PAIRWISE DECOMPOSITION WITH DEEP NEURAL NETWORKS AND MULTISCALE KERNEL SUBSPACE LEARNING FOR ACOUSTIC SCENE CLASSIFICATION

Erik Marchi[1,3]*, Dario Tonelli*[2]*, Xinzhou Xu*[1]*, Fabien Ringeval*[1,3]*, Jun Deng*[1]*, Stefano Squartini*[2]*, Björn Schuller*[1,3,4]

[1] University of Passau, Chair of Complex and Intelligent Systems, Germany
[2]A3LAB, Department of Information Engineering, Universitá Politecnica delle Marche, Italy
[3]audEERING GmbH, Gilching, Germany
[4]Imperial College London, Department of Computing, London, United Kingdom
erik.marchi@tum.de

ABSTRACT

We propose a system for acoustic scene classification using pairwise decomposition with deep neural networks and dimensionality reduction by multiscale kernel subspace learning. It is our contribution to the Acoustic Scene Classification task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2016). The system classifies 15 different acoustic scenes. First, auditory spectral features are extracted and fed into 15 binary deep multilayer perceptron neural networks (MLP). MLP are trained with the 'one-against-all' paradigm to perform a pairwise decomposition. In a second stage, a large number of spectral, cepstral, energy and voicing-related audio features are extracted. Multiscale Gaussian kernels are then used in constructing optimal linear combination of Gram matrices for multiple kernel subspace learning. The reduced feature set is fed into a nearest-neighbour classifier. Predictions from the two systems are then combined by a threshold-based decision function. On the official development set of the challenge, an accuracy of 81.4% is achieved.

Index Terms— Computational Acoustic Scene Analysis, Acoustic Scene Classification, Multilayer Perceptron, Deep Neural Networks, Multiscale Kernel Analysis

1. INTRODUCTION

Acoustic scene classisfication aims at recognising the acoustic background and goes under the field of Computational Auditory Scene Analysis (CASA) [1]. Acoustic scene analysis is a challenging task since a plethora of different overlapping sound sources are composing the acoustic mark of a certain scene, making it a complex combination of various acoustic events.

In the past years, we observed an increasing interest on intelligent audio-based systems able to recognise an environment around a device [2]. This has stimulated the research community to find more robust and efficient methods ranging from unsupervised approaches such as acoustic novelty detection [3, 4] to supervised approaches such as acoustic scene classification and sound even detection [5].

Several works on acoustic scene classification applied different spectral, energy and voicing-related features, in conjunction with neural networks [6]. A system for acoustic scene recognition is described and evaluated in [7]. That system uses several audio features and a nearest neighbour (NN) classifier. In [8], a system for acoustic scene classification is described. The approach relies on Support Vector Machines (SVM), embedded in a hierarchical or parallel framework. In [9], the detection and classification of acoustic events is evaluated by providing a testbed. In [10], it was shown how, in the case of small amounts of training data, new acoustic events can be *learned* by a system. In [11], large-scale acoustic features are used in combination with SVM for the task of acoustic scene analysis.

Acoustic scene classification is applicable in several fields such as intelligent user interfaces [12], serious games [13], automotive [14], and street routing [15], where the context can be recognised using acoustic scene classification techniques.

In the scene classification task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2016), systems for acoustic scene recognition are compared. The provided corpus is divided into a development set and a non-public evaluation set. The dataset is categorised into 15 different classes of acoustic scenes.

This contribution describes our investigated method for acoustic scene classification. From the recordings, auditory spectral features and a large number of spectral, cepstral, energy and voicing-related audio features are extracted. Fifteen binary deep MLP neural networks are trained in a 'one-against-all' fashion to perform a pairwise decomposition instead of simply training a multi-class neural network. In a second stage, multiscale Gaussian kernels are used for multiple kernel suspace learning in order to decrease the dimensionality of the feature space. The reduced feature set is fed into a nearest-neighbour classifier. Finally, the predictions from the two system are then combined with a threshold-based decision function. To our best knowledge, little research focuses on multikernel subspace learning for CASA. Thus, we aim at filling this white spot in the litterature in order to verify if this method can significantly improve the generalisation abilities of a system for acoustic scene classification.

On the official development set of the challenge, an accuracy of 81.4 % is achieved. The employed database, audio features and classification methods are described in Section 2. Experimental results are presented in Section 3, and conclusions are given in Section 4.

2. METHODOLOGY

2.1. Database

For evaluation of our system, we employ the official dataset of the IEEE AASP Challenge on Detection and Classification of Acoustic

Scenes and Events [5]. Thereby, we use only the data of the scene classification task. This dataset contains 30 s recordings of various acoustic scenes, categorised into fifteen different classes. For each of the fifteen classes, the database contains 39 minutes of recordings in the development set, summing up to 9 hours and 45 minutes total duration of the development set. In addition, for the challenge, the systems were evaluated with a non-public test set containing similar data. Sounds were recorded with a high-quality binaural recording system, whereby the portability and subtlety of the system allowed to obtain unobstructed everyday recordings with relative ease. Since the recordings were performed with binaural microphones on the ears of a person, the head-related transfer function (HRTF) of that person is intrinsically incorporated.

2.2. Acoustic features

Auditory Spectral Features (ASF) [16, 17] are computed by applying the Short Time Fourier Transformation (STFT) using a frame size of 40 ms and a frame step of 20 ms. Each STFT yields the power spectrogram which is converted to the Mel-Frequency scale using a filter-bank with 26 triangular filters obtaining the Mel spectrograms $M_{40}(n, m)$. Finally, to match the human perception of loudness, a logarithmic representation is chosen:

$$M_{log}^{40}(n, m) = log(M_{40}(n, m) + 1.0). \quad (1)$$

In addition, the positive first order differences $D_{40}(n, m)$ are calculated from each Mel spectrogram as follows:

$$D_{40}(n, m) = M_{log}^{40}(n, m) - M_{log}^{40}(n - 10, m), \quad (2)$$

with n being the frame index, and k the frequency bin index. Furthermore, the frame energy and the log frame energy are also included as a feature leading to a total number of 56 features. The features are extracted with our open-source audio feature extractor openSMILE [18].

We separately consider the feature sets of the 'emobase' configuration [19]. These features are obtained by extracting the following Low-Level Descriptors (LLDs): intensity, loudness, 12 MFCC, fundamental frequency (F0), probability of voicing, F0 envelope, 8 line spectral frequencies, zero-crossing rate. Statistical functionals are then applied to the LLDs and their first order differences. The following functionals have been used: max./min. value and respective relative position within input, range, arithmetic mean, two linear regression coefficients, and linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1–3, and 3 inter-quartile ranges resulting in a total of 988 features.

2.3. Pairwise Decomposition

Multi-class neural learning [20] can be implemented via several paradigms. One of those is the so called 'one-against-all' paradigm. It consists in decomposing an N-class pattern recognition problem into a system of $L > 1$ neural networks. The L neural networks are trained using a given data set with the assumption of using different class lables. A decision function is usually applied to fuse the results of L neural networks and provide the final system prediction. The 'one-against-all' modelling paradigm employs an ensamble of $L = N$ binary neural networks, $ANN_i, i = 1, ..., N$, each with one unit output layer Y_i with output function f_i (usually sigmoid function) that provides $f_i(\bar{x}) = 1$ or 0 whether the input vector $\bar{x}$ belongs to class i or does not belong to class i. In order to train the i-th neural network ANN_i, the training set S_{tr} is relabelled in two sets, $S_{tr} = {S_{tr}}^i \cup \bar{S}_{tr}^i$, where ${S_{tr}}^i$ consists of all the reference patterns belonging to class i (labelled as '1'), and $\bar{S}_{tr}^i$ consists of all the reference patterns belonging to remaining other classes (labelled as '0'). The decision module in this paradigm should be designed to face the following three output scenarios: The first scenario consists in obtaining $f_i = 1$, and $f_j = 0$ for all j given $i \neq j$. The decision function D can be easily implemented as $D(\bar{x}, f_1, f_2, ..., f_L) = argmax_{i=1,...,L}(f_i)$. The second scenario consists in obtaining all $f_i = 0$ for $i = 1, ..., L$. The third one consists in having more than one neural networks output '1'. In both last scenarios the system is uncertain and the decision function outputs the class label that corresponds to the neural network that shows the largest output value by the activation function at the output layer unit:

$$D(\bar{x}, y_1, y_2, ..., y_L) = argmax_{i=1,...,L}(y_i), \quad (3)$$

where y_i is the output of the activation function used in the output layer of the i-th neural network.

Since we used fully connected MLP feed-forward neural networks, we will refer to this approach as pairwise decomposition with MLP (PDMLP).

In our final system we applied an enhanced decision function that relies on an auxiliary system when the PDMLP is uncertain. The adopted decision function is described in Section 2.5.

2.4. Auxiliary Systems

A system of N binary neural networks trained with 'one-against-all' is indeed a more flexible system and allows for a better discrimination of one class. However, it has one major drawback, the system decision borders generated by the N binary neural networks are suffering from overlapping or uncovering regions in a feature space. In order to mitigate this drawback we introduced some auxiliary systems when the N binary neural networks are providing uncertain outputs. We applied three different auxiliary system: selected 'one-against-one' (OAO) classification, SVM, and Multiple kernel leaning.

2.4.1. Selected 'one-against-one'

A first auxiliary system decomposes an N-class pattern classification problem into $N(N-1)/2$ two-class classification problems using the OAO paradigm. Let's define the $N(N-1)/2$ two-class neural networks as $ANN_k(i, j)$, with $1 \leq k \leq L = N(N-1)/2$. An $ANN_k(i, j)$ represents a neural network trained to discriminate class i from class j , for $1 \leq i < j \leq N$. An $ANN_k(i, j)$ is trained with reference patterns of class i and j, and its output, $f_k(i, j)$, is indicating whether the input pattern $\bar{x}$ is either class i or j. In our case, we refer to selected OAO (sOAO) since we just use as an auxiliary system the binary classifier $A\bar{N}N_k(i, j)$ where i and j are the two more likely classes resulting from the output of the PDMLP in Section 2.3.

2.4.2. Support Vector Machines

As a second auxiliary system we applied the traditional SVM approach trained on the high dimensional feature set 'emobase'.

SVMs have shown to achieve good performances for the task of acoustic scene analysis [10, 11], and are used in this contribution as a comparison to a state-of-the-art method.

2.4.3. Multiple Kernel Learning

The third auxiliary system is based on MultiScale-Kernel Fisher Discriminant Analysis (MSKFDA). This methods was recently

proven to be effective in solving emotion recognition in speech [21]. To our best knowledge, little research focuses on multiscale representation in CASA and we believe that this method for subspace learning may significantly improve the generalisation of the system by learning more robust features. This method benefits from alternatively optimising two variables, namely the kernelised mapping directions and a nonnegative linear combination for kernels with different scaling parameters.

The research of MSKFDA provides the possibility of solving multiscale analysis of acoustic scene factors. For Gaussian kernels, it is easy to draw the multiscale case by regulating scaling parameters. The kernel transforming between samples x_i and x is shown in Eq. (4), with the parameters $\sigma_m > 0$, $m = 1, 2, \ldots, M$ and $i = 1, 2, \ldots, N$:

$$(\Omega_{xi})_m = \phi_m^T(x_i)\phi_m(x) = e^{-\frac{(x_i - x)^2}{\sigma_m^2}}, \quad (4)$$

where Ω_{xi} is the multiple kernel coordinate matrix, and $\phi_m(x)$ is the high dimensional form of x. Kernel methods are originally represented as high-dimensional space by adopting inner product forms in RKHS. However, it can be also assumed that kernel methods bring a dimension-limited feature transformation in graph embedding. This transformation constructs a new feature space for each sample by kernel functions and training samples. Thus, the relationship between a given sample and each training sample leads to the new features. Then, the scales of kernels are mainly determined by the parameters of respective kernels.

As is shown in Figure 1, for sample x, the original features x are transformed into new features $\Omega_x\beta$ by linearly combining multiscale kernels, where $\beta \in \Re^{M\times 1}$ is the column vector with corresponding elements $\beta_m \geq 0$ for kernel m. Then, for the new features of x, the dimensionality-reduced sample can be achieved by using $A^T\Omega_x\beta$ in bilateral ways, where A contains the kernel mappings.

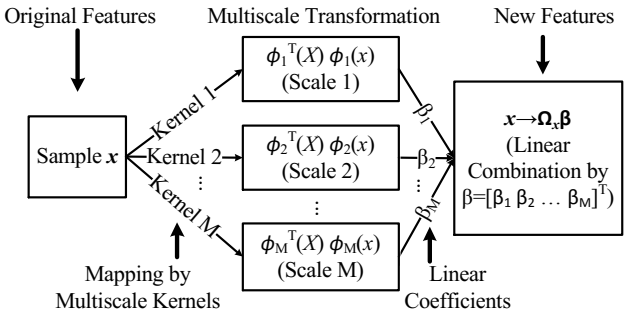

Figure 1: Schematic diagram of learning multiscale kernels. The original features x are transformed into new features $\Omega_x\beta$ by linearly combining multiscale kernels.

High-dimensional acoustic features inevitably include much interference resulting from the factors of background environment sound, speakers, etc., in spite of state-of-the-art feature acquisition ways. Therefore, CASA systems would benefit from the suggested novel feature reduction method in combination with the embedding graphs of FDA and multiple kernel learning. In addition, few parameters need to be regulated in FDA. For these reasons, we utilise MSKFDA as an auxiliary tool in order to obtain better performance as second stage of our algorithm.

2.5. Decision Functions

In our final system, the decision function is obtained by first applying a threshold to the output activations of a PDMLP. If the number of outputs above the threshold is 1, then the predicted class is the one corresponding to the neural network that generated that output. Otherwise, if more outputs are above the threshold or none of the outputs are above the threshold we only rely on the auxiliary system predicted class. The value of threshold is set to 0.3 and was optimised on the development set.

3. EXPERIMENTS

This section contains the experimental setup and the evaluation of different approaches on the development set of DCASE.

3.1. Setup

In the fifteen binary classifiers composing the PDMLP system, MLP were trained on 100 parallel sequences per batch, using Stochastic Gradient Descent with Adam by applying a fixed learning rate of 0.001 and a binary cross-entropy objective function. We used rectified linear units as activation function. Weights were initialized with Gaussian normal distribution (σ = 0.1, μ = 0). For better generalization, the networks were trained using early stopping on the corresponding test set per each fold. Furthermore, the early stopping criterion was applied considering the sum of all validation errors at each epoch of each network. In this way, we first reduced the training time by a factor of 3 and we also avoided potential overfitting. The training procedure stopped after a maximum number of 1000 epochs. The 'selected OAO' auxiliary networks were trained in the same fashion. In order to compare the proposed approaches with state-of-the-art methods, we also evaluated traditional multi-class MLP with exactly the same training algorithm and parameters, except that we used a multi-class cross entropy error as objective function. All networks were trained using Theano [22] and Lasagne[1]. We also evaluated SVM with a linear kernel and complexity value $C = 0.001, 0.01, 0.1, 1.0, 10.0$. SVM are trained with the sequential minimal optimisation (SMO) algorithm using the training data. The parameters in the MSKFDA are set as follows. The number of scales is set as $M = 21$, with the Gaussian scaling parameters σ_m ($m = 1, 2, \ldots, M$) selected as $0.0001n$, $0.0003n$, $0.0005n$, $0.0007n$, $0.001n$, $0.003n$, $0.005n$, $0.007n$, $0.01n$, $0.03n$, $0.05n$, $0.07n$, $0.1n$, $0.3n$, $0.5n$, $0.7n$, n, $3n$, $5n$, $7n$, and $10n$, respectively, where n is the number of original features. We employ openSMILE's '*emobase*' feature set which results in $n = 988$ here. The dimensions d of the dimensionality-reduced feature space are selected no larger than 21. The number of iterations is set as 7. A Nearest-Neighbour classifier is selected as the final decision maker.

3.2. Results

We first tested traditional multi-class approaches by using SVM and MLP in order to compare the performance of PDMLP with state-of-the-art methods. Table 1 reports performances on the development set using the 4-fold cross validation as specified in the challenge baseline. By applying PDMLP, we can observe an absolute improvement of 7% accuracy over the baseline of the DCASE challenge [5]. SVMs perform slightly better than the baseline with up to 74% accuracy (best performance obtained with $C = 0.1$).

[1] https://lasagne.readthedocs.io

Table 1: Comparison of performances between traditional multi-class systems and the proposed method with 15 binary classifiers (PDMLP). Multi-class classifiers: Gaussian Mixture Models (Baseline), Support Vector Machines (SVM), and Multi Layer Perceptron (MLP). For neural networks, the layout is indicated in parenthesis (*number of units* × *number of layers*). Results are given in terms of accuracy [%].

Method	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Mean**
Baseline [5]	67.2	68.9	72.2	81.9	72.5
SVM	72.2	73.9	77.1	72.4	74.0
MSKFDA	76.8	73.7	79.5	79.7	77.5
MLP (*54x3*)	78.5	70.8	77.7	75.9	75.9
MLP (*256x3*)	78.6	77.7	76.2	77.1	77.4
PDMLP (*54x3*)	82.6	76.9	77.5	77.0	78.5
PDMLP (*256x3*)	81.4	78.2	77.5	80.8	**79.5**

Table 2: Combination of the PDMLP system with different auxiliary systems. Best accuracy (%) obtained on the development set with 4-fold cross validation. Auxiliary systems: Support Vector Machines (SVM), selected 'one-against-one' classifier (sOAO), and MultiScale-Kernel Fisher Discriminant Analysis (MSKFDA) with nearest-neighbours. Results are given in terms of accuracy [%].

	Folds				
Method	1	2	3	4	**Mean**
PDMLP	81.4	78.2	77.5	80.8	79.5
PDMLP-SVM	79.4	75.8	78.6	80.0	78.4
PDMLP-sOAO	81.9	80.4	75.8	81.2	80.8
PDMLP-MSKFDA	81.5	79.8	81.5	82.9	**81.4**

MSKFDA performs significantly better than SVM with an auccuracy of 77.5%, corroborating our assumption that multikernel subspace learning is effective for acoustic scene classification. Multi-class MLP are evaluated using different layouts. For simplicity, we only report the best results obtained with three hidden layers composed by 54 units and 256 units. We can observe that, increasing the dimension of the hidden layer to 256 units brings better performances up to 77.4% accuracy, however, no more improvement was observed by further augmenting the dimensionality of the hidden layer. The same layout (256-256-256) also brought about an increase in performance in the PDMLP method of up to 79.5% accuracy. We kept this layout for the PDMLP as final first stage system and applied the auxiliary systems by adopting the enhanced decision function described in Section 2.5.

Table 2 shows the results obtained from the fusion with SVM, sOAO and MSKFDA. We observe that the combination with SVM is not fruitful and led to a decrease in performance down to 78.4% accuracy. However, combining PDMLP and sOAO seems to increase performances up to 80.8% with an absolute improvement of 1.3% accuracy. Further improvement is observed with the combination of PDMLP and MSKFDA up to 81.4% with an absolute improvement of 8.9% accuracy over the challenge baseline.

Table 3 shows the confusion matrix for the best-performing system, using PDMLP and MSKFDA. Some classes (*office*, *car*) are recognised with high accuracy, while for others (*park, restaurant, train*), low scores are obtained. Most confusions are made between the classes *park* and *residential area* or *city* and *train*. The recordings of the classes *park* and *residential area* are partly very similar.

Summing up, we believe that such a system is more robust to variation since it relies on two generalised system. In fact, PDMLP can be considered already a very flexible system given that it was tailored to discriminate one class against the rest. Additionally, we carefully trained the 15 MLP considering the overall validation error, avoiding individual training and subsequent overfitting. Furthermore, by selecting MSKFDA as auxiliary system we relied on another well-generalised model trained on a reduced and rubust feature set obtained via multi kernel subspace learning.

Table 3: Confusion Matrix of the development data for the proposed system, achieving an accuracy of 81.4 %.

	beach	*bus*	*cafe*	*car*	*city*	*forest*	*grocery*	*home*	*library*	*metro*	*office*	*park*	*resid.*	*train*	*tram*
beach	**62**	0	0	0	4	0	0	0	0	0	0	7	4	0	1
bus	0	**68**	0	6	0	0	0	0	0	0	0	0	0	2	2
cafe/rest.	0	0	**59**	0	0	0	6	2	0	7	0	1	0	0	3
car	0	4	0	**71**	0	0	0	0	0	0	0	0	0	3	0
city	0	0	0	0	**74**	0	1	0	0	0	0	0	3	0	0
forest	1	0	0	0	0	**73**	0	1	0	0	0	1	2	0	0
grocery	2	0	9	0	0	0	**56**	0	0	11	0	0	0	0	0
home	5	0	1	0	0	1	0	**67**	2	0	1	1	0	0	0
library	0	0	3	0	0	0	2	0	**71**	0	0	0	0	2	0
metro st.	0	0	0	0	0	0	3	5	0	**70**	0	0	0	0	0
office	0	0	0	0	0	0	0	0	5	0	**73**	0	0	0	0
park	4	0	0	0	2	1	0	0	4	0	0	**47**	20	0	0
res. area	2	0	0	0	1	4	0	0	1	0	1	15	**54**	0	0
train	0	9	6	0	15	0	1	0	1	0	0	0	0	**39**	7
tram	0	1	0	0	2	0	5	0	0	0	0	0	0	1	**69**

4. CONCLUSIONS

We presented and evaluated a system for acoustic scene classification. Combining pairwise decomposition with deep neural networks and dimensionality reduction by multiscale kernels, an accuracy of 81.4 % is obtained on the development set of the D-CASE challenge. A comparison with stat-of-the-art approaches showed that the pairwise decomposition can alone bring a significant (one-tailed z-test [23], $p<0.001$) improvement. Furthermore, we found that a dimensionality reduction via multiple kernel learning is also effective and outperforms the baseline significantly. The two methods seems to be complementary and thus – if combined – they provide a more robust system. Some acoustic scenes (*park, restaurant, train, city*) are difficult to recognise due to the high variability in the class and the similarity between the different classes. In future works, we will focus on new acoustic features and enhanced decision functions for the late fusion stage.

5. ACKNOWLEDGMENT

The research leading to these results has received funding from the EU's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 338164 (ERC Starting Grant iHEARu), and the EU's Horizon 2020 Programme agreements No. 645378 (RIA ARIA VALUSPA), and No. 688835 (RIA DE-ENIGMA).

6. REFERENCES

[1] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley interscience, 2006.

[2] 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 Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, Jan 2006.

[3] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A Novel Approach for Automatic Acoustic Novelty Detection Using a Denoising Autoencoder with Bidirectional LSTM Neural Networks," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. Brisbane, Australia: IEEE, Apr 2015, p. 5.

[4] E. Marchi, F. Vesperini, F. Weninger, F. Eyben, S. Squartini, and B. Schuller, "Non-Linear Prediction with LSTM Recurrent Neural Networks for Acoustic Novelty Detection," in *Proc. 2015 Int. Joint Conference on Neural Networks, IJCNN*, IEEE. Killarney, Ireland: IEEE, Jul 2015, p. 5.

[5] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *Proc. 24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, 2016.

[6] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 20, no. 1-2, pp. 61–79, 1998.

[7] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, FL, USA, 2002.

[8] H. Jiang, J. Bai, S. Zhang, and B. Xu, "Svm-based audio scene classification," in *Proc. Natural Language Processing and Knowledge Engineering (NLP-KE)*. IEEE, 2005, pp. 131–136.

[9] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans*. Springer, 2007, pp. 311–322.

[10] J. T. Geiger, M. A. Lakhal, B. Schuller, and G. Rigoll, "Learning new acoustic events in an hmm-based system using map adaptation," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 293–296.

[11] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-Scale Audio Feature Extraction and SVM for Acoustic Scene Classification," in *Proc. of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2013*, IEEE. New Paltz, NY: IEEE, October 2013, pp. 1–4.

[12] N. Sabouret, L. Paletta, B. Schuller, E. Marchi, H. Jones, and A. B. Youssef, "Intelligent User Interfaces in Digital Games for Empowerment and Inclusion," in *Proc. of the 12th Int. Conference on Advancement in Computer Entertainment Technology, ACE 2015*, ACM. Iskandar, Malaysia: ACM, November 2015, p. 8.

[13] B. Schuller, E. Marchi, S. Baron-Cohen, A. Lassalle, H. O'Reilly, D. Pigat, P. Robinson, I. Davies, T. Baltrusaitis, M. Mahmoud, O. Golan, S. Friedenson, S. Tal, S. Newman, N. Meir, R. Shillo, A. Camurri, S. Piana, A. Staglianò, S. Bölte, D. Lundqvist, S. Berggren, A. Baranger, N. Sullings, M. Sezgin, N. Alyuz, A. Rynkiewicz, K. Ptaszek, and K. Ligmann, "Recent developments and results of ASC-Inclusion: An Integrated Internet-Based Environment for Social Inclusion of Children with Autism Spectrum Conditions," in *Proc. 3rd Int. Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2015) as part of the 20th ACM IUI 2015*, ACM. Atlanta, GA: ACM, March 2015.

[14] I. Abdić, L. Fridman, E. Marchi, D. E. Brown, W. Angell, B. Reimer, and B. Schuller, "Detecting Road Surface Wetness from Audio: A Deep Learning Approach," *arxiv.org*, no. 1511.07035, p. 5, December 2015.

[15] B. Schuller, F. Pokorny, S. Ladsttter, M. Fellner, F. Graf, and L. Paletta, "Acoustic geo-sensing: Recognising cyclists' route, route direction, and route progress from cell-phone audio," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

[16] E. Marchi, G. Ferroni, F. Eyben, L. Gabrielli, S. Squartini, and B. Schuller, "Multi-resolution Linear Prediction Based Features for Audio Onset Detection with Bidirectional LSTM Neural Networks," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. Florence, Italy: IEEE, 2014, pp. 2183–2187.

[17] E. Marchi, G. Ferroni, F. Eyben, S. Squartini, and B. Schuller, "Audio Onset Detection: A Wavelet Packet Based Approach with Recurrent Neural Networks," in *Proc. 2014 Int. Joint Conference on Neural Networks (IJCNN) as part of the IEEE World Congress on Computational Intelligence (IEEE WCCI)*, IEEE. Beijing, China: IEEE, Jul 2014, pp. 3585–3591.

[18] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of the 21st ACM Int. Conference on Multimedia, MM 2013*, ACM. Barcelona, Spain: ACM, Oct 2013, pp. 835–838.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR–introducing the Munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops*. Amsterdam, The Netherlands: IEEE, 2009, pp. 576–581.

[20] G. Ou and Y. L. Murphey, "Multi-class pattern classification using neural networks," *Pattern Recognition*, vol. 40, no. 1, pp. 4 – 18, 2007.

[21] X. Xu, J. Deng, W. Zheng, L. Zhao, and B. Schuller, "Dimensionality reduction for speech emotion features by multiscale kernels," in *Proc. Annual Conference of the Int. Speech Communication Association (INTERSPEECH)*. Dresden, Germany: ISCA, 2015, pp. 1532–1536.

[22] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016.

[23] M. D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Proc. of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, ser. CIKM '07. New York, NY, USA: ACM, 2007, pp. 623–632.