CONFIDENCE REGULARIZED ENTROPY FOR POLYPHONIC SOUND EVENT DETECTION

Won-Gook Choi

Department of Electronic Engineering
Hanyang University, Seoul, Republic of Korea
onlyworld94@hanyang.ac.kr

*Joon-Hyuk Chang**

Department of Electronic Engineering
Hanyang University, Seoul, Republic of Korea
jchang@hanyang.ac.kr

ABSTRACT

One of the main issues of polyphonic sound event detection (PSED) is the class imbalance problem caused by the proportions of active and inactive frames. Since the target sounds occasionally appear, binary cross-entropy makes the model mainly fit on inactive frames. This paper introduces an effective objective function, confidence regularized entropy, which regularizes the confidence level to prevent overfitting of the dominant classes. The proposed method exhibits less overfitted samples and better detection performance than the binary cross-entropy. Also, we compare our method with the other objective function, the asymmetric focal loss also designed to solve the class imbalance problem in PSED. The two objective functions show different system characteristics. From an end-user perspective, we suggest choosing a proper objective function for the purposes.

***Index Terms*—** Polyphonic sound event detection, class imbalance problem

1. INTRODUCTION

Polyphonic sound event detection (PSED) is one of the acoustic classification and detection tasks that detects the target sound and timestamps in an audio signal. For many years, PSED has had following several challenges:

- Difficult to gather strongly labeled recordings.
- The subjectivity problem of manual labeling.
- Hard to find an analytic model that can cover the various sound patterns.
- The class imbalance problem due to the proportion of active and inactive frames.

The first problem has been solved with two approaches: semi-supervised learning approaches using both labeled and unlabeled data and training with synthetic audio mixed background noise and target sounds. The second problem could be relieved by choosing the metric when comparing the system with others [1]. And the third problem has been solved by deep neural netoworks using improved convolutional neural networks (CNNs), Transformer, attention mechanisms, etc [2].

This study focuses on the last problem of class imbalance. Most inactive frames (background sound) dominate an audio clip, so the problem arises when detecting a target sound frame by frame (Fig. 1). This phenomenon is not a problem presented only in PSED task; the image object detection also has suffered from the background-foreground class imbalance [3]. For the image object detection, a solution is the focal loss that controls the weight parameter of cross-entropy so that train the model well for the target object, but vice versa for the background images. Motivated by the focal loss, the previous study in PSED proposed asymmetric focal loss (AFL) [4] that could control the focal weights of entropies for the inactive and active terms, respectively. AFL successfully controlled the imbalance problem, but an adverse effect arose: the system detected repetitive impulsive sounds as a long-duration sound.

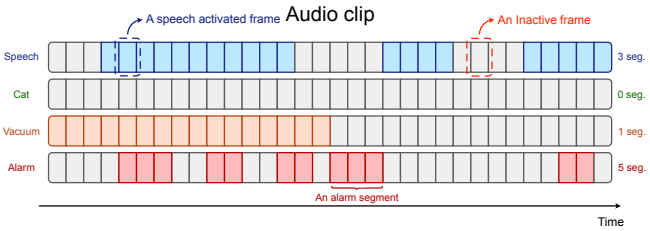

Figure 1: Simple illustration of an audio clip with target sounds.

In this study, we propose confidence regularized entropy (CRE), which set the confidence threshold to the binary cross entropy (BCE). When calculating the BCE, samples are eliminated for the backpropagation during training steps if the detected results are over the threshold. The proposed method keeps the samples less overfitted, especially for the inactive frames. Compared to the AFL, the proposed entropy resulted in a system that can detect the onsets and offsets of target sounds well. Both CRE and AFL relieved the class imbalance problem for PSED. However, they showed a different system characteristic: the CRE-based system was advantageous in detecting a target event's precise localization on frames, whereas the AFL-based system showed strength in detecting whether a target sound appeared. The details will be discussed in Section 5.2.

2. CLASS IMBALANCE PROBLEM WITH PSED

A class imbalance problem is one of the considerations for building and training a neural network. If the class imbalance problem remains unsolved, the model could remain ungeneralized [3]. When collecting data from real world, the target sound would appear intermittently rather than often; thus, one of the factors that cause class imbalance is the imbalance between the number of active and inactive frames [4] in dealing with the PSED tasks. Additionally, the imbalance among the target sounds could appear since each event's duration is entirely different, and the amount of recorded sound is

*corresponding author.

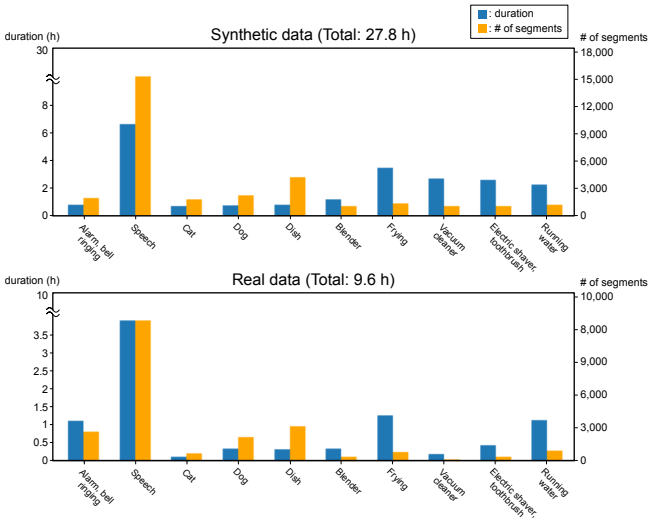

Figure 2: Bar graphs representing the duration of active sounds (blue) and the number of active segments (yellow). Considering the total audio length, the proportion of inactive frames is large.

also diverse according to the datasets.

The total duration and number of segments for each target sound composing the domestic environment sound event detection (DESED) dataset are shown in Fig. 2. In both synthetic and real data, inactive frames have large proportions and, all the events except speech have very low proportions.

Imoto *et al.* proposed asymmetric focal loss (AFL) [4] to control the entropies of active and inactive frames. The entropy between a ground truth $y_{n,c}$ and a model output $\hat{y}_{n,c}$ is described:

$$AFL = -\sum_{n,c=0}^{N,C} \{\underbrace{(1-\hat{y}_{nc})^{\gamma}\, y_{nc} \log(\hat{y}_{nc})}_{\text{Active term}} + \underbrace{(\hat{y}_{nc})^{\zeta}\, (1-y_{nc}) \log(1-\hat{y}_{nc})}_{\text{Inactive term}}\} \quad (1)$$

where γ and ζ denote the parameters to control the entropies of active and inactive frames in each, and N and C denote the number of frames and target sounds, respectively. If γ and ζ are set to 0, the entropy is same as the binary cross entropy, and the higher the values, the less focal. Imoto *et al.* set γ and ζ to 0.0625 and 1, respectively, which means that the objective function focuses more on the active frames.

3. CONFIDENCE REGULARIZED ENTROPY

Suppose that there are lot proportion of speech-activated and inactive frames among the datum. If then, the inactive points are converged earlier than the other target sounds (e.g., cat, vacuum cleaner, etc. in Fig. 2). Even if the training epoch is processed enough, the inactive points are still converging more closely to one or zero, whereas the network is less optimized for the other target sounds. To concentrate on training the network for the false positive and false negative data, we propose the confidence regularized entropy (CRE) that can regularize confidences so that they could not converge beyond the threshold.

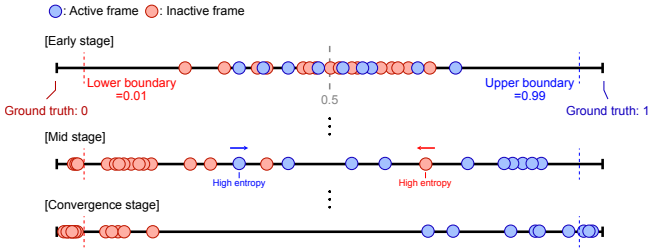

Figure 3: Training scenario when CRE is used for objective function.

$$CRE = -\frac{1}{NC}\sum_{n,c=0}^{N,C} \mathbb{I}_{|\hat{y}_{nc}-y_{nc}|>\gamma}(\hat{y}_{nc}) \cdot \{y_{nc} \cdot \log\hat{y}_{nc} + (1-y_{nc}) \cdot \log(1-\hat{y}_{nc})\}, \quad (2)$$

where $\mathbb{I}(\cdot)$ denotes an indicator function. We set γ to 0.01; the frames are excluded on each optimizing step, if those of confidence are either over the 0.99 or under 0.01. Generally, the mixup augmentation [5] is widely used for training the PSED network, and Eq. (2) also can be used whether the mixup is applied. If the mixup is used, confidences that are too close to the mixed labels are excluded during the training. In Section 5.1, the experimental results will demonstrate that a system that applied both CRE and mixup surpasses the system without either of them.

4. EXPERIMENTS

4.1. Dataset

To validate our proposed method, we used DESED database[1][6]. There were ten sound events that could occur in domestic environments. For the training set, there were 10,000 synthetic clips with strong annotations, 3,470 recorded clips with strong labels coming from the Audioset [7], 1,578 recorded clips with weak labels, and 14,412 unlabeled-recorded clips. For the evaluation set, there were 1,168 recorded clips. Each clip had a 10 s duration and was provided either 16 kHz or 44.1 kHz and single or dual channel. All clips were down-mixed to 16 kHz and extracted to log-mel spectrograms. For the details, window size and shift size were used 2048 and 255 samples, respectively, and 128 mel-filter banks.

4.2. CNN networks

The CNN architecture for the experiments is shown in Fig. 5. The group size of convolutional layer was 4, and output channel sizes were 32, 64, 128, 256, 256, 256 and 128, respectively. To reduce a temporal size of feature map without temporal pooling, we stacked frames in 4 layers. Also, we designed the axis-wise attention module (AWAM) inspired by parallel temporal-spectral attention [8] to improve the baseline model [9]. AWAM is a module that calculates the sigmoid-based score for each axis and adds to the input feature map, and it was adopted after the 2nd, 4th, and 6th convolutional blocks. The detail of AWAM architecture is shown in Figs. 4. The RNN network was same to the baseline CRNN introduced in [9].

[1]Strong labeled real recordings were newly released in DCASE 2022 challenge task 4. https://github.com/DCASE-REPO/DESED_task/tree/master/recipes/dcase2022_task4_baseline

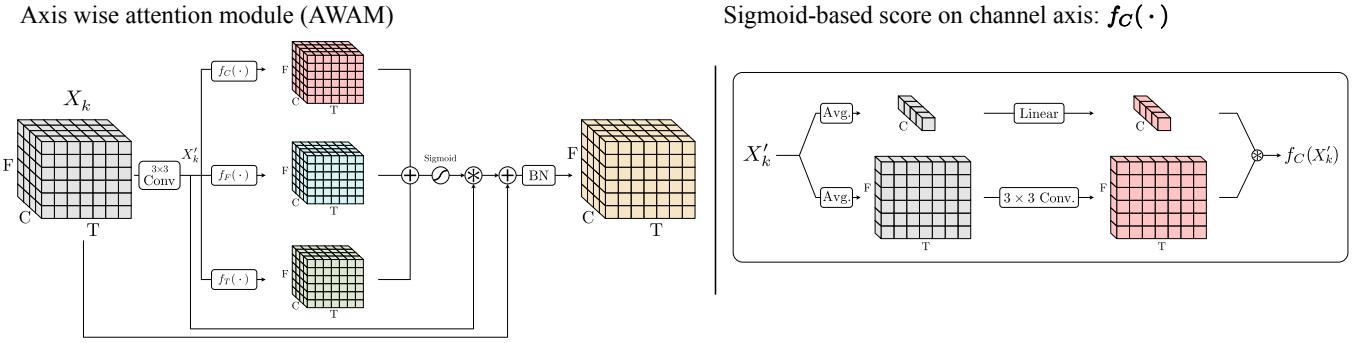

Figure 4: Axis wise attention module. X_k denotes the output of the k-th convolutional block in Fig. 5. The architecture of $f_F(\cdot)$ and $f_T(\cdot)$ are same to $f_C(\cdot)$, but are performed on frequency and time axis, respectively.

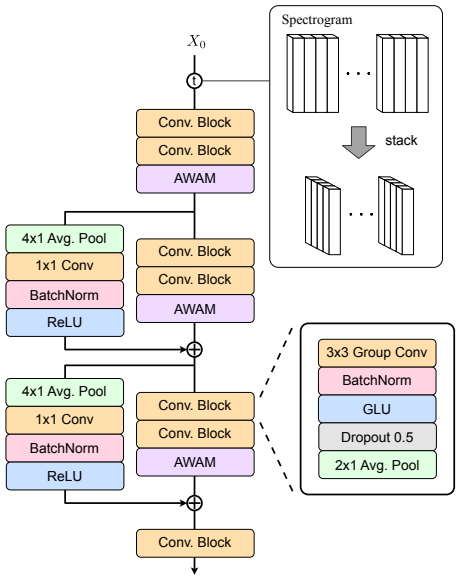

Figure 5: Block diagram of CNN architecture.

4.3. Experimental setup

We adopted the mean teacher [9, 10], one of the semi-supervised learning strategies to train the detection model using the unlabeled data. A minibatch consists of synthetic, strong, weak, and unlabeled-recorded clips with batch sizes of 8, 8, 4, and 40 each. All networks were optimized with the AdamW [11] optimizer and the cosine-annealing learning rate scheduler for 50 epochs after warming up the first 50 epochs from 0 to 0.001. Also, we set the weight decay and dropout to 0.001 and 0.5, respectively. We used event-based f1 score [12] and polyphonic sound detection score (PSDS) [13] for the evaluation metrics[2].

[2]The specific parameters settings for all metrics were same to the recent DCASE challenge.
https://dcase.community/challenge2022/task-sound-event-detection-in-domestic-environments

5. RESULTS AND DISCUSSION

5.1. Effect of confidence regularization

The experimental results according to the objective functions are compared in Table 1. If the other conditions are same except objective function, the systems built with CRE showed great performances under the event-f1 and PSDS1 metrics (CRE $>$ BCE $>$ AFL). Whereas, the systems built with AFL showed better performances under the PSDS2 metric (AFL $>$ BCE $>$ CRE), and the system with BCE showed medium performances for all metrics. Also, the proposed confidence regularization method was applicable with the mixup augmentation. The results demonstrate that if the detection performances of a mixup-applied system with BCE improved more than the system without the mixup, the mixup-applied system with CRE also improved. Although sounds and labels are mixed up, Eq. (2) keeps an output not too much fitting to the mixed label.

Confidence of detected sound event versus number of frames graphs are shown in Fig. 7. In the aspect of detection as the inactive frame, many frames with detection results close to 0 when BCE was used for the objective function. Most of the detection results of the CRE-based system were also close to 0 but more spread from 0 to 0.02 than the system with BCE. In other words, CRE made the model less overfitted to inactive frames, which shows the class imbalance problem was relieved. Whereas, the detection confidences of the AFL-based system were evenly distributed rather than biased towards zero.

In the aspect of detection as the active frame, all systems show similar results to each other but have a little difference. The peak of the CRE-based system's curve was left-biased due to the threshold; however, the peak of the AFL-based system's curve was right-biased since the focal weight was set to train well for the active frames. It demonstrates that the CRE-based network is trained well up to the regularization threshold and is prevented from overfitting when the confidences come over the threshold. On the other hand, according to the focal weights, the AFL-based network is trained well focused for the active frames but less focused for the inactive frames.

5.2. Discussion: system characteristics and PSDS

As shown in the experimental result, the system's scores are different according to the evaluation metrics. For instance, the system trained with CRE outperformed the system with AFL on PSDS1 but vice versa on PSDS2. Then which system should we choose or

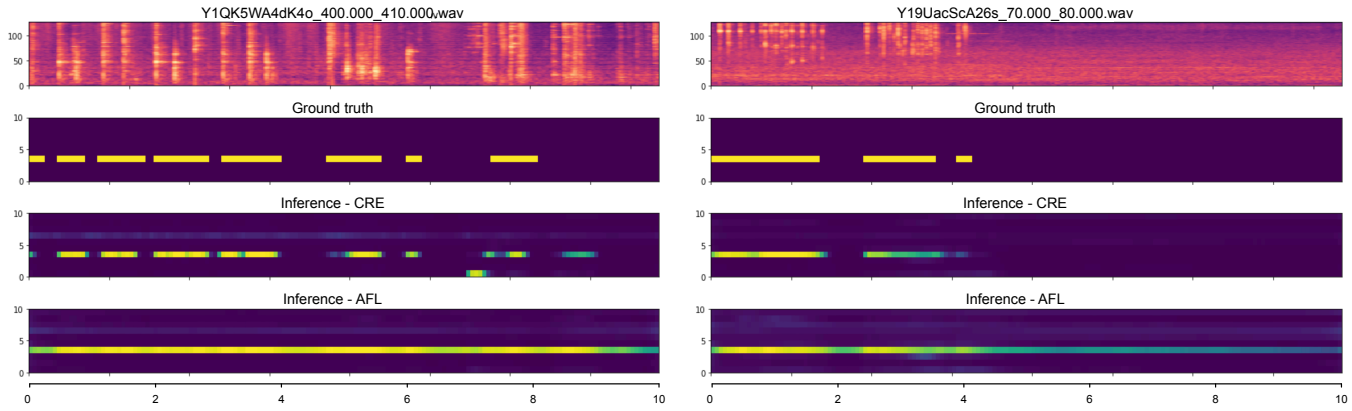

Figure 6: Two detection examples of CRE- and AFL-based systems. (From up to bottom: log-mel spectrogram, ground truth, CRE-based system, AFL-based system)

Table 1: Comparison of the detection performances among the different objective functions.

Network	Loss	Event-F1	PSDS1	PSDS2
CRNN w/o mixup	BCE	43.27	33.78	59.91
	AFL	40.63	31.99	**65.05**
	CRE	**45.50**	**33.98**	57.30
CRNN w/ mixup	BCE	44.93	34.79	58.60
	AFL	41.90	32.79	**60.92**
	CRE	**46.40**	**35.08**	57.82
CRNN+AWAM w/ mixup	BCE	49.17	37.51	66.34
	AFL	45.16	34.03	**66.88**
	CRE	**51.11**	**38.12**	64.33

which is better? As discussed in [1], PSDS1 is an effective metric for whether the system could detect the sound's timestamp correctly; whereas it has a severe problem related to the labeler's subjectivity. PSDS2 has strength in relieving the labeler's subjectivity; however it is hard to detect event localization precisely, and highly depends on the long-duration sounds.

For further description, we analyze the different detection patterns of the systems as shown in Fig. 6. Just as people label subjectively according to their background, systems have different characteristics under the objective functions. The CRE-based system tends to detect onsets and offsets precisely, whereas the AFL-based system tends to detect sound longer than the ground truth. In other words, the AFL-based system is proper to detect whether a sound appears in a clip rather than timestamps. The more general analysis is shown in the top graph of Fig. 7. The blue line shows that the confidences are more fitted to zero for inactive frames than green line. Instead, the green line is spread evenly without being biased to one side.

From the standpoint of user experience, the CRE-based system (system having high PSDS1 but low PSDS2) is required for users or environments that need to detect the target sound's onset and offset precisely. Whereas, the AFL-based system (system having high PSDS2 but low PSDS1) is more suitable in the environments for whether the appearance of target sounds is more important than detection resolution.

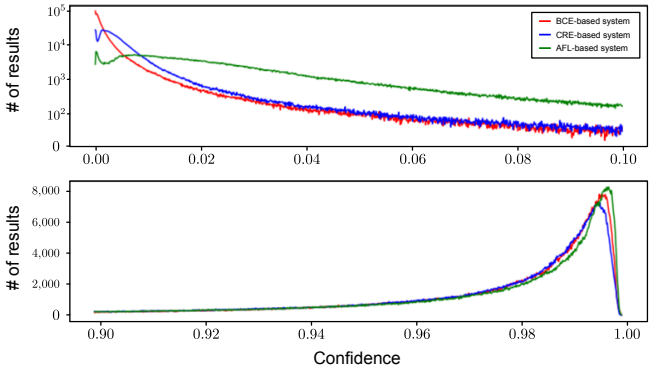

Figure 7: A graph of the number of frames over confidences of detected target sounds. Since there are a large number of inactive frames in the dataset, the graph of confidence near zero was log-scaled. (Top: confidence > 0.9, bottom: confidence < 0.1)

6. CONCLUSION

In this paper, we introduced the confidence regularized BCE that could avoid overfitting inactive frames. Compared to AFL, CRE performed better under the event-based f1 score and PSDS1. Furthermore, we suggested choosing the proper objective function according to the user's requirements. Of course, the system having good performance in both PSDS1 and PSDS2 is the best, but in a situation where you have to choose between the two, we can design a more suitable system by controlling the objective function.

7. ACKNOWLEDGEMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00456, Development of Ultra-high Speech Quality Technology for Remote Multi-speaker Conference System)

8. REFERENCES

[1] G. Ferroni, N. Turpault, J. Azcarreta, F. Tuveri, R. Serizel, Ç. Bilen, and S. Krstulović, "Improving sound event detection metrics: insights from dcase 2020," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 631–635.

[2] H. Sundar, M. Sun, and C. Wang, "Event specific attention for polyphonic sound event detection," in *Proc. Interspeech*, 2021, pp. 566–570.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[4] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, "Impact of sound duration and inactive frames on sound event detection performance," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 860–864.

[5] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[6] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 366–370.

[7] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[8] H. Wang, Y. Zou, D. Chong, and W. Wang, "Environmental sound classification with parallel temporal-spectral attention," in *Proc. Interspeech*, 2020, pp. 821–825.

[9] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, New York University, NY, USA, October 2019, pp. 253–257.

[10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[11] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

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