



Temporal Coding with Magnitude-Phase Regularization for Sound Event Detection

Sangwook Park¹, Sandeep Kothinti², Mounya Elhilali²

¹Department of Electronic Engineering, Gangneung-Wonju National University, Gangneung, South Korea

²Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA

spark2@gwnu.ac.kr, skothin1@jh.edu, mounya@jhu.edu

Abstract

Sound Event Detection (SED) is the challenge of identifying sound events into their temporal boundaries as well as sound category. With recent advances in deep learning, more effective SED techniques are investigated through the annual challenge of Detection and Classification of Acoustic Scenes and Events (DCASE). Most SED systems rely on data-driven learning where a deep neural network is trained to minimize the error between model prediction and the truth. While this framework is generally effective at identifying sound classes present in an audio recording, it results in unreliable estimates of temporal information for identifying sound boundaries. In order to heighten the temporal precision, this paper proposes a novel temporal coding of magnitude and phase for embedding vectors in an intermediate layer. This coding is reflected as a regularization term in the objective function for training the model. The regularization allows magnitude of embedding vectors to increase near event boundaries, which represent the onset and offset points. Simultaneously, each of the boundaries are distinguishable from others using phase difference between two neighboring vectors. This approach results in notable improvement in timing sensitivity compared to a baseline system tested on SED task in the context of DCASE2021 challenge.

Index Terms: sound event detection, coherence loss, time balanced focal loss, phase coding

1. Introduction

Sound Event Detection (SED) aims to identify sounds of interest in an audio recording by describing both *when* they happened as well as *what* types of sound they are [1]. SED techniques are critical for understanding the acoustic scene, and have been an integral part in a number of applications spanning video analytics, multimedia tagging, baby monitoring, or surveillance systems [2, 3, 4]. Given their broad impact, there has been increasing interest in developing effective SED systems, driven in large part by the annual challenge of Detection and Classification of Acoustic Scenes and Events (DCASE). Recent SED models considered as effective state of the art systems for this task use Convolutional Recurrent Neural Network (CRNN) as the main architecture leading to a family of powerful models [5, 6, 7].

The CRNN architecture combines a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN). Each part of CRNN could be conceptually interpreted as follows: The CNN projects time-frequency representations of audio inputs onto an embedding space to enhance discriminability among the sounds while the RNN learns temporal variation of

the soundscape. As such, the embedding vectors contain critical information to identify the sound events and their boundaries. This interpretation that the embeddings capture both aspects of a sound event (identity and boundaries) is not guaranteed since the model is generally trained to minimize the error between predictions of the CRNN and the truth. As such, the model leads to imprecise information regarding the temporal boundaries of sound events. In an earlier work, we have proposed a Temporal Contrastive Loss (TCL) [8] that extends the objective function of the SED system to enforce more precise temporal information in the CNN activation hence maintaining a degree of temporal coherence between CNN units that enforces more precise dynamics in the CNN activity.

With TCL, the CNN provides insights into the temporal boundaries of a sound event via the changes in magnitude of embedding vectors. According to the TCL formulation, the magnitude of embedding vector is forced to have big changes at the boundaries in euclidean sense, hence resulting in greater contrast *at* event boundaries. Conversely, *during* and *outside* the events, the changes in magnitude are forced to be minimal. This in turn results in heightened posteriors from the system output which ultimately leads to improved F-score metrics of the system. One of the drawbacks of this formulation is that there is no way to distinguish between onset and offset boundaries using the magnitude only. In addition, the objective function is unbounded leading to potential training problems. The current paper addresses these limitations by extending the concept of temporal contrastive loss to introduce more precise temporal markers. Specifically, the proposed method encodes the temporal information with respect to each of the boundaries of an event using both the magnitude and the phase of embedding vectors. This coding is formulated as a non-negative function to also address the potential unbounded constraints for network training. This temporal precision is integrated as a regularization term for training a SED network and tested in the context of DCASE2021 challenge. In this work, we investigate the effectiveness of the proposed method with two different objective functions: Binary Cross Entropy (BCE) and Time Balanced Focal Loss (TBFL), which was proposed to tackle imbalance issues on time and class [9]. With multiple metrics for evaluation, the proposed method results in notable improvement over the TCL as well as the challenge baseline.

The rest of this paper is structured as follows: related works are explored in Section 2. The details of the proposed method are described in Section 3. The next section gives all about experiment: database, evaluation metrics, experimental setting, and results. The conclusions and discussion are mentioned in the last section.

2. Related works

Current SED system employ a semi-supervised learning Mean-Teacher approach [10]. The systems employ a CRNN architecture trained on three types of datasets: strong labeled data that includes both sound category and its timestamps about onset and offset of sound events, weakly labeled data that includes sound categories only, and unlabeled data [11]. The objective function L_{base} is composed of a classification loss L^{cls} for labeled data and a regularization L^{reg} for all types of data defined as:

$$\begin{aligned} L_{base} &= L_{bcc}^{cls}(p, y) + \lambda L^{reg}(p, \hat{p}), \\ L_{bcc}^{cls}(p, y) &= BCE_{x \in S}(p, y^s) + BCE_{x \in W}(E_m[p], y^w), \\ L^{reg}(p, y) &= MSE_{x \in S, W, U}(p, \hat{p}), \end{aligned} \quad (1)$$

where S, W and U are set of strong labeled, weakly labeled, and unlabeled set, respectively. $BCE(p, y)$ and Mean Squared Error, $MSE(p, y)$ result a scalar value by averaging over the classes and frames. y^s and y^w is strong label and weakly label corresponding to an input x , respectively. p and \hat{p} are model prediction for the x by student and teacher network, respectively. E_m is an expectation operator over the frame.

Recent work proposed to incorporate the biological concept of temporal coherence in SED systems, which posits that embeddings driven by the same event should be coherently constrained, hence maximizing discriminability across events [8]. This formulation introduced a regularization term (TCL - temporal Contrastive Loss) formulated as:

$$\begin{aligned} L^{TCL}(z, y) &= -\alpha_1 \sum_{x \in S} \sum_{m=2}^M \mathbf{1}_{|y_m^{ts} - y_{m-1}^{ts}| > 0} |z_m - z_{m-1}|_2^2 \\ &+ \alpha_2 \sum_{x \in S} \sum_{m=2}^M \mathbf{1}_{|y_m^{ts} - y_{m-1}^{ts}| = 0} |z_m - z_{m-1}|_2^2 \end{aligned} \quad (2)$$

where y^{ts} is an integration of strong label over the classes as $y^{ts} = \sum_c y^s$ with class index c . z is an embedding vector produced by CNN part of the CRNN architecture for input x . $\mathbf{1}_A$ is a binary indicator, 1 if satisfaction of A , otherwise 0. $|\cdot|_2$ is an L_2 norm, and m is frame index. During training, this TCL is added to the baseline loss as another regularization term as: $L_{TCL} = L_{base} + L^{TCL}$. This formulation causes a big change between neighboring embedding vectors at boundaries while a small change could be observed in a non-edge region. Since the changes are reflected in the magnitude in the loss L_{TCL} , the following RNN is unable to distinguish between onsets and offsets. Besides, this formulation has a strict assumption that L^{TCL} should be a non-negative value during the training. Due to this assumption, the parameters α_1 and α_2 greatly affect a success of model training.

To address a separate issue of imbalance between the number of samples per class and sound event duration, a different training formulation was proposed to constrain the loss function using the concept of time-balanced focal loss - TBFL [9]. The TBFL function is defined as:

$$\begin{aligned} TBFL(p, y) &= - \sum_c w_c \{ y_c (1 - p_c)^\gamma \log(p_c) \\ &+ (1 - y_c) p_c^\beta \log(1 - p_c) \}, \quad (3) \\ w_c &\propto \frac{1 - \beta_c}{1 - \beta_c^{[k \times r_c]}}, \quad \sum_c w_c = C, \end{aligned}$$

where C is the number of target classes, c is a class index. m_c is the number of frames, $r_c = \frac{m_c}{\sum_c m_c}$ is a ratio of the number of frames in class c to total number of frames, and k is a hyper-parameter to convert from the ratio to the number of samples. Note that $TBFL(p, y)$ yields an average value over the classes and frames like BCE. The objective function for training with TBFL is denoted as:

$$\begin{aligned} L_{tbfl} &= L_{tbfl}^{cls}(p, y) + \lambda L^{reg}(p, \hat{p}) \\ L_{tbfl}^{cls}(p, y) &= TBFL_{x \in S}(p, y^s) + TBFL_{x \in W}(E_m[p], y^w) \end{aligned} \quad (4)$$

3. Proposed Method

Building on the concepts of both a binary cross entropy (BCE) or time-balanced focal loss (TBFL), the current work addresses the limitations of temporal precision in the network mappings. Here, we propose a Magnitude-Phase Regularization (MPR) for temporal coding of embedding vectors to resolve the issues of TCL. The MPR consists of 4-components for onset edge L_{on} , offset edge L_{off} , steady-state L_{steady} , and on-event L_{event} . With strong label y^{ts} like the TCL, the MPR is formulated to:

$$\begin{aligned} L^{MPR}(z, y) &= \alpha_1(L_{on} + L_{off}) + \alpha_2 L_{steady} + \alpha_3 L_{event} \\ L_{on} &= \sum_{x \in S} \sum_{m=2}^M \mathbf{1}_{(y_m^{ts} - y_{m-1}^{ts}) > 0} \frac{\sin^2(\theta_m)}{|z_m|_2^2} \\ L_{off} &= \sum_{x \in S} \sum_{m=2}^M \mathbf{1}_{(y_m^{ts} - y_{m-1}^{ts}) < 0} \frac{\cos^2(\theta_m)}{|z_m|_2^2} \\ L_{steady} &= \sum_{x \in S} \sum_{m=2}^M \mathbf{1}_{(y_m^{ts} - y_{m-1}^{ts}) = 0} |z_m|_2^2 \\ L_{event} &= \sum_{x \in S} \sum_{m=2}^M \mathbf{1}_{y_m^{ts} = 1} (\sin^2(\theta_m)) \end{aligned} \quad (5)$$

where θ_m is the angle of between neighboring embedding vectors, $\cos(\theta_m) = \frac{z_{m-1}^T z_m}{|z_{m-1}|_2 |z_m|_2}$, and $\sin^2(\theta_m) = 1 - \cos^2(\theta_m)$. Similar to TCL, this formulation evokes a large magnitude of embedding vectors in both boundaries; however, they are distinguishable in the angle between neighboring vectors, which would be in parallel/perpendicular to each other at onset/offset boundary. Additionally, neighboring vectors are sustained in parallel during the sound event while it has no constraint during background. As such, the formulation is aimed to facilitate recognition of sound events against background intervals. In the same manner with TCL, the training loss is defined by adding MPR to the objective function of Mean-Teacher approach.

4. Experiment

4.1. Database

The effectiveness of the proposed method is investigated with the DESED database that contains 10-sound events: Alarm/bell/ringing (A), Blender (B), Cat (C), Dishes (Di), Dog (Do), Electric shaver/toothbrush (E), Frying (F), Running Water (R), Speech (S), and Vacuum Cleaner (V) [12]. With Mean-Teacher approach, weakly labeled and unlabeled sets are used in training of a CRNN model in combination with strong labeled set that consists of 10,000 synthetic audios produced by *Scaper*. Then real validation set is used for evaluation.

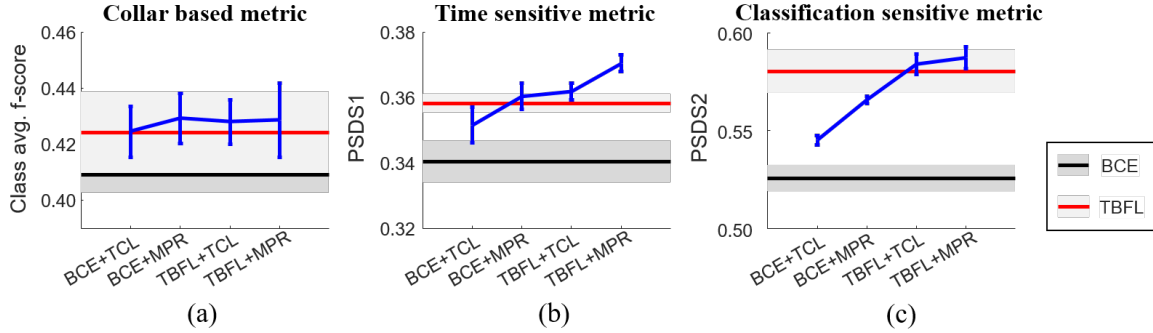


Figure 1: Performance Evaluation in (a) collar based f -score (b) time sensitive PSDS (c) classification sensitive PSDS. In each of panels, the performance of DCASE2021 baseline is marked as black line and dark-gray region while red and light-gray is for the performance of TBFL.

4.2. Experimental Setting

4.2.1. CRNN architecture

In all experiments, the architecture of CRNN is the same with the baseline of the DCASE2021 challenge task 4. In a nutshell, the model consists of 7 convolutional blocks and 2 bi-directional Gated Recurrent Units (GRU). Each of convolutional blocks is composed of a convolutional layer with 3×3 filter, Gated Linear Units (GLU) for non-linearity, and averaging pooling. In the first two blocks, the pooling is performed with 2×2 filter on time-frequency representation. On the other hand, 2×1 filter along to frequency axis is used for the pooling in the rest of the blocks. More details of the architecture are described in [11].

4.2.2. Post-processing

A post processing composed of thresholding and smoothing is performed on model prediction to detect sound intervals. To find the best threshold and smoothing length, each of the parameters is scanned on the range of 0.01 to 0.99 with 0.02 step for the threshold; and 64 ms to 576 ms with 64 ms steps for the length.

4.2.3. Evaluation metrics

Assessment is performed with collar based f -score and Polyphonic Sound Event Detection Score (PSDS) [13]. For f -score, a detected interval will be decided to true positive if it has matched to truth in time boundaries within 200 ms margin as well as sound class. As a moderate metric compared to the f -score, PSDS decides true positive if a ratio of intersection between detected interval and truth is above preset parameters determined depending on criteria. In here, two different criteria are considered. PSDS1 focuses on time accuracy of the intervals, on the other hand, PSDS2 focuses on classification among targets rather than time accuracy.

4.2.4. Performance comparison

Both the baseline of DCASE2021 challenge task 4 [11] and the TCL [8], which are respectively described in (1) and (2), are considered as counterparts to the proposed method. In addition, we explore the effectiveness of the additional regularization terms with a balanced classification loss L_{tbfl} in (4) instead of L_{base} . As a result, we separately evaluate 6-CRNNs trained with each of objective functions and parameters, described as follows:

- BCE: the challenge baseline as L_{base} in (1)

- BCE+TCL: previous work with BCE as $L_{base} + L^{TCL}$ where $\alpha_1 = 0.1$ and $\alpha_2 = 0.03$
- BCE+MPR: proposed method with BCE as $L_{base} + L^{MPR}$ where $\alpha_1 = 0.003$, $\alpha_2 = 0.001$, and $\alpha_3 = 0.001$
- TBFL: balanced loss as L_{tbfl} in (4) where $\gamma = 2.0$
- TBFL+TCL: previous work with TBFL as $L_{tbfl} + L^{TCL}$ where $\gamma = 2.0$, $\alpha_1 = 0.001$, and $\alpha_2 = 0.0001$
- TBFL+MPR: proposed method with TBFL as $L_{tbfl} + L^{MPR}$ where $\gamma = 1.0$, $\alpha_1 = 0.001$, $\alpha_2 = 0.0001$, and $\alpha_3 = 0.0003$

Note that we use the same parameters for k , β_c , and r_c with the ones described in the original paper [9]. And, other parameters such as α_i of TCL or MPR and γ of TBFL are heuristically optimized depending on classification loss. For each of the methods, the training and test are performed at least 3-times, and the results of the student network are summarized as mean and standard deviation over all iterations.

4.3. Experimental Results

Figure 1 shows the evaluation results in multiple metrics: f -score, PSDS1, and PSDS2. Each of panels shows the averages of BCE (DCASE2021 baseline) and TBFL as solid lines, black is for the baseline while red is for the other. The standard deviations are represented as dark-gray and light-gray regions for the baseline and the TBFL, respectively. In evaluation under the time sensitive criteria (Figure 1(b)), when BCE is applied to the classification loss in (1), the proposed method denoted as BCE+MPR shows a notable improvement compared to the TCL as well as the baseline. Additionally, the TBFL+MPR significantly improves PSDS1 compared to TBFL while the TBFL+TCL is comparable to TBFL. In Figure 1(c), the proposed method shows great improvement compared to the TCL as well as the baseline if BCE is used for classification loss (1). With the temporal coding which allows embedding vector to be distinguished between target event from non-target sound event, it could reduce a confusion among the target sound events. If BCE is replaced with TBFL, the proposed method TBFL+MPR shows a little improvement averagely comparable to the TBFL.

On the contrary, results of f -score show a different pattern to the results of PSDS (Figure 1(a)). One of potential reasons is a different criteria to decide on *true positive*. When a model detects multiple sound intervals for a long target sound, all detected intervals could be *true positives* in PSDS if a ratio

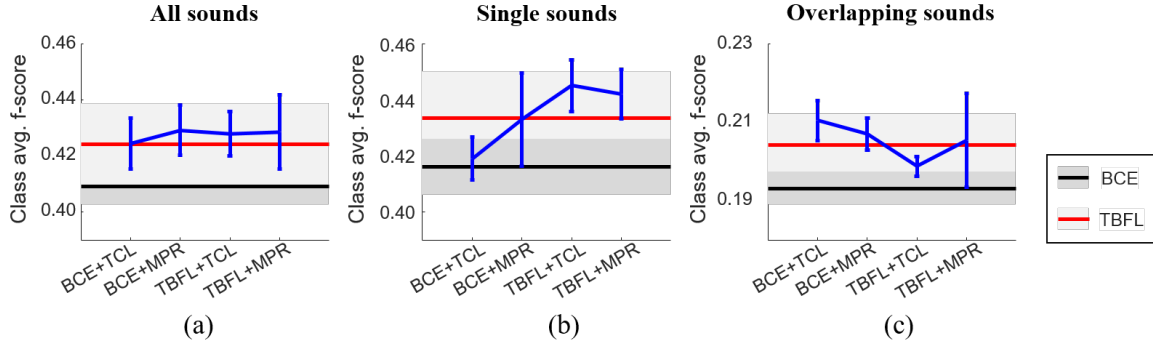


Figure 2: Performance Evaluation in collar based f-score for (a) all target sound events (b) isolated sound events (non-overlapping) (c) overlaid sound events with others. In each of panels, the performance of DCASE2021 baseline is marked as black line and dark-gray region while red and light-gray is for the performance of TBFL.

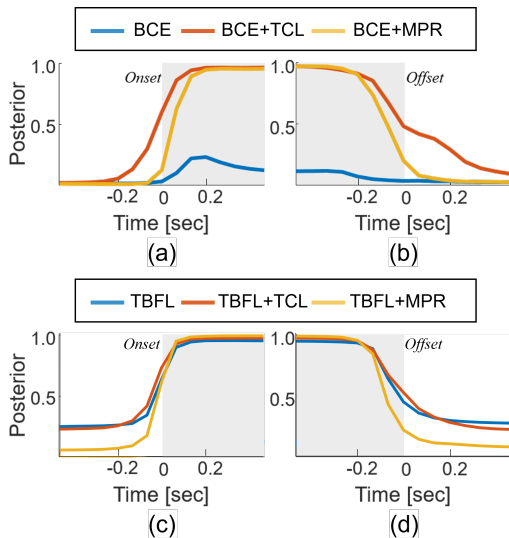


Figure 3: Posterior probabilities near sound boundaries for "YvATUVASx8ug_0.000_10.000.wav" from Real:Validation set in DESED database: (a) Onset boundary based on BCE, (b) Offset boundary based on BCE, (c) Onset boundary based on TBFL, (d) Offset boundary based on TBFL. Note that gray region represents an existence of target sound.

of intersection between those intervals and truth is above the threshold. But, all of detection would be determined as *false positives* in f-score. Similarly, when a model detects only a long sound interval for repeating sounds that are annotated in multiple truths, the long detection could be a *true positive* in PSDS while it makes one *false alarm* and multiple *false negatives* in f-score. But, the proposed method have no way to compensate these cases. To alleviate those impact, we separately evaluate the performance for single sounds or overlapping sounds since the overlapping sounds easily cause those cases (Figure 2). Note that, in the evaluation for single sounds, both truth and detection within the time intervals of overlapping sounds are discarded in f-score calculation. Similarly, both truth and detection within the time intervals of single sounds are discarded in the evaluation for overlapping sounds. In evaluation for single sounds (Figure 2(b)), the result of f-score shows a similar pattern to PSDSs. The TBFL-TCL and TBFL-MPR are comparable to each other and both are outstanding to the baseline. But, TBFL-TCL shows the worst result in evaluation for overlapping sounds as in Figure 2(c).

Figure 3 gives one example of network prediction near each

of sound boundaries. With BCE based training loss as in 1, the proposed method, BCE+MPR, shows the steepest transition in near the boundaries although it has affordable delay or ahead. Unlike the baseline, BCE+TCL shows reasonable posterior during the sound turned on, but it is vague to find sound boundaries due to an before and residual posterior. When TBFL is used for classification loss as in 4, TBFL+MPR shows good prediction to find sound boundaries compared to others.

5. Discussion

It is critical to find optimal parameters of α_i for a success of training with TCL and MPR. Since the regularization by TCL or MPR provide information with respect to sound boundaries, it is important to keep the balance among the terms of the overall objective function. In the current study, different parameters of α_i are used depending on the classification loss, either BCE or TBFL, because each loss has a different range to each other. This interdependence between the main loss objective and the regularization scaling needs to be further investigated to better understand how the network mappings evolve during training, instead of relying on heuristic testing to find optimal parameters.

6. Conclusions

Most of deep networks are trained to produce minimum error between the network prediction and truth. This is a typical framework, however, it has a potential issue that essential information toward the goal could be lost on the way of propagation in the model. With a CRNN model for SED task, this paper proposes a novel regularization that maintains temporal boundaries in embedding space produced by CNN. The regularization allows magnitude of embedding vector to get a big at boundaries, onset and offset. Simultaneously, each of boundaries are distinguishable from the other by the phase difference between two neighboring vectors. With the magnitude and phase of embedding vectors, the temporal information is forwarded to the next layer. The proposed method is applied to SED in the context of DCASE2021 challenge, and reports a notable improvement over the baseline, particularly in time sensitive metric.

7. Acknowledgements

This work was supported by NIH U01AG058532, ONR N00014-19-1-2014, and N00014-19-1-2689.

8. References

- [1] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound Event Detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 9 2021. [Online]. Available: <https://ieeexplore.ieee.org/document/9524590/>
- [2] K. Ahmad and N. Conci, "How deep features have improved event recognition in multimedia: A survey," *ACM Trans. on Multimedia Compu., Comm. and App.*, vol. 15, no. 2, 2019.
- [3] Y. Lavner, R. Cohen, D. Ruinskiy, and H. Ijzerman, "Baby cry detection in domestic environment using deep learning," in *IEEE International Conference on the Science of Electrical Engineering*. IEEE, 2017.
- [4] S. Park, W. Choi, D. K. Han, and H. Ko, "Acoustic event filterbank for enabling robust event recognition by cleaning robot," *IEEE Transactions on Consumer Electronics*, vol. 61, no. 2, pp. 189–196, 2015.
- [5] S. Park, A. Bellur, D. K. Han, and M. Elhilali, "Self-Training for Sound Event Detection in Audio Mixtures," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6 2021, pp. 341–345. [Online]. Available: <https://ieeexplore.ieee.org/document/9414450/>
- [6] D. De Benito-Gorron, D. Ramos, and D. T. Toledano, "A Multi-Resolution CRNN-Based Approach for Semi-Supervised Sound Event Detection in DCASE 2020 Challenge," *IEEE Access*, vol. 9, pp. 89 029–89 042, 2021. [Online]. Available: <https://ieeexplore.ieee.org/document/9453768/>
- [7] H. Dinkel, M. Wu, and K. Yu, "Towards Duration Robust Weakly Supervised Sound Event Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 887–900, 2021. [Online]. Available: <https://ieeexplore.ieee.org/document/9335265/>
- [8] S. Kothinti and M. Elhilali, "Temporal Contrastive-Loss for Audio Event Detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5 2022, pp. 326–330.
- [9] S. Park and M. Elhilali, "Time-Balanced Focal Loss for Audio Event Detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5 2022, pp. 311–315.
- [10] 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*, vol. 2017-Decem, no. Nips, 2017, pp. 1196–1205.
- [11] N. Turpault and R. Serizel, "Training Sound Event Detection On A Heterogeneous Dataset," in *DCASE workshop*, 2020.
- [12] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound Event Detection in Synthetic Domestic Environments," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5 2020, pp. 86–90. [Online]. Available: <https://ieeexplore.ieee.org/document/9054478/>
- [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)*, 2015, pp. 1–7. [Online]. Available: <http://ieeexplore.ieee.org/document/7280624/>