Convolutional Gated Recurrent Neural Network Incorporating Spatial Features for Audio Tagging

Yong Xu Qiuqiang Kong Qiang Huang Wenwu Wang Mark D. Plumbley
Center for Vision, Speech and Sigal Processing University of Surrey, Guildford, UK
Email: {yong.xu, q.kong, q.huang, w.wang, m.plumbley}@surrey.ac.uk

Abstract—**Environmental audio tagging is a newly proposed task to predict the presence or absence of a specific audio event in a chunk. Deep neural network (DNN) based methods have been successfully adopted for predicting the audio tags in the domestic audio scene. In this paper, we propose to use a convolutional neural network (CNN) to extract robust features from mel-filter banks (MFBs), spectrograms or even raw waveforms for audio tagging. Gated recurrent unit (GRU) based recurrent neural networks (RNNs) are then cascaded to model the long-term temporal structure of the audio signal. To complement the input information, an auxiliary CNN is designed to learn on the spatial features of stereo recordings. We evaluate our proposed methods on Task 4 (audio tagging) of the Detection and Classification of Acoustic Scenes and Events 2016 (DCASE 2016) challenge. Compared with our recent DNN-based method, the proposed structure can reduce the equal error rate (EER) from 0.13 to 0.11 on the development set. The spatial features can further reduce the EER to 0.10. The performance of the end-to-end learning on raw waveforms is also comparable. Finally, on the evaluation set, we get the state-of-the-art performance with 0.12 EER while the performance of the best existing system is 0.15 EER.**

I. Introduction

Audio tagging (AT) aims at putting one or several tags on a sound clip. The tags are the sound events that occur in the audio clip, for example, “speech”, “television”, “percussion”, “bird singing”, and so on. Audio tagging has many applications in areas such as information retrieval [1], sound classification [2] and recommendation system [3].

Many frequency domain audio features such as mel-frequency cepstrum coefficients (MFCCs) [4], Mel filter banks feature (MFBs) [5] and spectrogram [6] have been used for speech recognition related tasks [7] for many years. However, it is unclear how these features perform on the non-speech audio processing tasks. Recently MFCCs and the MFBs were compared on the audio tagging task [8] and the MFBs can get better performance in the framework of deep neural networks. The spectrogram has been suggested to be better than the MFBs in the sound event detection task [9], but has not yet been investigated in the audio tagging task.

Besides the frequency domain audio features, processing sound from the raw time domain waveforms, has attracted a lot of attentions recently [10], [11], [12]. However, most of this works are for speech recognition related tasks; there are few works investigating raw waveforms for environmental audio analysis. For common signal processing steps, the short time Fourier transform (STFT) is always adopted to transform raw waveforms into frequency domain features using a set of Fourier basis. Recent research [10] suggests that the Fourier basis sets may not be optimal and better basis sets can be learned from raw waveforms directly using a large set of audio data. To learn the basis automatically, convolutional neural network (CNN) is applied on the raw waveforms which is similar to CNN processing on the pixels of the image [13]. Processing raw waveforms has seen many promising results on speech recognition [10] and generating speech and music [14], with less research in non-speech sound processing.

Most audio tagging systems [8], [15], [16], [17] use mono channel recordings, or simply average the multi-channels as the input signal. However, using this kind of merging strategy disregards the spatial information of the stereo audio. This is likely to decrease recognition accuracy because the intensity and phase of the sound received from different channels are different. For example, kitchen sound and television sound from different directions will have different intensities on different channels, depending on the direction of the sources. Multi-channel signals contain spatial information which could be used to help to distinguish different sound sources. Spatial features have been demonstrated to improve results in scene classification [18] and sound event detection [19]. However, there is little work using multi-channel information for the audio tagging task.

Our main contribution in this paper includes three parts. First, we show experimental results on different features including MFBs and spectrogram as well as the raw waveforms on the audio tagging task of the DCASE 2016 challenge. Second, we propose a convolutional gated recurrent neural network (CGRNN) which is the combination of the CNN and the gated recurrent unit (GRU) to process non-speech sounds. Third, the spatial features are incorporated in the hidden layer to utilize the location information. The work is organized as follows: in Section II, the proposed CGRNN is presented for audio tagging. In section III, the spatial features will be illustrated and incorporated into the proposed method. The experimental setup and results are shown in Section IV and Section V. Section VI summarizes the work and foresees the future work.

II. Convolutional Gated Recurrent Network for audio tagging

Neural networks have several types of structures: the most common one is the deep feed-forward neural network. Another popular structure is the convolutional neural network (CNN),

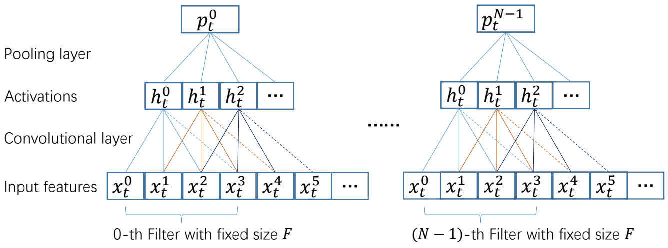
p_t^0
p_t^{N-1}
Pooling layer
Activations
h_t^0 h_t^1 h_t^2 ···
Convolutional layer
......
Input features
x_t^0 x_t^1 x_t^2 x_t^3 x_t^4 x_t^5 ···
0-th Filter with fixed size F
$(N-1)$-th Filter with fixed size F

Fig. 1. The structure of the one-dimension CNN which consists of one convolutional layer and one max-pooling layer. N filters with a fixed size F are convolved with the one dimensional signal to get outputs $p_t^i\{i = 0, \cdots, (N-1)\}$. x_t^i means the i-th dimension feature of the current frame.

which is widely used in image classification [20], [13]. CNNs can extract robust features from pixel-level values for images [13] or raw waveforms for speech signals [10]. Recurrent neural network is the third structure which is often used for sequence modeling, such as language models [21] and speech recognition [22]. In this section, we will introduce the convolutional neural network and the recurrent neural network with gated recurrent units.

A. One dimension convolutional neural network

Audio or speech signals are one dimensional. Fig. 1 shows the structure of a one-dimension CNN which consists of one convolutional layer and one max-pooling layer. N filters with a fixed size F are convolved with the one dimensional signal to get outputs $p_i^t\{i = 0, \cdots, (N-1)\}$. Given that the dimension of the input features was M, the activation h of the convolutional layer would have $(M-F+1)$ values. The max-pooling size is also $(M-F+1)$ which means each filter will give one output value. This is similar to speech recognition work [10] where CNN has been used to extract features from the raw waveform signal. The way of each filter producing one value can also be explained as a global pooling layer which is a structural regularizer that explicitly enforces feature maps to be confidence maps of meaningful feature channels [23]. So N activations are obtained as the robust features from the basic features. As for the input feature size M, a short time window, e.g., 32 ms, was fed into the CNN each time. The long-term pattern will be learned by the following recurrent neural network. As for the filter size or kernel size, a large receptive field is set considering that only one convolutional layer is designed in this work. For example, $F = 400$ and $M = 512$ are set in [10]. If the input feature was raw waveforms, each filter of the CNN was actually learned as a finite impulse response (FIR) filter [12]. If the spectrograms or mel-filter banks were fed into the CNN, the filtering was operated on the frequency domain [24] to reduce the frequency variants, such as the same audio pattern but with different pitches.

B. Gated recurrent unit based RNN

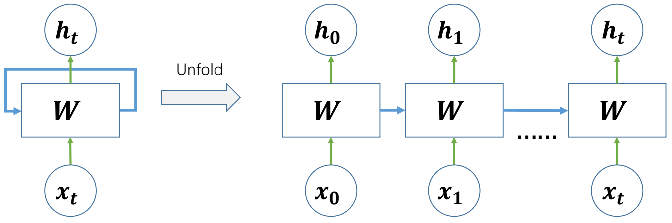

Fig. 2. The structure of the simple recurrent neural network and its unfolded version. The current activation $\mathbf{h}_t$ is determined by the current input $\mathbf{x}_t$ and the previous activation $\mathbf{h}_{t-1}$.

Recurrent neural networks have recently shown promising results in speech recognition [22]. Fig. 2 shows the basic idea of the RNN. The current activation $\mathbf{h}_t$ is determined by the current input $\mathbf{x}_t$ and the previous activation $\mathbf{h}_{t-1}$. RNN with the capability to learn the long-term pattern is superior to a feed-forward DNN, because a feed-forward DNN is designed that the input contextual features each time are independent. The hidden activations of RNN are formulated as:

$$\mathbf{h}_t = \varphi(\mathbf{W}^h\mathbf{x}_t + \mathbf{R}^h\mathbf{h}_{t-1} + \mathbf{b}^h) \tag{1}$$

However, a simple recurrent neural network with the recurrent connection only on the hidden layer is difficult to train due to the well-known vanishing gradient or exploding gradient problems [25]. The long short-term memory (LSTM) structure [26] was proposed to overcome this problem by introducing input gate, forget gate, output gate and cell state to control the information stream through the time. The fundamental idea of the LSTM is memory cell which maintains its state through time [27].

As an alternative structure to the LSTM, the gated recurrent unit (GRU) was proposed in [28]. The GRU was demonstrated to be better than LSTM in some tasks [29], and is formulated as follows [27]:

$$\mathbf{r}_t = \delta(\mathbf{W}^r\mathbf{x}_t + \mathbf{R}^r\mathbf{h}_{t-1} + \mathbf{b}^r) \tag{2}$$

$$\mathbf{z}_t = \delta(\mathbf{W}^z\mathbf{x}_t + \mathbf{R}^z\mathbf{h}_{t-1} + \mathbf{b}^z) \tag{3}$$

$$\tilde{\mathbf{h}}_t = \varphi(\mathbf{W}^h\mathbf{x}_t + \mathbf{r}_t \odot (\mathbf{R}^h\mathbf{h}_{t-1}) + \mathbf{b}^h) \tag{4}$$

$$\mathbf{h}_t = \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \tilde{\mathbf{h}}_t \tag{5}$$

where $\mathbf{h}_t$, $\mathbf{r}_t$ and $\mathbf{z}_t$ are hidden activations, reset gate values and update gate values at frame t, respectively. The weights applied to the input and recurrent hidden units are denoted as $\mathbf{W}^*$ and $\mathbf{R}^*$, respectively. The biases are represented by $\mathbf{b}^*$. The functions $\delta(\cdot)$ and $\varphi(\cdot)$ are the sigmoid and tangent activation functions. Compared to the LSTM, there is no separate memory cell in the GRU. The GRU also does not have an output gate, and combines the input and forget gates into an update gate $\mathbf{z}_t$ to balance between the previous activation $\mathbf{h}_{t-1}$ and the update activation $\tilde{\mathbf{h}}_t$ shown in Eq. (5). The reset gate $\mathbf{r}_t$ can decide whether or not to forget the previous activation (shown in Eq. (4)). $\odot$ in Eq. (4) and (5) represents the element-wise multiplication.

C. Convolutional Gated Recurrent Network for audio tagging

Fig. 3 shows the framework of a convolutional gated recurrent neural network for audio tagging. The CNN is regarded as the feature extractor along the short window (e.g., 32ms) from the basic features, e.g., MFBs, spectrograms or raw waveforms. Then the robust features extracted are fed into the GRU-RNN to learn the long-term audio patterns. For the audio tagging task, there is a lot of background noise, and acoustic events may occur repeatedly and randomly along the whole chunk (without knowing the specific frame locations). The CNN can help to extract robust features against the background noise by the max-pooling operation, especially for the raw waveforms. Since the label of the audio tagging task is at the chunk-level rather than the frame-level, a large number of frames of the context were fed into the whole framework. The GRU-RNN can select related information from the long-term context for each audio event. To also utilize the future information, a bi-directional GRU-RNN is designed in this work. Finally the output of GRU-RNN is mapped to the posterior of the target audio events through one feed-forward neural layer, with sigmoid output activation function. This framework is flexible enough to be applied to any kinds of features, especially for raw waveforms. Raw waveforms have lots of values, which leads to a high dimension problem. However the proposed CNN can learn on short windows like the short-time Fourier transform (STFT) process, and the FFT-like basis sets or even mel-like filters can be learned for raw waveforms. Finally, one-layer feed-forward DNN gets the final sequence of GRUs to predict the posterior of tags.

Binary cross-entropy is used as the loss function in our work, since it was demonstrated to be better than the mean squared error in [8] for labels with zero or one values. The loss can be defined as:

$$E = -\sum_{n=1}^{N} \|\mathbf{T}_n \log \hat{\mathbf{T}}_n + (1-\mathbf{T}_n)\log(1-\hat{\mathbf{T}}_n)\| \tag{6}$$

$$\hat{\mathbf{T}}_n = (1+\exp(-\mathbf{O}))^{-1} \tag{7}$$

where E is the binary cross-entropy, $\hat{\mathbf{T}}_n$ and $\mathbf{T}_n$ denote the estimated and reference tag vector at sample index n, respectively. The DNN linear output is defined as $\mathbf{O}$ before the sigmoid activation function is applied. Adam [30] is used as the stochastic optimization method.

III. Spatial features incorporated for audio tagging

Spatial features can often offer additional cues to help to solve signal processing problems. Many spatial features can be used for audio tagging, such as interaural phase differences or interaural time differences (IPD or ITD) [31], interaural level differences (ILD) [31]. The recordings of the audio tagging task of DCASE 2016 challenge are recorded in the home scenes. There are some TV, child speech, adult speech audio events. The spatial features potentially give additional information to analyze the content of the audio, e.g., recognizing the TV audio event by knowing the specific direction of the TV sound. The IPD and ILD are defined as:

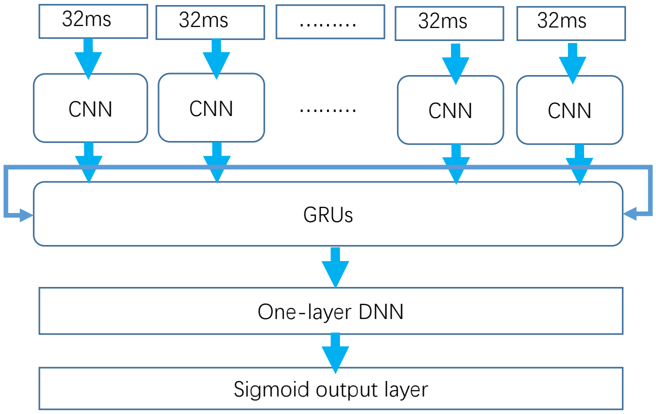

Fig. 3. The framework of convolutional gated recurrent neural network for audio tagging.

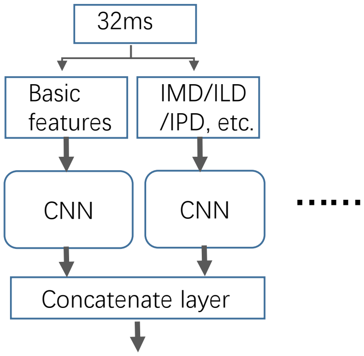

Fig. 4. The structure of incorporating the spatial features (IMD/ILD/IPD, etc.) using an additional CNN. Then the activations learned from the basic features and the activations learned from the spatial features are concatenated to be fed into the GRU-RNN shown in Fig. 3.

$$ILD(t,k) = 20\log_{10}\left|\frac{X_{\text{left}}(t,k)}{X_{\text{right}}(t,k)}\right| \tag{8}$$

$$IPD(t,k) = \angle\left(\frac{X_{\text{left}}(t,k)}{X_{\text{right}}(t,k)}\right) \tag{9}$$

where $X_{left}(t,k)$ and $X_{right}(t,k)$ denote the left channel and right channel complex spectrum of the stereo audio. The operator $|\cdot|$ takes the absolute of the complex value, and $\angle(\cdot)$ finds the phase angle. In this work, we also define interaural magnitude differences (IMD) which is similar to the ILD. IMD is defined in linear domain while ILD is defined in logarithmic domain.

$$IMD(t,k) = |X_{left}(t,k)| - |X_{right}(t,k)| \tag{10}$$

Fig. 4 shows the structure of incorporating the spatial features (IMD/ILD/IPD, etc.) using an additional CNN. Then the activations learned from the basic features and the activations learned from the spatial features are concatenated to be fed into the GRU-RNN plotted in Fig. 3. The audio files of the audio

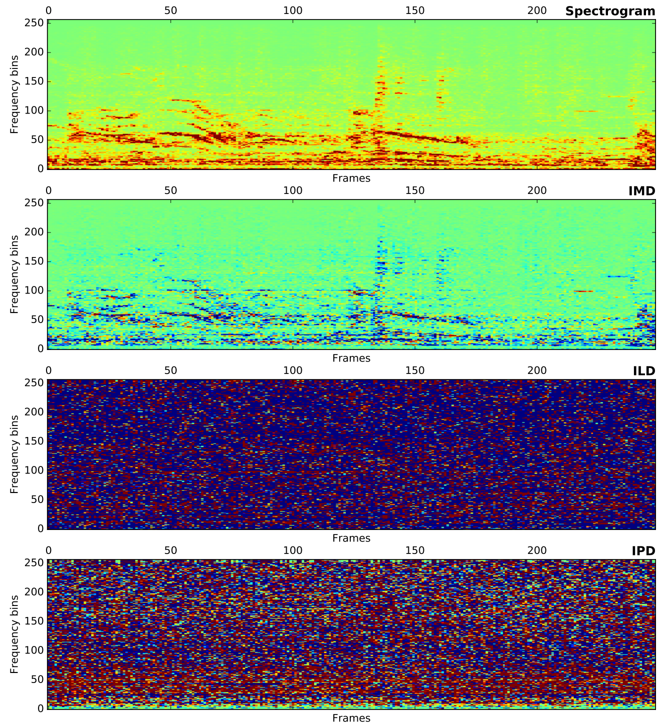

Fig. 5. The spectrogram, IMD, ILD and IPD of ac recording from the audio tagging task of DCASE 2016 challenge. The labels of this audio are "child speech" and "TV sound".

tagging task of the DCASE 2016 challenge are recorded in a domestic home environment. There are severe reverberation, high-level background noise and multiple acoustic sources. These factors might influence the effectiveness of IPD and ILD. Fig. 5 shows the spectrogram, IMD, ILD and IMD of one recording from the audio tagging task of DCASE 2016 challenge. The IMD appears to have some meaningful patterns while the ILD and the IPD seem to be random which would lead to the training difficulties of the classifier. From our empirical experiments, IPD and ILD do not appear to help to improve the classifier performance while IMD is beneficial. Similar results were reported in [19] where IPD was found not to be helpful for the sound event detection in home scenes but helpful for the event detection of residential areas. This may be because residential areas are open areas with less reverberation than indoor home environments. Hence we will use IMD as our spatial features in this work. The filter size of CNNs learned on the IMD is set the same with the related configuration for the spectrogram.

IV. Data set and experimental setup

A. DCASE 2016 audio tagging challenge

We conducted our experiments based on the DCASE 2016 audio tagging challenge [32]. This audio tagging task consists of the five-fold development set and the evaluation set, which are built based on the CHiME-home dataset [33]. The audio recordings were made in a domestic environment [34]. The audio data are provided as 4-second chunks at 48kHz sampling rates in stereo mode. We downsampled them into 16kHz sampling rate.

For each chunk, three annotators gave three labels, namely multi-label annotations. Then the discrepancies among annotators are reduced by conducting a majority vote. The annotations are based on a set of seven audio events as presented in Table I. A detailed description of the annotation procedure is provided in [33].

TABLE I
Seven audio events used as the reference labels.

audio event	Event descriptions
'b'	Broadband noise
'c'	Child speech
'f'	Adult female speech
'm'	Adult male speech
'o'	Other identifiable sounds
'p'	Percussive sound events, e.g. footsteps, knock, crash
'v'	TV sounds or Video games

TABLE II
The configurations for the five-fold development set and the final evaluation set of the DCASE 2016 audio tagging task.

Fold index of development set	#Training chunks	#Test chunks
0	4004	383
1	3945	442
2	3942	463
3	4116	271
4	4000	387
Evaluation set	4387	816

B. Experimental setup

In the experiments below, we follow the standard specification of the DCASE 2016 audio tagging task [32], On the development set, we use the official five folds for cross-validation. Table II shows the number of chunks in the training and test set used for each fold. The number of final evaluation configuration is also listed.

The parameters of the networks are tuned based on the heuristic experience. All of the CNNs have 128 filters or feature maps. Following [10], the filter size for MFBs, spectrograms and raw waveforms are 30, 200, and 400, respectively. These parameters can form a large receptive field for each type of basic features considering that only one convolutional layer was designed. The CNN layer is followed by three RNN layers with 128 GRU blocks. One feed-forward layer with 500 ReLU units is finally connected to the 7 sigmoid output units. We pre-process each audio chunk by segmenting them using a 32ms sliding window with a 16ms hop size, and converting each segment into 40-dimension MFBs, 257-dimension spectrograms and 512-dimension raw waveforms. For performance evaluation, we use equal error rate (EER) as the main metric which is also suggested by the DCASE 2016 audio tagging challenge. EER is defined as the point of equal false negative (FN) rate and false positive (FP) rate [35]. The source codes for this paper can be downloaded from Github[1].

C. Compared methods

We compared our methods with the state-of-the-art systems. Lidy-CQT-CNN [15] and Cakir-MFCC-CNN [16] won the first and the second prize of the DCASE2016 audio tagging challenge [32]. They both used convolutional neural networks (CNN) as the classifier. We also compare to our previous method [8] which demonstrated the state-of-the-art performance using de-noising auto-encoder (DAE) to learn robust features.

V. Experimental results and analysis

In this section, the IMD effectiveness will be firstly evaluated on the development set of Task 4 of the DCASE 2016 challenge among the different features, i.e., spectrograms, MFBs and raw waveforms. Then the final evaluation will be presented by comparing with the state-of-the-art methods on the evaluation set of Task 4 of the DCASE 2016 challenge.

A. The effectiveness of the IMD

Table III shows the EER comparisons on seven labels among the spectrogram, the raw waveform and the MFB systems with or without the IMD information, which are evaluated on the development set of the DCASE 2016 audio tagging challenge. Firstly, we can compare the proposed convolutional gated recurrent neural networks on spectrograms, raw waveforms and MFBs. Spectrograms are better than the MFBs perhaps because the spectrogram has more detailed frequency information compared with the MFB. For example, spectrograms are much better than MFBs on child speech (denoted as 'c') and female speech (denoted as 'f') where a lot of high frequency information exists. The raw waveforms are worse than the spectrograms and the MFBs. One possible reason is that the learned FIR filters are not stable when the whole training set is small (about 3.5 hours of audio in this work). The same explanation was given in [11] on the speech recognition task. [10] shows that raw waveforms can get better recognition accuracy with the mel-spectra on 2000 hours Google voice search data.

With the help of the IMD spatial features, the EER are improved compared to all of the corresponding basic features alone. The raw waveforms with IMD can even get comparable results with the spectrograms and the MFBs. The MFB-IMD combination is slightly better than Spec-IMD, which may be because the IMD is calculated from the left and right spectrograms. The IMD has some common information with the spectrograms which can be seen from Fig. 5. However, the IMD is more complementary for the MFBs and the raw waveforms. The previous best performance on the development set of the DCASE 2016 audio tagging challenge was obtained in our recent work using denoising auto-encoder [8] with 0.126 EER, but here we get better performance with 0.10 EER.

TABLE III
EER comparisons on seven labels among the spectrogram, the raw waveform and the MFB systems with or without the IMD information, which are evaluated on the **development set** of the DCASE 2016 audio tagging challenge.

Dev set	c	m	f	v	p	b	o	ave
Spec	0.121	0.085	0.155	0.025	0.138	0.017	0.231	0.110
RAW	0.157	0.085	0.156	0.028	0.139	0.059	0.263	0.127
MFB	0.145	0.086	0.167	0.024	0.133	0.037	0.239	0.119
Spec-IMD	**0.120**	**0.080**	0.143	**0.012**	0.115	0.023	0.232	0.104
RAW-IMD	0.135	0.085	0.164	0.014	**0.108**	**0.006**	0.231	0.106
MFB-IMD	0.125	0.086	**0.140**	**0.012**	0.110	0.011	**0.230**	**0.102**

TABLE IV
EER comparisons on seven labels among Lidy-CQT-CNN [15], Cakir-MFCC-CNN [16], DAE-DNN [8], and the proposed systems on the spectrogram, the raw waveform and the MFB systems with the IMD information, which are evaluated on the **final evaluation set** of the DCASE 2016 audio tagging challenge.

Eval set	c	m	f	v	p	b	o	ave
Cakir [16]	0.250	0.159	0.250	0.027	0.208	0.022	0.258	0.168
Lidy [15]	0.210	0.182	0.214	0.035	0.168	0.032	0.320	0.166
DAE [8]	0.210	0.149	0.207	**0.022**	0.175	0.014	0.256	0.148
RAW-IMD	0.189	0.152	0.200	0.053	0.156	**0.010**	**0.236**	0.142
Spec-IMD	0.166	0.165	**0.143**	0.024	**0.123**	0.034	0.250	0.129
MFB-IMD	**0.150**	**0.145**	**0.143**	0.031	0.135	0.013	0.248	**0.123**

B. Overall evaluations

Table IV presents the EER comparisons on seven labels among Lidy-CQT-CNN [15], Cakir-MFCC-CNN [16], our previous DAE-DNN [8], and the proposed systems on the spectrogram, the raw waveform and the MFB systems with the IMD information, which are evaluated on the final evaluation set of the DCASE 2016 audio tagging challenge. The denoising auto-encoder [8] was our recent work which can outperform the leading system in the DCASE 2016 audio tagging challenge, namely Lidy-CQT-CNN [15]. Our proposed convolutional gated recurrent neural network incorporating the IMD features in this work gives further improved performance. The MFB-IMD obtains the best performance with 0.123 EER which is the state-of-the-art performance on the evaluation set of the DCASE 2016 audio tagging challenge.

VI. Conclusion

In this paper, we propose a convolutional gated recurrent neural network (CGRNN) to learn on the mel-filter banks (MFBs), the spectrograms and even the raw waveforms. The spatial features, namely the interaural magnitude difference (IMDs), are incorporated into the framework and are demonstrated to be effective to further improve the performance. Spectrogram gives better performance than MFBs without the spatial features. However the MFBs with the IMDs can get the minimal EER, namely 0.102, on the development set of the DCASE 2016 audio tagging challenge. Raw waveforms give comparable performance on the development set. Finally, on the evaluation set of the DCASE 2016 audio tagging challenge, our proposed MFB-IMD system can get the state-of-the-art performance with 0.123 EER. It is still interesting to further explore why the MFB-IMD system is better than the Spec-IMD system in our future work. In addition, we will also

[1] https://github.com/yongxuUSTC/cnn_rnn_spatial_audio_tagging

investigate the proposed framework to model raw waveforms on larger training datasets to learn more robust filters.

Acknowledgment

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) of the UK under the grant EP/N014111/1. Qiuqiang Kong is partially supported by the China Scholarship Council (CSC).

References

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[2] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: an IEEE AASP challenge," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4.

[3] P. Cano, M. Koppenberger, and N. Wack, "Content-based music audio recommendation," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, 2005, pp. 211–212.

[4] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1, pp. 73–76.

[5] C. Nadeu, D. Macho, and J. Hernando, "Time and frequency filtering of filter-bank energies for robust hmm speech recognition," *Speech Communication*, vol. 34, no. 1, pp. 93–114, 2001.

[6] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech communication*, vol. 25, no. 1, pp. 117–132, 1998.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[8] 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," *arXiv preprint arXiv:1607.03681v2*, 2016.

[9] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 1, 2015.

[10] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. Interspeech*, 2015.

[11] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *Proc. of Interspeech*, 2014, pp. 890–894.

[12] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in lvcsr," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[14] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[15] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification and domestic audio tagging," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016)*. [Online]. Available: http://www.ifs.tuwien.ac.at/~schindler/pubs/DCASE2016b.pdf

[16] E. Cakır, T. Heittola, and T. Virtanen, "Domestic audio tagging with convolutional neural networks," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016)*. [Online]. Available: https://www.cs.tut.fi/sgn/arg/dcase2016/documents/challenge_technical_reports/Task4/Cakir_2016_task4.pdf

[17] S. Yun, S. Kim, S. Moon, J. Cho, and T. Kim, "Discriminative training of gmm parameters for audio scene classification and audio tagging," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016)*. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/documents/challenge_technical_reports/Task4/Yun_2016_task4.pdf

[18] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for dcase-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks," Tech. Rep., DCASE2016 Challenge, Tech. Rep., 2016.

[19] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," *IEEE Detection and Classification of Acoustic Scenes and Events workshop*, 2016. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/documents/workshop/Adavanne-DCASE2016workshop.pdf

[20] D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 1, 2011, p. 1237.

[21] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, 2010.

[22] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 6645–6649.

[23] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[24] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition." in *Proc. Interspeech*, 2013, pp. 3366–3370.

[25] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks." *ICML*, pp. 1310–1318, 2013.

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

[27] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5140–5144.

[28] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

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

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] J. Blauert, *Spatial Hearing: the Psychophysics of Human Sound Localization*. MIT press, 1997.

[32] http://www.cs.tut.fi/sgn/arg/dcase2016/task-audio-tagging.

[33] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. Plumbley, "CHiME-home: A dataset for sound source recognition in a domestic environment," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015, pp. 1–5.

[34] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proceedings of Interspeech*, 2010, pp. 1918–1921.

[35] K. P. Murohy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.