Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

# Training Onset-and-Offset-Aware Sound Event Detection on a Heterogeneous Dataset via Probabilistic Sequential Modeling

*Tomoya Yoshinaga*[1,2], *Yoshiaki Bando*[1], *Keitaro Tanaka*[2], *Keisuke Imoto*[3,4], *Masaki Onishi*[1], *Shigeo Morishima*[2]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2]Waseda University, Japan [3]Doshisha University, Japan [4]Kyoto University, Japan

tomozu@akane.waseda.jp, y.bando@aist.go.jp, keitaro@aoni.waseda.jp, keisuke.imoto@ieee.org, onishi-masaki@aist.go.jp, shigeo@waseda.jp

## Abstract

This paper presents a neural method to train onset-and-offset-aware sound event detection (SED) using heterogeneously labeled data. SED models are typically trained to predict frame-wise event activities, which have temporal fluctuations, resulting in unstable event boundaries. An end-to-end (E2E) method based on a hidden semi-Markov model (HSMM) has been proposed to improve performance by converting frame-wise predictions into event boundaries. This method, however, relies on temporal (strong) labels, which are costly to annotate. To overcome this limitation, we propose an E2E method to train an HSMM-based model from clip-level labels and unlabeled data. While the strong supervision was formulated to maximize event-wise posterior probabilities, we derive probabilistic objectives for such incompletely labeled data. Experimental results on the DESED dataset show that our method outperforms standard frame-wise methods. Our codes will be available at `https://ybando.jp/projects/sed-hsmm`.

**Index Terms**: sound event detection, hidden semi-Markov model, weakly-supervised and semi-supervised training

## 1. Introduction

Sound event detection (SED) aims to jointly detect classes of sound events in an audio recording and locate their onset and offset boundaries [1, 2]. SED forms the basis of many applications, including surveillance systems [3, 4] and multimedia retrieval [5, 6]. A typical method trains a deep neural network (DNN) by utilizing pairs of mixture signals and their temporal (strong) annotations of sound events. The DNN is trained in an end-to-end (E2E) manner to predict frame-wise sound event activities as posterior probabilities [2, 7–9].

One major focus in building an SED system is eliminating fine insertion and deletion errors caused by temporal fluctuations in frame-wise predictions [10, 11]. A popular solution is to apply post-processing, such as median filtering [9, 10, 12] and sound event bounding boxes (SEBBs) [13]. Since they are non-differentiable and prevent E2E training, efforts have also been made to design network architectures that directly output event-wise predictions [14–16]. Building on the recent success of frame-wise architectures, a differentiable frame-to-event mapping layer [17] has been proposed. This layer is based on a hidden semi-Markov model (HSMM) [18] that represents sound event sequences. The HSMM transforms frame-wise features (emission probabilities) predicted by a DNN into posterior probabilities of onset-and-offset-aware sound events (Fig. 1-(a)). This method, called DNN-HSMM, was reported to outperform an existing model with post-processing [17].

Another key technique for improving SED performance is to utilize heterogeneously labeled data. Since annotating strong labels is time-consuming, most frame-wise training also utilizes incompletely labeled data. For example, to utilize data with clip-level (weakly labeled) annotations, multiple instance learning [19, 20] has been adopted. In this approach, a DNN is trained to predict the posterior probabilities of clip-level labels by aggregating frame-wise predictions with pooling operations [21, 22]. Unlabeled data is also widely utilized in semi-supervised learning [23]. This method utilizes a teacher model constructed via the exponential moving average (EMA) of a student model. The teacher model generates pseudo labels for unlabeled data, and the student model is trained to produce frame-wise predictions consistent with those of the teacher model. While frame-wise training on a heterogeneous dataset is well established, the event-wise training of DNN-HSMM [17] has only been formulated for strong labels, limiting its scalability.

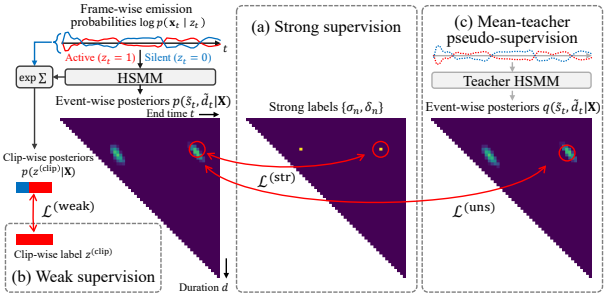



Figure 1: *The overview of DNN-HSMM and (a) its strong supervision, as well as our (b) weak- and (c) pseudo-supervision.*

In this paper, we derive probabilistic objective functions for training a DNN-HSMM with weak labels and unlabeled data in addition to strong labels. To utilize weak labels, we formulate weak supervision by deriving the probability mass of clip-level event existence from the frame-wise emission probabilities (Fig. 1-(b)). We show it can be efficiently calculated via a recursive formula that represents the probabilities for event sequences only with silence. To utilize unlabeled data, we redesign the EMA-based pseudo-supervision for the event-wise posterior probabilities (Fig. 1-(c)). We train a student model by encouraging consistency between the posterior probabilities of the student and teacher models. To further improve performance, we also extend Mixup [24], a popular augmentation in frame-wise supervision, for applying it to event-wise one.

The main contribution of this study is to extend the E2E event-wise training of the DNN-HSMM to work under practical conditions. Utilizing heterogeneous data is essential to build a high-performing system. We newly derive two objective functions for weak- and semi-supervision in a probabilistically


10.21437/Interspeech.2025-1642

grounded manner, making it easier to provide theoretical guarantees. We show that our weak- and semi-supervision improve the performance from strong-only supervision for the DESED dataset [25]. We also demonstrate that our event-wise Mixup is effective, and the resulting method outperforms existing frame-wise models with widely used post-processing methods.

## 2. Background

We review standard SED approaches with practical techniques and introduce the DNN-HSMM [17] for event-wise prediction.

### 2.1. Frame-wise sound event detection

Most SED systems train DNNs, such as convolutional recurrent neural networks (CRNNs) [26], to predict frame-wise sound event activities by minimizing a binary cross-entropy (BCE) loss. This training corresponds to the maximization of frame-wise posterior probabilities $q_\phi(z_{c,t} \mid \mathbf{X}) \in [0,1]$ as follows:

$$\mathcal{L}_\phi^{(\mathrm{fr})} = -\frac{1}{C}\sum_c \frac{1}{T}\sum_t \log q_\phi(z_{c,t} \mid \mathbf{X}), \tag{1}$$

where $\phi$ is a set of network parameters, $\mathbf{X} \triangleq [\mathbf{x}_1, \dots, \mathbf{x}_T] \in \mathbb{R}^{F\times T}$ is an $F$-dimensional input audio feature having $T$ frames, $z_{c,t} \in \{0,1\}$ is an event label for class $c = 1, \dots, C$ at frame $t$. Since the frame-wise predictions are often fluctuating, post-processing is important to obtain precise sound event boundaries [9, 11, 27]. While median filtering is popular [9, 28–30], SEBBs [13] improved performance by detecting the boundaries from the gradient of the frame-wise predictions.

Various techniques have improved performance by addressing data scarcity. One common technique is to use incompletely labeled data. Weak supervision is typically performed using a clip-level BCE loss, where the prediction is obtained by a pooling operation, such as attention pooling [21, 22]. Meanwhile, EMA-based semi-supervision for unlabeled data is often applied using a frame-wise mean squared error loss [23]. Another technique is data augmentation, as represented by Mixup [24]. Mixup generates new training samples by linearly combining two audio recordings and their labels. Finally, using models pre-trained on large-scale datasets, such as AudioSet [31], is crucial [1, 9]. The baseline system of DCASE 2023 Task 4 significantly improved a CRNN by using features extracted by a Transformer-based pre-trained model called BEATs [32].

### 2.2. Onset-and-offset-aware SED based on DNN-HSMM

The DNN-HSMM [17] was proposed to combine with a frame-wise neural model to predict onset-and-offset-aware sound events. This method represents sound event sequences by class-wise HSMMs. Hereinafter, we omit the class index $c$ without losing the generality. Let $n = 1, \dots, N$ be an index of sound events and $\mathcal{S} \triangleq \{0,1\}$ be a set of possible states (0: silent and 1: active). The HSMM is a probabilistic model that generates observed acoustic features $\mathbf{X} \in \mathbb{R}^{F\times T}$ from a sequence of latent events having states $s_n \in \mathcal{S}$ and durations $d_n \in \mathcal{D} \triangleq \{1, \dots, T\}$. The $n$-th event $s_n, d_n$ is assumed to be recursively generated from its previous event as follows:

$$s_n, d_n \mid s_{n-1} \sim p_\theta(s_n \mid s_{n-1}) p_\theta(d_n \mid s_n), \tag{2}$$

$$s_1, d_1 \sim p_\theta(s_1) p_\theta(d_1 \mid s_1), \tag{3}$$

where $\theta$ is a set of the model parameters. The initial probabilities $p_\theta(s_1)$ and transition probabilities $p_\theta(s_n|s_{n-1})$ are represented by learnable parameters $\rho_{s_1} \in [0,1]$ ($\sum_s \rho_s = 1$) and hyperparameters $\psi_{s_{n-1}s_n} \in [0,1]$ ($\sum_{s'} \psi_{ss'} = 1$), respectively. On the other hand, the duration probability of active states $p_\theta(d_n|s_n = 1)$ is represented parametrically as follows:

$$p_\theta(d_n|s_n = 1) = \begin{cases} (1-w)\cdot \mathcal{G}(d_n \mid a, b) & \text{if } d_n < T, \\ w & \text{if } d_n = T, \end{cases} \tag{4}$$

where $\mathcal{G}(x \mid a, b) \propto x^{a-1}\exp(-x/b)$ is a truncated gamma distribution with learnable parameters $a \in \mathbb{R}_+$ and $b \in \mathbb{R}_+$, and $w \in [0,1]$ is a weight parameter to represent an event spanning the clip length. The duration probability of silence $p_\theta(d_n \mid s_n = 0)$ is set to be uniform. Conditioned by a sequence of $N$ events $\{s_n, d_n\}_{n=1}^N$ generated by Eq. (2) and (3), the observation $\mathbf{X}$ is generated in a frame-wise manner:

$$\mathbf{X} \mid \{s_n\}_{n=1}^N, \{d_n\}_{n=1}^N \sim \prod_{t=1}^T p(\mathbf{x}_t \mid z_t), \tag{5}$$

where $z_t \in \mathcal{S}$ denotes the state $s_n$ of event $n$ covering frame $t$.

The HSMM is used to predict sound event activities as posteriors $p_{\theta,\phi}(z_t \mid \mathbf{X})$ from frame-wise features predicted by a DNN. Specifically, the DNN $g_\phi$ transforms an acoustic feature $\mathbf{X}$ into frame-wise log emission probabilities $\log p(\mathbf{x}_t \mid z_t = z) \triangleq y_{z,t} \in \mathbb{R}$ in Eq. (5) as follows:

$$\{\mathbf{y}_1, \dots, \mathbf{y}_T\} \leftarrow g_\phi(\mathbf{X}), \tag{6}$$

where $\mathbf{y}_t \triangleq \{y_{z,t}\}_{z\in\mathcal{S}}$ is an emission probability vector. The HSMM, then, calculates posteriors $p_{\theta,\phi}(\tilde{s}_t, \tilde{d}_t \mid \mathbf{X})$, where $\tilde{s}_t \in \mathcal{S}$ and $\tilde{d}_t \in \mathcal{D}$ are the state and duration of an event terminating at frame $t$, respectively. These posteriors are calculated by a forward-backward algorithm [18] as follows:

$$p_{\theta,\phi}(\tilde{s}_t, \tilde{d}_t \mid \mathbf{X}) \propto \alpha^{[t]}_{\theta,\phi}(\tilde{s}_t, \tilde{d}_t)\beta^{[t]}_{\theta,\phi}(\tilde{s}_t), \tag{7}$$

where $\alpha^{[t]}_{\theta,\phi} \in \mathbb{R}_+$ and $\beta^{[t]}_{\theta,\phi} \in \mathbb{R}_+$ are forward and backward variables, respectively. They are calculated recursively as:

$$\alpha^{[t]}_{\theta,\phi}(s,d) \triangleq p_\theta(d \mid s)\cdot \exp(\bar{y}_{s,t-d+1:t}) \sum_{s'\in\mathcal{S}}\sum_{d'=1}^{t-d} \psi_{s's}\cdot \alpha^{[t-d]}_{\theta,\phi}(s', d'), \tag{8}$$

$$\beta^{[t]}_{\theta,\phi}(s) \triangleq \sum_{s'\in\mathcal{S}}\sum_{d'=1}^{T-t} p_\theta(d' \mid s')\cdot \exp(\bar{y}_{s',t+1:t+d'}) \psi_{ss'}\cdot \beta^{[t+d']}_{\theta,\phi}(s'), \tag{9}$$

where $\bar{y}_{z,t_1:t_2}$ is $\sum_{t=t_1}^{t_2} y_{z,t}$. The initial values $\alpha^{[t]}_{\theta,\phi}(s, d{=}t)$ and $\beta^{[T]}_{\theta,\phi}(s)$ are $p_\theta(t|s)\exp(\bar{y}_{s,1:t})\rho_s$ and 1, respectively. Finally, the frame-wise posteriors $p_{\theta,\phi}(z_t \mid \mathbf{X})$ are obtained as:

$$p_{\theta,\phi}(z_t \mid \mathbf{X}) = \sum_{t'=t}^{T}\sum_{d=t'-t+1}^{t'} p_{\theta,\phi}(\tilde{s}_{t'} = z_t, \tilde{d}_{t'} = d \mid \mathbf{X}). \tag{10}$$

The model parameters $\{\theta, \phi\}$ can be trained in an event-wise manner by utilizing strongly labeled data. Let strong labels be $N$ pairs of event states $\{\sigma_n\}_{n=1}^N \in \mathcal{S}^N$ and durations $\{\delta_n\}_{n=1}^N \in \mathcal{D}^N$. The parameters are trained to maximize the event-wise posteriors $p_{\theta,\phi}(\tilde{s}_t, \tilde{d}_t \mid \mathbf{X})$. Specifically, the loss function is formulated as a negative log-posterior probability:

$$\mathcal{L}^{(\mathrm{str})}_{\theta,\phi} = -\frac{1}{N}\sum_{n=1}^N \log p_{\theta,\phi}(\tilde{s}_{t_n} = \sigma_n, \tilde{d}_{t_n} = \delta_n \mid \mathbf{X}), \tag{11}$$

where $t_n$ is the frame where the $n$-th event ends. Since this loss function is defined with differentiable operators, the parameters are optimized in an E2E manner. This method was reported

to suppress temporal fluctuations and significantly improve the performance of boundary estimation [17].

# 3. Weakly- and Semi-Supervised Training for DNN-HSMM-Based SED

We extend the strong supervision of DNN-HSMM to weak- and semi-supervision for handling heterogeneously labeled datasets. We also propose an event-wise extension of Mixup.

## 3.1. Weakly-supervised training

We first describe our weakly-supervised training to utilize clip-wise labels $z^{(\mathrm{clip})} \in \{0, 1\}$. Specifically, the model is trained to predict the label as the posterior probability $p_{\theta,\phi}(z^{(\mathrm{clip})} \mid \mathbf{X})$ by minimizing the clip-level BCE loss. This loss function is defined as a negative log-posterior probability as follows:

$$\mathcal{L}^{(\mathrm{weak})}_{\theta,\phi} = -\log p_{\theta,\phi}(z^{(\mathrm{clip})} \mid \mathbf{X}). \tag{12}$$

It is, however, difficult to directly calculate the posterior probability for the active label $p_{\theta,\phi}(z^{(\mathrm{clip})} = 1 \mid \mathbf{X})$ because the number of possible sequences where at least one event occurs is numerous. We, instead, calculate it as the complement of the probability that a non-active sequence occurs through the clip:

$$\begin{aligned} &p_{\theta,\phi}(z^{(\mathrm{clip})} = 1 \mid \mathbf{X}) \\ &\quad = 1 - \sum_{d_1,\ldots,d_N \in \mathscr{D}} p_{\theta,\phi}(\mathbf{s} = \mathbf{0}, d_1, \ldots, d_N \mid \mathbf{X}), \end{aligned} \tag{13}$$

where $\mathscr{D}$ denotes a set of possible sequences of durations $d_1, \ldots, d_N$ in the clip, and $\mathbf{s} = \mathbf{0}$ represents $s_n = 0$ for all $n = 1, \ldots, N$. This posterior probability can be factorized into the product of the observation model of Eq. (5) and prior probabilities of a silent clip by using Bayes' theorem as follows:

$$\begin{aligned} &\sum_{d_1,\ldots,d_N \in \mathscr{D}} p_{\theta,\phi}(\mathbf{s} = \mathbf{0}, d_1, \ldots, d_N \mid \mathbf{X}) \\ &\quad \propto \sum_{d_1,\ldots,d_N \in \mathscr{D}} \{p_\phi(\mathbf{X} \mid \mathbf{s}{=}\mathbf{0}, d_1, \ldots, d_N) \cdot p_\theta(\mathbf{s}{=}\mathbf{0}, d_1, \ldots, d_N)\}, \\ &\quad = \exp(\bar{y}_{0,1:T}) \cdot \sum_{d_1,\ldots,d_N \in \mathscr{D}} p_\theta(\mathbf{s} = \mathbf{0}, d_1, \ldots, d_N). \end{aligned} \tag{14}$$

Let $\gamma^{[T]}_\theta \triangleq \sum_{d_1,\ldots,d_N \in \mathscr{D}} p_\theta(\mathbf{s} = \mathbf{0}, d_1, \ldots, d_N)$ be the prior term in Eq. (14). The variable $\gamma^{[T]}_\theta$ can be calculated via the following recursive formula:

$$\gamma^{[t]}_\theta = \sum_{d=1}^{t} p_\theta(d \mid s = 0) \cdot \psi_{00} \cdot \gamma^{[t-d]}_\theta, \tag{15}$$

where the initial values $\gamma^{[0]}_\theta$ is set to $\rho_0/\psi_{00}$. In practice, since the duration probability of silence $p_\theta(d \mid s = 0)$ is assumed to be uniform (i.e., $1/T$), $\gamma^{[T]}_\theta$ becomes a simple form as follows:

$$\gamma^{[T]}_\theta = \frac{\rho_0}{T}\left(1 + \frac{\psi_{00}}{T}\right)^{T-1}. \tag{16}$$

## 3.2. EMA-based pseudo-supervised training

To train the model with unlabeled data, we propose EMA-based pseudo-supervision for the event-wise posterior probabilities. We transfer onset-and-offset knowledge from an EMA-based teacher model with parameters $\{\theta', \phi'\}$ to a student model with parameters $\{\theta, \phi\}$ by utilizing its event-wise predictions. The student model is trained such that its event-wise posteriors $p_{\theta,\phi}(\tilde{s}_t, \tilde{d}_t \mid \mathbf{X})$ are consistent with those of the teacher model $q_{\theta',\phi'}(\tilde{s}_t, \tilde{d}_t \mid \mathbf{X})$, as an probabilistic version of Eq. (11):

$$\begin{aligned} \mathcal{L}^{(\mathrm{uns})}_{\theta,\phi} = -\frac{1}{T}\sum_{s\in\mathcal{S}}\sum_{t=1}^{T}\sum_{d=1}^{t} B_d \Big\{ & \\ & q_{\theta',\phi'}(\tilde{s}_t = s, \tilde{d}_t = d \mid \mathbf{X}) \\ & \cdot \log p_{\theta,\phi}(\tilde{s}_t = s, \tilde{d}_t = d \mid \mathbf{X}) \Big\}. \end{aligned} \tag{17}$$

Here, we introduced a weight hyperparameter $B_d$ for each event. Since longer events are used for more time frames in Eq. (10), they should be trained with greater emphasis. We, thus, set $B_d$ to $d$ since event $\{\tilde{s}_t = s, \tilde{d}_t = d\}$ is used for $d$ times to calculate $p_{\theta,\phi}(z_t \mid \mathbf{X})$ for $t = 1, \ldots, T$ in Eq. (10).

## 3.3. Mixup for event-wise training

To improve the detection performance, we introduce a Mixup augmentation extended for event-wise training. Mixup [24] mixes two samples consisting of inputs $x_1, x_2$ and labels $z_1, z_2$ with a weight $r \in [0, 1]$ to make a new training sample $x, z$:

$$x = rx_1 + (1 - r)x_2, \quad z = rz_1 + (1 - r)z_2. \tag{18}$$

However, it is not trivial how to mix the event-wise label $\{\sigma_{c,n}, \delta_{c,n}\}_{n=1}^{N}$. We focus on the fact that the BCE loss for a new sample $\mathcal{L}^{(\mathrm{fr})}_\phi$ can be regarded as a weighted sum of the BCE losses for individual labels $\mathcal{L}^{(\mathrm{fr})}_{\phi,1}$ and $\mathcal{L}^{(\mathrm{fr})}_{\phi,2}$:

$$\mathcal{L}^{(\mathrm{fr})}_\phi = r\mathcal{L}^{(\mathrm{fr})}_{\phi,1} + (1 - r)\mathcal{L}^{(\mathrm{fr})}_{\phi,2}. \tag{19}$$

Based on this observation, we apply Mixup to training by mixing the loss functions for individual labels $\mathcal{L}^{(*)}_{\theta,\phi,1}$ and $\mathcal{L}^{(*)}_{\theta,\phi,2}$:

$$\mathcal{L}^{(*)}_{\theta,\phi} = r\mathcal{L}^{(*)}_{\theta,\phi,1} + (1 - r)\mathcal{L}^{(*)}_{\theta,\phi,2}, \tag{20}$$

where $*$ denotes one of the two loss functions (str and weak) for event-wise training. Note that $\mathcal{L}^{(\mathrm{uns})}_{\theta,\phi}$ doesn't need Mixup as there are no ground-truth labels.

# 4. Experimental Evaluation

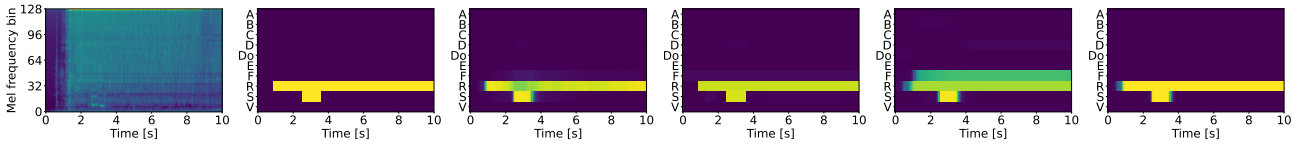
We evaluated our weak-supervision and semi-supervision for the DNN-HSMM on a real-recorded DESED dataset [25].

## 4.1. Dataset

The DESED dataset is a widely used standard benchmark and consists of 59.2 hours of 10-second real-recorded audio clips. The dataset consists of $C = 10$ classes in domestic environments: alarm/bell/ringing (A), blender (B), cat (C), dishes (D), dog (Do), electric shaver/toothbrush (E), frying (F), running water (R), speech (S), and vacuum cleaner (V). The training set consists of 3,470, 1,578, and 14,412 clips of strongly labeled, weakly labeled, and unlabeled audio data, respectively. The validation and evaluation sets consist of 1,168, and 692 clips. To supplement the shortage of real-recorded training data, we also used 10,000 synthetic mixture clips generated by Scaper [33]. This configuration follows that of DCASE Challenge 2023 Task 4. All the audio recordings were resampled at 16 kHz.

## 4.2. Experimental configuration

We utilized a DNN consisting of BEATs [32] and CRNN, referred to as BEATs-CRNN. This architecture followed the baseline system of DCASE Challenge 2023 Task 4. A 7-layer CNN transforms a log-mel spectrogram to 128-dimensional features. The pre-trained BEATs was frozen and used to obtain 756-




(a) *Input audios* (b) *Annotated labels* (c) *B3: DNN w/o pp.* (d) *D3: DNN + SEBBs* (e) *P1: Ours (str. only)* (f) *P3: Ours*

Figure 2: *Examples of prediction results by the baseline and proposed methods.*

dimensional features by 12-layer Transformer encoders with 8 heads. These features were concatenated and transformed into 256-dimensional features by a 2-layer bidirectional gated recurrent unit. The log emission probabilities $\mathbf{y}_t$ are obtained by two linear layers with batch normalization and rectified linear units (ReLUs). Following [17], the model is also designed to output the frame-wise predictions for stabilizing the training.

The models were trained by an Adam optimizer [34] and an exponential warmup increased the learning rate to $1.0 \times 10^{-3}$ after 50 epochs. We obtained log-mel spectrograms with $F = 256$ bins, 128 ms window, and 16 ms stride. As in [17], we used an HSMM mixture and had three hyperparameters regarding HSMM architecture. We set $\psi_{01}$ and $\psi_{11}$ to 0.01, and mixed 8 HSMMs. We applied Mixup [24] with 50% probability following the DCASE Challenge 2023 Task 4 baseline system. We clamped $N$ of Eq. (11) at 7 so that the loss for clips with frequent events did not become too small. The loss function for the HSMM $\mathcal{L}_{\theta,\phi}^{(\mathrm{hsmm})}$ was a sum of the strong-, weak-, and semi-supervision loss $\mathcal{L}_{\theta,\phi}^{(\mathrm{str})} + \mathcal{L}_{\theta,\phi}^{(\mathrm{weak})} + \alpha\mathcal{L}_{\theta,\phi}^{(\mathrm{uns})}$, where the weight $\alpha$ was increased to 2.0 from 0.0 with the rate of the warmup. Our total loss is a weighted sum of the loss $\mathcal{L}_{\theta,\phi}^{(\mathrm{hsmm})}$ and the frame-wise loss of Eq. (1): $0.1\mathcal{L}_{\theta,\phi}^{(\mathrm{hsmm})} + \mathcal{L}_{\phi}^{(\mathrm{fr})}$. These hyperparameters were determined by using the validation set.

We compared our method with the BEATs-CRNN-based systems with existing post-processing methods. The network architecture was the same as the Challenge baseline. As post-processing methods, we evaluated the standard median filtering and SEBBs [13]. We measured the detection performance with event-based F1 [35] (Eb-F1), intersection-based F1 [36] (Ib-F1), and two types of polyphonic sound detection scores (PSDS1 and 2) [37]. PSDS1 and PSDS2 mainly focus on temporal localization errors and class confusion, respectively [38]. We tuned the hyperparameters and thresholds of post-processing to maximize the Eb-F1 on the validation set. We reported average scores for 3 random seeds. As a reference, we also show the performance of the state-of-the-art (SOTA) model called ATST-SED [9]. Note that this model cannot be directly compared because it is built by using 7,384 strongly labeled clips from AudioSet in addition to the DESED dataset.

### 4.3. Experimental results

Table 1 shows the detection performance of the proposed and baseline methods using weakly labeled and/or unlabeled data in addition to strongly labeled data. Under all conditions, our BEATs-CRNN-HSMM (P*) significantly improved the detection performance of the BEATs-CRNN without post-processing (B*) across all the metrics. Our method also outperformed the median filtering (C*) and the recently proposed post-processing called SEBBs (D*) under all the conditions.

Table 1 also shows the effectiveness of each proposed technique. In the training of the BEATs-CRNN-HSMM, introducing the weak-supervision improved all metrics (P1→P2), and the performance was further improved by introducing the semi-supervision (P2→P3). Fig. 2 shows examples of prediction results. Our HSMM (Fig.2-(f)) and SEBB (Fig.2-(d)) successfully suppressed temporal fluctuations in predictions without post-processing (Fig.2-(c)). Our method (Fig.2-(f)) predicted event existence more precisely compared to the DNN-HSMM only with strong supervision (Fig.2-(e)). By applying our Mixup, each metric was improved or maintained (A*→P*), indicating its effectiveness. We can also see that weights $B_d$ for the semi-supervision of Eq. (17) improved the performance (F3→P3).

Table 1: *Detection performance for DESED.*

| ID | Method | PSDS1 | PSDS2 | Eb-F1 | Ib-F1 |
|---|---|---|---|---|---|
| B1 | BEATs-CRNN | 0.448 | 0.744 | 0.542 | 0.761 |
| B2 | + Weak-supervision loss | 0.533 | 0.775 | 0.567 | 0.813 |
| B3 | + Semi-supervision loss | 0.565 | 0.815 | 0.620 | 0.840 |
| C1 | BEATs-CRNN + Med. filt. | 0.456 | 0.772 | 0.552 | 0.772 |
| C2 | + Weak-supervision loss | 0.544 | 0.808 | 0.586 | 0.826 |
| C3 | + Semi-supervision loss | 0.576 | 0.845 | 0.635 | 0.856 |
| D1 | BEATs-CRNN + SEBBs | 0.496 | 0.759 | 0.628 | 0.788 |
| D2 | + Weak-supervision loss | 0.569 | 0.780 | 0.626 | 0.813 |
| D3 | + Semi-supervision loss | 0.608 | 0.821 | 0.681 | 0.858 |
| P1 | BEATs-CRNN-HSMM | 0.570 | 0.788 | 0.652 | 0.826 |
| A1 | w/o Mixup | 0.548 | 0.776 | 0.657 | 0.826 |
| E1 | w/o BEATs [17] | 0.458 | 0.594 | 0.540 | 0.702 |
| P2 | + Weak-supervision loss | 0.600 | 0.817 | 0.677 | 0.848 |
| A2 | w/o Mixup | 0.586 | 0.817 | 0.657 | 0.832 |
| P3 | + Semi-supervision loss | **0.630** | 0.847 | **0.690** | **0.863** |
| A3 | w/o Mixup | 0.570 | **0.851** | 0.668 | 0.839 |
| F3 | w/o weight $B_d$ | 0.611 | 0.820 | 0.660 | 0.845 |
| G3 | ATST-SED[1] [9] | 0.631 | 0.846 | – | – |

This study is to extend the CRNN-HSMM proposed in [17] to achieve a practical performance by extending the training framework (E1→P3). Our method (P3) achieved a performance comparable to the SOTA model called ATST-SED (G3). Note that ATST-SED was built by fine-tuning the pre-trained model ATST [7] with a CRNN and also used 7,384 extra strongly labeled clips for the fine-tuning.

## 5. Conclusion

This paper presented probabilistically formulated training of the onset-and-offset aware DNN-HSMM with weak and unlabeled data. For weak supervision, we derived a probabilistic formula to estimate clip-level event existence. Unlabeled data are handled by the event-wise extensions of EMA-based pseudo-supervision. To further improve performance, we extend Mixup for event-wise training. In experiments, our method outperformed the frame-wise baselines with widely used post-processing methods. Our future work includes investigating the effectiveness of our HSMM for various frame-wise models.

[1] https://github.com/Audio-WestlakeU/ATST-SED

# 6. References


[1] J. W. Kim, S. W. Son, Y. Song, H. K. Kim, I. H. Song, and J. E. Lim, "Semi-supervised learning-based sound event detection using frequency dynamic convolution with large kernel attention for DCASE Challenge 2023 Task 4," DCASE2023 Challenge, Tech. Rep., 2023.

[2] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1291–1303, 2017.

[3] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys*, vol. 48, no. 4, 2016, Art. no. 52.

[4] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 165–168.

[5] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[6] Q. Jin, P. F. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," in *Proc. Interspeech*, 2012, pp. 2085–2088.

[7] X. Li, N. Shao, and X. Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1336–1351, 2024.

[8] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2020, pp. 100–104.

[9] N. Shao, X. Li, and X. Li, "Fine-tune the pretrained ATST model for sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 911–915.

[10] Y. Guan, J. Han, H. Song, W. Song, G. Zheng, T. Zheng, and Y. He, "Contrastive loss based frame-wise feature disentanglement for polyphonic sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1021–1025.

[11] H. Dinkel and K. Yu, "Duration robust weakly supervised sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 311–315.

[12] S. Xiao, J. Shen, A. Hu, X. Zhang, P. Zhang, and Y. Yan, "Sound event detection with weak prediction for DCASE 2023 Challenge Task4A," DCASE2023 Challenge, Tech. Rep., 2023.

[13] J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, "Sound event bounding boxes," in *Proc. Interspeech*, 2024, pp. 562–566.

[14] Z. Ye, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, and K. Ouchi, "Sound event detection Transformer: An event-based end-to-end model for sound event detection," *arXiv preprint arXiv:2110.02011*, 2021.

[15] S. Venkatesh, D. Moffat, and E. R. Miranda, "You only hear once: A YOLO-like algorithm for audio segmentation and sound event detection," *Applied Sciences*, vol. 12, no. 7, 2022, Art. no. 3293.

[16] S. Bhosale, S. Nag, D. Kanojia, J. Deng, and X. Zhu, "DiffSED: Sound event detection with denoising diffusion," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 792–800.

[17] T. Yoshinaga, K. Tanaka, Y. Bando, K. Imoto, and S. Morishima, "Onset-and-offset-aware sound event detection via differentiable frame-to-event mapping," *IEEE Signal Processing Letters*, vol. 32, pp. 186–190, 2025.

[18] S.-Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, 2010.

[19] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 31–35.

[20] S.-Y. Tseng, J. Li, Y. Wang, J. Szurley, F. Metze, and S. Das, "Multiple instance deep learning for weakly supervised small-footprint audio event detection," in *Proc. Interspeech*, 2018, pp. 3279–3283.

[21] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 316–320.

[22] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 121–125.

[23] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[24] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–13.

[25] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2019, pp. 253–257.

[26] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 Task 4," DCASE2019 Challenge, Tech. Rep., 2019.

[27] L. Cances, P. Guyot, and T. Pellegrini, "Evaluation of post-processing algorithms for polyphonic sound event detection," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 318–322.

[28] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 86–90.

[29] T. K. Chan and C. S. Chin, "Multi-branch convolutional macaron net for sound event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2972–2985, 2021.

[30] Y. Guan, J. Xue, G. Zheng, and J. Han, "Sparse self-attention for semi-supervised sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 821–825.

[31] J. F. Gemmeke, D. P. W. 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 Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 776–780.

[32] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Int. Conf. Machine Learn.*, 2023, pp. 5178–5193.

[33] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 344–348.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

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

[36] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 61–65.

[37] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 1021–1025.

[38] F. Ronchini and R. Serizel, "A benchmark of state-of-the-art sound event detection systems evaluated on synthetic soundscapes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 1031–1035.