Performance Evaluation of Acoustic Scene Classification Using DNN-GMM and Frame-Concatenated Acoustic Features

Gen Takahashi*, Takeshi Yamada*, Nobutaka Ono† and Shoji Makino*
* University of Tsukuba, Japan
† National Institute of Informatics / SOKENDAI, Japan
Email: g.takahashi@mmlab.cs.tsukuba.ac.jp

***Abstract*—We previously proposed a method of acoustic scene classification using a deep neural network-Gaussian mixture model (DNN-GMM) and frame-concatenated acoustic features. It was submitted to the Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 Challenge and was ranked eighth among 49 algorithms. In the proposed method, acoustic features in temporally distant frames were concatenated to capture their temporal relationship. The experimental results indicated that the classification accuracy is improved by increasing the number of concatenated frames. On the other hand, the frame concatenation interval, which is the interval with which the frames used for frame concatenation are selected, is another important parameter. In our previous method, the frame concatenation interval was fixed to 100 ms. In this paper, we optimize the number of concatenated frames and the frame concatenation interval for the previously proposed method. As a result, it was confirmed that the classification accuracy of the method was improved by 2.61% in comparison with the result submitted to the DCASE 2016.**

I. Introduction

Attempts have been made to automatically recognize human behavior and surrounding circumstances. This technology is applicable to monitoring elderly people, the auto-tagging of multimedia contents and life-log collection. Acoustic event detection and acoustic scene classification have been focused on as fundamental technologies used in these systems. Acoustic event detection detects acoustic events including sound signals and their timestamps, where acoustic events are a single sound emitted from one sound source, such as door opening, coughing and mouse clicking. On the other hand, acoustic scene classification classifies the place or situation where the acoustic sound was recorded. The length of an acoustic sound is typically on the order of 10 s, and types of sounds include those of buses, parks and crowds. In this research, we focus on acoustic scene classification.

The Detection and Classification of Acoustic Scenes and Events (DCASE) 2013 [1] is a workshop on acoustic event detection and acoustic scene classification. Tasks involving acoustic scene classification and acoustic event detection are prepared at the DCASE 2013. In the methods proposed for acoustic scene classification [2][3], mel frequency spectral coefficients (MFCCs), the mel frequency spectrum and recurrence quantification analysis (RQA) have been used as features and a Gaussian mixture model (GMM) and a support vector machine (SVM) have been used as classifiers. The resulting classification accuracy was at most 70%.

On the other hand, deep neural networks (DNNs), which have a multilayer neural network, have recently been actively investigated. In general, a DNN tends to fall into a local solution and requires an unrealistic learning time. However, a pre-training method [4] that gives appropriate initial values and high-speed computation on a graphics processing unit (GPU) have been established. Because of this, DNNs are now being applied in various classification problems. For speech recognition, a DNN-hidden Markov model (HMM), which combines a DNN and an HMM, has been proposed [4]. The probability distribution in an HMM is approximately represented by a GMM. On the other hand, it is precisely represented by the DNN in a DNN-HMM. It was reported that the performance of speech recognition is markedly improved using a DNN-HMM [4].

Previously, we proposed a method of acoustic scene classification using a DNN-GMM and frame-concatenated acoustic features [5]. It was submitted to the DCASE 2016 Challenge [6] and was ranked eighth among 49 algorithms. In the proposed method, features in temporally distant frames were concatenated to capture their temporal relationship. The experimental results indicated that the classification accuracy is improved by increasing the number of concatenated frames. On the other hand, the frame concatenation interval, which is the interval with which the frames used for frame concatenation are selected, is another important parameter. In our method, the frame concatenation interval was fixed to 100 ms. In this paper, we optimize the number of concatenated frames and the frame concatenation interval of the previously proposed method.

II. Proposed method

A. Process flow of the proposed method

Figure 1 shows the process flow of the previously proposed method. We assume two input channels because the acoustic data in the DCASE 2016 had two channels. First, we compute acoustic features (the MFCCs and its first and second differences) in each time frame for both the left and right channels. MFCCs are a representation of frequency characteristics considering human auditory characteristics. The

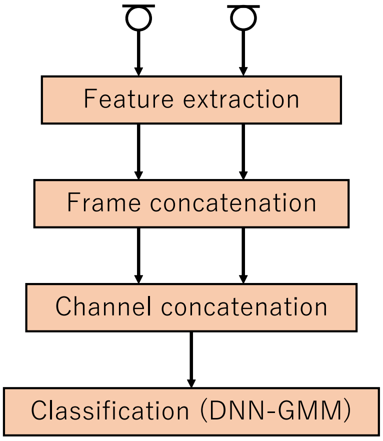


Fig. 1. Process flow.

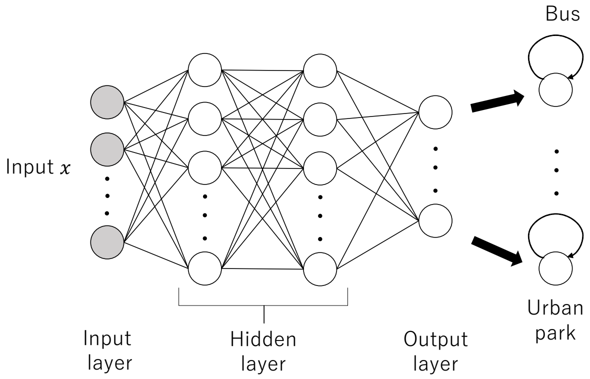


Fig. 2. Example of a DNN-GMM.

features used in our method are similar as those in the baseline system of the DCASE 2016. Next, we concatenate acoustic features in each frame and channel, which we will describe in Sect. II-C. Finally, we perform acoustic scene classification by inputting the high-dimensional acoustic features obtained in this manner into the DNN-GMM. We describe the DNN-GMM below.

B. DNN-GMM

Figure 2 shows an example of a DNN-GMM. A DNN-GMM consists of multilayered neural network and GMMs corresponding to individual acoustic scenes. A DNN-GMM is basically the same as a DNN-HMM but a GMM (a one-state HMM without a state transition) is used instead of an HMM.

Acoustic scene classification is the problem of classifying an acoustic scene $\hat{s}$ from a time series of features $\mathbf{X}$. $\hat{s}$ is expressed as

$$\hat{s} = \underset{s_k}{\operatorname{argmax}}\, P(s_k|\mathbf{X}), \tag{1}$$

where s_k is an individual acoustic scene. By transforming $P(s_k|\mathbf{X})$ in this equation using Bayes' theorem, the following equation is obtained:

$$P(s_k|\mathbf{X}) = \frac{P(\mathbf{X}|s_k)}{P(\mathbf{X})}P(s_k). \tag{2}$$

Since $P(\mathbf{X})$ is independent of s_k, $P(\mathbf{X})$ can be regarded as a constant. Also, if we assume that the probability of appearance of each acoustic scene has a uniform distribution, we also can ignore $P(s_k)$. Therefore, Equation (1) becomes

$$\hat{s} = \underset{s_k}{\operatorname{argmax}}\, P(\mathbf{X}|s_k). \tag{3}$$

A GMM is often used in acoustic scene classification as a model for solving Equation (3). Let $\boldsymbol{x}_t$ and s_k^t be the feature vector and the state at time frame t, respectively, then $P(\mathbf{X}|s_k)$ can be obtained using a GMM as follows:

$$P(\mathbf{X}|s_k) = \prod_t P(\boldsymbol{x}_t|s_k^t)P(s_k^t|s_k^{t-1}). \tag{4}$$

This represents the probability that the series of features $\mathbf{X}$ is generated from the GMM of the acoustic scene s_k. It is generally difficult to obtain the true distribution of the output probability $P(\boldsymbol{x}|s_k)$; thus, the GMM represents it by the following Gaussian mixture distribution:

$$p(\boldsymbol{x}|s_k) = \sum_{i=1}^{I} \pi_i N(\boldsymbol{x}|\mu_i, \Sigma_i), \tag{5}$$

where $N(\cdot)$ is a normal distribution, and μ_i and Σ_i are the mean and variance of the normal distribution, respectively. Also, I is the number of distributions used for mixing and π_i is the weight for each distribution. A DNN-GMM represents this output probability using a DNN instead of a Gaussian mixture distribution. To express the output probability using a DNN, $P(\boldsymbol{x}_t|s_k^t)$ in Equation (4) is transformed by Bayes' theorem to obtain

$$p(\boldsymbol{x}_t|s_k^t) = \frac{P(s_k^t|\boldsymbol{x}_t)}{P(s_k^t)}P(\boldsymbol{x}_t). \tag{6}$$

Since $P(\boldsymbol{x}_t)$ is independent of s_k, $P(\boldsymbol{x}_t)$ can be regarded as a constant. Also, if we assume that the probability of appearance of each acoustic scene has a uniform distribution, $P(s_k^t)$ can also be ignored. The DNN and GMM are integrated by setting the input vector of the DNN to $\boldsymbol{x}_t$ in Equation (6) and each node of the output layer to $P(s_k^t|\boldsymbol{x}_t)$.

We describe the learning method of the DNN-GMM adopted in this paper. First, we perform supervised learning of each acoustic scene model using a conventional GMM. Next, using this GMM, we make training data for DNN consisting of the features in each frame of each training data and the state number of the GMM. In the case of acoustic scene classification, one acoustic sound file corresponds to one acoustic scene, allowing this step to be simplified. Then, the initial parameters of the DNN are determined by unsupervised pre-training [7]. In the pre-training, each layer is regarded as a restricted Boltzmann machine (RBM) and the training is performed using the contrastive divergence (CD) method. Finally, we perform fine-tuning, which is supervised training

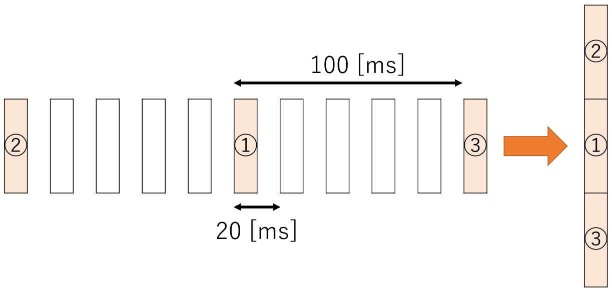


Fig. 3. Example of the time frame concatenation ($n = 3$ and $m = 100$ ms).

using training data. In the fine-tuning, we add a softmax layer initialized using a random seed and perform backpropagation using the stochastic gradient descent (SGD) method.

C. Feature concatenation

In the case of acoustic scene classification, the relationship between sounds that are more temporally distant than those involved in speech recognition is considered to be involved in the classification. Therefore, improvement in the classification accuracy can be expected by concatenating such features. Furthermore, we concatenate the features of the left and right channels in each time frame to capture the spatial information. We use the above features as the input into the DNN-GMM. Here we describe the concrete process of frame concatenation. We concatenate acoustic features in each frame with those in several temporally distant frames before and after the frame for both the left and right channels. There are two important parameters in this process. One is the number of concatenated frames n and the other is the frame concatenation interval m. We concatenate n frames before and after each frame at an interval of m ms, including the current frame. Figure 3 shows an example of the time frame concatenation. When n and m are set to 3 and 100 ms, the three frames are concatenated as shown in Figure 3. On the other hand, when m is changed to 20 ms, the eleven frames must be concatenated. The dimensions of the feature then increase about by four times. This would have a negative effect to train the classifier. By using n and m, we can capture the relationship of temporally distant frames without increasing the dimensions of the feature significantly.

III. Evaluation

A. Experimental conditions

In this experiment, we evaluated the effectiveness of our method using the development dataset and evaluation dataset provided by the DCASE 2016 Challenge. Table I gives an overview of this dataset. The dataset contains 15 acoustic scenes, for example, bus, restaurant and park. Each scene has 78 sound data in the development dataset and 26 sound data in the evaluation dataset, each of which is a stereo signal with a duration of 30 s. The evaluation dataset has open sound data, which are different from the sound data of the development dataset. The sampling frequency is 44.1 kHz and the number of quantization bits is 16.

TABLE I
Overview of the development dataset and the evaluation dataset.

# of scenes	15
# of sound data of development dataset	1170 (=15 scenes×78 data)
# of sound data of evaluation dataset	390 (=15 scenes×26 data)
data length	30 s
# of channels	2 (left and right)
sampling frequency	44.1 kHz
quantization bits	16 bit

TABLE II
Conditions of the acoustic features and the DNN-GMM.

feature	20th-order MFCCs $+\Delta + \Delta\Delta$ (60 dimensions) $\times n$ frames ×2 channels
frame length	40 ms
frame period	20 ms
# of concatenated frames n	1, 3, 5, 7
frame concatenation interval m (ms)	20, 100, 200, 500, 1000, 2000
# of hidden layers	2, 3, 4, 5
dimension of hidden layer	256, 512, 1024, 2048
dimension of input layer	$120\times n$
dimension of output layer	15

Table II shows the conditions of the acoustic features and the DNN-GMM. The features are based on a total of 60 dimensions of 20th-order MFCCs and their first and second differences. The time frame length and frame period in the frame analysis are 40 and 20 ms, respectively, which are the same as those used in the baseline system of the DCASE 2016 Challenge. By the concatenation of n frames and two channels, the dimensions of the features become $20 \times 3 \times n \text{ (frame)} \times 2 \text{ (ch)}$. The frames used for frame concatenation are selected at m ms intervals. In this experiment, the number of concatenated frames n is set to 1, 3, 5 or 7, and the frame concatenation interval m is set to 20, 100, 200, 500, 1000 or 2000 ms.

The number of hidden layers of the DNN (excluding the input layer and softmax layer) is set to 2, 3, 4 or 5, and the dimension of each hidden layer is constant and set to 256, 512, 1024 or 2048. We performed the pre-training processing on the RBM of the hidden layer using the CD-1 method and on the RBM of the input layer using the CD-2 method. We set the learning rate of the RBM to 0.4, the learning rate of the DNN to 0.008 and the dropout rate to 0.0 on the basis of the results of a preliminary experiment. In the training of the classifier, we first performed four fold cross-validation (data open, acoustic scene closed) on the development dataset and optimized the number of layers of the DNN and the dimension of the hidden

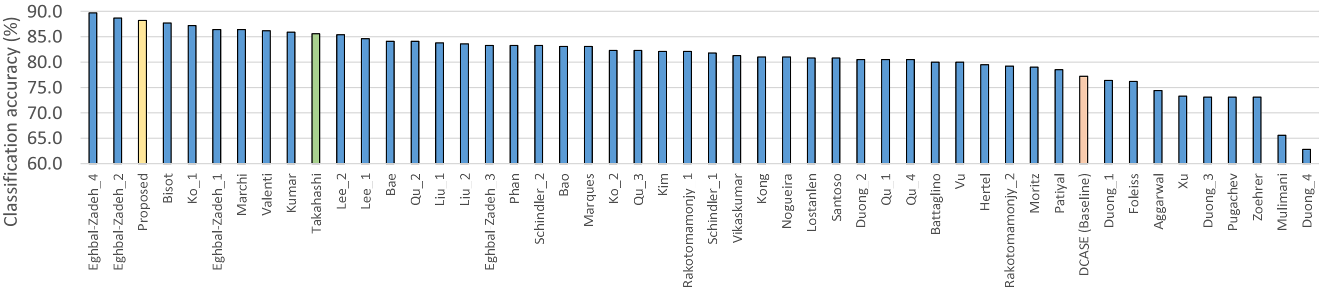


Fig. 4. Results of the DCASE 2016 Challenge [9].

TABLE III
CLASSIFICATION ACCURACY FOR EACH COMBINATION OF n AND m, WITH THE NUMBER OF HIDDEN LAYERS AND THE DIMENSION OF THE HIDDEN LAYERS OPTIMIZED FOR THE DEVELOPMENT DATASET IN PARENTHESES.

m \ n	1	3	5	7
20 ms	84.59 (3, 1024)	*85.13* *(2, 2048)*	*85.13* *(2, 1024)*	*85.47* *(3, 1024)*
100 ms	–	*86.28* *(2, 1024)*	*86.00* *(2, 2048)*	84.49 (3, 512)
200 ms	–	*85.26* *(2, 512)*	*86.62* *(2, 2048)*	*85.64* *(3, 2048)*
500 ms	–	*86.97* *(3, 2048)*	*85.23* *(2, 2048)*	84.92 (2, 1024)
1000 ms	–	*85.36* *(3, 512)*	82.69 (2, 256)	84.10 (2, 1024)
2000 ms	–	83.97 (2, 512)	83.49 (2, 1024)	82.69 (2, 256)

TABLE IV
CLASSIFICATION ACCURACY FOR EACH SOUND SCENE USING THE BASELINE SYSTEM, THE PREVIOUSLY PROPOSED METHOD ($n = 1$) AND THE OPTIMIZED METHOD ($n = 3$, $m = 500$ MS).

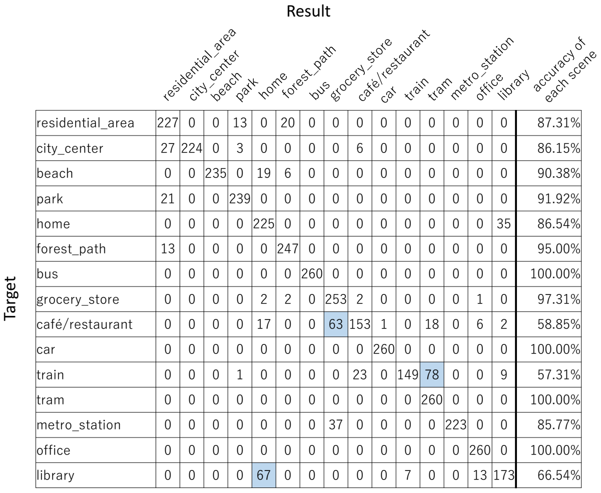
scene \ method	Previously proposed method ($n = 1$)	Optimized method ($n = 3, m = 500$)
residential_area	83.46%	87.31%
city_center	85.00%	86.15%
beach	89.62%	90.38%
park	91.92%	91.92%
home	88.46%	86.54%
forest_path	95.77%	95.00%
bus	96.54%	100.00%
grocery_store	89.23%	97.31%
café/restaurant	43.46%	58.85%
car	100.00%	100.00%
train	47.69%	57.31%
tram	100.00%	100.00%
metro_station	88.46%	85.77%
office	100.00%	100.00%
library	69.23%	66.54%

layer, with a fixed random seed for initialization of the softmax layer. Then, using the parameters, we trained the classifier of which we used all the acoustic data in the development dataset for the training, with each of ten different random seeds for initialization of the softmax layer. We evaluated the effectiveness of the method using the trained classifier and the evaluation dataset. We used the Kaldi toolkit [8] to build our system.

B. Results

Table III shows the results of the experiment. The columns are the time frame concatenation interval m and the rows are the number of concatenated frames n. Each classification accuracy is the average of the classification accuracies obtained for the ten different random seeds for initialization of the softmax layer. The standard deviation of the classification accuracies obtained for the ten different random seeds was at most 1.1% in each combination of n and m. The number of hidden layers and the dimension of the hidden layer optimized for development dataset are written in parentheses. We can see that the classification accuracy of our method varies with n and m. The highest average accuracy of 86.97% was obtained when $n = 3$ and $m = 500$ ms, which is 2.38% higher than that when $n = 1$ (no frame concatenation). This demonstrates the effect of using the concatenated acoustic feature. In Table III, classification accuracies of more than 85.0%, written in italics, were obtained when n was 5 or less and m was 500 ms or less. On the other hand, the classification accuracy was significantly reduced when n or m was too high.

Next, Figure 4 summarizes the results of the DCASE 2016 Challenge [9]. The orange, green and yellow bars in this figure represent the baseline system of the DCASE 2016 Challenge, the proposed method ($n = 5$, $m = 100$ ms) submitted to the DCASE 2016 Challenge (note that the features are irreversibly compressed) and the optimized method ($n = 3$, $m = 500$ ms), respectively. The classification accuracy of the optimized method ($n = 3$, $m = 500$ ms) was improved by 9.77% compared with the baseline system. This illustrates the effect of using both DNN-GMM and concatenated acoustic features. The best classification accuracy of our method ($n = 3$, $m = 500$ ms) in the ten random seeds was 88.21% and was improved by 2.61% compared with that of the proposed method ($n = 5$ and $m = 100$ ms), and was ranked third among the 49 algorithms.



Target \ Result	residential_area	city_center	beach	park	home	forest_path	bus	grocery_store	café/restaurant	car	train	tram	metro_station	office	library	accuracy of each scene
residential_area	227	0	0	13	0	20	0	0	0	0	0	0	0	0	0	87.31%
city_center	27	224	0	3	0	0	0	0	6	0	0	0	0	0	0	86.15%
beach	0	0	235	0	19	6	0	0	0	0	0	0	0	0	0	90.38%
park	21	0	0	239	0	0	0	0	0	0	0	0	0	0	0	91.92%
home	0	0	0	0	225	0	0	0	0	0	0	0	0	0	35	86.54%
forest_path	13	0	0	0	0	247	0	0	0	0	0	0	0	0	0	95.00%
bus	0	0	0	0	0	0	260	0	0	0	0	0	0	0	0	100.00%
grocery_store	0	0	0	0	2	2	0	253	2	0	0	0	0	1	0	97.31%
café/restaurant	0	0	0	0	17	0	0	63	153	1	0	18	0	6	2	58.85%
car	0	0	0	0	0	0	0	0	0	260	0	0	0	0	0	100.00%
train	0	0	0	1	0	0	0	0	23	0	149	78	0	0	9	57.31%
tram	0	0	0	0	0	0	0	0	0	0	0	260	0	0	0	100.00%
metro_station	0	0	0	0	0	0	0	37	0	0	0	0	223	0	0	85.77%
office	0	0	0	0	0	0	0	0	0	0	0	0	0	260	0	100.00%
library	0	0	0	0	67	0	0	0	0	0	7	0	0	13	173	66.54%

Fig. 5. Confusion matrix of the optimized method ($n = 3$, $m = 500$ ms) , which is generated by counting the classification results obtained for all the ten random seeds.

Finally, Table IV shows the classification accuracy for each sound scene. The column and row show the sound scene and the method, respectively. Our method ($n = 3$, $m = 500$ ms) is greatly improved for grocery_store, cafe/restaurant and train compared with the previously proposed method ($n = 1$). The sounds in grocery_store, cafe/restaurant and train are loud and unstationary with large temporal changes, including music and speech. Figure 5 is the confusion matrix of the optimized method ($n = 3$, $m = 500$ ms), which is generated by counting the classification results obtained for all the ten random seeds. The columns and rows show the target sound scene and the classified result, respectively. From Figure 5, many classification errors for cafe/restaurant, train and library can be observed. For example, the sound data in cafe/restaurant are often incorrectly classified as grocery_store, since the sound data in both sound scenes are similar. Therefore, improvement of the classification accuracy for these sound scenes is required in the future.

IV. Conclusion

In this paper, we optimized the number of concatenated frames and the frame concatenation interval for the previously proposed method. We carried out an experiment using the development dataset and the evaluation dataset of the DCASE 2016 Challenge. It was found that by setting the number of concatenated frames n to 5 or less and the frame concatenation interval m to 500 ms or less, high classification accuracy was obtained relatively stably. Also, the classification accuracy of sound scenes having large temporal changes was improved.

References

[1] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, M. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," Proc. WASPAA 2013, Oct. 2013.

[2] G. Roma, W. Nogueira, P. Herrera, "Recurrence quantification analysis features for auditory scene classification," Proc. WASPAA 2013, Oct. 2013.

[3] J. Nam Z. Hyung, K. Lee, "Acoustic scene classification using sparse feature learning and selective max-pool by event detection," Proc. WASPAA 2013, Oct. 2013.

[4] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury, "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.

[5] G. Takahashi, T. Yamada, S. Makino, N. Ono, "Acoustic scene classification using deep neural network and frame-concatenated acoustic feature," DCASE 2016 Challenge Technical Report, Sep. 2016, http://www.cs.tut.fi/sgn/arg/dcase2016/documents/challenge_technical_reports/Task1/Takahashi_2016_task1.pdf.

[6] http://www.cs.tut.fi/sgn/arg/dcase2016/.

[7] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," Momentum, Vol. 9, No. 1, 2010.

[8] http://kaldi-asr.org/.

[9] http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification.