SOUND EVENT LOCALIZATION AND DETECTION WITH VARIOUS LOSS FUNCTIONS

Technical Report

Sooyoung Park, Sangwon Suh, Youngho Jeong,

Electronics and Telecommunications Research Institute
Media Coding Group
218 Gajeong-ro, Yuseong-gu, Daejeon, Korea
{sooyoung, suhsw1210, yhcheong}@etri.re.kr

ABSTRACT

This technical report presents our system submitted to DCASE 2020 task 3. The goal of DCASE Task 3 is to detect a sound event and its location when a polyphonic sound event moves dynamically. We focus on designing loss functions to overcome the characteristics of the sub-task and imbalanced dataset. Temporal masking loss is used to overcome imbalance from zero labels of the silence frame. Soft floss is used for overcoming imbalance instances between class labels. A periodic loss function is proposed for regression that infers the periodic label in the direction of arrival estimation. Also, we take a feature pyramid network based network to overcome the information leakage occurred by the pooling layer in the CRNN.

Index Terms— Feature pyramid network, Temporal masking loss, Soft floss, Periodic loss

1. INTRODUCTION

In DCASE 2019 task 3, most of the participants used the state-of-the-art, CRNN-based networks. Multi-label classification is used to classify polyphonic sound events in the sound event localization and detection task [1]. In the direction of arrival estimation (DOAE) task, multi-label classification or multi-output regression is used to find the location of a polyphonic sound events.

CRNN-based models trained by binary cross-entropy (bce) loss were mainly used to solve sound event detection (SED) task. In the process of training the sound event detection network through such a system, it was often found that the binary cross-entropy loss and the binary accuracy for validation showed low values, even though the training was not fully performed. The first reason for these values is that the data used to detect sound events has a silence label between event labels and the number of instances of silence region is greater than the individual target class instance. The second reason is that the number of individual target class instance is different in given dataset [2]. For these reason, the given labels are imbalanced due to the configuration of the sound event data set. Therefore, we tried to overcome the imbalance from the silence region through the temporal masking loss function. By masking the loss value of the silence region, the neural network can more focus on the classification of the region where the sound event occurred. Also, we use soft floss [3, 4] to overcome performance degradation from the different number of target class instance in training dataset and to track training progress easily.

Distance loss, such as mean square error (MSE) or mean absolute error (MAE), is mainly used for regression analysis. The DCASE 2019 baseline uses a regression model with a polar coordinate label. In this case, the baseline cannot take into account the periodicity of the label. For example, the azimuth values of 180 and -180 degrees are in the same direction, but the distance losses mentioned above consider them to be completely different directions. Therefore, models trained with MSE or MAE do not use periodicity. Therefore, we propose a periodic loss function to use the periodicity of the label.

2. FEATURE

The development dataset consists of 4 channels of recording with the first-order ambisonic format and 4 channels of recording which are recorded from the tetrahedron microphone array. Each audio file was recorded with a 24 kHz sampling frequency. A total of 600 minutes recording was given for a development dataset with a length of 1 minute per file. Baseline system set 400 files for a training fold, 100 files for a validation fold, and 100 files for a test fold. For short-time Fourier transform, we use the Hanning window with 2048 nfft, 0.02s window length, and 0.01s hop length. For training, 200 frames (T) of data were used and the overlap is set to 50 frames (0.5s).

We use logmel energy, harmonic percussive separation (hpss), and intensity vector. For sound event detection, average logmel energy (1ch) and average logmel hpss (2ch) were used. For the direction of arrival estimation, logmel energy (4ch) and logmel intensity vector (6ch) including active and reactive parts were used. 256 mel bins were used for feature extraction.

3. NETWORK ARCHITECTURE

3.1. Proposed network

As mentioned in the introduction, it is difficult to determine the time or frequency pooling size when using CRNN. Setting too large pooling size causes loss of information, and setting too small pooling size can be a burden in training the network. We wanted to be free from the effort to determine the pooling size by using U-shape's feature pyramid network (FPN) [5]. PoolNet [6], one of the FPN based network, is a state-of-the-art network for salient object detection. PoolNet created a better feature representation

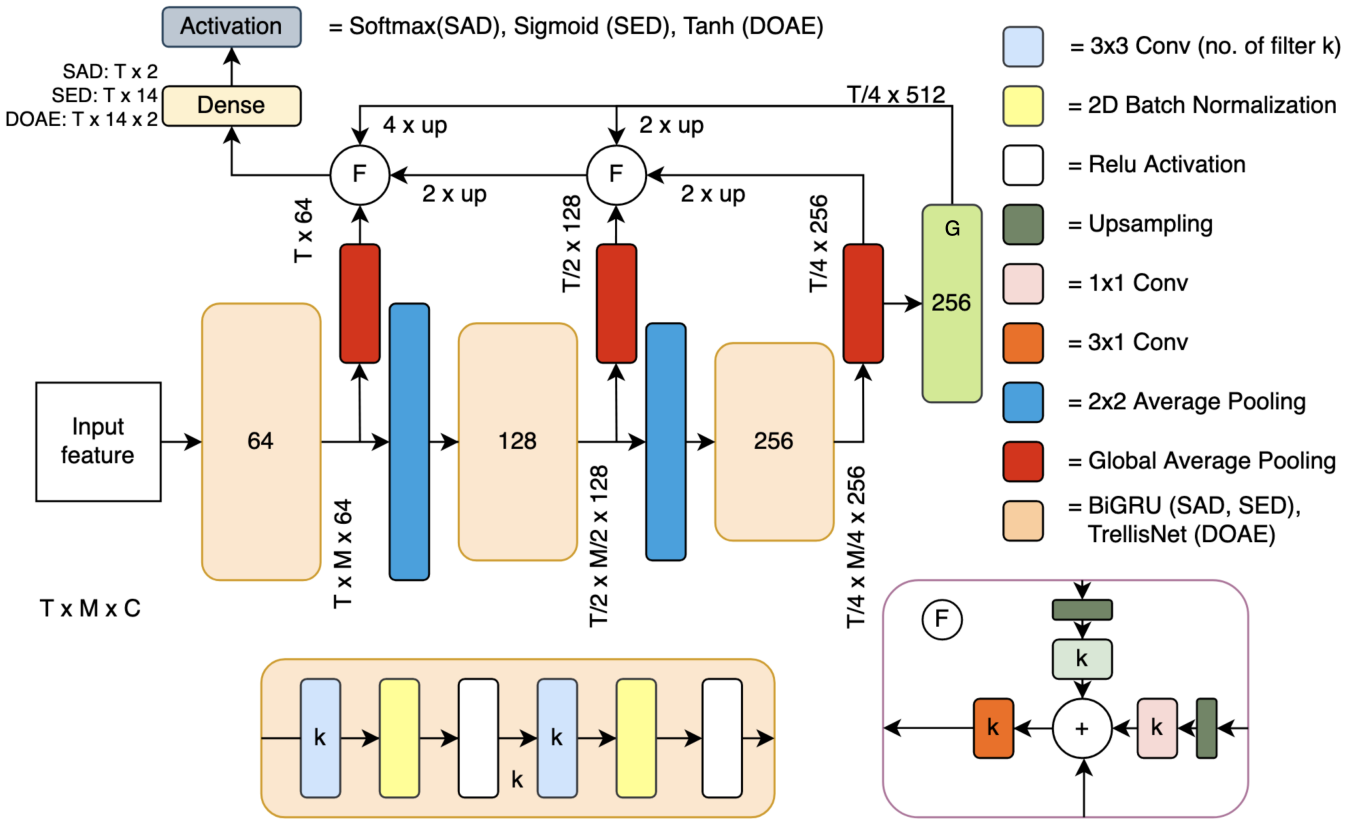

Figure 1: Proposed system for DCASE 2020 task 3; T: time, M: number of mel-bin, C: channel, G: global guidance module

using deep and shallow features.

We create a better intermediate-level feature representation using deep features compressed by pooling layers and those not. A structure that performs detection using the generated intermediate-level feature was adopted. PoolNet targets 2d output, so the network has been converted to 1D sequence output format. Our proposed system is shown in Figure 1.

Intermediate-level features are obtained by adding deep features and shallow features with upsampling. When creating an intermediate-level feature, the feature depth is set to be equal to the shallow feature through 1x1 convolution layer. We use 3x3 convolution layer to reduce the aliasing effect from upsampling. When creating an intermediate-level feature, a global guidance module is added so that the information of deep feature is not diluted. PoolNet use a pyramid pooling module as global guidance module, but our system use temporal network as global guidance module.

Global guidance module consists of BiGRU and TrellisNet [7]. Bidirectional GRU (BiGRU) was applied to sound activity detection (SAD) and sound event detection networks. On the other hand, TCN-based TrellisNet has strengths in DOAE [8]. We use TrellisNet for global guidance module in DOAE network.

3.2. Training Procedure

The training process of the proposed system is composed of three sub-network: SAD, SED, and DOAE. The SAD network solves the binary problem of whether or not a target sound event occurs. Our system uses temporal mask loss for training SED network. Temporal mask loss does not calculate the loss for the silence area which has no target sound event label. Therefore, the SAD network detects the silence label area. In the case of DOAE, only the direction loss of the active event is considered for multi-output regression training using the ground truth label.

3.2.1. Sound Activity Detection

In sound activity detection, a binary decision is made using a dense layer and softmax layer. These layers use a feature representation synthesized with the structure of the U-shape of the proposed network. The average logmel energy and the average logmel hpss were used as the input features for training. Since sound activity detection is a problem that infers time-varying patterns, we use BiGRU for global guidance module. The number of parameters of this network is 2,143,170. We use Ranger optimizer [9] with 200 epochs. The initial learning rate is 0.01. This learning rate is decayed by the cosine decay warm restart technique [10].

3.2.2. *Sound Event Detection*

In sound event detection, a multi-label classification is performed using a dense layer and sigmoid layer. The number of parameters of this network is 2,143,950. The rest of the settings are the same as the SAD network.

3.2.3. *Direction of Arrival Estimation*

In direction of arrival estimation, a multi-regression is performed using a dense layer and tanh layer. In this case, we use TrellisNet for the global guidance module. The number of parameters is 2,360,796. This network is trained with 300 epochs. The rest of the settings are the same as above.

3.3. Inference

The inference is performed in the following process using the three networks mentioned above. First, a frame with the probability of target sound activity greater than 0.5 is detected through the SAD network. Multi-label classification is performed using the SED network only in frames where the target sound is detected. Multi-regression for DOAE is performed using the DOAE network. The final result is derived by matching the DOA value corresponding to the SED result.

4. LOSS FUNCTIONS

Binary cross-entropy loss is mainly used for classification. A distance-based loss like mean square error or mean absolute error is mainly used for regression. In sound event detection datasets, there are many zero labels due to the silence area and one-hot encoding. Therefore the validation loss value is low enough from the beginning of training. This makes it difficult to track how much the network has been trained. Due to the gap between the loss and the metric, methods [3, 4, 11, 12] using a metric function as loss for training have been proposed.

The problem of distance-based loss function used in regression is that it is difficult to use periodic label characteristics. Therefore, a method of using a quaternion output [13] is proposed to take advantage of this periodic characteristic, but there are some difficulties in training. Therefore, we propose a periodic loss function to overcome this problem.

4.1. Soft floss

The F1 score used as a metric function in DCASE 2020 task 3 is calculated from true positive (TP), false positive (FP), and false negative (FN) expressed in equation (1) for all test data. y_k is the binary value obtained by thresholding the model output $\hat{y}_k$. Therefore, the F1 score is a non-differentiable function.

$$\begin{aligned} TP(Y,Y) &= \sum_k 1(y_k == 1 \text{ and } \quad y_k == 1), \\ FP(Y,Y) &= \sum_k 1(y_k == 0 \text{ and } \quad y_k == 1), \\ FN(Y,Y) &= \sum_k 1(y_k == 1 \text{ and } \quad y_k == 0), \\ F(Y,Y) &= \frac{2 \quad TP}{2 \quad TP + FN + FP} \end{aligned} \tag{1}$$

The F1 score function is modified as shown in equation (2) to enable differentiation.

$$\begin{aligned} TP(\hat{Y},Y) &= \sum_k \hat{y}_k \quad y_k, \\ FP(\hat{Y},Y) &= \sum_k \hat{y}_k \quad (1 \quad y_k), \\ FN(\hat{Y},Y) &= \sum_k (1 \quad \hat{y}_k) \quad y_k \end{aligned} \tag{2}$$

Assuming that TP, FP, and FN have occurred according to the model prediction probability distribution, each component can be transformed into a differentiable formula. Soft floss L_F [3, 4] is calculated as shown in equation (3) through the differentiable TP, FP, and FN.

$$\begin{aligned} L_F(\hat{Y},Y) &= 1 \quad \frac{2 \quad TP}{2 \quad TP + FN + FP} \\ &= 1 \quad F \end{aligned} \tag{3}$$

4.2. Temporal masking loss

There are several areas where the sound event does not occur due to the characteristic of the sound event dataset. So the number of silence instances is more than the number of individual target class instances. It causes an imbalance training dataset. We use temporal loss function which excludes the silence area for loss calculation. It is designed to erase the effect of the silence area in loss function as shown in equation (4). The network trained with temporal mask loss has not trained silence frames, so detecting silence frames, such as SAD network, is needed.

$$\begin{aligned} TP_{mask}(\hat{Y},Y) &= \sum_k \hat{y}_k \quad y_k \quad M_k, \\ FP_{mask}(\hat{Y},Y) &= \sum_k \hat{y}_k \quad (1 \quad y_k) \quad M_k, \\ FN_{mask}(\hat{Y},Y) &= \sum_k (1 \quad \hat{y}_k) \quad y_k \quad M_k \end{aligned} \tag{4}$$

We set the mask M_k to 0 for silence frames and 1 for frames with sound events. Therefore, equation (4) can exclude the silence frame in the loss calculation.

4.3. Sinusoidal loss

We propose a periodic distance loss function to consider periodic label value. A simple sinusoidal function is used to make a loss function considering the periodicity of distance as shown in equation (5). Equation (5) is designed considering the periodicity for

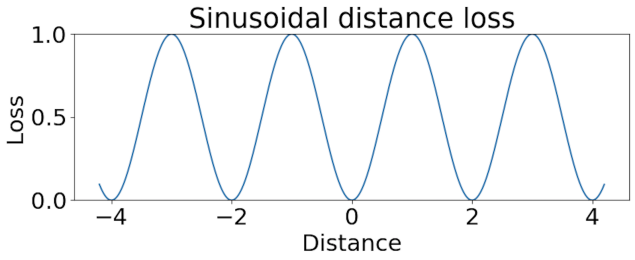

Figure 2: Proposed sinusoidal loss function

tanh output. Therefore, the period of proposed loss function is 2 as shown in Figure 2.

$$L_s(\hat{y}, y) = \sin^2\left(\frac{\pi}{2}\ \hat{y}\ \ y\right) \quad (5)$$

5. POST PROCESSING

We improve the result by applying post processing techniques to prediction result of our system. The median filtering is applied for SAD and SED. The number of filters in the median filter is set based on the average duration per class. In the case of DOAE, there are three possible angular speeds for moving sound events. So we use this prior knowledge to improve the result of the DOAE network. When the amount of change for the position of the sound event is constant, the result is modified using the given speed.

6. RESULT

The results of our proposed system are shown in Table 1. Our proposed system has better performance than the baseline system. The performance improvement was significantly improved at baseline for ER_{20°, F_{20°, and LR_{CD}. Our system improves about 2 degrees for LR_{CD}. Since LR_{CD} is calculated between the predictions and references of the same class, our system has been improved for more instances.

Table 1: Experimental results for development dataset

System	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
baseline (foa)	0.72	37.4 %	22.8°	60.7 %
baseline (mic)	0.78	31.4 %	27.3°	59.0 %
our system	0.59	52.8 %	21.0°	74.8 %
our system + DOA postprocessing	0.58	54.1 %	20.3°	74.3 %
our system + SED postprocessing	0.55	54.6 %	20.5°	76.0 %
our system + SED/DOA postprocessing	**0.54**	**55.6 %**	**20.0°**	**76.0 %**

7. SUBMISSION

The submission systems are constructed by changing the training split or adding an ensemble technique or augmentation technique.

ETRI_1: Proposed system trained using the same training split as the baseline system.

ETRI_2: Proposed system trained using the whole development dataset.

ETRI_3: Proposed system trained using the whole development dataset and snapshot ensemble technique [14].

ETRI_4: Proposed system trained using the whole development dataset, time stretch data augmentation, and snapshot ensemble technique.

8. ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00050, Development of Human Enhancement Technology for auditory and muscle support)

9. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, March 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8567942

[2] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv e-prints: 2006.01919*, 2020. [Online]. Available: https://arxiv.org/abs/2006.01919

[3] K. Zhao, S. Gao, W. Wang, and M. Cheng, "Optimizing the f-measure for threshold-free salient object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8848–8856.

[4] T. Tanaka and T. Shinozaki, "F-measure based end-to-end optimization of neural network keyword detectors," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1456–1461.

[5] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[6] J. Liu, Q. Hou, M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3912–3921.

[7] S. Bai, J. Z. Kolter, and V. Koltun, "Trellis Networks for Sequence Modeling," in *ICLR, International Conference on Learning Representations*, 2019, pp. 1–17.

[8] S. Park, "Trellisnet-based architecture for sound event localization and detection with reassembly learning," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 179–183.

[9] L. Wright, "New deep learning optimizer ranger: Synergistic combination of radam + lookahead for the best of both," 2019.

[10] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *ICLR*, 2017.

[11] Z. Fan, Z. Bai, X. Zhang, S. Rahardja, and J. Chen, "Auc optimization for deep learning based voice activity detection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6760–6764.

[12] E. Eban, M. Schain, A. Mackey, A. Gordon, R. Rifkin, and G. Elidan, "Scalable Learning of Non-Decomposable Objectives," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. Fort Lauderdale, FL, USA: PMLR, 20–22 Apr 2017, pp. 832–840. [Online]. Available: http://proceedings.mlr.press/v54/eban17a.html

[13] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Improvement of doa estimation by using quaternion output in sound

event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 244–247.

[14] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get M for free," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=BJYwwY9ll