TEMPORAL CONTRASTIVE-LOSS FOR AUDIO EVENT DETECTION

Sandeep Kothinti, Mounya Elhilali

Laboratory for Computational Auditory Perception, Johns Hopkins University, Baltimore, USA

ABSTRACT

Temporal coherence is a feature-binding mechanism that ensures features that evolve together in time belong to the same object or event. Coherence has been extensively studied in biological systems, demonstrating how our brain leverages this mechanism to perform complex tasks in real environments and facilitate segregation of complex sensory signals (or wholes) into individual objects (or parts), following Gestalt principles. Although intuitive and computationally tractable, these concepts have rarely been leveraged in audio technologies. Audio event detection is an application that specifically deals with identifying sound events in an audio recording; hence is a natural avenue to explore principles of temporal coherence. In this study, we propose coherence-based learning, formulated as a contrastive loss, to train event detection models whereby embeddings driven by acoustic events are coherently constrained to maximize discriminability across events. This approach results in improved detection performance with no additional computational cost and a very small overhead during the training procedure.

Index Terms— Audio event detection, temporal coherence, contrastive learning, DCASE challenge.

1. INTRODUCTION

Audio event detection (AED) is an audio processing task of identifying the presence of environmental sound objects with the specification of precise event boundaries. Audio event detection (AED) has been gaining traction in the past few years, with numerous large-scale datasets and community-based workshops accelerating the progress. AED has found applications in information retrieval [1], smart homes [2] and smart cities [3] to name a few. The annual DCASE challenges have ranged from few rare sounds classes to more general sound classes [4] and recent iterations have addressed several important challenges with AED for real-life recordings, such as limited or non-availability of strongly annotated datasets [5, 6]. These efforts have led to the development of methods that leverage various big data and deep learning techniques and pushed the boundaries of event detection.

One of the overlooked aspects in most current AED techniques is the dynamic nature of soundscapes, where changes in sound events give rise to variations across acoustic features that inform the perception of the appearance or disappearance of new sources in the environment. Specifically, conventional methods employing a multi-instance classification objective presume temporal independence across all time points and rely on supervised labels to assign correspondences between time frames and class posteriors. This, in turn, often leads to reasonable classification performances but less-than-satisfactory event detection performances. In this work, we focus on distinct information during events vs. event boundaries; i.e. specific transition moments where the representation of the audio signal switches as a new sound source emerges or disappears from the scene. Here, we propose that learned mappings need to differentially favor event boundaries where embedding spaces need to transition across different event loci (or 'states'). Specifically, we leverage the concept of temporal coherence [7, 8] which posits that features that belong to the same event tend to co-vary together over time, a concept that forms a basic mechanism described by Gestalt psychology as a guiding principle as to how the brain organizes complex auditory and visual scenes into individual objects or events [9]. This principle has been shown to provide computationally tractable solutions to problems of sound separation and scene analysis [10, 11]. In the present work, we explore this principle for audio event detection and propose an extension of training objectives whereby embedding features of the system are constrained differently within and across event boundaries hence formulating a time-dependent loss function that modulates learned mappings from the data.

The coherence formulation proposed in this work is adapted from temporal coherence for video applications [12], where the video representations are optimized with coherence constraints. This formulation exploits both samples from the same class and different classes similar to contrastive learning approaches. Contrastive learning methods such as the triplet-loss [13] have been used for learning audio representations utilizing object class and temporal proximity to improve scene classification and event retrieval performance [14]. In this work, we use a similar contrastive-loss function to enforce temporal coherence, where only consecutive samples are considered for contrasting each time point. This formulation allows flexibility for representations for a given object from different scenes while constraining only within event representations. We hypothesize that the proposed contrastive-loss helps improve object identification and tracking and thereby improves boundary identification.

The details of AED problem formulation and the proposed method are described in Section 2. Experimental validation of the proposed method and evaluation metrics are detailed in Section 3. Section 4 gives the particulars of the results of the experiments followed by conclusions and potential future directions in Section 5 and relevant references in Section 6.

2. SOUND EVENT DETECTION USING TEMPORAL CONTRASTIVE-LOSS

2.1. Baseline system

One of the most common approaches to audio event detection is based on multi-instance classification where a classification model indicates the presence or absence of an audio object at each time-point. When strongly labeled data is available, this can be formulated as a classification problem. Let $(\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^{N}$ be a dataset of audio segments, where $\mathbf{X}_i \in R^{TF}, \mathbf{Y}_i \in \{0,1\}^{T'C}$ are a pair of in-

This work was supported by NIH U01AG058532, ONR N00014-19-1-2014, and N00014-19-1-2689.

put sequence and label sequence respectively. Here $T, T' : T \geq T'$ represent the time samples, F represents the number of input dimensions at each time in $\mathbf{X}_i$ and C is the number of classes in the data. Typically $\mathbf{X}_i$ is a time-frequency representation such as Mel-spectrogram. a classification model $f : \mathbf{X}_i \xrightarrow{f} \mathbf{p}(\mathbf{Y}_i) \in [0,1]^{T'C}$ can be trained to minimize classification error averaged over all segments and time-points within each segment.

$$\mathcal{L}_{entp} = \sum_{i=1}^{N} \sum_{t=1}^{T'} \sum_{c=1}^{C} H(y_{i,t,c}, p(y_{i,t,c})) \tag{1}$$

$$H(a, b) = -a \log(b) - (1 - a) \log(1 - b) \tag{2}$$

Here, H is binary cross-entropy which measures the distance between two Bernoulli distributions. Thus the objective in (1) finds an optimal map from audio input to the probability of an audio class. During inference, a threshold and some post-processing can be applied on columns of $\mathbf{p}(\mathbf{Y_i})$ to find contiguous segments of the audio with the presence of different classes.

2.2. Coherence loss

The loss function in (1) considers all time-points as equally important or informative. This assumption effectively ignores the particular relevance of instances near event boundaries which are expected to facilitate boundary detection; while instances within events are expected to reflect more stationary or coherent behavior in feature representation. By ignoring this temporal sensitivity, the models have poorer detection performance when the evaluation criteria require precise event boundaries along with event labels. To overcome this limitation, we propose a modification to the loss function that enforces higher contrast across event boundaries and higher coherence within event boundaries. Let $\mathbf{z}_{i,t}$ be an intermediate representation within the classification model at time t for an input $\mathbf{X}_i$ and $\mathbf{y}_{i,t}$ be a row vector of the label matrix $\mathbf{Y}_i$ at time t. The loss function is now defined as:

$$\mathcal{L}_{mod} = \mathcal{L}_{coh} + \mathcal{L}_{entp}$$

$$\mathcal{L}_{coh} = -\alpha_1 \sum_{i=1}^{N} \sum_{t=2}^{T'} \mathbf{1}_{>0}(||\mathbf{y}_{i,t} - \mathbf{y}_{i,t-1}||_1)(||\mathbf{z}_{i,t} - \mathbf{z}_{i,t-1}||_2^2)$$
$$+ \alpha_2 \sum_{i=1}^{N} \sum_{t=2}^{T'} \mathbf{1}_{=0}(||\mathbf{y}_{i,t} - \mathbf{y}_{i,t-1}||_1)(||\mathbf{z}_{i,t} - \mathbf{z}_{i,t-1}||_2^2) \tag{3}$$

Here, $\mathbf{1}_A(x)$ is an indicator function and $||.||_p$ is an L_p norm. α_1 and α_2 are hyperparameters that control the contribution of the additional loss terms. The additional loss terms are designed to maximize the distance between two consecutive samples that belong to different classes and minimize the distance when they are from the same class composition. Note that the addition or removal of a single class is considered for maximization, irrespective of changes in other classes. Figure 1 details the loss mechanisms with an example input and output.

3. EXPERIMENTAL SETUP

3.1. Dataset

To validate the benefits of the proposed method, partial data from the audio event detection dataset from DCASE2021 Task4 [6] is employed. Since the proposed method requires labels along time, only the synthetic dataset is used for training the models. The training set consists of 10,000 audio segments of 10s each, sampled at 16kHz with mono-channel audio. These segments include data from 10 different sound classes commonly encountered in domestic environments, such as alarm, speech, cat, blender, etc. This synthetic training dataset is created using Scaper [15] by mixing foreground segments of classes of interest with various background mixtures with randomization in the positioning of the events and duration of events. For each segment in this training set, class-wise annotations are determined during the synthesis process.

For evaluation of models, two evaluation datasets, namely *validation* and *public* are adopted from the DCASE2021 Task4 dataset. While *validation* set has 1168 audio segments which are chosen from AudioSet [16] dataset, *public* set has 692 audio segments which consist of samples from YouTube and Vimeo. Unlike the synthetic dataset, these evaluation sets are taken from real audio recordings and the temporally strong labels for these datasets are manually labeled by annotators.

For training the event detection models, the Mel-spectrogram of each audio segment is used as the time-frequency representations. 128-dimensional Mel-spectrograms are extracted with a window length of 2048 samples and a hop-size of 160 samples resulting in 100Hz frames with Mel filters applied on magnitude spectra. A global mean-variance normalization of the feature frames is performed with statistics computed on the synthetic dataset.

After validating the coherence based models on the synthetic dataset, the effectiveness of the coherence loss function was tested using the DCASE 2021 challenge baseline setup [6]. Since the *weak* and *unlabel-in-domain* parts of the dataset do not have groundtruth labels, a cross-entropy based objective function with teacher-student methodology was used as per the DCASE baseline setup. The model architecture, feature representations, training and evaluation procedures were not modified except for the addition of the coherence loss function for the synthetic subset.

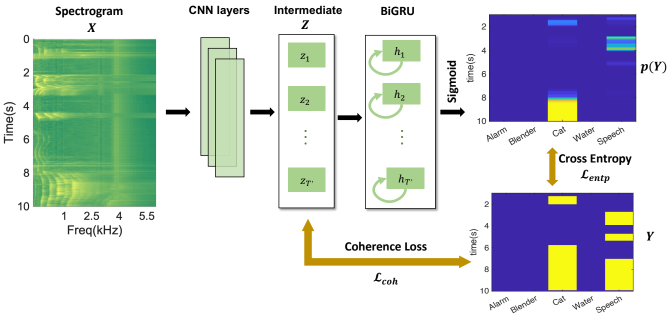

Fig. 1. Schematic of the model and two loss mechanisms: cross entropy and coherence loss The output of the CNN is used as the intermediate representation on which the coherence loss is applied. The model used in this paper consisted of 5 2D-CNN layers and 2 Bidirectional RNN layers.

3.2. Evaluation

For evaluating the performance of the different models, event-based F-scores [17] is used as the evaluation criteria. The F-score, computed as the harmonic mean of precision and recall, is a balanced measure of hits and false alarms. The event-based F-scores penalize errors in event onsets, offset and labels and have been used extensively to evaluate event detection models in the past iterations of

DCASE challenge [6, 5]. Macro F-score which is average class-wise F-score, is used as the primary evaluation criterion and micro F-scores, which ignore class-wise performance, are used as a secondary metric. F-scores are computed and reported for *validation* and *public* datasets separately. The tolerance parameter for the onsets is fixed at 200ms and is chosen as the maximum of 200ms or 20% of the duration of the event.

3.3. Model parameters

Convolutional recurrent neural networks (CRNN) are used as classification models in this work. The CRNN consists of 5 layers of 2D-CNNs with a kernel size of 3x3 and a stride of 1x1. Number of filters in the CNNs were [16, 32, 64, 128, 128] and pooling ratios for the layers are [[2, 2], [2, 2], [2, 2], [1, 4], [1, 4]]. The pooling ratios for the frequency axis are chosen to result in a single output dimension for each filter after the 5-layers. Gated linear units (GLU) [18] were used as the activation for the CNN layers. The downsampling from pooling in time results in a sampling rate of 12.5Hz. The output of the CNN is fed to two layers of bidirectional GRU (BiGRU) with 128 units in each layer. A fully connected layer with sigmoid units is used as the output layer with an output size of 10 to correspond with the 10 classes in the data. A dropout rate of 0.5 is applied during the training in all the layers. The architecture of the CRNN is fixed for all the experiments. The models are implemented with PyTorch toolkit [19], and a fixed manual seed is used to give similar initialization to all the models.

The CRNN models are trained with segments from the synthetic dataset using Adam optimizer with minibatch-based gradient descent with a minibatch size of 5 with 50 epochs of training over the synthetic dataset. A learning rate scheduling is performed with a decay factor of 0.8 for every 5 epochs. For the baseline model without coherence, the loss function from (1) is minimized. Among the trained models, the model with the best performance on *validation* is chosen as the final model. This model selection is performed to alleviate mismatch between the synthetic data used for training and the real data used for evaluation.

For the coherence models, the loss function is replaced with (3) with coherence loss applied on the output of the CNN layers as shown in Fig. 1. For stable convergence, the coherence loss is added to the classification loss with annealing over 5 epochs. To find optimal values for α_1 and α_2, a grid search for both the hyperparameters is performed in logarithmic scale with performance on the *validation* set as the criteria.

During the evaluation, posteriors from the CRNN models are thresholded with a cut-off of 0.5, and median filtering of window 5 samples (0.4s) is applied. These parameters are found to give consistently best performance across all models.

4. RESULTS

Table 1 shows macro and micro F-scores for both the baseline and coherence models for *validation* and *public* datasets. Since *validation* is utilized for selecting the best model, performance on this set served as maximum achievable improvement within the training conditions, and performance on *public* serves as the generalization measure of the models to unseen data. As can be seen from the table, the coherence model improves on both *validation* and *public* datasets with 2% absolute improvement in macro F-score, which indicates better detection across classes. To test statistical significance of the improvements, the models were trained for 5 trials with random initialization. Two-sample t-tests indicated significant improvement for both the validation (p=0.05) and public (p=0.007) datasets.

Improvement in micro F-score, an average measure on all the events ignoring imbalances in the number of events across classes, indicates the coherence model performs better on an average event. The best performance with the coherence loss is achieved using $\alpha_1 = 0.1$, $\alpha_2 = 0.03$. The higher value of α_1 can be attributed to the lower number of time points which are event boundaries compared to time points that belong to within event regions.

Model	*validation*		*public*	
	Macro	Micro	Macro	Micro
baseline	20.56	27.83	25.06	32.75
coherence	22.54	29.18	27.36	33.12

Table 1. Macro and micro F-scores for the baseline and coherence model. The coherence model was trained with $\alpha_1 = 0.1, \alpha_2 = 0.03$

.

To further analyze the observed improvements from coherence, the F-score is broken down with classes and presented in Fig. 2. Alarm, Blender, Cat, and Vacuum classes have significant improvement across both *validation* and *public* datasets. Dog, Water, and Speech classes show a slight degradation in F-scores. Overall, improvements are seen for classes that tend to have quasi-stationary spectral profiles. Classes that have broadband or noise-like profiles have no improvements or reduction in performance. This difference could be attributed to the underlying assumption of coherence that, within event boundaries, the audio object does not change too much.

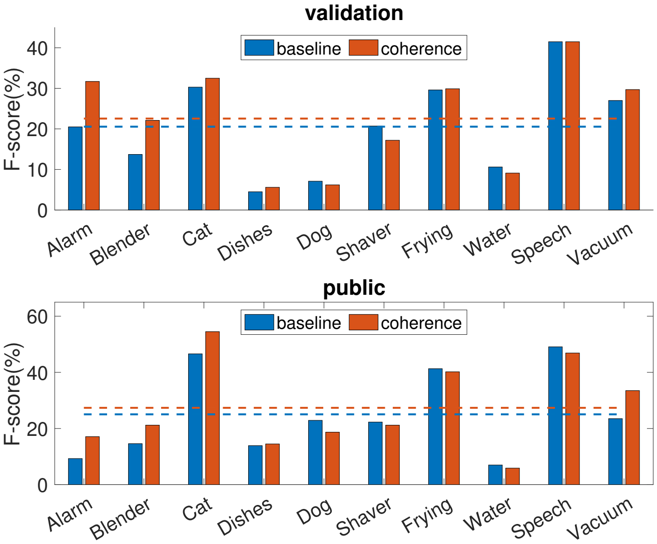

Fig. 2. Class-wise breakup of event based F-scores for baseline and coherence models across *validation* and *public* datasets. Horizontal dashed lines in blue and red indicate the macro F-score for baseline and coherence models respectively

The benefits of the coherence could come from better identification of classes, improved detection, and response to event boundaries, or both. To delineate the contribution from these factors, further analysis on the classes with improved F-scores (Alarm, Blender, Cat, Vacuum) is performed and presented in Fig. 3. Segment-based F-scores shown in Fig. 3a are a measure of accuracy in the classification of feature segments from within event boundaries. Looking at segment-based F-scores, it can be inferred that the coherence model identifies these classes better compared to the baseline model.

The segment-based F-scores may not necessarily translate to improvements in event-based F-scores as any misses around boundary segments or any noise in posterior within the event could lead to penalties in event-based F-scores. A measure of "change within the event" is computed as the L1-norm of derivative in the representation $\mathbf{z}$ for the select classes. A histogram of the changes shown Fig. 3b, gives different distributions for the two models, with the coherence model showing reduced changes both in terms of the variance and the mean of the distribution when compared to the baseline model, which can lead to stable posteriors and reduced false alarms.

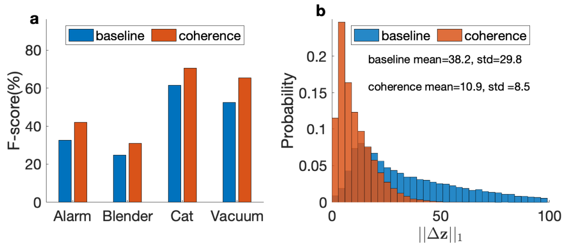

Fig. 3. (a) Segment based F-scores for select classes. (b) Histogram of change in representation within event boundaries for different models.

Event-based F-scores penalize a detection if the predicted boundaries do not match reference boundaries even if the model prediction matches the stable event regions. Thus the model posteriors must change at reference onsets and offsets for better event-based F-scores. To analyze this, a subset of events that are reliably detected within stable event regions are selected for further analysis. For events in this subset, the posteriors from the models from the samples around the onsets and offsets are collected to check the average posterior profile curves around event boundaries. Fig. 4a,b compare these average posteriors with reference onsets and offsets. As can be seen from these plots, the posterior probabilities for the coherence model have steeper slopes and reach higher/lower values within 200ms (tolerance used for detection) of the reference onsets/offsets compared to the baseline model. Thus even when both models detect the stable regions similarly, the boundaries are better represented by posteriors from the coherence model. The effect of such sharper posteriors can be seen in the example shown in Fig. 4c where the posterior probabilities of the Alarm class are shown for both baseline and coherence models. For this example, the coherence-based model has 3 hits, missing only the event from 5.5-6.5s, whereas the baseline model has no hits. This example demonstrates the benefits of having stable posteriors and sharper onsets.

Model	Macro F-score
DCASE 2021 baseline	41.5 $\pm$ 1
coherence	43.5 $\pm$ 1

Table 2. F-score for DCASE 2021 baseline and coherence models

Table 2 shows the macro averaged F-scores of the teacher models for DCASE 2021 baseline and a coherence model trained using the DCASE baseline setup. The reported results are average F-scores from 3 trials and their standard deviations. The coherence model, which modifies the loss function only for the synthetic dataset, was found to improve F-scores on a standard event detection task.

Thus coherence, formulated as a temporal contrastive loss, is found to achieve the objectives of finding stable representations within event regions while enhancing the edges around event boundaries. As hypothesized, these additional constraints lead to better event detection performance as evidenced by both event-based and segment-based F-scores. While the formulation discussed in this work requires strong temporal labels, it can be adapted to weakly labeled data either by using pseudo-labels or some other semi-supervised approach, which will be explored in the future.

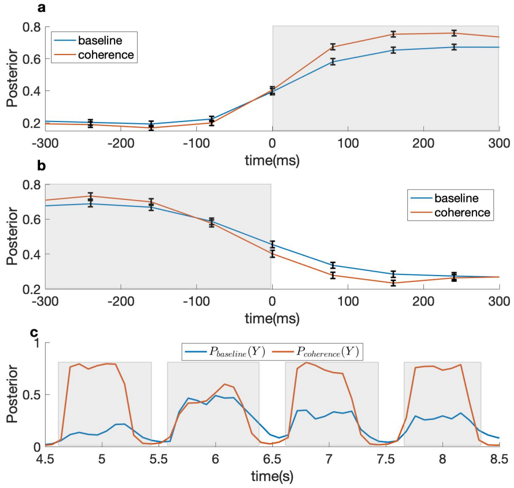

Fig. 4. (a) Average posterior probability anchored around event onsets. The shaded region indicates a post-onset event. (b) Average posterior probability anchored around event offsets. The shaded region indicates a pre-offset event. Error bars indicate ±1 standard error. (c) Example of an alarm class posterior probabilities from baseline and coherence models along with the reference event regions shaded in gray.

5. CONCLUSION

In the present work, we proposed a temporal coherence-based loss function for audio event detection. The proposed method exploits strongly labeled data to model event boundaries in a contrastive learning framework. By utilizing consecutive samples within each audio as both positive and negative samples, a temporal contrastive-loss-based objective function is proposed, which can be optimized using a gradient descent algorithm in a deep learning framework to train models. When trained using the synthetic subset of the DCASE2021 Task4 dataset, the coherence loss improved the event-based F-score of a CRNN model trained only using classification loss. By analyzing the representations within events and at event boundaries, we demonstrated the proposed objective function achieves both better event identification and sharper event boundaries. Given the simplicity of the formulation, the proposed method can be adapted to other event detection tasks or different model architectures.

6. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Sangwook Park for his help with setup of coherence loss for the DCASE 2021 baseline.

7. REFERENCES

[1] Qin Jin, Peter F. Schulam, Shourabh Rawat, Susanne Burger, Duo Ding, and Florian Metze, "Event-based video retrieval using audio," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2012, vol. 3.

[2] Christian Debes, Andreas Merentitis, Sergey Sukhanov, Maria Niessen, Nikolaos Frangiadakis, and Alexander Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, 2016.

[3] Juan Pablo Bello, Charlie Mydlarz, and Justin Salamon, "Sound analysis in smart cities," in *Computational Analysis of Sound Scenes and Events*. 2017.

[4] Annamaria Mesaros, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen, "Sound Event Detection in the DCASE 2017 Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 6 2019.

[5] Romain Serizel and Nicolas Turpault, "Sound Event Detection from Partially Annotated Data: Trends and Challenges," in *IcETRAN conference*, Srebrno Jezero, Serbia, 6 2019.

[6] Nicolas Turpault, Romain Serizel, Justin Salamon, and Ankit Parag Shah, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis," in *DCASE Workshop*. 2019, pp. 253–257, New York University.

[7] Shihab A. Shamma, Mounya Elhilali, and Christophe Micheyl, "Temporal coherence and attention in auditory scene analysis," *Trends in neurosciences*, vol. 34, no. 3, pp. 114–23, 3 2011.

[8] Shihab Shamma and Mounya Elhilali, "Temporal Coherence Principle in Scene Analysis," in *The Senses: A Comprehensive Reference*, B. Fritzsch, Ed., pp. 777–790. Elsevier, 2nd edition, 2020.

[9] A S Bregman, *Auditory scene analysis: the perceptual organization of sound*, MIT Press, Cambridge, Mass., 1990.

[10] Lakshmi Krishnan, Mounya Elhilali, and Shihab Shamma, "Segregating complex sound sources through temporal coherence.," *PLoS computational biology*, vol. 10, no. 12, pp. e1003985, 12 2014.

[11] Debmalya Chakrabarty and Mounya Elhilali, "A Gestalt inference model for auditory scene segregation," *PLOS Computational Biology*, vol. 15, no. 1, pp. e1006711, 1 2019.

[12] Hossein Mobahi, Ronan Collobert, and Jason Weston, "Deep learning from temporal coherence in video," in *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, New York, New York, USA, 2009, pp. 1–8, ACM Press.

[13] Kilian Q. Weinberger and Lawrence K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, 2009.

[14] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel P.W. Ellis, Shawn Hershey, Jiayang Liu, R. Channing Moore, and Rif A. Saurous, "Unsupervised Learning of Semantic Audio Representations," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, vol. 2018-April.

[15] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 10 2017, vol. 2017-Octob, pp. 344–348, IEEE.

[16] Jort Gemmeke, Daniel Ellis, Frydman Freedman, Aren Jansen, Wade Lawrence, Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proceedings of ICASSP*, 2017.

[17] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences (Switzerland)*, vol. 6, no. 6, 2016.

[18] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," in *34th International Conference on Machine Learning, ICML 2017*, 2017, vol. 2.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, vol. 32.