Strong Labeling of Sound Events Using Crowdsourced Weak Labels and Annotator Competence Estimation

Irene Martín-Morató, *Member, IEEE*, and Annamaria Mesaros, *Member, IEEE*

***Abstract*—Crowdsourcing is a popular tool for collecting large amounts of annotated data, but the specific format of the strong labels necessary for sound event detection is not easily obtainable through crowdsourcing. In this work, we propose a novel annotation workflow that leverages the efficiency of crowdsourcing weak labels, and uses a high number of annotators to produce reliable and objective strong labels. The weak labels are collected in a highly redundant setup, to allow reconstruction of the temporal information. To obtain reliable labels, the annotators' competence is estimated using MACE (Multi-Annotator Competence Estimation) and incorporated into the strong labels estimation through weighing of individual opinions. We show that the proposed method produces consistently reliable strong annotations not only for synthetic audio mixtures, but also for audio recordings of real everyday environments. While only a maximum 80% coincidence with the complete and correct reference annotations was obtained for synthetic data, these results are explained by an extended study of how polyphony and SNR levels affect the identification rate of the sound events by the annotators. On real data, even though the estimated annotators' competence is significantly lower and the coincidence with reference labels is under 69%, the proposed majority opinion approach produces reliable aggregated strong labels in comparison with the more difficult task of crowdsourcing directly strong labels.**

***Index Terms*—Strong labels, Sound event detection, Crowdsourcing, Multi-annotator data.**

I. Introduction

ANNOTATED data is a key player in the development of machine learning methods. While advanced methods may be capable of learning from data without or with only partial annotations, evaluation of their performance does require annotated data. The degree of difficulty and effort necessary for producing annotated audio datasets varies depending on the task. Some tasks require classification of audio at a coarse temporal level, such as the general-purpose audio tagging of Freesound content [1] or AudioSet [2]. On the other hand, tasks like sound event detection (SED) [3] or sound event localization and detection (SELD) [4] require a fine temporal resolution output, to indicate the onset and offset of sound event instances.

The textbook case for training a SED system is based on strongly annotated data, in which textual labels, onsets and offsets are provided for the sound event instances [5]. Such annotation requires a significant effort and as a consequence strongly-labeled datasets are small in size, if they are real-life recordings. Synthetic strongly-labeled data can be easily created [6], [7], but often lacks the complexity and variability of real acoustic environments, which creates a mismatch for methods expected to be used in practical situations. On the other hand, weakly-annotated data that contains only textual labels to indicate the presence of different sound events requires less annotation effort and has become the predominant type of data in the field.

Research on SED and SELD is continuously developing, but the acute lack of strongly annotated datasets steers the approaches towards learning based on weak labels [8], [9] and semi-supervised methods [10]. There is also a large body of work that has produced powerful, highly-performing approaches that use semi-supervised methods, such as student-teacher learning paradigm, to compensate for the weak labels in learning [11], [12], [13]. For example training is possible using unlabeled training data together with smaller amounts of weakly-labeled data, and possibly strongly-labeled synthetic data, as proposed by Turpault et al., [11]. However, there is always a need for strongly-labeled data for evaluation, and this is often manually annotated.

The measured system performance is dependent on the quality of the evaluation data, since the reference annotations of the evaluation dataset define what is considered correctly and erroneously detected in the system output. It is therefore important that these reference annotations are reliable, in order for the measured performance to reflect reality. It is widely accepted that the manual annotations are highly subjective, which manifests in variability of textual labels (when annotators are required to provide them) [14] and inaccurate timestamps for the event instances [5]. Sound event detection is evaluated with respect to the temporal location of reference event instances [15], [16], which creates a strong dependence of the system performance on the quality of the annotations.

An alternative method to manual annotation is automatic content analysis with added human verification of the proposed labels, a method that has mostly been employed for weak

Manuscript received 18 February 2022; revised 19 August 2022 and 5 December 2022; accepted 27 December 2022. Date of publication 13 January 2023; date of current version 9 February 2023. This work was supported by Academy of Finland under Grant 332063, Teaching machines to listen. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wenwu Wang. *(Corresponding author: Irene Martín-Morató.)*

The authors are with the Computing Sciences, Tampere University, 33720 Tampere, Finland (e-mail: irene.martinmorato@tuni.fi; annamaria.mesaros@tuni.fi).

Digital Object Identifier 10.1109/TASLP.2022.3233468

labeling [2], [17]. For example the FSD50 k dataset labels were proposed based on the tags provided by users and then verified manually by expert and non-expert annotators [17].

Crowdsourcing offers a more efficient method for annotation of large amounts of data. Even though mostly used for weak labeling, attempts to collect strong labels using crowdsourcing exist [18], [19]. Cartwright et al. [18] employed the classical annotation approach, requiring the annotators to provide onset, offset, and a textual label to all event instances; the task was simplified by providing the annotators with a list of labels to choose from. In our previous work [19], the annotation was formulated as weak labeling of overlapping temporal segments, and the strong labels were reconstructed with a 1 s resolution; similar to [18], a preselected list of textual labels was provided, to simplify the annotation task.

One important factor in using crowdsourced data is the availability of multiple opinions, and the way they are aggregated. The aggregation of annotator opinions is typically based on simple strategies like majority vote (consensus) [20], [21]. Using multiple expert annotators is more common in medical imaging than audio, with different strategies employed for aggregating the expert annotator opinions. Simple aggregation methods include, similar to the audio studies, intersection and union [22]; more complex strategies estimate an optimal ground truth using expectation-maximization as done in STAPLE [23] or maximizing the joint agreement between annotators [24]. A review of these approaches indicates that the method used to estimate the ground truth has a significant effect on the evaluated performance of the system, with STAPLE causing underestimation of performance when only few annotations are available, and consensus overestimating it [25]. In our previous work, [26] we proposed an extended version of MACE - Multi-Annotator Competence Estimation [27] to predict the "true" labels for multi-labeled audio data using models of the annotators' competence. The method weighs the annotator opinions based on their competence, in contrast to majority voting which trusts and weighs all annotators equally. This approach was incorporated in the strong label estimation proposed in [19], and shown to produce better estimates than the majority vote procedure.

In this paper, we present two key contributions to the problem of strongly-labeling audio data for SED. First, we propose a method for estimating strong labels, using crowdsourcing of weak labels and a processing stage to reconstruct the temporal information. While the method has been introduced in our previous work [19], we now extensively test its effectiveness on real-life recordings, to understand its applicability in practical situations. Second, we propose a novel aggregation method that we call "majority opinion," applied directly to the weak labels as provided by the annotators. This approach operates on the raw data obtained from annotators instead of estimating the tags for each annotated segment, as done in [19], and uses annotator competence to weigh the individual opinions. All the previous work on crowdsourcing strong labels has been done on synthetic data, and the methods have not been tested on real-life audio. In this work, we investigate the crowdsourced annotation outcome on two known real-life SED datasets, and also compare the outcome of our proposed method with the approach of Cartwright et al. [18]. Finally, we investigate the effect of the reference annotation generation method on the evaluated performance of a SED system, to understand what proportion of the measured errors are due to incomplete reference data.

The remainder of this paper is organized as follows: Section II presents the related work in more detail, and the novel elements of the proposed approach; Section III presents the crowdsourcing annotation procedure, annotator competence estimation and the proposed strong label estimation method. Section IV introduces the datasets and the annotator competence analysis. The experimental results for the labels estimation are presented in Section V, which includes analysis of the resulting weak labels, strong label estimation, the comparison to direct strong annotation and discusses the sound event detection experiments using estimated labels. Finally, Section VI presents conclusions and future work.

II. Related Work

Manual annotation is the most obvious approach to obtaining strong labels. Because the annotation task is difficult and time-consuming, most datasets containing strong labels are very small, for example TUT Sound events 2016 [28] and TUT Sound events 2017 [3] datasets contain only about 2 h of data each, in files of length 3-5 minutes. Their reference annotation was produced by two annotators that listened to the audio and could inspect the spectrogram, and had to provide a textual label composed of noun and verb (object and action), and onset and offset for all audible sound event instances [28]. The obtained set of labels was later manually processed to merge some classes, and the most frequent ones were selected and provided with the data. Similarly, the MAVD-traffic datased for SED in Urban environments [29] was manually annotated using the ELAN software, displaying the audio waveform, the video, and the spectrogram of the audio signal. The dataset consists of 4 h of data in files of approximately 5 minutes, and contains 21 classes.

The largest strongly-labeled dataset to date is a portion of AudioSet consisting of around 120 K files that were manually annotated. The annotation process consisted of several steps, in which a first-pass labeling was reviewed by a different annotator who could adjust the temporal boundaries. The verification/adjustment step was repeated, but even with 5 stages this process rarely converged to consensus [30], which implies that the annotators did not seem to agree on the boundaries. While very large in terms of classes and size, the audio files in this dataset have a length of only 10 s, which makes it very different from the aforementioned ones which are more representative of the overlaps and sequentiality of sound events in everyday environments.

As mentioned earlier, crowdsourcing is a very effective way to collect or curate data because it provides immediate access to a large number of nonexpert annotators. For example in FSD50 k dataset, selected clips were automatically assigned labels based on the tags provided by the users, mapped onto the AudioSet ontology [17]. A specifically created tool, the Freesound Annotator (FSA), was then used to curate the data: volunteer users were asked to validate that a certain sound is

present in the audio or not. The sound classes were divided according to an estimated level of difficulty and only the easy and medium difficulty classes were validated publicly through FSA. The classes considered difficult to annotate were validated by a pool of hired raters. Crowdsourcing was used to collect annotations for many notable image datasets such as ImageNet and Microsoft COCO, and a number of recent audio datasets, for example Clotho [31] and Open-MIC [20].

When multiple opinions are available for one annotated item, they are commonly aggregated through a majority vote. As a consequence, the expertise and diligence of the annotators in the annotation task influences the result. Our previous work addressed the problem of analysing annotator behavior for generating a reliable reference annotation based on their aggregated opinions [26]. A pool of 133 annotators was used to annotate 3930 audio recordings, providing 3-5 opinions per file. Aggregation based on annotator competence estimation was found to provide the best set of labels, evaluated using annotator agreement metrics. A second experiment using synthetic data, for which the ground truth was available, confirmed that the competence-based aggregation approach is superior to majority vote, validating the connection between annotator competence and reliability of the aggregated annotation [19].

Crowdsourcing of strong labels has been studied by Cartwright et al. in a controlled experiment that aimed to investigate the effect of visualizations and complexity on the crowdsourced annotations [18]. The study used 3000 synthesized soundscapes which were 10 s in duration, each containing up to 9 sound events, and a maximum polyphony of 4. The aggregated annotation was obtained by converting the annotations to a frame-based time-series representation using a frame size of 100 ms, and majority vote: a time frame was marked as active if at least half of the participants marked it as active. The study observed a sharp increase in quality of the estimated aggregated annotation for the first 5 annotators, followed by more subtle improvements as the number of annotators considered in the aggregation increased.

Our previous work introduced an alternative to the crowdsourcing of strong labels by breaking the annotation task into weak labeling of consecutive audio segments, followed by postprocessing to recover the temporal connection between the labeled events [19]. Aggregation based on annotator competence was also incorporated into the strong label estimation process. The study was based on 20 synthetic soundscapes containing a maximum number of six sound event classes and a maximum polyphony of 2. The comparison of the resulting estimated strong annotation with the reference generated with the data showed that the proposed method successfully reconstructs about 80% of the ground truth information.

In this work, we continue exploring the method in [19] and propose a novel aggregation method that uses directly the segment labels as provided by the annotators, instead of estimating the true labels with MACE. The aggregation starts from the raw data and takes into account annotator competence directly in the estimation of strong labels. For the synthetic data, we perform additional analyses of the labels with respect to signal-to-noise ratio and polyphony of sounds in the audio. Most importantly, we investigate the proposed method's applicability to real-life data, which is much more complex in terms of acoustic content than the synthetically generated one. In addition, we compare the outcome of the proposed method with the strong annotation approach from [18], to understand the tradeoff between cost-effectiveness and labeling process outcome.

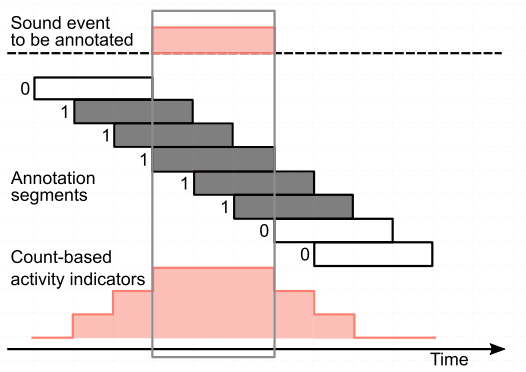

Fig. 1. Estimating event activity from overlapping weakly-labeled segments.

III. Crowdsourcing Annotations

A simple and well-defined annotation task is the key for successful and consistent behavior of the annotators. The typical annotation process for creating strong labels requires the annotator to listen to an audio excerpt, recognize the target sound events, and annotate their presence by marking the temporal boundaries for each instance of the target classes. Oftentimes this requires repeatedly listening to the audio example to annotate sounds that overlap, or to make corrections to the already marked temporal boundaries. Selection of the temporal boundaries is subjective, and different annotators tend to disagree on their exact location [30], which indicates that the strong labeling annotation task is a difficult one.

A. Annotation Procedure

We propose a procedure that simplifies the annotation task by dividing it into unit tasks that require only weak labeling. The files to be annotated are segmented into short, overlapping segments, which are to be annotated with weak labels by indicating binary activity of sound events within the entire segment. The list of target sound classes is selected in advance and presented to the annotator, making the labeling task as simple as possible. The proposed method is illustrated in Fig. 1. A sliding "annotation window" goes over the length of the audio file, with a high rate of overlap between consecutive segments covered by this window. The temporal sequence of these annotated segments provides the temporal activity of the sounds within the original long file, by aggregating activity indicators at each time step. If all weak annotations are correct, therefore all annotators have indicated correctly that a sound is active or not, the event boundaries correspond to the boundaries of the maximum-valued region in the count-based activity indicators.

To facilitate accurate recognition of sound sources in the audio segments provided to annotator, we choose a segment

length of 10 s. This length is motivated by studies that examined the recognition by humans of a list of 42 different sounds, and concluded that listeners need a maximum of 6.8 s to accurately identify the sounds of the studied categories [32]. A hop of one second between the segments will provide a one second resolution in the temporal reconstruction of the events activity, which is in line with the diffuse labels created in [30] and the segment length used in the evaluation of most SED systems [3]. We formulate the annotation task as a single-pass multi-label annotation, as done in [21] and [26]. As a consequence of this procedure, the presence of a sound is explicitly indicated by selecting the corresponding label, while the absence is implicit by the label not being selected.

B. Annotator Competence and Ground Truth Estimation

When working with non-expert annotators, it is important to be able to trust their answers. We employ MACE [27] to estimate how reliable these annotators are. The method allows identification of trustworthy annotators and provides a prediction for the ground truth based on aggregation of the annotators opinions. MACE does not necessarily require that all annotators provide answers on all data, but requires at least that a large pool of annotators annotate partially the same data, in order to learn from redundant annotations.

The model, as originally introduced by Hovy et al. [27], considers that annotator j produces label A_{ij} on instance i. The annotated label depends on the true label T_i, and whether annotator j is spamming (spamming means that the annotator is selecting the answer at random). Annotator behavior is modeled by binary variable S_{ij} drawn from a Bernoulli distribution with parameter $(1-\theta_j)$. The behavior assumes that when an annotator is not spamming on instance i ($S_{ij}=0$), the annotation A_{ij} corresponds to the true label. When the annotator is spamming, $S_{ij}=1$, A_{ij} is sampled from a multinomial distribution with parameter vector ξ_j. The annotations A_{ij} are observed, the true labels T_i and the spamming indicators S_{ij} are unobserved. The model parameter θ_j specifies the probability of trustworthiness for annotator j, while ξ_j determines the spamming behavior of annotator j.

The model parameters are estimated using the expectation maximization algorithm, to maximize the probability of the observed data:

$$P(\mathbf{A};\theta,\xi)=\sum_{T,S}\left[\prod_{i=1}^{N}P(T_i)\prod_{j=1}^{M}P(S_{ij};\theta_j)P(A_{ij}|S_{ij},T_i;\xi_i)\right] \quad (1)$$

where $\mathbf{A}$ is the matrix of annotations, $\mathbf{S}$ is the matrix of competence indicators, and $\mathbf{T}$ is the vector of true labels. Here N refers to the number of instances i that are annotated, and M to the number of annotators j that provide an opinion for instance i. The method was shown to produce predicted labels very accurately in comparison with ground truth data on a few tasks. At the same time, the model's θ_j was shown to correlate strongly with annotator proficiency [27].

Because MACE was originally defined for single-labeled items, we extend the representation of our multi-labeled data such that each file is assigned a set of binary *yes*/*no* labels, each corresponding to one target sound class. This implies that each (file, sound label) pair is considered an independently annotated item, equivalent to a multiple-pass binary annotation [21]. The difference is that in a multi-pass binary annotation both the present (*yes*) and absent (*no*) labels would be explicitly provided by the annotator, while in the single-pass multi-label annotation such as our task, the absence is implicit. We consider that the tagging task is easy enough to allow changing the data representation without introducing significant errors. We therefore explicitly represent as absent the items that were not explicitly marked as present by the annotators.

TABLE I
ANNOTATION MATRIX EXAMPLE WITH EXPLICIT/IMPLICIT ANNOTATIONS PRODUCED BY m ANNOTATORS

items (file, label)	1	2	3	4	..	m
scape-00, car-horn	1	1	0	1	...	-
scape-00, children-voices	0	0	0	1	...	-
scape-00, dog-bark	-	-	-	-	...	1
...	...	...	...	...	...	
scape-01, car-horn	1	1	1	-	...	-

The annotations are represented as a matrix containing the answers of all annotators per file and per label, as illustrated in Table I. Each row refers to a (file, sound label) item, and each column represents the answer of one annotator in the format $[0, 1, -]$, marking the presence (1, explicit) or absence (0, implicit) of this label within the audio file; "$-$" indicates that this file was not assigned to this specific annotator.

Using this representation, we estimate the annotators' competence and predict the aggregated weak labels using MACE. It is important to note that MACE does not discard annotators, but weighs their opinion based on their competence, which results in a different procedure than majority voting which trusts and weighs all annotators equally. In some experiments, we also eliminate the most unreliable annotators based on their estimated competence, to study if relying on a smaller pool of better annotators is more advantageous than using a higher number of annotators wherein low-competence annotators are also present.

C. Strong Label Estimation Based on Majority Opinion

The illustration in Fig. 1 takes into account one weak label for each segment and reconstructs the temporal activity pattern of a sound event as a count-based activity at hop-size resolution. Having multiple annotators per segment allows for estimation of this weak label using MACE. The count-based activity indicators are then binarized to obtain the maximum-valued regions that corresponds to the estimated temporal boundaries of the sound event instances. In [19], a threshold of 80% was used instead of the maximum, in order to allow for possible incorrect answers from the annotators.

We propose a novel method of estimating the strong labels, in which we consider directly the labels provided by the individual annotators. This way, the method takes into account the fine-grained differences in annotators' opinions instead of

transforming them first into an estimated weak label per segment. Given the procedure explained above, we consider all annotator opinions in each hop-size segment t and aggregate them such that the vote of each annotator (sound event active or not active) is weighed by his/her estimated competence. The individual competence associated to each annotator is the model parameter θ_j estimated using (1), in other words the probability of trustworthiness for annotator j. We calculate the activity indicators using the following expression:

$$a_t = \frac{\sum_{j=1}^{M} (\theta_j \cdot v_j)}{\sum_{j=1}^{M} \theta_j} \tag{2}$$

where a_t is the activity level a for one class in segment t, M is the number of available opinions for that segment, θ_j is the competence of annotator j, and v_j indicates the annotator's opinion, being 1 for the presence and 0 for the absence of the label. The estimation is done independently for each class.

This formulation is a generalization of the majority vote: if we consider all annotators as equally and perfectly competent, their competence level θ_j is 1. With the opinions being 0 or 1, normalizing the sum of opinions by the sum of the annotators' competence results in a value higher than 0.5 only when over half of the annotators have indicated a sound as being active. If the annotator competence is not always 1, the resulting value is still a number between 0 and 1, but it can be higher than 0.5 when less than half of the annotators indicated a sound as active, given that these annotators are the most trustworthy ones. This is still a consensus-based aggregation, but instead of majority vote (over half the annotators voting 1) we are considering the *majority opinion*, i.e. enough weight brought by the trustworthiness of annotators.

IV. Datasets Annotation Task Setup

In the experiments we use both synthetic and real audio recordings. The synthetic data offers the possibility of performing a detailed analysis of how the polyphony and SNR levels of the sound events present in the soundscape affect the outcome of the annotation, and allows the comparison of the method outcome with the correct reference annotation that is generated at the same time with the audio mixtures. On the other hand, the real recordings are more complex than synthetic data due to the unrestricted and uncontrolled sounds distribution and overlap, and present a difficult task to the annotators. To the best of our knowledge, this is the first experiment to attempt crowdsourcing strong annotations for real recordings, and the detailed analysis of its outcome will allow us to understand how the estimation of the annotators and annotations reliability translates from the highly controlled and simplified synthetic case to a real-world situation.

A. Datasets

1) Synthetic Data: The synthetic dataset used in this study is MAESTRO Synthetic (Multi-Annotator Estimated STROng labels) [33], which was created using a slightly modified version of Scaper [34]. Soundscapes were generated by iteratively placing sound events at random intervals until the desired maximum polyphony of 2 is obtained. Intervals between two consecutive events were selected at random between 2 and 10 seconds. The sound event classes and sound instances were chosen uniformly, and mixed with a signal-to-noise ratio (SNR) randomly selected between 0 and 20 dB over a Brownian noise background. The mixing procedure did not allow two overlapping sounds of the same class.

The dataset contains the following classes: *car horn*, *children voices*, *dog bark*, *engine idling*, *siren*, and *street music*. The isolated sound event instances were extracted from the UrbanSound dataset [6] based on their temporal boundaries which were manually annotated by the dataset authors (*children playing* label from the UrbanSound dataset was renamed to *children voices* for the annotation task, as often the audio examples contained childrens' laughter). Only sounds marked as being in the foreground were used. The selection of target classes was based on the intention to mimic the content of the street scenes annotated in our previous study [26] and from the real-life TUT Sound Events 2016 and 2017 datasets. MAESTRO Synthetic dataset consists of 20 audio files, each having a length of 3 minutes. The reference annotation of this dataset is created at the same time with the audio mixtures. We consider this reference annotation as correct and complete, because of the way it is produced. Dataset statistics are presented in Table II.

2) Real-Life Data: The real life-recordings used in this study include a subset of the TUT Sound Events 2016 [28] and a subset of TUT Sound Events 2017 [35]. We use the *residential area* acoustic scene from TUT Sound Events 2016, and select six target classes: *bird singing*, *car*, *children*, *people speaking*, *people walking*, and *wind blowing* (i.e. we do not consider the *object banging* class of the dataset). From TUT Sound Events 2017 we use the recordings corresponding to the *city center* acoustic scene, with target classes *brakes squeaking*, *car*, *children*, *large vehicle*, *people speaking* and *people walking*. We will refer to the strong annotations produced by the described method as MAESTRO Real and publish them for further study.[1] The reference annotation for MAESTRO Real is the annotation provided with the original datasets, which was obtained through manual annotation performed by two expert annotators that each annotated half of the data [28]. While these manually annotated data cannot be considered correct and complete due to the complexity of the acoustic content, our purpose is to understand the differences between different methods to produce annotations, therefore we use these reference annotations to evaluate how the different crowdsourced versions coincide with expert opinions. We accept the fact that the expert annotations are also subjective, and analyze the effect of different annotation procedures on the produced labels and on the evaluation of SED systems. The statistics of the data are presented in Table II. The two acoustic scenes (*city center* and *residential area*) are treated separately in all our experiments.

[1] MAESTRO-Real, 10.5281/zenodo.7244360

TABLE II
SOUND EVENTS CLASS DISTRIBUTION IN MAESTRO SYNTHETIC AND MAESTRO REAL DATASETS, ACCORDING TO THE REFERENCE ANNOTATION

MAESTRO Synthetic		MAESTRO Real residential area		MAESTRO Real city center	
event class	count	event class	count	event class	count
car horn	41	bird singing	56	brakes squeaking	39
children voices	63	car	45	car	55
dog bark	76	children	17	children	13
engine idling	102	people speaking	19	large vehicle	30
siren	69	people walking	32	people speaking	48
street music	40	wind blowing	16	people walking	41

MAESTRO Synthetic reference annotation is generated together with the audio mixtures, MAESTRO Real reference annotation is provided with the original datasets, and was produced by human annotators.

B. Crowdsourcing Task Setup

As explained in Section III, the audio soundscapes were cut into 10 s segments with 1 s offsets. Each individual 10 s segment was considered as an independent annotation task, provided on Amazon Mechanical Turk as one HIT (Human Intelligence Task). In order to prevent the same worker annotating overlapping segments, the data was organized into batches containing segments located at least 15 seconds apart in the original audio. The batches were launched one at a time, and workers that already performed at least 50 hits in previous batch(es) were disqualified from working on the task. A payment of $0.10 was offered per HIT. Worker qualification was requested as at least 1000 completed HITs with average approval rating of at least 85%.

One HIT consisted of listening to the provided audio excerpt and indicating which sounds are present in it, from the given list of classes or "none of the above". The number of playbacks allowed was not limited. No visualization (e.g. spectrogram) was provided. Workers were instructed to complete the task using headphones, and in a quiet environment. Before the job, they were also provided short descriptions for every class, and four example audio excerpts that contained sounds from all target classes. Each 10 s segment was annotated by 5 workers. While MTurk requires reviewing the assignments in order to approve or reject the answers submitted by workers, we approved all assignments, irrespective of the quality of the answers, in order to study the annotator behavior.

C. Annotators Competence Analysis

Annotators' competence analysis performed with MACE is shown in Fig. 2. This analysis considers only the weak labels provided by the annotators to the 10 s segments, and the audio segments are considered as independently annotated items. The synthetic data was annotated by a pool of 680 workers, while the real data was annotated by 861 and 717 workers for the residential area and city center scenes, respectively. Each set consisted of approximately 20 thousand HITs.

Most annotators seem to have high competence for the synthetic data, with about one third of the annotators in the highest tier (competence 0.9 to 1.0). Competence of the annotators on the real data shows a completely different situation: the values are

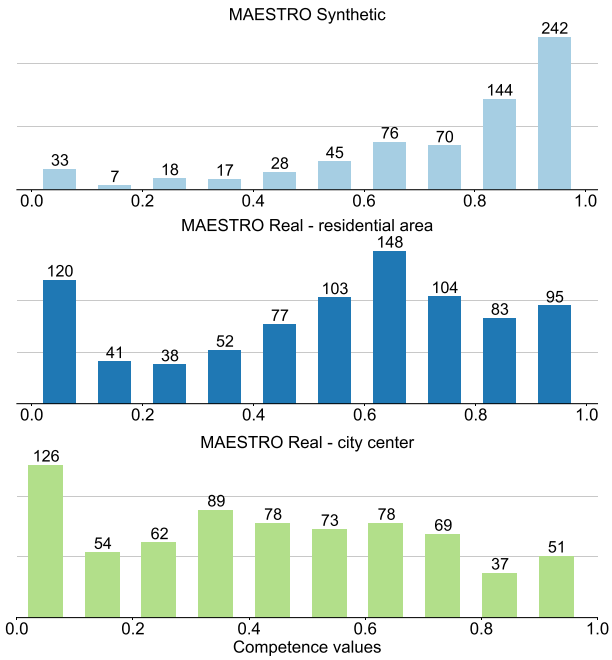

Fig. 2. Annotator competence estimated using MACE.

distributed over the entire range, and a high number of annotators have extremely low competence (17% for city center and 14% for residential area have an estimated competence of under 0.1). We did expect to see a deterioration of overall competence for the annotation of the real soundscapes, but such a pronounced difference was surprising. This in itself is a very good indicator of the task difficulty for a non-expert annotator.

It is important to note that the annotators of the real and synthetic soundscapes are different, and individual annotators were limited to maximum 50 HITs. The competence estimation is therefore applied to a large pool of annotators, and the result can be seen as an indicator of the task complexity. Previous works that studied annotation procedures all used synthetic data to draw their conclusions [18], [19], while the difficulty and subjectivity of annotating real data was always mentioned and accepted as true [5]. Fig. 2 shows histograms of the estimated annotators' competence on the different datasets. These results are the first that demonstrate in a quantifiable way that real data

TABLE III
STATISTICS OF THE THREE DATASETS AND THEIR CORRESPONDING KRIPPENDORFF'S ALPHA VALUES

Data	# Workers	# HITs	α_{all}	$\alpha_{\text{comp}>0.6}$
MAESTRO Synthetic	680	20520	0.56	0.73
MAESTRO Real-RA	861	20574	0.25	0.54
MAESTRO Real-CC	717	21264	0.20	0.54

RA and CC stand for Residential area and City center, respectively.

is much more difficult to annotate than synthetically generated one.

Inter-annotator agreement was calculated using Krippendorff's alpha, and is presented in Table III along with more details about the annotation task. In the table, α_{all} represents Krippendorff's alpha for the entire set of annotations. The values show how difficult it is for annotators to agree on annotation of the real data, compared to the synthetic data. Removing the less competent annotators increases the inter-annotator agreement: using annotators with competence higher than 0.6 results in a 30% relative increase in agreement for the synthetic data; for the real data, the relative increase is 116% and 170%, respectively, to $\alpha_{\text{comp}>0.6}$ of 0.54.

When removing annotators based on their competence values ($\alpha_{\text{comp}} > 0.6$), the number of annotators left for the agreement calculation is considerably reduced, being 532, 430 and 228 workers, respectively. As a consequence, the number of HITs that the agreement is calculated on is reduced to 20514, 19164 and 16164, respectively. The most affected subset is the city center data: 4770 of the 21264 annotated items are left without annotations, because 489 annotators have an estimated competence below 0.6.

V. EXPERIMENTAL RESULTS

The weak and strong labels estimation methods are analyzed by comparing their output with the reference annotation. We evaluate the quality of the resulting weak labels using precision, recall, and F1, and the strong labels using the most common metrics from SED.

A. Weak Label Estimation

Considering the annotated segments individually, the annotation process output is evaluated by comparing the audio tags with the reference tags for each segment. The reference tags per segment were generated based on the reference strong labels by assigning a label to a segment if the sound is active at any time within that segment.

The multiple annotations were aggregated for each segment using three different methods: union, majority vote and MACE. Union assigns a label to an item if at least one of the annotators has assigned it to that item; majority vote assigns a label to an item if most annotators have assigned it (in this case at least 3 of the 5 annotators). MACE uses the estimated competence of the annotators to predict the labels for each item, as explained in Section III. The comparison to the reference labels is done using F1, precision and recall. The results are presented in Table IV.

TABLE IV
WEAK LABEL ESTIMATION COMPARED TO THE REFERENCE ANNOTATION USING THREE DIFFERENT AGGREGATION METHODS

Dataset	Aggregation method	F1 [%]	P [%]	R [%]
MAESTRO Synthetic	Union	78.7	70.2	89.4
	Majority vote	68.8	98.2	52.9
	MACE	86.3	97.5	77.4
MAESTRO Real Residential area	Union	29.7	33.8	26.5
	Majority vote	29.8	51.5	21.0
	MACE	32.1	47.5	24.3
MAESTRO Real City center	Union	42.9	58.8	33.8
	Majority vote	32.8	75.7	20.9
	MACE	42.6	72.2	30.3

For the synthetic data, the best F1 is obtained using MACE: 86%, with 97% precision and 77% recall. Recall values show that many sounds are not annotated: with the majority vote, only slightly over half of the tags are found, while taking into account all opinions through union aggregation brings recall close to 90%. MACE produces a good compromise between a high precision and a good recall.

Looking at the real data, the metrics behavior is very similar, although the actual values are much lower: aggregation through union produces the best recall, while majority vote produces the best precision, and MACE raises the recall level while slightly lowering precision. It is worth noting that, for the real audio recordings, the reference annotations should not be considered as being absolutely correct, since even though they were produced by expert annotators, they were produced by a single person for each file. It is however discouraging that the aggregated opinion of multiple annotators overlaps only so little with the original annotator's opinion. The results nevertheless show that MACE is the best aggregation method for both types of data, synthetic and real. For this reason, we will focus on MACE-based aggregation approaches for the remainder of the experiments.

1) Polyphony Analysis: We analyze the influence of the polyphony on the aggregated weak labels using the synthetic data, for which such details are available. The synthetic data has been designed to have maximum two overlapping sound events at a given time. However, a 10 s segment may have more than two labels assigned, depending on its content. We use the term "polyphony" broadly to mean the number of events present in one 10 s segment, not necessarily all overlapping in time. We also calculated the average gini-polyphony introduced in [18] and defined based on the sound event polyphony at 100 ms time intervals throughout the soundscape. Interpreted as a measure of soundscape complexity, with zero representing maximal equality (low soundscape complexity) and one representing maximal inequality (high soundscape complexity) [18], the average gini-polyphony of the data is 0.74, which shows that the complexity of the soundscapes is generally high.

Table V presents the F1, precision and recall for segments of different polyphony, along with the number of these segments in the data. According to the reference, there are no segments containing no events, 95% of the segments contain two or three events, while only 4 segments have a polyphony of 5. On the other hand, in the MACE output almost 90% of the segments

TABLE V
MAESTRO SYNTHETIC LABEL ESTIMATION FOR DIFFERENT LEVEL OF POLYPHONY

PL	N_s GT	N_s MACE	F1 [%]	P [%]	R[%]
1	10	1000	95.2	90.9	100.0
2	2202	2051	87.9	96.8	80.6
3	1041	351	84.1	98.6	73.3
4	163	4	83.6	98.4	72.7
5	4	0	72.7	100.0	57.1

Pl stands for Polyphony level, determined based on the reference and N_s for the number of segments.

TABLE VI
MAESTRO SYNTHETIC AUDIO TAGGING METRICS FOR DIFFERENT RANGE SNR

SNR interval [dB]	N_s	F1 [%]	P [%]	R[%]
0-5	1737	82.8	97.2	72.1
5-10	1719	86.4	97.8	77.4
10-15	1428	88.0	98.4	79.6
15-20	1778	85.5	97.4	76.3

N_s is the number of segments where at least one event has the SNR within the given range.

have one or two events, indicating a large number of missing labels, therefore explaining the lower recall.

Table V indicates the number of segments with different degree of polyphony, with column 2 corresponding to the reference labels, and column 3 to the labels estimated using MACE. The metrics in columns 5–7 compare the reference and MACE output with respect to the number of segments in the reference (N_sGT). For the 10 segments of polyphony 1, all labels were correctly estimated (R = 100), but some of them were assigned more than one label (PR = 90.9). For the case of maximum polyphony, only half of the labels (R = 57.1) within the four segments where labeled correctly (PR = 100).

As expected, precision and recall vary with the polyphony, with recall decreasing at a high rate when polyphony increases. A similar annotator behavior was observed by Cartwright et al. [18] in the case of strong annotations: when more than two sounds overlapped, annotators failed to recall all the concurrent sounds. This may be because it is more difficult to identify sound events when there are more than two, but it may also show a tendency of the annotators to only identify one sound, and annotating a second one only if it was clearly identifiable. Additionally, the listening conditions play a role in identification too: in reality humans have better capabilities to distinguish overlapping sound events because of using both ears, therefore the spatial perception plays a role in the process, while listening to a mono recording in headphones does not provide the necessary spatial cues for disambiguation.

2) SNR Analysis: We investigate the effect of the SNR on the precision and recall of the sound events by annotators. Because each sound instance has been randomly assigned an SNR level when creating the synthetic mixtures, in most cases the 10 s segments contain more than one sound with different values of SNR. We group the segments by considering a segment into a specific SNR range if at least one of the sounds in the segment has the SNR within that range, and all other sounds in the segment have SNR within that range or lower (e.g. a segment with a sound at 7 dB, one at 3 dB, and one at 4 dB is in the [5–10] dB range).

The results, presented in table VI, show that recall is increasing with SNR, with a 7% absolute increase for sounds in the [10–15] dB range compared to those in the [0–5] dB range. The lower values for the [15–20] dB are observed due to the definition of these groups: based on the statistics in Table V, most segments have 2 or 3 sounds, so most of the 1778 segments with sounds in [15–20] dB interval also have some other sounds that are at lower SNR and are missed, hence the lower F1, P and R. Of the 1778 segments, 1660 segments (93%) have events with SNR lower than 15 dB. In this case inter-event ratios also play an important role: when two events occur simultaneously, the louder one will be masking strongly or at least partially the other one, making it harder to identify. For the other ranges this relative ratio is smaller (40% for [5–10] dB, 60% for [10–15] dB), resulting in less chances for masking. According to [32], identification accuracy and speed depends on the type of sound, therefore identifying the concurrent sounds will depend not only on the relative prominence of the sounds in a scene, but also on the degree of overlap, and the familiarity of the annotator with the sounds to be annotated.

If we consider only segments where all the sound events present have the SNR within the same interval, the number of evaluated segments decreases to about 20% of the total. In this case F1 for range [0-5] dB is 88.92% (202 segments), while for range [15-20] dB is 95.71% (118 segments), demonstrating the ease of annotating sound events that are relatively loud compared to the background.

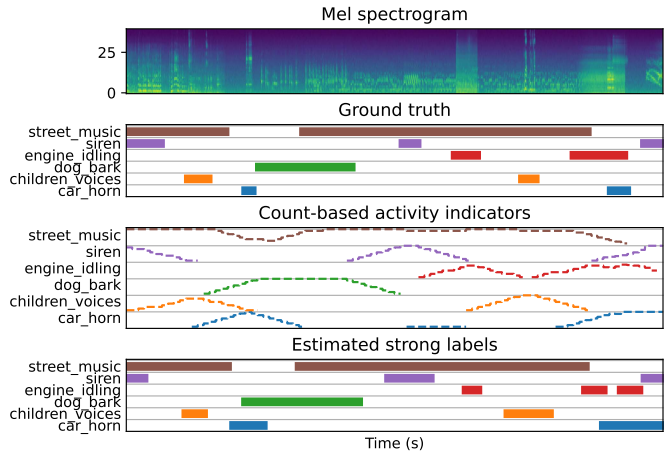

Fig. 3. Estimation of strong labels based on the weak labels of consecutive, overlapping, audio segments.

B. Strong Label Estimation

Following the scheme for temporal activity reconstruction of the sound events described in Section III-C, we stack the annotated segments in their original order and combine the multiple annotator opinions using the proposed majority opinion. An example of estimating the strong labels from the count-based activity curves is presented in Fig. 3. The annotation task produces

TABLE VII
SOUND EVENT DETECTION METRICS CALCULATED BETWEEN THE ESTIMATED STRONG LABELS AND THE GROUND TRUTH

Dataset	labels based on	ER_{1s}	S	D	I	$F1_{1s}$ [%]	P [%]	R [%]	$F1_{dtc=0.7}$ [%]	$F1_{dtc=0.1}$ [%]
MAESTRO Synthetic	(1) comp $>$ 0.6	0.44	0.02	0.36	0.05	72.6	88.9	61.3	44.0	81.3
	(2) MACE	**0.37**	0.02	0.20	0.15	**80.0**	82.3	77.9	41.1	90.1
	(3) majority opinion	0.46	0.03	0.14	0.29	77.4	72.2	83.4	39.3	92.2
	(4) strong annotation	0.46	0.04	0.40	0.02	69.1	90.0	56.1	40.1	70.9
MAESTRO Real Residential area	(1) comp $>$ 0.6	0.65	0.05	0.57	0.03	52.1	82.4	38.1	14.4	42.8
	(2) MACE	0.59	0.09	0.43	0.07	58.9	75.6	48.3	12.8	51.7
	(3) majority opinion	**0.53**	0.10	0.33	0.10	**64.3**	74.3	56.8	18.0	58.6
	(4) strong annotation	0.86	0.04	0.81	0.01	25.0	73.9	15.0	11.3	31.4
MAESTRO Real City center	(1) comp $>$ 0.6	0.64	0.03	0.61	0.00	52.5	92.2	36.6	22.9	55.1
	(2) MACE	0.54	0.06	0.47	0.02	61.5	86.8	47.6	22.9	57.6
	(3) majority opinion	**0.46**	0.07	0.36	0.03	**68.1**	84.4	57.1	25.0	61.6
	(4) strong annotation	0.88	0.02	0.86	0.00	21.0	85.9	11.9	10.4	30.9

S, D and I stand for substitutions, deletions and insertions, respectively.

5 opinions per 10 s segment, which translates into 50 opinions per 1 s segment, due to the staggered annotation procedure. According to the estimation method explained in Section III-C, the temporal location of an event instance corresponds to the region in which all annotators have considered it active in the weakly-labeled segments. To accommodate possible incorrect answers from the annotators, in [19] we used a threshold of 80% for binarizing this representation, i.e. a sound event was considered active in a 1 s segment if at least 80% of the opinions available for that segment considered it active [19].

We compare the proposed method with the MACE estimate as presented in [19], considering that MACE provided the best estimation of the reference weak labels; in addition, we also compare it with the aggregation of data from annotators with a competence higher than 0.6. The results are presented in VII in the following order: (1) using only annotators with a competence higher than 0.6; in this case, low-competence annotators are eliminated, resulting in a varying number of opinions per 1 s segment (on average 37, 20, and 13 annotators for synthetic, real-residential and real-city-center, respectively); (2) using the labels estimated with MACE; in this case, each 10 s segment is assigned the labels estimated by MACE, and there is only one opinion per 10 s segment (the MACE output), which translates into 10 opinions per 1 s segment, due to the staggered annotation procedure; (3) majority opinion. For cases (1) and (2) we use the 80% threshold to binarize the count-based activity, as explained above. For majority opinion, we binarize the activity at the mid-point of 0.5, according to the definition in Section III-B.

Table VII presents the SED scores between the reference annotations and the estimated strong labels based on the three described approaches, using segment-based F1 and ER [36] and intersection-based F-score as defined for the Polyphonic Sound Detection Score (PSDS) [16]. PSDS is evaluated for two scenarios, as defined in DCASE 2021 Challenge Task 4.[2] The two metrics are evaluated using the following parameters: $F1_{dtc=0.7}$ uses a detection tolerance criterion (DTC): 0.7; ground truth intersection criterion (GTC): 0.7; cost of instability across class (α_{ST}): 1; cost of CTs on user experience (α_{CT}): 0; maximum false positive rate (e_max): 100. $F1_{dtc=0.1}$ uses a detection tolerance criterion (DTC): 0.1; ground truth intersection criterion (GTC): 0.1; cost of instability across class (α_{ST}): 1; cross-trigger tolerance criterion (cttc): 0.3; cost of CTs on user experience (α_{CT}): 0.5; maximum false positive rate (e_max): 100. For details on the parameters and their effect, we refer the reader to [16].

The error rate (ER) consists of deletions (D), events present in the reference which are missed in the output, insertions (I), events erroneously marked as present in the output, and substitutions (S), events that are mislabeled in the output compared to the reference. We observe that a large proportion of errors in ER are deletions. This means many of the sound events were not identified, which is expected based on the previously observed recall rates in the weak labels analysis. Deletions (and implicitly ER) are very high for the real data, being about twice as many in comparison to the synthetic data, for a similar amount of annotated segments. This, in particular, indicates the high difficulty in identifying the target sounds in real-life mixtures.

The strong annotations estimated for the real-life data compare rather poorly with the reference annotation. The use of MACE has a clear effect on increasing recall, with the proposed majority opinion aggregation (3) providing the best outcome. However, the higher recall is reflected in a lower precision and a significant increase in insertions, even though the overall ER decreases. The best precision is obtained by using a selected proportion of highly competent annotators according to method (1), but this means discarding large amounts of raw data, in particular for the real audio recordings.

F1 values show a similar trend, with MACE helping improve the scores significantly. The majority opinion approach provides by far the best F1 for the real data, for all three calculated versions. For the synthetic data, the proposed method does not always provide the better strong label estimates.

Here one should not forget that the synthetic data comes with correct and complete reference annotations for the sound event instances, while the real recordings were manually annotated and therefore are prone to labeling errors that arise from subjective perception of each annotator. While the superiority of the proposed method can only be demonstrated numerically on the synthetic data, this does not diminish its importance; on

[2][Online]. Available: http://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments#evaluation

TABLE VIII
SEGMENT BASED METRICS (100 MS) OF AGGREGATED ANNOTATION IN THE STRONG ANNOTATION PROCEDURE ON MAESTRO SYNTHETIC

Aggregation	F1 [%]	P [%]	R [%]	ER
Majority vote	68.3	89.6	55.2	0.47
Union	66.8	54.6	85.9	0.77

the contrary, it shows that the proposed competence-weighted aggregation provides consistent results across different types of datasets, and may be used as an objective and reproducible procedure for creating strong annotations.

One scenario in which this method fails is when two events of the same class follow each other at short intervals, within a 10 s segment. In this case, correctly indicating presence of the sound event class in all segments that overlap any of the instances will create a situation where there are no gaps, leading to the estimation of a continuous, single instance.

C. Comparison to Direct Strong Annotation

For comparison, we reproduced the annotation method of Cartwright et al. [18] which provided workers with the spectrogram visualization along with the audio, and required annotators to produce strong annotations. We used the exact same annotation protocol through Amazon Mechanical Turk, using the code provided by the authors,[3] to collect five annotations for each audio file. We provided the visualization as a spectrogram, and explained to the annotators how it can be interpreted. The workers for this task were selected to have at least 95% accepted jobs.

Aggregation of the multiple annotations was done following the same procedure as in [18]: each annotation was transformed into a discrete sequence of 100 ms length segments; for each 100 ms segment, an event was considered active if the majority of the annotators (in this case 3 of 5) have annotated it as active. The resulting aggregated strong labels are compared with the ground truth (for synthetic data) or with the reference annotation (for real-life data).

Table VIII shows information retrieval measures in 100 ms segments for the synthetic data, for comparison with the work in [18]. The F1 of 68.3% is much lower than the approximately 93% in [18] in the case of 5 annotators. We hypothesize that this large difference is due to the annotation task being more difficult: our soundscapes have a length of 3 minutes, and may exhaust the worker's attention, in comparison with a short 10 s one. While for our experiment the precision and recall are 89.6% and 55.2%, respectively, the same metrics for the 10 s soundscapes in [18] are 98% and 95%. As an attempt to increase recall to the maximum possible, we verify the outcome of a union-based aggregation instead of consensus on the 100 ms segments, and obtain a recall of 85.9%. The method does however deteriorate precision, leading also to a much higher error rate.

In line with the other experiments presented in this paper, we calculate the SED metrics between the reference annotation and the aggregated strong annotation. The results are presented in Table VII as approach (4). For the synthetic data, the segment-based F1, P and R calculated in 1 s segments are in the same range with the same metric in 100 ms segments (Table VIII). In comparison with the evaluation of the other approaches in Table VII, we can conclude that this method provides very poor results, in particular on the real data. While precision values are comparable among the four approaches, the recall in the direct strong labeling approach is very low, also visible in the high proportion of deletions. An example of how our proposed method based on weak labeling and majority opinion behaves better than the direct strong annotation and majority vote aggregation shown in Fig. 4.

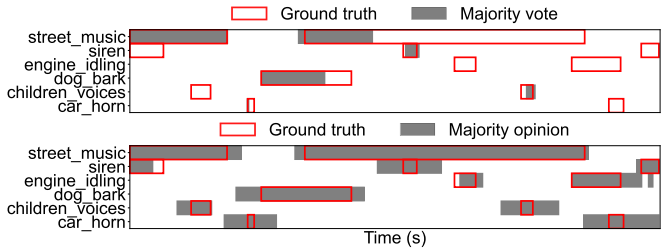

Fig. 4. Visual comparison of estimated labels to reference annotation using two different annotation methods for synthetic soundscapes; upper panel: direct strong annotation and majority vote aggregation; lower panel: proposed method consisting of weak labeling and majority opinion aggregation.

While we conclude that the strong annotation crowdsourcing as studied in [18] does not seem to be suitable for minutes-long real recordings, we have to mention a peculiar behavior of the annotators: the number of annotated event instances for the real data was very high, with a visible tendency of "filling up" the length of the audio. As can be seen in Fig. 5, the spectrogram visualization of a synthetic soundscape has more prominent segments corresponding to individual sound instances that are easily noticeable on the background, which may elicit a different annotator behavior. The complexity of the spectrogram for real data, brought by the unconstrained presence and overlapping of non-target sounds might give the impression that there is always something happening that needs to be annotated. In light of this, providing the spectrogram in the annotation task may have been detrimental to the quality of annotations instead of aiding the process.

In terms of time and cost, the required annotation effort for the two methods is quite different. While the weak labels are faster to annotate, the HITs were published in batches, and the average time for completing all batches of one dataset was 4 hours. In comparison, the strong annotation took on average 7 h for each set. Cost-wise, the tagging HITs were paid $0.10, while the strong annotation HITs were paid $5 each, which resulted in a 4 times higher cost for tagging than for the strong labeling. However, we observed that many workers in the tagging task completed the maximum allowed HITs, while for the strong annotations most workers completed only one HIT, indicating that they considered the work load too high. This reinforces our intuition that simple unit annotation tasks like tagging are preferable to ones requiring the annotator to take complex decisions such as is the case for strong labeling.

[3][Online]. Available: https://github.com/CrowdCurio/audio-annotator

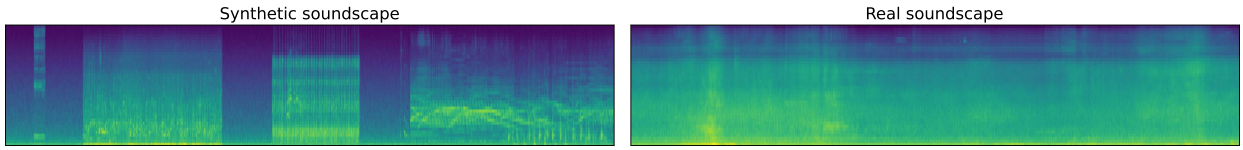

Fig. 5. Example of spectrogram visualization provided for synthetic and real soundscapes during strong annotation.

D. Sound Event Detection Using Estimated Labels

As an additional experiment, we investigate how the reference annotations influence the evaluated performance of a SED system. Traditionally, algorithms are trained and evaluated using the reference produced through manual annotation. Accepting that such annotation is subjective means that the reference is not necessarily complete, and the differences and disagreement between annotators may be the cause of some of the measured errors. Moreover, it is difficult to create a consistent annotation protocol that results in similar annotator differences for different datasets. As a consequence, testing a method across datasets will be affected by errors caused not only by the acoustic content mismatch but also by the mismatch in the labeling procedure.

In order to observe how this mismatch in the labeling procedure affects the evaluated performance, we train a SED system using the official reference labels, and evaluate its output against differently produced strong labels. We consider as baseline the system trained and evaluated using the reference annotation (generated for the synthetic data, annotated by experts for the real data). The experiments follow a leave-one-out setup in order to use as much as possible data for training the model. For each of the three datasets, the training/test procedure uses one soundscape for testing, one for validation, and the rest of the soundscapes for training the model. All training/test experiments were run first, and the evaluation was performed on the entire data at once, to avoid possible imbalances due to averaging over file-wise results [36].

We use PANNs [37], specifically the wavegram-logmel-CNN14 model,[4] consisting of six convolutional (conv) blocks. Each conv block contains two 2-D convolutional layers with a 3x3 kernel and batch normalization, followed by a ReLU nonlinearity layer. After each conv block 2x2 average pooling and a dropout layer with 0.2 rate are applied. The input to the PANNs is the concatenation of the log-mel spectogram features and the wavegram. Wavegram is a feature representation proposed by the authors in [37], which is learnt by a CNN block from the raw audio file. Using both mel spectrogram and wavegram as input features has been shown to improve the performance significantly compared to mel only [37]. The model is pretrained using AudioSet, with all audio converted to mono, resampled to 32 kHz, and padded to 10 s.

To fine-tune the model for our experiment, a fully-connected layer consisting of six units was added to the pre-trained conv layers of the selected PANNs model, after which the complete model was further trained for a few epochs (maximum number of epochs 50) to learn to classify the given six classes. The fine-tuning was done separately for each subset.

The audio files available for fine-tuning the model were resampled to 32 kHz, to be similar to the data used for pre-training, then cut into 10 s segments that are individually processed. The mel spectrogram was calculated using a window size of 1024 with a hop length of 320 samples, and 64 mel filter banks, with the lower and upper frequencies set to 50 and 14 kHz, resulting in a 1001×64 feature matrix for an audio clip of 10 s. The same parameters were applied to the wavegram. SED was performed using the inference method provided by Kong et al. [37], which outputs frame-based activity within the processed segment for the target sound classes. Postprocessing of the individual outputs was done to create a single list of detected event instances corresponding to the entire analyzed audio file.

The results are presented in Table IX and show that the training and evaluation process is not highly robust to the mismatch in the reference label production method: the system trained with the reference annotation and evaluated with the estimated strong labels has the highest ER, and especially high for the real data. Compared against complete and correct annotations, we obtained at best ER_{1s} of 0.45 and $F1_{1s}$ 71.7% for a set of synthetic soundscapes.

The significant decrease observed in the metrics when the system output is evaluated against the crowdsourced annotations (second line in Table IX for each dataset) compared to the official reference annotation (first line in IX) is not connected to the actual performance of the system, since we evaluate the exact same system output, but against a differently produced reference. Based on the analysis presented earlier, we know that the estimated strong labels are only a subset of the reference, therefore this evaluation across annotation methods is blending the effect of the errors in the system output and the errors (missed or substituted events) in the crowdsourced annotations. When the system is trained and evaluated using the crowdsourced annotations (third line in IX), the measured performance is similar to the one obtained using the official reference for each dataset.

This experiment highlights one main challenge in the development of robust SED systems: when the data available for system development (training) is annotated using a differently defined procedure or a different pool of experts than the data expected at deployment stage (evaluation), a large drop in performance does not necessarily indicate a failure to generalize to different acoustic conditions, but hides a mismatch in the labeling process, which can be caused by differences in the definition of the labeling procedure itself and differences in annotators' opinion. Even though its output is not entirely accurate, the method proposed in this paper is based on multiple opinions per item and a very large

[4][Online]. Available: https://zenodo.org/record/3987831

TABLE IX
EVALUATION AGAINST DIFFERENT SETS OF STRONG LABELS FOR THE THREE DATASETS USING PANNS

Dataset	eval. reference	ER_{1s}	$F1_{1s}$ [%]	$F1_{dtc=0.7}$ [%]	$F1_{dtc=0.1}$ [%]
MAESTRO Synthetic	GT	0.30	81.0	57.5	72.6
	estim. majority opinion	0.46	70.8	45.2	67.8
	train&eval majority opinion	0.45	71.7	44.3	61.6
MAESTRO Real Residential area	GT	0.59	61.2	47.5	58.1
	estim. majority opinion	0.71	60.7	48.6	54.0
	train&eval majority opinion	0.50	69.8	49.9	63.3
MAESTRO Real City center	GT	0.58	63.1	70.1	58.6
	estim. majority opinion	0.77	62.3	65.6	54.2
	train&eval majority opinion	0.48	67.9	72.9	80.2

pool of annotators, therefore it has the potential of producing the labels in a more objective and reproducible manner.

VI. CONCLUSION

While crowdsourcing has been repeatedly used as a fast method to collect large amounts of labeled data, the specific format of strong labels for sound event detection is still difficult to crowdsource. In addition to the complexity of the task itself, the outcome is affected by subjectivity of the annotators in perceiving the sounds. Collecting multiple opinions alleviates the subjectivity, but comes with the question on how to aggregate the multiple annotations for the best outcome.

This paper presented two key contributions to the research problem of crowdsourcing strong labels. First, we introduced a novel workflow in the crowdsourcing task which breaks the strong annotation process into two stages: weak labeling and reconstruction of temporal information based on the weak labels. The weak labeling task is much simpler than strong labeling, therefore expected to produce consistent quality labels. Second, we proposed a novel method for aggregating multiple annotator opinions, using annotator competence estimation tools. Given that some users produce more reliable annotations than others, replacing the majority vote aggregation with a majority opinion scheme was expected to produce higher quality outcome.

Results have shown that weighing the annotators' opinions by their estimated competence produces better strong labels than any other method, including direct strong annotation. In addition, the results show that the proposed majority opinion approach produces reliable aggregated strong labels in comparison with a manually annotated reference produced by an expert annotator. Using a SED experiment, we have also shown how a model's evaluated performance is linked to the selected reference annotation. Annotations produced manually by different annotators reflect their personal biases and are prone to annotator-dependent errors, which are not separable from the system-produced errors when evaluated against. The proposed method uses multiple annotators in a crowdsourced manner and a data-independent processing chain for producing the strong labels, therefore has the advantage of being objective and reproducible, even though the produced annotations were shown to be incomplete.

Future research may investigate incorporating additional knowledge into the workflow. The main advantage of the proposed approach is its streamlined and reproducible setup, but the drawback is its high level of redundancy. For a more efficient method, it would be useful to preprocess the audio to select regions of interest, so that only the parts expected to contain the target events are annotated with high redundancy.

REFERENCES

[1] E. Fonseca et al., "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2018, pp. 69–73.

[2] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, 2017, pp. 776–780.

[3] A. Mesaros et al., "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 992–1006, Jun. 2019.

[4] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 684–698, 2021.

[5] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Process. Mag.*, vol. 38, no. 5, pp. 67–83, Sep. 2021.

[6] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Conf. Multimedia*, New York, NY, USA, 2014, pp. 1041–1044.

[7] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, New York City, NY, United States, 2019, pp. 253–257.

[8] 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 Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2450–2460, 2020.

[9] H. Dinkel, M. Wu, and K. Yu, "Towards duration robust weakly supervised sound event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 887–900, 2021.

[10] S. Park, A. Bellur, D. K. Han, and M. Elhilali, "Self-training for sound event detection in audio mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 341–345.

[11] N. Turpault and R. Serizel, "Training sound event detection on a heterogeneous dataset," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, Tokyo, Japan, 2020, pp. 200–204.

[12] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning for weakly-labeled semi-supervised sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 626–630.

[13] A. H. Cheung, Q. Tang, C.-C. Kao, M. Sun, and C. Wang, "Improved student model training for acoustic event detection models," in *Proc. 6th Detection Classification Acoust. Scenes Events Workshop*, Barcelona, Spain, 2021, pp. 181–185.

[14] C. Guastavino, "Everyday sound categorization," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. Ellis, Eds. Berlin, Germany: Springer, 2018, ch. 7, pp. 183–213.

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

[16] 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.

[17] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2021.
[18] M. Cartwright et al., "Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations," in *Proc. ACM Hum.-Comput. Interact.*, 2017, pp. 1–21.
[19] I. Martín-Morató, M. Harju, and A. Mesaros, "Crowdsourcing strong labels for sound event detection," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021, pp. 246–250.
[20] E. J. Humphrey, S. Durand, and B. McFee, "OpenMIC-2018: An open data-set for multiple instrument recognition," in *Proc. Int. Soc. Music Inf. Retrieval*, 2018, pp. 438–444.
[21] M. Cartwright, G. Dove, A. E. M. Méndez, J. P. Bello, and O. Nov, "Crowdsourcing multi-label audio annotation tasks with citizen scientists," in *Proc. Conf. Hum. Factors Comput. Syst. - Proc.*, 2019, pp. 1–11.
[22] T. Kauppi et al., "Fusion of multiple expert annotations and overall score selection for medical image diagnosis," in *Proc. Scand. Conf. Image Anal.*, 2009, pp. 760–769.
[23] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.
[24] J.-K. Kamarainen, L. Lensu, and T. Kauppi, "Combining multiple image segmentations by maximizing expert agreement," in *Proc. Mach. Learn. Med. Imag.*, 2012, pp. 193–200.
[25] T. A. Lampert, A. Stumpf, and P. GanÇarski, "An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2557–2572, Jun. 2016.
[26] I. Martín-Morató and A. Mesaros, "What is the ground truth? reliability of multi-annotator data for audio tagging," in *Proc. 29th Eur. Signal Process. Conf.*2019, pp. 76–80.
[27] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning whom to trust with MACE," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2013, pp. 1120–1130.
[28] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. 24th Eur. Signal Process. Conf.*, Budapest, Hungary, 2016, pp. 1128–1132.
[29] P. Zinemanas, P. Cancela, and M. Rocamora, "MAVD-traffic dataset," 2019. [Online]. Available: https://doi.org/10.5281/zenodo.4741232
[30] S. Hershey et al., "The benefit of temporally-strong labels in audio event classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 366–370.
[31] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 736–740.
[32] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds," *J. Exp. Psychol.: Hum. Percep. Perform.*, vol. 19, no. 2, pp. 250–267, 1993.
[33] I. Martín-Morató, M. Harju, and A. Mesaros, "MAESTRO synthetic - multi-annotator estimated strong labels," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5126478
[34] 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.
[35] A. Mesaros et al., "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. Workshop Detection Classification Acoust. Scenes Events*, 2017, pp. 85–92.
[36] A. Mesaros, T. Heittola, and D. Ellis, "Datasets and evaluation," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. Ellis, Eds. Cham, Switzerland:Springer, 2018, ch. 10, pp. 147–179.
[37] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

Irene Martín-Morató (Member, IEEE) received the bachelor's degree (with Hons.) and the M.Sc. degree in telecomunications, and the Ph.D. degree in information technology, communications and computing under the University Faculty Training Program (FPU) from the Universitat de València, València, Spain, in 2014, 2016, and 2019, respectively. She is currently a Postdoctoral Research Fellow with Tampere University, Tampere, Finland. Her research interests include the field of acoustic signal processing, machine learning, and audio event detection and classification.

Annamaria Mesaros (Member, IEEE) received the Ph.D. degree in signal processing from the Tampere University of Technology, Tampere, Finland, in 2012. She is currenty an Assistant Professor with Tampere University, Tampere, Finland. She is also the Coordinator with International Evaluation Challenge on Detection and Classification of Acoustic Scenes and Events. Her research focuses on sound event detection in real-world multisource environments, and includes more than 40 scientific publications on this topic and many open datasets. She is the Academy of Finland Research Fellow for Teaching Machines to Listen and the Member of the Audio and Acoustic Signal Processing Technical Committee of IEEE Signal Processing Society.