INTEGRATED BOTTOM-UP AND TOP-DOWN INFERENCE FOR SOUND EVENT DETECTION

Technical Report

Sandeep Kothinti[1], *Gregory Sell*[2], *Shinji Watanabe*[1], *Mounya Elhilali*[1]

[1] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA.
[2] Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA.

ABSTRACT

While supervised methods have been highly effective at defining boundaries of sound events, the characteristics of the acoustic scene itself can provide complementary information about the changing profile of the scene and presence of new events. This work explores an integrated supervised and unsupervised approach to weakly labeled sound event detection by complementing a class-based inference system with a bottom-up, salience-based analysis. The two systems work conjointly in two ways: 1) Class information from the supervised model is used to tune the parameters of the bottom-up salience detection; and 2) Salience-based boundaries are leveraged to create pseudo-labels for weakly labeled data to generate more samples of strongly annotated data. These operations reflect the interplay between stimulus driven analysis and semantic driven analysis. The proposed method gives an absolute improvement of 11% on macro-averaged F-score on the development set.

***Index Terms*—** Audio event detection, saliency, restricted Boltzman machines, conditional restricted Boltzman machines, mean-teacher student.

1. INTRODUCTION

Emerging audio technologies require robustness in various noise conditions similar to how humans face challenging environments in daily life. Humans parse the complex soundscapes presented to them to focus on the sources that are most interesting or informative. Audio technologies can benefit from emulating processing observed in human auditory system to become robust to harsh environmental noise conditions.

Modern technologies leverage deep learning methodologies to achieve robustness in tasks like speech recognition [1] [2], speaker recognition [3], audio tagging [4] etc, Most of these approaches rely on supervised training methods using large amounts of training data. Acquiring large amounts of labeled data for tasks with different goals is difficult in practice. This presents an interesting scientific challenge of leveraging large amounts of unlabeled data to overcome this limitation.

Task4 of DCASE challenge 2018 focused on the event detection task with a large set of unlabeled data along with a small set of weakly labeled data. Semi-supervised approaches tackled this problem either using pseudo-labelling [5, 6] or by using multi-task based transfer learning methods [7]. Pseudo-labelling methods used a two step training process in which the model from the first stage is used to label the unlabelled data, which is subsequently used in the second stage of training. Since there is very limited labeled data, this method can add bias to the model which might limit the performance gains. Multi-task training, as in [7], presented an interesting alternative as it used the unlabeled data to add regularization terms to the cost function. As an alternative to the semi-supervised approaches, our submission [8] used an unsupervised boundary detection and supervised event labelling to leverage the unlabeled data. Task4 of DCASE 2019 expands on previous edition by adding a small amount of synthetic data labeled with timestamps of events. The objective of this edition is to find if synthetic data can be leveraged to improve the accuracy of event detection.

While leveraging machine learning methods for audio event detection, it is important to incorporate findings from studies on the human auditory system. Event labels are annotated by humans and any biases added by human behavior can be accounted with such knowledge. Salience-based event detection is a step in this direction and previous works using this method [8, 9] showed promising results. In the current work, we aim to improve on the salience-based event detection method of [8] by utilizing top-down information from the supervised system to improve the boundary detection. Since boundaries predicted by this method are reliable when an event label is known, we aim to use pseudo-labels for the weakly labeled data with event timestamps. By integrating information from the complementary subsystems we hope to move towards an event detection system that integrates top-down and bottom-up information seamlessly. Section 2 provides details of the proposed method and delineates how this integration is performed. Experimental setup and observed results are detailed in Section 3. Section 4 summarizes findings of this work and discusses potential directions for further work.

2. PROPOSED METHOD

The proposed method divides the event detection task into two sub-problems: 1) finding event boundaries; and 2) labelling the detected event boundaries. Similar to [10], we use salience derived from unsupervised methods to find event onsets. A supervised deep neural network is used for the labelling. These subsystems are explained in detail in the following sections.

2.1. Bottom-up Event Boundary Detection

Event onsets are derived using salience-based onset detection. Leveraging on previous studies on salience[11, 12], we use change detection as a measure of salience. In [10], the derivative based approach of [11] was used for task 4 of DCASE 2018 challenge. We expand on this approach by replacing the derivative with a Kalman filter prediction error similar to [12]. Onset analysis consists of three stages, an acoustic analysis stage that projects the audio data to a high dimensional space, followed by a salience analysis either using a derivative or a Kalman prediction and a peak detection method that outputs event onsets from the salience map. These stages are briefly discussed below.

2.1.1. Acoustic analysis

Acoustic analysis is divided into 3 steps. First, a biomimetic auditory spectrogram [13] $S(t, f)$, which is a time(t)-frequency(f) representation extracted from the input audio. 3 consecutive frames of $S(t, f)$ are stacked to produce a temporal context of 30ms and are used as input to a Restricted Boltzmann Machine (RBM) [14] to capture local spectro-temporal dependencies of the incoming audio signal. RBM weights (**W**) and hidden bias (**b**) are used to transform input data (**v**) as given in (1).

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

The next stage processes RBM outputs (**h**) using an array of 10 conditional RBMs(CRBMs)[15] with temporal contexts from 30ms to 300ms, hence capturing global dynamics in the signal and tracking events with different temporal characteristics. The weights (**W**, **A**) and biases (**b**) of the CRBM array are used as an affine transform to generate a high-dimensional representation of the acoustic signal, as given in (2).

$$\begin{aligned} d_i(t) &= \sum_j h_j(t-1)A_{ji} + b_i \\ y_i(t) &= \sum_j h_j(t)W_{ji} + d_i(t) \end{aligned} \tag{2}$$

We perform PCA whitening on the output of each of the CRBMs to get components that capture 95% of variance.

2.1.2. Derivative-based approach

In this method, changes in features are used as a measure of salience. The derivative is performed on the PCA outputs($z(t)$). The absolute values of the derivatives are summed across dimensions and are smoothed to give a single measure of salience ($s(t)$). Since different events have different temporal dynamics, we use class specific smoothing (τ_c) as given in (3).

$$s(t) = \sum_{t-\tau_c/2}^{t+\tau_c/2} \sum_i |z_i(t) - z_i(t-1)| \tag{3}$$

For each input audio, the class with highest average posterior is assumed to be the event in that audio. Optimal time-constants(τ_c) for different classes are chosen empirically to achieve best performance on the development set.

2.1.3. Kalman filter-based approach

The derivative-based approach doesn't account for the varying statistics of features over time. To account for this uncertainty, we use a Kalman filter-based prediction model which predicts the next sample of the PCA outputs($z(t)$) given the history. We use an autoregressive prediction model of order p as given in (4) and (5). Here, ϵ_x is the process noise and ϵ_z is the measurement noise. Covariance of $\epsilon_x(Q)$ and variance of $\epsilon_z(R)$ are chosen empirically.

$$x(t+1) = Cx(t) + \epsilon_x,\ \epsilon_x \sim N(0, Q) \tag{4}$$

$$z(t) = Hx(t) + \epsilon_z, \epsilon_z \sim N(0, R) \tag{5}$$

State transition matrix(C) and output matrix (H) are given by

$$\begin{aligned} C &= \begin{bmatrix} \frac{e^{-\alpha}}{S} & \frac{e^{-2\alpha}}{S} & ... & \frac{e^{-p\alpha}}{S} \\ 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ \vdots & \vdots & ... & \vdots \end{bmatrix}, S = \sum_{k=1}^{p} e^{-k\alpha} \\ H &= [1, 0, ..., 0]^T \end{aligned} \tag{6}$$

Similar to τ_c in the derivative method, p is chosen empirically for different classes. This system predicts $z(t+1)$ as an decaying exponential sum of history with decay factor α which is chosen empirically. Kalman prediction is performed independently on each feature dimension and the prediction error is computed for each sample. Prediction errors from different feature dimensions are summed to produce a measure of salience as given in (7).

$$s(t) = \sum_i |z_i(t) - \hat{z}_i(t)| \tag{7}$$

2.1.4. Onset and offset detection

Onsets are detected as peaks in the salience $s(t)$ computed by the derivative based approach or the Kalman filter based approach. Detected onsets are post-processed such that no two consecutive onsets are within 200ms. Offsets are detected by thresholding the short term energy (STE) of the audio signal. Each onset is paired with an offset that immediately follows the onset. We chose the thresholds for STE empirically for each class.

2.2. Top-down Event Labeling

We use the baseline system of DCASE 2019 challenge with few modifications as the top-down system. The first modification is to replace the softmax attention layer with linear softmax attention layer. Linear softmax attention was shown to perform better [16] for event localization compared to other forms of attention. To expand the training set with time stamp labels, a subset of weakly labeled files with events of only one class(984 files) are added to the synthetic dataset. For these files, event timestamps are added using the boundary detected from the bottom-up event boundary detection.

To label the acoustic event detected by the bottom-up approach, we use the class posteriors computed by this top-down system. For each detected onset-offset pair, the class with highest average posterior probability is assigned as the event label.

3. EXPERIMENTS

3.1. Dataset

For evaluating the proposed method, we used the dataset provided in task 4 of DCASE 2019 challenge. Training data consists of real recordings with a weakly labeled subset (1578 files) and an unlabeled in-domain subset (14412 files). Along with this real data, an additional synthetic dataset(2045 files) with events with timestamps is also provided. The development set (1168 files) on which the performance is evaluated, is annotated with time boundaries for each event. In our systems, we used only weakly labeled and unlabeled in-domain training data for unsupervised models. For the supervised systems, we used synthetic data along with the real data.

3.2. Evaluation metric

Event detection is evaluated using the macro average of event-based F-scores. Macro average is computed as the average of class-wise F-scores. The sed_eval toolbox [17] is used to compute F-scores. Onsets are evaluated with a collar tolerance of 200ms. Tolerance for offsets is computed per event as the maximum of 200ms or 20% of event length.

3.3. System Description

For training the supervised top-down systems we use the training script provided with DCASE 2019 baseline. The baseline system is a CRNN with 3 CNN layers, 1 BiGRU layer and 1 dense layer. The training is similar to that of the best performing system of DCASE 2018 task 4 [7] which uses a multi-task training method that includes a frame level cost for synthetic data, segment level cost for weakly labeled data and a consistency cost for unlabeled data. The consistency cost is computed on the prediction of a student model and a mean-teacher which has similar architecture as the student model. Parameters of the student model are updated using gradient descent and parameters of the teacher model are computed as the exponential moving average of the student parameters. 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. The dense layer has a softmax function applied over the time dimension which provides an attention mechanism that can used to computed a weighted average of the posteriors.

For training RBM-CRBM systems for bottom-up acoustic analysis, we used constrastive divergence [18] with 10 steps of alternating Gibbs sampling. This objective maximizes likelihood of the input data. Both the top-down and bottom-up systems are trained on NVIDIA RTX-2080Ti GPUs.

3.4. Results

Since we have different variations of both top-down and bottom-up systems, first we discuss results of variations in top-down systems, followed by variations in bottom-up systems and finally combinations of top-down and bottom-up variations.

3.4.1. Top-down systems

As discussed in the proposed method, we experimented with the attention layer and pseudo-labelling the weakly-supervised data. Table 1 shows macro-averaged F-scores on two test sets: Eval2018 (which is the evaluation set used for DCASE 2018) and Validation (which includes Eval2018 and development set of DCASE 2018). It can be seen that using linear softmax attention and pseudo labelled data provides gains in performance. Since these two modifications are unrelated, using both of these modifications improves F-score by 7%.

Table 1: Macro F-score for different top-down systems

Method	Eval 2018	Validation
Baseline	22.59%	24.18%
Linear softmax(LS)	24.48%	26.30%
Pseudo label(PL)	26.50%	28.61%
PL+LS	29.70%	31.29%

3.4.2. Bottom-up systems

As explained in section 2.1, event boundary detection is performed either using derivative or Kalman prediction error. Since these bottom-up systems do not have any class information, we compare their performance using oracle labels. For each of the detected boundaries, we use the label from the reference event list based on event proximity. Table 2 shows performances of these two variations with oracle labels and a third system with boundaries from PL+LS system paired with oracle labels. Both Derivative based system and Kalman filter based system have more accurate boundaries than PL+LS boundaries. These results indicate that the bottom-up systems can improve performance of the top-down systems as these oracle results are higher than purely top-down systems. Kalman prediction based system performs slightly poorly compared to the derivative based method.

Table 2: Macro F-score for bottom-up systems with oracle labels

Method	Eval 2018	Validation
Derivative	44.16%	45.32%
Kalman Filter	43.25%	41.52 %
PL+LS boundaries	36.05%	37.27%

3.4.3. Integrated systems

We combine event boundaries from the bottom-up with two variations of top-down systems. First the top-down system using pseudo labels and linear softmax(PL+LS) is used for the supervision. Next we use an ensemble of baseline, pseudo label and PL+LS system using majority voting criterion. Table 3 shows the results of these various combinations. As indicated by these results, using boundaries from a bottom-up system improves the performance of PL+LS system by 3%. The ensemble system gives an additional gain of 0.5%.

Table 3: Macro F-score for integrated systems

Bottom-up	Top-down	Validation
Derivate	PL+LS	34.61%
Derivate	Ensemble	35.31%
Kalman Filter	PL+LS	34.37 %
Kalman Filter	Ensemble	35.00 %

4. CONCLUSION

In this work we presented a hybrid event detection method that integrates boundaries from a bottom-up analysis system and event labels from a top-down prediction system. We improve the boundary detection by taking advantage of class information. Boundary detection is used to expand strongly labeled data by adding boundary information to weakly labeled data. A Kalman prediction based approach was introduced as a parallel to the derivative based salience. Although the Kalman prediction based method performs slightly worse than the derivative based method, we hope further investigation in to such predictive models will expand the analysis to use memory or statistics of the history. The proposed system improves F-score by 11% compared to the baseline system.

5. REFERENCES

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *CoRR*, vol. abs/1610.05256, 2016. [Online]. Available: http://arxiv.org/abs/1610.05256

[2] G. Saon, T. Sercu, S. J. Rennie, and H. J. Kuo, "The IBM 2016 english conversational telephone speech recognition system," *CoRR*, vol. abs/1604.08242, 2016. [Online]. Available: http://arxiv.org/abs/1604.08242

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5329–5333.

[4] I.-Y. Jeong and H. Lim, "Audio tagging system for dcase 2018: focusing on label noise, data augmentation and its efficient learning," Tech. Rep., DCASE2018 Challenge, Tech. Rep., 2018.

[5] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Iterative knowledge distillation in r-cnns for weakly-labeled semi-supervised sound event detection," 11 2018.

[6] D. Wang, L. Zhang, C. Bao, K. Xu, B. Zhu, and Q. Kong, "Weakly supervised CRNN system for sound event detection with large-scale unlabeled in-domain data," *CoRR*, vol. abs/1811.00301, 2018. [Online]. Available: http://arxiv.org/abs/1811.00301

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

[8] S. Kothinti, K. Imoto, D. Chakrabarty, S. Gregory, S. Watanabe, and M. Elhilali, "Joint acoustic and class inference for weakly supervised sound event detection," DCASE2018 Challenge, Tech. Rep., September 2018.

[9] Z. Podwinska, I. Sobieraj, B. M. Fazenda, W. J. Davies, and M. D. Plumbley, "Acoustic event detection from weakly labeled data using auditory salience," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 41–45.

[10] S. Kothinti, K. Imoto, D. Chakrabarty, G. Sell, S. Watanabe, and M. Elhilali, "Joint acoustic and class inference for weakly supervised sound event detection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 36–40.

[11] N. Huang and M. Elhilali, "Auditory salience using natural soundscapes," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, p. 2163, 2017. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.4979055http://www.ncbi.nlm.nih.gov/pubmed/28372080

[12] E. M. Kaya and M. Elhilali, "Investigating bottom-up auditory attention," *Frontiers in Human Neuroscience*, vol. 8, p. 327, 2014. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2014.00327/abstracthttp://www.ncbi.nlm.nih.gov/pubmed/24904367http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4034154

[13] 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.

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

[15] G. W. Taylor, L. Sigal, D. Fleet, and G. E. Hinton, "Dynamical binary latent variable models for 3d human pose tracking," Proc. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp. 631–638, 06 2010.

[16] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 31–35.

[17] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 1–17, 2016.

[18] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.