Bag-of-Deep-Features: Noise-Robust Deep Feature Representations for Audio Analysis

Shahin Amiriparian*†, Maurice Gerczuk*, Sandra Ottl*, Nicholas Cummins*, Sergey Pugachevskiy*, Björn Schuller*‡

*Chair of Embedded Intelligence for Health Care and Wellbeing, Univeristy of Augsburg, Germany.
†Machine Intelligence and Signal Processing Group, Technische Universität München, Germany.
‡GLAM – Group on Language, Audio and Music, Imperial College London, UK.

shahin.amiriparian@tum.de

Abstract—In the era of deep learning, research into the classification of various components of the acoustic environment, especially in-the-wild recordings, is gaining in popularity. This is due in part to the increasing computational capacities and the expanding amount of real-world data available on social multimedia. However, the noisy nature of this data can add an additional complexity to the already complex deep learning systems. Herein, we tackle this issue by quantising deep feature representations of various in-the-wild audio data sets. The aim of this paper is twofold: 1) to assess the feasibility of the proposed feature quantisation task, and 2) to compare the efficacy of various feature spaces extracted from different fully connected deep neural networks to classify six real-world audio corpora. For the classification, we extract two feature sets: i) DEEP SPECTRUM features which are derived from forwarding the visual representations of the audio instances, in particular mel-spectrograms through very deep task-independent pre-trained Convolutional Neural Networks (CNNs), and ii) Bag-of-Deep-Features (BoDF) which is the quantisation of the DEEP SPECTRUM features. Using BoDF, we show the suitability of quantising the deep representations for noisy in-the-wild audio data. Finally, we analyse the effect of early and late fusion of the CNN features and models on the classification results.

I. INTRODUCTION

Within the field of machine understanding, unsupervised representation learning is gaining widespread interest as a highly effective alternative to using conventional 'hand-crafted' feature sets [1]–[4]. In computer audition, the understanding of audio and soundscapes by machines, and unsupervised representation learning based on deep learning techniques have been used for a diverse range of tasks, including environmental audio tagging [5], acoustic scene classification [6], and urban sound classification [7]. One particular challenging aspect of computer audition is soundscape perception and classification of audio recorded in real-world environments. Such audio samples typically contain a variety of confounding effects such as non-stationary noise and less than ideal microphone placements. Despite the wide use of unsupervised feature representation, their advantages have yet to be fully realised in such adverse conditions. In this regard, this paper explores if a bag-of-audio-words approach [8] can be used to improve the robustness of the state-of-the-art DEEP SPECTRUM feature representation [9]–[11]. To help ensure authenticity, we use various data sets which have been sourced entirely from YouTube with our Cost-efficient Audio-visual Acquisition via Social-media Small-world Targeting (CAS²T) toolkit for efficient large-scale big data collection [12].

Our proposed feature representation, herein referred to as Bag-of-Deep-Features (BoDF), are generated via quantising (bagging) DEEP SPECTRUM features. DEEP SPECTRUM features are generated by forwarding audio spectra through pre-trained image Convolutional Neural Networks (CNNs) such as AlexNet [13], VGG16 and VGG19 [14], and GoogLeNet [15]. Despite not being trained for audio, the deep convolutional operations of these networks extract salient audio features from spectrograms. The versatility of DEEP SPECTRUM features has been shown in a range of audio classification tasks, including autism severity detection [16], snore sound recognition [10], [17], audio-based sentiment analysis [9], and speech-based emotion detection [11].

The aim of extending DEEP SPECTRUM features into BoDF representations is to further increase robustness to noise related adverse and confounding effects. BoAW have also been shown to be a useful feature representation [18]–[20]. BoAW is a histogram representation of the original feature space generated by first assigning features to their 'nearest' representations in a predetermined dictionary and then counting the final number of assignment to each dictionary element [8]. This quantisation step can be considered to be quasi-filtering against small amounts of noise present in the original feature space. We therefore hypothesise that bagging will help increase the robustness of a DEEP SPECTRUM feature space.

The rest of this paper is laid out as follows. Section II introduces our databases. Section III outlines our machine learning methods for extracting the deep representations from the audio files. The classification experiments and the evaluation metrics are outlined in Section IV. The obtained results are given in Section V, before concluding the paper in Section VI.

II. DATABASES

For our classification experiments, we choose 6 unique audio databases containing different human speech and vocalisation types [12]:

1) *Freezing:* 785 recordings picked from videos in which the speech is produced by an individual shivering with cold.

TABLE I: *Specifications of each data set.* l_{total}*: the total length of the data set;* l_{min} *and* l_{max}*: the minimum and maximum lengths of the audio recording;* σ*: standard deviation;* $\#n$*: the number of all audio recordings in each set.* $\#s$ *: the number of 0.5 s segments, i. e. the number of frames of input mel-spectrograms, denoted in parentheses.* c_{ratio}*: the class ratio for each data set (target class:'normal speech').*

| | Train | | | | | Evaluation | | | | |
Tasks	l_{total}	l_{min}/l_{max}	σ	$\#n\ (\#s)$	c_{ratio}	l_{total}	l_{min}/l_{max}	σ	$\#n\ (\#s)$	c_{ratio}
Freezing	75.9 m	2.0 s/29.4 s	5.8 s	614 (8 813)	2 : 1	22.4 m	2.0 s/28.6 s	5.9 s	171 (2 595)	1.1 : 1
Intoxication	139.7 m	2.0 s/29.9 s	6.5 s	1069 (16 200)	0.9 : 1	16.7 m	2.0 s/24.8 s	5.3 s	152 (1 930)	1.8 : 1
Screaming	53.6 m	2.0 s/29.9 s	7.6 s	375 (6 192)	1.2 : 1	22.0 m	2.1 s/29.9 s	5.5 s	189 (2 505)	1.4 : 1
Threatening	106.6 m	2.0 s/29.8 s	7.4 s	652 (12 360)	6 : 1	45.8 m	2.0 s/29.2 s	5.2 s	441 (5 271)	0.6 : 1
Coughing	94.3 m	0.5 s/28.8 s	3.5 s	2 088 (10 336)	2.9 : 1	63.9 m	0.5 s/23.2 s	2.7 s	1 571 (6 935)	2.2 : 1
Sneezing	6.7 m	0.5 s/8.0 s	1.3 s	238 (691)	0.9 : 1	9.2 m	0.5 s/9.3 s	1.4 s	291 (967)	1 : 1
Σ	476.8 m	–	–	5 036 (54 592)	——	180 m	–	–	2 815 (19 933)	——

2) *Intoxication:* 1 221 language independent recordings picked from videos in which the speech is produced under the influence of drugs.
3) *Screaming:* 564 recordings picked from videos in which people are screaming when they are scared.
4) *Threatening:* 1 093 language independent recordings picked from videos in which the speech is perceived by our annotators to be of a threatening manner.
5) *Coughing:* 3 659 recordings picked from videos in which people are coughing during a conversation or a talk.
6) *Sneezing:* 529 recordings picked from videos in which people are sneezing during a conversation or a talk.

These datasets are based on the concept of acoustic surveillance [21]. The first four topics are related to audio-based surveillance for security purposes in noisy public places. The latter two topics, related to the monitoring of everyday activity – in terms of, e. g. personal health – in common, relatively quiet environments such as home or office [21]. All corpora offer a two-class classification problem, i. e. they have a target class, e. g. *freezing* or *intoxication* and a 'normal speech' class which contains audio samples that are not affected by the target class. All audio data has a sample rate of 44.1 kHz and the audio channel is mono. For full details on the construction of theses datasets the interested reader is referred to [12]. For details on the data see Table I.

III. DEEP FEATURE REPRESENTATIONS

Before starting to classify the databases, we first create mel-spectrograms from chunked audio recordings (cf. Section III-A). We then send these visual representations through various image classification CNNs (cf. Section III-B) to extract the DEEP SPECTRUM features (cf. Section III-C). Afterwards, we create Bag-of-Deep-Features (BoDF) by quantising the DEEP SPECTRUM features in order to cope with the amount of noise in the audio recordings (cf. Section III-D).

A. Mel-Spectrograms

The mel-Spectrograms are computed with a window size of 2 048 and an overlap of 1 024 from the log-magnitude spectrum by dimensionality reduction using a mel-filter. We

TABLE II: Overview of the architectural similarities and differences between the three of the CNNs used for the extraction of DEEP SPECTRUM features, AlexNet, VGG16, and VGG19. *conv* denotes convolutional layers and *ch* stands for channels. The table is adapted from [10].

AlexNet	VGG16	VGG19
input: RGB image		
1×conv size: 11; ch: 96; stride: 4	2×conv size: 3; ch: 64; stride: 1	
maxpooling		
1×conv size: 5; ch: 256	2×conv size: 3; ch: 128	
maxpooling		
1×conv size: 3; ch: 384	3×conv size: 3; ch: 256	4×conv size: 3; ch: 256
maxpooling		
1×conv size: 3; ch: 384	3×conv size: 3; ch: 512	4×conv size: 3; ch: 512
maxpooling		
1×conv size: 3; ch: 256	3×conv size: 3; ch: 512	4×conv size: 3; ch: 512
maxpooling		
fully connected *fc6*, 4 096 neurons		
fully connected *fc7*, 4 096 neurons		
fully connected, 1 000 neurons		
output: soft-max of probabilities for 1 000 object classes		

apply 128 filter banks equally spaced on the mel-scale defined in Equation (1):

$$2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \tag{1}$$

The mel-scale is based on the frequency response of the human ear that has better resolution at lower frequencies. We also display the mel-spectrogram on this scale. A sample mel-spectrogram plot for a member of the sneezing class can be seen in Figure 1. For the mel-spectrogram plots, we use two different colour mappings: *viridis*, and *magma*. It is during testing (cf. Section V) that we identify the optimal colour map for the spectral feature spaces. In Figure 4, we highlight the audio similarities and differences that potentially exist between different classes in our corpora by showing an example mel-
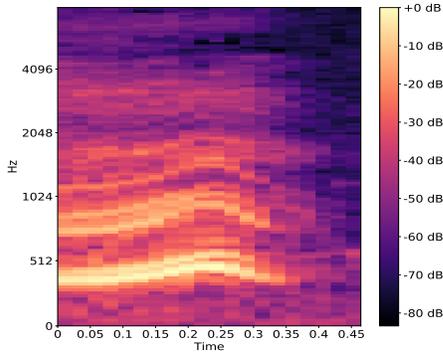
Fig. 1: A mel-spectrogram plot of an audio-sample from the *sneezing* class with the colour map *magma*. The colour bar to the right shows the colour changes associated with increasing spectral energy.
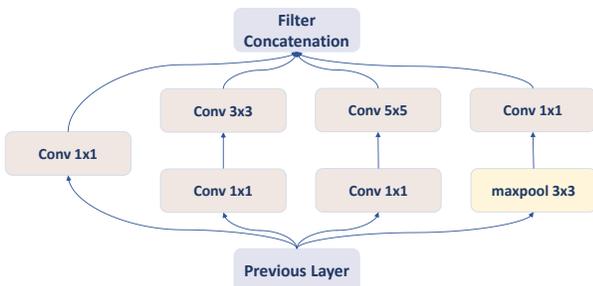

Fig. 2: An inception module used in the GoogLeNet architecture. Small 1×1 convolutions reduce the dimensionality, and filters of different patch sizes are concatenated to combine information found at different scales.

spectrogram from each target class and from a 'normal speech' which was not affected by the target classes.

B. Deep Feature Extractors

We use four different architectures of image CNNs to extract deep representations from the mel-spectrograms described in Section III-A. All four networks have been trained for the task of object categorisation on the large ImageNet [22] corpus which provides more than 1 million images labelled with 1 000 object classes. The architectural differences and similarities of AlexNet, VGG16, and VGG19 are given in Table II. The GoogLeNet's architecture is depicted in Figure 2.

1) AlexNet: AlexNet's architecture consists of 5 convolutional layers followed by 3 fully connected layers [13]. Overlapping maxpooling is used between the first, second, and third convolutional layer, and a rectified linear unit (ReLU) non-linearity is applied to improve generalisation capabilities. We use the 4 096 activations of AlexNet's seventh layer (commonly denoted as *fc7*) as features.

2) VGG16/VGG19: In contrast to AlexNet, both VGG16 and VGG19 utilise small 3×3 receptive fields in all of their convolutional layers [14]. Both architectures include 2

additional maxpooling layers and are deeper than AlexNet at 16 and 19 layers. Similar to AlexNet, ReLUs are applied for response normalisation. For both networks, the activations of the second fully connected layer are considered as feature vectors of size 4 096.

3) GoogLeNet: The fourth and one of the strongest image CNN models we applied as a feature extractor in our experiments is GoogLeNet [15] which bases its architecture on the so-called inception modules (cf. Figure 2). These modules aggregate information extracted at different scales by combining the activations of convolutional filters of different size. In the overall network, pooling layers are applied after these inception modules and a fully connected layer is used for the actual ImageNet classification task. We consider the activations of the last pooling layer as features for our classification tasks.

C. DEEP SPECTRUM *Features*

We use a state-of-the-art system based on the introduced CNN image descriptors (cf. Section III-B). The basic system architecture (before quantisation) is shown in the left part of Figure 3. We extract the DEEP SPECTRUM features as follows. First, mel-spectrograms are created from the chunked (each 0.5 s) audio recordings using the audio and music analysis library *librosa* [23]. We choose mel-spectrograms since they have been successfully applied for a wide range of audio recognition tasks [7], [24]–[26]. The mel-spectrograms are then transformed to images by creating colour mapped plots. The second step consists of feeding the created plots to the pre-trained CNNs and extracting the activations of a specific layer from each CNN as large feature vectors. These features are a high-level representation of the plots generated from low-level audio features. Initial experiments indicated that the *viridis* spectrograms worked better for AlexNet and for the other CNNs, the *magma* spectrograms led to better performance.

D. Bag-of-Deep-Features

The last important component of our system is the feature quantisation block (cf. Figure 3). In this stage, we bag (quantise) the extracted DEEP SPECTRUM features which we described in Section III-C to analyse the denoising effect of the deep feature quantisation. In order to achieve this, we generate a fixed length histogram representation of each audio recording. This is done by first identifying a set of 'deep audio words' from some given training data, and then quantising the original feature space, with respect to the generated codebook, to form the histogram representation. The histogram shows the frequency of each identified deep audio word in a given audio instance [8], [18], [19].

We normalise the features to $[0, 1]$ and random sample a codebook with fixed size from the training partition. Afterwards, each input feature vector (from training and evaluation partitions) is applied a fixed number of its closest vectors from the codebook. We then use logarithmic term-frequency weighting to the generated histograms.

The size of the codebook and the number of assigned codebook words (*cw*) are optimised
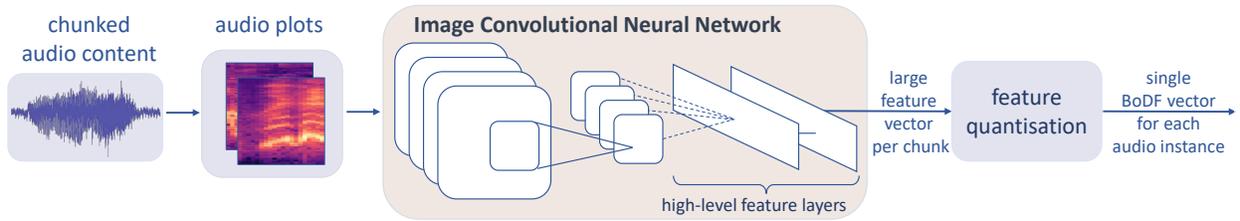
Fig. 3: Mel-spectrograms are generated from the chunked audio files. They are then used as input for the image CNN networks and the activations of a specific layer are extracted as DEEP SPECTRUM features. This results in a large feature vector for each chunk of an audio clip. Finally, for each clip, the extracted chunk-level vectors are bagged to form a single BoDF vector. For the last component, we use openXBOW, our open-source toolkit for the generation of bag-of-words representations [8].

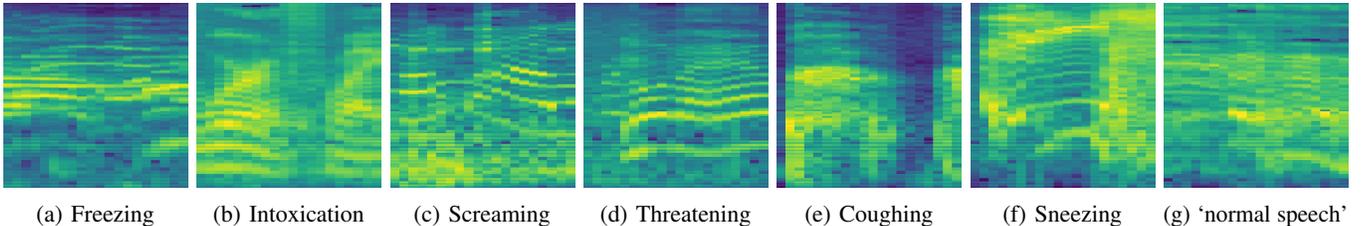| (a) Freezing | (b) Intoxication | (c) Screaming | (d) Threatening | (e) Coughing | (f) Sneezing | (g) 'normal speech' |

Fig. 4: Example mel-spectrograms (a – f) extracted from the target classes contained in the six different corpora. The last example mel-spectrogram is from an audio sample considered to be a 'normal speech' utterance which was not affected by the target classes. The range of the horizontal (time) and vertical (frequency) axes are [0 - 0.45] s and [0 - 4 096] Hz (cf. Figure 1). We observe relatively high f0 of the *Screaming* class, wide band spectra for *Coughing* and *Sneezing* classes, narrow band spectra for the *Freezing* and *Threatening* classes, and for the *Intoxication* class we see that the lower frequencies are more dense (have higher amplitude).

with $size \in \{10, 20, 50, 100, 200, 500, 1\,000\}$, $cw \in \{1, 10, 25, 50, 100, 200, 500\}$ and evaluated on the evaluation partition using a the linear SVM classifier. For this purpose, the classifier's complexity parameter is optimised on a logarithmic scale between 10^{-9} and 10^0 with a step size of 10^1. The best performing codebook is then applied for evaluation on the test set.

IV. CLASSIFIER AND EVALUATION METRICS

In order to predict the class labels for the audio instances in each of the six corpora, we train a linear Support Vector Machine (SVM) classifier. The evaluation metric is unweighted average recall (UAR) as this measure gives equal weight to all classes and is accordingly more suitable than a weighted metric (e. g. accuracy) for our datasets which have imbalanced class distribution (cf. Table I).

For our classifier we use the open-source linear SVM implementation provided in the *scikit-learn* machine learning library [27]. For the extracted DEEP SPECTRUM features (cf. section III) we do not apply standardisation, i.e. subtracting the mean and dividing by the standard deviation and normalisation as they have been found to negatively impact classifier performance. Moreover, we preferred to apply SVM over Deep Neural Network (DNN) as the classifier for two reasons: first, the data sets are too small for a DNN, and second, BoDF is a sparse feature representation and SVM are effective at handling sparse data.

V. RESULTS

An extensive series of experiments has been conducted to evaluate the performance of the extracted deep feature representations (cf. Section III) using the proposed classifier (cf. Section IV). First, we obtain the classification results for the non-quantised DEEP SPECTRUM features (cf. Section V-A). We then evaluate the robustness and the performance of quantising the representations (BoDF) for all four CNN-descriptors (cf. Section V-B). Finally, we perform early (feature) and late (model) fusion for various combinations of the CNN-descriptors (cf. Section V-C).

A. DEEP SPECTRUM *Features (non-quantised)*

For the non-bagged DEEP SPECTRUM features extracted from the mel-spectrograms of two colour maps using four different pre-trained CNNs, we applied majority voting to obtain the prediction for a whole audio recording from its chunk-level results. We observe that the results – despite being strong – are behind the best baseline (cf. column 1, and 2 in Table III) and almost all BoDF results. This is mainly due to the existing amount of noise in the feature set. We handled this issue by generating BoDF.

B. *Bag-of-Deep-Features*

We generated the BoDF representations for all databases and optimised the BoDF parameters codebook size and number of assigned codebook words *cw* (cf. Section III-D). The results

TABLE III: *Classification results of each paralinguistic task from the baseline paper [12] by the funtionals (Func.) of the IS09-emotion feature set,* BoAW, *and Convolutional Neural Network (CNN) compared with our results from the non-quantised* DEEP SPECTRUM *features (nqDS) and Bag-of-Deep-Features (*BoDF*). For the non-quantised* DEEP SPECTRUM *features we used majority voting to obtain the prediction for a whole audio recording from its chunk-level results. The best result for each corpus is highlighted with a light grey shading. The chance level for each task is 50.0 % UAR.*

% UAR	IS09-emotion (Ref. [12])			MFCCs (Ref. [12])		AlexNet		VGG16		VGG19		GoogLeNet	
	Func.	BoAW	CNN	BoAW	CNN	nqDS	BoDF	nqDS	BoDF	nqDS	BoDF	nqDS	BoDF
Freezing	70.2	67.5	56.9	65.6	51.0	62.5	70.4	71.3	**72.9**	67.9	69.1	67.3	71.6
Intoxication	64.7	72.6	66.8	66.7	67.5	60.3	61.9	58.2	64.7	55.4	71.3	63.1	**73.6**
Screaming	89.2	97.0	89.2	94.0	87.3	94.9	**98.5**	96.8	96.7	94.7	98.2	89.8	94.3
Threatening	73.8	66.3	71.9	67.0	70.3	72.2	76.4	70.7	73.9	70.6	70.3	70.5	**77.3**
Coughing	95.4	96.7	95.4	**97.6**	93.6	94.3	95.3	94.5	95.3	94.2	95.2	91.0	92.0
Sneezing	79.3	76.4	**85.2**	79.8	80.2	74.0	74.6	71.8	74.9	76.8	79.4	64.0	71.8

TABLE IV: *Performance of early and late fusion strategies for the CNN-descriptors using linear SVM classifiers on our corpora. UAR is used as the measure. For early fusion, the linear SVM classifier's complexity parameter is optimised on a logarithmic scale between 10^{-9} and 10^{0} with a step size of 10^{1}. For late fusion, we employ a majority vote on the test set using the best individual models obtained during previous experiments. We denote AlexNet as A., VGG16 as V16, VGG19 as V19, and GoogLeNet as G. The fusion results for each corpus which are better than the results given in Table III are highlighted with a light grey shading. The chance level for each task is 50.0 % UAR.*

% UAR	Early fusion						Late fusion				
	A.+V16	A.+V19	A.+G.	V16+G.	V19+G.	All	A.+V16+V19	A.+V16+G.	A.+V19+G.	V16+V19+G.	All
Freezing	69.2	71.3	68.4	74.1	70.4	71.2	70.4	**76.3**	70.4	71.2	68.5
Intoxication	65.7	73.0	67.1	68.4	67.8	**73.8**	68.3	67.7	68.8	63.9	60.9
Screaming	98.5	**99.1**	97.8	97.1	**99.1**	98.2	98.0	98.0	98.2	98.2	98.9
Threatening	74.8	72.8	76.0	73.2	72.3	73.1	73.1	76.3	75.2	72.2	68.9
Coughing	96.0	95.5	95.4	95.6	95.3	96.5	96.3	96.3	95.2	95.2	95.3
Sneezing	76.3	77.7	76.3	75.6	78.0	79.8	77.0	77.7	79.1	79.8	74.1

in Table III show that quantisation improves the results for all CNN-descriptors. Our results also demonstrate the strength of the BoDF outperforming the best baseline results for the acoustic surveillance tasks *Freezing*, *Intoxication*, *Threatening*, and *Screaming* corpora. We observed that BoDF worked best on the longer audio chunks as opposed to the short ones from the *Coughing* and *Sneezing* data sets. We assume this is due to the the lack of discriminating information in the shorter chunk spectrograms leading to weaker DEEP SPECTRUM representations. It is worth noting that for both *Coughing* and *Sneezing*, BoDF consistently outperforms DEEP SPECTRUM features adding evidence to that quantisation improving system robustness.

C. Fusion Experiments

We apply both early and late fusion schemes to our Bag-of-Deep-Feature systems (cf. Section III) in order to investigate their complementarity. For early (feature-level) fusion, we combine DEEP SPECTRUM features extracted using different CNN-architectures from the chunked (0.5 s) audio recordings.

Here, for each CNN-descriptor (AlexNet, VGG16, VGG19, and GoogleNet), we use the same plots and colourmaps as in our non fusion experiments, i. e. for AlexNet *viridis* mel-spectrograms and for the other three networks *magma* mel-spectrograms build the basis for feature extraction. Afterwards, we build BoDF representations of those features analogous to the non-fusion systems outlined in Section III. We again optimise the BoDF parameters on the evaluation partition. As before, a linear SVM is used for the classification task, and its cost parameter is optimised on a logarithmic scale between 10^{-9} and 10^{0} with a step size of 10^{1}. Our late fusion scheme on the other hand, combines the predictions of the best BoDF models for each dataset obtained in previous experiments in a majority vote. The results achieved by different configurations of these two fusion schemes on all six databases are displayed in Table IV. For *Sneezing* and *Coughing*, results are slightly improved over non-fusion systems but still do not reach the baseline performance in Table III. We further denote small performance boosts of the early fusion models over non-fusion systems for the *Intoxication* and *Screaming*

datasets and a larger increase in UAR for the *Freezing* dataset using a late fusion system of BoDF systems based on AlexNet, VGG16, and GoogleNet DEEP SPECTRUM features. However, as there is no consistent pattern, it is difficult to interpret the differences between the early and late fusion results of the CNN models. We observe that the early fusion of all features for all tasks, except for *Screaming*, lead to stronger performance than combining all models by late fusion. Based on these finding we assume that fusing high-level shift-invariant CNN features can lead to stronger performance than fusing the predictions of the trained models.

VI. CONCLUSIONS AND FUTURE WORK

Despite representation learning with deep neural networks having shown superior performance over expert-designed feature sets in a range of machine learning recognition tasks, such approaches have not been widely explored within the domain of noisy, in-the-wild, audio classification. In this regard, our results indicate that state-of-the-art image classification CNNs are capable of providing strong feature sets on real-world audio recognition. Further, we demonstrated the strength of bagging the DEEP SPECTRUM features as a means to cope with the amount of noise available in the corpora. We showed that using BoDF, it is possible to improve upon almost all results obtained from non-quantised DEEP SPECTRUM features. These results give strong evidence that the quantising step when bagging features can be viewed as a quasi-filtering process which, in general, improves system robustness. Finally, we showed that both early and late fusion can still increase the BoDF classification results. These findings imply that features and models obtained from the applied CNN-descriptors are in most cases complementary.

In the future work, we will be testing the BoDF with very deep residual networks such as ResNet [28]. We also want to explore the benefits of fine-tuning the pre-trained networks on larger in-the-wild databases like *AudioSet* [29] or data sets for Acoustic Scene Classification and Sound Event Detection challenges [6], [30], [31].

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognition Letters*, vol. 42, pp. 11 – 24, 2014.

[3] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *arXiv preprint arXiv:1712.04382*, 2017.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. Jackson, and M. D. Plumbley, "Unsupervised feature learning based on deep models for environmental audio tagging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1230–1241, 2017.

[6] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[7] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 171–175.

[8] M. Schmitt and B. Schuller, "openxbow – introducing the passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, 2017, 5 pages.

[9] S. Amiriparian, N. Cummins, S. Ottl, M. Gerczuk, and B. Schuller, "Sentiment analysis using image-based deep spectrum features," in *Proceedings of the 2nd International Workshop on Automatic Sentiment Analysis in the Wild, WASA 2017, held in conjunction with the 7th biannual Conference on Affective Computing and Intelligent Interaction, ACII 2017*, AAAC. San Antonio, TX: IEEE, October 2017, pp. 26–29.

[10] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 3512–3516.

[11] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM International Conference on Multimedia, MM 2017*. Mountain View, CA: ACM, Oct. 2017, pp. 478–484.

[12] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," in *Proceedings of the 7th biannual Conference on Affective Computing and Intelligent Interaction, ACII 2017*. San San Antonio, TX: IEEE, Oct. 2017, 340–345.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012, vol. 25, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computing Research Repository (CoRR)*, vol. abs/1409.1556, 2014.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA, USA: IEEE, 2015, pp. 1–9.

[16] A. Baird, S. Amiriparian, N. Cummins, A. M. Alcorn, A. Batliner, S. Pugachevskiy, M. Freitag, M. Gerczuk, and B. Schuller, "Automatic classification of autistic child vocalisations: A novel database and results," in *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 849–853.

[17] M. Freitag, S. Amiriparian, N. Cummins, M. Gerczuk, and B. Schuller, "An 'End-to-Evolution' Hybrid Approach for Snore Sound Classification," in *Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*, ISCA. Stockholm, Sweden: ISCA, August 2017, pp. 3507–3511.

[18] S. Pancoast and M. Akbacak, "Softening quantization in bag-of-audio-words," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1370–1374.

[19] ——, "Bag-of-audio-words approach for multimedia event classification." in *Proceedings of INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*. Portland, OR, USA: ISCA, 2012, pp. 2105–2108.

[20] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using lbp-hog based bag-of-audio-words feature representation," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] A. Härmä, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Proc. of the Int. Conf. on Multimedia and Expo*. Amsterdam, The Netherlands: IEEE, July 2005, no pagination.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[23] B. McFee, M. McVicar, O. Nieto, S. Balke, C. Thome, D. Liang, E. Battenberg, J. Moore, R. Bittner, R. Yamamoto, D. Ellis, F.-R. Stoter, D. Repetto, S. Waloschek, C. Carr, S. Kranzler, K. Choi, P. Viktorin, J. F. Santos, A. Holovaty, W. Pimenta, and H. Lee, "librosa 0.5.0," Feb. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.293021

[24] S. Panwar, A. Das, M. Roopaei, and P. Rad, "A deep learning approach for mapping music genres," in *System of Systems Engineering Conference (SoSE), 2017 12th*. IEEE, 2017, pp. 1–5.

[25] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "Dcase 2016 acoustic scene classification using convolutional neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 95–99.

[26] S. Amiriparian, N. Cummins, M. Gerczuk, S. Pugachevskiy, S. Ottl, and B. Schuller, ""are you playing a shooter again?!" deep representation learning for audio-based video game genre recognition," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. PP, 218, submitted, 10 pages.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 776–780.

[30] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Munich, Germany, Nov 2017.

[31] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference (EUSIPCO 2016)*. Budapest, Hungary: IEEE, Aug 2016, pp. 1128–1132.