TRANSFER LEARNING BASED DNN-HMM HYBRID SYSTEM FOR RARE SOUND EVENT DETECTION

Jian-Fei Wang, Wei-Qiang Zhang, Jia Liu

Department of Electronic Engineering,
Tsinghua University,
Bejing 100084,China
wangjf15@mails.tsinghua.edu.cn, {wqzhang,liujia}@tsinghua.edu.cn

ABSTRACT

In this paper, we propose an improved Deep Neural Network-Hidden Markov Model (DNN-HMM) hybrid system for rare sound event detection. The proposed system leverages transfer learning technology in the neural network training stage. Experiment results indicate that transfer learning is more efficient when the training samples are insufficient. We use the Multi-Layer Perception (MLP) system and standard DNN-HMM system as the baseline. The performance was evaluated on the DCASE2017 task 2 development dataset show that our proposed system outperforms the MLP and DNN-HMM baselines, and finally achieves an average error rate (ER) of 0.38 and 78.3% F1-score on the event-based evaluation. The average error rate of proposed system is 15% and 8% absolutely lower than the MLP and DNN-HMM systems, respectively.

Index Terms— Rare sound event detection, transfer learning, deep neural network, hidden Markov model

1. INTRODUCTION

The research of sound event detection (SED), also named as acoustic event detection (AED), became popular in this decade. The applications of SED technology are needed on a number of occasions, including acoustic surveillance, environmental context detection, automatic audio indexing/retrieving, smart house, diseases detection, urban planning, acoustic ecology, organisation/navigation of sound archives, scene understanding, audio source segment [1–4], and so on.

There are two main categories in the SED research. One is the detection of sound events in a particular scene, called overlapping sound events detection, which also called polyphonic event detection (PED) [5] [6]. This task requires events to be related to the scene. However, there is no restriction on the number of event classes and when the events occur. Its purpose is to make research tasks closer to reality. The other category is the detection of specific sound events in different scene environments, called monophonic event detection, such as the rare SED task in DCASE2017 challenge [7]. This task is mainly used to monitor important events, and it can be used to find the interesting audio events on the Internet.

The research methods of SED includes designing the sound event features and constructing the acoustic models. The designed sound features are usually high-level features according to the requirements of the task. Valenzise et al. proposed a fusion of traditional features to detect screams and guns in different event-to-background ratio (EBR) environments [8]. These methods are followed by a good classifier as usual. The support vector machine (SVM) is one of the most popular choices for its good performance and ease of use [9]. Acoustic model building is to find the best way to describe the audio events based on statistical significance. The training is supervised, and the inputs are always the basic features. Mesaros built a SED systems based on HMM for SED in everyday environment [10]. Zhuang et al. used artificial neural network (ANN-HMM) to model the sound events [11]. DNN is one of the most popular topics in recent years. It has not only achieved a high performance with simple structures, more importantly, it can learn knowledge from the data. Kong et al. used DNN to combine feature extraction and model construction [12]. Espi et al. used the auto-encoder to complete the unsupervised feature extraction [13]. The McLoughlin team used DNN as the classifier with time-frequency spectral image feature (SIF) [14]. In addition, Convolution neural network (CNN) and Long Short-Term Memory (LSTM) were introduced into SED as well [15–17]. However, most of these methods need a lot of data to train a robust model.

In this paper, we proposed a new DNN-HMM hybrid system for rare SED task. The proposed system leverages the transfer learning technology in the neural network training stage. We used DNN-HMM as the main structure because the DNN-HMM system has good stability and adaptability in Auto-Speech Recognition (ASR), which is similar to SED. For rare SED, the training sample is not sufficient. So we make an attempt to utilize transfer learning to overcome the data sparse problem.

The rest of paper is organized as follows: Section 2 presents the concept and strategy of transfer learning technology. Section 3 describes the proposed method in detail. Section 4 introduces the experiments and analyzes the results. Finally, we conclude this paper and discuss future work in Section 5.

2. TRANSFER LEARNING

DNN is a data-driven multi-layer self-learning network structure. The classification and self-learning are implemented by forward and backward propagation of the network. Specifically to the nodes on each layer, the propagations are shown in Eq. (1):

$$y_i = f(\sum_{j=1}^{n} W_{ij} x_j), \tag{1a}$$

$$W_l = W_l - \alpha \Delta W_l, \tag{1b}$$

This work was supported by the National Natural Science Foundation of China under Grant No. 61370034, No.61403224, and in part supported by China Scholarship Council. The corresponding author is W.-Q.Zhang.

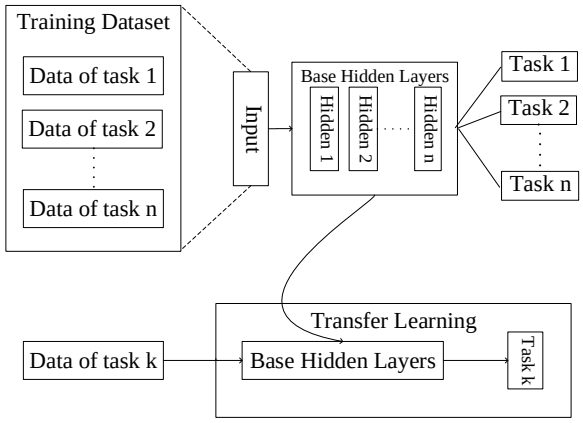

Figure 1: Transfer Learning for SED

Table 1: Feature Configuration

	fbank	MFCC
use energy	false	true
frame length	40 ms	
frame shift	25 ms	
num mel bins	40	
num ceps	-	20
low freq	300 Hz	
high freq	22050 Hz	

Table 2: EAD Configuration

energy base (E_0)	3
mean scale (S)	0.625
frames context (N)	5
proportion threshold	0.6

where, x_j is the value of the *j*th nodes, y_i is the value of the *i*th nodes at the next layer, W_{ij} is the weight connecting those two nodes above, f is the activation function, α is the learning rate.

DNN will eventually converge to a local optimal value after multiple iterations and is prone to be over-fitting when the training data is insufficient. A small amount of training data may be distinguished by insignificant differences easily, and that leads to ignore the main features. The accuracy of the training dataset will be higher, but that of the evaluation dataset is not good.

One solution is to apply Restricted Boltzmann Machine (RBM) to pre-training a Deep Belief Network (DBN). RBM is a bi-directional network, by minimizing the the total energy, it can get better parameters to describe the map between input and output comprehensively. Through fine-tuning, a better DNN can be obtained [18].

Another solution is pre-train a neural network with large out-domain data and use transfer learning to adapt the network to in-domain data. The transfer learning can be used to solve three problems: handle new tasks with existing models from other domains. Update the old models using a little bit of new data. Build a new models borrowing data from related domains.

Transfer learning has been proved to be beneficial in many examples in knowledge engineering [19]. For rare SED, we used the strategy of transfer learning as illustrated in Fig. 1.

Like human speech, the different sounds have a lot of common properties, such as tone and beats. Therefore, we can use the sound of other tasks to train a common Base Hidden Layers (BHL). The amount of training data could be increased linearly. Meanwhile, a certain degree of complementarity between events can make the extraction better. Based on BHL, we learn each task separately again to make the model fit better.

3. PROPOSED METHOD

3.1. Data preparation

The DCASE2017 Challenge provided a TUT Rare Sound Events 2017 dataset, a set of isolated sound events for detecting, including babycry, glassbreak and gunshot, and everyday acoustic scenes to serve as background. Source code for creating mixtures at different EBR was provided as well. and we used the provided code to generate a bigger training dataset. In order to simulate the possible different EBR environments, we set the EBR to 3 cases, -6dB, 0dB and 6dB. The whole length of background is 9.34h. In each EBR case, we generated 3000 mixtures wav files per target. In addition, we used background noises directly as the negative samples. The training dataset has a total of 27844 samples. After that, we also need to do some pre-processing for these synthesised audio: 1) Clipping. Find the location of the event, cut off other parts. 2) Formatting. Convert the audio to 16-bit 44.1kHz mono wav. 3) Scaling. Scale the amplitude of the audio linearly to [-255,255]. We only apply step 2 for step 3 above for background noises.

3.2. Feature extraction

Our system extracted two features, log mel-band energy features (fbank) and Mel Frequency Cepstrum Coefficient (MFCC). Fbank can keep more original information. We used it to train the DNN. MFCC was extracted based on fbank for further Discrete Cosine Transform (DCT). We used MFCC to construct the Gaussian Mixture Model (GMM) with diagonal variances. It was used for the state level label alignment later. The configuration of the feature is shown in the Table. 1.

After the feature extraction, we did Sound Event Activity Detection (SEAD) using the energy feature. We considered the lowest part of an audio as a useless part, which belongs to neither the event nor scene. It is simple and effective for AED, however, this part may affect the accuracy of model training sometimes. We removed it in the training process. The decision score is calculated as Eq. (2).

$$E_\theta = E_0 + \frac{S}{T}\sum_{t=0}^{T-1} E_t, \tag{2a}$$

$$A_t = \frac{1}{2N+1}\sum_{t-N}^{t+N}(E_t > E_\theta), \tag{2b}$$

We set an energy threshold E_θ. It is calculated by an energy base E_0 and a mean scale S. E_t is the log energy of the t-th frame. The activity score of the t-th frame A_t is determined by N frames context together. We compared the activity score A_t with the proportion threshold to decide the activity of the t-th frame. The configuration of the SEAD is shown in Table. 2. There is an example of the effect of the SEAD illustrated in Fig. 2. The upper

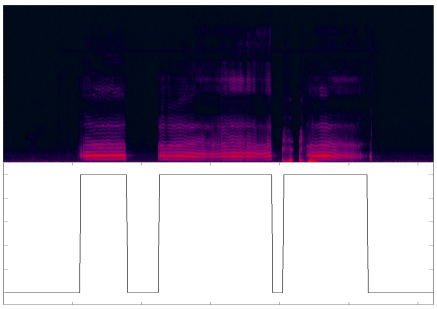

Figure 2: Effect example of EAD

figure shows the spectrum of an original audio, the lower figure shows the corresponding SEAD effect. We can use it to eliminate the low energy pause in the event to improve the accuracy of frame-level tags.

We did the Cepstral Mean Normalization (CMN) for each training data after SEAD to eliminate multiplicative noises. The feature vectors were calculated using Kaldi toolkit [20].

3.3. Acoustic model

Each event corresponds to an acoustic model. And we generally describe the acoustic models from two perspectives, the temporal structure and the spectral structure. Here, we chose DNN-HMM structure to simulate the acoustic models.

3.3.1. Temporal structure

Temporal structure mainly refers to the transition of the event with time. The process of transition in each category are different. These process can be roughly divided into several cases. Depending on whether the event is periodic, it can be divided into periodic events and aperiodic events. According to the symmetry of events, it can be divided into symmetry events and asymmetric events. The state number of each event is not the same, which is related to the average length of the event in one period.

Firstly, there are two special states. The first state S0 is the event-start state (ESS), and the last state is the event-end state (EES), which play the role of connecting different events. The ESS can only transfer to state S1, and the EES can not transfer.

Except for the ESS and EES, we designed 4 different state transfer topology for corresponding situations. As illustrated in Fig. 3, babycry event belongs to symmetry periodic category. Because of symmetry, the back-end states are similar to the front states. That will cause confusion during training. So we set the rule that for symmetry periodic category, adjacent states can transfer between each other, and the start state S1 can transfer directly to the EES. The glassbreak and gunshot events belong to asymmetric periodic category. Different from symmetry events, the states in different are different too. So we set the second rule that for asymmetric periodic category, transfer between stats can only be done from front to back, except for the last state S4. Only S4 can transfer back to the start state S1 to keep the periodicity. The regularity of scenes is not obvious, thus, we used a simple 3 states structure. The silence model is used to absorb the stationary random noise and impulse noise. It also belongs to period structure.

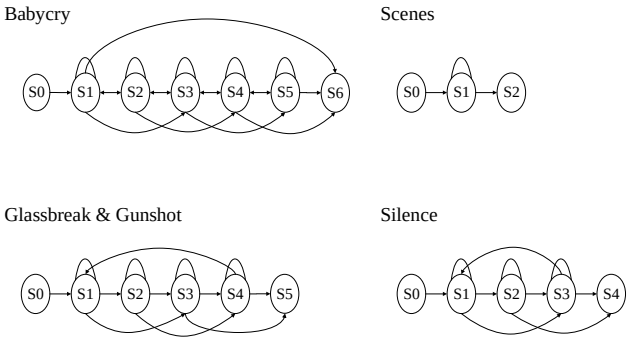

Figure 3: HMM structure

The HMMs were trained using HTK [21]. The state emission probability was generated by GMMs. Therefore, we chose the MFCC features as the input. As a by-product of training HMM, we could get the state-level labels of events. GMM-HMM training was semi-supervised. We only provided the content of the audio, without specifying the location where the event occurred. The cluster of the state was completely driven by data. This can avoid manmade annotated errors. But it will also lead to the problem that the location of label is not accurate, and sometimes marking wrong. Thus, we should use the existing event-level label to do a calibration alignment.

3.3.2. Spectral structure

Spectral structure describes the distribution of energy in different frequency bands. It is just like a classifier. We used the 5 frame-context fbank features as the input, and obtained the probability of each state:

$$P(s_t = i|x_t) = \frac{\exp(a_{i,t})}{\sum_{i=0}^{K} \exp(a_{i,t})} \tag{3}$$

where the x_t is the feature vector of the t-th frame, and the s_t is the state of the t-th frame, and $a_{i,t}$ is the output of activation function of t-th frame to the i-th state. We used a DNN model to simulate the spectral structure with a softmax layer at the end. In the training stage, we used state-level labels generated above as the output. The DNN model has 2 hidden layers. The dimension of input is 200, each hidden layer has 400 nodes.

To avoid the over-fitting, transfer learning strategy was used. We used all of the training data to train the BHL regarding the whole states of three events as the output, which has 32 dimensions. DBN was trained for pre-training preparation. The BHL was adapted on the base of DBN. We used the sigmoid as the activation function and cross-entropy as the cost function. Considering the problem of imbalance between positive and negative samples, we used different weights for each type of data. The weights update formula are shown below:

$$W_l = W_l - P^t_{weight} \alpha \Delta W_l, \tag{4}$$

where the P^t_{weight} is the weight of the t-th frame. Then, we used the data of each task to do the adaptation on the BHL with a dropout rate 0.2. The configuration in details is listed in Table. 3.

Table 3: DNN Configuration

	DBN	DNN
learning rate	0.4/0.01	0.001
nnet depth	2	
hidden dimension	400	

Table 4: Results compared to baselines

	Class-based average	
	ER	F1[%]
MLP(Baseline)	0.53	72.7
DNN-HMM (standard)	0.46	72.2
DNN-HMM+BHL (Our method)	**0.39**	**77.6**

4. EXPERIMENT

We evaluated our proposed method on the DCASE2017 task2 development dataset. Each event has 500 test audio, these audio are also synthesized by the same way as the training dataset. The EBR of the synthetic audio could be -6dB, 0dB or 6dB, and only half of them have a real target event. The length of each test audio last 30s. The evaluation metric for the experiment is event-based error rate calculated using onset-only condition with a collar of 500 ms. Additionally, event-based F-score with a 500 ms onset-only collar will be calculated [22].

4.1. Comparing with baseline

We compared our proposed system with two baseline systems. One is a multi-layer perception architecture system using fbank features. The neural network contained two dense layers of 50 hidden units per layer and 20% dropout was trained for 200 epochs for each class. The detection was decided by a median filter based on a single output neuron of sigmoid type [23]. The other is a standard DNN-HMM without transfer learning. The results are shown in Table. 4. Our method achieved the best results among three systems. The ER metrics of DNN-HMM system are better than single MLP system, but the F1 score of standard DNN-HMM are lower than MLP baseline. The transfer learning can help the DNN-HMM system more powerful.

A detailed description of results of each events is shown in Table. 5. MLP system mainly used the information of the spectral structure. Therefore, the performance is good when the spectral structure is more important than the temporal structure, such as glassbreak. Conversely, it can not describe the feature of babycry with an obvious temporal structure. The DNN-HMM system increased the ability to characterize timing. The result of babycry detection were much better than that of MLP. And the result of glassbreak detection was improved a little. However, the result of gunshot detection was worse then others. On the one side, the temporal and spectral structure of the gunshot is not easy to learn. On the other side, we found that the insertion rate is 0.08, but the deletion rate of gunshot detection is 0.63. The recalled parts of the gunshot are almost right, but more than half of gunshot can not be recalled. That means the DNN-HMM system did not fully capture the features of gunshot, which belongs to an over-fitting phenomenon. From the result, the transfer learning technology can help reduce the degree of over-fitting by the gunshot model training. The deletion rate of DNN-HMM+BHL for gunshot decreased to 0.49, and the insertion rate decreased to 0.04. However, the performance of detecting the glassbreak became worse. The deletion rate changed from 0.14 to 0.22. That means, for events whose features are easy to learn, the transfer learning technology may reduce the performance. At the same time, for those features not obvious, transfer learning can help a lot.

Table 5: Results of each event

	Babycry		Glassbreak		Gunshot	
	ER	F1[%]	ER	F1[%]	ER	F1[%]
MLP	0.67	72.0	0.22	88.5	0.69	57.4
DNN-HMM	0.46	76.5	**0.19**	**89.7**	0.72	50.5
DNN-HMM +BHL	**0.39**	**79.9**	0.25	86.0	**0.54**	**65.1**

Table 6: Results of fusion system

	Class-based average	
	ER	F1[%]
Fusion system	0.38	78.3

4.2. Fusion experiment

Through the above analysis, we found that the transfer learning is good at learning the hard events but can not help the easy-learning events. Thus, we selectively chose the transfer learning for different events. The result is shown as Table. 6. The error rate result decreased to 0.38.

5. CONCLUSION

We proposed a new application of transfer learning in rare SED base on a DNN-HMM hybrid system, and applied it to the DCASE2017 challenge task 2. We compared our proposed method to MLP system and standard DNN-HMM sytem. Our proposed method has the best performance, and achieved an average error rate of 0.38 (F-score of 78.3%) on the event-based evaluation.

In future work, we will try to improve the system performance in three ways. First, the fbank features will still lose some useful information, we should find more suitable features for the sound event detection, such as bottleneck features. Second, DNN is not suitable for modeling data with timing structure. We need to try more types of network structures to describe the spectral features, such as RNN, LSTM. At last, the restrictions of HMM on state jumps are not necessary, we need to find a more efficient way to describe the temporal structure.

6. REFERENCES

[1] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *2016 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 6460–6464.

[2] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6440–6444.

[3] J. Schr, J. Anemiiller, S. Goetze, *et al.*, "Classification of human cough signals using spectro-temporal gabor filterbank features," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6455–6459.

[4] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2742–2746.

[5] E. Benetos, M. Lagrange, M. D. Plumbley, *et al.*, "Detection of overlapping acoustic events using a temporally-constrained probabilistic model," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6450–6454.

[6] E. Miquel, M. Fujimoto, and T. Nakatani, "Acoustic event detection in speech overlapping scenarios based on high-resolution spectral input and deep learning," *IEICE TRANSACTIONS on Information and Systems*, vol. 98, no. 10, pp. 1799–1807, 2015.

[7] 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)*, November 2017, submitted.

[8] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*. IEEE, 2007, pp. 21–26.

[9] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[10] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Signal Processing Conference, 2010 18th European*. IEEE, 2010, pp. 1267–1271.

[11] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[12] Q. Kong, I. Sobieraj, W. Wang, and M. D. Plumbley, "Deep neural network baseline for dcase challenge 2016," *Proceedings of DCASE 2016*, 2016.

[13] M. Espi, M. Fujimoto, Y. Kubo, and T. Nakatani, "Spectrogram patch based acoustic event detection and classification in speech overlapping conditions," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*. IEEE, 2014, pp. 117–121.

[14] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540–552, 2015.

[15] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[16] 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.

[17] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "Blstm-hmm hybrid system combined with sound activity detection network for polyphonic sound event detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 766–770.

[18] 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.

[19] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[21] http://htk.eng.cam.ac.uk.

[22] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162

[23] http://www.cs.tut.fi/sgn/arg/dcase2017/.