the images are fed into the pre-trained VGG16 as input and the features are extracted

[1] http://image-net.org/challenges/LSVRC/2014/results



(a) STFT

(b) *bump* wavelet

(c) *morse* wavelet

Figure 2: The STFT spectrogram and two types of scalograms for the acoustic scenes. All of the images are extracted from the first audio sequence of DCASE2017's "*a001_0_10.wav*" with a label *residential area*.

from the activations on the first fully connected layer *fc6*, which includes 4096 neurons. Therefore, we extract deep features with 4096 attributes from all segmented audios.

2.3. Gated Recurrent Neural Networks

Similar to LSTM-RNNs, GRNNs are able to learn sequence information as a special type of RNN. The GRNNs contain a gated recurrent unit (GRU) [23], which consists of two gating units (reset gate r and update gate z), an activation h, and a candidate activation $\tilde{h}$, as shown in Figure 3. Different from LSTM, the information flows inside the GRU without separate memory cells so that a GRU needs less parameters. Hence, GRNNs converge faster than LSTM, i. e., they need less epochs during training iterations to obtain the best model [23].

Based on the deep sequence features extracted by pre-trained CNNs, we design a two-layer GRNN, which is followed by a fully connected layer and a softmax layer (see Figure 1). Therefore, we obtain the classification predictions from the softmax layer for the three different feature sets.

2.4. Decision Fusion

To improve the performance of our system, we apply a decision fusion method on the three classification results from different feature sets. The Margin Sampling Value (MSV) method is introduced in [30] as the difference of the first and second highest posteriori probabilities for each predicted label of the test sample. We obtain the final label by selecting the model which has the maximum MSV, which is the most confident among the three models.

Table 1: Configurations of the VGG16 convolutional neural networks. 'conv' denotes convolutional layers, size means receptive field size, and 'ch' stands number of channels.

Input: 224×224 RGB image
2×conv size: 3; ch: 64 Maxpooling
2×conv size: 3; ch: 128 Maxpooling
3×conv size: 3; ch: 256 Maxpooling
3×conv size: 3; ch: 512 Maxpooling
3×conv size: 3; ch: 512 Maxpooling
Fully connected layer *fc6* with 4096 neurons Fully connected layer *fc7* with 4096 neurons Fully connected layer with 1000 neurons
Output: softmax layer of probabilities for 1000 classes

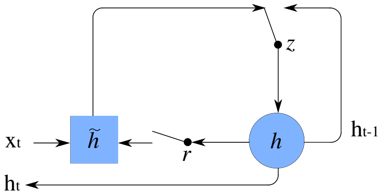


Figure 3: Illustration of a Gated Recurrent Unit (GRU) [23]. r and z represent the reset and update gates separately, and h and $\tilde{h}$ are the activation and the candidate activation.

3. EXPERIMENTS

3.1. Setup

Each audio file is first segmented into a sequence of 19 audio samples with 50 % overlap and 1 000 msec duration. Next, we apply the pre-trained VGG16 model provided by the MATLAB toolbox MatConvNet [29] on the STFT spectrogram, *bump* and *morse* wavelet scalograms, and features are extracted from activations in the first layer *fc6* of the VGG16 model. Then, we use two-layer GRNNs (120–60), followed by a fully connected layer and a softmax layer. We implement this architecture in TensorFlow[2] and TFLearn[3] with a fixed *learning rate* of 0.0002 (the optimiser is 'rmsprop'), and train it for 30, 50, and 70 *epochs*. Finally, the margin sampling value decision fusion method described in Section 2.4 is selected to combine the three neural networks to obtain the final predictions.

3.2. Results

We train the three models in parallel but end at different epochs. Table 2 presents the performances of the 4-fold evaluation on the development set and their mean accuracies. We can see that, all

[2] https://github.com/tensorflow/tensorflow

[3] https://github.com/tflearn/tflearn

Table 2: Performance comparison of different epochs (*epoch*∈{30, 50, 70}) of GRNNs on features extracted by CNNs from STFT spectrogram, *bump*, and *morse* scalograms. The GRNNs are implemented in two layers with 120 and 60 GRU cells in each layer and a *learning rate*=0.0002.

accuracy [%]	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Mean**
(a) STFT					
epoch 30	77.9	72.5	73.1	79.3	75.7
epoch 50	79.2	74.7	74.3	77.7	**76.5**
epoch 70	77.1	75.8	72.9	77.4	75.8
(b) *bump* wavelet					
epoch 30	74.5	75.4	73.9	77.2	**75.2**
epoch 50	73.6	72.9	73.6	73.2	73.3
epoch 70	69.7	73.4	72.6	72.1	72.0
(c) *morse* wavelet					
epoch 30	74.5	75.4	73.9	77.2	**75.2**
epoch 50	73.6	72.9	73.6	73.2	73.3
epoch 70	69.7	73.4	72.6	72.1	72.0

Table 3: Performance comparison of different combinations of the three feature sets by decision fusion on the multi-class classifier GRNNs. The GRNNs are implemented in two-layers (120-60), *learning rate*=0.0002, *epoch*∈{30, 50, 70,*the best epoch* (STFT: 50, *bump*: 30, *morse*: 30)}. All of the models are first trained independently, and then combined to make a final decision by the MSV method.

	accuracy [%]	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Mean**
epoch 30	STFT+bump	80.9	79.9	77.5	82.2	80.1
	STFT+morse	79.8	79.4	76.8	81.5	79.4
	bump+morse	76.7	77.5	76.0	77.5	76.9
	STFT+bump+morse	80.9	80.1	78.7	81.7	**80.3**
epoch 50	STFT+bump	81.9	78.4	77.6	80.9	79.7
	STFT+morse	81.5	80.4	76.4	79.7	79.5
	bump+morse	76.7	77.5	76.0	77.5	76.9
	STFT+bump+morse	82.1	79.9	78.4	80.0	**80.1**
epoch 70	STFT+bump	80.3	78.8	76.6	81.7	79.4
	STFT+morse	80.3	78.3	76.1	80.9	78.9
	bump+morse	72.8	75.9	72.5	76.1	74.3
	STFT+bump+morse	80.2	79.1	77.2	82.7	**79.8**
best epoch	STFT+bump	82.6	79.5	77.5	80.9	80.1
	STFT+morse	81.1	80.0	76.5	81.5	79.8
	bump+morse	76.7	77.5	76.0	77.5	76.9
	STFT+bump+morse	82.6	80.7	78.7	81.5	**80.9**

Table 4: Confusion matrix of the development set for the proposed method, in which the values are averaged in the 4-fold cross validation. Our proposed approach achieves an accuracy of 80.9%.

		Prediction														
		beach	bus	cafe/rest.	car	city cent.	forest path	groc. store	home	library	metro st.	office	park	resid. area	train	tram
Actual	beach	**68**	0	0	0	2	1	0	1	0	0	0	2	4	0	1
	bus	0	**74**	0	1	0	0	0	1	0	0	0	0	0	2	1
	cafe/rest.	0	0	**59**	0	1	0	6	5	1	1	2	0	0	0	3
	car	0	1	0	**74**	0	0	0	0	0	0	0	0	0	1	3
	city cent.	0	0	0	0	**66**	0	1	0	0	1	0	2	8	0	0
	forest path	1	0	1	0	3	**71**	0	0	0	0	0	0	2	0	0
	groc. store	1	0	5	0	0	0	**61**	0	2	6	1	1	0	0	1
	home	0	2	2	0	0	0	1	**61**	5	0	9	0	0	0	0
	library	1	0	1	0	0	3	2	3	**58**	4	3	0	1	2	0
	metro st.	0	0	0	0	1	0	3	0	4	**71**	0	0	0	0	0
	office	0	0	0	0	0	0	0	4	1	0	**73**	0	0	0	0
	park	4	0	0	0	5	0	0	0	2	0	1	**48**	19	0	0
	resid. area	2	0	0	0	7	3	0	1	1	1	1	13	**50**	1	0
	train	0	7	3	2	8	0	0	1	1	1	0	0	3	**45**	8
	tram	0	0	1	2	1	0	2	0	0	1	0	1	0	3	**69**

performances from the three models are comparable with the baseline of the DCASE2017 challenge. The results of STFT and *bump* wavelets perform slightly better than the baseline. We find that the best accuracy of each model is obtained at different epochs. For the STFT spectrogram, we observe the best performance (76.5%) at *epoch* 50, but both for *bump* and *morse* wavelet (75.2% and 72.6%) at *epoch* 30.

Therefore, we apply late-fusion to the three GRNNs results to obtain the final results, as shown in Table 3. We observe that epochs affect the performances substantially. The similarity, however, is that, the best performances in all epochs are achieved when combining results of all three models, corroborating our assumption that scalograms are efficient for acoustic scene classification. A further improvement is observed for the combination of the best epoch from each feature representation (STFT: 50, *bump*: 30, *morse*: 30): up to 80.9% accuracy with a significant improvement of 6.1% accuracy over the baseline of the DCASE2017 challenge ($p < .001$ in a one-tailed z-test [31]). Table 4 shows the confusion matrix for the best performance, combining the best epoch for the neural network of each model. Some classes, such as *beach* and *car*, are classified with high accuracies, while others, such as *park* and *residential area*, are not easy to be recognised.

To sum up, our proposed scalograms are helpful to improve the performance on acoustic scene classification, and the presented approach which connects sequence learning with pre-trained CNNs can increase the accuracy on this task.

4. CONCLUSIONS

We proposed a method for classifying acoustic scenes that relies on the ability of deep pre-trained CNNs to extract useful features from STFT and wavelet representations. Using our deep image spectrum features on GRNNs as a sequential learning method, we were able to improve the performance significantly on the official development set of the DCASE2017 challenge in a 4-fold cross validation, achieving an accuracy of 80.9% ($p < .001$ in a one-tailed z-test). In our experiments, we found that wavelet features are helpful to increase the accuracy when combining with STFT spectrogram representations. In future works, we will investigate which CNNs infer the best representations from our audio representations, and experiment with data augmentations of the training data.

5. 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).

6. REFERENCES

[1] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proc. of ICASSP*, vol. 2. IEEE, 2002, pp. II–1941.

[2] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Proc. of WMCSA*. IEEE, 1994, pp. 85–90.

[3] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am i? scene recognition for mobile robots using audio features," in *Proc. of ICME*. IEEE, 2006, pp. 885–888.

[4] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the ACM MM*, Barcelona, Catalunya, Spain, 2013, pp. 835–838.

[5] 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, in press.

[6] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of ACM MM*. ACM, 2013, pp. 835–838.

[7] G. Gwardys and D. Grzywczak, "Deep image features in music information retrieval," *International Journal of Electronics and Telecommunications*, vol. 60, no. 4, pp. 321–326, 2014.

[8] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proc. of INTERSPEECH*. Stockholm, SE: ISCA, 2017, 5 pages.

[9] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "Dcase 2016 acoustic scene classification using convolutional neural networks," in *Proc. of DCASE Workshop*, 2016, pp. 95–99.

[10] K. Qian, C. Janott, Z. Zhang, C. Heiser, and B. Schuller, "Wavelet features for classification of vote snore sounds," in *Proc. of ICASSP*, Shanghai, China, 2016, pp. 221–225.

[11] K. Qian, C. Janott, J. Deng, C. Heiser, W. Hohenhorst, M. Herzog, N. Cummins, and B. Schuller, "Snore sound recognition: on wavelets and classifiers from deep nets to kernels," in *Proc. of EMBC*, 2017, pp. 3737–3740.

[12] K. Qian, C. Janott, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1731–1741, 2017.

[13] 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*, Munich, Germany, 2017, in press.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. of NIPS*, 2012, pp. 1097–1105.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of CVPR*, 2015, pp. 1–9.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016, pp. 770–778.

[19] J. Schluter and S. Bock, "Improved musical onset detection with convolutional neural networks," in *Proc. of ICASSP*. IEEE, 2014, pp. 6979–6983.

[20] J. Deng, N. Cummins, J. Han, X. Xu, Z. Ren, V. Pandit, Z. Zhang, and B. Schuller, "The university of passau open emotion recognition system for the multimodal emotion challenge," in *Proc. of CCPR*. Springer, 2016, pp. 652–666.

[21] D. P. Mandic, J. A. Chambers, *et al.*, *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. Wiley Online Library, 2001.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[24] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. of DCASE Workshop*, 2016.

[25] S. Abdullah, J. Choi, J. Giacomin, and J. Yates, "Bump extraction algorithm for variable amplitude fatigue loading," *International Journal of Fatigue*, vol. 28, no. 7, pp. 675–691, 2006.

[26] S. C. Olhede and A. T. Walden, "Generalized morse wavelets," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2661–2670, 2002.

[27] E. Sejdić, I. Djurović, and J. Jiang, "Time–frequency feature representation using energy concentration: An overview of recent advances," *Digital Signal Processing*, vol. 19, no. 1, pp. 153–183, 2009.

[28] I. Daubechies, *Ten lectures on wavelets*. SIAM, 1992.

[29] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proc. of ACM MM*. ACM, 2015, pp. 689–692.

[30] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Proc. of IDA*. Springer, 2001, pp. 309–318.

[31] M. R. Spiegel, J. J. Schiller, R. A. Srinivasan, and M. LeVan, "Probability and statistics," 2009.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-4042-4