IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A multi-resolution CRNN-based approach for semi-supervised Sound Event Detection in DCASE 2020 Challenge

DIEGO DE BENITO-GORRÓN, DANIEL RAMOS, DOROTEO T. TOLEDANO
AUDIAS Research Group, Escuela Politécnica Superior, Universidad Autónoma de Madrid (Madrid, Spain) (e-mail: diego.benito@uam.es)

Corresponding author: Diego de Benito-Gorrón (e-mail: diego.benito@uam.es).

Work developed under project DSForSec (RTI2018-098091-B-I00), funded by the Ministry of Science, Innovation and Universities of Spain and FEDER.

ABSTRACT Sound Event Detection is a task with a rising relevance over the recent years in the field of audio signal processing, due to the creation of specific datasets such as Google AudioSet or DESED (Domestic Environment Sound Event Detection) and the introduction of competitive evaluations like the DCASE Challenge (Detection and Classification of Acoustic Scenes and Events). The different categories of acoustic events can present diverse temporal and spectral characteristics. However, most approaches use a fixed time-frequency resolution to represent the audio segments. This work proposes a multi-resolution analysis for feature extraction in Sound Event Detection, hypothesizing that different resolutions can be more adequate for the detection of different sound event categories, and that combining the information provided by multiple resolutions could improve the performance of Sound Event Detection systems. Experiments are carried out over the DESED dataset in the context of the DCASE 2020 Challenge, concluding that the combination of up to 5 resolutions allows a neural network-based system to obtain better results than single-resolution models in terms of event-based F1-score in every event category and in terms of PSDS (Polyphonic Sound Detection Score). Furthermore, we analyze the impact of score thresholding in the computation of F1-score results, finding that the standard value of 0.5 is suboptimal and proposing an alternative strategy based in the use of a specific threshold for each event category, which obtains further improvements in performance.

INDEX TERMS Sound Event Detection, Multi-Resolution, DCASE 2020 Task 4

I. INTRODUCTION

UNDERSTANDING the acoustic environment is an ongoing challenge for artificial intelligence which has motivated several research fields. While some of them are focused in the retrieval of information from specific kinds of acoustic signals, such as automatic speech recognition [1], [2], language or speaker identification [3], [4] (for speech signals) or music information retrieval [5], [6] (for musical signals), other tasks aim to determine the categories which an audio recording belongs to, among a set of target classes (e.g. human voice, vehicle, musical instruments) [7]. These categories can either refer to different environments where a recording can be obtained (e.g. inside a house or in a crowded street) or to different actions or sources which produced the obtained acoustic signals. In the former case, we talk about acoustic scene classification [8], while in the latter the problem at hand is sound event classification or detection [9].

Sound events can be defined as acoustic signals that have a direct correspondence with particular occurrences in the near environment. Hence, by hearing these sounds people can infer that the event is happening somewhere around them. Sound event classification and sound event detection (SED) aim to solve this problem for machine perception. In the case of sound event classification, signals are expected to belong to one among a set of target categories, while the temporal boundaries of the events are not relevant [10]. If multiple categories can be assigned to each recording, the task is called audio tagging [11]. The task that aims to find the time limits of each event is sound event detection, which can be monophonic (if only one event can be present at a given time)

or polyphonic (if different events can overlap in time). In every case, the set of target event categories is usually defined by the field of application, ranging from a single target event (e.g. speech activity detection) to potentially hundreds of categories.

Training and developing modern systems for the aforementioned tasks requires the use of large-scale labeled audio event datasets. In the field of computer vision, the research in object recognition was notably impulsed by the creation of ImageNet, a large-scale, hierarchical image corpus [12]. This motivated the creation of Google AudioSet, a large-scale audio dataset consisting of more than two million ten-seconds audio recordings, annotated according to an ontology of more than 500 sound events [13]. In the recent years, research has been carried out not only aiming to detect every category in AudioSet, but also focusing on smaller, application-oriented subsets of event classes. Recent editions of DCASE Challenge (Detection and Classification of Acoustic Scenes and Events), one of the most relevant international evaluations in this field, have employed subsets of the recordings specified in AudioSet for tasks such as audio tagging and sound event detection [14]–[16].

Regarding the creation of audio event datasets, audio recordings are relatively easy to obtain from web resources like YouTube[1], Vimeo[2] or Freesound[3]. However, it is costly to annotate them with human-verified event labels, therefore it is common for large-scale audio datasets to include only weak labels (i.e. indications of the presence or absence of each event in a recording, without time boundaries), usually obtained in a semi-automatic manner. A certain amount of label noise is likely to appear in the process of annotation, due to involuntary omission or insertion of labels in the ground truth [10]. Hence, additional challenges arise in the learning process, such as developing algorithms which are robust to label noise [17] and inferring the temporal locations of events from weak labels [18]. Moreover, validating the performance of the systems requires verified annotations and, in the case of sound event detection, strong labels (indicating temporal onsets and offsets, in opposition to weak labels). For this purpose, smaller datasets have been curated with human-revised annotations [19].

Over the recent years, most works in Sound Event Detection have employed deep neural network (DNN) models, being particularly common those with convolutional and recurrent stages [20]. These systems usually take as input time-frequency representations of audio signals based in the Short-Time Fourier Transform (STFT). The most frequent type of audio feature in this task is the mel-spectrogram, a two-dimensional representation of audio which uses the Mel-frequency scale. For a given audio sample frequency (f_s), the temporal and frequency resolution of such representation is defined by the parameters of the feature extraction process: the size of the FFT, the temporal window of the STFT and the number of Mel filters.

We hypothesize that, due to the different temporal and spectral characteristics of different kinds of acoustic events, employing several resolution points in the feature extraction process would improve the performance of sound event detection systems. Following this idea, in this paper we propose a multi-resolution approach for the task of sound event detection.

The use of multiple input resolutions has been already explored in several deep learning applications. One of them is the task of object detection in the computer vision field, which can be considered an analogous problem to sound event detection. However, multi-resolution has different properties when dealing with image data or audio features. In a picture, multiple resolutions can be helpful to recognize objects at different scales [21], [22], but the desired benefit when using more than one resolution in audio applications is to exploit different details of the feature maps with each resolution point. For instance, the use of two different resolutions has been proposed to improve automatic speech recognition in reverberant scenarios [23], in which a wide-context window gives information about the acoustic environment and reverberation, whereas a narrow-context window provides finer detail about the content of the speech signal. This is possible due to the existence of a trade-off between time resolution and frequency resolution in the extraction of Fast Fourier Transform-based audio features [24] such as the mel-spectrogram, which is also the base for the analysis proposed in this work.

Our multi-resolution approach relies on two key aspects: one of them is the choice of the time-frequency resolution points to be considered, while the other one is the method employed to combine the different resolutions. Thus, the intermediate stages between feature extraction and the combination of resolutions (e.g. the topology of the neural network models) are not affected by this approach and can be considered as black-boxes to be used by the multi-resolution system.

Considering that multi-resolution can be implemented independently of the underlying sound event detection systems, the potential improvements in performance are complementary to those that could be obtained by optimizing the hyperparameters of the neural networks. Therefore, through this approach, multi-resolution could be added to other DNN-based sound event detection systems in a similar manner.

The proposed analysis is tested using a state-of-the-art system, the baseline for DCASE 2020 Challenge Task 4 "Detection and Separation of Sound Events in Domestic Environments" [25]. The aim of this challenge is to make use of unlabeled and weakly-labeled recordings, together with strongly-labeled synthetic audio clips, to train systems that predict the temporal locations of ten different event categories in audio recordings. Furthermore, an additional contribution of this paper is an exploration of the impact of different score thresholding strategies in the performance of

[1]http://youtube.com/
[2]http://vimeo.com/
[3]http://freesound.org/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088949, IEEE Access

IEEE Access

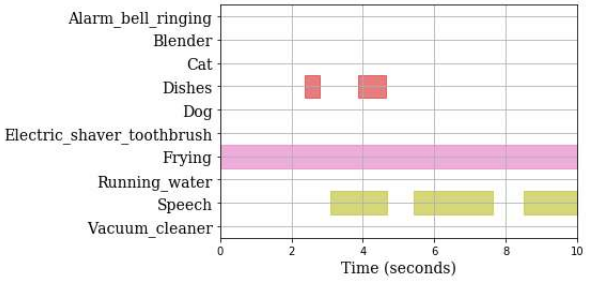

FIGURE 1. Ground truth event annotations provided for an audio segment of the Validation set. The horizontal bars represent the time intervals where each event category is active. Several categories can be active at the same time.

the systems.

The rest of the paper is structured as follows: Section II presents the evaluation metrics, the dataset and the most common approaches of the DCASE Challenge Task 4. Section III introduces the multi-resolution analysis, explaining its motivation and giving details about the definition of different resolution points. In Section IV the experimental framework is described, discussing the details of the models employed, the model fusion process, the F1-score thresholding and the post-processing of scores. Section V contains the results for the different experiments, discussing the impact of multi-resolution and thresholding in the performance of sound event detection, as well as a study of the behavior of overlapped events, the impact in execution times, and an analysis of the relationship between the results for each category and the characteristics of the audio. Finally, Section VI highlights the conclusions of this work.

II. SOUND EVENT DETECTION IN DCASE 2020 CHALLENGE

A. DCASE 2020 TASK 4: "DETECTION AND SEPARATION OF SOUND EVENTS IN DOMESTIC ENVIRONMENTS"

The goal of DCASE Challenge 2020 Task 4, "Detection and Separation of Sound Events in Domestic Environments," is to explore the use of both labeled and unlabeled data to build systems for sound event detection, considering a set of ten event categories drawn from the AudioSet ontology. The target categories describe acoustic events typically found in domestic acoustic scenes: *Speech*, *Dog*, *Cat*, *Alarm/bell/ringing*, *Dishes*, *Frying*, *Blender*, *Running water*, *Vacuum cleaner* and *Electric shaver/toothbrush*. The task consists on determining the starting and ending time of each event found in the audio segments, considering that more than one event category can be active at the same time. An example is provided in Fig. 1.

Systems are evaluated by means of the F1-score metric, widely used to measure performance in Sound Event Detection tasks [26]. In order to compute F1-score, some intermediate metrics have to be computed: True Positives (TP), False Positives (FP), and False Negatives (FN). Different definitions of these statistics lead to either event-based or segment-based F1 metrics.

For event-based metrics, each instance of an event in the ground truth and each event predicted by the system are considered in order to count TPs, FPs, and FNs. Usually, a collar-based approach is taken, considering some tolerance (collar) for the estimations of onset and offset times. A prediction is considered correct if the difference between the predicted time and the ground truth is equal or lower than the collar for both the onset and the offset times. The value of collars in DCASE 2020 Task 4 is 200ms for onsets and $\max(200\text{ms}, 0.2\times\text{event length})$ for offsets, hence the offset collar is more tolerant for longer events, which often present more diffuse endings.

Segment-based metrics, on the other hand, compare the ground truth with the system predictions in short time intervals. Each interval can be counted as a TP, a FP, or a FN depending on its ground truth label and the system prediction. While event-based metrics give the same importance to each event, in segment-based metrics longer events are considered more relevant, as they contain more time intervals. Segment-based metrics are more robust to short pauses between events that may not be reflected in the ground truth labelling.

The F1-score for a given category is then obtained from the number of TPs, FPs, and FNs.

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

The global F1-score can be obtained in two different manners. On the one hand, micro-averaged F1-score gives equal weight to each event ocurrence, thus the predominant categories in the dataset are given more importance. In contrast, macro-averaged F1-score gives the same weight to each category, independently of the number of occurrences.

Event-based, macro-averaged F1-score is the primary metric in DCASE 2020 Task 4, whereas PSDS (Polyphonic Sound Detection Score) is proposed as a complementary measure of performance [27]. PSDS aims to solve several issues of the F1-score as a performance metric for Sound Event Detection:

- **Single operating point.** F1-score is defined using a single decision threshold for each event category. Moreover, such decision threshold is usually set at 0.5 for every category by default, with no evidence of this value being optimal. On the contrary, PSDS considers a set of thresholds linearly distributed between 0 and 1, averaging the performance of the system in each one of them. Thus, PSDS is independent of the choice of the decision threshold.
- **Subjectivity in ground truth.** Human annotators can label the starting and ending time of events in each audio segment with sufficient precision, however, these labels are not objective because the same recording could be correctly labelled in several different ways. For instance, a short event that happens three times in a brief lapse of time could be labelled by some annotators as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088949, IEEE Access

three different occurrences of the event, but as a single occurrence by others. When evaluating a system with a collar-based metric (e.g. event-based F1-score), none of the possible system outputs would be correct for both labeling options. In order to overcome this problem, PSDS takes a different approach to the comparison of predictions and labels, based on intersections rather than time collars.

- **Importance of cross-triggers.** When training a sound event detector with multiple target categories, some of the false positive predictions can match a different event class. These are called cross-triggers, and taking them in consideration can provide a better understanding of errors in the detectors. Cross-triggers are more usual between acoustically similar categories, therefore they might indicate a bias in data rather than a flaw in the model.

PSDS introduces two criteria to define TPs and FPs. The Detection Tolerance Criterion (DTC) sets a minimum intersection between a prediction and ground truth labels of the same class for such prediction to be considered relevant. Non-relevant predictions are counted as FPs. On the other hand, Ground Truth intersection Criterion (GTC) controls the minimum percentage of a ground truth label that must be covered by relevant predictions of its class to be considered as a TP. A third rule, the Cross-Trigger Tolerance Criterion (CTTC) is defined to set the necessary intersection between a non-relevant prediction and ground truth labels of a different class for the prediction to be considered a cross-trigger.

For each of the three criteria, a parameter ρ defines the corresponding ratio of intersection. In DCASE 2020 Task 4, the value of these three parameters is fixed to $\rho_{DTC} = 0.5$, $\rho_{GTC} = 0.5$, and $\rho_{CTTC} = 0.3$.

Moreover, two cost parameters are introduced. α_{ct} defines the cost of cross-triggers in the PSDS score, while α_{st} penalizes the instability of TP rates across different classes. By combining the value of these parameters, three PSDS configurations are defined as follows:

- PSDS ($\alpha_{ct} = 0$, $\alpha_{st} = 0$)
- PSDS Cross-Trigger ($\alpha_{ct} = 1$, $\alpha_{st} = 0$)
- PSDS Macro ($\alpha_{ct} = 0$, $\alpha_{st} = 1$)

B. DESED DATASET

The dataset used in DCASE 2020 Task 4 is DESED (Domestic Environment Sound Event Detection) [28], [29]. DESED is composed of real and synthetic audio recordings. Real recordings are obtained from AudioSet segments, extracted from YouTube, while synthetic recordings are generated by overlapping foreground event clips from the target categories over background recordings of domestic environments. The generation of synthetic audio clips is performed with the Scaper library [30], using foreground audios from Freesound and backgrounds from the SINS dataset [31].

The DESED dataset for DCASE 2020 Task 4 is divided into different subsets:

- **Synthetic training set** (2584 clips). Synthetic recordings with strong labels.
- **Weakly-labeled training set** (1578 clips). Real recordings from AudioSet with weak labels.
- **Unlabeled training set** (14412 clips). Real recordings from Audioset which contain events from the set of target categories, with no labels provided.
- **Validation set** (1168 clips). Real recordings from AudioSet with human-annotated strong labels.
- **2020 Evaluation set.** Real recordings from YouTube and Vimeo with human-annotated strong labels (for system ranking), and synthetic recordings with strong labels (for result analysis). Ground truth labels are not publicly available, but results can be obtained by sending automatic annotations to the organizers of the evaluation.

C. EXISTING APPROACHES TO SOUND EVENT DETECTION

Existing trends in Sound Event Detection systems can be described by observing the submissions to the DCASE Task 4 evaluations over the recent years. Since 2018, the same set of sound event categories is used, as well as a similar organization for the dataset.

Taking into account that Sound Event Detection aims to infer the temporal locations of events in a given audio recording, in general terms the input to the system is some representation of the audio segment, while the final output is a list of predictions indicating the starting and ending times and the category of the event detected. Hence, the Sound Event Detection task can be interpreted as an independent two-class classification problem (presence or absence) for each target event category, as described in Fig. 2.

In particular, DCASE Task 4 proposes an scenario where only a small portion of the audio corpus is annotated. Moreover, these annotations were exclusively weak labels in DCASE 2018 [20], while in 2019 an additional subset with strongly-labeled, synthetic recordings was introduced [16]. To deal with the lack of strong labels, several semi-supervised learning methods have been proposed. Pseudo-labeling [32] was the most popular approach until the success of the mean-teacher scheme [33]. Pseudo-label trains a first system using only the labeled data, and uses such system to generate labels for the unlabeled recordings. Then, a final system is trained using the labeled and pseudo-labeled data. On the other hand, mean-teacher involves a single training process, with a student model and a teacher model which uses the exponential moving average of the student model weights. In addition to the usual classification cost, a consistency cost is defined to learn from unlabeled data, encouraging the system to provide consistent outputs when the input is corrupted with a slight amount of noise [34].

Sound event detection systems usually consist on convolutional neural networks (CNN), recurrent neural networks (RNN), or neural networks combining convolutional and recurrent stages (CRNN): among the top-10 submissions to

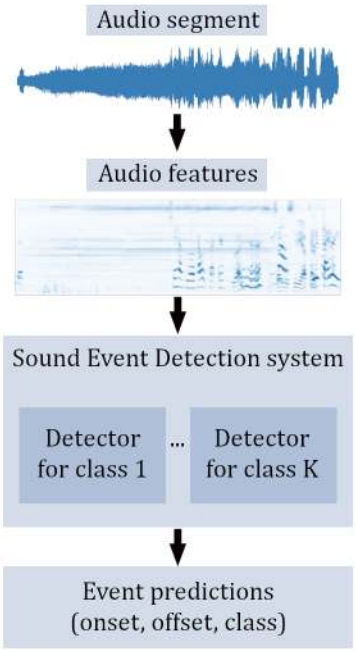

FIGURE 2. Block diagram describing the general pipeline of a Sound Event Detection system with *K* target categories.

the last three editions of the DCASE Challenge Task 4, every participation employed at least one of the listed models [16], [20], [35].

The most recurring type of input feature is the mel-spectrogram, a time-frequency representation of the audio which is widely used for many sound analysis tasks. The extraction process starts with a Short-Time Fourier Transform (STFT), which applies a Discrete-Time Fourier Transform to a moving temporal window of the audio signal, resulting on a bidimensional representation of the evolution of the spectrum in time. A bank of mel-filters is then used to map the spectra to the auditory Mel-scale, obtaining the mel-spectrogram. Almost every submission to DCASE Task 4 in its last three editions uses mel-spectrogram features, while some participants used other types of representations such as MFCC (Mel-Frequency Cepstral Coefficients), Δ features, CQT (Constant-Q Transform), or the raw waveform.

Some research has been carried out trying to apply different resolutions at some point of the detection process. For instance, controlling the size of a median filter during the post-processing according to the average temporal duration of the target category [36], [37]. Other existing approach is to process the audio segments with two different temporal resolutions, one aiming to optimize audio tagging performance and the other trying to specialize in temporal localization of events [38]. However, these approaches use input features that limit the audio representation to a particular time-frequency resolution.

In order to determine the temporal boundaries (onset and offset) for each prediction, binary decisions have to be made. Considering that the usual output of neural networks for two-class classification problems is a sigmoid-based score bounded between 0 and 1, a decision-making criterion has to be defined. The standard approach is to set a threshold value $th \in (0, 1)$, so that the presence of an event is predicted when the score is above th. In the systems proposed for DCASE 2018 and 2019 Challenge Task 4, the value for this threshold is usually set to $th = 0.5$, without further justification for choosing such value. Nevertheless, some different thresholding strategies have been proposed, such as double-thresholding [39] or dynamic thresholding [40].

III. MULTI-RESOLUTION ANALYSIS

The main hypothesis for the experiments presented in this paper is that sound event detection systems can benefit from using different time and frequency resolutions in the feature extraction process, instead of using a single resolution point, which is the most common approach in previous works. This idea is motivated by the fact that acoustic events can present very different temporal and spectral characteristics. A similar approach was proposed recently for the task of automatic speech recognition [41], obtaining modest but consistent improvements despite the types of sounds to be classified (human phones) were much more similar.

The distribution of the time durations of the examples in each class have been computed over the Synthetic Training set and are presented in Fig. 3 as a histogram for each category. The figure shows that the distribution of time durations vary very significantly depending on the event class. While some categories tend to have very short examples (*Alarm bell/Ringing*, *Cat*, *Dishes*, *Dog*, or *Speech*), others present more diverse lengths (*Electric shaver/Toothbrush*, *Frying*, or *Running water*).

During the extraction of mel-spectrogram audio features, a particular time-frequency resolution point is defined by the set of parameters used, given the sample rate (f_s) of the audio segment. Such parameters are the number of samples of the DTFT (N), the type of window used in the STFT, its length (L) and hop size (R), and the number of filters of the Mel filter bank (n_{mel}). Varying the values of these parameters, the temporal and the frequency resolutions of the resulting features will be different. There is a compromise between the time and frequency resolutions, as increasing one of them implies decreasing the other one.

In order to illustrate the convenience of using multiple time-frequency resolutions to represent different sound events, Fig. 4 provides an example where two mel-spectrograms are displayed for the same audio segment which belongs to the class *Electric shaver/Toothbrush*, using two different time-frequency resolution points. The acoustic event presents some frequential components that remain constant in time, resulting in horizontal lines in the mel-spectrogram. Such lines are much better captured by the second mel-spectrogram, which offers a higher frequency resolution.

A second example is provided in Fig. 5, representing an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088949, IEEE Access

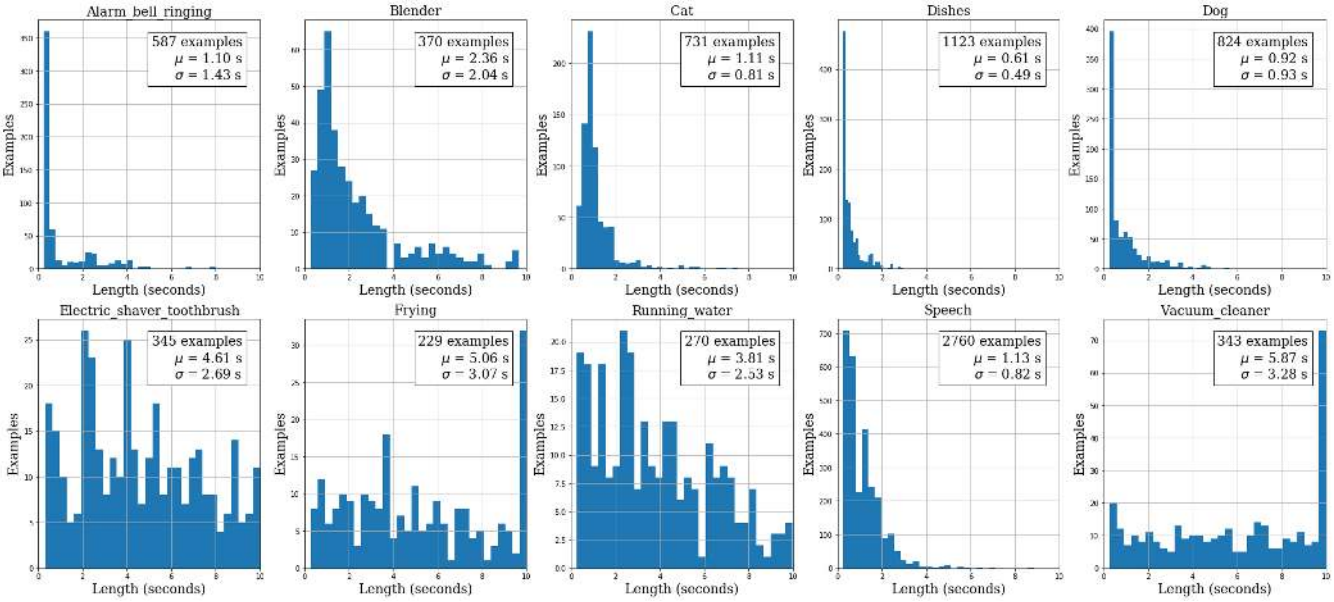

FIGURE 3. Histograms of the durations of each event category in the Synthetic Training set.

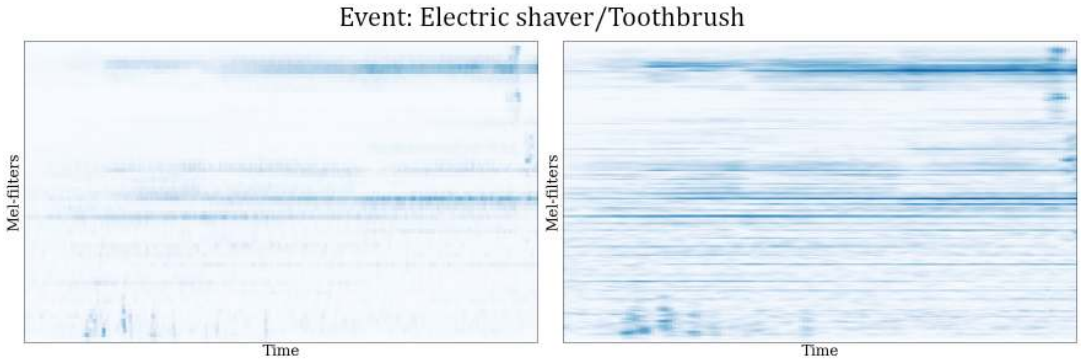

FIGURE 4. Two representations of the same audio segment belonging to the category *Electric shaver/Toothbrush*. Each representation is a mel-spectrogram extracted using a higher temporal resolution (left) and a higher frequency resolution (right).

Alarm bell/Ringing audio segment using two different resolutions. At the beginning of the segment, some repetitions of a tone can be observed, which could be a representative aspect of this event category. However, the different repetitions are better represented in the left mel-spectrogram, which offers a higher temporal resolution.

Taking the resolution of the baseline system as a reference, we define four additional resolution points. The five working points used share in common with the baseline the use of a sample rate of $f_s = 16000$ Hz and the use of a Hamming window, while the other parameters (N, L, R, n_{mel}) are modified to increase the time resolution or the frequency resolution. The configuration of each resolution point is described below, and the values of the parameters are presented in Table 1.

1) **BS** (Baseline). The baseline uses an analysis window of length $L = 128$ ms and a window hop of $R = 15.94$ ms (255 samples). Both parameters are related to the temporal resolution of the analysis. On the other hand, the frequency resolution is limited by the width of the main lobe of the Hamming window, $8\pi/(L-1) = 8\pi/2047$ rad/sample, which corresponds to a frequency resolution of $4/2047 \times 16000 \approx 31$ Hz. However, this frequency resolution is later more limited in a non-linear way by the use of the Mel filterbank with 128 filters.
2) **T++** (Twice better time resolution). We halve the analysis window to a length of $L = 64$ ms and the window hop to $R = 8$ ms, which essentially doubles the time resolution. We also halve the number of Mel filters, which along with the previous changes roughly halves the frequency resolution.
3) **F++** (Twice better frequency resolution). We double the analysis window length to $L = 256$ ms and the window hop to $R = 32$ ms, which essentially halves the time resolution. We also double the number of Mel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088949, IEEE Access

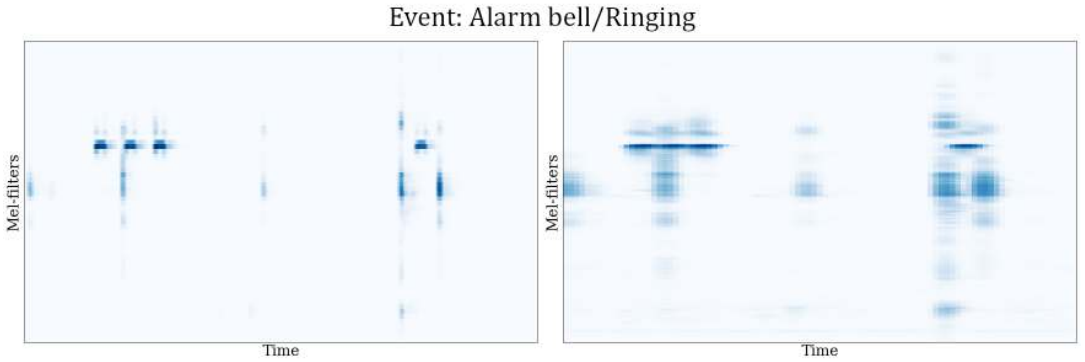

FIGURE 5. Two representations of the same audio segment belonging to the category *Alarm bell/Ringing*. Each representation is a mel-spectrogram extracted using a higher temporal resolution (left) and a higher frequency resolution (right).

TABLE 1. FFT length (N), window length (L), window hop (R) and number of Mel filters of the five proposed time-frequency resolution working points.

Resolution	T_{++}	T_{+}	BS	F_{+}	F_{++}
N	1024	2048	2048	4096	4096
L	1024	1536	2048	3072	4096
R	128	192	255	384	512
n_{mel}	64	96	128	192	256

N, L, and R are reported in samples, using a sample rate $f_s = 16000Hz$.

TABLE 2. Dimensions of the max-pooling layers in the convolutional stage, adapted for each resolution point.

Resolution	n_{mel}	Pooling sizes [time, mel]
T_{++}	64	[2, 1], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 2]
T_{+}	96	[2, 1], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 3]
BS	128	[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 2]
F_{+}	192	[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 3]
F_{++}	256	[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 4]

There is a total of seven max-pooling layers in the model, one after each convolutional layer. The total pooling factor is always $[4, n_{mel}]$.

filters, which along with the previous changes roughly doubles the frequency resolution.

4) **T+** (Intermediate point between BS and T_{++}). Analysis window of length $L = 96$ ms, window hop $R = 12$ ms. An intermediate number of Mel filters is used ($n_{mel} = 96$).
5) **F+** (Intermediate point between BS and F_{++}). Analysis window of length $L = 192$ ms, window hop $R = 24$ ms. An intermediate number of Mel filters is used ($n_{mel} = 192$).

IV. EXPERIMENTAL FRAMEWORK

Our experiments have been performed using a Convolutional Recurrent Neural Network (CRNN) based upon the Baseline System of DCASE 2020 Task 4. As in the baseline system, the features are extracted from the audio signals without a pre-processing stage such as a noise reduction module. In order to incorporate multi-resolution analysis into the model described by the Baseline, first we adapt the model to each resolution point, training a model for each resolution, and then we perform model fusion with the resulting models.

A. MODEL STRUCTURE AND TRAINING

Following the configuration of the Baseline System, the models are trained by means of the Mean Teacher method (described in Section II-C). The mel-spectrogram features are fed to the convolutional stage of the model, formed by seven 2D-convolutional layers with kernels of size 3×3. The number of filters is 16 for the first layer and is doubled in each layer until a maximum of 128. The activation function is the Gated Linear Unit (GLU).

Each convolutional layer is followed by a max-pooling. In order to reduce the different input mel-frequency sizes to a single dimension as done in the baseline system, the pooling size in the mel-frequency dimension is modified in the networks used for the different time-frequency resolution points, as specified in Table 2. The total pooling factor of the convolutional stage is 4 in the time dimension and n_{mel} in the mel dimension.

At the end of the convolutional stage, the input features have been shaped into a temporal sequence with length L which is fed to the recurrent stage of the model. Such recurrent stage is formed by two layers of bidirectional gated recurrent units (bi-GRU) with 128 units each. Finally, an attention pooling layer is applied with sigmoid activation, obtaining a temporal score sequence for each target category.

B. MODEL FUSION

We define a model fusion method that allows us to combine the scores generated by several CRNN models before further post-processing, such as thresholding or median filtering, with no additional parameter tuning. These models have to be trained individually beforehand, and they can employ different input features, namely the mel-spectrograms computed at different time-frequency resolution points as described in Section III, or could be as well trained with the same input features but different configurations.

For a given category i, a sound event detection system performs a classification between classes $\{\theta_{i,0}; \theta_{i,1}\}$, meaning absence or presence of event i, respectively. For each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088949, IEEE Access

classification task (i.e. for each target category), a detector j generates a score $s_i^{(j)}$ as a time series with a given time resolution or frame rate. By convention, lower scores show a stronger support to $\theta_{i,0}$, while higher scores support $\theta_{i,1}$. In the proposed systems, each of the scores is taken from the output of a sigmoid layer trained with a cross-entropy criterion. The scores are then between 0 and 1, and can be interpreted as the posterior probability of the presence of the event category i, $P(\theta_{i,1}|x) = 1 - P(\theta_{i,0}|x)$, where x is the observation of the audio segment.

The fusion procedure is performed as follows. In order to handle the different frame rates of score sequences generated with different feature resolutions, the first step is an interpolation. Let J be the number of models to combine, each score sequence $s_i^{(j)}$ is interpolated along the temporal dimension to fit a target frame rate, obtaining a set of J score sequences $(t_i^{(1)}, \ldots, t_i^{(J)})$. The target frame rate is chosen as the highest frame rate of the J model outputs to be combined. Then, the resulting sequences have N time frames each, being N the number of frames of the sequence with the highest frame rate: $t_i^{(j)} = (t_{i,1}^{(j)}, \ldots, t_{i,N}^{(j)})$.

Afterwards, the logit operator is applied frame-wise, in the following way:

$$l_{i,n}^{(j)} = \text{logit}(t_{i,n}^{(j)}) \equiv \log \frac{t_{i,n}^{(j)}}{1 - t_{i,n}^{(j)}} \tag{2}$$

Then, the logit scores of the model fusion, l_i are computed frame by frame as the average of the logit scores from each model j.

$$l_{i,n} = \frac{1}{J} \sum_{j=1}^{J} l_{i,n}^{(j)} \tag{3}$$

As a final step for the model fusion process, the sigmoid operator is applied to the resulting logit score sequences l_i, obtaining the final score sequences s_i for each category i. Then, these sequences are post-processed as described in Section IV-C, in order to obtain temporal predictions. The whole process is described in Fig. 6.

C. SCORE POST-PROCESSING

Evaluating a system by means of F1-score requires converting the score sequences s_i into timestamps which mark the start and the end of each event. For such purpose, defining a threshold th is necessary in order to obtain a set of predictions. The most common value for such threshold, used by the Baseline System, is $th_i = 0.5$ for every category i.

In the case that the posteriors $P(\theta_{i,1}|x)$ were properly computed (i.e. calibrated) and the prior probabilities of the evaluation set were $P(\theta_{i,1}|x) = P(\theta_{i,0}|x) = 0.5$, the previous approach would be the optimal decision in a Bayesian scenario. However, in the scenario of DCASE Challenge Task 4 the prior information of the 2020 Evaluation set was not known and could not be estimated reliably. Moreover, the cost of Bayes decisions and the F1-score are not comparable metrics, and therefore, even with an optimal Bayesian decision scenario, it is not guaranteed that the F1-score will be optimized. As a consequence, there is no reason whatsoever to support $th = 0.5$ as an adequate decision threshold.

Aiming at choosing a more optimal threshold for decision-making under the F1-score criterion, we tested two options:

1) Applying $th_i = 0.5$ for every event category, as done in the Baseline
2) Choosing the optimal threshold for each category empirically, as that which maximizes $F1$ over the Validation set.

It is worth noting that he choice of specific thresholds for each category does not apply in the case of PSDS metrics, since PSDS is not dependent on the threshold.

As a final stage, median filtering is applied to the binary scores with a window length of 450 ms. The purpose of median filtering is to clean impulsive peaks which are not representative of the presence or absence of acoustic events. The filtered binary vectors can be considered temporal predictions over which F1 or PSDS metrics can be computed.

V. RESULTS AND DISCUSSION

A. SINGLE-RESOLUTION RESULTS

In the first place, experiments were carried out using single-resolution systems with no model fusion involved. A model was trained five times with different random initializations for each one of the resolution points described in Table 1. The models are based on the Baseline System of DCASE 2020 Challenge Task 4, adapting the pooling sizes according to the time-frequency resolution as specified in Table 2. Taking into account that the BS resolution point coincides with the baseline system of DCASE Challenge 2020 Task 4, the results obtained using this resolution constitute the common benchmark for the aforementioned task.

The results of the single-resolution models are presented in Table 3 as the mean and the standard deviation of the F1-scores obtained with the five trainings. Observing the performances for each event category, it can be noted that different resolution points hold the best average result for different classes, supporting our hypothesis that different time-frequency resolutions are better suited to detect certain types of events.

Some classes achieve better results when employing higher temporal resolutions, for example *Dog*, *Blender*, or *Running water*, while the clearest tendency to achieve a better performance with a higher frequency resolution is shown by the category *Electric shaver/toothbrush*. The BS resolution point, used by the Baseline System, achieves the best result only for the category *Cat*.

B. MULTI-RESOLUTION RESULTS

The model fusion process described in Section IV-B allows to combine models trained at different time-frequency resolution points, thus obtaining multi-resolution systems. The goal is to achieve better performance thanks to the complementary information supplied by the different resolutions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088949, IEEE Access

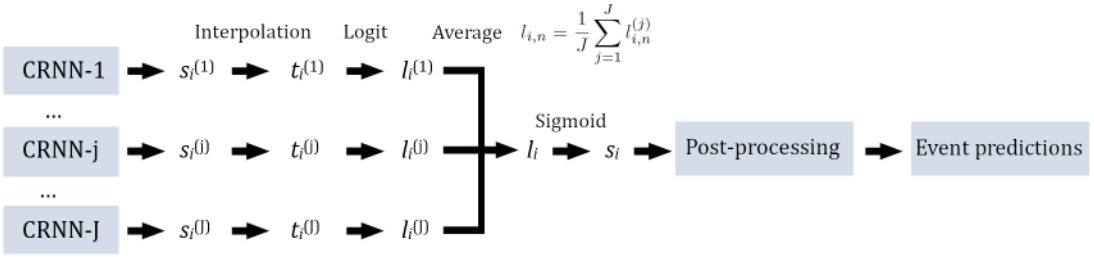

FIGURE 6. Block diagram describing the model fusion procedure for class i with J models. After the interpolation step, each sequence $t_i^{(j)}$ is N time frames long, where N is the length of the longest sequence $s_i^{(j)}$. The logit, average, and sigmoid operations are performed for each time frame $n \in [1, N]$. Therefore, the resulting sequence s_i is also N time frames long.

TABLE 3. Event-based F1-score (%) over the Validation set for each event category obtained with different time-frequency resolution working points.

	T_{++}	T_{+}	BS	F_{+}	F_{++}
Alarm bell / ringing	42.1 ± 1.5	**43.8** ± 2.1	42.0 ± 1.4	42.2 ± 3.1	41.0 ± 2.0
Blender	**32.9** ± 3.2	32.3 ± 1.4	27.4 ± 1.6	30.0 ± 2.6	30.9 ± 3.9
Cat	38.4 ± 1.8	40.0 ± 1.8	**41.0** ± 2.1	39.3 ± 3.9	34.7 ± 2.3
Dishes	20.8 ± 1.5	21.9 ± 1.1	20.8 ± 2.1	**22.6** ± 1.7	21.0 ± 1.2
Dog	15.1 ± 0.7	**17.1** ± 2.6	16.5 ± 1.0	12.3 ± 1.1	12.8 ± 2.7
Electric shaver / toothbrush	32.8 ± 4.2	35.5 ± 4.7	37.2 ± 2.9	36.2 ± 5.4	**41.1** ± 2.9
Frying	23.5 ± 2.2	**23.9** ± 2.3	20.9 ± 4.8	**23.9** ± 2.2	22.2 ± 2.6
Running water	**31.7** ± 3.3	29.8 ± 2.2	30.4 ± 2.6	27.6 ± 1.8	27.2 ± 1.6
Speech	42.7 ± 3.1	**47.1** ± 2.9	45.2 ± 1.5	46.2 ± 2.6	46.3 ± 1.8
Vacuum cleaner	40.1 ± 1.7	39.9 ± 2.3	38.9 ± 3.3	**44.5** ± 4.1	40.1 ± 5.0
Total macro	32.0 ± 1.3	**33.1** ± 0.9	32.0 ± 1.1	32.5 ± 1.5	31.7 ± 1.0

Mean ± standard deviation computed across 5 trainings with random initializations.

Following this idea, two multi-resolution models are proposed: a three-resolution model (*3res*) which combines the baseline resolution BS with the resolution points T_{++} and F_{++}, and a five resolution model (*5res*) combining the five resolution points defined (BS, T_{++}, F_{++}, T_{+}, and F_{+}). To obtain these models, the model fusion procedure is employed with one model per resolution point. Thus, *3res* is a combination of three single-resolution models, and *5res* is a combination of five single-resolution models. As shown in Table 4, both multi-resolution systems outperform the single-resolution systems in terms of macro-averaged F1 over the Validation set, with *5res* obtaining a higher performance (39.8%) than *3res* (38.2%).

However, it is necessary to determine whether the improvements in performance are due to the combination of several resolutions rather than solely to the combination of different models. In order to achieve this, an additional combined model is proposed which performs a model fusion with five models trained with the BS resolution point and different initializations (*5×BS*). Such combined model was found to outperform the individual models trained with the same resolution in terms of macro-averaged F1 over the Validation set, but with a lower performance (36.9%) than *3res* or *5res*. Therefore, we conclude that, although model fusion allows to improve performance even in a single-resolution setting, the multi-resolution approach is able to obtain further improvements in terms of macro F1-score.

Aiming to compare the results of single-resolution and multi-resolution models, the F1-scores obtained by the combined models (*3res*, *5res*, and *5×BS*) over the Validation set are presented in Table 4, next to those obtained using the BS resolution point (previously presented in Table 3). The improvements are consistent in every category when increasing the number of resolution points involved.

In terms of the mean macro-averaged F1-score computed across five random initializations of each system, the *3res* model obtains 6.2 points more than the single-resolution model BS. In the case of the *5res* model, an improvement of 1.6 points is observed with respect to *3res*, which makes a total improvement of 7.8 points with respect to the macro-averaged F1-score achieved by the BS model. Moreover, some categories seem to benefit of the multi-resolution analysis more than others: it is the case of *Blender*, which obtains 27.4% with BS and 44.6% with *5res*, or *Vacuum cleaner*, which obtains 38.9% with BS and 54.6% with *5res*. According to the single-resolution results shown in Table 3, the BS resolution is the least fitted to the detection of these two types of events, which would explain a higher impact of multi-resolution in these categories.

Regarding the PSDS metrics, the multi-resolution analysis has been found to achieve improvements as well. The PSDS has been computed for the single-resolution models trained

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088949, IEEE Access

TABLE 4. Event-based F1-score (%) results over the Validation set.

Event Category	BS	3res	5res	5×BS
Alarm bell / ringing	42.0±1.4	46.0±0.8	**46.9**±0.8	43.7±1.1
Blender	27.4±1.6	43.2±4.3	**44.6**±3.2	37.6±0.6
Cat	41.0±2.1	43.7±1.9	**44.8**±0.5	41.6±0.6
Dishes	20.8±2.1	23.4±1.5	**24.7**±1.0	23.3±0.9
Dog	16.5±1.0	18.1±1.5	**18.9**±1.7	18.3±0.7
E.shaver/toothbrush	37.2±2.9	43.8±1.8	**45.8**±3.2	41.2±0.5
Frying	20.9±4.8	28.0±3.8	**29.6**±2.5	23.6±3.5
Running water	30.4±2.6	36.7±3.1	**38.5**±1.4	36.0±1.4
Speech	45.2±1.5	47.3±1.5	**48.9**±1.1	47.5±0.6
Vacuum cleaner	38.9±3.3	51.4±3.5	**54.6**±2.9	44.7±1.6
Total macro	32.0±1.1	38.2±1.1	**39.8**±0.8	35.8±0.7

Mean ± standard deviation computed across 5 random initializations of each system.

TABLE 5. PSDS, PSDS cross-trigger, and PSDS macro results over the Validation set.

	α_{ct}	α_{st}	BS	3res	5res	5×BS
PSDS	0	0	0.584	0.657	**0.666**	0.635
PSDS cross-trigger	1	0	0.498	0.595	**0.609**	0.564
PSDS macro	0	1	0.400	0.467	**0.479**	0.451

α_{ct} is the weight related to the cost of cross-trigger. α_{st} is the weight related to the cost of instability across classes.

with the BS resolution points as well as for the combined models *3res*, *5res*, and *5×BS*, using the configuration proposed for the DCASE Challenge 2020 Task 4. The results over the Validation set are shown in Table 5, showing that the PSDS scores obtained are higher when using more different resolution points. Such effect is observed in the three configurations: PSDS, PSDS cross-trigger, and PSDS macro. Moreover, the combination of five models with the same resolution (*5×BS*) achieves a higher PSDS performance than each individual model (BS) on its own, but does not reach the results of multi-resolution models *3res* and *5res*.

C. SCORE THRESHOLDING RESULTS

Whereas the experiments described in sections V-A and V-B employ a threshold $th = 0.5$ for every category, additional experiments have been carried out aiming to determine the adequacy of such approach for F1-score decisions. For this purpose, we have taken the *5res* model as a starting point, and we have studied its performance in each category in terms of event-based F1-score using different values for the threshold, considering 50 values linearly distributed from $th = 0$ to $th = 1$. It should be noted that this analysis does not affect the PSDS performance, which is not dependent on the threshold value.

Following this procedure, we obtain the results presented in Fig. 7. Observing the F1 curves, it can be observed that the optimal value of the threshold usually differs from $th = 0.5$. Moreover, the election of th affects the performance differently in each category.

Aiming to improve the performance by tuning the values of the thresholds for each category, we define a new model, *5res-thr*, which is based upon the *5res* model but uses a specific threshold th_i for each category, instead of a global threshold $th = 0.5$. The thresholds are chosen as those which maximize the F1 performance for each category over the Validation set. Following this criterion, the resulting thresholds are listed in Table 6. The performances of *5res* and *5res-thr* over the Validation set are compared in Table 7, where it is shown that the choice of specific thresholds allows to increase the performance in every category, and up to 2.5 points in F1 macro. However, it should be noted that the results of *5res-thr* over the Validation set represent the best-case scenario, where we know the optimal thresholds, whereas these optimal values differ from one dataset to another. In order to test the threshold tuning approach in a more realistic scenario, we have compared the performances of *5res* and *5res-thr* over a different dataset, the Public Evaluation set, showing the results in Table 8. Although the performance does not increase in every category, the *5res-thr* achieves a higher overall F1-score. Additionally, *5res* and *5res-thr* models were submitted to the DCASE Challenge 2020 Task 4, both outperforming the Baseline System and also obtaining a higher overall F1 performance by using class-specific thresholds [42].

TABLE 6. Binarization thresholds used in the *5res-thr* system.

	Threshold
Alarm bell / ringing	0.31
Blender	0.49
Cat	0.65
Dishes	0.31
Dog	0.69
E. shaver / toothbrush	0.61
Frying	0.29
Running water	0.45
Speech	0.83
Vacuum cleaner	0.65

TABLE 7. Event-based F1-score (%) results of a multi-resolution model with global thresholding $th = 0.5$ (5res) and with specific thresholds for each class (5res-thr) over the Validation set.

Event Category	5res	5res-thr
Alarm bell / ringing	47.2	**48.2**
Blender	49.5	**50.0**
Cat	45.2	**47.3**
Dishes	23.9	**25.2**
Dog	18.6	**22.3**
E. shaver / toothbrush	46.8	**49.0**
Frying	29.7	**34.3**
Running water	39.6	**41.6**
Speech	49.9	**55.6**
Vacuum cleaner	58.7	**61.0**
Total macro	40.9	**43.4**

D. EVENT OVERLAP ANALYSIS

When tackling the problem of polyphonic sound event detection, it is possible to encounter multiple event categories coinciding in time. Generally, such overlap constitutes a particularly challenging scenario for sound event detectors, because one of the events can mask the others, making them more difficult to recognize and thus producing False Negative

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088949, IEEE Access

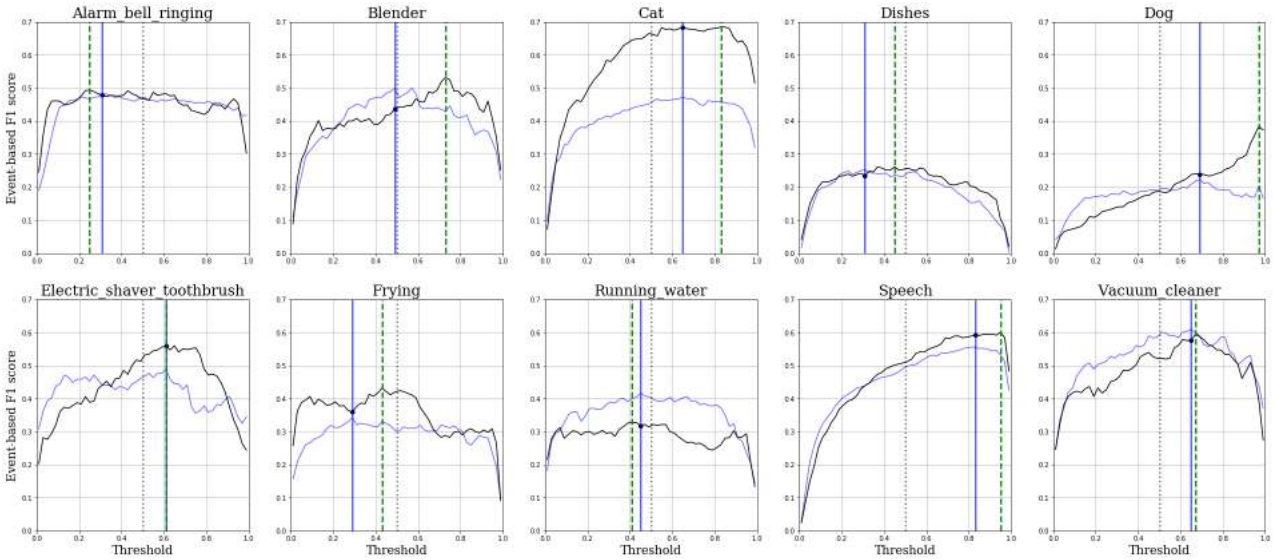

FIGURE 7. F1-scores obtained by the *5res* model in each category with different threshold values. The blue curve represents the F1 performance over the Validation set, marking the best result with a blue vertical line. The black curve represents the performance in the Public evaluation set. The intersection of the blue line and the black curve indicates the performance over the Public evaluation set using the threshold chosen with the Validation set. The optimal threshold for the Public Evaluation set is marked with a green vertical dashed line, whereas the default threshold (0.5) is marked with a grey vertical dotted line.

TABLE 8. Event-based F1-score (%) results of a multi-resolution model with global thresholding $th = 0.5$ (5res) and with specific thresholds for each class (5res-thr) over the Public Evaluation set (eval 2019).

Event Category	5res	5res-thr
Alarm bell / ringing	46.5	**47.8**
Blender	**44.2**	43.7
Cat	66.2	**68.4**
Dishes	**25.4**	23.5
Dog	18.7	**23.9**
E. shaver / toothbrush	53.2	**56.0**
Frying	**40.6**	36.0
Running water	31.8	**31.9**
Speech	51.3	**59.3**
Vacuum cleaner	52.8	**57.6**
Total macro	43.0	**44.8**

errors.

However, taking into account that different resolutions are more fitted to capture certain event classes, the proposed multi-resolution approach should be able to improve the results for overlapped events.

In order to compare the performance of our systems in the scenario of event overlap, we have divided the DESED Validation set into two subsets: a non-overlapped subset, containing the events that do not coincide in time with any other class, and an overlapped subset, which contains the events which occur at the same time than other categories in their entirety. Given that the main problem of overlapped events is the appearance of False Negative errors, we have studied the event-based Recall metric (R), which is the component of the event-based F1-score that is affected by false negatives:

$$R = \frac{TP}{TP + FN} \tag{4}$$

The results of macro-averaged recall over the Validation set and both subsets are presented in Table 9. It can be observed that the recall metric is consistently lower for the overlapped subset than for the complete Validation set. On the other hand, the recall metric when considering only non-overlapped events is very similar to that of the Validation set, which was the expected result considering that the non-overlapped subset constitutes the majority of the Validation set. In every case, the multi-resolution approach provides better results considering the mean recall of five systems with different random initializations.

In terms of relative improvement, the impact of multi-resolution is more accentuated in the overlapped subset. With the *3res* model, the mean recall increases from 10.6% to 13.5%, which constitutes a 27.8% relative improvement. The mean recall over this subset experiments a further increase when using five resolutions, reaching 14.8%, a 40.0% relative improvement with respect to the BS system.

In contrast, the relative improvement is considerably lower when considering the whole validation set (13.1%) or only the non-overlapped subset (12.6%). This fact seems to suggest a larger impact in the correct detection of overlapped events.

However, these results need to be taken with caution, since the analysis was limited by the lack of overlapped events in most of the classes, as shown in Table 10. Moreover, the improvement in mean recall in the case of overlapped events does not seem statistically significant, given the standard deviations found in the multi-resolution results with overlapped events.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3088949, IEEE Access

TABLE 9. Event-based recall (%) results of the systems over the Validation set, the Non-overlapped Validation subset, and the Overlapped Validation subset. Relative improvement is given with respect to the BS model. The highest relative improvement for each system is highlighted in bold.

Model	Macro-averaged Recall (%)					
	Validation	Rel. Imp. %	Non-overlapped	Rel. Imp. %	Overlapped	Rel. Imp. %
BS	36.1±0.8	-	36.4±0.5	-	10.6±0.9	-
3res	39.4±1.1	9.2	39.4±1.1	8.4	13.5±5.6	**27.8**
5res	40.9±0.8	13.1	40.9±0.7	12.6	14.8±5.6	**40.0**

Mean ± standard deviation computed across 5 random initializations of each system.

TABLE 10. Number of events included in the Validation set, the Non-overlapped Validation subset, and the Overlapped Validation subset for each target category.

Event Category	Validation	No overlap	Overlap
Alarm bell / ringing	420	412	3
Blender	96	88	0
Cat	341	337	1
Dishes	567	403	162
Dog	570	551	16
E. shaver / toothbrush	65	54	0
Frying	94	92	0
Running water	237	226	3
Speech	1754	1396	338
Vacuum cleaner	92	82	1
Total	4236	3641	524

E. RESOURCE ANALYSIS

In Section V-B, it has been shown that the combination of multiple resolutions is able to provide improvements in terms of F1-scores. However, it would be relevant to know the impact of multi-resolution in terms of execution times.

For this reason, we have measured the time required by the baseline system (BS) and by the multi-resolution systems *3res* and *5res* to perform the feature extraction process and generate predictions for the DESED Validation set (181 minutes of audio).

Each system has been run five times, using 15 CPU cores for feature extraction and a Nvidia GeForce RTX 2080 GPU for the forward pass of the neural networks. Averaging the five executions, we have computed a $0.02\times$ real time factor for the BS model. In the same manner, we have measured a $0.04\times$ factor for the *3res* model and a $0.07\times$ factor for the $5res$ model.

It can be observed that the increase of the execution time is lower than the number of resolutions. This is due to the existence of two different stages in the test process. The first stage is repeated for each resolution, and consists of the mel-spectrogram feature extraction and the forward pass of the CRNN, after which the score sequences for each resolution are obtained. Afterwards, the scores from each model are averaged and binarized by means of a threshold, and the F1 metrics are computed: this process is performed only once, regardless of the number of resolutions involved.

F. FEATURE ANALYSIS

Aiming to give insights into the different temporal and spectral characteristics that make a given event category more adequate for a certain resolution point, we have studied the variations in time and frequency of the mel-spectrogram features in the DESED Validation set.

As a first step for this analysis, we have selected the mel-spectrogram features of the events that are not overlapped in time with any other in-domain event, in order for overlapped events not to interfere in the analysis of other categories. Rather than the entire audio segments, we have only considered the relevant time interval for each event, i.e., from the onset time to the offset time.

For these mel-spectrograms, we have obtained the first differences (Δ-features) in each axis, which indicate the change of energy with respect to the previous time step (in the time axis) or the adjacent mel-filter (in the frequency axis). Let them be called Δt and Δf. In order to obtain the most reliable measures in each axis, Δt has been computed from the T_{++} resolution with a time step of four frames, which corresponds to a 50% overlap of the analysis window of feature extraction, and Δf has been computed from the F_{++} resolution.

We have computed the standard deviation of the Δ-features between consecutive temporal frames ($\sigma_{\Delta t,i}$) and between adjacent mel filters ($\sigma_{\Delta f,i}$) for each event category i. A higher standard deviation means that the variations in the corresponding axis are larger. Aiming to obtain a measure which determines whether the variations in time or in frequency are predominant for a certain category, we have computed the ratio r_i between the mean values of the standard deviations for each category i:

$$r_i = \frac{\bar{\sigma}_{\Delta t,i}}{\bar{\sigma}_{\Delta f,i}} \tag{5}$$

A higher r_i implies that the corresponding event category has its predominant variations in the time axis. Therefore, a higher time resolution should, in principle, be able to capture such events with more detail. In Table 11, the ratios r_i for each category are presented. When comparing these values with the best performing resolution point for each category in Table 3, it can be observed that the only category that performs best with F_{++}, *Electric shaver/toothbrush*, presents the lowest ratio (0.31). Additionally, the class with the highest ratio, *Dog* (0.71), obtains its best performance using the T_{+} resolution.

However, the relationship between the ratio value and the best performing resolution point is not perfect for every category. In fact, the analysis can be complemented with the average length of each event category, which was already described in Section III and Fig. 3. Those categories that

TABLE 11. Value of the ratio $r_i = \bar{\sigma}_{\Delta t,i}/\bar{\sigma}_{\Delta f,i}$ for each category.

Event Category	r_i
Alarm bell / ringing	0.56
Blender	0.59
Cat	0.52
Dishes	0.61
Dog	0.71
E. shaver / toothbrush	0.31
Frying	0.57
Running water	0.59
Speech	0.58
Vacuum cleaner	0.58

present long events, such as *Frying* or *Vacuum cleaner*, obtain better results with the F_+ resolution, while short-duration categories like *Speech* or *Alarm bell/ringing* perform better with T_+.

VI. CONCLUSION

In this work we present a method to better modelling the different temporal and spectral characteristics of sound events in the task of Sound Event Detection. We hypothesize that features extracted using different time-frequency resolution parameters are able to represent certain event categories in a more recognizable way. Hence, in contrast to most current approaches which use a single time-frequency resolution during the feature extraction process, we propose combining the information from several resolution points to improve the performance of the detectors.

In order to test our hypothesis, we take as a starting point the Baseline System of DCASE 2020 Task 4, which consists of a Convolutional-Recurrent Neural Network that is trained using mel-spectrogram features. By training this system with different feature resolutions, we observe that each sound event category obtains a higher performance at different time-frequency resolution points. This supports our idea that different resolutions are more suited to represent certain sound event classes.

Afterwards, aiming to combine the information of each resolution point into a multi-resolution system, a model fusion procedure is defined that operates over the scores of the CRNNs. We obtain the final scores as the average of the scores of each individual model, without additional trainable parameters. Such process could be applied to other score-based systems.

Using the DESED Validation set to test the performance of the systems, we find that multi-resolution models are able to outperform single-resolution models in every category in terms of event-based F1-score, and also in terms of the PSDS metric, with longer execution times, but still much faster than real-time performance.

Additionally, we have explored the impact of the threshold used to define the event predictions, finding that its usual value $th = 0.5$ is not necessarily the optimal setting. We are able to improve the performance of a multi-resolution model by choosing specific thresholds for each category using the DESED Validation set. Although the optimal thresholds change when using a different dataset, the specific thresholds obtain a better result in terms of macro-averaged F1-score over the DESED Public Evaluation set.

REFERENCES

[1] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, 2019. [Online]. Available: https://www.mdpi.com/2073-8994/11/8/1018

[2] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19 143–19 165, 2019.

[3] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[4] A. Lozano-Diez, O. Plchot, P. Matejka, and J. Gonzalez-Rodriguez, "DNN Based Embeddings for Language Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5184–5188.

[5] J. Nam, K. Choi, J. Lee, S. Chou, and Y. Yang, "Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from Bach," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 41–51, 2019.

[6] M. Schedl, "Deep learning in music recommendation systems," *Frontiers in Applied Mathematics and Statistics*, vol. 5, p. 44, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fams.2019.00044

[7] D. de Benito-Gorron, A. Lozano-Diez, D. T. Toledano, and J. Gonzalez-Rodriguez, "Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, p. 9, 2019.

[8] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[9] T. K. Chan and C. S. Chin, "A comprehensive review of polyphonic sound event detection," *IEEE Access*, vol. 8, pp. 103 339–103 373, 2020.

[10] 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 *Proc. IEEE ICASSP 2019*, Brighton, UK, 2019.

[11] 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.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR09*, 2009.

[13] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[14] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 69–73. [Online]. Available: https://arxiv.org/abs/1807.09902

[15] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," in *Submitted to DCASE2019 Workshop*, NY, USA, 2019.

[16] N. Turpault, R. Serizel, A. Parag 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*, New York City, United States, Oct. 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[17] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2014.

[18] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*, 2016. [Online]. Available: http://dx.doi.org/10.1145/2964284.2964310

[19] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proceedings of the 18th Interna-*

tional Society for Music Information Retrieval Conference (ISMIR 2017), Suzhou, China, 2017, pp. 486–493.

[20] R. Serizel and N. Turpault, "Sound event detection from partially annotated data: Trends and challenges," in *IcETRAN conference*, Srebrno Jezero, Serbia, Jun. 2019. [Online]. Available: https://hal.inria.fr/hal-02114652

[21] W. Zhang, G. Zelinsky, and D. Samaras, "Real-time accurate object detection using multiple resolutions," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[22] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 241–254.

[23] S. Park, Y. Jeong, and H. S. Kim, "Multiresolution CNN for reverberant speech recognition," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–4.

[24] A. V. Oppenheim, "Speech spectrograms using the Fast Fourier Transform," *IEEE Spectrum*, vol. 7, no. 8, pp. 57–62, 1970.

[25] N. Turpault and R. Serizel, "Training sound event detection on a heterogeneous dataset," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 200–204.

[26] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, May 2016. [Online]. Available: http://dx.doi.org/10.3390/app6060162

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

[28] N. Turpault, R. Serizel, A. Parag 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*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[29] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, 2020. [Online]. Available: https://hal.inria.fr/hal-02355573

[30] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.

[31] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, B. Van den Bergh, 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 *Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, 2017.

[32] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013.

[33] L. JiaKai, "Mean teacher convolution system for DCASE 2018 task 4," DCASE2018 Challenge, Tech. Rep., September 2018.

[34] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[35] DCASE 2020 task 4: Sound event detection and separation in domestic environments - Challenge results, http://dcase.community/challenge2020/task-sound-event-detection-and-separation-in-domestic-environments-results.

[36] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 task 4," 2019.

[37] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning convolution system for DCASE 2019 task 4," in *Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, p. 134.

[38] J. Yan, Y. Song, L. Dai, and I. McLoughlin, "Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 326–330.

[39] S. Lee and M. Kim, "Waveform-based end-to-end deep convolutional neural network with multi-scale sliding windows for weakly labeled sound event detection," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2020, pp. 182–186.

[40] Y. Liu, J. Tang, Y. Song, and L. Dai, "A capsule based approach for polyphonic sound event detection," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1853–1857.

[41] D. T. Toledano, M. P. Fernández-Gallego, and A. Lozano-Diez, "Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on timit," *PloS one*, vol. 13, no. 10, 2018.

[42] D. de Benito-Gorron, D. Ramos, and D. T. Toledano, "A multi-resolution approach to sound event detection in DCASE 2020 task4," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 36–40.

DIEGO DE BENITO-GORRÓN obtained the Telecommunication Technology and Service Engineering Degree in 2017 and the ICT Research and Innovation Master's Degree in 2018, both at the Escuela Politécnica Superior of Universidad Autonóma de Madrid (UAM), Spain. He has been recipient of the award to the best academic record in Telecommunication Technology and Service Engineering.

Since 2017, he has been working as a Research Assistant within AUDIAS Group at UAM, where he is currently pursuing his Ph.D. degree, focusing on automatic speech recognition and polyphonic sound event detection.

DR. DANIEL RAMOS finished his Ph.D. in 2007 in Universidad Autonoma de Madrid (UAM), Spain. From 2011, he is an Associate Professor at the UAM. He is a staff member of AUDIAS Group. During his career, he has visited several research laboratories and institutions around the world, including the Institute of Scientific Police at the University of Lausanne (Switzerland), the School of Mathematics at the University of Edinburgh (Scotland), the Electrical Engineering school at the University of Stellenbosch (South Africa), and more recently the Netherlands Forensic Institute and the Computational and Biological Learning Lab of the University of Cambridge. He has been visiting professor at the Universidad de Buenos Aires in 2019.

His research interests are focused on forensic evaluation of the evidence using Bayesian techniques, probabilistic calibration, validation of forensic evaluation methods, speaker and language recognition and, more generally, signal processing and pattern recognition. Dr. Ramos is actively involved in the research of development of different aspects of forensic science, including the statistical evaluation of speech and chemical evidence (mainly glass). He has been invited by the NIST to several workshops, including the OSAC standardization initiative. He is author of multiple publications in national and international journals and conferences, some of them awarded. He has also participated in several international competitive evaluations of speaker and language recognition technology, since 2003. Recently, he is working on signal processing and machine learning for industrial applications in the energy sector. Dr. Ramos is regularly a member of scientific committees in different international conferences, and he is often invited to give talks in conferences and institutions.

PROF. DOROTEO T. TOLEDANO received the M.S. degree in 1997, and the Ph.D. in Electrical and Electronic Engineering in 2001, both from Universidad Politecnica de Madrid, Spain. He has been recipient of several academic awards, such as the First National Bachelor Award of Spain, the best academic record in Electrical and Electronic Engineering and a Ph.D. Dissertation Award from the Spanish Association of Telecommunication Engineers.

After his Ph.D., he joined M.I.T. as Postdoctoral Research Associate in the Spoken Language Systems Group (2001-2002), under the supervision of Profs. Victor Zue and James Glass. He has also experience working in the industry, in particular in the Speech Technology Division of Telefonica R&D, where he worked from 1994 to 2001 and also in 2003. His trajectory as professor in signal processing starts in 2004, when he joined Universidad Autonoma de Madrid, where he is currently Full Professor. Prof. Toledano has over 25 years of experience in speech processing, over 100 scientific publications. He has participated in 6 EU research projects and in over 40 national projects (in 10 of them as principal investigator). He has participated in over 15 technological competitive evaluations (mainly NIST evaluations) and has organized three. He was General Co-Chair and main organizer of IberSPEECH 2012, and organizer and session chair of several other conferences. Prof. Toledano current research is focused on audio, speech, speaker and language recognition. Since July 1st 2018 he is the new Director of the AUDIAS research group.