Deep scalogram representations for acoustic scene classification

Zhao Ren, Kun Qian, Yebin Wang, Zixing Zhang, Vedhas Pandit, Alice Baird, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Ren, Zhao, Kun Qian, Yebin Wang, Zixing Zhang, Vedhas Pandit, Alice Baird, and Björn Schuller. 2018. "Deep scalogram representations for acoustic scene classification." *IEEE/CAA Journal of Automatica Sinica* 5 (3): 662–69. https://doi.org/10.1109/jas.2018.7511066.

Deep Scalogram Representations for Acoustic Scene Classification

Zhao Ren, Kun Qian, *Student Member, IEEE,* Zixing Zhang, *Member, IEEE,* Vedhas Pandit, Alice Baird, *Student Member, IEEE,* and Björn Schuller, *Fellow, IEEE*

***Abstract*—Spectrogram representations of acoustic scenes have achieved competitive performance for acoustic scene classification. Yet, the spectrogram alone does not take into account a substantial amount of time-frequency information. In this study, we present an approach for exploring the benefits of deep scalogram representations, extracted in segments from an audio stream. The approach presented firstly transforms the segmented acoustic scenes into bump and morse scalograms, as well as spectrograms; secondly, the spectrograms or scalograms are sent into pre-trained convolutional neural networks; thirdly, the features extracted from a subsequent fully connected layer are fed into (bidirectional) gated recurrent neural networks, which are followed by a single highway layer and a softmax layer; finally, predictions from these three systems are fused by a margin sampling value strategy. We then evaluate the proposed approach using the acoustic scene classification data set of 2017 IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE). On the evaluation set, an accuracy of 64.0% from bidirectional gated recurrent neural networks is obtained when fusing the spectrogram and the bump scalogram, which is an improvement on the 61.0% baseline result provided by the DCASE 2017 organisers. This result shows that extracted bump scalograms are capable of improving the classification accuracy, when fusing with a spectrogram-based system.**

***Index Terms*—Acoustic scene classification (ASC), (bidirectional) gated recurrent neural networks ((B) GRNNs), convolutional neural networks (CNNs), deep scalogram representation, spectrogram representation.**

I. Introduction

ACOUSTIC scene classification (ASC) aims at the identification of the class (such as 'train station', or 'restaurant') of a given acoustic environment. ASC can be a challenging task, since the sounds within certain scenes can have similar qualities, and sound events can overlap one another [1]. Its applications are manifold, such as robot hearing or context-aware human-robot interaction [2].

In recent years, several hand-crafted acoustic features have been investigated for the task of ASC, including frequency, energy, and cepstral features [3]. Despite such year-long efforts, recently, representations automatically extracted from spectrogram images with deep learning methods [4], [5] are shown to perform better than hand-crafted acoustic features when the number of acoustic scene classes is large [6], [7]. Further, compared with a Fourier transformation for obtaining spectrograms, the wavelet transformation has the ability to incorporate multiple scales, and for this reason locally can reach the optimal time-frequency resolution [8] concerning the Heisenberg uncertainty of optimal time and frequency resolution at the same time. Accordingly, wavelet features have already been applied successfully for many acoustic tasks [9]−[13] , but often, the greater effort in calculating a wavelet transformation is considered not worth the extra effort if gains are not outstanding. In the theory of wavelet transformation, the scalogram is the time-frequency representation of the signal by wavelet transformation, where the brightness or the colour can be used to indicate coefficient values at corresponding time-frequency locations. Compared to spectrograms, which offer (only) a fixed time and frequency resolution, a scalogram is better suited for the task of ASC due to its detailed representation of the signal. Hence, a scalogram-based approach is proposed in this work.

We use convolutional neural networks (CNNs) to extract deep features from spectrograms or scalograms, as CNNs have proven to be effective for visual recognition tasks [14], and ultimately, spectrograms and scalograms are images. Several specific CNNs are designed for the ASC task, in which spectrograms are fed as an input [7], [15], [16]. Unfortunately, those approaches are not robust and it can also be time-consuming to design CNN structures manually for each dataset. Using pre-trained CNNs from large scale datasets [17] is a potential way to break this bottleneck. ImageNet[1] is a suited such big database promoting a number of CNNs each year, such as 'AlexNet' [18] and 'VGG' [19]. It seems promising to apply transfer learning [20] through extracting

This work was supported by the German National BMBF IKT2020-Grant (16 SV7213) (EmotAsS), the European-Unions Horizon 2020 Research and Innovation Programme (688835) (DE-ENIGMA), and the China Scholarship Council (CSC).

Citation: Z. Ren, K. Qian, Z. X. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA J. of Autom. Sinica*, vol. 5, no. 3, pp. 662−669, May 2018.

Z. Ren, V. Pandit, and A. Baird are with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany (e-mail: {zhao.ren, vedhas.pandit, alice.baird}@informatik.uni-augsburg.de).

K. Qian is with the Machine Intelligence and Signal Processing Group, Technische Universität München, Germany, and also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany (e-mail: andykun.qian@tum.de).

Z. X. Zhang is with the Group on Language, Audio and Music (GLAM), Imperial College London, UK (e-mail: zixing.zhang@imperial.ac.uk).

B. Schuller is with the Group on Language, Audio and Music (GLAM), Imperial College London, UK, and also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany (e-mail: schuller@ieee.org).

[1] http://www.image-net.org/

features from these pre-trained neural networks for the ASC task — the approach taken in the following.

As to handling of audio besides considering 'images' (the spectrograms and/or scalograms) by pre-trained deep networks, we further aim to respect its nature as a time-series. In this respect, sequential learning performs better for time-series problems than static classifiers such as support vector machines (SVMs) [21] or extreme learning machines (ELMs) [17]. Likewise, hidden Markov models (HMMs) [22], recurrent neural networks (RNNs) [23], and in the more recent years in particular long short-term memory (LSTM) RNNs [24] are proven effective for acoustic tasks [25], [26]. As gated recurrent neural networks (GRNNs) [27] — a reduction in computational complexity over LSTM-RNNs — are shown to perform well in [13], [28], we not only use GRNNs as the classifier rather than LSTM-RNNs, but also extend the classification approach with bidirectional GRNNs (BGRNNs), which are trained forward and then backward within a specific time frame. Likewise, we are able to capture 'forward' and 'backward' temporal contexts, or simply said the whole sequence of interest. Unless moving with the microphone or changes of context, acoustic scenes in the real-world usually prevail for longer amounts of time, however, with potentially highly varying acoustics during such stretches of time. This allows to consider static chunk lengths for ASC, despite modelling these as a time series to preserve the order of events, even though being only interested in the 'larger picture' of the scene than in details of events within that scene. In the data considered in this study based on the dataset of 2017 IEEE AASP Challenge on Detection and Classification of Acoustic Scenes And Events (DCASE), the instances have a (pre-)specified duration (10 s per sample in the [29]).

In this article, we make three main contributions. First, we propose the use of scalogram images to help improve the performance of only a single spectrogram extraction for the ASC task. Second, we extract deep representations from the scalogram images using pre-trained CNNs, which is much faster and more efficient in terms of conservative data requirements than manually designed CNNs. Third, we investigate the performance improvement obtained through the use of (B) GRNNs for classification.

The remainder of this paper is structured as follows: some related work for the ASC task is introduced in Section II; in Section III, we describe the proposed approach, the pipeline of which is shown in Fig. 1; the database description, experimental set up, and results are then presented in Section IV; finally, conclusions are given in Section VI.

II. Related Work

In the following, let us outline point by point related work to the points of interest in this article, namely using spectrogram-type images as network input for audio analysis, using CNNs in a transfer-learning setting, using wavelets rather or in addition to spectral information, and finally the usage of memory-enhanced recurrent topologies for optimal treatment of the audio stream as time series data.

Extracting spectrograms from audio clips is well known for the ASC task [7], [30]. This explains why a lion's share of the existing work using non-time-signal input to deep network architectures and particularly CNNs use spectrograms or derived forms as input. For example, spectrograms were used to extract features by autoencoders in [31]. Predictions were obtained by CNNs from mel spectrograms in [32], [33]. Feeding analysed images from spectrograms into CNNs has also shown success. Two image-type features based on a spectrogram, namely covariance matrix, and a secondary frequency analysis were fed into CNNs for classification in [34].

Further, extracting features from pre-trained CNNs has been widely used in transfer learning. To name but two examples, a pre-trained 'VGGFace' model was applied to extract features from face images and a pre-trained 'VGG' was used to extract features from images in [17]. Further, in [6], deep features of audio waveforms were extracted by a pre-trained 'AlexNet' model [18].

Wavelet features are applied extensively in acoustic signal classification, but in fact, in their history they were broadly used also in other contexts such as for electroencephalogram (EEG), electrooculogram (EOG), and electrocardiogram (ECG) signals [35]. Recent examples particularly in the domain of sound analysis include for example successful application for snore sound classification [10], [11] , besides wavelet transform energy and wavelet packet transform energy having also been proven to be effective in the ASC task [12].

Various types of sequential learning are repeatedly and frequently applied for the ASC task. For example, in [36], experimental results have shown superiority when employing RNNs for classification. There are also some special types of RNNs that have been applied for classification in this context. As an example, LSTM-RNNs were combined with CNNs using early-fusion in [25]. In [37], GRNNs were utilised as the classifier, and achieved a significant improvement using a Gaussian mixture model (GMM).

To sum the above up, while similar methods mostly use spectrograms or mel spectrograms, minimal research has been done about the performance of scalogram representations extracted by pre-trained CNNs on sequential learning for audio analysis. This work does so and is introduced next.

III. Proposed Methodology

A. Audio-to-Image Pre-Processing

In this work, we first seek to extract the time-frequency information which is hidden in the acoustic scenes. Hence, the following three types of representations are used in this study, which is a foundation of the following process.

1) Spectrogram: The spectrogram as a time-frequency representation of the audio signal is generated by a short-time Fourier transform (STFT) [38]. We generate the spectrograms with a Hamming window computing the power spectral density by the dB power scale. We use Hamming windows of size 40 ms with an overlap of 20 ms.

2) 'Bump' Scalogram: The bump scalogram is generated by the bump wavelet [39] transformation, which is defined by

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

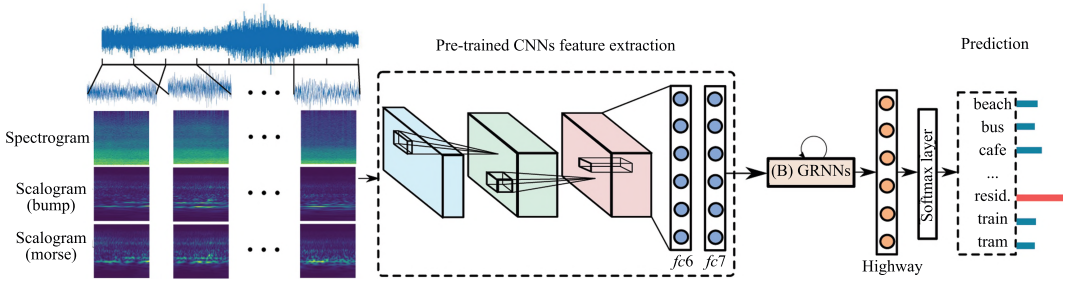

Fig. 1. Framework of the proposed approach. First, spectrograms and scalograms (bump and morse) are generated from segmented audio waveforms. Then, one of these is fed into the pre-trained CNNs, in which further features are extracted at a subsequent fully connected layer $fc7$. Finally, the predictions (predicted labels and probabilities) are obtained by (B) GRNNs with a highway network layer and a softmax layer with the deep features as the input.

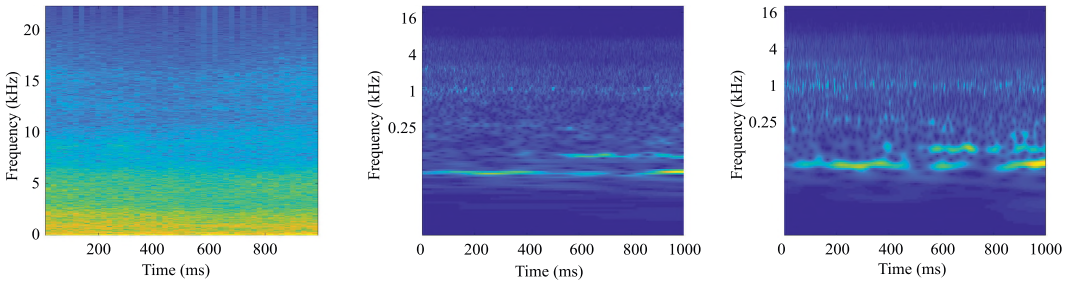

Fig. 2. The spectrogram and two types of scalograms are extracted from the acoustic scenes. All of the images are extracted from the first audio sequence of DCASE2017's 'a001_10_20.wav' with a label 'residential area'.

where s stands for the scale, μ and σ are two constant parameters, in which σ affects the frequency and time localisation, and $\Psi(s\omega)$ is the transformed signal.

3) 'Morse' Scalogram: The morse scalogram [40] generation is defined by

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

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

The three image representations of one instance are shown in Fig. 2. While the STFT focuses on analysing stationary signals and gives a uniform resolution, the wavelet transformation is good at localising transients in non-stationary signals, since it can provide a detailed time-frequency analysis. In our study, the training model is proposed based on the above three representations and comparisons of them are provided in the following sections.

B. Pre-Trained Convolutional Neural Networks

By transfer learning, the pre-trained CNNs are transfered to our ASC task for extracting the deep spectrum features. For the pre-trained CNNs, we choose 'AlexNet' [18], 'VGG16', and 'VGG19' [19], since they have proven to be successful in a large number of natural image classification tasks, including the ImageNet Challenge[2]. 'AlexNet' consists of five convolutional layers with [96, 256, 384, 384, 256] kernels of size [11, 5, 3, 3, 3], and three maxpooling layers. 'VGG' networks have 13 ([2, 2, 3, 3, 3], 'VGG16'), or 16 ([2, 2, 4, 4, 4], 'VGG19') convolutional layers with [64, 128, 128, 256, 256] kernels and five maxpooling layers. All of the convolutional layers in the 'VGG' networks use the common kernel size 'three'. In these three networks, the convolutional and maxpooling layers are followed by three fully connected layers $\{fc6, fc7, fc\}$, and a soft-max layer for 1000 labelled classifications according to the ImageNet challenge, in which $fc7$ is employed to extract deep features with 4096 attributes. More details on the CNNs are given in Table I. We obtain the pre-trained 'AlexNet' network from MATLAB R2017a[3], and 'VGG16' and 'VGG-

TABLE I
CONFIGURATIONS OF THE CONVOLUTIONAL NEURAL NETWORKS. 'ALEXNET', 'VGG16', AND 'VGG19' ARE USED TO EXTRACT DEEP FEATURES OF THE SPECTROGRAM, 'BUMP', AND 'MORSE' SCALOGRAMS. 'CONV' STANDS FOR THE CONVOLUTIONAL LAYER

AlexNet	VGG16	VGG19
input: RGB image		
1×conv11-96	2×conv3-64	2×conv3-64
maxpooling		
1×conv5-256	2×conv3-128	2×conv3-128
maxpooling		
1×conv3-384	3×conv3-256	4×conv3-256
maxpooling		
1×conv3-384	3×conv3-512	4×conv3-512
maxpooling		
1×conv3-256	3×conv3-512	4×conv3-512
maxpooling		
fully connected layer $fc6$-4096		
fully connected layer $fc7$-4096		
fully connected layer fc-1000		
output: soft-max		

[2]http://www.image-net.org/challenges/LSVRC/
[3]https://de.mathworks.com/help/nnet/ref/alexnet.html

19' from MatConvNet [41]. As outlined, we exploit the spectrogram and two types of scalograms as the input for these three CNNs separately and extract the deep representations from the activations on the second fully connected layer $fc7$.

C. (Bidirectional) Gated Recurrent Neural Networks

As a special type of RNNs, GRNNs contain a gated recurrent unit (GRU) [27], which features an update gate u, a reset gate r, an activation h, and a candidate activation $\tilde{h}$. For each ith GRU at a time t, the update gate u and reset gate r activations are defined by

$$u_t^i = \sigma(W_u x_t + U_u h_{t-1})^i \tag{3}$$

$$r_t^i = \sigma(W_r x_t + U_r h_{t-1})^i \tag{4}$$

where σ is a logistic sigmoid function, W_u, W_r, U_u, and U_r are the weight matrices, and h_{t-1} stands for the activation function. At time t, the activation function and candidate activation function are defined by

$$h_t^i = (1 - u_t^i)h_{t-1}^i + u_t^i \tilde{h}_t^i \tag{5}$$

$$\tilde{h}_t^i = \tanh(W x_t + U(r_t \odot h_{t-1}))^i. \tag{6}$$

The information flows inside the GRU with gating units, similarly to, but with separate memory cells in the LSTM. However, there is not an input gate, forget gate, and output gate which are included in the LSTM structure. Rather, there are a reset and an update gate, with overall less parameters in a GRU than in a LSTM unit so that GRNNs usually converge faster than LSTM-RNNs [27]. GRNNs have been observed to be comparable and even better than LSTM-RNNs sometimes in accuracies, as shown in [42]. To gain more time information from the extracted deep feature sequences, bidirectional GRNNs (BGRNNs) are an efficient tool to improve the performance of GRNNs (and in fact of course similarly for LSTM-type RNNs), as shown in [43], [44]. Therefore, BGRNNs are used in this study, in which context interdependences of features are learnt in both temporal directions [45]. For classification, a highway network layer and a softmax layer follow the (B) GRNNs, as highway networks are often found to be more efficient than fully connected layers for very deep neural networks [46].

D. Decision Fusion Strategy

It was found in a recent work that the margin sampling value (MSV) [47] method, which is a late-fusion method, was effective for fusing training models [48]. Hence, based on the predictions from (B) GRNNs for multiple types of deep features, MSV is applied to improve the performance. For each prediction $\{L_j, p_j\}$, $j = 1, \ldots, n$, in which L_j is the predicted label, and p_j is the probability of the corresponding label, n is the total number of models, MSV is defined by

$$L = \left\{ L_k | d_k = \max_{j=1}^{n} \left(p_j^1 - p_j^2 \right) \right\} \tag{7}$$

where p_j^1 and p_j^2 are the first and second highest probabilities, d_k is the MSV of the kth model, which is the most confident for the corresponding sample.

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

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

IV. Experiments and Results

A. Database

As mentioned, our proposed approach is evaluated on the dataset provided by the DCASE 2017 Challenge [29]. The dataset contains 15 classes, which include 'beach', 'bus', 'cafe/ restaurant', 'car', 'city centre', 'forest path', 'grocery store', 'home', 'library', 'metro station', 'office', 'park', 'residential area', 'train', and 'tram'. As further mentioned above, the organisers split each recording into several independent 10 s segments to increase the task difficulty and increase the number of instances. We train our model using a cross validation on the officially provided 4-fold development set, and evaluate on the official evaluation set. The development set contains 312 segments of audio recordings for each class and the evaluation set includes 108 segments of audio recordings for each class. Accuracy is used as the final evaluation metric.

B. Experimental Setup

First, we segment each audio clip into a sequence of 19 audio instances with 1000 ms and a 50 % overlap. Then, two types of representations are extracted: hand-crafted features for comparison, and deep image-based features, which have been described in Section III. Hand-crafted features are as follows:

Two kinds of low-level descriptors (LLDs) are extracted due to their previous success in ASC [29], [49], including Mel-frequency cepstral coefficient (MFCC) $1-14$ and logarithmic Mel-frequency band (MFB) $1-8$. According to feature sets provided in the INTERSPEECH COMPUTATIONAL PARALINGUISTICS CHALLENGE (COMPARE) [50], in total 100 functionals are applied to each LLD, yielding $14 \times 100 = 1400$ MFCCs features and $8 \times 100 = 800$ log MFBs features. The details of hand-crafted features and the feature extraction tool openSMILE can be found in [3].

These representations are then fed into the (B) GRNNs with 120 and 160 GRU nodes respectively with a 'tanh' activation, followed by a single highway network layer with a 'linear' activation function, which is able to ease gradient-based training of deep networks, and a softmax layer. Empirically, we implement this network using TensorFlow[4] and TFLearn[5] with a fixed learning rate of 0.0002 (optimiser 'rmsprop') and a batch size of 65. We evaluate the performance of the model at the kth training epoch, $k \in \{23, 30, \ldots, 120\}$. Finally, the MSV decision fusion strategy is applied to combine the (B) GRNNs models for the final predictions.

C. Results

We compute the mean accuracy on the 4-fold partitioned development set for evaluation according to the official protocols. Fig. 3 presents the performance of both GRNNs and BGRNNs on different feature sets when stopping at the multiple training epochs. From this we can see that, the accuracies of both GRNNs and BGRNNs on MFCCs, and log MFBs features are lower than the baseline. However, the performances of deep

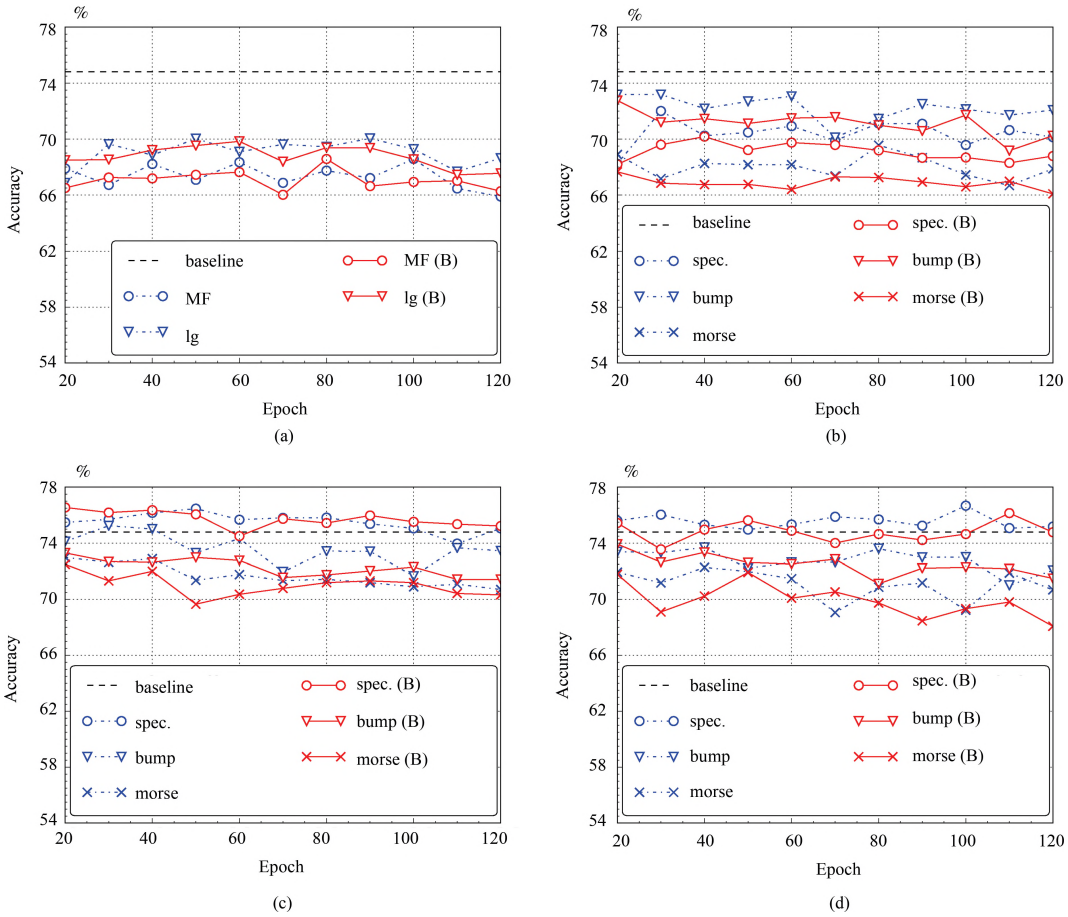

Fig. 3. The performances of GRNNs and BGRNNs on different features. (a) MFCCs (MF) and log MFBs (lg) features. The performances of features from the spectrogram and scalograms (bump and morse) extracted by three CNNs. (b) AlexNet. (c) VGG16. (d) VGG19.

features extracted by pre-trained CNNs are comparable with the baseline result, especially the representations extracted by the 'VGG16' and the 'VGG19' from spectrograms. This indicates the effectiveness of deep image-based features for this task.

Table II presents the accuracy of each model from each type of feature. For the development set, the accuracy of each type of feature is denoted as the highest one of all epochs. For the evaluation set, we choose the consistency epoch number of the development set. We find that the accuracies after decision fusion achieve an improvement based on a single spectrogram or scalogram image. In the results, the performances of BGRNNs and GRNNs are comparable on the development set but the accuracies on the BGRNNs are slightly higher than those of the GRNNs on the evaluation set, presumably because the BGRNNs cover the overall information in both the forward and backward time direction. The best performance of 84.4 % on the development set is obtained when extracting features from the spectrogram and the bump scalogram by the 'VGG19' and classifying by GRNNs at epoch 20. This is an improvement of 8.6 % over the baseline of the DCASE 2017 challenge ($p < 0.001$ by a one-tailed z-test). The best result of 64.0 % on the evaluation set is also obtained when extracting features from the spectrogram and bump scalogram by the 'VGG19', but classifying by BGRNNs at epoch 20. The performance on the evaluation set is also an improvement upon the 61.0 % baseline.

V. Discussion

The proposed approach in our study improves on the baseline performance given for the ASC task in the DCASE 2017 Challenge for sound scene classification and performs better than (B) GRNNs based on a hand-crafted feature set. The accuracy of (B) GRNNs on deep learnt features from a spectrogram, bump, and morse scalograms outperform MFCC and log MFB in Fig. 3. The performance of fused (B) GRNNs on deep learnt features is also considerably better than on hand-crafted features in Table II. Hence, the feature extraction method based on CNNs has proven itself to be efficient for the ASC task. We also investigate the performance when combining different spectrogram or scalogram representations. In Table II, the bump scalogram is validated as being capable of improving the performance of the spectrogram alone.

Fig. 4 shows the confusion matrix of the best results on the evaluation set. The model performs well on some classes, such as 'forest path', 'home', and 'metro station'. Yet, other classes such as 'library' and 'residential area' are hard to recognise. We think this difficulty is caused by noises or that the waveforms have similar environments within the acoustic scene.

To investigate the performance of each spectrogram or scalogram on different classes, a performance comparison of

TABLE II
PERFORMANCE COMPARISONS ON THE DEVELOPMENT AND THE EVALUATION SET BY GRNNS AND BGRNNS ON HAND-CRAFTED FEATURES (MFCCS (MF) AND LOG MFBS (LG)) AND FEATURES EXTRACTED BY PRE-TRAINED CNNS FROM THE SPECTROGRAM (S), BUMP SCALOGRAM (B), AND MORSE SCALOGRAM (M)

	GRNNs						BGRNNs					
	Development set			Evaluation set			Development set			Evaluation set		
	MF	lg	MF + lg	MF	lg	MF + lg	MF	lg	MF + lg	MF	lg	MF + lg
acc (%)	68.6	70.0	75.6	49.3	56.0	56.9	68.6	69.8	74.7	48.7	53.7	52.1
acc (%)	AlexNet	VGG16	VGG19	AlexNet	VGG16	VGG19	AlexNet	VGG16	VGG19	AlexNet	VGG16	VGG19
S	72.0	76.5	76.7	56.3	57.7	57.3	70.2	76.5	76.1	54.3	60.3	56.2
B	73.2	75.2	73.7	52.1	48.8	50.4	72.7	73.3	73.9	50.9	53.9	52.0
M	69.5	73.0	72.3	46.1	51.1	49.0	67.6	72.5	71.9	46.1	50.4	49.7
S + B	78.9	**84.4**	82.3	55.9	**61.7**	61.4	78.0	81.9	**83.4**	58.5	**64.0**	59.4
S + M	76.8	82.6	81.5	54.6	61.0	57.8	76.5	82.4	82.1	57.2	60.7	59.5
B + M	76.1	77.4	80.1	47.5	54.1	54.8	73.7	76.8	78.6	48.5	53.4	53.0
S + B + M	79.7	82.6	83.7	56.5	60.7	61.3	78.1	81.3	82.8	57.1	62.2	59.0

TABLE III
PERFORMANCE COMPARISONS ON THE EVALUATION SET FROM BEFORE AND AFTER LATE-FUSION OF BGRNNS ON THE FEATURES EXTRACTED FROM THE SPECTROGRAM (S) AND THE BUMP SCALOGRAM (B)

Precision (%)	beach	bus	cafe	car	city	forest	groc.	home	library	metro	office	park	resid.	train	tram
S	54.6	30.6	52.8	64.8	51.9	81.5	62.0	69.4	35.2	83.3	88.0	48.1	58.3	71.3	52.8
B	10.2	62.0	61.1	47.2	65.7	88.0	36.1	98.1	25.0	87.0	17.6	24.1	49.1	88.0	49.1
S + B	40.7	55.6	66.7	58.3	63.0	88.0	54.6	92.6	30.6	89.8	74.1	41.7	59.3	88.0	57.4

the spectrogram and the bump scalogram from the best result on evaluation set is shown in Table III. We can see that, the spectrogram performs better than the bump scalogram for 'beach', 'grocery store', 'office', and 'park'. However, the bump scalogram is optimal for the 'bus', 'city', 'home', and 'train' scenes. After fusion, the precision of some classes is improved, such as 'cafe/restaurant', 'metro station', 'residential area', and 'tram'. Overall, it appears worth using the scalogram as an assistance to the spectrogram, to obtain more accurate prediction.

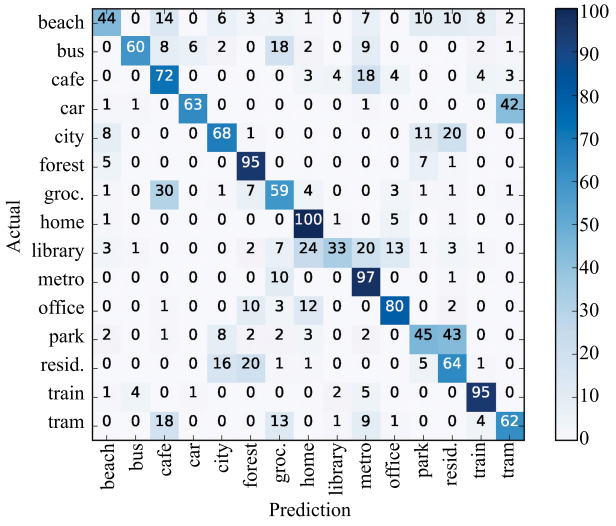

Fig. 4. Confusion matrix of the best performance of 64.0 % on the evaluation set. Late-fusion of BGRNNs on the features extracted from the spectrogram and the bump scalogram by 'VGG16'.

The result from the champion on the ASC task of the DCASE challenge 2017 is 87.1 % on the development set and 83.3 % on the evaluation set [51], using a generative adversarial network (GAN) for training set augmentation. There is a significant difference between the best result reached by the methods proposed herein which omit data augmentation, as we focus on a comparison of feature representations, and this result of the winning DCASE contribution in 2017 ($p < 0.001$ by one-tailed z-test). We believe that in particular the GAN part in combination with the proposed method shown herein holds promise to lead to an even higher overall result. Hence, it appears to be highly promising to re-investigate the proposed method in combination with data augmentation before training in future work.

VI. CONCLUSIONS

We have proposed an approach using pre-trained convolutional neural networks (CNNs) and (bidirectional) gated recurrent neural networks ((B) GRNNs) on the spectrogram, bump, and morse scalograms of audio clips, to achieve the task of acoustic scene classification (ASC). This approach is able to improve the performance on the 4-fold development set of the 2017 IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), achieving an accuracy of 83.4 % for the ASC task, compared with the baseline of 74.8 % of the DCASE challenge ($P < 0.001$, one-tailed z-test). On the evaluation set, the performance is improved from the baseline of 61.0 % to 64.0 %. The highest accuracy on the evaluation set is obtained when combining models from both the spectrogram and the scalogram images; therefore, the scalogram appears helpful to improve the performance reached by spectrogram images for the task of ASC. We focussed on the comparison of feature types in this contribution, rather than trying to reach overall best results by combination of 'tweaking on all available screws' such as is usually done

by entries into challenges. Likewise, we did for example not consider data augmentation by generative adversarial networks (GANs) or similar topologies as for example the DCASE 2017 winning contribution did. In future studies on the task of ASC, we will thus include further optimisation steps as the named data augmentation [52], [53]. In particular, we also aim to use evolutionary learning to generate adaptive 'self-shaping' CNNs automatically. This avoids having to hand-pick architectures in cumbersome optimisation runs.

REFERENCES

[1] E. Marchi, D. Tonelli, X. Z. 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. Detection and Classification of Acoustic Scenes and Events* , Budapest, Hungary, 2016, pp. 65−69.

[2] W. He, Z. J. Li, and C. L. P. Chen, "A survey of human-centered intelligent robots: Issues and challenges," *IEEE/CAA J. of Autom. Sinica*, vol. 4, no. 4, pp. 602−609, Oct. 2017.

[3] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 835−838.

[4] L. Li, Y. L. Lin, N. N. Zheng, and F. Y. Wang, "Parallel learning: A perspective and a framework," *IEEE/CAA J. of Autom. Sinica*, vol. 4, no. 3, pp. 389−395, Jul. 2017.

[5] F. Y. Wang, N. N. Zheng, D. P. Cao, C. M. Martinez, L. Li, and T. Liu, "Parallel driving in CPSS: A unified approach for transport automation and vehicle intelligence," *IEEE/CAA J. of Autom. Sinica*, vol. 4, no. 4, pp. 577−587, Oct. 2017.

[6] 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. INTERSPEECH 2017: Conf. Int. Speech Communication Association*, Stockholm, Sweden, 2017, pp. 3512−3516.

[7] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016*, Budapest, Hungary, 2016, pp. 95−99.

[8] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 961−1005, Sep. 1990.

[9] V. N. Varghees and K. I. Ramachandran, "Effective heart sound segmentation and murmur classification using empirical wavelet transform and instantaneous phase for electronic stethoscope," *IEEE Sens. J.*, vol. 17, no. 12, pp. 3861−3872, Jun. 2017.

[10] K. Qian, C. Janott, Z. X. Zhang, C. Heiser, and B. Schuller, "Wavelet features for classification of vote snore sounds," in *Proc. 2016 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 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. 39th Ann. Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, Seogwipo, South Korea, 2017, pp. 3737−3740.

[12] K. Qian, C. Janott, V. Pandit, Z. X. 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 multifeature analysis," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1731−1741, Aug. 2017.

[13] K. Qian, Z. Ren, V. Pandit, Z. J. Yang, Z. X. Zhang, and B. Schuller, "Wavelets revisited for the classification of acoustic scenes," in *Proc. Detection and Classification of Acoustic Scenes and Events 2017*, Munich, Germany, 2017, pp. 108−112.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. H. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211−252, Dec. 2015.

[15] J. Schlüter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *Proc. 2014 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Florence, Italy, 2014, pp. 6979−6983.

[16] G. Gwardys and D. Grzywczak, "Deep image features in music information retrieval," *Int. J. Electron. Telecomm.*, vol. 60, no. 4, pp. 321−326, Dec. 2014.

[17] J. Deng, N. Cummins, J. Han, X. Z. Xu, Z. Ren, V. Pandit, Z. X. Zhang, and B. Schuller, "The University of Passau open emotion recognition system for the multimodal emotion challenge," in *Proc. 7th Chinese Conf. Pattern Recognition (CCPR)*, Chengdu, China, 2016, pp. 652−666.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, 2012, pp. 1097−1105.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations*, San Diego, CA, USA, 2015.

[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345−1359, Oct. 2010.

[21] W. Y. Zhang, H. G. Zhang, J. H. Liu, K. Li, D. S. Yang, and H. Tian, "Weather prediction with multiclass support vector machines in the fault detection of photovoltaic system," *IEEE/CAA J. of Autom. Sinica*, vol. 4, no. 3, pp. 520−525, Jul. 2017.

[22] S. Young, G. Evermann, D. Kershaw, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, UK: Cambridge University Engineering Department, 2002.

[23] D. P. Mandic and J. A. Chambers, *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability*. New York, USA: Wiley Online Library, 2002.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735−1780, Nov. 1997.

[25] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016*, Budapest, Hungary, 2016, pp. 11−15.

[26] D. Yu and J. Y. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA J. of Autom. Sinica*, vol. 4, no. 3, pp. 396−409, Jul. 2017.

[27] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS 2014 Deep Learning and Representation Learning Workshop*, Montreal, Canada, 2014.

[28] Z. Ren, V. Pandit, K. Qian, Z. J. Yang, Z. X. Zhang, and B. Schuller, "Deep sequential image features for acoustic scene classification," in *Proc. Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, 2017, pp. 113−117.

[29] 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. Workshop on Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, 2017, pp. 85−92.

[30] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. 2017 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, 2017, pp. 131−135.

[31] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. Detection and Classification of Acoustic Scenes and Events 2017*, Munich, Germany, 2017, pp. 17−21.

[32] E. Fonseca, R. Gong, D. Bogdanov, O. Slizovskaia, E. Gomez, and X. Serra, "Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks," in *Proc. Detection and Classification of Acoustic Scenes and Events 2017*, Munich, Germany, 2017, pp. 37−41.

[33] A. Vafeiadis, D. Kalatzis, K. Votis, D. Giakoumis, D. Tzovaras, L. M. Chen, and R. Hamzaoui, "Acoustic scene classification: From a hybrid classifier to deep learning," in *Proc. Detection and Classification of Acoustic Scenes and Events 2017*, Munich, Germany, 2017, pp. 123−127.

[34] S. Park, S. Mun, Y. Lee, and H. Ko, "Acoustic scene classification based on convolutional neural network using double image features," in *Proc. Detection and Classification of Acoustic Scenes and Events 2017*, Munich, Germany, 2017, pp. 98−102.

[35] R. N. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-

extraction algorithm," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 121−131, Jan. 2011.

[36] T. H. Vu and J. C. Wang, "Acoustic scene and event recognition using recurrent neural networks," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016*, Budapest, Hungary, 2016.

[37] M. Zöhrer and F. Pernkopf, "Gated recurrent networks applied to acoustic scene classification and acoustic event detection," in *Proc. Detection and Classification of Acoustic Scenes and Events 2016*, Budapest, Hungary, 2016, pp. 115−119.

[38] E. Sejdić, I. Djurović, and J. Jiang, "TimeCfrequency feature representation using energy concentration: an overview of recent advances," *Digit. Signal Process.*, vol. 19, no. 1, pp. 153−183, Jan. 2009.

[39] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, Pa, USA: SIAM, 1992.

[40] S. C. Olhede and A. T. Walden, "Generalized morse wavelets," *IEEE Trans. Signal Process.*, vol. 50, no. 11, pp. 2661−2670, Nov. 2002.

[41] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, Australia, 2015, pp. 689−692.

[42] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. 32nd Int. Conf. Machine Learning*, Lille, France, 2015, pp. 2342−2350.

[43] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learning Representations 2015*, San Diego, CA, USA, 2015.

[44] Z. C. Yang, D. Y. Yang, C. Dyer, X. D. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," in *Proc. NAACL+HLT 2016*, San Diego, CA, USA, 2016, pp. 1480−1489.

[45] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673−2681, Nov. 1997.

[46] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," arXiv preprint, arXiv: 1505.00387, 2015.

[47] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in *Proc. 4th Int. Conf. Advances in Intelligent Data Analysis*, Porto, Portugal, 2001, pp. 309−318.

[48] K. Qian, Z. X. Zhang, A. Baird, and B. Schuller, "Active learning for bird sound classification via a kernel-based extreme learning machine," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1796, Oct. 2017.

[49] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. 24th European Signal Processing Conf.* , Budapest, Hungary, 2016, pp. 1128−1132.

[50] 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. 14th Ann. Conf. Int. Speech Communication Association*, Lyon, France, 2013, pp. 148−152.

[51] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," in *Proc. Detection and Classification of Acoustic Scenes and Events 2017*, Munich, Germany, 2017, pp. 93−97.

[52] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 2672−2680.

[53] K. F. Wang, C. Gou, Y. J. Duan, Y. L. Lin, X. H. Zheng, and F. Y. Wang, "Generative adversarial networks: introduction and outlook," *IEEE/CAA J. of Autom. Sinica*, vol. 4, no. 4, pp. 588−598, Oct. 2017.

Zhao Ren (S'17) received the master degree in computer science and technology from Northwestern Polytechnical University (NWPU), China, 2017. Currently, she is a Research Assistant and working on the Ph.D. degree at the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, where she is involved with the German national BMBF IKT2020-Grant project EmotAsS, for emotion analysis based on speech. Her research interests mainly lie in transfer learning, unsupervised learning, and deep learning for the application in health care and wellbeing.

Kun Qian (S'14) received the master degree in signal and information processing from the Nanjing University of Science and Technology (NUST), China, 2014. Currently, he is working on the Ph.D. degree in electrical engineering and information technology at Technische Universität München (TUM), Munich, Germany. He was sponsored by scholarships to conduct cooperative research at the Nanyang Technological University (NTU), Singapore, the Tokyo Institute of Technology (Tokyo Tech), Japan, and the Carnegie Mellon University (CMU), USA. His research interests include signal processing, machine learning, biomedical engineering, and deep learning in high performance computing systems.

Zixing Zhang (M'15) received the master degree in physical electronics from Beijing University of Posts and Telecommunications, China, 2010, and the Ph.D. degree in engineering from the Machine Intelligence and Signal Processing group at Technische Universität München (TUM), Munich, Germany, 2015. He is currently a Research Associate at Imperial College London, UK. He has authored more than fifty publications in peer-reviewed journals and conference proceedings. His research interests mainly lie in semi-supervised learning, active learning, and deep learning for the application in affective computing.

Vedhas Pandit (S'11) received the master degree from Arizona State University (ASU) in USA, 2010, in electronic and mixed signal circuit design (EECE) with his thesis on mathematical modelling of a-Si:H SOI transistors. After working for Intel as a Graphics Hardware Engineer, he worked as a Researcher at the Indian Institute of Technology Bombay (IITB) developing tools for automated music information retrieval. Since February 2015, he has been working on the Ph.D. degree at the University of Passau, Germany, and the University of Augsburg, Germany. His research interests include music information retrieval, speech and virtual instrument synthesis, deep learning strategies in machine learning, and biomedical signal processing.

Alice Baird is a Research Assistant at the ZD.B Chair of Embedded Intelligence for Healthcare and Wellbeing, University of Augsburg, Germany, where she is involved with the Horizon 2020 project DE-ENIGMA, for analysis of vocal and linguistic cues. Alice has recently been awarded a ZD.B Ph.D. Fellowship (2018−2021), in which she will research speech monitoring and soundscape synthesis. Alice has an (S'16) M.FA in Sound Arts from Columbia University, Computer Music Center, and a (S'13) B.A. in Music Technology from London Metropolitan University. Alice works across an array of disciplines, predominately in the realm of paralinguistic speech and intelligent audio analysis. Her research focus is towards applications of computing for health and wellbeing, with consideration to methodologies for 'in the wild' data collection.

Björn Schuller (M'06-SM'15-F'18) received his diploma in 1999, his doctoral degree for his study on automatic speech and emotion recognition in 2006, and his habilitation and adjunct teaching professorship in the subject area of signal processing and machine intelligence in 2012, all in electrical engineering and information technology from Technische Universität München (TUM), Germany. He is a tenured Full Professor heading the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, and a Reader (Associate Professor) in Machine Learning heading GLAM — the Group on Language, Audio and Music, Department of Computing at the Imperial College London in London, UK. Dr. Schuller is elected member of the *IEEE Speech and Language Processing Technical Committee*, Editor in Chief of the *IEEE Transactions on Affective Computing*, President-emeritus of the AAAC, Fellow of the IEEE, and Senior Member of the ACM. He (co-)authored 5 books and more than 700 publications in peer reviewed books, journals, and conference proceedings leading to more than 17 000 citations (h-index 64).