COMBINED SOUND EVENT DETECTION AND SOUND EVENT SEPARATION NETWORKS FOR DCASE 2020 TASK 4

Technical Report

You-Siang Chen[1], *Zi Jie Lin*[1], *Shang-En Li*[2], *Chih-Yuan Koh*[1], *Mingsian R. Bai*[1], *Jen-Tzung Chien*[2], *Yi-Wen Liu*[1]

[1] National Tsing Hua University, Hsinchu 30013, Taiwan, s108033851@m108.nthu.edu.tw, jasonlin.851206@gmail.com, jimmy133719@gapp.nthu.edu.tw, msbai@pme.nthu.edu.tw and ywliu@ee.nthu.edu.tw

[2] National Chiao Tung University, Hsinchu 30010, Taiwan, {w0860239.cm08g, jtchieng}@nctu.edu.tw

ABSTRACT

In this paper, we propose a hybrid neural network (NN) to handle the tasks of sound event separation (SES) and sound event detection (SED) in Task 4 of DCASE 2020 challenge. The convolutional time-domain audio separation network (Conv-TasNet) is employed to extract the foreground sound events defined in DCASE challenge. By comparing the baseline SED network with various training strategies, we demonstrate that the SES network is capable of enhancing the SED performance effectively in terms of several event-based performance metrics including macro F1 and poly-phonic sound detection score (PSDS).

Index Terms— Sound event separation, Conv-TasNet, Sound event detection

1. INTRODUCTION

Sound event detection (SED) is recently an active research topic in the areas of signal processing for machine learning. DCASE task 4 [1] aims to classify not only the sound event classes but also the event time boundaries. The baseline SED approach inspired by the mean-teacher model [2] relies on convolutional-recurrent neural network (CRNN) that has shown good capability of identifying the sound events with weakly labelled and unlabeled training data. In order to explore the possibility of improvement due to source separation, the participants are encouraged to develop a SED system combined with a sound event separation (SES) network. For separation task, Conv-TasNet was shown to yield significant improvement on speech separation [3]. Similar architecture of the network is recently used in universal source separation including the sound event separation [4]. Therefore, this study adopts the Conv-TasNet as the means of feature extraction intended for the baseline SED network. By using the proposed framework, the performance of SED is further enhanced. Different training strategies are also evaluated.

2. SOUND EVENT SEPARATION

Conv-TasNet is a single-channel deep neural network (DNN) consisting of learnable time-domain transformation and time-dilated convolutional network (TDCN). Based on the best separation result reported in [4], the learnable base window is chosen with 2.5 ms and the non-causal network is employed with the hyper-parameters summarized in Table 1.

Table 1. Hyper-parameters applied in the Conv-TasNet

Symbol	Parameter	Description
N	512	Filters in autoencoder
L	40	Filter length (Samples)
B	128	Channels in bottleneck
H	512	Channels in convolutional blocks
P	3	Kernel size in convolutional blocks
X	8	Convolutional blocks in each repeat
R	3	Number of repeats

The training objective is to maximize the scale-invariant source-to-noise ratio (SI-SNR) [5, 6] defined as

$$\begin{cases} \mathbf{s}_{\text{target}} = \dfrac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2} \\ \mathbf{e}_{\text{noise}} = \hat{\mathbf{s}} - \mathbf{s}_{\text{target}} \\ \text{SI-SNR} = 10 \log_{10} \dfrac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{noise}}\|^2} \end{cases} \tag{1}$$

in which $\hat{\mathbf{s}}$ and $\mathbf{s}$ are the extracted source events and the reference audio signals, respectively. The extracted outputs of the SES network include two branches of the desired foreground events in Domestic Environment Sound Event Detection (DESED) datasets [7, 8] and the background event mixtures in Free Universal Sound Separation (FUSS) dataset [9]. The permutation invariant training [10] was not employed to ensure the specific output of the network would be the desired separated audio events.

3. WORKFLOW

The workflow of the sound event separation and detection is depicted in Figure 1. Here, the SES network is the Conv-TasNet [3] trained in advance by the mixture of the DESED and FUSS dataset.

The SED network exploits the baseline model which is essentially a modified mean-teacher approach [11]. After the SES network, most of the labeled sound events are extracted at one specific output. The extracted signals are then averaged with the original audio mixture in the time domain to enhance the features of the desired sound events.

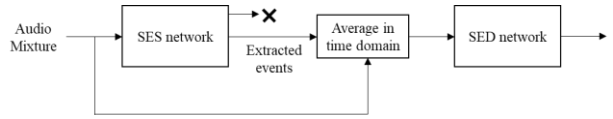

Figure 1. The workflow of the combined sound event separation and detection network.

4. CONSTRUCTION OF DATASETS

There are two portions for the dataset preparation. The SES and SED networks are trained separately and combined with the aforementioned framework. The sampling rate of the input audio is 16 kHz for both networks.

4.1.1. Dataset for the training of SES network

The SES dataset was created with Scaper [12], a soundscape synthesis and augmentation library. It is composed by the DESED foreground events and the mixture of the FUSS dataset served as the background noise. The duration of the audio file is 3-second long segment with maximum 4 and minimum 1 foreground events. In total, 9,600 training mixtures (8 hours) and 2,400 validation mixtures (0.67 hour) were generated. The goal of the SES network is to extract the desired foreground event mixtures and exclude the irrelevant background events and noise. The SES network was trained with 200 epochs and batch size 5. Adam [13] was used as the optimizer with the learning rate 10^{-3}. The model was selected with the best SI-SNR 8.5 dB in the validation set.

4.1.2. Dataset for the training of SED network

To further enhance the performance of the SED network, the original training set of DESED dataset including 1,578 weakly labeled, 14,412 unlabeled and 2,584 synthetic strong label data were fed into the SES network and averaged with the original audio signals in time domain. Next, the baseline SED model was trained by using the processed training set with the same architecture and parameters, where window size = 2048, hop size = 255, maximum frequency 8000 Hz, and 128 Mel bins are assumed. The network was initialized with the pre-trained weights of the original training set, followed by another 200 epochs of training.

To check if the preprocess of the separation framework is helpful to the SED network, we compared the baseline SED performance with three other different scenarios. The detailed descriptions of the framework and training strategies are summarized in Table 2. In case B, the validation set was processed by the SES network and further estimated by the original baseline SED network. On the other hand, in the case C and D, the original SED network was trained with the preprocessed training set using SES network. The main difference of these two cases is whether or not the validation set is processed by the SES network.

Table 2. Descriptions of different training scenarios and evaluations of the SED frameworks.

Case	Model	Description of dataset usage	
		Training set	Validation set
A	Baseline SED only	Original DESED set	Original DESED set
B	SES + baseline SED	Original DESED set	DESED set processed by SES
C	SES + trained baseline SED	DESED set processed by SES	Original DESED set
D	SES + trained baseline SED	DESED set processed by SES	DESED set processed by SES

5. SUMMARY OF RESULTS

The evaluation of the SED is based on the event-based measures with a 200 ms collar on onsets and a 200 ms of the events length collar on offsets. The averaged macro F1 score and PSDS [14, 15] are calculated with the validation set in different scenarios and the results are summarized in Table 3. In case B where the input signals are simply passed through the SES network, the F1 score has been slightly increased from 34.2 % to 35.5 %. In case D, with the DESED training set being preprocessed by the SES network, the performance can be further improved to 36.7%, which indicates that the SED task does benefit from the preprocessing of the SES network. In addition, in case C where we evaluate the non-processed validation set with the trained baseline SED, the F1 score also attains similar value, as compared with model D. This suggests that the SES network can alternatively serve as a means of data augmentation for the SED system.

Table 3. Class-wise averaged macro F1 score and PSDS macro score in different training strategies.

Cases with different training strategies	F1 (%)	PSDS (%)
Case A	34.2	59.6
Case B	35.5	59.7
Case C	36.4	61.9
Case D	**36.7**	**62.0**

Table 4. Class-wise F1 score for each event type in different cases.

Class	Case			
	A	B	C	D
Cat	43.8	43.1	**44.1**	43.2
Running water	32.6	30.9	**34.3**	31.6
Alarm/bell/ringing	**39.8**	39.4	37.6	38.3
Dishes	24.1	**25.7**	25.1	25.6
Speech	47.4	**49.4**	44.7	48.0
Electric shaver /toothbrush	36.1	40.0	**44.6**	38.6
Dog	20.4	**23.1**	19.8	21.6
Frying	23.1	22.4	**26.6**	25.1
Vacuum cleaner	44.9	45.6	**55.5**	**55.5**
Blender	30.3	35.5	32.1	**39.3**

The F1 score for each event type in different cases are summarized in Table 4. For Cases C and D, the major improvement in comparison to the baseline SED network lies in the detection of vacuum cleaner that produces a long and persistent noise. However, it turns out that sound events of this kind are easily handled in SES network as background noise separated into the unwanted branch. Therefore, averaging the mixture with separated signals will increase the contrast between the persistent sound event and short-period sound event, which increases the mean-teacher model's ability in recognizing long-duration events.

6. CONCLUSIONS

In this paper, we have presented a learning framework that combines the SES and SED networks by averaging the separated signals with the original input mixtures. As shown in the macro F1 and PSDS scores, the SES network helps to enhance the performance of the SED network. This suggests that the SES unit can be either integrated to the SED network, or alternatively be used as a data augmentation means to increase the diversity of the input features.

7. ACKNOWLEDGMENT

The work was supported by the Add-on Grant for International Cooperation (MAGIC) of the Ministry of Science and Technology (MOST) in Taiwan, under the project numbers 107-2221-E-007 -039 -MY3 and the MOST AI project with number 109-2634-F-009-024.

8. REFERENCES

[1] http://dcase.community/challenge2020/task-sound-event-detection-and-separation-in-domestic-environments.

[2] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for dcase 2019 task 4," *Technical Report*, Orange Labs Lannion, France, June 2019.

[3] Y. Luo and N. Mesgarani, "Tasnet: Surpassing ideal time-frequency masking for speech separation," *arXiv preprint arXiv:1809.07454*, 2018.

[4] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 175–179. October 2019.

[5] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," Interspeech 2016, pp. 545–549, 2016.

[6] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[7] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*. New York City, United States, October 2019.

[8] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing*. Barcelona, Spain, 2020.

[9] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, 411–412., 2013.

[10] M. Kolbak, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 10, pp. 1901-1913, Oct. 2017.

[11] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, 2017, pp. 1195–1204.

[12] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: a library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 344–348. New Paltz, NY, USA, Oct. 2017.

[13] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[14] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovi´c, "A framework for the robust evaluation of sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61–65, 2020.

[15] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, 6(6):162, 2016.