# ZHENG_USTC TEAM'S SUBMISSION FOR DCASE2021 TASK4 - SEMI-SUPERVISED SOUND EVENT DETECTION

Technical Report

*Xu Zheng, Han Chen, Yan Song*

National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China.
{zx980216, ch1180}@mail.ustc.edu.cn, songy@ustc.edu.cn

## ABSTRACT

In this technical report, we present our submitted system for DCASE2021 Task4: sound event detection and separation in domestic environments. Specifically, three main techniques are applied to improve the performance of the official baseline system with both synthetic and real data (weakly labeled and unlabeled). Firstly, in order to improve the localization ability of CRNN model, we propose to use the selective kernel(SK) unit. By stacking the SK unit, each neuron can adaptively adjust its receptive field for both short- and long- duration events. Secondly, based on the fact that detection outputs are dominated by the high-confidence predictions(lower than 0.1 or higher than 0.9), we propose to use soft detection output by setting proper temperature parameter in sigmoid, which can effectively improve the PSDS2 score. Thirdly, several data augmentation techniques and score fusion mechanisms are applied to improve the stability and robustness of the system performance. Experiments on the DCASE2021 task4 validation dataset demonstrate the effectiveness of the techniques used in our system. Specifically, PSDS scores of 0.45 and 0.78 are achieved for scenario1 and scenario2 respectively, outperforming the result of 0.34 and 0.53 in baseline system.

***Index Terms—*** Sound Event detection, Semi-supervised learning, mean teacher, PSDS, selective kernel

## 1. PROPOSED METHOD

### 1.1. Network architecture with SK unit

The selective kernel(SK) networks [1], where multiple branches with different kernel sizes are selected by a softmax attention mechanism, can capture target objects with different scales. In our previous works [2], we applied the SK unit for semi-supervised SED, significant SED performance are achieved for different event classes. Specifically, it achieves more than 20% improvement in event-based F1 score for short-duration events (Cat and Dishes), and over 10% improvement in event-based F1 for long-duration events (Frying and Running_water).

In this year's competition, we extend our previous work and apply the SK unit as the building block for CRNN model. For CNN part, the first stage is constructed by VGG block, then followed by 4-stage bottleneck resblocks with SK unit. And there are 2 resblocks in each stage. The number of filters for 5-stage CNN are respectively [32, 64, 128, 128, 128], and frequency pooling rate for each stage is set to 2, whereas the total temporal pooling rate is set from 2 to 8. The RNN part is the same as the baseline, and for localization module, linear softmax [3] pooling function is used for aggregating the detection output into audio tagging output.

### 1.2. Backend processing with temperature

In our experiments, we count the distribution of our system's detection output on unlabeled data, and the results are presented in Table 1. From the results, we see that our system provides too many high-confident predictions, where $96.59\%$ of our predictions are below 0.1 or higher than 0.9. However, given a threshold, too many predictions higher than 0.9 may cause many false-positive samples, and we believe detection performance may be benefited from softer predictions.

Table 1: Distribution for detection output on unlabeled data.

| detection output | Probability % |
|---|---|
| [0,1] | 100 |
| [0,0.4] or [0.6,1] | 99.81 |
| [0,0.3] or [0.7,1] | 98.84 |
| [0,0.2] or [0.8,1] | 98.04 |
| [0,0.1] or [0.9,1] | 96.59 |

Inspired by Knowledge distillation [4], where a temperature parameter $T$ is used to soften the softmax output. We apply the temperature parameter in the Sigmoid function to soften the detection output.

$$y_i = Sigmoid(z_i/T) = \frac{1}{1 + exp(-z_i/T)} \tag{1}$$

where $z_i$ ,$y_i$ donate logit and softened detection output for event class i.

It's worth noting that the purpose of setting temperature in our system or Knowledge distillation differs. In Knowledge distillation, the higher temperature produces a softer probability distribution over classes, whereas in our system, the temperature is applied to reduce the polarized distribution among detection output. Meanwhile, the implementation for setting temperature in our system also differs from that in Knowledge distillation. In Knowledge distillation, the higher temperature is only set in training stage, However, in our system, the temperature $T$ is set to 1 in training stage and higher $T$ is set during inference stage.

### 1.3. Data augmentation

Three data augmentations are applied in our system, namely spec-augmentation [5], shift consistency training(SCT) [6] and interpolation consistency training(ICT) [7].

In spec-augmentation [5], there are three data augmentation methods, warping the features, frequency masking, and temporal masking. In our implementation, only frequency masking is applied, which means that entire mel frequency bands are consecutively masked.

In time-shifting or shift consistency training(SCT) [6], predictions of time-shifted input are encouraged to be consistent with the time-shifted target. In our implementation, the synthetic strongly-labeled clips and the real data with pseudo-targets by teacher are shifted with size is set to 2s.

The interpolation consistency training(ICT) [7] encourages the prediction at an interpolation of unlabeled points to be as close as the interpolation of soft-prediction at those points.

## 2. EXPERIMENTS

### 2.1. Dataset and Feature Extraction

All experiments are conducted on the DCASE 2021 domestic environment sound event detection (DESED) [8] dataset, which is composed of real soundscapes and synthesized soundscapes. For real soundscapes, data can be divided into 5 subsets: weakly-labeled (1578 clips), unlabeled-in-domain (14412 clips), synthetic dataset(10000 clips), validation(1168 clips) and evaluation. The input features used in the proposed system are log-mel spectrograms, which are extracted from the audio signal resampled to 16000 Hz. The log-mel spectrogram uses 2048 STFT windows with a hop size of 313 and 128 Mel-scale filters. As a result, each 10-second sound clip is transformed into a 2D time-frequency representation with a size of (512×128).

### 2.2. Experimental Settings

The neural networks are trained using the Adam optimizer [9], with a maximum learning rate of 0.001, and a learning rate rampup during the first 20 epochs. Reduction rate r in the SK unit is set to 8. In our experiments,we save the best models for PSDS1 and PSDS2 separately, which can be further used for model ensembling.

## 3. RESULTS AND ANALYSIS

### 3.1. Evaluation for SK unit

Firstly, We evaluate the performance of SK unit, and the experimental results are shown in Table 2. As we can see, our CRNN model with fixed kernel outperforms the official CRNN model both in PSDS1 and PSDS2. After the SK unit is added, the performance of our CRNN model is further improved compared with the official baseline with PSDS1 increasing from 0.353 to 0.4082 and PSDS2 increasing from 0.553 to 0.6338. Therefore, SK unit is effective in this task.

### 3.2. Effectiveness for temperature parameter

Then, we evaluate the effect of temperature parameter on the SED performance, and the results are shown in Table 3. After the introduction of temperature parameter, the PSDS2 is improved, and the results shows that the larger this parameter is, the greater the performance improvement is. When the temperature value is set as 20, the PSDS2 reaches the maximum, which increases from 0.6175 to 0.7389 compared with the case without temperature parameter.

Table 2: Evaluation for our SK unit baseline model.

| Model | PSDS1 | PSDS2 |
|---|---|---|
| CRNN(official) | 0.353 | 0.553 |
| CRNN(ours)<br>with fixed kernel | 0.3624 0.3729 | 0.6030 0.6004 |
| CRNN(ours)<br>with SK unit | 0.4082 0.3951 | 0.6181 0.6338 |

Table 3: Evaluation for temperature parameter $T$.

| temperature $T$ | PSDS2 |
|---|---|
| 1 | 0.6175 |
| 2 | 0.6275 |
| 4 | 0.6961 |
| 6 | 0.7209 |
| 8 | 0.7314 |
| 10 | 0.7353 |
| 20 | 0.7389 |

### 3.3. Evaluation for data augmentation

Finally, we evaluate the effect of data augmentation on our model performance. The baseline system is the Mean teacher system, and spec-augmentation is added to the student model only. Table 4 lists the changes of PSDS1 after adding SCT and ICT. It can be seen that both SCT and ICT can improve PSDS1 at different temporal down-sampling rates. It is worth mentioning that the introduction of ICT has a little greater effect on performance improvement than SCT. When the temporal down-sampling rate is increased from 2 to 8, PSDS1 score decreases gradually, indicating that some useful information for detection is lost with large temporal down-sampling rate.

Table 4: Evaluation for data augmentation.

| Training<br>Method | Temporal<br>down-sampling rate | PSDS1 |
|---|---|---|
| Mean Teacher(MT) | 2 | 0.4082 0.3951 |
| MT+SCT | 2 | 0.4221 0.4126 |
| MT+ICT | 2 | 0.4207 0.4228 |
| MT | 4 | 0.3737 0.3988 |
| MT+SCT | 4 | 0.3939 0.3994 |
| MT+ICT | 4 | 0.4095 0.4102 |
| MT | 8 | 0.3460 0.3442 |
| MT+SCT | 8 | 0.3485 0.3507 |
| MT+ICT | 8 | 0.3671 0.3627 |

Table 5: Description for our submitted system .

| system id | model count | temperature T | PSDS1 | PSDS2 |
|---|---|---|---|---|
| 1 | 3 | 3 | 0.4535 | 0.6714 |
| 2 | 9 | 3 | 0.4539 | 0.6802 |
| 3 | 10 | 20 | 0.3972 | 0.7879 |
| 4 | 9 | 20 | 0.4021 | 0.7858 |

## 4. SUBMISSION SYSTEM

The systems we submitted are shown in Table 5. The four systems adopt the model fusion strategy. System 1 and system 2 integrate the models that are conducive to improving PSDS1 to obtain the maximum PSDS1, while system 3 and system 4 aim to obtain the maximum PSDS2. The largest PSDS1 is 0.4539, and the largest PSDS2 is 0.7879.

## 5. REFERENCES

[1] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[2] X. Zheng, Y. Song, I. McLoughlin, L. Liu, and L.-R. Dai, "An improved mean teacher based method for large scale weakly labeled semi-supervised sound event detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 356–360.

[3] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.

[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[5] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[6] C.-Y. Koh, Y.-S. Chen, Y.-W. Liu, and M. R. Bai, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 376–380.

[7] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *arXiv preprint arXiv:1903.03825*, 2019.

[8] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," June 2019, working paper or preprint. [Online]. Available: https://hal.inria.fr/hal-02160855

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.