# JOINT ACOUSTIC AND CLASS INFERENCE FOR WEAKLY SUPERVISED SOUND EVENT DETECTION

Technical Report

*Sandeep Kothinti*[1*], *Keisuke Imoto*[2*], *Debmalya Chakrabarty*[1], *Gregory Sell*[3],
*Shinji Watanabe*[1], *Mounya Elhilali*[1]

[1] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA.
[2] College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan.
[3] Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA.

## ABSTRACT

Sound event detection is a challenging task, especially for scenes with simultaneous presence of multiple events. Task4 of the 2018 DCASE challenge presents an event detection task that requires accuracy in both segmentation and recognition of events. Supervised methods produce accurate event labels but are limited in event segmentation when training data lacks event time stamps. On the other hand, unsupervised methods that model acoustic properties of the audio can produce accurate event boundaries but are not guided by the characteristics of event classes and sound categories. In this report, we present a hybrid approach that combines an acoustic-driven event boundary detection and a supervised label inference using a deep neural network. This framework leverages benefits of both unsupervised and supervised methodologies and takes advantage of large amounts of unlabeled data, making it ideal for large-scale weakly labeled event detection. Compared to a baseline system, the proposed approach delivers a 15% absolute improvement in F1-score, demonstrating the benefits of the hybrid bottom-up, top-down approach.

***Index Terms—*** Sound event detection, unsupervised learning, weakly labeled data, restricted Boltzmann machine, conditional restricted Boltzmann machine, convolutional recurrent neural network

## 1. INTRODUCTION

Sounds in everyday soundscapes present a real challenge for audio technologies in order to parse the changing nature of the scenes and detect relevant events in the environment. With growing interest in smart devices, smart assistants and interactive technologies, there are increased efforts to develop robust ambient sound analysis systems able to analyze soundscapes, detect and track different sound sources and identify events of interest.

Parsing a scene to identify important events is a nontrivial task. Even humans exhibit a notable degree of variability in detecting occurrences of salient events when presented with realistic busy scenes [1]. Machine audition has tackled the problem of sound event detection by leveraging labeled data that allow machine learning algorithms to 'learn' characteristics of sound events, hence allowing the system to detect them whenever they occur. This supervised approach yields a reasonable performance especially in constrained settings where the nature of sound events and background sounds is well captured by the labeled data available for training. In reality, however, a fully supervised approach has limited scalability especially when dealing with everyday sound environments that can vary drastically depending on the setting and density of sources present. Acquiring large amounts of fully-labeled data in unconstrained environments is practically unfeasible especially considering that the kind of labels required for event detection involves not only identifying sound events in a scene, but also accurately labeling time stamps of occurrence of such events.

This in turn raises the question of potential benefits of unlabeled data to augment supervised training methods. There is a growing number of corpora that represent various urban soundscapes, domestic or workplace environments as well as everyday sounds. The abundance of such unlabeled datasets can enrich our ability to tackle ambient sound analysis provided the right kinds of tools are available to take advantage of both labeled and unlabeled data. DCASE 2018 task4[2] focuses on scenarios with large amount of unlabeled data along with a small set of labeled data. Past approaches to using unlabeled data to supplement supervised training have taken advantage of novel ideas to data augmentation to train machine learning systems that yield more robust event detection accuracies [3]. In parallel, unsupervised techniques have also been proposed to infer characteristics of sound events hence taking into account the dynamics of sound classes [4].

In the current work, we aim to leverage both the power of machine learning using a combination of labeled and unlabeled data to learn characteristics of event classes, as well as our knowledge of the physical and perceptual attributes of sounds that can help guide the segmentation of sound events as they occur in a scene. The latter approach employs principles from bottom-up auditory attention models where we know changes in sound structure are flagged by the human perceptual system as salient events that attract our attention for further processing. Detecting the onset and offset of these events of interest provides an anchor to our event labeling system that eliminates discontinuities in event labels hence resulting in notable improvement over a pure label-guided classification system. Section 2 describes the proposed system for event detection. Section 3 describes the experimental setup used to validate the proposed system. Finally, Section 4 discusses concluding remarks and future

[*] Contributed equally to this work

directions.

## 2. PROPOSED ARCHITECTURE

The proposed system combines a bottom-up (acoustic-driven) and top-down (label-guided) approach to detect sound events. The bottom-up approach relies solely on the acoustic characteristics of the audio signal to flag changes over time as captured in a high-dimensional mapping of the signal. The top-down approach is a supervised label-driven characterization of the sound labels derived from a deep neural network.

### 2.1. Event boundary detection

The acoustic-driven approach employs a generative framework to extract a rich mapping of the acoustic waveform that captures both local and global spectro-temporal regularities in the signal. The output of this representation is a rich array of activations in a high-dimensional space which allow tracking auditory events with different characteristics. This mapping is learned in an unsupervised fashion on weakly labeled and unlabeled in-domain training data.

This acoustic analysis is structured as a hierarchical system with 2 main stages. First, a Restricted Boltzmann Machine (RBM) [5] is employed to capture local spectral and temporal dynamics in the audio input over contexts of 30ms. The RBM, trained using Contrastive Divergence(CD), takes as input a biomimetic auditory spectrogram [6] and learns a mapping to Gaussian-Bernoulli units that best reproduces the signal spectrum. After training, RBM weights (**W**) and hidden bias (**b**) are used to transform input data (**v**) as given in (1).

$$h_i = \sum_j v_j W_{ji} + b_i \tag{1}$$

The next stage in the acoustic mapping further processes RBM outputs (**h**) using an array of 10 conditional RBMs [7, 8]. The cRBM array further analyzes the output of the first stage along a range of temporal contexts from 30ms to 300ms, hence capturing global dynamics in the signal and tracking events with different characteristics. The cRBM layer also employs Gaussian-Bernoulli visible-hidden units and is trained using CD. The weights (**W**, **A**) and biases (**b**) of the cRBM array are used as an affine transform to generate a final high-dimensional representation of the acoustic signal, as given in (2).

$$\begin{aligned} b_i^t &= \sum_j h_j^{t-1} A_{ji} + b_i \\ c_i^t &= \sum_j h_j^t W_{ji} + b_i^t \end{aligned} \tag{2}$$

The activations across the nodes of each cRBM network are further processed using Principal Component Analysis (PCA) [9] to get directions of maximal variance and reduce dimensionality to 16 dimensions per cRBM. The PCA outputs are then processed through first order difference and smoothed using a moving average with window length inversely proportional to the cRBM context length. The smoothed derivative from all the dimensions are summed to produce a measure of activity in time. We flag local maxima in this activity to indicate notable changes in the acoustic signal and hence a likely index of new acoustic events. The closest proceeding sample at 25% of the detected peak is marked as the onset point. Local minima of the short-term energy in the signal immediately following the detected onsets are flagged as offsets to the corresponding onsets. All parameters are tuned to maximize the F-score on the development set.

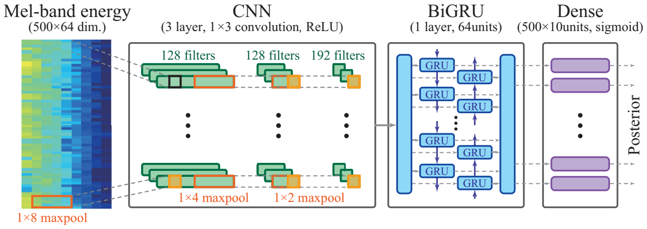

Figure 1: Convolutional recurrent neural network for acoustic event labeling

### 2.2. Event labeling

To label the acoustic event detected by the bottom-up approach, we employ a supervised deep neural network. This neural network outputs a posterior of acoustic event in each time frame, which is combined with the event boundary detection results for the class inference of acoustic events.

#### 2.2.1. *Convolutional recurrent neural network*

For the classification of acoustic events, we apply a convolutional recurrent neural network (CRNN), which is used as the baseline system of the task 4 of DCASE 2018. The acoustic features used in this system consist of 64 dimensional log mel-band energy extracted in 40 ms Hamming windows with 50% overlap. The log mel-band energy is then fed to the CRNN, which has 3 convolutional layers followed by a bi-directional gated recurrent units (BiGRU). The network structure and parameters are shown in Fig. 1. To keep the time resolution, the pooling is not performed along the time axis in CNN layers. The CRNN is trained using Adam optimizer and an early stopping technique.

#### 2.2.2. *Label inference*

An acoustic event label isgiven to the unlabeled event by calculating an average posterior of each acoustic event in the active duration. In this report, we regard an acoustic event that has the highest average posterior in the active duration as the classification result.

## 3. EXPERIMENTAL EVALUATION

### 3.1. Dataset

Audio data for Task4 of the DCASE Challenge 2018 is a subset of Audioset[10] drawn from Youtube videos and consisting of various sound classes occurring in domestic context. Training data includes a small number of audio files labeled at the sound clip level along with a large set of unlabeled files containing both in-domain and out-of-domain data. Test data contains a development set which is annotated with event boundaries at event level and an evaluation set which is used to evaluate the submission. In our system, we used

Table 1: F-score and error rate in event-based metrics

| **Method** | **Macro average** | | **Micro average** | |
|---|---|---|---|---|
| | F-score | Error rate | F-score | Error rate |
| Baseline | 14.87% | 1.52 | 8.87% | 1.41 |
| System 1 | 29.31% | 1.40 | 34.11% | 1.22 |
| System 2 | 29.83% | 1.44 | 33.46% | 1.22 |
| System 3 | 24.56% | 1.46 | 27.19% | 1.30 |
| Ensemble | **30.05%** | **1.36** | **34.12%** | **1.19** |

only weakly labeled and unlabeled in-domain training data for both unsupervised and supervised models.

### 3.2. Evaluation metric

Event detection is evaluated event-by-event using macro average and micro average of F1-scores. Macro average is computed as the average of class-wise F-scores and micro average is the F-score of all events irrespective of classes. Error rate (ER) is used as a secondary metric to assess errors in terms of insertions, deletions and substitutions. sed_eval toolbox[11] is used to compute F1-scores and ER. Onsets are evaluated with collar tolerance of 200ms. Tolerance for offsets is computed per event as the maximum of 200ms or 20% of event length. An event is considered to be a hit only when the predicted label matches with the ground truth and event boundaries correspond to the annotated boundaries. Hence any mismatch in either the labels or boundaries will result in a false positive and a false negative.

### 3.3. Baseline system

The baseline system is a CRNN with 3 CNN layers and 1 BiGRU layer, trained in two stages. During the fist stage, weakly labeled data is used for training with an objective of predicting the label at clip level. Unlabeled in-domain data is labeled using the first trained model and is used in the second stage of training. Training progress is monitored using a held-out validation set. During the first stage of training 20% of the weakly labeled data is used as the validation set and during the second stage of training, the entirety of the weakly labeled data is used as the validation set. 64 dimensional log Mel-band magnitudes are used as input features and the whole sound clip is given as the input to the CRNN which uses 2-D convolution in time and frequency. During test time, strong labels are assigned based on the posterior probabilities and smoothed using a median filter of length 1s. Performance of the baseline system is given in Table 1.

### 3.4. Classification results

The classification results for the test set of the development dataset are shown in Table 1. In System 1, the event labels are predicted using a CRNN trained using only the weakly labeled data. In System 2, CRNN is trained using weakly labeled data (1,578 clips) and augmented data (1,080 clips) which are generated by mixing multiple weakly labeled clips. System 3 uses predictions from the DCASE 2018 baseline model for Task4. Ensemble system uses majority vote on predictions from Systems 1-3. As seen in the table, the ensemble system achieves the best performance. Class-wise performance is shown in Table 2.

Table 2: Class-wise F-score

| **Class** | **Baseline** | **Ensemble** |
|---|---|---|
| Alarm/Bell/Ringing | 5.0 | 34.9 |
| Blender | 17.8 | 20.3 |
| Cat | 0.0 | 31.2 |
| Dishes | 0.0 | 17.8 |
| Dog | 0.0 | 48.1 |
| Electric shaver/toothbrush | 35.1 | 22.6 |
| Frying | 29.4 | 10.5 |
| Running water | 10.3 | 33.3 |
| Speech | 0.0 | 36.2 |
| Vacuum cleaner | 51.1 | 45.5 |

## 4. CONCLUSION

In this work, we propose a segmentation and recognition method for sound event detection based on unsupervised and semi-supervised methods. This method combines the acoustic-driven event boundary detection and the supervised acoustic event classification to annotate sound events in complex acoustic scenes. The proposed method makes use of unlabeled data enabling large-scale acoustic event detection. Experiments on the DCASE challenge 2018 dataset showed that the proposed method outperforms the baseline system and it achieves 30.05% in the event-based macro average metric.

## 5. REFERENCES

[1] N. Huang and M. Elhilali, “Auditory salience using natural soundscapes,” *The Journal of the Acoustical Society of America*, vol. 141, no. 3, p. 2163, mar 2017.

[2] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. Parag Shah, “Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments,” July 2018, submitted to DCASE2018 Workshop. [Online]. Available: https://hal.inria.fr/hal-01850270

[3] Z. Zhang and B. Schuller, “Semi-supervised learning helps in sound event classification,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 333–336.

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

[5] G. E. Hinton, “Learning multiple layers of representation,” *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[6] T. Chi, P. Ru, and S. A. Shamma, “Multiresolution spectrotemporal analysis of complex sounds,” *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[7] G. W. Taylor, G. E. Hinton, and S. T. Roweis, “Modeling human motion using binary latent variables,” in *Advances in neural information processing systems*, 2007, pp. 1345–1352.

[8] G. W. Taylor, L. Sigal, D. Fleet, and G. E. Hinton, “Dynamical binary latent variable models for 3d human pose tracking,” pp. 631–638, 06 2010.

[9] K. P. F.R.S., “Liii. on lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901. [Online]. Available: https://doi.org/10.1080/14786440109462720

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

[11] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.