Enhancing Conformer-Based Sound Event Detection Using Frequency Dynamic Convolutions and BEATs Audio Embeddings

Sara Barahona, Diego de Benito-Gorrón, Doroteo T. Toledano, and Daniel Ramos

***Abstract*—Over the last few years, most of the tasks employing Deep Learning techniques for audio processing have achieved state-of-the-art results employing Conformer-based systems. However, when it comes to sound event detection (SED), it was scarcely used after it won the DCASE Challenge 2020 Task 4. In previous research, we found that Conformer-based systems achieved a higher performance in terms of sound events classification compared to other architectures frequently employed, such as Convolutional Recurrent Neural Networks (CRNNs). Given that the second scenario proposed for the Polyphonic Sound Detection Score (PSDS2) is focused on avoiding confusion between classes, in this paper we propose to optimize a Conformer-based system to maximize the performance on this scenario. For this purpose, we performed a hyperparameter tuning and incorporated recently proposed Frequency Dynamic Convolutions (FDY) to enhance its classification properties. Additionally, we employed our previously proposed multi-resolution approach not only to enhance the performance but also to gain a deeper understanding of the Conformer architecture for SED, analyzing its advantages and disadvantages, and finding possible solutions to them. Additionally, we explored the integration of embeddings from the pre-trained model BEATs, an iterative framework to learn Bidirectional Encoder representation from Audio Transformers. By concatenating these embeddings into the input of the Conformer blocks, results were further improved, achieving a PSDS2 value of 0.813 and considerably outperforming SED systems based on CRNNs.**

***Index Terms*—Sound event detection, conformer, DCASE challenge, PSDS, multi-resolution, BEATs.**

I. Introduction

SOUND event detection (SED) is an active research topic that focuses on localizing and classifying relevant sound events within an audio clip. In recent years, there has been a growing interest in this field due to its diverse range of applications in various domains, including bioacoustics [1], [2], medical support [3], [4], urban sound monitoring [5] or industrial monitoring [6]. The progress in SED techniques has been significantly facilitated by notable contributions, such as the publication of Google AudioSet [7] and the yearly challenges and workshops organized by the DCASE community [8]. Among the diverse tasks that constitute the DCASE Challenge, it is of special interest Task 4A titled "Sound Event Detection with Weak Labels and Synthetic Soundscapes". This task involves evaluating SED systems by employing both real and synthetic recordings which contain 10 sound event classes that can be found in a domestic environment. Besides, it also addresses the challenge of working with unlabeled data as well as two different types of annotations: strong labels that provide temporal information (timestamps) along with the sound event category, and weak labels that solely indicate the category.

Originally, SED systems were evaluated by employing the event-based and the segment-based F1-score [9]. However, these metrics have limitations due to their reliance on a single operating point and susceptibility to human subjectivity when labeling sound events in time, which can significantly impact a model's performance since collar-based metrics rely on the onset and offset of sound events. To address these issues and achieve a more robust evaluation, the Polyphonic Sound Detection Score (PSDS) [10] was introduced, by measuring the intersection between detected and annotated sound events. Additionally, PSDS employs different decision thresholds to evaluate SED systems, obtaining a fairer metric as it is calculated as the area under the Polyphonic Sound Detection Receiver Operating Characteristic (PSD-ROC) curve. An additional advantage of this metric is that its parameters can be adjusted to evaluate different properties of a SED system. As a result, the DCASE Challenge Task 4A proposes two evaluation scenarios. Whereas the first one (PSDS1) emphasizes a fast reaction upon a sound event requiring a highly accurate localization, the second scenario (PSDS2) aims to minimize confusion between classes and is less strict about timing errors.

The Convolutional-Augmented Transformer (Conformer) [11] has achieved state-of-the-art (SOTA) results in several audio-related tasks including automatic speech recognition (ASR) [12] and automatic speaker verification (ASV) [13]. In the field of SED, it achieved promising results by winning the DCASE Challenge Task 4 in 2020 [14], obtaining the highest F1-score. Although this architecture was further explored with the 2020 DCASE Challenge setup [15], subsequent editions

Manuscript received 1 February 2024; revised 14 June 2024; accepted 2 August 2024. Date of publication 15 August 2024; date of current version 2 September 2024. This work was supported in part by FPI under Grant PRE2022-104808, from FSE+ and in part through Project PID2021-125943OB-I00 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE from the Spanish Ministerio de Ciencia e Innovacion and Fondo Europeo de Desarrollo Regional. The associate editor coordinating the review of this article and approving it for publication was Dr. romain serizel. *(Corresponding author: Sara Barahona.)*

The authors are with the AUDIAS Research Group, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Madrid, Spain (e-mail: sara.barahona@uam.es; diego.benito@uam.es; doroteo.torre@uam.es; daniel.ramos@uam.es).

Digital Object Identifier 10.1109/TASLP.2024.3444490

revealed performance inconsistencies, particularly when non-target events were incorporated into synthetic data and the Polyphonic Sound Detection Score (PSDS) was introduced as the primary evaluation metric. Consequently, current SED systems most often employ Convolutional Recurrent Neural Networks (CRNNs) as architecture backbone [16] with the addition of the recently proposed Frequency Dynamic Convolutions (FDY) [17]. A simple and straightforward technique for enhancing PSDS2 is to perform weak predictions over the entire sequence [18], by using the full duration of the audio as timestamps. However, this approach significantly reduces performance in the first PSDS scenario, obtaining values of PSDS1 close to 0, indicating that the localization of the sound events is completely missed. Furthermore, there is a growing popularity in employing embeddings from pre-trained models in AudioSet to improve performance. By this means, for the DCASE Challenge 2023 Task 4A, a baseline that exploits embeddings from the pre-trained model BEATs [19] was introduced.

This work explores in depth the advantages and disadvantages of the Conformer model for SED compared to the reference architecture in SED, the CRNN. Acknowledging the Conformer's difficulties in providing accurate timestamps (i.e. optimizing the PSDS1) [20], we decided to optimize the Conformer architecture only for PSDS2. In this paper, we present the different steps that we have followed to optimize this network, as well as a comprehensive analysis over the two scenarios proposed for PSDS, to get insights into the benefits and disadvantages of the Conformer over the most common architectures. This optimization has enhanced our understanding of the Conformer's capabilities and limitations. However, we found that by combining the Conformer architecture with our previously proposed multi-resolution approach [21], [22] we could gain a better understanding of the reasons behind the limited time resolution of the outputs of the Conformer model, as well as possible ways to compensate for it. Additionally, employing our multi-resolution approach along with the attention matrices at different depths of the Conformer allowed us to check if the temporal resolution is progressively being lost across the network, and have tried combining the outputs of the different layers to reduce this effect. All these experiments, along with the use of Frequency Dynamic Convolutions and the integration of BEATs embeddings, have allowed us not only to better understand the Conformer in the task of SED, but also to obtain a final PSDS2 comparable to the best non-ensemble systems in the DCASE Challenge 2023 Task 4A.

The structure of this paper is as follows. In Section II, we introduce an overview of the Conformer block and the main techniques employed for enhancing its classification properties and analyzing its insights. Our proposed optimization and the analysis over different resolution points are described in Section III, as well as the experimental setup followed. In Section IV, we present the ablation study of our optimization, analyzing and discussing the results using our multi-resolution approach. Additionally, we conduct a comparative analysis of the advantages and disadvantages of our optimized Conformer and CRNNs within the context of the DCASE Challenge Task 4A, highlighting the benefits of using BEATs embeddings. Lastly, conclusions are presented in Section V.

II. Related Work

A. Conformer Block

Sound events own diverse temporal and spectral characteristics. Therefore, exploiting both local and temporal features is crucial for accurately detecting and classifying every class event. Convolutional Neural Networks (CNNs) [23] excel in extracting local patterns but lack at exploring global representations. In contrast, Transformer [24] networks are proficient in modeling long-term dependencies thanks to the utilization of self-attention but lack sensitivity to local details. For this purpose, CNNs and Transformers are combined into a single architecture named Conformer (Convolutional-Augmented Transformer), to capture both local and global features.

The attention function employed in the Transformer and Conformer is defined by three vectors (queries, keys and values), usually of the same dimension (d_k), obtained by linearly projecting the input. Whereas the query vector represents the position from the sequence for which we seek information, the key vector indicates the relevance of each position in the input sequence to a specific query. By computing the dot product between queries and keys, we can obtain the similarity score of this interaction, known as the attention score. This score is scaled by the key dimension d_k and transformed into a probability distribution using softmax, as shown in (1). These attention scores are used to weight the value vectors, which represent the actual values of the input sequence. Higher attention scores imply greater importance of the corresponding values. By employing this mechanism, the model can consider the entire input sequence while focusing on the relevant parts determined by the attention scores.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Equation (1) describes the attention mechanism used in both the Transformer and Conformer. In many cases queries, keys and values are all just the input vectors. In these cases the attention mechanism is called self-attention. This attention mechanism is rarely employed alone. It is more common to apply (1) to different linear projections, giving rise to the Multi-Head Self Attention (MHSA) mechanism [24] which allows to simultaneously learn various attention functions by employing multiple heads. By this means, each head can attend to different patterns in the input sequence, allowing the model to capture complex relationships between different keys and queries. Additionally, instead of employing absolute positional encoding as in the original Transformer [24], the Conformer integrates a relative positional encoding into the MHSA module, as proposed in the Transformer-XL [25]. This encoding strategy injects the temporal information in the attention score of each function, by calculating the relative distances between each key and query, from which also absolute positions can be recovered. This method improves the generalization of MHSA across sound event sequences of varying lengths, making the model more robust to different utterance durations.

The convolution module is constructed using a combination of point-wise and 1D depth-wise convolution blocks, which helps

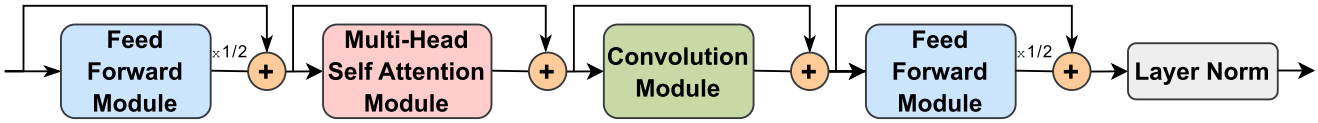


Fig. 1. Structure of Conformer block. The sandwiched structure is created by dividing the feed-forward module into two blocks with half-step residual connections, encapsulating both multi-head self-attention and convolution modules. A layer normalization is added at the end of the block.

decrease the computational cost while achieving the same effects of traditional convolution operation. The feed-forward module of the Transformer is divided into two half-step modules as in a Macaron-Net [26]. By this means, both the MHSA and the convolution modules are sandwiched by the two feed-forward blocks as represented in Fig. 1. The last feed-forward module is followed by a layer normalization and every module employs residual connections.

B. Frequency Dynamic Convolution

Convolutional Neural Networks are widely employed for extracting local features in many audio-related tasks, including sound event detection. The raw waveform of an audio recording represents low-level characteristics of this signal. Consequently, a common procedure in SED is extracting features employing the mel scale. By this means, a mel-spectrogram representing the frequency distribution of energy over time can be obtained.

Originally, 2D convolutions were designed for the image domain [23], in which it is crucial for visual data that the extracted patterns are shift-invariant in both axes. When moving to the audio domain and specifically to SED and mel-spectrograms, this property is not desirable. The identification of a specific sound event relies on its spectral and temporal characteristics. If the pattern of a sound event is shifted along the time axis, we will still be able to recognize it. Conversely, if the same pattern is shifted along the frequency dimension, it would yield to a significantly different sound, as the frequency components that define a particular sound event class would change. Consequently, 2D convolutions that impose translation equivariance on both axes of a mel-spectrogram are unsuitable for the task in question.

To address this issue, Frequency Dynamic Convolutions (FDY) [17] were introduced as a method to enhance the performance of SED systems. The key objective is to maintain translation equivariance along the temporal dimension while removing it along the frequency axis. This is achieved by extracting frequency-adaptive attention weights and combining them through a weighted sum with basis kernels. By this means, frequency-adaptive kernels are obtained and employed as the kernels of a vanilla 2D convolution.

C. Multi-Resolution Approach

Sound events can be identified by their particular temporal and spectral characteristics, exhibiting substantial dissimilarities across different categories. The extraction of mel-spectrogram audio features requires defining a unique time-frequency resolution point, which may not adapt to the whole variability of sound events. Therefore, in previous research [20], [21], [22] we have analyzed the benefits of employing a multi-resolution approach to enhance the results through the utilization of various time-frequency resolution points.

Our multi-resolution technique involves the modification of the parameters that define the time-frequency resolution points. Considering the trade-off between time and frequency resolution of the Short Time Fourier Transform (STFT), we can enhance the resolution in the time axis by decreasing the number of samples employed in the STFT (N), the window length (L) and its hop length (R). These changes involve a decrease in time resolution and therefore, a reduced number of mel filters (n_{mel}) is required. Conversely, by increasing these four parameters we can obtain the opposite effect, enhancing the resolution in the spectrum at the expense of a reduction in time resolution. By varying these parameters during feature extraction, we create different single-resolution models.

Comparing the performance of these single-resolution models allows us to further analyze the behavior of our system at different resolution points, offering valuable insights into its strengths and limitations. Moreover, we have observed that each proposed PSDS scenario for this task benefits from specific time-frequency resolution points [22]. While increasing the resolution in time, in general, helps with obtaining a better localization of sound events, enhancing the resolution in frequency generally improves the recognition of sound events, avoiding confusion between classes.

Additionally, multi-resolution systems can be obtained by averaging frame-wise the output score for class c at time t ($s_{c,t}$) of N different single-resolution models, as follows:

$$s_{c,t}^{(comb)} = \frac{1}{N} \sum_{n=1}^{N} s_{c,t}^{(n)} \tag{2}$$

Since modifying the hop size leads to different sequence lengths, we linearly interpolate single-resolution scores to the maximum length, enabling their frame-wise combination.

D. BEATs

Self-supervised learning (SSL) techniques have witnessed a huge success in the audio domain, obtaining robust representations for both speech and non-speech signals, all accomplished with the utilization of unlabeled data. SSL models of speech [27], language [28], and vision [29] commonly adopt discrete label prediction as pre-training objective, mimicking the audio understanding skills of humans by extracting and clustering high-level semantics. However, audio SSL models still employ an acoustic feature reconstruction loss as pre-training task [30], [31], which seems to give priority to the accuracy of low-level time-frequency attributes but fails to consider the higher-level abstraction of audio semantics. In the pursuit of

achieving an audio SSL model based on discrete label prediction, Bidirectional Encoder representations from Audio Transformers (BEATs) [19] were proposed as an iterative audio pre-training framework, in which an acoustic tokenizer and an audio SSL model are optimized by iterations.

Initially, a random-projection tokenizer [32] is used to generate discrete labels which are then employed to train the audio SSL model using a mask and discrete label prediction. When the audio SSL model has converged, it is used as teacher to train a new acoustic tokenizer with knowledge distillation. This iterative learning process is repeated until convergence, allowing the model to learn relevant semantic information from iterations. The Vision Transformer [33] is used as the backbone network of the SSL model and Masked Audio Modeling (MAM) is proposed as pre-training task, in which the model learns to predict the patch-level discrete labels generated by the acoustic tokenizer. BEATs has notably achieved state-of-the-art results in audio classification tasks such as in the AudioSet-2M [7] and ESC-50 [34] benchmarks, outperforming previous SOTA audio SSL models such as MAE-AST [31].

E. Polyphonic Sound Detection Score

The evaluation of polyphonic sound event detection systems considers the difference in the location between the detections and the ground-truths for each sound event category. Timestamps can be highly influenced by the perception of the annotator, especially for brief, continuous sound events that can be interpreted either as distinct individual occurrences or as a singular unified event.

However, traditional evaluation collar-based metrics based on the proximity between onset and offset timestamps do not adequately address this inherent problem of subjective interpretation. Additionally, these metrics rely on a fixed operating point for making binary decisions with the output probabilities, giving a limited vision of a system's performance as it fails to consider the full range of potential operating conditions and decision trade-offs.

In order to build a more robust framework for the evaluation of SED systems, Bilen et al. [10] proposed the Polyphonic Sound Detection Score, which relies on the intersection between sound event detections and ground truths, without strict collar-based requirements. A detection is considered a true positive (TP) if it fulfills two criteria:

- The **Detection Tolerance Criterion (DTC)** discards detections from being counted as TP if the intersection with the ground truth over the length of the detected event is less than ρ_{DTC}.
- The **Ground-truth Tolerance Criterion (GTC)** measures the intersection between the ground truth and the set of detections that have accomplished the DTC. If this intersection normalized by the length of the ground-truth event is at least ρ_{GTC}, the detection will be counted as TP.

The detections that do not accomplish both criteria are considered false positive (FP) detections. In addition, the set of FP detections that intersect with ground-truth events from a different category are considered Cross Triggers (CT). For this purpose, the **Cross-Trigger Tolerance Criterion (CTTC)** is defined to account for CTs given a ρ_{CTTC}. Their penalty cost on user experience is regulated by the weighting parameter α_{CT}. Despite cross-triggers being more identifiable than false positives, they can lead to a more negative user experience than FPs.

The overall PSD-ROC curve is calculated by averaging the different class-dependent ROC curves, considering a cost of instability across classes denoted as α_{ST}. Finally, the PSDS score is obtained by integrating the PSDS-ROC curve from 0 to a specified parameter $e_{\max}$, which is measured in events per hour and defines the maximum false positive rate at which the operating points remain relevant.

By modifying the different parameters that define this metric, diverse characteristics of a SED system can be evaluated. Thus, in the DCASE Challenge Task 4A, two scenarios are proposed. The first one (PSDS1) focuses on a fast reaction upon sound events which translates into a highly accurate detection of timestamps. As shown in Table I, for achieving a high percentage of intersection both ρ_{DTC} and ρ_{GTC} are defined with large values. Whereas the cost of cross triggers is not considered for this scenario, it is crucial for the second one (PSDS2), which focuses on avoiding confusion between classes. For this purpose, the accurate localization of timestamps is not of high importance, but the evidence of cross-triggers penalizes the final score. For both scenarios, it is considered the highest cost for the instability across classes and an $e_{\max}$ value of 100 events/hour.

TABLE I
PARAMETER CONFIGURATION FOR THE TWO PSDS SCENARIOS PROPOSED

PSDS	ρ_{DTC}	ρ_{GTC}	ρ_{CTTC}	$\alpha_{\mathbf{CT}}$	$\alpha_{\mathbf{ST}}$	$\mathbf{e_{max}}$
Scenario 1	0.7	0.7	0.0	-	1.0	100
Scenario 2	0.1	0.1	0.3	0.5	1.0	100

III. METHOD DESCRIPTION

A. Proposed Conformer-Based System

Our Conformer model is based on the winning system of the DCASE 2020 Task 4 [14]. The system is composed of a first stage where a 7-layer CNN is utilized to extract high-level features from a mel-spectrogram, while also reducing dimensionality. Subsequently, 3 Conformer blocks are stacked. The MHSA module of each block is defined with 4 attention heads and an encoder dimension of 144. A linear layer is then employed to obtain the final posteriors by performing a position-wise classification. Additionally, it introduces a tagging token similar to the classification token used in the successful BERT language model [28] for summarizing the weak label predictions through the attention layers. By this means, this token is attached to the first frame of the feature sequence and consequently, weak predictions will be obtained from this first frame.

To enhance the performance over the second scenario of the PSDS, we incorporate Frequency Dynamic Convolutions [17] into the CNN-based feature extractor, to which we will refer as FDY-CNN. By this means, we were able to asses whether the improvements observed in the classification of non-stationary sound events in CRNN systems were also obtained with

Conformer-based ones. For the FDY-CNN we employ context gating as the activation function and define a time resolution reduction of 8, which was the value set in the original configuration. Additionally, we conduct a hyperparameter tuning using the value of PSDS2 as the objective score. For this purpose, we investigate the impact of varying the number of conformer blocks, attention heads and the pooling factor.

Data augmentation techniques are crucial for leveraging the robustness of a system. To improve regularization while enhancing the PSDS2 value, we investigate the combination of three different techniques. Due to the scarcity of labeled data, Mixup [35] is widely employed for obtaining new samples and their corresponding labels from the vicinity distribution of the original training data. Additionally, for improving robustness in the localization of sound events, frame-shift is applied over the sequence. Lastly, we experiment as well with the recently proposed FilterAugment [36]. This method adds robustness by modeling different acoustic environments, which is achieved through a random amplification of the energy at some frequency bands of the mel-spectrogram while reducing it in other bands. All these data augmentation methods were applied to the training data with a probability of 50%.

Finally, we exploit embeddings extracted from pre-trained model BEATs [19] to analyze its benefits along with the Conformer system. Following the same procedure proposed for the DCASE 2023 Challenge Task 4A baseline, one-dimensional embeddings of length 718 are extracted from our input audio and concatenated at frame-level to the output of the FDY-CNN feature extractor as shown in Fig. 2, requiring both sequences to have the same temporal dimension. To ensure this alignment and, following the baseline proposal again, we employ adaptive average pooling to adjust the length of the BEATs embeddings to match the one obtained at the output of our FDY-CNN.

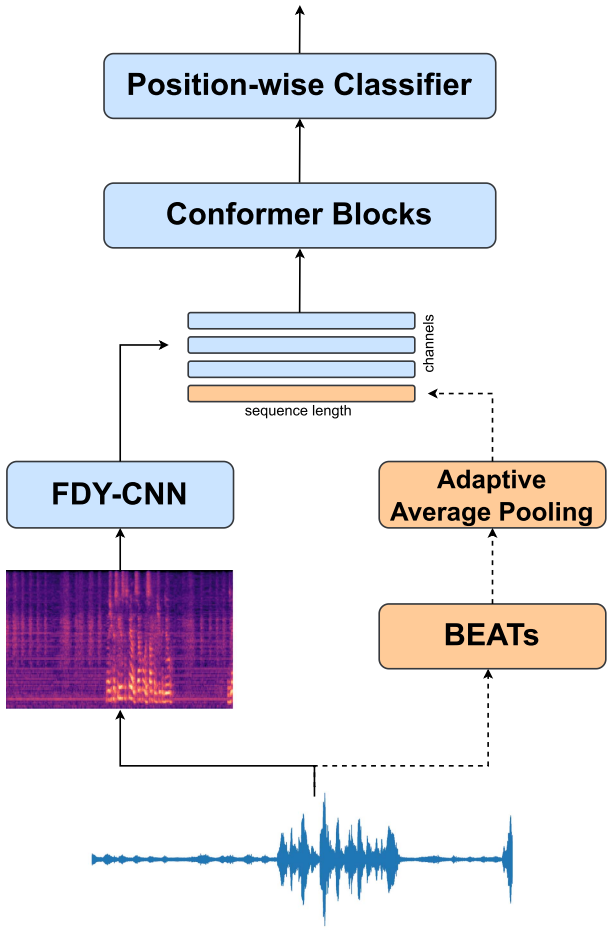


Fig. 2. Proposed Conformer-based system for Sound Event Detection. Frequency Dynamic Convolutional Neural Network is employed for subsampling and extracting features from mel-spectrograms. When incorporating pre-trained model BEATs, adaptive average pooling is applied to concatenate at frame-level the extracted embedding with FDY-CNN features, which is then fed as input to the Conformer blocks.

B. Multi-Resolution Analysis

In previous studies [20] we observed that CRNNs outperformed considerably the performance of the Conformer in terms of PSDS1. This suggests that the Conformer model may suffer from a deficiency in temporal precision, which complicates the accurate localization of onsets and offsets within a sequence. This particular limitation appears to be the primary weakness of the Conformer model.

To address this limitation, we propose to analyze our optimized Conformer employing the multi-resolution approach introduced in Section III-B. By performing this, we can examine whether an enhancement in time resolution of the employed features could potentially solve its inherent issues. Additionally, this method allows us to gain a deeper understanding of this architecture's behavior across different resolution points.

In Table II we introduce the five time-frequency resolution points we have defined for evaluating the performance of the Conformer across different feature representations. We have set the feature extraction configuration utilized in the baseline of the DCASE 2023 Task 4A (referred to as BS) as starting point. From this one, we define four additional resolution points. Among these, two are designed to double the resolution in frequency (F_{++}) and in time (T_{++}), whereas the remaining two are halfway points between BS and F_{++} (F_{+}) or T_{++} (T_{+}).

TABLE II
PARAMETER CONFIGURATION OF THE FIVE RESOLUTION POINTS EMPLOYED FOR THE FEATURE EXTRACTION, USING AS SAMPLE RATE $f_s = 16000$ HZ

Resolution	T_{++}	T_{+}	BS	F_{+}	F_{++}
N	1024	2048	2048	4096	4096
$L(ms)$	64.0	96.0	128.0	192.0	256.0
$R(ms)$	8.0	12.0	16.0	24.0	32.0
n_{mel}	64	96	128	192	256

The window length (L) and the hop size (R) are reported in Milliseconds (ms).

Additionally, we evaluated the benefits of using multi-resolution systems by fusing the output of different single-resolution models. We examined combinations involving 3, 4, and all resolutions presented in Table II, which we will refer to as *3res*, *4res* and *5res* in Section IV. All these combinations include the BS configuration, and depending on whether the remaining resolution points prioritize time or frequency enhancement,

we denote them with a T or a F, respectively. For instance, the multi-resolution system designated as *3res-T* comprises the BS resolution and the two resolutions enhanced in time (T_+ and T_{++}).

C. Experimental Setup

We conducted the different experiments employing the DESED (Domestic Environment Sound Event Detection) dataset [37], as it is the one employed for the DCASE Challenge Task 4A. The training set is composed of 10,000 synthetic audio clips with strong labels, 1,578 real recordings with weak labels and 14,412 real unlabeled recordings. While the real data is taken from AudioSet [7], the synthetic one is generated using the Scaper soundscape synthesis and augmentation library [38]. The new synthetic audios are obtained by mixing foreground events with background sounds. Foregrounds events have been extracted from FSD50K (Freesound Dataset 50k) [39] verifying that the sound event class belongs to the DESED dataset and is dominant in the audio clip. Background sounds that contain non-target classes are taken from the SINS dataset [40], MUSAN dataset [41] or Youtube. For selecting the best model during the training procedure, the synthetic validation set (2,500 clips) together with 10% of the weakly-labeled set is employed. For testing, we employ the validation set, which was constructed to match the clip-per-class distribution of the weakly labeled training set. It is composed of 1,168 real audio clips annotated with strong labels. Therefore, the different results provided in the following section are given over this real validation set and employing the recently proposed threshold-independent PSDS [42] as our evaluation metric.

As aforementioned, to optimize the Conformer architecture, we initially adopted the network configuration that won the DCASE 2020 Task 4 [14]. Therefore, for a consistent ablation study comparison, we employ the same configuration for feature extraction as the original system. This involved extracting mel-spectrograms with 64 frequency bins using a Hamming window with a size of 1024 samples (64.0 ms) and a hop length of 323 samples (20.2 ms). However, for our multi-resolution analysis, we utilize the parameter setting presented in Table II, where the BS resolution aligns with the one employed in the DCASE Challenge 2023 Task 4A baseline system, based on CRNNs. This approach allowed us to analyze both architectures using the same time-frequency resolution points to determine whether they can be complementary or if one can enhance both PSDS scenarios.

For dealing with both labeled and unlabeled data, we adopt semi-supervised learning. The Mean-Teacher method [43] is a prevalent technique in the SED field, utilizing two identical models: a student and a teacher, with the teacher's weights being an exponential moving average of the student's weights. By minimizing a consistency cost between the predictions of the student and teacher, the model learns to generate targets from unlabeled data. We employed Mean Squared Error as the loss function for minimizing this consistency cost, whereas Binary Cross Entropy is employed for the classification cost. Additionally, we have included in each batch the three types of annotations: strong, weak and unlabeled data with a ratio of 1:1:2. Models are trained employing the Adam optimizer with an exponential warm-up, setting a total of 200 epochs. Generally, the teacher model achieves a more consistent learning trajectory across epochs leading to a superior performance during testing. Thus, the model selection is performed over the teacher network.

TABLE III
EFFECTS OF EMPLOYING FDY-CNN ALONG WITH CONFORMER BLOCKS OVER THE DESED VALIDATION SET

Model Architecture	PSDS1	PSDS2
Conformer	0.221	0.554
FDY-Conformer	**0.268**	**0.618**

TABLE IV
EFFECTS OF INCREASING THE NUMBER OF CONFORMER BLOCKS OVER THE DESED VALIDATION SET

Conformer blocks	PSDS1	PSDS2
3	0.268	0.618
5	**0.269**	0.646
7	0.254	**0.648**
9	0.243	0.602

IV. EXPERIMENTS AND RESULTS

A. Optimization of Conformer System

As a first step, we explored the integration of Frequency Dynamic Convolutions within the CNN-based feature extractor module preceding the Conformer blocks. As shown in Table III, removing the translation equivalence in the frequency axis when extracting high-level features from mel-spectrograms considerably enhances the performance across both specified scenarios.

To investigate the impact of the number of Conformer blocks used in conjunction with the FDY-CNN feature extractor, we systematically varied the number of blocks from 3 to 9. The results presented in Table IV indicate that PSDS2 values benefit from a higher number of stacked blocks. The peak PSDS2 performance was achieved with the use of 7 blocks (0.648), surpassing the one obtained with the initial configuration of 3 blocks (0.618). However, it is important to note that increasing the number of blocks seems to have a negative impact on the temporal precision, as evidenced by the reduced PSDS1 values. We hypothesize that this behavior results from self-attention layers, which have access to the entire input sequence. Stacking a higher number of blocks might create an accumulative effect on the output of the Conformer, tending to be less local and more global, thus reducing the time precision. In the next section, we will further explore this effect by visualizing how the attention matrices vary along the Conformer blocks.

Considering a fixed encoder dimension, each head within the self-attention module specializes in focusing on specific segments of the input sequence. Therefore, modifying the number of attention heads inherently alters the span of attention for each head. We conducted experiments by varying the number of attention heads within the 7-block FDY-Conformer system to assess its impact on the two PSDS scenarios. As presented

TABLE V
EFFECTS OF MODIFYING THE NUMBER OF HEADS EMPLOYED IN THE MULTI-HEAD SELF-ATTENTION MODULE OVER THE DESED VALIDATION SET

Attention Heads	Dim/Head	PSDS1	PSDS2
2	72	**0.264**	0.614
4	36	0.254	**0.648**
8	18	**0.264**	0.626

TABLE VI
EFFECTS OF POOLING FACTOR EMPLOYED IN THE FDY-CNN FEATURE EXTRACTOR OVER THE DESED VALIDATION SET

Net Pooling	PSDS1	PSDS2
2	**0.343**	0.543
4	0.328	0.582
8	0.254	**0.648**
16	0.167	0.645

TABLE VII
EFFECTS OF DATA AUGMENTATION ON BOTH PSDS SCENARIOS OVER THE DESED VALIDATION SET

Mixup	Frame Shift	FilterAugment	PSDS1	PSDS2
X			0.254	0.648
X	X		0.273	0.637
X		X	0.274	0.650
X	X	X	**0.281**	**0.678**

in Table V, PSDS2 achieved its highest value with the default number of heads (4), while PSDS1 improved with alternative head counts (2 or 8), although the differences are small.

The CNN-based feature extractor includes a net pooling factor along the time axis, determining the length of the input sequence processed by the Conformer blocks. By reducing this factor, we can obtain longer sequences with higher temporal resolution. However, this also entails increased computational costs. Given that the Conformer baseline uses a default net pooling factor of 8, we conducted experiments to evaluate the effects of modifying this factor. The results in Table VI indicate that lowering the pooling factor improves temporal precision within the Conformer system, and thus PSDS1, but negatively impacts PSDS2 performance. In contrast, doubling the net pooling factor does not seem to improve the PSDS2 value but considerably affects the results of PSDS1. Therefore, a net pooling factor of 4 seems to achieve a reasonable balance between the two performance criteria. Considering that our primary focus was optimizing this architecture for the second scenario, we continued to use a net pooling factor of 8in subsequent experiments. However, given the performance obtained for both scenarios when increasing the input's length, we speculate that longer sequences might be impacting global attention, reducing its effectiveness as attending to distant frames becomes more difficult. In the next section, we will further analyze this effect on the attention function using our multi-resolution approach.

Finally, we explored the effects of various data augmentation techniques on our best-performing PSDS2 Conformer system, which was initially trained using only mixup [35]. As presented in Table VII, we examined the combination of mixup with other methods such as frame-shifting and FilterAugment [36]. The results show that employing frame-shifting alone improves performance, primarily in terms of PSDS1, while FilterAugment enhances results in both scenarios. Notably, when both techniques are used along with Mixup, there is a substantial overall improvement in performance.

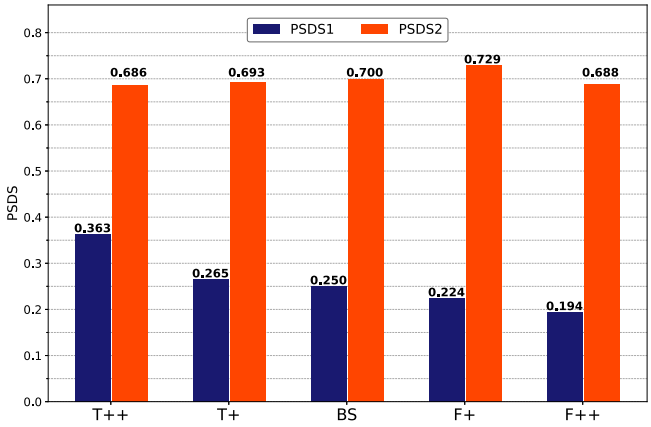


Fig. 3. Analysis on both PSDS scenarios of training the optimized Conformer system across the different time-frequency resolution points defined in Table II. Results are presented over the DESED Validation set.

B. Multi-Resolution Analysis

Given the results presented in the previous section, it is evident that the Conformer system exhibits relatively lower performance in terms of PSDS1, indicating that its key limitation is its lack of temporal precision. This observation underscores the need for further exploration and refinement of the Conformer architecture, particularly in addressing temporal aspects of sound event detection. For this purpose, we have employed our multi-resolution approach (see Section III-B) to analyze whether different time-frequency resolution points could alleviate this issue.

In Fig. 3 we present the results obtained when training our optimized Conformer with the different time-frequency resolution points defined in Table II. Observing the PSDS1 values, we can assert that the Conformer system is very dependent on the time resolution, as there is a notable variance between the results obtained with different settings compared to PSDS2, for which results present higher stability. Enhancing the time resolution significantly improves the results for PSDS1 without a dramatic drop in the PSDS2 performance. Additionally, an enhancement in frequency resolution appears to be more beneficial for PSDS2 but results in a lower score for PSDS1.

For a more in-depth analysis of the results, Fig. 4 presents the predictions of the Conformer output scores for two example audio clips at different time-frequency resolution points. For this purpose, we have selected the BS configuration, F_{+} which yielded the highest PSDS2 score and T_{++}, which enhances PSDS1. It is noticeable that enhancing the resolution in time improves the precision of onsets and offsets predictions, particularly when dealing with brief, continuous sound events such as the cat's meows depicted in Fig. 4(b). For such events, the

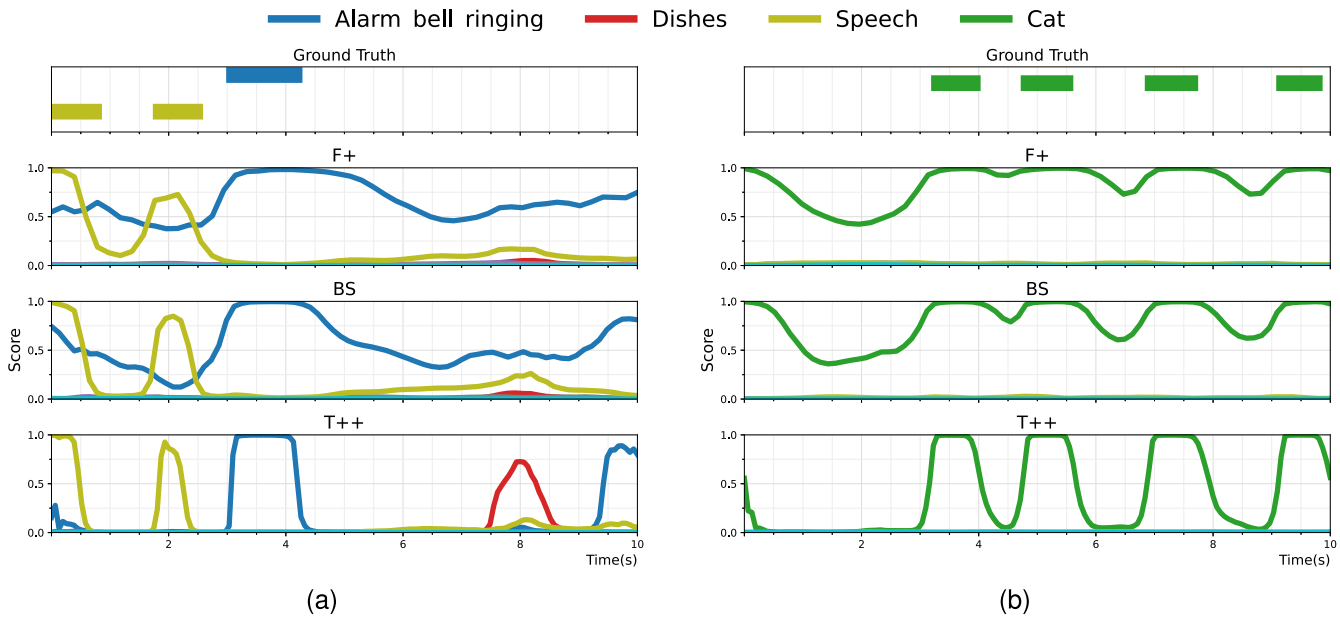


Fig. 4. Comparison of Conformer model predictions across three time-frequency resolution points defined in Table II. This comparison was conducted for two audio samples: (a) *Y39PfIe5CrPU_30.000_40.000.wav* and (b) *YPB5FosTwM8s_10.000_20.000.wav*, both extracted from the DESED Validation set.

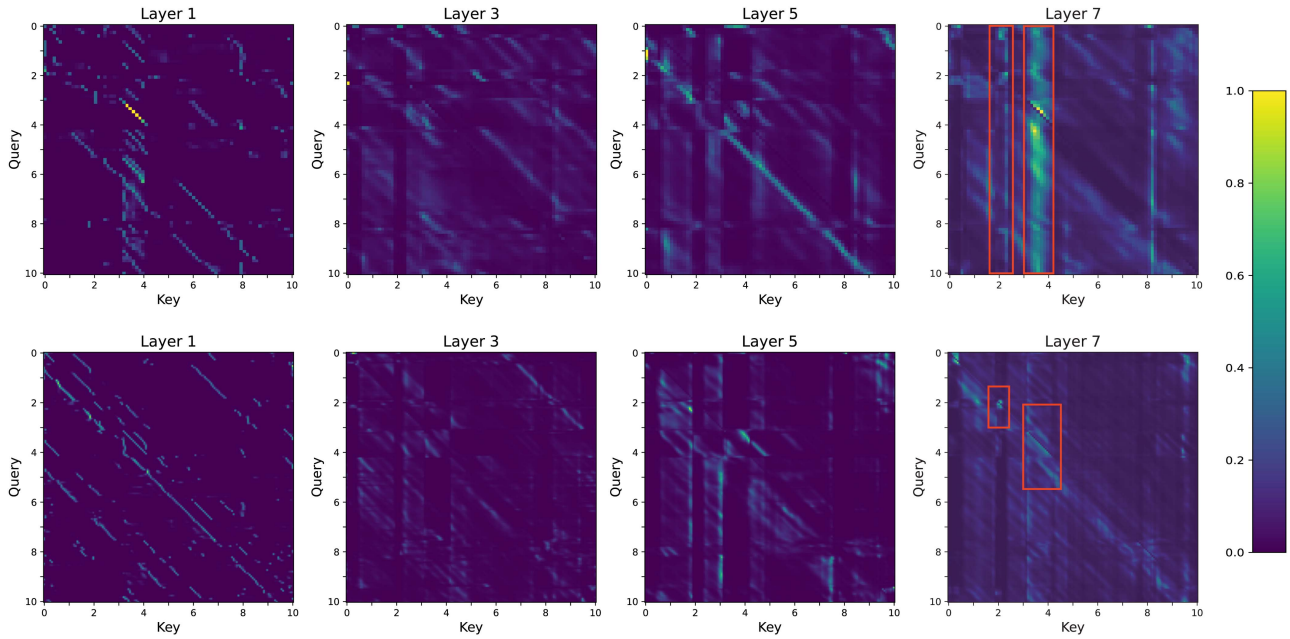


Fig. 5. Attention scores obtained for audio Y39PfIe5CrPU_30.000_40.000.wav. The top row displays attention scores obtained with BS resolution and the bottom row with T++ resolution. The attention matrices represent the average scores from the 4 heads in the MHSA module, visualizing only the even layers. Both axes denote time, facilitating a better comparison between the two resolution points.

Conformer tends to group all events into a single one, but the time enhancement helps to correctly locate each one separately. However, it is worth mentioning that this grouping feature of the Conformer also helps to avoid false predictions of short events caused by noise, which can be detected when time resolution is enhanced, as observed with the *Dishes* event in Fig. 4(a).

Considering the accurate time boundary predictions obtained when increasing the resolution in time, it is worth studying how the attention evolves along the Conformer layers employing different time-frequency resolution points. For this purpose, in Fig. 5 we present the averaged attention matrices of the 4 heads employed in the MHSA module for the example in Fig. 4(a). Observing the matrices obtained for the BS resolution (top row), it seems that the attention patterns appear randomly in the first layers. However, when increasing the number of blocks, they become more localized in the segments containing relevant

TABLE VIII
EFFECTS OF INPUTTING THE AGGREGATION OF ALL THE CONFORMER BLOCKS OUTPUTS IN THE POINT-WISE CLASSIFIER

Resolution	Aggregation	PSDS1	PSDS2
BS	No	0.250	**0.700**
	Yes	0.307	0.666
T_{++}	No	**0.363**	0.686
	Yes	0.356	0.628

Results are reported over the desed validation set.

audio events, achieving in the output layer vertical lines that tend to dominate over diagonal ones. In our example, the first red rectangle in the seventh layer represents the second speech event, whereas the following red rectangle is the alarm event. These vertical lines are associated with global attention, where all queries in the sequence (i.e. all output positions) attend with similar intensity to the same keys (i.e. to the same input locations). Consequently, as output frames at all positions focus on the alarm event, its time boundaries are not correctly localized and a fade-out effect is produced in the prediction. These global patterns may be adequate for long stationary sound events and avoiding cross-triggers, but they harm localization. This observation aligns with the results presented in Table IV, concluding that stacking a higher number of Conformer blocks increases the accumulative effect of self-attention layers, spanning the attention over the whole output sequence and leading to higher PSDS2 scores, but harming PSDS1 values.

Considering the evolution of attention patterns along the Conformer blocks and the impact of predominately global attention on results, it is worth investigating whether combining features from different layers could benefit our task. Based on the approach proposed for the Multi-scale Feature Aggregation Conformer (MFA-Conformer) [13], we concatenated the output of every Conformer layer and used it as input to the point-wise classifier. By this means, we could analyze whether low-level feature maps could also contribute to the accurate localization of sound events. As presented in Table VIII, aggregating the outputs of all layers for the baseline configuration improves the PSDS1 score compared to using only the output of the last layer. This suggests that adding representations from previous blocks partially compensates for the predominance of global attention patterns. However, this produces the opposite effect in the second scenario (PSDS2), where global attention seems to be more important.

In the case of the resolution enhanced in time, aggregating the outputs degrades the metric in both scenarios, suggesting that the output of the last block was a better representation. Analyzing the attention matrices of this resolution point, bottom row of Fig. 5, it can be observed that these vertical lines are still present but with a lower intensity and more discontinuous, implying that queries will pay higher attention to neighboring keys in the input and not to the whole sequence. Therefore, it seems that using enhanced time resolution or longer input sequences reduces the global attention effect, forcing it to be more localized and thereby, facilitating the identification of time boundaries.

TABLE IX
RESULTS OF COMBINING DIFFERENT TIME-FREQUENCY RESOLUTION POINTS OVER THE DESED VALIDATION SET

Resolutions		PSDS1	PSDS2
1res	T_{++}	**0.363**	0.686
	BS	0.250	0.700
3res	F_{+}, BS, T_{+}	0.263	0.738
3res-F	F_{++}, F_{+}, BS	0.240	0.736
3res-T	BS, T_{+}, T_{++}	**0.333**	0.732
4res-F	F_{++}, F_{+}, BS, T_{+}	0.264	0.742
4res-T	F_{+}, BS, T_{+}, T_{++}	0.293	0.740
5res	F_{++}, F_{+}, BS, T_{+}, T_{++}	0.306	**0.745**

In light of these observations, it appears challenging to identify a single resolution point that optimally enhances both scenarios simultaneously. Therefore, we explore different combinations of time-frequency resolution points to make the most of them and obtain a more robust performance. In Table IX we present some of the combinations explored along with their corresponding results. Analyzing the scores obtained for PSDS1, none of the combinations surpasses the performance achieved with T_{++}. This suggests that averaging this time-enhanced resolution point with other settings featuring lower temporal resolution adversely impacts the temporal precision, thereby affecting PSDS1 results. Conversely, multi-resolution systems outperform single-resolution models in terms of PSDS2 while increasing the PSDS1 performance of the resolution points not enhanced in time, obtaining systems more robust.

C. Conformer System in the DCASE 2023 Challenge Task 4 A

In this section we present part of our submission to the DCASE 2023 Challenge Task 4 A, where we introduced our PSDS2-optimized Conformer [44]. As detailed in Section III-C, to analyze the benefits of each architecture for the two scenarios proposed, we employ the time-frequency resolution used in the baseline of this task.

To assess the advantages of employing Conformer blocks for SED, we first compared the baseline system with our Conformer-enhanced model, without the use of Frequency Dynamic Convolutions. This initial comparison revealed a significant PSDS2 performance improvement resulting from the incorporation of Conformer blocks, as demonstrated in Table X. Our analysis further highlights the positive impact of Frequency Dynamic Convolutions on both the baseline and Conformer-based systems, particularly benefiting PSDS2. However, it is noteworthy that even with this enhancement, our Conformer system consistently outperforms CRNN-based systems in the second PSDS scenario. Regarding the number of parameters, our Conformer system has almost four times as many as the CRNN model. This difference is especially noticeable when adding the FDY-CNN module to both architectures, causing a greater relative increase in parameters for the CRNN compared to its baseline.

TABLE X
COMPARISON OF BASELINE AND PROPOSED ARCHITECTURES FOR SED OVER THE DESED VALIDATION SET

Model Architecture	PSDS1	PSDS2	Parameters
CRNN [1]	**0.367**	0.544	1.1M
Conformer	0.231	**0.584**	4.2M
FDY-CRNN	**0.385**	0.642	9.7 M
FDY-Conformer	0.250	**0.700**	12.6M
CRNN + BEATs [1]	**0.487**	0.752	1.2M
FDY-CRNN + BEATs	0.474	0.780	9.8 M
Conformer + BEATs	0.354	0.807	4.3M
FDY-Conformer + BEATs	0.390	**0.813**	12.9M

Additionally, we evaluated whether employing embeddings extracted from the pre-trained BEATs model was as well beneficial for Conformer blocks, as the incorporation of BEATs embeddings into the baseline yielded notably superior results. Following the same procedure, we concatenated the BEATs embeddings to both the output of our CNN and FDY-CNN modules, which subsequently serve as input to the Conformer blocks. In line with previous findings, the inclusion of Conformer blocks does not yield notable benefits for PSDS1 results, but it does enhance PSDS2 values, resulting in even more favorable outcomes than those achieved with multi-resolution systems. While incorporating FDY to the Conformer system with BEATs significantly improves the temporal localization of sound event timestamps, its impact on PSDS2 is less pronounced compared with previous observations. Therefore, it appears that the inclusion of BEATs embeddings already makes a significant contribution to the accurate classification of sound events, thereby diminishing the impact of FDY on the predictions. However, it is noteworthy that although using FDY-CRNNs along with BEATs increases the performance in PSDS2, results for PSDS1 are slightly worse. Additionally, in terms of efficiency, concatenating BEATs embeddings adds slightly more parameters to the model while boosting considerably the performance, making it a better approach than using only FDY.

V. CONCLUSION

In this paper, we have analyzed the advantages and limitations of the Conformer architecture for Sound Event Detection, in comparison with the most common architecture in this field, the CRNN, within the context of the DCASE 2023 Challenge Task 4 A. Given that the main limitation of the Conformer architecture relies on its temporal precision, we decided to optimize the Conformer for the PSDS2. Through this optimization process, we gained several important insights.

For instance, we observed that increasing the number of Conformer layers tends to improve PSDS2 while worsening PSDS1, suggesting that additional Conformer layers tend to smooth the outputs over time. To gain more insights into this problem, we combined the Conformer architecture with our previously proposed multi-resolution approach. By using features with enhanced temporal resolution, we partially compensated for the Conformer's lack of temporal precision, reaching similar PSDS1 results to those obtained with CRNN-based models, with minimal degradation in the PSDS2 results. This multi-resolution study has also allowed us to analyze the temporal resolution of the conformer outputs, confirming that this architecture tends to smooth the output in time and that this effect can be alleviated by using input features with better time resolution. We hypothesize that this effect could be due to the attention mechanism in the Conformer, which has access to the whole input sequence to generate the output at each time. Therefore we studied the attention matrices of the different Conformer blocks, observing that they tend to include vertical lines in the last block. This indicates that the prediction for every position in the output sequence tends to be computed based on only limited parts of the input sequence, where the sound events are located, producing therefore larger predictions over time. Given that the output of the different Conformer blocks seemed to contain different temporal precision, we also tried combining them, obtaining small improvements for PSDS1.

All these experiments have allowed us to better understand the benefits and drawbacks of the Conformer architecture for SED and how to optimize it for PSDS1 or PSDS2. This optimization process has also allowed us, by combining the conformer with Frequency Dynamic Convolutions and with the addition of BEATs embeddings, to obtain a very competitive PSDS2 result (0.813), outperforming the best non-ensemble PSDS2 score achieved by the winning team (0.807) [45] in the DCASE Challenge 2023 Task 4A.

REFERENCES

[1] J. Juodakis and S. Marsland, "Wind-robust sound event detection and denoising for bioacoustics," *Methods Ecol. Evol.*, vol. 13, no. 9, pp. 2005–2017, 2022.

[2] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Robust sound event detection in bioacoustic sensor networks," *PLoS One*, vol. 14, no. 10, Sep. 2019, Art. no. e0214168.

[3] S. Das, S. Pal, and M. Mitra, "Acoustic feature based unsupervised approach of heart sound event detection," *Comput. Biol. Med.*, vol. 126, Nov. 2020, Art. no. 103990.

[4] X. Zheng et al., "A CRNN system for sound event detection based on gastrointestinal sound dataset collected by wearable auscultation devices," *IEEE Access*, vol. 8, pp. 157892–157905, 2020.

[5] F. Angulo, S. Essid, G. Peeters, and C. Mietlicki, "Cosmopolite sound monitoring (CoSMo): A study of urban sound event detection systems generalizing to multiple cities," in *Proc. ICASSP 2023-2023 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[6] T. Lin et al., "Sound-based intelligent detection of FOD in the final assembly of rocket tanks," *Machines*, vol. 11, no. 2, 2023, Art. no. 187.

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

[8] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.

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

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

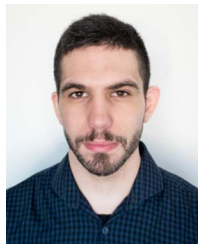
[11] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[12] J. Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Trans. Signal Inf. Process.*, vol. 11, no. 1, 2022, Art. no. e8.

[13] Y. Zhang et al., "MFA-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *Proc. Interspeech*, 2022, pp. 306–310.

[14] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation," in *Proc. Detection Classification Acoust. Scenes Events 2020 Workshop*, Nov. 2020, pp. 100–104.
[15] H. Zhang, S. Li, X. Min, S. Yang, and L. Zhang, "Conformer-based sound event detection with data augmentation," in *Proc. 2022 Int. Conf. Knowl. Eng. Commun. Syst.*, 2022, pp. 1–7.
[16] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech Lang. Proc.*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.
[17] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 2763–2767.
[18] H. Nam et al., "Heavily augmented sound event detection utilizing weak predictions," DCASE2021 Challenge, Tech. Rep.,Jun. 2021.
[19] S. Chen et al., "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. 40th Int. Conf. Mach. Learn.*, Jul. 23–29, 2023, vol. 202, pp. 5178–5193.
[20] D. de Benito-Gorron, S. Barahona, S. Segovia, D. Ramos, and T. Doroteo, "Multi-resolution combination of CRNN and conformers for DCASE 2022 task 4," DCASE2022 Challenge, Tech. Rep., Jun. 2022.
[21] D. De Benito-Gorrón, D. Ramos, and D. T. Toledano, "A multi-resolution CRNN-based approach for semi-supervised sound event detection in DCASE 2020 challenge," *IEEE Access*, vol. 9, pp. 89029–89042, 2021.
[22] D. de Benito-Gorron, S. Segovia, D. Ramos, and D. T. Toledano, "Multiple feature resolutions for different polyphonic sound detection score scenarios in DCASE 2021 task 4," in *Proc. 6th Detection Classification Acoust. Scenes Events 2021 Workshop*, Nov. 2021, pp. 65–69.
[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[24] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 5998–6008.
[25] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2019, pp. 2978–2988.
[26] Y. Lu et al., "Understanding and improving transformer from a multi-particle dynamic system point of view," 2019, *arXiv:1906.02762*.
[27] W.-N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, vol. 1, pp. 4171–4186.
[29] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Representations*, 2022.
[30] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-supervised audio spectrogram transformer," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, pp. 10699–10709.
[31] A. Baade, P. Peng, and D. Harwath, "MAE-AST: Masked autoencoding audio spectrogram transformer," in *Proc. Interspeech 2022, 23rd Annu. Conf. Int. Speech Commun. Assoc.*, H. Ko and J. H.L. Hansen, Eds., Sep. 2022, pp. 2438–2442.
[32] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 3915–3924.
[33] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[34] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd Annu. ACM Conf. Multimedia*, 2015, pp. 1015–1018.
[35] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. 6th Int. Conf. Learn. Representations*, Apr. 30, 2018.
[36] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *Proc. 2022 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 4308–4312.
[37] 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. Detection Classification Acoust. Scenes Events 2019 Workshop*, Oct. 2019, pp. 253–257.
[38] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "SCAPER: A library for soundscape synthesis and augmentation," in *Proc. 2017 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 344–348.
[39] 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, 2022.
[40] G. Dekkers et al., "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proc. Detection Classification Acoust. Scenes Events 2017 Workshop*, Nov. 2017, pp. 32–36.
[41] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, *arXiv:1510.08484v1*.
[42] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores," in *Proc. 2022 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 1021–1025.
[43] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.
[44] S. Barahona, D. de Benito-Gorron, S. Segovia, D. Ramos, and D. T. Toledano, "Multi-resolution conformer for sound event detection: Analysis and optimization," in *Proc. 8th Detection Classification Acoust. Scenes Events 2023 Workshop*, Sep. 2023, pp. 11–15.
[45] J. W. Kim, S. W. Son, Y. Song, H. K. Kim, I. H. Song, and J. E. Lim, "Semi-supervised learning-based sound event detection using frequency dynamic convolution with large kernel attention for DCASE challenge 2023 task 4," Detection and Classification of Acoustic Scenes and Events DCASE2023 Challenge, Tech. Rep., Jun. 2023.

Sara Barahona received the bachelor's degree in sound and image engineering from Universidad de Málaga, Málaga, Spain, in 2021, and the master's degree in deep learning for audio and video signal processing from Escuela Politécnica Superior, Universidad Autónoma de Madrid (UAM), Madrid, Spain, in 2022. She is currently working toward the Ph.D. degree in polyphonic sound event detection and source separation with AUDIAS Research Group, UAM. She has been working as a Research Assistant with the AUDIAS Research Group, UAM, since 2022.



Diego de Benito-Gorrón received the bachelor's degree in telecommunication technology and service engineering, the master's degree in ICT research and innovation, and the Ph.D. degree from Universidad Autónoma de Madrid (UAM), Madrid, Spain, in 2017, 2018, and 2023, respectively. Since 2018, he has been a part of the AUDIAS Research Group, UAM, focusing his research on deep-learning-based polyphonic sound event detection and leading the participation of AUDIAS in the DCASE (Detection and Classification of Acoustic Scenes and Events) international challenges in the years 2020, 2021, and 2022. In 2021, he joined the Speech@FIT Research Group, Brno University of Technology, Brno, Czechia, for a research internship under the supervision of Dr. Katerina Zmolikova, extending his investigation on sound event detection by incorporating source separation techniques. In 2017, he was the recipient of the UAM/Escuela Politécnica Superior award to the best academic record in Telecommunication Technology and Service Engineering.

Doroteo T. Toledano received the M.S. and Ph.D. degrees in electrical and electronic engineering from the Universidad Politécnica de Madrid, Madrid, Spain, in 1997 and 2001, respectively. He has experience of working in industry, in particular with the Speech Technology Division, Telefonica Research and Development, from 1994 to 2001, and in 2003. After his Ph.D. degree, he joined the Massachusetts Institute of Technology, Cambridge, MA, USA, as a Postdoctoral Research Associate with the Spoken Language Systems Group, from 2001 to 2002, under the supervision of Prof. Victor Zue and Prof. James Glass. In 2004, his trajectory as a Professor in signal processing starts when he joined the Universidad Autónoma de Madrid, Madrid, where he is currently a Full Professor. Since 2018, he has been the new Director of the AUDIAS Research Group. He has more than 25 years of experience in speech processing and more than 100 scientific publications. He has participated in six EU research projects and in more than 40 national projects (in ten of them as a Principal investigator). He has participated in more than 15 technological competitive evaluations (mainly NIST evaluations) and has organized three. His research interests include audio, speech, speaker, and language recognition. He was the recipient of several academic awards, such as the First National Bachelor Award of Spain, Best Academic Record in electrical and electronic engineering, and Ph.D. Dissertation Award from the Spanish Association of Telecommunication Engineers. He was the General Co-Chair and main Organizer of IberSPEECH 2012, and organizer and the session chair of several other conferences.

Daniel Ramos received the Ph.D. degree from the Universidad Autónoma de Madrid (UAM), Madrid, Spain, in 2007. Since 2011, he has been an Associate Professor with UAM. During his career, he has visited several research laboratories and institutions around the world, including the Institute of Scientific Police, University of Lausanne, Lausanne, Switzerland, School of Mathematics, The University of Edinburgh, Edinburgh, Scotland, Electrical Engineering School, Stellenbosch University, Stellenbosch, South Africa, more recently The Netherlands Forensic Institute, and Computational and Biological Learning Laboratory, University of Cambridge, Cambridge, U.K. He has been a Visiting Professor with the Universidad de Buenos Aires, Buenos Aires, Argentina, since 2019. He is currently a Staff Member with the AUDIAS Research Group, UAM. He 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 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. He has been invited by NIST to several workshops, including the OSAC standardization initiative. He has authored multiple publications in national and international journals and conferences, some of them awarded. His research interests include forensic evaluation of the evidence using Bayesian techniques, probabilistic calibration, validation of forensic evaluation methods, speaker and language recognition, signal processing, and pattern recognition. He is regularly a member of scientific committees in different international conferences, and often invited to give talks in conferences and institutions.