DNN-BASED SOUND EVENT DETECTION WITH EXEMPLAR-BASED APPROACH FOR NOISE REDUCTION

Inkyu Choi, Kisoo Kwon, Soo Hyun Bae and Nam Soo Kim

Seoul National University
Department of Electrical and Computer Engineering and INMC
Gwanak P.O.Box 34, Seoul 151-744, Korea
{ikchoi, kskwon, shbae}@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

In this paper, we present a sound event detection system based on a deep neural network (DNN). Exemplar-based noise reduction approach is proposed for enhancing mel-band energy feature. Multi-label DNN classifier is trained for polyphonic event detection. The system is evaluated on IEEE DCASE 2016 Challenge Task 2 Datasets. The result on the evaluation set yields up to 0.787 and 0.3660 in terms of F-Score and error rate on segment-based metric, respectively.

Index Terms— Sound event detection, deep neural network, exemplar-based noise reduction

1. INTRODUCTION

Sound event detection (SED) plays an important role in computational auditory scene analysis, with a specific purpose of detecting meaningful sounds, generally referred to sound events. Detecting sound events such as speech, footstep and door slam provides fundamental information for understanding the situation using acoustic signal. Furthermore, SED could be utilized in many applications, including automated surveillance systems, information retrieval, smart home systems and military applications.

Many previous works on SED were based on conventional speech recognition techniques. The most common approach is to use a system based on spectral features such as Mel-Frequency Cepstral Coefficients (MFCCs) and Hidden Markov Models (HMMs) for sound event classification [1, 2]. In recent works, approaches based on Support Vector Machine (SVM) [3, 4, 5] or non-negative matrix factorization (NMF) [6, 7, 8] were also proposed for SED. Most of the previous works were monophonic SED, which focused on detecting a single event at the same time. However, more than two events can happen simultaneously in real environments. In this case, conventional monophonic SED approaches may not be suitable for detecting overlapping events. Polyphonic SED aims to detect multiple sound events in the same time instance of the sound data. A polyphonic AED system that used MFCC for feature and HMMs as classifiers with consecutive passes of the Viterbi algorithm was proposed [9]. In [10], Generalized Hough transform (GHT) voting system has been used to recognize overlapping sound events. In another work, NMF-based approach was used for source separation and then events were detected from each stream [11]. Deep neural networks (DNNs) have shown good performance for polyphonic SED by modeling overlapping sound events in a natural way [12].

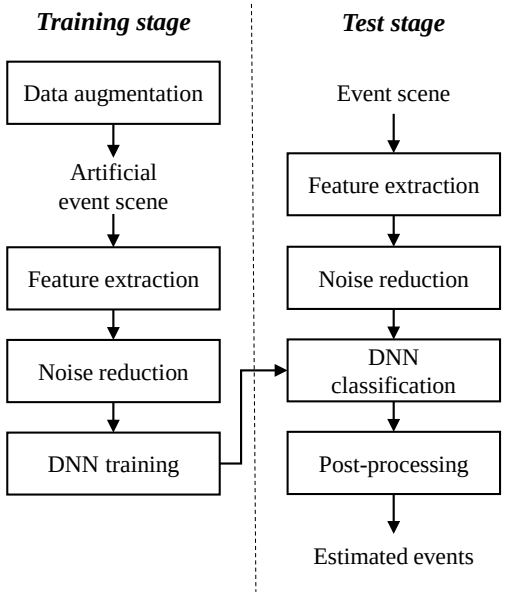

Figure 1: Flowchart of the proposed system

In this paper, we propose a DNN-based SED system. In the proposed system, data augmentation is performed to deal with data sparsity problem in small training dataset and generate polyphonic event examples. Exemplar-based noise reduction algorithm is proposed for feature enhancement. DNN classifier is trained for polyphonic event detection and adaptive thresholding algorithm is applied as a post-processing for robust event detection in noisy condition.

2. THE PROPOSED SED SYSTEM

The proposed system consists of 4 main processing stages. The overall system is illustrated in Fig. 1. First, data augmentation is performed to generate artificial sound event scenes which are used for training the classifier. In the second stage, mel-band energy features are extracted and enhanced by exemplar-based noise reduction. Third, the enhanced feature is fed to a DNN classifier. The features from artificial sound event scene are used for training the DNN classifier. In final stage, the sound events are detected by filtering and thresholding the output of the DNN classifier.

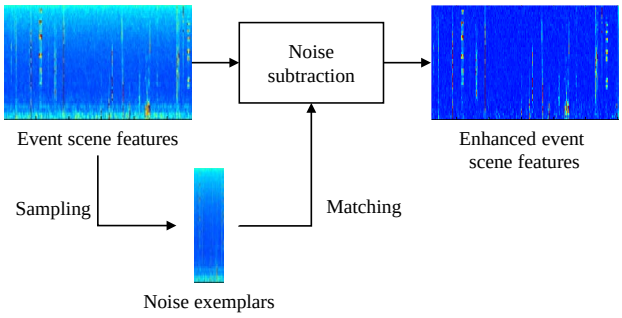

Figure 2: Exemplar-based approach for noise reduction

2.1. Data augmentation

DNNs have shown good performance as classifiers in many applications. When the training data is large, the DNN could learn from the variations presented in the training data under the same labels and make classifications that are robust to intra-class variations. However if the training data from each class is not sufficient to cover its intra-class variations, the DNN classifier trained with the data may have poor generalization ability, leading to low classification performance for test samples. In [13], data augmentation approach was used for training DNNs to deal with the data sparsity problem.

Unlike speech datasets which usually consist of hours of data or more, conventional sound event dataset is not sufficiently long enough to train a robust DNN classifier. Under this condition, data augmentation can help to enhance the performance of the DNN classifier by improving the generalization ability of the neural network. In recent research, data augmentation approach was performed for better performance in polyphonic SED [14]. In this paper, to construct the diverse sound event data from a small dataset, artificial event scenes are generated using data augmentation. In the artificial event scenes, events are overlapped to each other or manipulated by time stretching and power modification for diversity of dataset. These event scenes are corrupted by white, blue and pink noises.

2.2. Exemplar-based approach for noise reduction

In real life recordings, various noises exist and make it difficult to detect sound events correctly. To alleviate the effect of the noises, noise reduction is performed for feature enhancement. Since we assume that the test noise conditions are unknown, model adaptation-based approaches for noise robustness may not be suitable. In order to suppress unseen noises in test conditions, exemplar-based noise reduction approach is proposed. In this approach, noise exemplars are selected from the event scene features, then noise is directly subtracted from the event scene features by using the noise exemplars.

For each event scene, mel-band energy features are extracted and the features that have L1 norm corresponding to the lower 30% are considered to be noise candidates. From the candidates, K noise frames are selected randomly or using K-means algorithm for noise exemplars. For each frame, best matching noise exemplar that minimizes the noise estimation error, defined as in (1), is selected.

$$E_k = ||max(X_t - N_k, 0)||_1 + \alpha \cdot ||max(N_k - X_t, 0)||_1 \ . \quad (1)$$

E_k is the noise estimation error of a noise exemplar N_k and X_t is a feature vector at time index t. Noise estimation error E_k is the summation of under estimation error and over estimation error with ratio α. The selected noise exemplar is subtracted from the frame feature for noise reduction. The proposed noise reduction process is illustrated in Fig. 2.

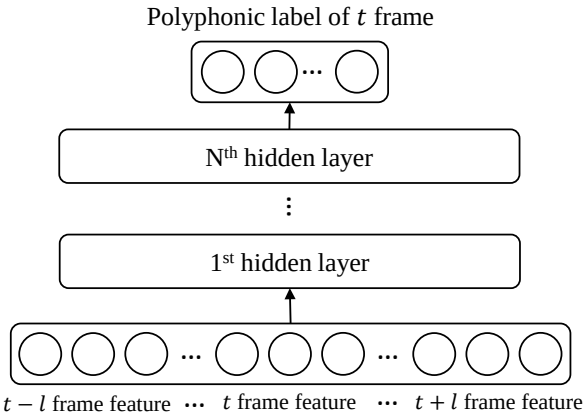

Figure 3: A DNN structure for the proposed SED system

2.3. DNN Classifier

In this paper, we trained a DNN-based classifier for SED. Unlike speech, sound events come from different physical sources so they possess unique characteristics that are distinct from one another. The DNN structure is employed to successfully represent distinct sound events in a single model. The DNN system for SED is illustrated in Fig. 3. The DNN consists of an input layer, a few hidden layers and an output layer which are fully connected to their adjacent layers. As for the input, the mel-band energy features enhanced by the proposed noise reduction approach are used. To consider temporal information, several adjacent frame features are concatenated for a single frame input. The output of the DNN is the estimated labels for input frames. The number of output unit is the same as that of the event classes, and each output unit is matched to each class. When the event exists in input frame, the output unit of the class is set to 1, otherwise it is set to 0. We used rectified linear activation function for hidden layers and sigmoid function units for the output layer.

Artificial event data generated by data augmentation is used for training the DNN classifier. In the fine-tuning stage, backpropagation algorithm with the minimum mean squared error (MMSE) function between the correct label and the estimated label is employed to train to the DNN. A stochastic gradient descent algorithm is performed in mini-batches to improve learning convergence. To deal with overfitting problem, we used the dropout technique which has already proved its regularization capability for training DNN [15].

2.4. Post-processing

The output of the DNN classifier is filtered for robust event detection. An averaging filter may help to remove outliers, but also discourage precise detection in onset or offset period of an event due to non-event periods nearby. For precise onset and offset detection, we used two filters: one of which is a sigmoid function and the other

is the former reflected about the y-axis. The former one is sensitive to the onset and the latter one is sensitive to the offset of an event. To detect both onset and offset of an event correctly, larger values of the output of two filters are taken from both output of the filters.

Generally, static threshold value is used for detection. However, in noisy event scenes, static threshold value can lead to high false detection error rate when the noise has similar characteristic with the events. To consider the noise effect on detection, adaptive threshold value is used as in (2) ,

$$T_i = T_{base} + \beta \cdot S_i \tag{2}$$

where T_i is an adaptive threshold value of class i, T_{base} is a base threshold value, S_i is mean of the DNN output of the class i in the event scene which reflects noise similarity with class i, and β is ratio value for S_i. When noise characteristic is similar to class i, T_i gets higher and reduces false detection error rate of class i.

3. EXPERIMENTAL RESULT

In order to evaluate the performance of the proposed system, we conducted a SED experiment on IEEE DCASE 2016 Challenge Task 2 Train/Development Datasets [16]. The training dataset was composed of mono recordings of isolated acoustic events typically found in an office environment. 11 classes were available: clearthroat, cough, doorslam, drawer, keyboard, keys, knock, laughter, pageturn, phone, speech and each class was represented by 20 recordings in training dataset. The development dataset consisted of 18 two-minute recordings in various noise and event density conditions. Only training dataset and noises sampled from probability density functions were used for learning the system and development dataset was used for evaluation.

Data augmentation was performed for generating the training event scene. Each sound event scene was about two-minute long. All events in the training dataset were normalized to have the same power and 30 of them were randomly selected for one event scene. To diversify the training data, half of the events were manipulated by stretching the time at a $\pm 10\%$ rate and modifying the power in the range of $50\% - 200\%$. One third of the events were overlapped to each other for polyphonic event examples. To consider the effect of noise on events, white Gaussian noise at signal-to-noise ratio (SNR) levels 6 to 18 dB and pink noise and blue noise at SNR level 12 dB were mixed. Total 110 artificial event scenes were generated for training the system.

We used mel-band energy as input features. Instead of original frequency 44.1 kHz, we used the sampling frequency of 30 kHz, spanning 50 bands between 100 Hz and 15 kHz. We used a hamming window with a frame length of 30 ms and a frame shift of 10 ms for frame segmentation. For noise reduction, $K = 100$ noise exemplars are selected and α is set to 0.5. As training data and test data may have power mismatch, the features extracted from each event scene are normalized.

For training the DNN-based classifier, 50-dimensional mel-band energy features were used as input. The input layer for DNN was formed by applying a context window of 11 frames, having 550 visible units for the network. The DNN had 3 hidden layers with 768 hidden rectified linear units with in each layer and the final sigmoid output layer had 11 units, each corresponding to the event classes. The parameters of the network were initialized by random values sampled from zero-mean normal distribution. The fine-tuning of the network was performed using mean squared error as the loss function by error back propagation supervised by the correct label of frames. The mini-batch size for the stochastic gradient descent algorithm was set to be 128. The learning rate was initially set to be 0.015 and exponentially decayed over each epoch with decaying factor 0.99 after fifth iteration. The momentum was set to be 0.7. The training was stopped after 80 epochs. The dropout percentage of 20% was applied for regularization.

Table 1: Average detection results on IEEE DCASE 2016 Challenge Task 2 Development Dataset

Metrics	Segment-based	Event-based
Precision	0.9311	0.7553
Recall	0.9211	0.8367
F-score	0.9261	0.7939
Substitutions	0.0091	0.0152
Deletions	0.0698	0.1481
Insertions	0.0590	0.2559
ER	0.1379	0.4192

Table 2: Average detection results on IEEE DCASE 2016 Challenge Task 2 Evaluation Dataset

Metrics	Segment-based	Event-based
F-score	0.787	0.671
ER	0.3660	0.6178

In post-processing stage, two 21-tap sigmoid shape filters are applied for smoothing output of the DNN. Larger values are taken from both output of the filters and thresholded for event detection. We set T_{base} to 0.6 and α to 0.5 for adaptive thresholding. Same events within 200 ms gap are concatenated and events shorter than 100 ms are removed.

As evaluation measures the F-Score and the error rate (ER) are used on Segment-based level. The F-Score F is the harmonic mean of precision P and recall R. The ER is the total number of insertions I, deletions D and substitutions S relative to the number of reference events N.

$$F = \frac{2P \cdot R}{P + R}\,, \quad ER = \frac{S + D + I}{N} \tag{3}$$

The results on the development dataset, averaged over the 18 synthetic audio event scenes are shown in Table 1. F-score and ER on segment-based metrics are 0.9261 and 0.1379, respectively. On event-based overall metrics, F-score and ER are 0.7939 and 0.4192, respectively. For DCASE 2016 task 2 challenge evaluation, both training data and development data are used for training DNN classifier. In Table 2, the results on the evaluation dataset are shown. F-score and ER on segment-based metrics are 0.787 and 0.3660, respectively. On event-based overall metrics, F-score and ER are 0.671 and 0.6178, respectively.

4. CONCLUSION

We presented a SED system based on a DNN. We used data augmentation to deal with data sparsity problem and exemplar-based approach for noise reduction. We trained a DNN for classification and filtering and adaptive thresholding are used for detecting events.

The proposed system has shown promising results on IEEE DCASE 2016 Challenge Task 2 Datasets.

5. ACKNOWLEDGMENT

This research was supported in part by the National Research Foundation of Koera (NRF) grant funded by the Korea government (MEST) (NRF-2015R1A2A1A15054343), and by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP(Institute for Information & communications Technology Promotion).

6. REFERENCES

[1] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, "HMM-based acoustic event detection with AdaBoost feature selection," *in Classification of Events, Activities and Relationships Evaluation and Workshop*, 2007.

[2] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," *in Proc. European Signal Processing Conference*, 2010.

[3] A. Temko and C. Nadeu, "Classification of acoustic events using SVM-based clustering schemes," *Pattern Recognition*, vol. 39, pp. 682–694, 2006.

[4] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, pp. 1281–1288, 2009.

[5] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro,, "Non-speech audio event detection," *ICASSP*, 2009.

[6] M. Chin and J. Burred, "Audio event detection based on layered symbolic sequence representations," *ICASSP*, 2012.

[7] J. F. Gemmeke, L. Vuegen, B. Vanrumste, and H. Van hamme, "An exemplar-based NMF approach for audio event detection," *in Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2013.

[8] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," *ICASSP*, 2015.

[9] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–13, 2013.

[10] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised hough transform," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1085–1093, 2013.

[11] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," *ICASSP*, 2013.

[12] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multilabel deep neural networks," *in Int. Joint Conf. on Neural Networks (IJCNN)*, 2015.

[13] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *ICASSP*, 2014.

[14] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," *ICASSP*, 2012.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] IEEE DCASE 2016 Challenge, `http://www.cs.tut.fi/sgn/arg/dcase2016/`, 2016.