Multi-frame Concatenation for Detection of Rare Sound Events Based on Deep Neural Network

Jun Wang

Beijing University of
Posts and Telecommunications
wangjun19930314@bupt.edu.cn

Shengchen Li

Beijing University of
Posts and Telecommunications
shengchen.li@bupt.edu.cn

ABSTRACT

This paper proposes a Sound Event Detection (SED) system based on Deep Neural Network (DNN). Three DNN-based classifiers are trained for detecting three target sound events including baby cry, glass break and gun shot from the audio streams provided. This paper investigates the influence of different frame concatenation when detecting sound events. Our results illustrate that the number of frames concatenated affects the accuracy of SED. The SED system proposed is tested by Development Datasets provided by Detection of Rare Sound Events in DCASE Challenge 2017. The average accuracy of the detection is that F-score and Error Rate (ER) on event-based metrics are 84.98% and 0.28, respectively.

Index Terms— DNN, sound event detection，frame concatenation

1. INTRODUCTION

With the development of Internet technology, large amounts of audio and other multimedia data are uploaded to website. These data, such as environmental recordings and speech, contain different kinds of sounds from a variety of sources. It can provide important information to detect sound events such as baby cry, glass break, footstep, gun shot and door slam for helping people analyze acoustic signal in life. There are many applications of sound event detection including multimedia indexing [1], intelligent monitoring system in living environment [2, 3, 4], scene classification and recognition [5, 6, 7], automatic audio tagging [8], audio segmentation [9], and health care [10], etc.

The most common approaches of sound event detection are based on conventional techniques of speech recognition. Sound event detection often includes two parts including feature extraction and classification. In previous works, classical features including Mel-Frequency Cepstral Coefficients (MFCCs) [11] and Constant Q Transform (CQT) [12] are used for representing audio signal. The classifiers used for SED include Hidden Markov Models (HMMs) [13] and Gaussian Mixture Models (GMMs) [14]. More recently, deep learning have shown good performance for sound events detection. Choi et al [15] propose a noise reduction approach to enhance mel-band energy feature in DNN-based system. An approach of data augmentation was proposed for acoustic modeling using DNNs [16]. Hayashi et al [17] propose a new method of recognizing daily human activities based on a DNN classifier and this method has achieved good performance in identifying different types of daily activities.

When DNN is used as a classifier, the input of DNN is a fixed number of frame concatenation, which is shown in Table 1. There lacks investigations on how the number of frames concatenated affects the accuracy of SED.

Table 1: Number of frames concatenated in previous works

Work	Sampling Rate	Frame Size	Overlap	Consecutive frames
[15]	30KHz	30ms	1/3	11
[16]	8KHz	-	-	9
[17]	16KHz	1s	0%	11
[18]	-	25ms	40%	10
[19]	16KHz	25ms	60%	60
[20]	-	-	-	10
[21]	44.1KHz	40ms	50%	5

Table 1 reviews the number of frames concatenated taken as the input of classifier in previous works. Although frame concatenation is a common method of SED, the way to select the number of frames for concatenation is yet to be investigated.

In this paper, we discuss the influence of frame concatenation by building a DNN-based SED system for Detection of Rare Sound Events which is the Task 2 of DCASE 2017 Challenge [21] and try different numbers of frame concatenation.

Detection of Rare Sound Events uses artificially created mixtures. This specific use of data will allow creating mixtures of background everyday audio and sound events of to be detected. In this task, the term *rare* refers to that every target sound event to be detected could occur at most once within a half-minute period. There are three target sound events to be detected: baby cry, gun shot and glass break. In every mixture, there is only one known target sound event selected from baby cry, gun shot and glass break to be detected. For each of the three target sound event classes, three separate DNN-based systems are develop to detect the temporal occurrences of these events.

The reminder of the paper is organized as follows. Second 2 discusses the development dataset synthesized. Section 3 describes the feature extraction. Section 4 describes the deep DNN structure and parameters for training. Section 5 provides three post-processing methods. Section 6 describes related experiment and result about influence of concatenation. Second 7 draws conclusion of our work.

2. DATASET

The audio material used includes source files for creating mixtures of sound events with background audio. Artificially created mixtures is used to develop the three DNN-based SED systems for each type of events. Sound events which are known are of the following classes: baby cry (106 instances for training, 42 instances for test), glass break (96 instances for training, 43 instances for test), gun shot (134 instances for training, 53 instances for test). The statistics data for the duration of the sound event in development datasets is shown in the Table 2.

Table 2: The statistical data the duration of sound events

Class	Baby cry		Glass break		Gun shot	
Usage	Train	Test	Train	Test	Train	Test
Mean	2.41s	1.85s	1.36s	0.72s	1.43s	1.04s
Maximum	5.1s	4.24s	4.54s	1.82s	4.4s	3.68s
Minimum	0.66s	0.78s	0.26s	0.3s	0.24s	0.3s
Median	2.33s	1.67s	1.29s	0.7s	1.21s	0.76s
Standard deviation	0.98s	0.83s	0.75s	0.3s	0.86s	0.83s

From the Table 2 we can see that the duration of the target event is usually short comparing with the background noise each lasting 30 seconds. Datasets for training and testing are generated respectively. Mixtures are generated with source data provided including background noise and target events by software package. In this paper, there are 500 mixtures generated each with 30 ms for every target class in both training dataset and test dataset. The probability of the existence of target event is 0.5, that is, 250 synthetic audios have target events and other 250 synthetic audios do not have target events. The isolated sound events and background noise selected randomly are synthesized at different event-to-background ratios (EBR) which was defined as a ratio of average Root Mean Squared Error (RMSE) value calculated over the duration of the event and the corresponding background segment on which the event will be mixed, respectively. The background instance, the event instance, the event timing in the mixture and EBR value in the synthetic audio are selected randomly.

3. FEATURE EXTRACTION

In order to analyze the influence of different frame concatenation when detecting sound events, Mel features are extracted to represent audio signal in this paper. Due to the development of Mel band filter [22], Mel features are widely used in the classification and detection of audio events because of carrying time domain and frequency domain information. Then, 40 log mel-band energies are extracted within 40 millisecond (ms) frames with 20 ms overlap in this paper.

4. CLASSIFICATION

To investigate the influence of different frame concatenation for different sound events, three different DNN-based classifier are trained respectively for detecting three types of target sound events. DNN-based system [15, 23] has been widely employed to detection of sound event. The DNN is a kind of fully connected network and the output of the former layer is used as the input of latter layer. Three DNN systems of this paper all consist of an input layer, two hidden layers and an output layer. The structure of DNN system is illustrated in Figure 5.

The input of the DNN is taken as a concatenation of sever adjacent frame features, which includes the center frame, the several preceding frames, and several succeeding frames. Different numbers of adjacent frame features is taken for different target sound events [24]. The output of the DNN is the estimated labels for input frames. Three different DNN systems all have two hidden layers, each with 50 neurons. Rectified Linear Unit (ReLU) activation function [25] is used in hidden layers. For all DNN systems, output layer has two neurons with softmax used as activation function. Here, output layer has two neurons instead of one neuron because the target class active or inactive are seen as two states, and the two states are mutually exclusive (only one state is valid at any time). Compared with one neuron in output layer, to some extent, the use of two neurons plays a role in the expansion of the features.

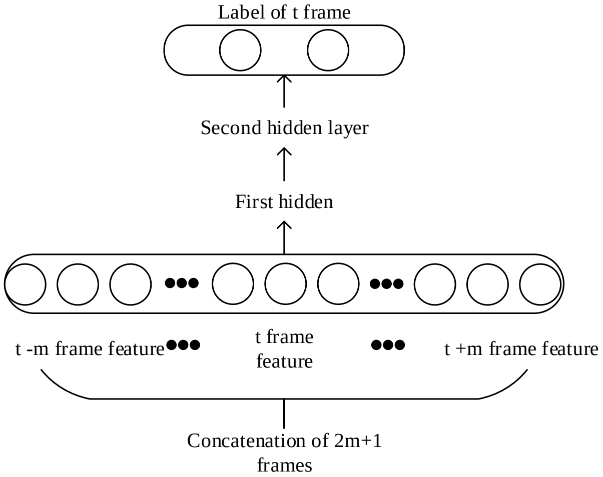

Figure 5: A DNN structure for the proposed SED system

In the training stage, all development datasets are used in sound event detection to improve the performance of the proposed system. For all development datasets, 90% of the training datasets are used for training and 10% of training datasets are used for validation to prevent overfitting; while all test datasets are used for testing the performance of the SED system. In order to make full use of the computing resources, "tensorflow" framework is used to build deep neural networks in this paper. The parameters of the network were initialized by random values sampled from zero-mean normal distribution. Three DNN-based systems are trained by using back-propagation with cross-entropy loss function, correct labels and estimated labels. A stochastic gradient descent algorithm [26, 27] is performed using Adam algorithm optimization [28] in mini-batches to improve learning convergence. Dropout technique [29] is used to prevent overfitting problem.

5. POST-PROCESSING

While occurrence of outliers result in poor precise detection in onset or offset period. The existence of outliers is not good to research the influence of different frame concatenation as the input of SED system when detecting sound events. The effective post-processing method can reduce the influence of outliers and improve the accuracy of detection. In this paper, we propose a three-stage post-processing to deal with predicted results of DNN classifier. The first stage is to filter the predicted values through dynamic threshold to reduce the difficulties of classifier learning due to imbalance between positive and negative samples. The next stage is median filtering for smoothing the output of dynamic threshold. The last method of post-processing is end to end, considering that each target sound event occurs at most once within a half-minute audio for reducing insertions of system output that are not correct.

A. Dynamic Threshold

Considering that target sound events occur at most once within a half-minute recording, the target event lasted for a short period of time, which resulting in unbalance between positive and negative samples. The percentages of positive and negative samples for three target sound events are shown in Table 3.

Table 3: Percentages of positive and negative

Target event	Positive	Negative
Baby cry	4.13%	95.87%
Glass break	2.47%	97.53%
Gun shot	2.47%	97.53%

Tabel 3 illustrates that there is a huge unbalance between the positive and negative samples, so that the predicted value of the output is sometimes small. If the influence of the predicted value of the output is not taken into account and the static threshold is set unreasonably, it is difficult for DNN-based system to detect target sound event. Dynamic threshold [15] is a method to solve that problem when average value of system output is seen as a kind of parameter considered. The dynamic threshold is expressed as follows:

$$T_i = T_{base} + \beta \cdot S_i \tag{3}$$

Where T_i is a dynamic threshold value of audio i, and T_{base} is a basic threshold value, and S_i is the average value of the DNN-based system output of the audio i and β is ratio value for S_i.

B. Median Filtering

When a target sound event in a continuous audio stream is to be detected, the target sound event is continuous, will neither occur alternately, and nor jump from one audio event type to another suddenly. So the detection results of some sound events in the continuous audio stream can be smoothed and processed by median filtering to further improve the detection accuracy. For example, if detection result in a segment of continuous audio stream is 1-1-0-1-1, the results will be smoothed to 1-1-1-1-1 through the median filter, where 0 indicates the target class inactive and 1 indicates the target class active.

In addition to smoothing the results of the threshold filter, median filtering has an influence on the onset and offset of the event. In the predicted results, several values could appear before the onset of event or after the offset of event, and these values far away from the onset or offset of the event cause that the prediction results are unreasonable. Median filter will filter out these values seen as outliers to ensure that the prediction results are close to the real values.

C. End to end

After median filter can filter out some outliers, the SED system considers the discontinuous results from median filter as the occurrence of multiple events, while there is only one and continuous target event to be detected. There is a contradiction between the output of median filter and the real situation. Then end to end is proposed to solve this problem. Each target sound event in mixture is taken as one time period in the prediction process. In the label, the location of target event in accordance with one audio feature is continuous 1. So, in the prediction process, all the numbers between the first one and the last are seen 1.

6. EXPERIMETAL RESULTS

Our system is on the basis of baseline system provided by DCASE2017 Challenge [21]. Three DNN systems are trained separately for three different target sound events.

For evaluation of system performance, the proposed system includes evaluation of results using event-based error rate (ER) and event-based F-score as metrics [30]. Event-based metrics compare system output and corresponding reference event by event. The F-Score 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.

The calculation of ER and F-score are as follows:

$$\mathrm{ER} = \frac{S+D+I}{N} \tag{4}$$

$$\mathrm{F} = \frac{2P \cdot R}{P+R} \tag{5}$$

where $\mathrm{P} = \frac{TP}{Tp+FP}$, $\mathrm{R} = \frac{TP}{TP+FN}$, TP represents correctly detected events, FP represents events in the system output that are not correct according to the definition, FN represents events in the ground truth that have not been correctly detected according to the definition, I represents events in system output that are not correct nor substitutions, D represents events in ground truth that are not correct nor substituted and s represents events in system output that have correct temporal position but incorrect class label.

In this paper, we investigate the influence of different frame concatenation. Different numbers of frame concatenation have been tried in experiments. The numbers of frame concatenation we choose various from 1 frame to 33 frames. Then, we evaluate the performance of different frame concatenation. The result of experiments is described in Table 4.

Table 4 Numbers of frame concatenation and performance

	Baby cry		Glass break		Gun shot	
	ER	F-Score	ER	F-score	ER	F-score
1	0.28	85.53%	0.21	88.98%	0.78	48.40%
3	0.29	84.06%	0.23	87.28%	0.67	55.47%
5	0.30	85.03%	0.20	89.46%	0.60	66.07%
9	0.27	86.07%	0.17	90.75%	0.56	66.67%
11	0.32	84.19%	0.14	93.07%	0.53	68.56%
13	0.29	86.63%	**0.12**	**93.98%**	0.52	69.77%
15	0.30	85.09%	0.15	92.28%	0.55	68.89%
17	0.28	86.12%	0.13	93.36%	0.51	71.04%
19	0.34	87.76%	0.15	92.02%	0.49	72.52%
21	**0.26**	**86.42%**	0.18	90.91%	0.50	71.00%
23	0.30	84.87%	0.15	92.24%	0.53	69.30%
25	0.34	83.06%	0.14	92.56%	**0.46**	**74.55%**
27	0.31	83.47%	0.19	89.98%	0.50	71.53%
29	0.44	78.19%	0.18	90.49%	0.48	72.77%
31	0.29	84.95%	0.19	90.11%	0.47	73.10%
33	0.40	79.84%	0.20	88.94%	0.46	73.61%

It is obvious that different numbers of frame concatenation are sensitive to different events. For baby cry, the performance is best at the frame concatenation of 21; glass break preforms best at the frame concatenation of 13; gun shot performs best at the frame concatenation of 25. Because the properties of three target sound events are different, baby cry often have long duration and the energy distributes in the duration of event; glass break is a type of impulsive sound that duration is short and the energy is mainly distributed under 10kHz; gun shot is also a kind of impulsive sound, and energy is mainly concentrated under 1.5kHz. Considering different properties of sound event, different sound events have different effects when using different number of frame combination. So different events should use different number of frames for concatenation. It is unreasonable to use the same number of frame combinations in previous work.

The final results of our DNN-based SED system on the development dataset are shown in Table 5 and the final results of baseline [21] provided by organizer of DCASE 2017 Challenge are show in Table 6.

Table 5: Results of our DNN-based SED system

	ER	F-score
Baby cry	0.26	86.42%
Glass break	0.12	93.98%
Gun shot	0.46	74.55%
Average	0.28	84.98%

Table 6 Results of baseline

	ER	F-score
Baby cry	0.67	72%
Glass break	0.22	88.5%
Gun shot	0.69	57.4%
Average	0.53	72.7%

Table 5 illustrates that the DNN classifier has a significant difference in the performance for detecting different sound events. For detecting glass break, the DNN classifier shows the best performance. The next is performance of detecting baby cry. The DNN classifier gets the worst performance when detecting gun shot. We assume that the DNN classifier is not suitable for detecting all sound events and different sound events have different sensitivities for classifier models.

Comparing Table 5 and Table 6, we found that performance of our system is better than the performance of the baseline system provided. Fixed consecutive frames are used as the input of classifier in baseline system while different consecutive frames are used as the input of classifier in our system. In case that DNN structure of our system is the same as the baseline system, the fact that our system outperforms baseline system further illustrates that different numbers of frame concatenation should be suggested as the input of SED system when detecting different sound events

7. CONCLUSIONS

In this paper, we first build a DNN-based system for Detection of Rare Sound Events to detect three sound events which are baby cry, glass break, and gun shot. Then, we investigate the performance of different frame concatenation as the input of SED system when detecting sound events. The results confirm our assumption that different events have different performance under different consecutive frames. So we should take number of frame concatenation into consideration when detecting different sound events.

8. REFERENCES

[1] D. Zhang and D. Ellis, “Detecting sound events in basketball video archive,” *Dept. Electron. Eng. Columbia Univ*, 2001.

[2] D. Stowell and D. Clayton, “Acoustic event detection for multiple overlapping similar sources,” *2015 IEEE Work. Appl. Signal Process. to Audio Acoust. WASPAA 2015*, pp. 1–12, 2015.

[3] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J. E. Appell, and F. Wallhoff, “Acoustic monitoring and localization for social care,” *J. Comput. Sci. Eng.*, vol. 6, no. 1, pp. 40–50, 2012.

[4] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Context-dependent sound event detection,” pp. 1–13, 2013.

[5] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Advances in neural information processing systems*, 2009, pp. 1096–1104.

[6] S. Allegro, M. Büchler, and S. Launer, “Automatic sound classification inspired by auditory scene analysis,” *Consistent Reliab. Acoust. Cues Sound Anal. CRAC oneday Work. Aalborg Denmark Sunday Sept. 2nd 2001 directly before Eurospeech 2001*, vol. 2005, no. 18, pp. 1–4, 2001.

[7] B. Picart, S. Brognaux, and S. Dupont, “Analysis and automatic recognition of Human BeatBox sounds: A comparative study,” *2015 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 4255–4259, 2015.

[8] M. Shah, B. Mears, C. Chakrabarti, and A. Spanias, “Lifelogging: Archival and retrieval of continuously recorded audio using wearable devices,” *2012 IEEE Int. Conf. Emerg. Signal Process. Appl. ESPA 2012 - Proc.*, pp. 99–102, 2012.

[9] G. Wichern *et al.*, “Segmentation , Indexing , and Retrieval for Environmental and Natural Sounds,” vol. 18, no. 3, pp. 688–707, 2010.

[10] J. Schroeder, S. Wabnik, P. W. J. van Hengel, and S. Goetze, “Detection and Classification of Acoustic Events for In-Home Care,” *Ambient Assist. Living*, pp. 181–195, 2011.

[11] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, “HMM-based acoustic event detection with adaboost feature selection,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4625 LNCS, pp. 345–353, 2008.

[12] A. Dessein, A. Cont, and G. Lemaitre, “Real-time detection of overlapping sound events with non-negative matrix factorization.” pp. 341–371, 2013.

[13] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, “Acoustic event detection in real-life recordings,” *18th Eur. Signal Process. Conf.*, pp. 1267–1271, 2010.

[14] L. Vuegen, B. Van Den Broeck, P. Karsmakers, J. F. Gemmeke, and B. Vanrumste, “An MFCC-GMM approach for event detection and classification,” *IEEE*

AASP Chall. Detect. Classif. Acoust. Scenes Events, no. 2, pp. 2–4, 2013.

[15] I. Choi, K. Kwon, S. H. Bae, and N. S. Kim, "Dnn-Based Sound Event Detection With Exemplar-Based Approach for Noise Reduction," no. September, 2016.

[16] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Detection," 2016.

[17] T. Hayashi, M. Nishida, N. Kitaoka, and K. Takeda, "Daily activity recognition based on DNN using environmental sound and acceleration signals," *Signal Process. Conf. (EUSIPCO), 2015 23rd Eur.*, pp. 2306–2310, 2015.

[18] 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)*, 2016, pp. 6460–6464.

[19] A. Gorin, N. Makhazhanov, and N. Shmyrev, "Dcase 2016 Sound Event Detection System Based on Convolutional Neural Network," no. October, pp. 11–14, 2016.

[20] Q. Kong, I. Sobieraj, W. Wang, and M. Plumbley, "Deep Neural Network Baseline for DCASE Challenge 2016," no. September, pp. 3–7, 2016.

[21] A. Mesaros *et al.*, "DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System," no. November, 2017.

[22] L. Deng and O. Abdel-hamid, "A Deep Convollutional Neural Using Heterogeneous Pooling For Trading Acoustic Invariance With Phonetic Confusion," *Acoust. Speech Signal Process. (ICASSP), 2013 IEEE Int. Conf.*, pp. 6669–6673, 2013.

[23] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 international joint conference on neural networks (IJCNN)*, 2015, pp. 1–7.

[24] A. Scenes, "Deep Neural Network Baseline for DCASE Challenge 2016," no. September, pp. 3–7, 2016.

[25] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proc. 27th Int. Conf. Mach. Learn.*, no. 3, pp. 807–814, 2010.

[26] R. Johnson and T. Zhang, "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction," *Nips*, vol. 1, no. 3, pp. 315–323, 2013.

[27] P. Abbeel, "Policy Gradient," *Control*, no. i, pp. 1–6, 2008.

[28] A. T. Hadgu, A. Nigam, and E. Diaz-Aviles, "Large-scale learning with AdaGrad on Spark," *Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015*, vol. 2, pp. 2828–2830, 2015.

[29] G. Hinton, "Dropout : A Simple Way to Prevent Neural Networks from Overfitting," vol. 15, pp. 1929–1958, 2014.

[30] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for Polyphonic Sound Event Detection," *Appl. Sci.*, vol. 6, no. 6, p. 162, 2016.