MULTICHANNEL AUDIO CLASSIFICATION WITH NEURAL NETWORKS USING SCATTERING TRANSFORM

S. Y. Ezra[1], *Y. Gershon*[1], *U. Levi*[1], *M. Palatin*[1], *A. Raveh*[1], *S. Sheer*[1], *Y. Doweck*[1], *and A. Amar*[2]

1. Signal Processing Department, National Research Center, Haifa, Israel
2. Faculty of Electrical Engineering, Technion, Haifa, Israel

ABSTRACT

This technical paper presents an approach for the 2018 acoustic scene classification challenge (DCASE 2018) task 5. A sequence of audio segments are observed by an array with 4 microphones. The task is to suggest a multichannel processing to classify the audio signals to one of 9 pre-defined classes. The proposed approach combines a deep neural network with scattering transform. Each audio segment is first represented by two layers of scattering transform. The 4 denoised transforms of each of the two layers are combined together. Each of the fused layers are processed in parallel by two neural networks (NN) architectures, RESNET and long short-term memory (LSTM) network, with a joint fully connected layer.

Index Terms: Scattering transform, neural networks, RESNET, LSTM.

I. INTRODUCTION

Acoustic scene classification (ASC) is a subtask of the more general computational auditory scene analysis area of research [1]–[3]. It attracts the attention of researchers in machine learning and has been applied into surveillance, navigation and context-aware services, where computational algorithms try to outperform humans when discriminating between sound scenes. Recently, there is an increased interest in smart environments, where the aim is automatically to understand the home scene using different types of sensors including microphones .

In the DCASE 2018 challenge task 5 [4], the goal is to classify multi-channel audio 10 seconds segments observed by an array with four microphones, which are located at an unknown position in a room, into one of pre-defined classes. The classes are daily activities (e.g., cooking, working, watching TV, etc.). In the data set no overlapping of activities are present, that is, each segment contains only one of the classes. Since the positions of the 4 microphone arrays at the test set are unknown in advance, and may be different from the positions of the 7 microphone arrays in the training set, classifying based on the absolute locations of the sounds may not be beneficial. Therefore, the focus of the task is to exploit other multichannel processing approaches which are independent of the sensors' locations.

The basic idea of our proposed approach is to represent the audio segment acquired by each microphone using scattering transform, and then combine the transforms of all channels as an input to a neural network (NN). The widely used representation of audio signals is the Mel Frequency Cepstral Coefficients (MFCC) [5], usually used for speech signals. The features are based on averaging the signal in the Short Time Fourier Transform (STFT) with Mel scale filters which are logarithmically spaced in the frequency domain, and then applying a cosine transform. However, the MFCC representation is based on the assumption that the signal is stationary over a short time interval (usually, between 20 msec to 40 msec). Thus, it lacks the ability to focus on the non-stationarity of the audio over larger time frames. The scattering transform [6], [7] overcomes this shortcoming by considering a much larger time frames for processing the data using 2 layers of wavelets filters, where each layer is followed by the modulus operator. The MFCC features are similar to the coefficients of the first order of the scattering transform. Therefore, the representation of signals using the first and the second orders of the scattering transform extends the MFCC representation.

Mathematically speaking, denote by $x_k(t)$ the observed sound signal by the kth microphone expressed as

$$x_k(t) = s_k(t) + v(t)\,, 0 \le t \le T\,, k = 1, \ldots, 4 \qquad (1)$$

where $s_k(t)$ is the audio segment associated with a specific (but unknown) class, $v(t)$ is the additive noise and interference, and $T = 10$ [sec] is the observation interval. The observed signal is processed in overlapping time frames of length T' producing the signals $y_k(t)$, and then processed by the scattering transform which is a cascade of wavelet convolutions and modulus operators. The first scattering transform is given by

$$S_{1,k}(t, \lambda_1) = |y_k(t) * \psi_{\lambda_1}| * \phi(t) \qquad (2)$$

where $\phi(t)$ is a low pass filter with a frequency bandwidth $2\pi/T'$. The wavelet filter bank is defined by

$$\psi_\lambda(t) = \lambda \psi(\lambda t) \qquad (3)$$

where the wavelet $\psi(t)$ is a band pass filter with a central frequency normalized to one, $\lambda = 2^{j/Q}$, j is a positive

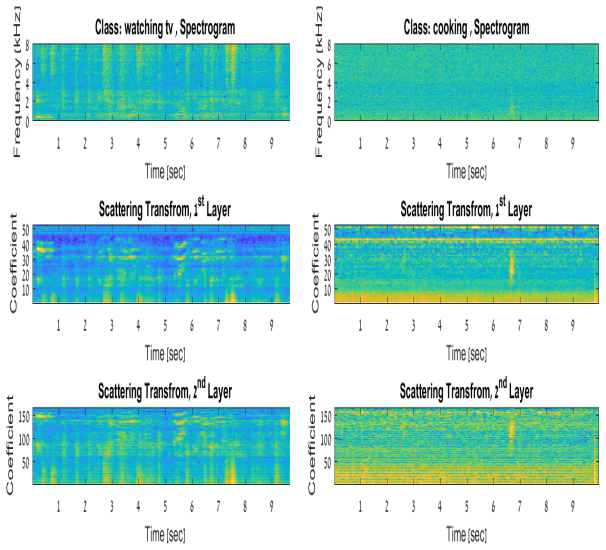

Fig. 1. An example of the spectrogram of an audio segment (top), the first layer of the scattering transform (center), and second layer of the scattering transform (bottom). The left plots are associated an example of an audio segment of the class "working", while the right plots are associated with an example of an audio segment of the class "cooking".

integer, and Q is the number of wavelets per octave. The bandwidth of the wavelet is in the order of $1/Q$. As a result, the filters are logarithmically spaced in he frequency domain. The second order scattering transform is

$$S_{2,k}(t,\lambda_1,\lambda_2) = ||y_k(t) * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi(t) \qquad (4)$$

The top plot in Figure 1 presents the spectrogram of the an audio segment associated with the "Watching TV" class and "Cooking class". The sampling frequency of the signal is 16 [KHz] and the duration of the segment is 10 [sec]. The first and second layers of the the scattering transform of the audio segment are also presented.

The output of this representation contain 4 matrices, $\{\mathbf{S}_{1,k}\}_{k=1}^4$ associated with the first order scattering transform, and 4 matrices (with the same number of columns but with larger number of rows) $\{\mathbf{S}_{2,k}\}_{k=1}^4$ associated with the second order scattering transform. The 4 matrices associated with each of the two scattering transform layers are fused by averaging their values, i.e.,

$$\bar{\mathbf{S}}_1 = \frac{1}{4}\sum_{k=1}^{4}\mathbf{S}_{1,k} \ , \ \bar{\mathbf{S}}_2 = \frac{1}{4}\sum_{k=1}^{4}\mathbf{S}_{2,k} \qquad (5)$$

An optional step is to denoise these matrices using singular value decomposition (SVD), or any other denoising procedure. For example, the matrix associated with the first layer

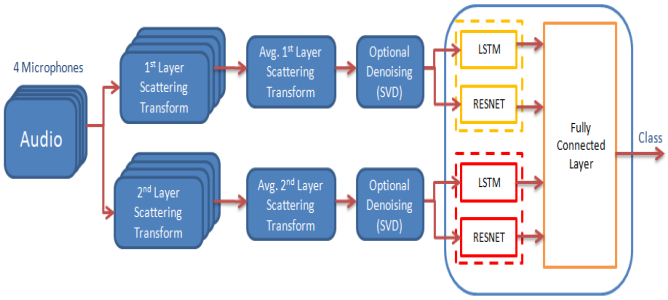

Fig. 2. The block diagram of the proposed approach.

is approximated using SVD as (the matrix associated with the second layer is approximated similarly)

$$\bar{\mathbf{S}}_1 \cong \sum_{i=1}^{K}\sigma_i^2\mathbf{u}_i\mathbf{v}_i^T$$

where $\mathbf{u}_i$, $\mathbf{v}_i$, and σ_i^2 are the left singular vectors, the right singular vectors and the singular values of the decomposition, respectively, and the K is the number of the dominant singular values determined as

$$K = \text{argmin}_k \frac{\sum_{i=1}^{k}\sigma_i^2}{\sum_{k=1}^{K}\sigma_k^2} < \eta \qquad (6)$$

where η is a predefined threshold (e.g., $\eta = 0.9$). These averaged matrices are used as inputs to NNs as explained next. The block diagram of the proposed approach is presented in Fig. 2.

II. NEURAL NETWORK ARCHITECTURE

Each of the averaged scattering transforms $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$ are used as input to two branches of the NN. The first branch is a modified RESNET architecture [9] with 32 residual convolutional layers with batch normalization (Fig. 3). The second branch is an LSTM architecture which sequentially learns the context between time frames in a given scattering transform representation of an audio segment. A time averaging is performed following the LSTM layer. The complete architecture of the proposed ANN is detailed in Fig. 3 and in Fig. 4. The four NN branches have two joint fully connected layers. The final classification is the output of a single softmax layer.

We also propose second architecture with a small modification to the RESNET branch. replace the first 7×7 convolutional layer with 1D convolution layer over the feature dimension, this modification causes a reduction of the 3×3 convolutional layer to 1×3 .

III. EVALUATION RESULTS

The development data set contains $72,984$ audio segments each of 10 seconds length, partitioned into 4 different training and test groups (termed as folds). The data set consists of

	Total (without denoising)		Total (with denoising)	
class	1d	2d	1d	2d
Absence	83.59	83.95	83.12	84.73
Cooking	95.93	95.47	94.46	94.20
Dishwashing	82.10	78.00	76.64	73.31
Eating	88.87	89.68	86.16	87.13
Other	55.36	55.88	52.35	54.22
Social	94.99	93.97	94.55	94.06
Vacuum	100.0	100.0	100.0	100.0
Tv	99.74	99.40	99.73	99.77
Working	84.61	85.22	83.94	85.09
F_1	**87.24**	**86.84**	**85.67**	**85.82**

Table I. Classification F_1 score for 1d convolutional layer and for 2d convolutional layer with and without denoising.

Class	Absence	Cooking	Dishwashing	Eating	Other	Social	Vacuum	Tv	Working
Absence	**0.853**	0.001	0	0.007	0.171	0.003	0	0	0.093
Cooking	0	**0.968**	0.022	0	0.013	0	0	0	0
Dishwashing	0.001	0.024	**0.902**	0.003	0.004	0	0	0	0
Eating	0	0	0.017	**0.834**	0.008	0	0	0.001	0.001
Other	0.007	0.007	0.025	0.024	**0.456**	0.002	0	0	0.013
Social	0.001	0	0.02	0	0.017	**0.968**	0	0	0
Vacuum	0	0	0	0	0	0	**1**	0	0
Tv	0	0	0	0	0	0.015	0	**0.998**	0
Working	0.138	0	0.014	0.101	0.331	0.012	0	0	**0.892**

Table II. Confusion matrix of the development set for fold_1 for the case of a 2d convolutional layer .

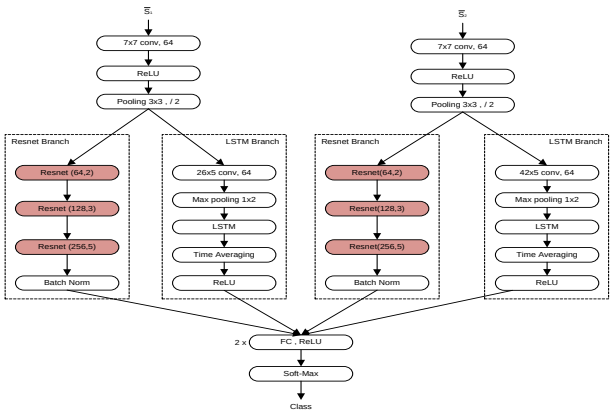

Fig. 3. Schematic description of the proposed neural network architecture.

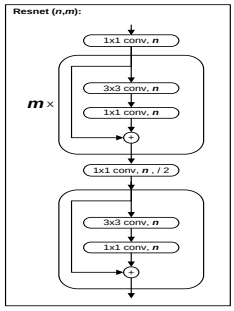

Fig. 4. Description of the building block of the RESNET branch detailing the blocks marked in color in Fig. 3.

9 different classes which are Cooking, Dishwashing, Eating, Social activity, Vacuum cleaning, Watching TV, Working (typing, mouse click, ...), Absence (nobody present in the room), and other (a person is present but not doing any relevant activity). The model is evaluated according to a four-fold cross-validation scheme. Per-class F_1 scores are computed on the test set for each audio segment independently. Finally, the overall F_1 score is calculated by averaging the F_1 scores of the different four folds.

The confusion matrix for fold_1 is presented in Table II. As can be seen, most classes are classified with high accuracy except class "Other", which its associated segments are also classified to the class "Absence" and the class "working". All the audio segments of the classes "TV" and "Social" are correctly classified, which emphasize that as expected scattering transform succeeds in classifying these kind of speech signals, but also succeeds to classify other non-speech audio signals.

IV. CONCLUSION

We proposed a method to classify audio segments observed by an array with 4 microphones into one of 9 predefined classes. The method represents the data segment using two layers of scattering transform. The transforms are processed in parallel by RESNET and LSTM neural networks with a joint fully connected layer.

REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. Plumbley, Detection and Classification of Acoustic Scenes and Events, IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 17331746, 2015.

[2] M. Crocco, M. Cristani, A. Trucco, and V. Murino, Audio surveillance: A systematic review, ACM Computing Surveys (CSUR), 2016.

[3] K. J. Piczak, Environmental sound classification with convolutional neural networks IEEE International Workshop on Machine Learning for Signal Processing, 2015.

[4] G. Dekkers, P. Karsmakers, and L. Vuegen, "Monitoring of domestic activities based on multichannel acoustics", http://dcase.community, 2018.

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

[6] J. Anden and S. Mallat, Deep scattering spectrum, IEEE Transactions on Signal Processing, vol. 62, no. 16, pp. 4114–4128, 2014.

[7] S. Mallat, Group invariant scattering, Communications on Pure and Applied Mathematics, vol. 65, no. 10, pp. 1331–1398, 2012.

[8] S. H. Bae, I. Choi, and N. S. Kim, Acoustic scene classification using parallel combination of LSTM and CNN, in De- tection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), Budapest, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.