Augmented Strategy For Polyphonic Sound Event Detection

Bolun Wang*, Zhong-Hua Fu*†, Hao Wu*
* School of Computer Science, Northwestern Polytechnical University, Xi'an, China,
† Xi'an IFLYTEK Hyper Brain Information Technology Co., Ltd.
E-mail: blwang@mail.nwpu.edu.cn, mailfzh@nwpu.edu.cn, 107023224@mail.nwpu.edu.cn

***Abstract*—Sound event detection is an important issue for many applications like audio content retrieval, intelligent monitoring, and scene-based interaction. The traditional studies on this topic are mainly focusing on identification of single sound event class. However, in real applications, several sound events usually happen concurrently and with different durations. That leads to a new detection task on polyphonic sound event classification along with event time boundaries. In this paper, we propose an augmented strategy for this task, which faces challenges of a large amount of unbalanced and weakly labelled training data. Specifically, the strategy includes data augmentation to enrich training set to eliminate data unbalance, a new loss function that combines cross entropy and F-score, and model fusion to integrate the powers of different classifiers. The performance of the strategy is validated on DCASE2019 dataset, and both the event and segment detections are significantly improved over the baseline system.**

***Index Terms*—Sound event detection, Data augmentation, Model fusion**

I. Introduction

Sound event detection (SED) is a research area in Computational Auditory Scene Analysis [1]. The goal of the task is to detect the onset and offset time of some specific events in sound clip. With the development of the audio interaction, SED is becoming more and more popular. SED systems have many applications: health care [2], environmental surveillance [3], self awareness of embedded systems [4] and segments detection in videos based on audio [5] [6]. All the SED systems can be classified into two main categories: monophonic and polyphonic systems. Monophonic systems aim at classifying the isolated sounds [7], while polyphonic systems aim at classifying the isolated sounds as a sequence [8] and recognising all the overlapping sound events [9].

Many SED systems have been proposed to explore the possibility of improvement of event detection. [10] used mel-frequency cepstral coefficients (MFCC) features and hidden Markov models (HMMs) as classifiers. Non-negtive matrix factorization[NMF] was proposed to do the task in [11]. It can classify the prominent events, overlapping events, however, can not be detected properly. Coupled NMF was used to overcome the overlapping problem in [12].

With the development of the Deep learning technologies, more and more state of the art performance were obtained by neural networks. In [13], log mel-band based features and DNN model were used. Afterwards, RNN-LSTM networks were used for multi label classification in [14], which got better performance than DNN model.

Detection and Classification of Acoustic Scenes and Events(DCASE) is a challenge for events and scenes detection since 2013 [15], and sound event detection task is a sub task of DCASE since the very begining of the DCASE, and get more realistic since DCASE 2016 [16]. The goal of the task is to evaluate systems for the detection of sound events, especially with a combined data set containing strong labeled data, weakly labeled data and unlabeled data, which is considered as a semi-supervised task, usually simulates better for real scenarios. There are many methods proposed to try to solve this semi-supervised problem. The state of the art of DCASE 2018 challenge proposed mean teacher system to use the large amount of weakly labeled data [17].

The following part of the paper is organised as follows: We describe the polyphonic sound event detection in section II. Section III describe the proposed augmented strategy methods for multiple sound event detection. Section IV gives a detail of our experiment on DCASE 2019 dataset and in section V we make conclusions for our work.

II. Sound Event Detection

A. Sound Event Classification

Sound event classification is typically a classification task, it outputs a classification result of given event set. In the classification tasks, the output is a event label in clip level, and the time boundaries of the event is ignored.

B. Sound Event Detection

Along with classification of the occurrences of events, the biggest difference between event detection and classification is that we have to estimate the onset and offset of the events. This means we do the event classification in segment level, instead of clip level as what we do in sound event classification. Monophonic sound event detection systems output a monophonic event class at each time frame. While polyphonic sound event detection can be seen as a frame level multi label classification system, it can output multiple event class at each time frame. Fig 1 shows the difference between monophonic sound event detection and polyphonic sound event detection.

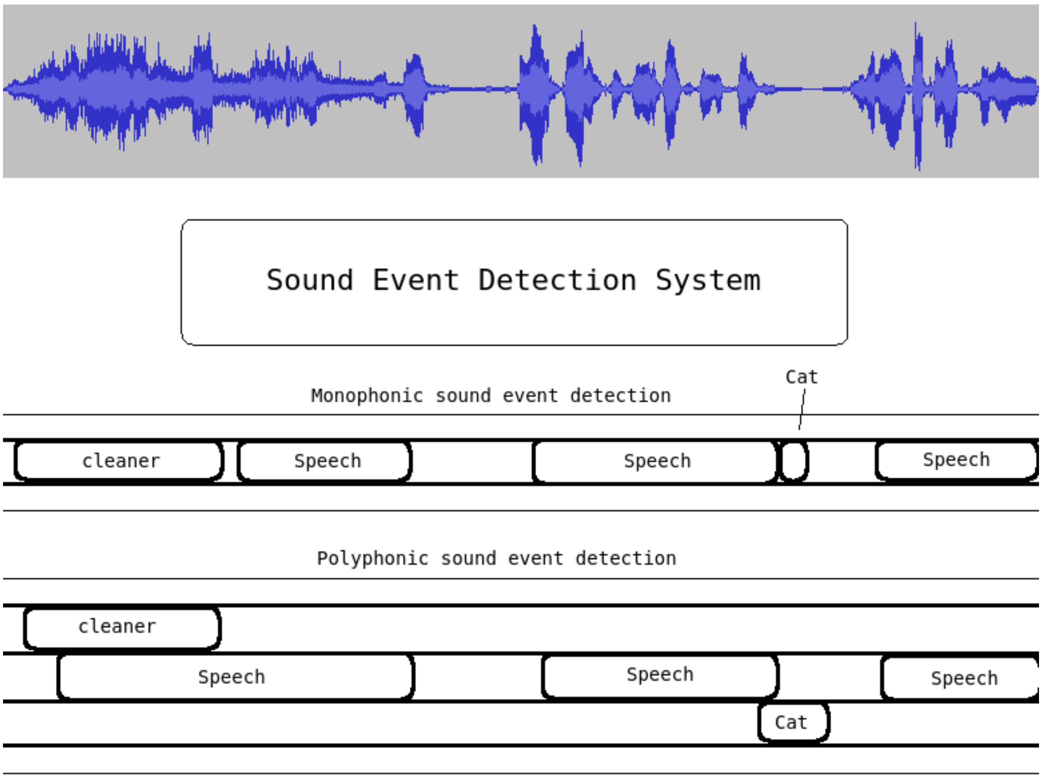
Sound Event Detection System
Monophonic sound event detection
Cat
cleaner
Speech
Speech
Speech
Polyphonic sound event detection
cleaner
Speech
Speech
Speech
Cat

Fig. 1. Sound Event Detection diagram

C. Dataset Limitation Problem

SED tasks require accurate onset and offset time stamps, which can be seen as a frame level multi label classification task. Because of the fact that it is usually too time consuming to annotate real data manually, SED systems are usually trained with a very small weakly annotated training set to get a better detection performance as possible as we can. The biggest challenge of this task is exploring the possibility to make use of a large amount of unbalanced and unlabeled training data with limited labelled dataset.

III. Methods

In this section, we present several methods, including data augmentation, and another optimization methods. Because the data in the challenge is limited, we try to use these methods to reduce the impact. First we do an event augmentation to solve the data unbalance problem, then follow an audio tagging augmentation. What's more, a new loss function was proposed to give ad direct optimization of sound event detection task and finally, we do a classwise model fusion. These methods are as follows:

A. Event Augmentation

Data augmentation method has been proven to be very useful for improving the generalization of the neural network [18], especially when the training data is extremely unbalanced. Events in the targeting set have different spectral properties, which have controversial optimization directions for neural network. The large amount data of specific event mislead the classifier to a wrong direction so that impact the overall performance. According to the idea in [19], we proposed an event augmentation method. With this method, we can mix up [18] events to the original training set as much as possible so that the data unbalance problem can be resolved. There are two steps of the event augmentation

- Event Extraction: Extract the events with the given time stamp to get multiple audio segments with labels
- Event Addition: Add the separated segments into the original training data with a random scaling factor α

$$y = \alpha x + (1 - \alpha)e \qquad (1)$$

Where $\alpha \in (0, 1)$ is the scaling factor, y is the augmented data, x is the original training data, e is the extracted event segment, all in time domain. With this mix up method, the target events are augmented comparing with other non-target segments in the same audio clip.

B. Audio Tagging Augmentation

Data driven methods' performance deeply depends on the training data in practice, but the data set in real life is extremely limited, especially for some specific tasks. To solve this issue, it is necessary to enrich the training set to use unlabeled data as possible as we can. There is a two stage method to do this [20].

- Audio tagging stage

- Sound event detection stage

Firstly a audio tagging model is trained using the original training data. During this stage the onset and offset time stamps are ignored, the only goal of this stage is to classify the event in clip level. Then the model is used to do the audio tagging for a large amount of unlabeled data, making an extra training set. To avoid too many wrong labeled training data are added in this stage, we set a very high classification threshold(0.9 in our experiment), and only those cases with probability over the threshold can be added to the training set. In the second stage, the training procedure is similar to the first stage, but instead of original training data, the model is trained with the combine of original training data and tagged data, which means we can get a better performance.

C. F1 Based Loss Function Augmentation

Cross entropy was successfully applied in classification tasks in the past research. The SED is a multi label classification task, so it was natural to apply it to SED. However the final performance evaluation criteria of the SED tasks is macro F1 score. To better optimize the networks, we proposed a new loss function to combine F1 score and traditional cross entropy with a weight factor:

$$J = \beta J_F + (1 - \beta) J_B \tag{2}$$

Where J_B is the binary cross entropy, which is usually used in multiple label classification tasks, and J_F is the proposed F1 loss function. This new loss function makes it possible to focus on either classification or F1 for different training data, and leads to a better performance for SED tasks. We did some experiments for different parameter β in following section.

D. Classwise Fusion

The classes differ in spectrum, it is normal to have totally different results for different classes, with the same model. So we ensemble the classifiers for different events. In this idea, one model is trained to classify one class in the worst case, then we combine all models together to get the final result.

IV. Experiment Setup

A. Dataset of Task4 in DCASE 2019

In DCASE 2019 challenge, the training set of task4 consists of tree parts: weakly labeled data, strong labeled data and unlabeled data. The target of task4 is 10 classes in domestic environment.

Weakly labeled data set contains 1578 clips, with 2244 class occurrences. The unlabeled data contains 14412 clips, The clips are selected such that the distribution per class is close to the distribution in the labeled set. The strong labeled data consists of 2045 clips, with 6032 class occurrences.

B. Audio Preprocessing

Firstly we resample the audio clips to 32000Hz, extract the log mel-spectrogram feature from the clips by a 64 mel bands, a 1024 hanning windows, 500 hop size, so we get 64 frames per second [21]. The overall pipeline of the system is as follows:

- Resample wave files to 32000Hz
- Extract events with strong time stamps of synthetic data
- Add events randomly into original training data in wave domain
- Extract log mel-spectrogram features

C. Network Configurations

The neural network is not our main focus for this work, so little modification is done. For most experiments, we use the network configurations in Table I, which is as similar as the network described in [21].

First we place two convolutional layers with kernel size 3x3, and number of filter 64, concatenated with ReLU activation and a batch normalization layer. After this, we add a pooling layer with size 2x2. Then we repeat these structure three times with different number of filters. Finally we concatenate a convolutional layer to the end of the network without pooling layer.

TABLE I
Network Configurations

item	config
Conv1	(3x3@64, BN, ReLU)x2
Pool1	2x2 pooling
Conv2	(3x3@128, BN, ReLU)x2
Pool2	2x2 pooling
Conv3	(3x3@256, BN, ReLU)x2
Pool3	2x2 pooling
Conv4	(3x3@512, BN, ReLU)x2

D. Evaluation Metrics

The performance of SED is evaluated by F-score(F1), both in segment level and event level. The submissions rankings is based on the event level F-score, all these are described in [22]. Precision(P) and Recall(R) are core evaluation statistics of SED tasks, which are defined as:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

Where TP, FP, FN are defined as true positive, false positive and false negative cases, respectively. P and R are the preferred metrics for information retrieval. In multi label classification, they are called calculated to get the final F-score:

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

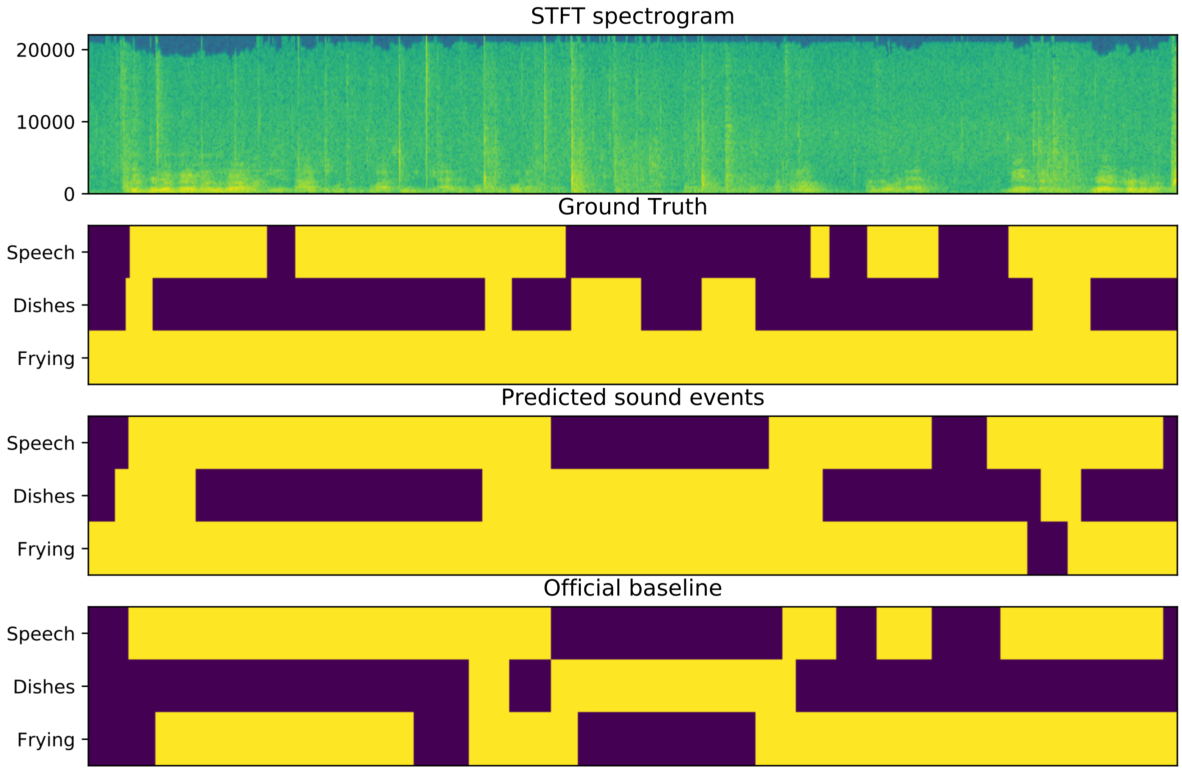

Fig. 2. Example of proposed SED system

TABLE II
F-SCORE FOR DIFFERENT METHODS

method	event F1	segment F1
Baseline	0.237	0.552
Event augmentation	0.247	0.603
Tagging augmentation	0.200	0.569
CRNN	0.187	0.560
F1 loss function	0.250	0.611
Classwise fusion	0.319	0.605

TABLE III
F-SCORE FOR DIFFERENT β

β	event F1	segment F1
1.0	0.208	0.527
0.9	0.250	0.611
0.7	0.214	0.598
0.5	0.240	0.602
0.3	0.220	0.60

E. Results and Discussion

Table II shows the of F-score result on validation set of DCASE 2019 challenge. Event augmentation outperforms baseline in event based F-score, which has no augmentation methods. Because of the event addition usually lasts only several segments, instead of whole clip, the augmentation result is even be better than baseline in terms of segment based F-score.

Tagging augmentation focus on audio tagging, which can be thought as same classification as SED task for unlabeled data, results show that audio tagging does positive contribution in segment based F-score, but not in event based. because there are too much mismatch in the event level prediction for tagging augmentation.

CRNN is a network concatenate RNN cells after normal CNN blocks. We can see that comparing with normal CNN, CRNN can not do too much contribution in segment F-score. What's worse, it does too much negative effect in terms of event based F-score, so we did all experiments on CNN afterwards.

Instead of traditional Binary Cross Entropy(BCE) loss for multi label classification, we propose a new loss function, which was combined by normal BCE loss and F1 loss with a weighting factor. The neural network was trained to optimize final performance criteria F1 directly. Results shows a giant improvement in both event based F1 and segment based F1.

Finally we made a classwise fusion, making the models to

TABLE IV
EVENT BASED F-SCORE OF FUSION

class	event F1
Speech	0.449
Dog	0.145
Cat	0.361
Alarm_bell_ringing	0.188
Dishes	0.132
Frying	0.385
Blender	0.295
Running_water	0.262
Vacuum_cleaner	0.46
Electric_shave_toothbrush	0.516
Overall	0.319

classify what they ”good at classifying”. Comparing with all other methods, we get a best performance in validation set of DCASE 2019, and this is also very close to the best result in DCASE 2018 [17], in terms of F1 score, Fig 2. shows an randomly selected case from the validation set of DCASE 2019 challenge.

Table III shows the impact of β on performance. We use a fixed weight in this experiment, the result shows the best parameter of β is 0.9. This leads a improvement than baseline.

Table IV shows event F-score for all class in the final fusion scheme. We can see that performance varies in all classes, so it is very necessary to do the fusion. The fusion method can optimize every class as better as possible simultaneously.

V. CONCLUSIONS

We proposed an augmentation strategy for Sound Event Detection task in this paper. In real life, data set is usually limited and unbalanced, which makes it difficult to get a satisfactory performance in sound event detection. This problem becomes even worse for those classes totally different in spectral. If we want to apply the model in real scenario instead of research, this issue must be taken consideration. The experiments show that data augmentation can handle this issue to get a better performance. The methods lead to different performance for classes, there is no such thing as a general model to fit well for all environments. Instead of dealing with all data in a huge classifier, model fusion is a method to do some several classifications in different branches, which have similarity in spectral, then fusion all results together. The experiment results show that our strategy makes a great improvement in validation data set of DCASE 2019.

ACKNOWLEDGMENT

The research work is supported by the National Key Research and Development Program of China under Grant No. 2017YFB1002102.

REFERENCES

[1] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[2] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. Tsai, “Healthcare audio event classification using hidden markov models and hierarchical hidden markov models,” in *2009 IEEE International Conference on Multimedia and Expo*. IEEE, 2009, pp. 1218–1221.

[3] S. Ntalampiras and I. Potamitis, “Detection of human activities in natural environments based on their acoustic emissions,” in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 1469–1473.

[4] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, “Where am i? scene recognition for mobile robots using audio features,” in *2006 IEEE International conference on multimedia and expo*. IEEE, 2006, pp. 885–888.

[5] R. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, and L.-H. Cai, “A flexible framework for key audio effects detection and auditory context inference,” *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 3, pp. 1026–1039, 2006.

[6] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, “Audio keywords generation for sports video analysis,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 4, no. 2, p. 11, 2008.

[7] H. D. Tran and H. Li, “Sound event recognition with probabilistic distance svms,” *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 6, pp. 1556–1568, 2010.

[8] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events,” *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[9] 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.

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

[11] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, “Supervised model training for overlapping sound events based on unsupervised source separation,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8677–8681.

[12] O. Dikmen and A. Mesaros, “Sound event detection using non-negative dictionaries learned from annotated overlapping events,” in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[13] 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)*. IEEE, 2015, pp. 1–7.

[14] 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.

[15] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. Plumbley, “Detection and classification of acoustic scenes and events,” *Multimedia, IEEE Transactions on*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[16] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, “Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[17] L. JiaKai, “Mean teacher convolution system for dcase 2018 task 4,” DCASE2018 Challenge, Tech. Rep., September 2018.

[18] Y. Sakashita and M. Aono, “Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions,” DCASE2018 Challenge, Tech. Rep., September 2018.

[19] M. Lasseck, “Acoustic bird detection with deep convolutional neural networks,” DCASE2018 Challenge, Tech. Rep., September 2018.

[20] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, “Large-scale weakly labeled semi-supervised sound event detection in domestic environments,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 19–23. [Online]. Available: https://hal.inria.fr/hal-01850270

[21] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, “Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems,” *arXiv preprint arXiv:1904.03476*, 2019.

[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