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

WAVELETS REVISITED FOR THE CLASSIFICATION OF ACOUSTIC SCENES

Kun Qian[1,2,3], *Zhao Ren*[2,3], *Vedhas Pandit*[2,3], *Zijiang Yang*[2,3], *Zixing Zhang*[3], *Björn Schuller*[2,3,4]

[1] MISP Group, MMK, Technische Universität München, Germany
[2] Chair of Embedded Intelligence for Health Care & Wellbeing, Universität Augsburg, Germany
[3] Chair of Complex & Intelligent Systems, Universität Passau, Germany
[4] GLAM – Group on Language, Audio & Music, Imperial College London, UK

andykun.qian@tum.de, schuller@ieee.org

ABSTRACT

We investigate the effectiveness of wavelet features for acoustic scene classification as contribution to the subtask of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2017). On the back-end side, gated recurrent neural networks (GRNNs) are compared against traditional support vector machines (SVMs). We observe that, the proposed wavelet features behave comparable to the typically-used temporal and spectral features in the classification of acoustic scenes. Further, a late fusion of trained models with wavelets and typical acoustic features reach the best averaged 4-fold cross validation accuracy of 83.2 %, and 82.6 % by SVMs, and GRNNs, respectively; both significantly outperform the baseline (74.8 %) of the official development set ($p < 0.001$, one-tailed z-test).

Index Terms— Acoustic Scene Classification, Wavelets, Support Vector Machines, Sequence Modelling, Gated Recurrent Neural Networks

1. INTRODUCTION

Acoustic scene classification (ASC) is defined as classification of the environment in which a recording has been made [1], which is a subfield of Computational Auditory Scene Analysis [2]. It is based on the assumption that, various acoustic scenes can be distinguished from one another by their general acoustic properties due to general characterisations of a location or situation [1]. In practice, ASC is a challenging task since a certain scene is usually similar to others, and shares commonalities of sound sources all across [3]. In recent years, there is an increasing interest in finding more robust and efficient ASC methods to be applied into multimedia searching [4], smart mobile devices [5], and intelligent monitoring systems [6, 7]. Previous DCASE Challenges in 2013 [8], and 2016 [9] attracted numerous teams from across the world working on this uprising topic. A recent overview on the ASC literature is found in [10]. The acoustic features used for ASC include mel-frequency cepstral coefficients [5, 11], histograms of sounds [12], and histogram of gradients learnt from time-frequency representations [13]. In terms of the classifiers, hidden Markov models (HMMs) [12], Gaussian mixture models (GMMs) [11], and support vector machines (SVMs) [13, 14] have been popular. More recently, a series of methods using deep learning are applied to ASC tasks [3, 15–18].

In this contribution, we investigate the effectiveness of wavelet features for the ASC task, which had been proven to be successful in our previous work in snore sound classification [19–21]. A large scale typical acoustic feature vector extracted by openSMILE [22] is compared with, and combined in a late fusion process with the proposed wavelet features. As to machine learning models, SVMs [23], and gated recurrent neural networks (GRNNs) [24] are implemented and compared. To train both models, we first extract *low level descriptors* (LLDs) on the frame level; then, we apply different functionals over *clips*. It is worth noting that the *clip* for the SVM model refers to a long segment, whereas for the GRNNs, it denotes a short episode which is *sequentially* segmented from a long segment in a fixed duration length. Both for the SVMs and the GRNNs, the models trained independently with wavelets and typical acoustic features, are fused later to make the final decision by a *margin sampling* strategy [25].

In comparison to the enormous focus on cepstral and other spectral features that do not optimise the Heisenberg-alike time-frequency trade-off, there is little attention on the effectiveness of wavelet features for the ASC task. This work thus explores avenues towards accordingly optimised novel features which are efficient in classification of acoustic scenes, and to investigate their performances by the popular classifiers such as SVMs, and state-of-the-art machine learning techniques like GRNNs.

This paper is organised as follows: Section 2 will give a description of the database and the methodology we used. The experimental results will be shown in Section 3 before a conclusion is made in Section 4.

2. METHODOLOGY

2.1. Database

To evaluate the proposed systems, we use the official dataset of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2017) [1]. This dataset is accessible through the challenge website [1]. The development set contains 312 segments of 10 seconds in each of the 15 classes. The total duration of the development set is 13 hours. The fifteen acoustic scene classes needed to be recognised in this task are: *beach*, *bus*, *cafe/restaurant*, *car*, *city centre*, *forest path*, *grocery store*, *home*, *library*, *metro station*, *office*, *park*, *residential area*, *train*, and *tram*. The organisers split the data such that all segments from the same recording are put in either the train or the test partition, for all of the four folds they generated for cross-validation. This is done to evaluate robustness of the proposed systems. The correct recognition *accuracy* is used as the evaluation metric.

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/

Table 1: Parametres for wavelet features.

	Wavelet Function	$J_{\mathbf{max}}$	# of LLDs
WPTE	'rbio3.3'	7	255
WEF	'db7'	7	287

2.2. Wavelet Features

We earlier introduced wavelet features into the area of snore sound classification in [19]. Wavelets were found to be effective in localising different snore sounds, and performed better than the other widely-used spectral feature-types such as *mel-frequency cepstral coefficients*, *formants*, *fundamental frequency*, etc. One most recent work on deep wavelet features for ASC was presented in [26]. Firstly, we use the wavelet packet transform energy (WPTE) feature extracted by wavelet packet transformation (WPT) [27]. In contrast to the discrete wavelet transformation (DWT) [28], WPT further decomposes 'detail' components to obtain their own 'approximation'. We use the normalised bank filter energy in [29] as our WPTE LLDs, which is defined as:

$$\mathbf{E}_{\mathbf{\Omega}_{j,k}} = \log \sqrt{\frac{\sum_{n=1}^{N_{j,k}} (\mathbf{w}_{j,k,n})^2}{N_{j,k}}}, \quad (1)$$

where $\mathbf{w}_{j,k,n}$ are the coefficients calculated by WPT from the analysed signal at the subspace $\mathbf{\Omega}_{j,k}$. $N_{j,k}$ is the total number of wavelet coefficients in the k-th subband at the j-th decomposition level. The scale of k is $0, 1, 2, \ldots, 2^j - 1$. Totally, $2^{\boldsymbol{J_{max}}+1} - 1$ WPTE based LLDs are generated. The $\boldsymbol{J_{max}}$ is the maximum level for wavelet decomposition by a certain *wavelet function*.

In addition, another wavelet feature set based on DWT is defined as:

$$\widetilde{\mathbf{E}}_{\mathbf{\Omega}_j} = \frac{(\mathbf{w}_j)^2}{\sum_{j=1}^{\boldsymbol{J_{max}}} (\mathbf{w}_j)^2} \times 100, \quad (2)$$

where $\mathbf{w}_j$ are the coefficients generated by DWT at the j-th decomposition level. Furthermore, the *mean*, *variance*, *waveform length* (the sum of the absolute differences), and *entropy* are calculated from the vector (see Eq. 2) as LLDs. In total, for a j-th decomposition of DWT, this procedure generates $4 \times (\boldsymbol{J_{max}} + 1)$ LLDs.

We combine the features extracted according to Eq. 1 with Eq. 2, and refer to them as wavelet energy features (WEFs) as in [21]. Subsequently, four statistical *functionals*, i. e., *maximum*, *mean*, *minimum*, and *bias* of the estimated linear regression on the frame-level features are applied to the LLDs of WPTE, and WEF. These four selected *functionals* are shown to be efficient in [21]. The *wavelet function* was selected empirically based on initial experiments, which are shown in Table 1, where the $\boldsymbol{J_{max}}$ and dimensions of LLDs of wavelets are included as well. The wavelet function names and the decomposition scripts are based on the Wavelet Toolbox [2] of Matlab by MathWorks.

2.3. Temporal and Spectral Features

We use our toolkit openSMILE [22] to extract the large scale temporal and spectral features, which has been proven to be efficient on the ASC task in DCASE2016 [3]. In this study, we chose the INTERSPEECH ComParE feature set [30]. This feature set contains the 'usual suspects' of most popular acoustic features like *mel-frequency cepstral coefficient* (MFCC), *root mean square* (RMS) energy, *harmonics to noise ratio* (HNR), and others. The LLDs are presented in Table 2, and some statistical *functionals* are applied to these LLDs, which generate 6 373 features in total from one audio sample. Details of the features can be found in [31].

[2] http://www.mathworks.com/products/wavelet/

Table 2: ComParE acoustic feature set: 65 low-level descriptors (LLDs). MFCC: Mel-Frequency Cepstral Coefficient; RASTA: Relative Spectral Transform; HNR: Harmonics to Noise Ratio; RMS: Root Mean Square. Refer to [31] for more details.

55 spectral LLDs	**Group**
MFCC 1–14	Cepstral
Psychoacoustic sharpness, harmonicity	Spectral
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	Spectral
Spectral energy 250–650 Hz, 1 k–4 kHz	spectral
Spectral flux, centroid, entropy, slope	Spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	Spectral
Spectral variance, skewness, kurtosis	spectral
6 voicing related LLDs	**Group**
F_0 (SHS and Viterbi smoothing)	Prosodic
Probability of voicing	Voice quality
log HNR, jitter (local and δ), shimmer (local)	Voice quality
4 energy related LLDs	**Group**
RMS energy, zero-crossing rate	Prosodic
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic

2.4. Support Vector Machines

As a popular classifier, SVMs [23] are chosen as the baseline learning models in our experiments. Both the wavelet features and the ComParE features are fed into SVM models in the format as *functionals*. The original feature values are standardised before the training phase, and the information of the *training* sets are applied to the *test* sets.

2.5. Gated Recurrent Neural Networks

Zöhrer et al. introduced gated recurrent neural networks (GRNNs) for the ASC task in [32]. GRNNs are built with blocks of gated recurrent units (GRUs, see Figure 1) [24], which is a simple alternative to long short term memory networks (LSTMs) [33]. GRNNs and LSTMs share the common characteristic of learning temporal information from the sequence as recurrent neural networks (RNNs) [34]. In particular, GRNNs need fewer parameters than LSTMs, when reaching a comparable performance. Figure 1 shows the flow diagram of one GRU in which the *z*, and *r* are *update*, and *reset* gates, governing the network to learn temporal information from an input sequence. Details on GRNNs can be found, e. g., in [24].

When feeding features to GRNNs, our first step is to segment the audio file (of 10 seconds) into *episodes* (of 1 second) sequenced by time steps of 0.5 seconds. Then, the features are extracted from the *episodes* in the format of *functionals*. Finally, features (standardised) are fed into GRNNs as the same index in the sequence of *episodes*.

2.6. Decision Fusion

To improve the efficiency and robustness of our proposed systems, we use a decision fusion process (see Figure 2) for combining mod-

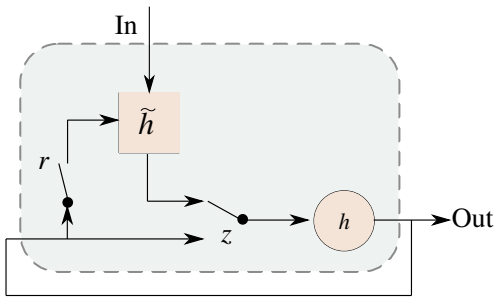

Figure 1: Diagram of a Grated Recurrent Unit (refer to [24] for more details).

els trained independently on varied feature sets. A *margin sampling value* (MSV) is defined as the difference between the first and second largest *posteriori probabilities* estimated by the trained classifier for the given test sample [25]. The final label will be given by the model which has the maximum *MSV* for the given test sample, which means this model is more 'confident' than others when making this decision.

3. EXPERIMENTS

3.1. Setup

The SVM models are implemented by the toolkit LIBSVM [35]. We select the SVMs with a *linear kernel*, and the complexity value *C* is optimised by searching within the grids spanned by $10^{-5}, 10^{-4}, 10^{3}, \ldots, 10^{3}, 10^{4}, 10^{5}$. The GRNNs models are implemented by TensorFlow [3]. We use a two-layer (120-60) GRNNs structure, and empirically set the *learning rate*, the *drop out rate*, and the *epoch* as 0.0002, 0.1, and 50 respectively. The LLDs are extracted from the frame-level of the audio signals within a frame length of 40 ms, and an overlap of 20 ms as the set in the official baseline [1]. We combine predictions given independently using different features; the final decision is made by considering the *margin sampling* strategy mentioned in Section 2.6.

3.2. Results

We can see from both Tables 3 and 4 that, wavelet features (WPTE, WEF) are comparable to ComParE features for the ASC task in this study. Among wavelets, WEFs perform slightly better than WPTE (77.8 % vs 75.7 % on SVMs, and 76.0 % vs 72.6 % on GRNNs). Furthermore, by combining the models trained by wavelets, the final performance can be improved. In particular, by a late fusion of ComParE features with wavelet features, both models (SVMs and GRNNs) can reach an averaged accuracy of more than 81.0 %, which significantly ($p < 0.05$, one-tailed z-test [36]) outperforms the best performance (77.9 %) achieved by the model (SVMs) trained with a single feature set (ComParE). SVMs lead to a slightly better performance than GRNNs considering the overall best performance (83.2 % vs 82.6 %). Both methods considerably outperform the official baseline (74.8 %) at a significance level of $p < 0.001$ in a one-tailed z-test.

Table 5, and Table 6 show the confusion matrices of the two best-performing systems using SVMs, and GRNNs respectively. It

[3] https://www.tensorflow.org/

Table 3: Performance comparison obtained by different feature sets applied to the original segments (of 10 seconds). Classifier: Support Vector Machines (SVMs) with *linear kernel*. *C*-value is set to 0.01, 10 and 0.1 for ComParE, WPTE, and WEF, respectively. All the models are trained independently, and combined to make the final decision by *margin sampling values* generated by each model.

accuracy [%]	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Mean**
ComParE	76.8	76.8	75.7	82.5	77.9
WPTE	76.1	75.9	72.8	78.3	75.7
WEF	79.9	79.0	75.2	77.1	77.8
ComParE+WPTE	80.6	82.3	79.9	85.5	82.1
ComParE+WEF	82.3	83.9	81.7	83.7	82.9
WPTE+WEF	80.1	79.8	76.4	80.0	79.1
ComParE+WPTE+WEF	82.4	83.9	81.7	84.7	**83.2**

Table 4: Performance comparison between different feature sets (*Sequentially Learnt*). Classifier: Gated Recurrent Neural Networks (GRNNs). The GRNNs are structured as two-layer (120-60) topology, *learning rate*: 0.0002, *drop out rate*: 0.1, *epoch*: 50 for both ComParE, WPTE, and WEF. All models are trained independently, and combined to make the final decision by *margin sampling values* generated by each model.

accuracy [%]	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Mean**
ComParE	79.3	74.8	77.0	81.0	78.0
WPTE	73.6	71.8	71.1	74.1	72.6
WEF	77.7	76.6	73.1	76.8	76.0
ComParE+WPTE	82.1	79.0	80.1	84.8	81.5
ComParE+WEF	83.2	81.2	81.3	84.7	**82.6**
WPTE+WEF	78.5	77.2	74.3	77.6	76.9
ComParE+WPTE+WEF	82.6	81.8	81.0	85.0	**82.6**

is common for both SVMs and GRNNs that, some acoustic scenes like *office* and *metro station* are recognised with a high accuracy, while others like *park*, *residential area*, and *train* are difficult to be distinguished. The SVMs considerably outperform the GRNNs in classifying *city centre*, *park*, *residential area*, and*train* while the GRNNs perform better at classifying *beach* than the SVMs.

Note that, in our experiments, we intended to combine the best two models, i. e., SVMs and GRNNs trained with wavelets and ComParE features. However, the result (73.1 %) is below the achieved best performances – in fact even lower than the official baseline. One of the possible future directions is to find an efficient way to fuse the various well-trained models preserving their strengths more efficiently.

Overall, the proposed wavelet features can help improve the final recognition performance of the trained models. Both the two learning models (SVMs and GRNNs) were found to be efficient on the ASC task.

4. CONCLUSION

We found our proposed wavelet features can perform well, and help to improve the performance of typical acoustic features in classification of different acoustic scenes. Popular SVMs, and the state-of-the-art GRNNs were compared as learning algorithms. In this work, SVMs slightly outperformed GRNNs by measuring the best averaged accuracy of 4-fold cross validation on the DCASE 2017 development set (83.2 % vs 82.6 %). Both the best models significantly outperformed the official baseline (an averaged accuracy of

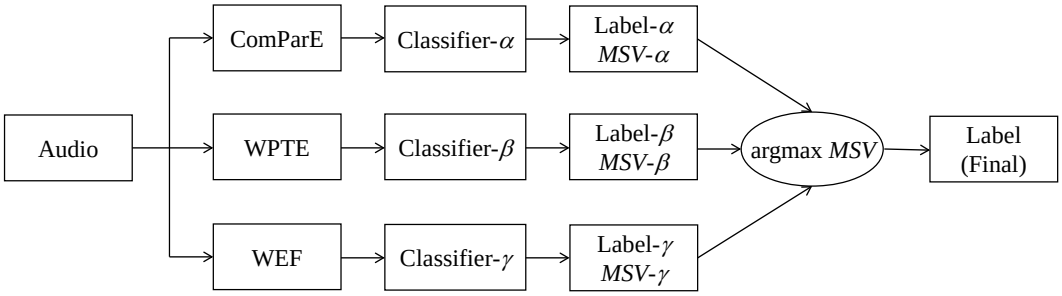

Figure 2: The diagram of a decision fusion process. The input audio files fed to SVMs, and GRNNs are original segments (of 10 seconds), and episodes (of 1 second) rowed as a sequence with time step of 0.5 second, respectively.

Table 5: Confusion matrix of the development set by decision fusion of SVMs models trained independently on the ComParE, WPTE, and WEF feature sets. Values are averaged by 4-fold cross validation on development sets.

Predicted ->	*beach*	*bus*	*cafe/rest.*	*car*	*city cent.*	*forest path*	*groc. store*	*home*	*library*	*metro st.*	*office*	*park*	*resid. area*	*train*	*tram*
beach	**61**	0	0	1	4	2	1	1	0	0	0	3	6	0	1
bus	0	**71**	0	3	0	0	0	0	0	0	0	1	0	1	3
cafe/rest.	0	0	**63**	0	0	0	3	5	1	2	2	1	1	1	1
car	0	1	0	**71**	0	0	0	0	0	0	0	1	0	2	5
city cent.	0	0	1	0	**73**	0	0	0	0	0	0	1	3	0	0
forest path	1	0	1	0	0	**66**	0	2	0	1	2	1	5	1	0
groc. store	1	0	3	0	0	0	**65**	3	0	6	0	0	1	0	0
home	2	0	1	0	0	2	0	**57**	14	1	4	0	0	0	0
library	1	0	0	0	0	1	0	5	**61**	2	3	0	3	3	0
metro st.	0	0	1	0	0	0	2	0	2	**73**	1	0	0	0	0
office	0	0	0	0	0	1	0	3	0	0	**75**	0	0	0	0
park	2	0	1	0	2	1	0	0	1	0	0	**57**	13	1	0
resid. area	3	0	1	0	2	2	0	0	0	0	0	11	**59**	0	0
train	2	3	3	1	3	0	1	0	1	1	0	0	0	**57**	8
tram	2	1	1	1	0	0	1	0	0	1	0	0	0	5	**67**

Table 6: Confusion matrix of the development set by decision fusion of GRNN models trained independently on ComParE, WPTE, and WEF feature sets. Values are averaged by 4-fold cross validation on development sets.

Predicted ->	*beach*	*bus*	*cafe/rest.*	*car*	*city cent.*	*forest path*	*groc. store*	*home*	*library*	*metro st.*	*office*	*park*	*resid. area*	*train*	*tram*
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**

74.8 %, $p < 0.001$, one-tailed z-test). Some acoustic scenes, e. g., *park*, *residential area*, and *train* are difficult to be distinguished in our study. In future works, we will focus on feature selection and enhancement.

5. ACKNOWLEDGMENT

This work was partially supported by the China Scholarship Council (CSC), the European Union's Seventh Framework under grant agreements No. 338164 (ERC StG iHEARu), and the EU's Horizon 2020 Programme through the Innovation Action No. 645094 (SEWA).

6. 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. DCASE Workshop*, Munich, Germany, 2017, in press.

[2] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[3] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini, and B. Schuller, "Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification," in *Proc. DCASE Workshop*, Budapest, Hungary, 2016, pp. 65–69.

[4] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 1151–1154.

[5] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[6] D. Stowell and D. Clayton, "Acoustic event detection for multiple overlapping similar sources," in *Proc. WASPAA*, New Paltz, NY, US, 2015, no pagination, 5 pages.

[7] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering*, vol. 6, no. 1, pp. 40–50, 2012.

[8] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *Proc. WASPAA*, New Paltz, NY, US, 2013, no pagination, 4 pages.

[9] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. EUSIPCO*, Budapest, Hungary, 2016, pp. 1128–1132.

[10] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[11] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[12] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio context recognition using audio event histograms," in *Proc. EUSIPCO*, Aalborg, Denmark, 2010, pp. 1272–1276.

[13] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 142–153, 2015.

[14] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experiments on the DCASE Challenge 2016: Acoustic scene classification and sound event detection in real life recording," in *Proc. DCASE Workshop*, Budapest, Hungary, 2016, pp. 20–24.

[15] Q. Kong, I. Sobieraj, W. Wang, and M. Plumbley, "Deep neural network baseline for DCASE Challenge 2016," in *Proc. DCASE Workshop*, Budapest, Hungary, 2016, pp. 50–54.

[16] Y. Xu, Q. Huang, W. Wang, and M. D. Plumbley, "Hierarchical learning for DNN-based acoustic scene classification," in *Proc. DCASE Workshop*, Budapest, Hungary, 2016, pp. 105–109.

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

[18] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. DCASE Workshop*, Budapest, Hungary, 2016, pp. 11–15.

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

[20] 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. EMBC*, Jeju, Korea, 2017, pp. 3737–3740.

[21] 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.

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

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[24] 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, 9 pages.

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

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

[27] R. R. Coifman, Y. Meyer, and V. Wickerhauser, "Wavelet analysis and signal processing," in *Wavelets and their Applications*. Sudbury: MA: Jones and Barlett, 1992, pp. 153–178.

[28] S. Mallat, *A wavelet tour of signal processing: the sparse way*. Burlington, MA, USA: Elsevier, 2009.

[29] R. N. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 121–131, 2011.

[30] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 148–152.

[31] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*. Switzerland: Springer International Publishing, 2015.

[32] M. Zöhrer and F. Pernkopf, "Gated recurrent networks applied to acoustic scene classification," in *Proc. DCASE Workshop*, Budapest, Hungary, 2016, pp. 115–119.

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

[34] D. P. Mandic and J. A. Chambers, *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. West Sussex, PO19 1UD, England: John Wiley & Sons, Ltd, 2001.

[35] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[36] M. R. Spiegel, J. J. Schiller, R. A. Srinivasan, and M. LeVan, *Probability and Statistics*. New York, NY, USA: McGraw-Hill, 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