AN ENSEMBLE SYSTEM FOR DOMESTIC ACTIVITY RECOGNITION

Technical Report

Huaping Liu[1], *Feng Wang*[1], *Xinzhu Liu*[2], *Di Guo*[1], *Fuchun Sun*[1]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing, China,
[2] Department of Computer Science and Technology, Jilin University, Jilin, China,

ABSTRACT

As one of the most important sensing modalities, the acoustics is becoming more and more popular in achieving our smart environment nowadays. In this challenge, we try to tackle the task of monitoring domestic activities based on multi-channel accoustics. Several acoustic features including the Mel-Spectrograms, MFCC and VGGish features are extracted from the raw audio and fused to train different deep neural networks. An ensemble system is then established with the trained models. The experimental results on the development dataset demonstrate that the proposed system has shown superior performance in recognizing domestic activities.

Index Terms— Mel-Spectrograms, MFCC, VGGish, ensemble

1. TASK AND DATASET DESCRIPTION

This technical report describes our implementation of "Monitoring of domestic activities based on multi-channel acoustics", Task5, D-CASE2018 Challenge[1] [2].

A person lives alone in a vacation home for a period of one week, where a net work of 13 microphone arrays are distributted. Each microphone array is annotated as a node in this report. There are 4 linearly arranged microphones in a node. With the collected continuous recordings, the problem of acoustic scene recognition from multi-channel audio segments is tackled in this task. It is focused on the daily dometic activities and a complete list of activities can be found in Table 1.

An overview of the problem is demonstrated in Figure 1. The input of the system is the multi-channel audio segments from a node and the system outputs the predicted domestic activity. For this task, 7 nodes in the combined living room and kitchen area are used. Figure 2 shows the floorplan of the recorded environment along with the position of the used nodes.

The dataset used in this task is a derivative of the SINS dataset[3][4]. The continuous recordings were split into audio segments of 10s. Each segment contains 4 channels from the 4 microphones in a node. The segment is provided with activity label and each segment only contains one activity class. As development set, approximately 200 hours of labeled data from 4 nodes are used. The position information of the nodes is unknown. As for the evaluation set, unlabeled data is provided from all nodes.

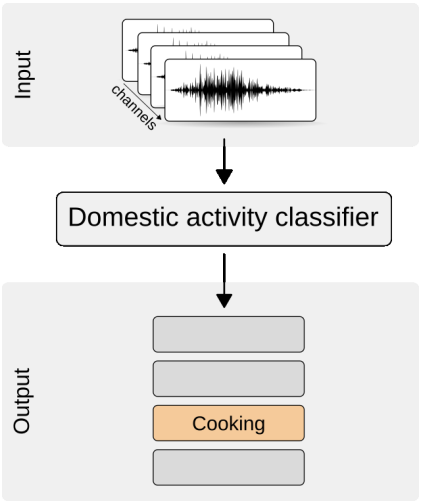

Figure 1: An overview of the task.

2. APPROACH

2.1. Feature extraction

In the task of acoustic scene classification (ASC), there are multiple methods to extract features from recorded audios. In this work, the length of the original audio is 10 senconds with a sample rate of 16 kHz. We combine the Mel-Spectrograms, MFCC and a VGGish features together in the proposed classification system.

- Mel-Spectrograms feature: The Mel-Spectrograms feature is calculated for each 10s audio segment with a window length of 0.04 senconds, a hop length of 0.02 seconds and 40 Mel-bands. Therefore, a 40×501 Mel-Spectromgrams feature matrix can be obtained.
- MFCC feature: The MFCC feature has been very popular in ASC for years. We calculate the MFCC feature with a hop length of 320 and 40 MFCC coefficients are selected. Then a MFCC feature matrix with the size of 40×501 is obtained for each audio segment.
- VGGish feature: Because of the superior performance of deep networks in audio classification task, we also employ a VGGish model [5] to extrac high level features from the audio segments. The 10s audio is equally divided into 10 parts and each has 1 second. The Mel-Spectrograms are regarded as the input of the model. To fit the input size of the VGGish model, a 64×96 Mel-Spectromgrams feature matrix is calculated for each 1 second audio segment, where 64 is the number of Mel-bands. The output of the model is a 128-dimension embedding.

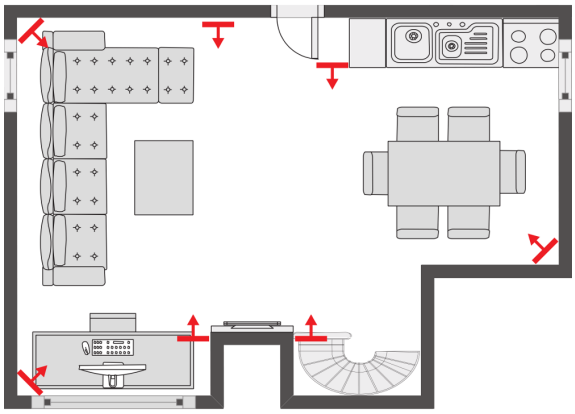

Figure 2: 2D floorplan of the combined kitchen and living room with the used nodes.

Table 1: Domestic Activities List

Activity
Absence (nobody present in the room)
Cooking
Dishwashing
Eating
Other (present but not doing any relevant activity)
Social activity (visit, phone call)
Vacuum cleaning
Watching TV
Working (typing, mouse click, ...)

2.2. Model architecture

Different classifiers are trained to recognize different domestic activities regarding to different features. For the Mel-Spectrograms and MFCC features, a CNN network is used, while an LSTM network is used for the VGGish feature.

- CNN network: The architecture of the CNN network can be seen in Figure 3.The CNN network is mainly composed of two convolutional layers followed by a fully connected layer. The input of the network is the Mel-Spectrograms and MFCC feature matrix, which is in the size of 40×501. The first convolutional layer contains 64 kernels with the size of 40×5 and the second convolutional layer contains 128 kernels with the size of 1×3. After a global max pooling layer, a 128 dimension feature vector is obtained. After a fully connected layer, the output of the network is the possibilities of the 9 activities annotated as $P_1 \in R^9$ for the Mel-Spectrograms and $P_2 \in R^9$ for the MFCC feature.
- LSTM network: The LSTM network is indicated in Figure 4, which has a many-to-one architecture. Since we have divided the original audio into equally 10 segments chronologically, the feature of each segment is fed into the network step by step. The output of the last LSTM is connected to a fully connected layer and the possibilities of the 9 activities annotated as $P_3 \in R^9$ can be obtained.

Figure 3: The architecture of the CNN network.

To obtain the ensemble results from above models, we average their possibilities as follows:

$$P = \frac{1}{3}\sum_{i=1}^{3} P_i \tag{1}$$

In this way, the activity with the largest possibility is selected as the predicted one.

3. EXPERIMENT

The development dataset is used for the training process. There are 4 channels (C_1, C_2, C_3, C_4) within one node. We manually average the 4 channels to yield C_5, where $C_5 = (C_1 + C_2 + C_3 + C_4)/4$. It is believed to better fuse the 4 channels and augment the dataset.

In the traing phase, the audio segment from each channel is regarded as independent samples. Also, three models with the three different acoustic features are trained independently. In the testing phase, an ensemble result of the three models is obtained to predict the activity label for the given audio segment. The given evaluation setup is used to evaluate the performance of the proposed system. The results from different evaluation folds are demonstrated as follows.

Table 2 indicates the recognition accuracy results. The results reveal that the ensemble model has a better performance in most cases than using only single model.

The confusion matrix (see Figure. 5) is also calculated for the ensemble model among 9 activities listed in Table 1. It can be concluded that the proposed model can correctly recognize most activities except for the "other" activity, which is often wrongly classified as "absence" or "working". By hearing the audio segments by ourselves, we find these three activities audio do have some similar patterns in common. The "vaccumm cleaning" and "watching TV" have the best performances amony all the activities.

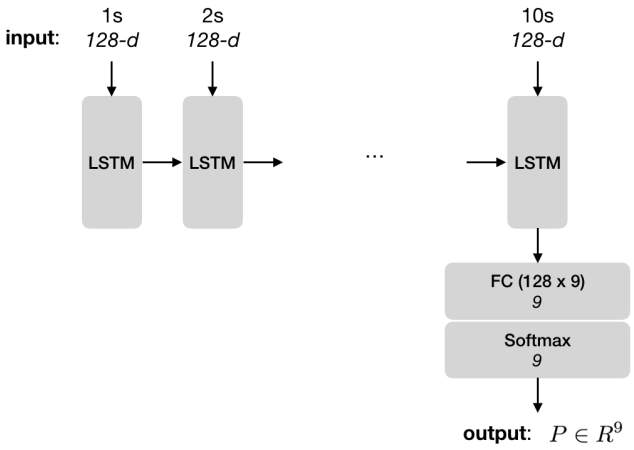

Figure 4: The architecture of the LSTM network.

Table 2: Recogniton accuracy results

Method	Fold 1	Fold 2	Fold 3	Fold 4
a	0.9164	0.9124	**0.9088**	0.9474
b	0.9134	0.9176	0.8531	0.9490
c	0.8947	0.9118	0.8282	0.9404
a + b + c	**0.9214**	**0.9257**	0.8841	**0.9534**

a = Mel-Spectrogram, b = MFCC, c = VGGish

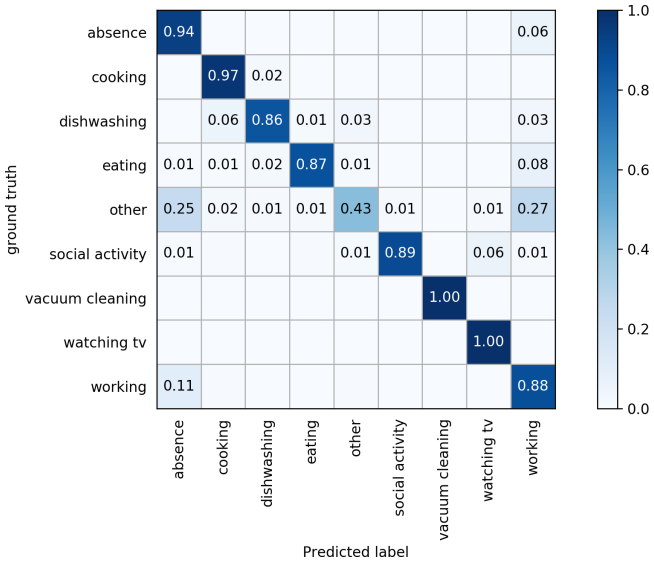

Figure 5: The confusion matrix among different activities.

4. CONCLUSION

In this challenge, an ensemble system is proposed to tackle the task of monitoring domestic activities based on multi-channel accoustics. The Mel-Spectrograms, MFCC and VGGish features are extracted from the raw audio and fused to train different deep neural networks. The experimental results on the development dataset demonstrate that the proposed system has shown superior performance in recognizing domestic activities.

5. REFERENCES

[1] http://dcase.community/challenge2018/.

[2] http://dcase.community/challenge2018/task-monitoring-domestic-activities.

[3] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 32–36.

[4] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," KU Leuven, Tech. Rep., 2018. [Online]. Available: https://arxiv.org/abs/1807.11246

[5] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.