Analysis of weak labels for sound event tagging

Nicolas Turpault, Romain Serizel, Emmanuel Vincent

Analysis of weak labels for sound event tagging

Nicolas Turpault Romain Serizel Emmanuel Vincent
Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

***Abstract*—Weak labels are a recurring problem in the context of ambient sound analysis. While multiple methods using neural networks have been proposed to address it, limited attention has been given to the analysis of the problem to have a better understanding of it. Many of these methods seem to improve detection or tagging performance, but they have been evaluated in scenarios where other problems such as unreliable labels, overlapping sound events, or class unbalance also occur. Therefore, it is difficult to conclude whether the observed improvement is due to solving the problem of weak labels or not. In this article, we provide for the first time a detailed analysis of the impact of weak labels independently of other problems on a sound event tagging system. We show that, in order to limit the negative impact of weak labels on the performance, the training clips must be at least as long as the test clips and longer training clip durations have a minor impact. We also show that good temporal aggregation can help to reduce this impact at test time and provide insight on the annotation granularity needed depending on the targeted scenario.**

***Index Terms*—Audio tagging, weak labels, temporal aggregation.**

I. Introduction

The interest for ambient sound analysis research has grown in the recent years because of its potential applications like assisted living, urban monitoring, or animal tracking [1]. This has led to the release of ambient sound datasets collected from various sources and targeting different applications [2]–[6]. Given a single-channel audio clip, one of the ultimate goals is to find which sound events have happened and when. This task is called *sound event detection* (SED). In principle, it can be addressed in a fully supervised fashion given *strongly labeled* training data, whose labels indicate not only the sound event classes which are present in each training clip but also the corresponding timestamps. However, such data is rare because labeling the timestamps corresponding to each sound event class is time-consuming.

To overcome this issue, sound class labels without timestamps, also known as *weak labels*, are often used instead. Weak labeling is 3 to 15 times faster than strong labeling [7] and, contrary to strong labeling which requires special expertise, it can be achieved via crowdsourcing [8]. Weakly labeled training data can be used for *weakly supervised learning*, either alone or in combination with strongly labeled training data. In the case of SED, this includes inferring a segmentation model from training clips that do not provide a reference segmentation. Multiple weakly supervised learning approaches have been proposed in various domains [9]–[14]. They often rely on multiple instance learning (MIL) [15], [16], various kinds of temporal aggregation [17], [18], attention pooling [19], or self-attention [20]. The review in [21] reported the performance of different SED systems trained on unlabeled and weakly labeled data for different segmentation tolerance collars. However, it did not analyze the impact of weak labels at training time.

Sound event tagging (SET) is the task of finding which sound events are present in an audio clip regardless of when. It is a simpler task than SED, since the desired system outputs at test time are of the same type as the weak labels available at training time. A few works have tried to exploit weak labels for SET based on MIL [22], Gaussian filters [23], or attention [24]. To the best of our knowledge, only two works so far have analyzed the impact of weak labels on SET. Shah et al. [25] made an attempt in that direction. However, they relied on Audioset [2] which exhibits many other problems than weak labels: unreliable or missing labels [26], overlapping sound events [27], variable event-to-background ratio [28], class unbalance [29], and long-term dynamics [30]. Also, they used 10 s weakly labeled audio clips and expanded them into 30 or 60 s clips by taking a longer piece of audio from the original Youtube videos. Their analysis did not take into account the duration of the target sound event in the original clip, which can drastically affect the impact of weak labels [31], nor the fact that additional instances of that sound event may be present in the expanded clip. To overcome these drawbacks, in [31], we created a suitable dataset focusing on the problem of weak labels independently of other problems and we assessed the impact of weak labels on the resulting embeddings and classifier. This preliminary work faced several limitations. First, part of the study focused on sound events truncated to 200 ms which is too short to effectively recognize some of the sound event classes of interest. Second, it did not study the role of temporal aggregation. Third, it reported summary results across all sound event classes which are hard to interpret due to their different event durations.

In this article, we study the impact of weak labels on an SET system, especially the impact of the aggregation of frame-level quantities into clip-level scores. To do so, we created a new dataset, where sound events are truncated not only to 200 ms but also to longer durations, and we consider a larger set of clip durations. We measure the tagging performance as a function of the duration of the training and test clips with respect to that of the events and assess the impact of the aggregation function. We consider several temporal aggregation functions (mean, max, softmax) and we propose to use L_p aggregation as a parametric aggregation function whose behavior depends

The authors are with Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France (e-mail: nicolas.turpault@inria.fr, romain.serizel@loria.fr, emmanuel.vincent@inria.fr). This work was made with the support of the French National Research Agency, in the framework of the project LEAUDS "Learning to understand audio scenes" (ANR-18-CE23-0020) and the French region Grand-Est. High Performance Computing resources were partially provided by the EXPLOR centre hosted by the University de Lorraine.

on p.

The article is organized as follows. In Section II, we define the problem of weak labels. Section III describes the choice of aggregation methods and the hypothesized impact of weak labels. In Section IV, we define the experimental setup and in Section V we conduct a series of experiments to validate our hypotheses regarding the impact of weak labels. We conclude in Section VI.

II. Problem definition

Let us denote by $\mathbf{X}$ the time-frequency representation of an audio clip. $\mathbf{X}$ is a matrix of size $F \times T$ where F is the number of frequency bins and T is the number of time frames. Let us further denote by $\mathscr{C} = \{1, \dots, C\}$ the set of sound event classes of interest. We assume that $\mathbf{X}$ contains one or more instances of sound events in $\mathscr{C}$ on top of a stationary background.

The ground-truth strong labels for this clip can be represented in the form of a matrix $\mathbf{Y}$ of size $C \times T$ where C is the number of classes and every $y_{c,t} \in \{0, 1\}$ encodes the absence or presence of class c at time t. The vector $\mathbf{y}_t = [y_{1,t}, \dots, y_{C,t}]^T$ of size $C \times 1$ represents the activity of all classes at time t.

Similarly, the ground-truth weak labels for this clip can be represented in the form of a vector $\mathbf{w} = [w_1, \dots, w_C]^T$ of size $C \times 1$ whose entries w_c satisfy

$$w_c = \max_t(y_{c,t}). \tag{1}$$

This is because, when a class is present in at least one time frame, it is present in the clip and the corresponding weak label should be set to 1. This is represented in Fig. 1. Both $\mathbf{Y}$ and $\mathbf{w}$ are assumed to be correct, i.e., there are no labeling errors.

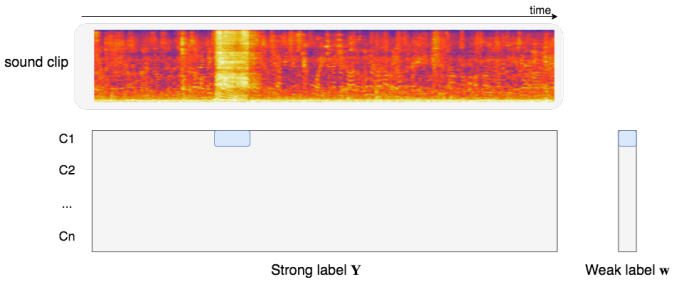


Fig. 1: Illustration of the difference between strong and weak labels.

Since we are interested in weak labels, we define the event density as follows:

$$D_c = \frac{1}{T} \sum_{t=1}^{T} y_{c,t}. \tag{2}$$

The longer an event is present in the clip, the higher D_c. When the event density is equal to 0 or 1, it actually corresponds to the ground-truth weak label w_c. Otherwise, weak labels introduce a certain amount of noise in the learning process, which can be quantified as $w_c - D_c$.

SET is the task of detecting which events appear in an audio clip, regardless of how many times they appear and when. In other words, it consists of estimating a vector $\widehat{\mathbf{w}}$ of clip-level class activity scores from $\mathbf{X}$. These clip-level scores are obtained by estimating frame-level quantities[1] encoding the activity of the sound event classes and aggregating them along the time axis using a fixed or parametric aggregation function. Depending whether the training data are strongly or weakly labeled, this aggregation stage can happen only at test time or both at training and test time, and the model used to estimate frame-level quantities can be trained according to frame-level ($\mathbf{Y}$) or clip-level ($\mathbf{w}$) targets. Also, the parameters of the aggregation function can be trained or fixed.

To study the impact of weak labels on SET, we propose to vary the event density and analyze the impact of the noise introduced by weak labels.

III. Expected impact of temporal aggregation

Obtaining clip-level scores $\widehat{\mathbf{w}} = [\widehat{w}_1, \dots, \widehat{w}_C]^T$ from frame-level quantities requires some form of temporal aggregation. In this section we present popular aggregation methods and a new L_p aggregation method, and we discuss their expected impact depending the event density. The studied methods are illustrated in Fig. 2, where $\mathbf{O}$ denotes the $N \times T$ frame-level representation obtained at the penultimate model layer, and $\widehat{\mathbf{Y}}$ denotes the $C \times T$ matrix of frame-level posterior probabilities $\widehat{y}_{c,t} = P(y_{c,t} = 1|\mathbf{X})$, whose dimension is identical to $\mathbf{Y}$ and which is derived from $\mathbf{O}$ by applying a time-distributed linear layer with trainable weights and biases $\{\mathbf{W}_1, \mathbf{b}_1\}$ followed by a sigmoid activation.

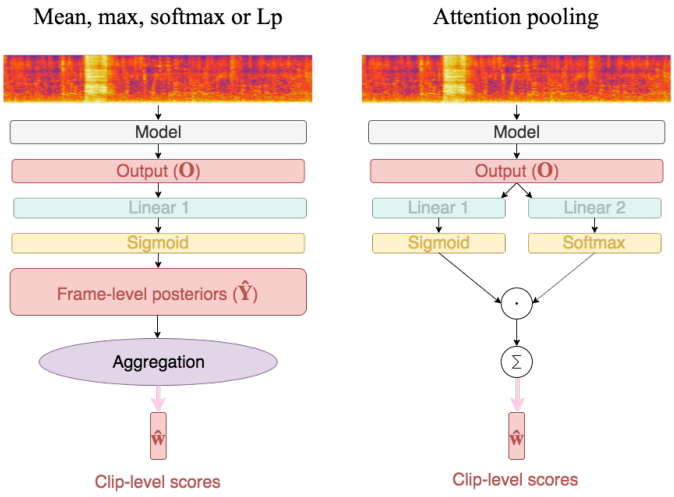


Fig. 2: Considered aggregation methods.

A. Mean pooling

Mean pooling is a common aggregation method that averages the frame-level posteriors over time:

$$\widehat{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathbf{y}}_t. \tag{3}$$

[1]For notation simplicity, we assume that the model inputs and outputs have the same time resolution indexed by $t \in \{1, \dots, T\}$. In reality, the time resolutions of inputs and outputs may be different.

When $D_c \neq 0$, the drawback of this method is that, the lower D_c, the poorer the clip-level score $\widehat{w}_c$ will be since it is expected to behave like D_c. To understand this, let us consider the case when a system provides perfect frame-level posteriors $\widehat{y}_{c,t} = y_{c,t}$, so that $\widehat{w}_c = D_c$. We denote by τ the decision threshold, such that only the events for which $\widehat{w}_c \geq \tau$ are tagged as present. In that case, when $D_c < \tau$, the event is wrongly tagged as absent (i.e., the false negative rate is nonzero), despite the fact that the frame-level estimates are perfect. In practice, the value of D_c below which this happens varies due to the fact that frame-level posteriors are not perfect, but this observation remains valid on average for unbiased estimates, as shown in Fig. 3. Lowering the threshold τ decreases the false negative rate but increases the false positive rate. Therefore, mean pooling is not a good choice except for very long sound events with a high density $D_c \simeq 1$.

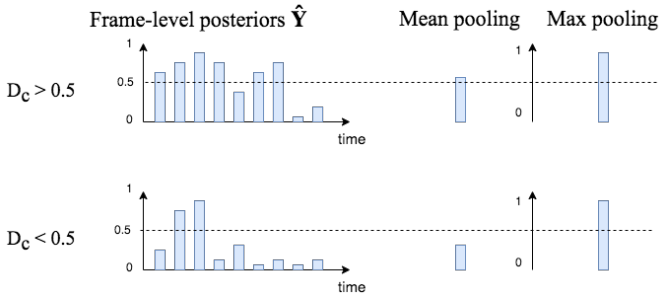


Fig. 3: Mean and max pooling ($\tau = 0.5$).

B. Max pooling

Max pooling is defined as

$$\widehat{\mathbf{w}} = \max_{t \in \{1,\dots,T\}} \widehat{\mathbf{y}}_t \tag{4}$$

where the maximum is computed elementwise, i.e., the clip-level score for each event is assumed to be the highest frame-level posterior, as shown in Fig. 3. While this aggregation method seems appropriate since it matches the way ground-truth weak labels are defined in (4), it suffers from two drawbacks in practice. First, it is very sensitive to false positives in the frame-level posteriors, which systematically translate into false positives at the clip level. Second, the output is not differentiable with respect to all inputs: at training time, the gradient is only backpropagated with respect to the maximum input (subgradient), which can be an issue as explained by McFee et al. [17].

C. Softmax pooling

Softmax pooling is defined as

$$\widehat{\mathbf{w}} = \sum_{t=1}^{T} \widehat{\mathbf{y}}_t \odot \frac{\exp(\widehat{\mathbf{y}}_t)}{\sum_{t'=1}^{T} \exp(\widehat{\mathbf{y}}_{t'})} \tag{5}$$

where $\odot$ denotes elementwise multiplication, and the exponential and division operations are also computed elementwise. This method can be seen as a tradeoff between mean and max pooling: it is differentiable with respect to all inputs like mean pooling while being closer to max pooling in the way it aggregates frame-level posteriors. Despite its wide use, it lacks flexibility since this tradeoff cannot be controlled. When $D_c \simeq 1$, we expect softmax pooling to work well similarly to mean pooling. When $D_c < \tau$, softmax pooling will continue to give correct clip-level scores in many cases, while mean pooling fails. However, when D_c is low, softmax pooling will also start failing while max pooling could work. When using softmax pooling, false positive frames with high posteriors can have a big impact on the clip-level score (like for max pooling), but many false positive frames with small posteriors can also have an adverse impact (like for mean pooling).

D. L_p aggregation

L_p aggregation is inspired by the computation of the L_p norm. The difference stands in taking the average of the entries raised to the power of p instead of the sum. It can be seen as another tradeoff between mean and max pooling. This form of aggregation was previously used to pool scores across multiple signals when performing SED after signal separation [32]. We now propose to use it as an aggregation function over time:

$$\widehat{\mathbf{w}} = \left(\frac{1}{T} \sum_{t=1}^{T} \widehat{\mathbf{y}}_t^p \right)^{\frac{1}{p}} \tag{6}$$

where $p > 0$ and the exponentiations are computed elementwise. When $p = 1$, L_p aggregation is equivalent to mean pooling. When $p \to \infty$, it behaves like max pooling but it remains differentiable with respect to all inputs. This solution offers some flexibility since we can vary p depending on the sound class of interest and the targeted application. In Fig. 4, we represent the L_p aggregation function for different values of p, and show that softmax is close to L_p aggregation with $p = 2$. We expect a small value of p to be preferable when D_c is large and vice-versa.[2]

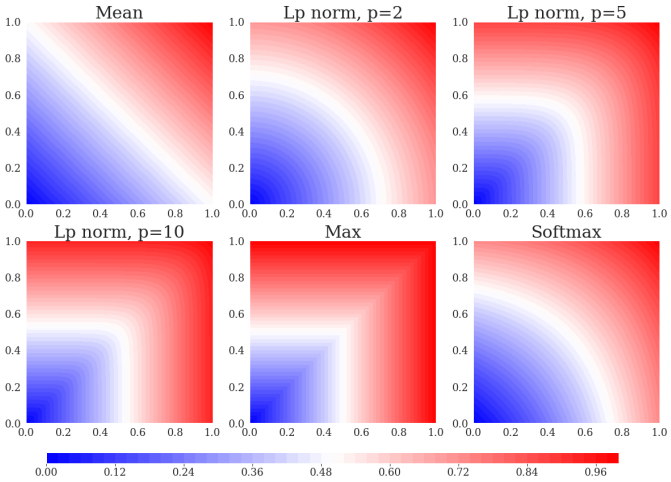


Fig. 4: Comparison of L_p aggregation and other fixed aggregation functions.

[2]Note that p could also be trained similarly to the auto-pooling method in [17]. This is out of the scope of this article since we aim to assess the impact of weak labels depending on the value of p.

E. Attention pooling

Attention pooling has become a popular aggregation method in recent works [16], [19], [24], [33], and it has many variants. For simplicity, we focus on the most common variant:

$$\widehat{\mathbf{w}} = \sum_{t=1}^{T} \widehat{\mathbf{y}}_t \odot \text{softmax}_t(\mathbf{A}) \tag{7}$$

where

$$\text{softmax}_t(\mathbf{A}) = \frac{\exp(\mathbf{a}_t)}{\sum_{t'=1}^{T} \exp(\mathbf{a}_{t'})}, \tag{8}$$

all operations are performed elementwise, and $\mathbf{a}_t = \mathbf{W}_2\mathbf{o}_t + \mathbf{b}_2$ is obtained from a time-distributed linear layer whose weights and biases $\{\mathbf{W}_2, \mathbf{b}_2\}$ are different from those used to obtain $\widehat{\mathbf{y}}_t$.

By contrast with the above aggregation functions which have no trainable parameter,[3] attention pooling has a sizeable number of trainable parameters, namely $\{\mathbf{W}_2, \mathbf{b}_2\}$. Because of this, it is expected to perform well in a wider range of situations and values of D_c [19], [24].

IV. Experimental setup

To assess the impact of weak labels and that of the aggregation function on an SET system, we created a new dataset which makes it possible to assess this impact independently of other problems such as multiple labels, unbalanced classes, or overlapping events. We describe this dataset below, as well as the considered baseline system and evaluation metric.

A. Datasets

1) WAA dataset: In [31], we introduced the Weak Annotation Analysis (WAA) dataset, which consists of audio clips generated by mixing a single sound event from Freesound [34] with a random background from SINS [35] or MUSAN [36] at a random signal-to-noise ratio (SNR) between 6 dB and 30 dB. The 10 considered event classes are: alarm/bell/ringing, blender, cat, dishes, dog, electric shaver/toothbrush, frying, running water, speech, and vacuum cleaner. The training, validation, and evaluation sets contain 2,700, 300, and 750 audio clips generated from 909, 100, and 314 unique events, respectively, and they are balanced, i.e., they contain the same number of clips for each class.

The duration of the clips in the WAA dataset is $d_{\text{clip}} = 10$ s. In order to assess the impact of the clip duration on the SET performance, we cut each clip into five shorter clips with respective duration $d_{\text{clip}} = 0.2$, 0.5, 1, 3, or 5 s.[4] The start and end times of each shorter clip within the original 10 s clip are such that the shorter clip contains the entire sound event (when the event duration is shorter than d_{clip}) or the event spans the entire clip (when the event duration is longer than d_{clip}). The shorter the clip, the higher the event density D_c.

[3]other than the weights and biases $\{\mathbf{W}_1, \mathbf{b}_1\}$ used to derive $\widehat{\mathbf{Y}}$ from $\mathbf{O}$, which are not part of the aggregation process itself

[4]In [31], we considered three clip durations only: 0.2, 1, or 10 s.

2) TWAA dataset: While the WAA dataset preserves the original duration of all events, the fact that these durations are different and their distributions are highly class-dependent influences the ensuing analysis. In order to address this issue, we introduce the Truncated Weak Annotation Analysis (TWAA) dataset. This dataset consists of five subsets, each of which has the same structure and is generated in the same way as the WAA dataset, except that all sound events whose duration is longer than a maximum duration d_{max} are truncated to d_{max} before they are mixed with the backgrounds, and all sound events whose duration is lower than a minimum duration d_{min} are discarded. The five truncated event duration intervals $[d_{\text{min}}\ d_{\text{max}}]$ are [0 0.5], [0.5 1], [1 3], [3 5], and [5 10] s.[5]

The resulting 10 s clips are then cut into five shorter clips with duration d_{clip} in the same way as above. Again, the start and end times of each shorter clip are chosen such that it contains the entire truncated sound event (when the truncated event duration is shorter than d_{clip}) or the truncated event spans the entire clip (when the truncated event duration is longer than d_{clip}). As a result, the event density satisfies $D_c < d_{\text{max}}/d_{\text{clip}}$. In the situation when $d_{\text{max}} < d_{\text{clip}}$, the event density becomes $D_c < 1$ for all clips. This situation will be marked with dashed lines in the plots below.

B. Baseline system

As the starting point for our experiments, we use a well-known system: the baseline system for Task 4 of the DCASE 2019 and 2020 Challenges [7], [37]. This system is inspired from previous DCASE submissions [38]–[40], and it has been extensively studied and compared to state-of-the-art systems [21], [37], [41]. While it was initially designed for SED, we use it for SET hereafter.

The baseline system comprises a convolutional neural network (CNN) followed by a bidirectionnal recurrent neural network (BiRNN), forming an architecture known as CRNN (see Fig. 5). The output $\mathbf{O}$ of the BiRNN is transformed into frame-level posteriors $\widehat{\mathbf{Y}}$ by means of a time-distributed linear layer followed by a sigmoid. These frame-level posteriors are then converted into clip-level scores using an aggregation function. This sytem was trained on the DESED dataset [7] which contains heterogeneous (unlabeled, weakly labeled, and strongly labeled) data from both recorded and generated soundscapes.

C. Metric

To quantify the SET performance, we use an SET metric: the clip-level F1-score. The 90% confidence interval on this score is computed using the bootstrap method [42] using 200 iterations with 80% of the data per iteration.

V. Experimental results and analysis

Starting from the above baseline system, we now perform a series of experiments on the original WAA dataset and the newly proposed TWAA dataset in order to assess the impact

[5]This contrasts with the choice of a single, short truncation duration $d_{\text{max}} = 200$ ms in [31].

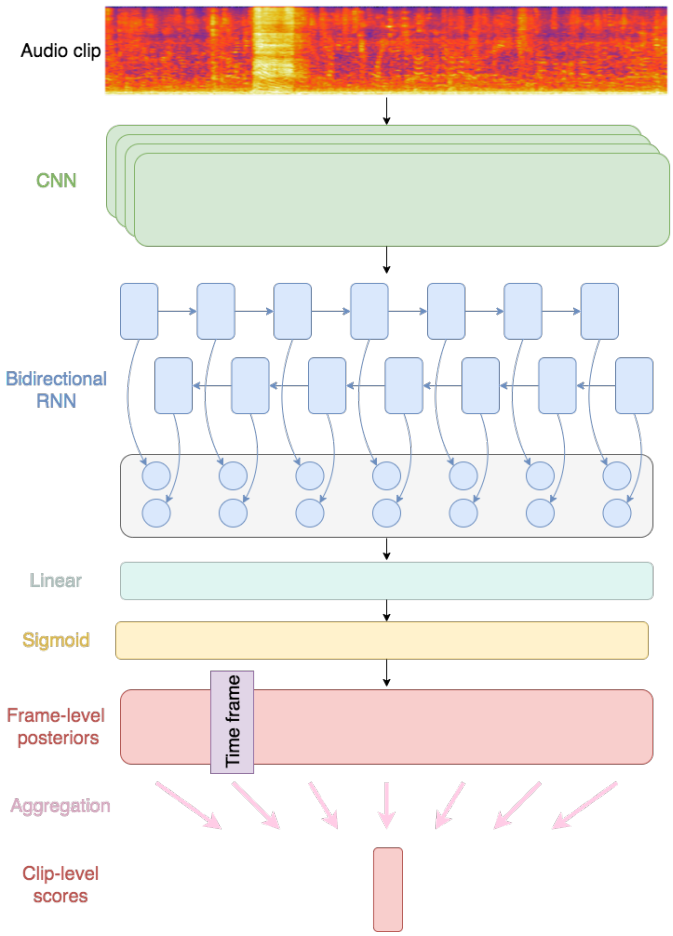


Fig. 5: Architecture of the baseline system.

of various aggregation functions at training and test time as a function of the clip and event durations, and we derive suitable system design choices.

A. Impact of test-time aggregation on the pretrained baseline

Our first experiment uses the frame-level posteriors output by the baseline system in Section IV-B without retraining. It focuses on the aggregation of these frame-level posteriors into clip-level scores at test time on the TWAA dataset.

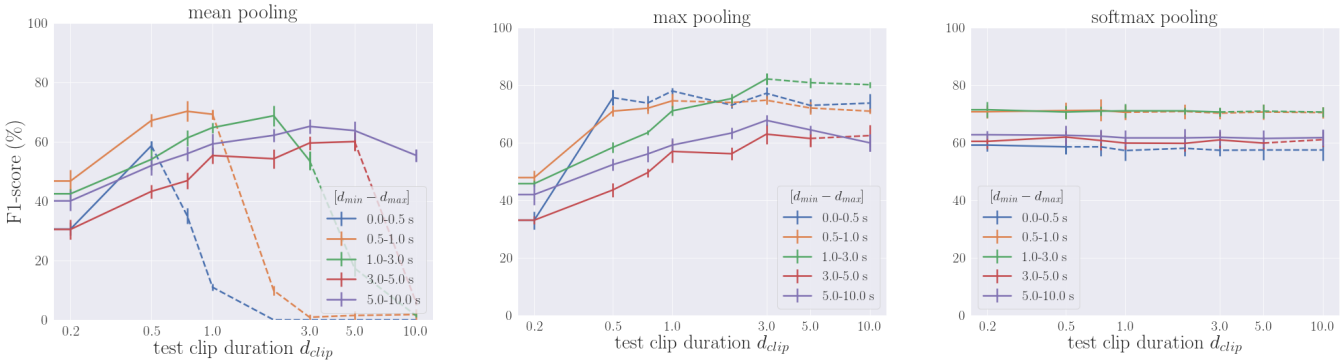
Figure 6 evaluates the three classical aggregation functions, namely mean, max, and softmax pooling, as a function of the test clip duration and the truncated event duration. The results confirm the expected behavior discussed in Section III. Mean pooling (6a) is not efficient when $D_c < 1$, as indicated by the dashed lines. Max pooling is very efficient in this scenario since it is performed at test time only (i.e., its drawback at training time explained in Section III-B does not arise). The best SET performance is achieved for $d_{\text{clip}} = 3$ s: for longer clips, the chance that the maximum is a false positive increases. Softmax pooling is almost not impacted by the clip duration. This makes it an interesting aggregation function in this scenario, however its performance is not always as good as that obtained with max pooling, especially for events shorter than 3 s.

Figure 7 compares various aggregation functions for a fixed clip duration. The first one, denoted as "weak" in the figure, is the attention pooling layer of the baseline system. Since the baseline system was initially designed for SED, this layer was trained on weakly labeled data only (10% of the DESED training set) as opposed to the entire DESED training set. As a result, it can be seen to perform poorly. The other aggregation functions are L_p aggregation with p ranging from 1 (equivalent to mean pooling) to 100 (close to max pooling), and max pooling. The proposed L_p aggregation turns out to provide a good tradeoff between mean and max pooling and allows us to discuss the impact of the aggregation function depending on the clip duration. Figures 7a and 7b show that, when $d_{\text{clip}} \lesssim 0.5$ s, aggregation has little or no impact, which is natural since $D_c \simeq 1$ for most clips. Also, $d_{\text{clip}} = 0.2$ s appears to be too short to correctly recognize the sound events of interest. In Fig. 7c corresponding to 1 s clips, we can observe the same behavior, except for clips containing sound events from 0 to 0.5 s ($D_c < 0.5$ with an average of $D_c = 0.32$) for which a low value of p degrades the performance. When the clip duration becomes longer, the value of p starts to matter even more, with short sound events requiring a larger value of p due to their low event density. From all the plots, we can see that using $p = 5$ or $p = 10$ is a good alternative to the maximum.

B. Impact of segmented data on training

The ideal scenario is a scenario whithout weakly labeled data (where $D_c = 1$). To achieve this, a natural way would be to segment the data in order to have only the sound event of interest in each clip. However, in this scenario, we would need to train a model with inputs having different number of time frames. Technically, this makes it hard to create a batch of tensors with different sizes. An alternative scenario is to have a batch size of 1 but this poses a problem with batch normalisation (to be accumulated). Training without batch normalisation could be an option but the performance is greatly decreased, this is presented in the first row of table I. The closest scenario to the ideal case is to apply a mask on the input and on the output (predictions). Analysing the masking of the input and the predictions separately could help to understand which part of the model is more impacted by weak labels. Masking can also be applied at training, at test or both.

In table I we present the results about the impact of batch normalisation and segmentation. In this table, masking refers to masking the input and the predictions (Masking (I/O)). We notice that using segmented data and accumulated gradients has a negative impact compared to using a masking scenario with batch normalisation. The difference in performance between the scenario using masks with and without batch normarlisation and the one without batch is surprising because it means that for our model, batch normalisation is very important. However, since our model is a mean teacher with a rampup on the learning rate, the convergence duration of training is important that could explain this difference in performance. From these experiments, we conclude that we can use masking of the input and output with batch normalization as our scenario without weak labels.



(a) (b) (c)

Fig. 6: F1-score achieved by the pretrained baseline system on the WAA test set for classical test-time aggregation functions, as a function of the test clip duration. Dashed lines indicate that $D_c < 1$ for all test clips.

	No batch normalization	Batch normalization
Masking (I/O)	54.9 ± 1.2	70.2 ± 1.2
Segmented	40.1 ± 1.4	X

TABLE I: Impact of segmentation and batch normalisation.

	Masking (I/O)	10 s clips
Clip-level loss (mean aggregation)	70.2 ± 1.2	62.2 ± 1.3
Frame-level loss	66.0 ± 1.1	56.5 ± 1.5

TABLE II: Impact of clip-level and frame-level loss for a scenario where $D_c = 1$ and one where $D_c \leq 1$. Training done with 5k generated clips.

			F1-score
Mask	input + output (ideal)	Training	33.8 ± 0.9
		Test	69.8 ± 1.2
		Training + Test	70.2 ± 1.2
	Output only	Training	44.5 ± 1.4
		Test	67.6 ± 1.1
		Training + Test	70.2 ± 1.0

TABLE III: Experiments varying the way the strong annotations are used to mask the data

Since we are using strong labels to segment the clips in the ideal scenario, another way to define the ideal scenario could be to use the strong labels to compute a frame-level loss at training time. Table II presents the results of the impact of the segmentation and the use of a clip-level or frame-level loss at training time.We experiment using frame-level and clip-level loss on clips masked on the input and the output and on 10 s clips to see the impact of the aggregation at training time. Using a clip-level loss means adding an aggregation function to aggregate the frame-level predictions, in this case we use the mean function. Figure 8 illustrates the different scenarios using segmentation, masking, frame-level and clip-level loss. Results show that using this aggregation function at training time is beneficial on both masked clips or 10 s clips. This result can be explained by the noise introduced in the loss. While using a clip-level loss, we try to optimize the loss after aggregation, meaning we do not focus on a specific frame. While using a frame-level loss, we want to minimize the loss for every frame. This makes it harder for the model which needs to differentiate every single frame even if some sound events from different classes can have very similar frames. The difference in performance between clip-level loss and frame-level clips is more important for 10 s clips than for masked clips. This shows that not only the optimisation problem is harder but the bias introduced by surrounding noise is increased. From these experiments we conclude that we can define our *ideal* scenario as training a model with clips masked on the input and the predictions while using a clip-level loss.

1) Masking and ideal scenario at training time vs test time: In these experiments, we are interested in applying the ideal scenario or the scenario masking only the prediction at training or at test time. In Table III we present the different ways of applying masking (input and output mask) at training and test time and how we apply masking on the output only at training or test time. We notice that the best scenario is to mask the input and the output at training and test time. Unexpectidely, masking only the output at training and test time performs as well as having an ideal scenario when we use weak labels during training. This means that the bias introduced by the background noise in the input of the surrounding event does not seem to be a problem. It is important to also notice that taking one of these scenarios (output masking only or ideal) at training time while doing a different scenario at test time is degrading extensively the performances. These results were expected, even if usually not shown quantitatively in papers. This proves that for an application point of view, the most important is to have well segmented data at test time. Indeed, if the data are clean or segmented at training time the model cannot recognize a sound event surrounded by noise at test time. From an application point of view, this means that if

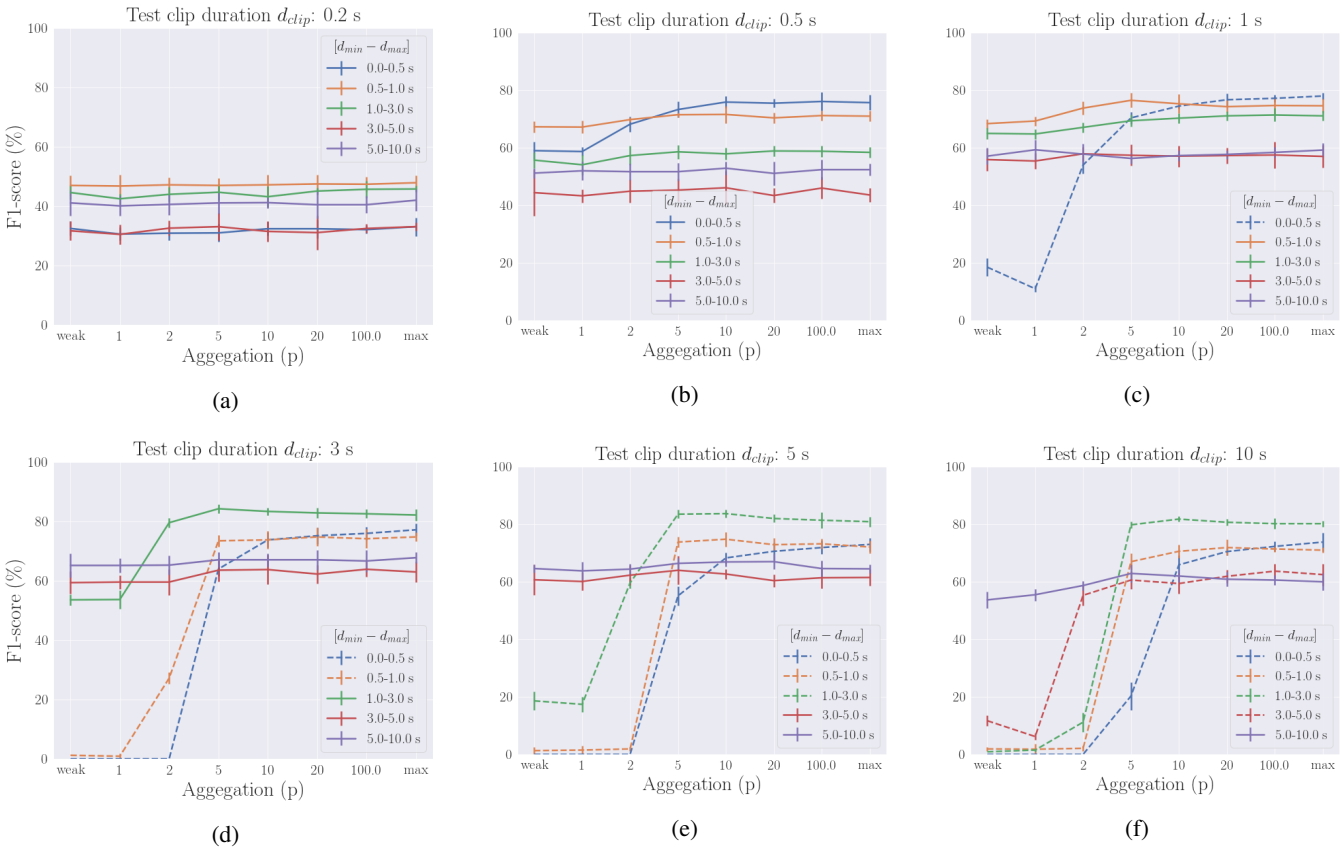


Fig. 7: F1-score achieved by the pretrained baseline system on the WAA test set for six different test clip durations, as a function of the event duration and the test-time aggregation function (weakly-trained attention pooling, L_p aggregation with $p = 1$ to 100, or max pooling). Dashed lines indicate that $D_c < 1$ for all test clips.

someone wants to train a model on FSD50k [3] for example (which contains pretty clean recordings in terms of weak labels), it will probably be hard at test time to recognize sound events in long recordings with few events (e.g. continuous recordings with few events which are common in real life).

The difference between masking only the output and the ideal scenario is the amount of data that has been seen at training time by the model. In the ideal scenario, the model only takes decision based on the sound event contained in the clips. In the case of output masking only, the model is presented sound events surrounded by noise and the loss is computed only on the part containing the sound events. This means that predictions were still computed containing the information of the surrounded noise present in the temporal context of the CRNN. This information used during training appears to be beneficial at test time on clips containing noise surrounding sound events as stated in lines "Train" of Table III. When applying some masking only at test time, it is more beneficial to use the ideal scenario than the output masking only which tends to indicate that segmentation is very important at test time no matter the training method used. The ideal scenario at training and test time have been used to define the number of training data generated. We have identified that above 1000 clips generated the training is almost similar even if more training data seems to make more stable predictions. From these experiments, we decided to generate and use 5000 training clips for the experiments of this paper.

	10 seconds		ideal	
p during training	F1-score (Validation)	F1-score (Test)	F1-score (Validation)	F1-score (Test)
1	82,5 ± 1,0	62,2 ± 1,3	80,8 ± 0,9	70,2 ± 1,2
2	80,8 ± 1,1	61,8 ± 1,2	80,5 ± 0,9	71,1 ± 1,0
5	80,3 ± 0,9	61,4 ± 1,4	80,1 ± 1,3	73,8 ± 1,3

TABLE IV: Results depending on the aggregation used during training on 10 s clips (neither segmented or masked) and the ideal scenario.

2) Aggregation during training: In Table IV, we present experiments where we vary p of the L_p aggregation applied during training. We show performances on the validation set and the test set for the model trained on 10 s clips (not segmented nor masked) or using the ideal scenario and vary p. In Table IV we can see that the aggregation during training does not have a significant impact. We also have to note that the best performance at validation is not the best performance on the test set. Training a model with $p > 5$ does not work because it introduces Nan in the loss due to numerical instability. In the following experiments of this paper, we are taking the best value at validation which means we are training

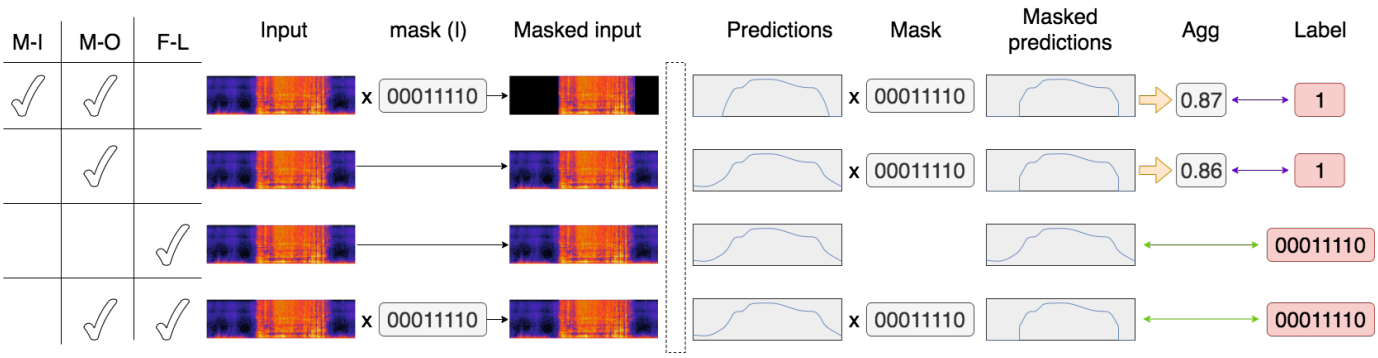


Fig. 8: Explanation of the scenarios when segmented or masking the input or masking the predictions. M-I stands for masked input, M-O stands for masked predictions, F-L stands for frame-level loss and Agg stands for aggregation. The dashed rectangle represents the model. The yellow arrows means the aggregation uses only the frames within the mask, the purple double arrows represents the clip-level loss and the green double arrows represent the frame-level loss.

the model with $p = 1$. The performance on the 10 s clips on the test set seems to indicate that either the segmentation is well made or it indicates that we predict many FP frames of the sound event class present in the class.

C. Impact of the clip duration

In these experiments, we want to know the impact of the the clips duration on the classifier. To understand what happens when dealing with weak labels, we are presenting multiple conditions. First, we are varying the duration of the training clips and the test clips while not truncating the events to understand how the model is impacted by the duration of the clips. We have previously seen in V-B2 that the model generalises better when we use the L_p aggregation and $p = 1$ during training so we train a model using this configuration. At test, we use $p = 5$ since in V-A we showed that this value is close enough to the max and still not too prone to false positives.

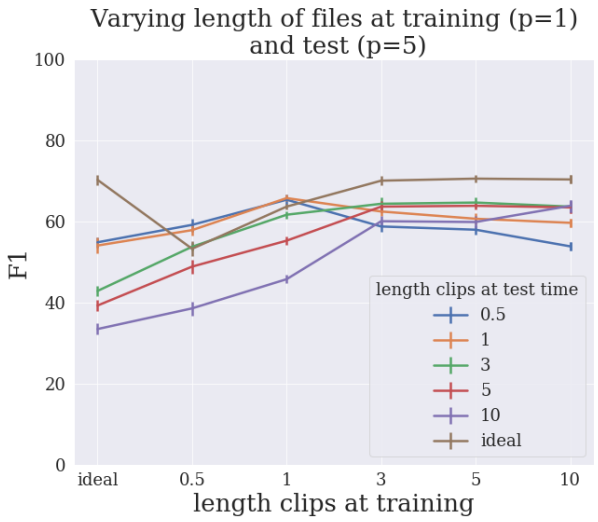
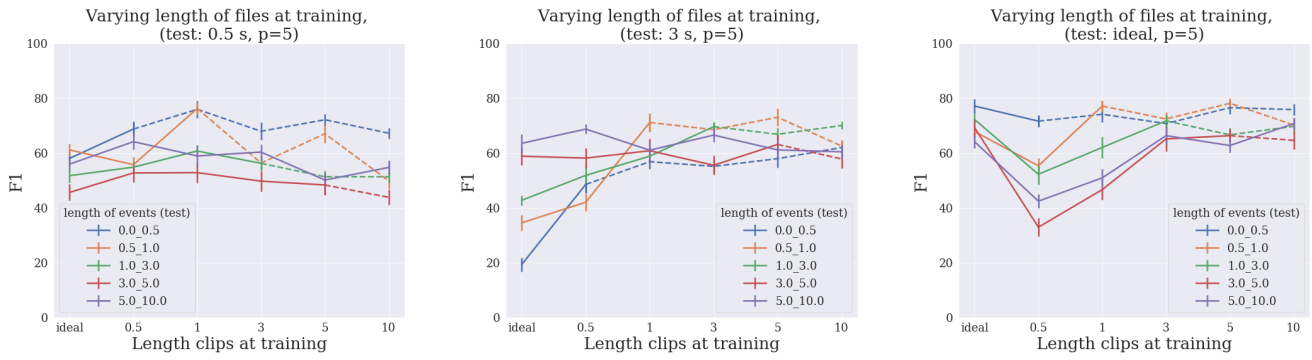


Fig. 9: F1-score depending of the duration of clips at training time and test time. Experiments made on WAA.

In Figure 9, we present the results when varying the duration of the clips at training time and test time while using WAA dataset. We remind that WAA has non truncated events, so the amount of background noise (weak labels) is varying when increasing the duration of the clip when a sound event is shorter than the duration of the clip chosen (D_c not constant between classes), but at the same time, while we increase the clip duration, we increase the amount of information about long sound event available to the model (sound events longer than the sound clip). In Figure 9, the first point on x-axis corresponds to the ideal scenario at training time. The performances are the lowest except when testing in the ideal scenario as well. This indicates that training a model on a well segmented dataset does not generalize well to test conidtions with surrounding background noise or where we see only a portion of a long event. On this figure only, it is difficult to choose one of these conclusions. This will be analysed further in the next experiments. In this figure, we can also notice that having close duration of files at training and test is beneficial. Finally, we can see an overall inflation point around 3 s which is a result that was identified by Salamon et al [43]. They showed that 4 s of an event is long enough to recognize it, that is why they designed UrbanSound8k dataset with 4 s segments. However, it has to be noted that when we train on very short segments of events (0.5 s), the results are degrading for test clips longer than 1 s.

To have a better understanding of the impact of weak labels during training and test, we focus on clips durations of 0.5 s, 3 s and a clip duration equal to the sound event (ideal), this is presented in Figure 10. We present the performance depending on the duration of sound events for test clips of a fix duration to study the lower performance for short clips duration at training was coming from noise around sound events (short sound events) or from truncated sound events (long sound events).

If we first focus in the Figure 10a, we can see that when we test a model on short clips, the duration of the training clip does not have an important impact. This is not the case when testing on longer clips (fig. 10b10c). We can see that if we train on short clips, and test on longer clips the model is



(a) (b) (c)

Fig. 10: F1-score depending of the duration of clips at training time and duration of events. Each subfigure represents a duration of file at test time (a line in Fig. 9), Dashed lines correspond to events being only weak labels at training.

failing to recognize the sound events but this is also affected by the event duration. If we focus on Figure 10b, we can see that short events ($<$ 1 s) are very hard to recognize in a 3 s clips if we have trained our model with only sound events (ideal) or 0.5 s clips which means that having background noise at test ($D_c << 1$) when it was not the case at training ($D_c \simeq 1$) is degrading performances. When training clips are longer than 1 s during training, the performances tend not to vary anymore until 10 s clips duration at training. Finally, Figure 10c indicates that if we test on an ideal scenario, having an ideal scenario at training is the best. Having weak labels (dashed lines) at training tend not to be a problem when testing on segmented sound events but seeing the complete sound event at training is very important, this is represented by all the curves improving before reaching the point of clip duration during training being longer than the sound event duration. In these 3 figures we can compare the position of the curves relative to each other. If we focus on the two first points of each figures representing a clip duration at training being segmented or 0.5 s. We notice that short events ($< 1s$) benefit from testing with 0.5 s (fig. 10a) or segmented (fig. 10c) clips. However, when testing with 3 s clips we can notice that the curves order is changed meaning it is hard to recognize short events and a lot simpler to recognize long events. Overall, our observations tend to indicate that when $D_c < 1$, the more D_c decreases during test while it remains high during training, the more we decrease the performances. Overall, when we have longer clips at test than training, the results tend to decrease, this can be explain by the model which is not suited for weaker labels at test time (even with a good aggregation method) than at training time and that the model has a hard time recognizing an event when it has been trained on part of events.

D. Impact of D_c

To understand more the impact of the amount of background noise (weak labels) at training time when background noise is present during test, we use TWAA dataset and vary the maximum duration of the event, so we have control over the maximum value of D_c which was not the case in previous experiments where some clips could always have $D_c = 1$. The

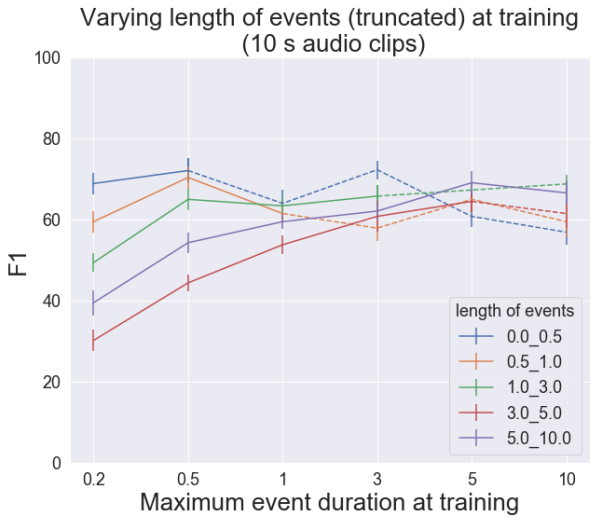


Fig. 11: F1-score depending of the duration of truncated events at training time and the duration of events at test time using TWAA. An L_p aggregation with $p = 1$ is used at training and $p = 5$ during test.

clips are 10 s long at training and test time, only the duration of target sound events is truncated at training to a maximum event duration allowed. Figure 11 presents the F1-score for varying maximum event duration at training time (the sound event can still be shorted than this duration) and present the results per duration of the original sound event (before truncation). In this figure, we can see the contrast between short events, shorter than 0.5 s and the other events. The short events benefit a lot from very short durations at training time. Longer events are very hard to recognize when truncated to 0.2 s, this is because there is not enough information to recognize them as we have seen in V-C. This also confirms the hypothesis that our previous work [31] using 0.2 s sound events was too short to make a good analysis of weak labels. Having better performances for short events when all sound events are truncated to 0.5 s ($D_c \simeq 0.05$) can identify that it is difficult to segment short events in long audio files, so biasing the

system at training by showing only short events makes them easier to recognize since the segmentation problem is the same for all events (we always expect 0.5 s sound events). It could as well be because the shorter the duration of the truncated sound events are, the longer the background noise is around a sound event introducing weak labels. The blue and yellow curves (events < 1 s) are mostly decreasing when increasing the maximum event duration indicating that showing longer events to the model makes it harder to recognize short events (segmentation, unbalanced problem). However, we can note that it is better to have background noise (weak labels) than truncating too much the sound events.

Finally, we propose to analyse the impact of weak labels independently of the duration of each event (D_c fixed) to validate that the impact we have observed in the previous experiments are not dependant of the sound event classes chosen in these experiments, their duration or the relative amount of noise introduced. We have previously seen that truncating all the sound events to a short time duration is not suitable. To overcome this problem while having a D_c fixed for all sound events, we work on segmented sound events and we choose to add background noise depending on the size of the event. In table V, we are presenting the results on the TWAA dataset for which we fix the max sound event duration to 3 seconds. In this experiment we only vary the density D_C of the event in the clip. It is the only experiment where it is possible to have $D_c = 0.1$ for all events independently of their duration.

		Test		
		Ideal ($D_c = 1$)	Biased $\times 2$ ($D_c = 0.5$)	Biased $\times 10$ ($D_c = 0.1$)
Train	Ideal ($D_c = 1$)	**66.2** $\pm$ 1.2	49.5 $\pm$ 1.9	33.2 $\pm$ 1.6
	Biased $\times 2$ ($D_c = 0.5$)	60.6 $\pm$ 1.2	55.7 $\pm$ 1.2	39.5 $\pm$ 1.1
	Biased $\times 10$ ($D_c = 0.1$)	60.1 $\pm$ 1.0	**61.5** $\pm$ 1.0	54.8 $\pm$ 0.9

TABLE V: Experiments made on a dataset with TWAA of 3 s.

In Table V, we present experiments for D_c varying from 0.1 to 1 both at training and test. Looking at the rows, we can confirm that having a smaller D_c at test time than at training time is degrading the performances at the exception of the last row where the model has been trained on clips with $D_c = 0.1$ for which having $D_c = 0.5$ or $D_c = 1$ have around similar performance. An explanation could be that the model needs to segment well the sound event to be able to classify it at training because only 10% of the frames have it. It makes it easier to extract a sound event in background at test time. If we focus on the columns, the scenario consists in having a rough idea of the density of events at test time and define how to annotate and use the training dataset to best suit our scenario. It can also be seen as how much budget we need to annotate training data depending on our scenario defined during inference. If we have a perfect segmentation of events at test time (by hand or automatically) corresponding to the first column of the table, the best case would be to also have segmented data at training time. We can notice that annotating weakly the training set is partly degrading performances but there is no difference between $D_c = 0.5$ and $D_c = 0.1$. This indicates that if we have the budget only for weak annotations, we can still target a resolution of ten times the envent duration which would be speeding the process. As an example for events which are usually around 3 s, there is no difference between an annotation of a 6 s clips or 30 s clips which is very interesting to know to be able to reduce the labeling effort. Interestingly, if we do not have access to a perfect segmentation at test time (or if we need a coarse time resolution), it is not worth labeling sound events with boundaries. Moreover, the best perdormance is obtained for $D_c = 0.1$ which could reduce greatly the labeling effort.

VI. Conclusion

In this article, we presented a study on the impact of weak labels on SET models. We defined the problem of weak labels, proposed a way to create a dataset which isolate the problem of weak labels and suggested a scenario where the events are segmented or not to present the different experiments with or without weak labels. We first showed that to deal with weak labels, the aggregation function at inference is crucial, and we proposed to use the L_p aggregation as an alternative to the max or $softmax$ function since it is both differentiable at training and flexible. We also shown that aggregation matters more at test than at training. Regarding training segmentation, we showed that it is better to have training clips that are too long than too short whatever the scenario at testing. Finally we gave insight on the granularity of annotation needed at training time depending on the scenario targeted at test time. It would be interesting to conduct this study to SED systems, and discuss the impact of weak labels to the sound events segmentation at test.

References

[1] T. Virtanen, M. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Springer, 2017.

[2] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[3] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50k: an open dataset of human-labeled sound events," *arXiv preprint arXiv:2010.00475*, 2020.

[4] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2018, pp. 9–13.

[5] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[6] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019, pp. 209–213.

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

[8] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 486–93.

[9] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[10] D. Little and B. Pardo, "Learning musical instruments from mixtures of audio with weak labels." in *International Society for Music Information Retrieval Conference (ISMIR)*, 2008, pp. 127–132.

[11] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly supervised learning of deep ConvNets for image classification, pointwise localization and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5957–5966.

[12] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao, "Learning from weak and noisy labels for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 486–500, 2017.

[13] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang, "CurriculumNet: Weakly supervised learning from large-scale web images," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 139–154.

[14] Q. Huang, A. Jansen, L. Zhang, D. P. W. Ellis, R. A. Saurous, and J. Anderson, "Large-scale weakly-supervised content embeddings for music recommendation and tagging," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 8364–8368.

[15] V. Morfi and D. Stowell, "Deep learning for audio event detection and tagging on low-resource datasets," *Applied Sciences*, vol. 8, no. 8, p. 1397, 2018.

[16] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, "Weakly labelled AudioSet tagging with attention neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1791–1802, 2019.

[17] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

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

[19] S. Adavanne, H. Fayek, and V. Tourbabin, "Sound event classification and detection with weakly labeled data," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019, pp. 15–19.

[20] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with CNN-transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[21] R. Serizel and N. Turpault, "Sound event detection from partially annotated data: Trends and challenges," in *IcETRAN conference*, 2019.

[22] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *ACM International Conference on Multimedia*, 2016, pp. 1038–1047.

[23] T.-W. Su, J.-Y. Liu, and Y.-H. Yang, "Weakly-supervised audio event detection using event-specific Gaussian filters and fully convolutional networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 791–795.

[24] B. Kim and S. Ghaffarzadegan, "Self-supervised attention model for weakly labeled audio event classification," in *European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[25] A. Shah, A. Kumar, A. G. Hauptmann, and B. Raj, "A closer look at weak label learning for audio events," *arXiv preprint arXiv:1804.09288*, 2018.

[26] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 21–25.

[27] V. Bisot, S. Essid, and G. Richard, "Overlapping sound event detection with supervised nonnegative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 31–35.

[28] E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley, "Detection of overlapping acoustic events using a temporally-constrained probabilistic model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6450–6454.

[29] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 121–125.

[30] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 171–175.

[31] N. Turpault, R. Serizel, and E. Vincent, "Limitations of weak labels for embedding and tagging," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 131–135.

[32] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, "Improving sound event detection in domestic environments using sound separation," *arXiv preprint arXiv:2007.03932*, 2020.

[33] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2018, pp. 188–192.

[34] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *ACM International Conference on Multimedia*, 2013, pp. 411–412.

[35] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017, pp. 32–36.

[36] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv:1510.08484v1*, 2015.

[37] N. Turpault and R. Serizel, "Training sound event detection on a heterogeneous dataset," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 200–204.

[38] L. JiaKai, "Mean teacher convolution system for DCASE 2018 Task 4," DCASE 2018 Challenge, Tech. Rep., 2018.

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

[40] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning convolution system for DCASE 2019 Task 4," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019, pp. 134–138.

[41] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 86–90.

[42] T. J. DiCiccio and B. Efron, "Bootstrap condence intervals," *Statistical Science*, vol. 11, pp. 189–228, 1996.

[43] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *ACM International Conference on Multimedia*, 2014, pp. 1041–1044.