Sound Event Detection Based on Multi-States Transition Model

Jian-Fei Wang, Wei-Qiang Zhang*, Jia Liu
Media Cognitive Computing Research Center
Department of Electronic Engineering, Tsinghua University
Beijing 100084, China
wangf15@mails.tsinghua.edu.cn, {wqzhang.liuj}@tsinghua.edu.cn

Abstract—Deep Neural Network Hidden Markov Model (DNN-HMM) is a popular sound event modeling framework in recent years, which shows excellent performance in various sound event detection tasks. However, the traditional DNN-HMM framework shows vast performance differences for the detection of different types of events. Multi-states transition models can efficiently solve the problem that the simple states transition topologies cannot adapt to multiple types of sound events. We evaluated the performance on the DCASE2017 task2 development dataset, which shows that proposed DNN-HMM system outperforms the baseline. The error rate (ER) is 23% lower than the Multiple Layers Perception (MLP) system, and 14% lower than the traditional DNN-HMM system.

Keywords—sound event detection; multi-states transition model; hidden Markov model; transfer learning; multi-task learning

I. INTRODUCTION

Sound event detection (SED) has received extensive attention from academia in recent years. The detection consists of two parts: detecting the position of an event fragment and figuring out the correct event class from the detected fragment. The detected events are defined according to the needs. Applications of sound event detection have been carried out in many fields, including public safety monitoring [1], intelligent living [2], medical research [3], sports competition [4] and so on.

The research direction of SED mainly focuses on the design of the acoustic model. Systems based on the Hidden Markov model (HMM) are widely introduced into the field of SED [5,6,7]. The hmm-based system can partition sound events into multi-states sequences according to the characteristics of the temporal structure. The multi-states transition model transforms the problem of modeling an event into modeling the states of the event with particular states transition structures, which can efficiently reduce the difficulty of describing sound events. Traditional HMM systems use a Gaussian Mixture Model (GMM) to model a single state and generate the likelihood probability. Given the occurrence probability of a known state, we can calculate the posterior probabilities of event frame corresponding to the state can by using Bayesian formula. However, because the occurrence probability of each state is difficult to be obtained accurately, the estimation of the GMM model would deviate easily. Deep Neural Network (DNN) is a multi-layer discriminant model. Each layer of the network has multiple nodes, and nodes in adjacent layers are connected to each other by a particular activation function. DNN does well in simulating the nonlinearity relationship. Thus it can be used to estimate the states labels as well, and output the exact posterior probability directly, which has improved the performance significantly in Automatic Speech Recognition (ASR) tasks [8]. Therefore, we chose DNN-HMM as the SED framework.

The model objects of DNN-HMM system used in ASR is the sub-word units such as phoneme. The phonemes have a short duration, and the timing structure is relatively simple, so it can be well modeled by using simple states transition structure. However, the event states last usually longer and change richly. The simple states transition model cannot meet the needs at all.

In this paper, we proposed an improved DNN-HMM SED system. Aiming at improving the capability of the state transition model in HMM, a more reasonable multi-states transition model is flexibly designed.

The rest of paper is organized as follows: Section 2 introduces the system framework briefly and describes the proposed multi-states transition model. Section 3 presents the experiments and analyzes the results. Section 4 concludes this paper and discusses the future work.

II. METHOD

A. System Overview

DNN-HMM system consists of 6 parts: data preparing, feature extracting, state aligning, DNN training, DNN testing, and decoding. Fig. 1 shows the framework of the SED system.

In data preparing stage, data argumentation is used to expand the dataset. Then in feature extracting module, frame-level acoustic features are extracted to obtain a feature matrix $X \in \mathbb{R}^{F \times T}$, where $F \in \mathbb{N}$ is the dimension of the feature, and $T \in \mathbb{N}$ is the number of frames in the acoustic signal. In the state aligning module, we decomposed event-level labels into state-level labels sequences $S$ and used the new state labels to generate corresponding states transition models. DNN models are trained and tested in training stages and testing stages. The

This work was supported by National Natural Science Foundation of China under Grant No. 61370034. The corresponding author is W.-Q. Zhang.
outputs of the DNN are the posterior probability of states, which will be the inputs of the HMM decoder. In decoder module, our system will calculate the optimal states transition sequence for each sample, as in (1). We import \( \delta(i) \) to define the maximum probability of the state sequence \((i_t, i_2, ..., i_{t-1})\) with \( i = i_t \) is the state in time \( t \). Then we can calculate the \( i_t \) state \( \psi(i) \) in the maximum probability sequence \((i_t, i_2, ..., i_{t-1}, i)\). \( \lambda \) is the parameter of the state model, \( a_i \) is the transition probability from state \( j \) to state \( i \), and the \( L \) is the number of the state classes. According to the relationship between state label and event label, we can get the location of the event in the audio sample directly.

\[
\delta(i) = \max_{i_{t-2} \ldots i_1} P(i = i_t, i_{t-1}, ..., i_1, o_1, ..., o_L | \lambda), i = 1, 2, ..., L.
\]

\[
\psi(i) = \arg \max_{i_{t-2} \ldots i_1} \delta(i), i = 1, 2, ..., L.
\]

We clustered the state-level labels and trained multi-states transition model parameters by using the Baum-Welch algorithm. The transfer learning [9] and multi-task learning [10] technologies were applied to eliminate the overfitting problems.

**B. Acoustic Features**

The acoustic features used in this work are MFCC and log Mel-band energies. The features between 300Hz and 22050Hz cutoff frequencies were reserved, and the frame length was set to 40ms with 50% overlap. The MFCC features added the first-order and second-order differences, totally 60 dimensions, as input to the state alignment module. The dimension of log Mel-band energies feature was 440, which contained 11 frames contextual information. Each feature was then normalized independently to zero mean and unit variance by using statistics calculated from the training data.

**C. Multi-states Transition Architecture**

We summarized the multi-states transition architecture used in three cases: Left-To-Right topology (LTR), Symmetric Transition Model topology (STM) and Full Connection Skip topology (FCS), as in Fig. 2.

The states transition model is designed by setting the states transition topologies and the number of state classes. The number of state types is mainly related to the duration and complexity of events. And the states transition model can be designed by expert knowledge. The right topology facilitates the clustering of states and stable description of events.

Conversely, unsuitable topologies make it difficult to reasonably cluster states and hard to describe the event.

LTR topology corresponds to a linear transition structure. The state can only jump from the left (s1) to the right (s3), to meet the events whose energy feature changes in a progressive way, except that we designed the transition from s3 to s1 for the audio events those repeated periodically. LTR is the most commonly used states transition model. LTR is suitable for the glass-breaking-like events. STM topology is a new kind of state transition model. Its symmetric states transition structure is good at describing the fluctuation characteristics of some events. In this topology, only adjacent states are allowed to jump to each other, and the last state must return to the initial state s1. This setting makes up for the shortcomings that LTR can only describe the linear structure. STM is suitable for the baby-crying-like events. In FCS topology, the state can exchange among all state classes of the event. That will increase the difficulty of states clustering. However, it can learn about any states transition structure with no limits as well. So this topology is designed for sound events without clear transition structure, sometimes, using LTR and STM cannot solve.

**III. Evaluation**

We evaluated our proposed method on the DCASE2017 task2 development dataset.

**A. Acoustic Material**

The DCASE2017 dataset includes three categories of events to be detected: gunshot, baby crying and glass breaking, and 15 types of everyday acoustic scenes such as the city center, beach, park, train station, and so on, for a total of 10 hours. To simulate the event happens in many kinds of scenes, we added the event records randomly to the background scene audio with several Event-to-Background Ratio (EBR). There is only one event or no event in each synthesized audio sample [11]. The length of each sample is fixed at the 30s. In data preparing stage, we synthesized 45000 training data and 8936 validation data at -5dB, 0dB, 5dB, 10dB and 15dB EBRs. The testing data are synthesized in the same way, except that the EBRs are among -6dB, 0dB, and 6dB, and only half of the testing samples have a real target event in it. 1496 testing samples are synthesized.
B. Baseline

We compared our proposed system with two baseline systems. One is a multi-layer perceptron (MLP) system. The other is a DNN-HMM hybrid system based on the LTR states transition model.

- MLP used the log Mel-band energies features as the input. For gathering more extended temporal information, the features were concatenated for five consecutive frames, resulting in a feature vector with 200 dimensions. The output layer produced the probability of detected event per frame. The neural network contained two dense layers of 50 hidden units. 20% dropout was applied to both the inputs and hidden state outputs. And model parameters were trained for 200 epochs. The final detection result was generated through a median filter. [12]

- DNN-HMM hybrid system based on a single states transition model was almost the same with proposed DNN-HMM system except that only LTR states transition model was used.

C. Evaluation Metric

The event-based error rate (ER) with onset tolerance of 500ms was applied. ER is the sum of insertion (I), deletion (D) and substitution (S) rates, as (2). N represents the number of the event segments. [13]

\[
ER = \frac{S + D + I}{N}.
\]  

D. Results and Analyses

To verify the validity of the multi-states models, we tested three sound events separately using three state transition structures. The DNN model was a three-layer network structure; the hidden unit number was set to 400. We trained the DNN model for each event independently. The outputs of the network were the posterior probabilities of event states and background scene states. Adam algorithm was adopted to optimize the model parameters [14]. The learning rate was updated by adopting the decay strategy. The initial learning rate was 0.01. Each time the loss ceased to drop, the learning rate was halved until the loss had not decreased by three times and the training ended. The results of all experiments are given in Tab. 1.

The systems of different topology are named as LTR, FCS, and STM. It is easy to find that using LTR structure is more effective in detecting events with linear states such as glass breaking event, and it is more suitable to use STM structure for detecting the events with symmetric states transition structure such as baby crying. FCS shows better performance for gunshot event because the states transition structure of gunshot is not apparent enough. Fusion system DNN-HMM_f combines the advantages of the three topologies above, having a 4% performance gain compared with single using of LTR.

In addition to the topology design, the number of state classes also has a significant impact on the performance of the model. In previous experiments, the number of state classes of each event was set to 3. Therefore, we added seven sets of experiments, setting the number of state classes to 3, 5, 7, 9, 11, 13 and 15 respectively. Fig. 3 shows the ER of detection of three events from different data sets under multiple numbers of state classes. The vertical axis is the ER index and the horizontal axis the number of state classes. The solid lines represent the results of the development dataset, and the dotted lines represent the testing dataset results. The results show that due to the differences in event duration and complexity, the performance of the system will be severely affected by the number of state classes. We tend to use more state classes when the event length is longer, or the state transition structure is complicated. The duration of baby crying event is long, and the timing structure is relatively complicated, so it needs more state classes. Gunshot event also has a long-term duration, so multiple state classes also play a specific effect. However, the sound of glass breaking is usually sharp, using more kinds of states cannot help to reduce detecting the error. Finally, we set the number of state classes of baby crying to 15, the gunshot event seven state classes and keep the glass breaking three state classes. The multiple states system DNN-HMM_s reduces the ER by 7%.

From the experimental results in Fig. 3, it is easy to find that over-fitting phenomenon exists in DNN-HMM system, of which the detection of gunshot is the most serious. So we apply transfer learning and multi-task learning to rectify that defect. Here we needed to redesign DNN training. We used all the training samples of the three events to train a standard network first, called based neural network (BNN). The outputs of BNN are the posterior probabilities of states of all events. It can make the BNN learn more distinguishing features from multiple types of event states samples by expending the training dataset. Moreover, we use the hidden layer parameters of BNN, which represent the distinguishing feature extracting model, to initiate each event model respectively. At last, we update the three event models independently to achieve a stronger distinction. The results show that the transfer learning system DNN-HMM_t further reduces the ER by 2%. The performances of detection of glass breaking and gunshot event

<table>
<thead>
<tr>
<th>Systems</th>
<th>BabyCry</th>
<th>GlassBreak</th>
<th>Gunshot</th>
<th>Class-Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>0.67</td>
<td>0.22</td>
<td>0.69</td>
<td>0.53</td>
</tr>
<tr>
<td>LTR</td>
<td>0.49</td>
<td>0.18</td>
<td>0.64</td>
<td>0.44</td>
</tr>
<tr>
<td>FCS</td>
<td>0.55</td>
<td>0.24</td>
<td>0.58</td>
<td>0.46</td>
</tr>
<tr>
<td>STM</td>
<td>0.45</td>
<td>0.22</td>
<td>0.62</td>
<td>0.43</td>
</tr>
<tr>
<td>DNN-HMM_L</td>
<td>0.45</td>
<td>0.18</td>
<td>0.58</td>
<td>0.40</td>
</tr>
<tr>
<td>DNN-HMM_s</td>
<td>0.25</td>
<td>0.18</td>
<td>0.57</td>
<td>0.33</td>
</tr>
<tr>
<td>DNN-HMM_t</td>
<td>0.26</td>
<td>0.16</td>
<td>0.50</td>
<td>0.31</td>
</tr>
<tr>
<td>DNN-HMM_m</td>
<td>0.28</td>
<td>0.13</td>
<td>0.49</td>
<td>0.30</td>
</tr>
</tbody>
</table>
are obviously improved, but that of baby crying event declines instead. From the Fig. 3, we can find that the detection of baby crying does not show any apparent over-fitting phenomenon. Therefore no rectifying training was required. It indicates that transfer learning plays a more prominent role in correcting over-fitting phenomena, whereas the detection of non-overfitting events may be slightly disturbed.

DNN can achieve good learning effects for simple classification tasks, while BNN undertook classification tasks for all events resulting in more difficult studying challenge. To help the network learning toward the desired direction, we added a simple 2-class classification task to inspire the training of the network. This task regarded all events including baby crying, glass breaking, and gunshot as one class named event class, and unified all kinds of scene noise into a non-event class. The task was relatively simple and easy to learn. The purpose of this design is to let the network distinguish the event from non-event sound first and then subdivide the various types of events in the event class. Moreover, the outputs of BNN increased two-dimensions for event class and non-event class correspondingly. The system with multi-task learning processing DNN-HMM_m showed the best detection performance making ER reduced to 0.3. This 2-class classification task focuses on enhancing the detection of events by only using robust frame features. Therefore, it can effectively improve the detection performance for gunshot and glass breaking events. However, at the same time, as the secondary task ignored the temporal features, the detection of baby crying event had been weakened. We concluded that multi-task learning could efficiently help the network training, while the secondary task should be designed based on the characteristics of the events carefully.

IV. CONCLUSIONS

We proposed a new DNN-HMM hybrid system based on multi-states transition model, exploring the impact of state transition topologies and number of state classes on the performance of SED. And we redesigned the DNN training strategy by using transfer learning and multi-task learning to handle the over-fitting phenomenon. The performance of the proposed method was compared with the MLP system and standard DNN-HMM system based on LTR states transition model. Our proposed method had the best performance and achieved an average error rate of 0.3 on the event-based evaluation.

In future, we will try to improve the system performance in three ways. First, the log Mel-band energies feature still loses some useful information, and we should find more suitable features for the sound event detection. Second, DNN is not suitable for modeling audio data with strong temporal structure. We need to try to use more types of network structures to describe the spectral features, such as CNN, LSTM. At last, the HMM can only characterize the temporal characteristics to a certain extent, while the way of choosing state transition model still appears to be not flexible enough, we need to find a more efficient way to describe the temporal structure.

REFERENCES