ABNORMAL SOUND EVENT DETECTION USING TEMPORAL TRAJECTORIES MIXTURES

Debmalya Chakrabarty *Mounya Elhilali*

Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore, MD, USA.

ABSTRACT

Detection of anomalous sound events in audio surveillance is a challenging task when applied to realistic settings. Part of the difficulty stems from properly defining the 'normal' behavior of a crowd or an environment (e.g. airport, train station, sport field). By successfully capturing the heterogeneous nature of sound events in an acoustic environment, we can use it as a reference against which anomalous behavior can be detected in continuous audio recordings. The current study proposes a methodology for representing sound classes using a hierarchical network of convolutional features and mixture of temporal trajectories (MTT). The framework couples unsupervised and supervised learning and provides a robust scheme for detection of abnormal sound events in a subway station. The results reveal the strength of the proposed representation in capturing non-trivial commonalities within a single sound class and variabilities across different sound classes as well as high degree of robustness in noise.

Index Terms— Anomalous sound events, Hierarchical network, Convolutional feature representation, Mixture of temporal trajectory models

1. INTRODUCTION

Defining 'abnormal' behavior in an audio recording is a challenging task. First of all, there is no universal definition of what *abnormality* means. Second, even what is *normal* cannot be easily defined given the complex nature of sound sources in realistic scenarios. To date, most research efforts in anomaly detection have mainly focused on detection of isolated events in continuous recordings such as shouts [1], screams [2], laughs [3], gunshots and explosions [5] etc. However, for setting up a surveillance system in an environment like a train or subway station, detecting abnormalities based on examining isolated events becomes highly inefficient since collections of such isolated events can overlay normal behavior. Instead, we consider the problem of obtaining a good model representation of normal behavior in the environment. We are particularly interested in models that can capture non-trivial commonalities across various sound events as well as their interactions in the context of a complex scene.

Modeling acoustic scene behavior ultimately reduces to a choice of feature representation and learning model that can best characterize the myriad events that can be encountered in acoustic scenes. Mel-Frequency Cepstral Coefficients (MFCC) are the most widely used representation in acoustic event detection tasks. They provide a compact and efficient mapping of the spectral characteristics of simple scenes [6, 7, 8]. Unfortunately, their performance does not generalize to real world environments which are inherently dynamic and often corrupted by noise. In order to accurately report the intricacies of such realistic scenarios, it is imperative that any modeling of acoustic characteristics captures both spectral and temporal nuances of the signal over multiple resolutions and time-constants [9, 10]. Work in this direction has often employed two-dimensional time-frequency filter-banks using Gabor filters, localized Fourier bases or even biomimetic spectro-temporal receptive fields [11]. In [12], Lee *et al.* reported a localized and rich tiling of the spectro-temporal space of sound classes derived from unsupervised learning of unlabeled data in the context of *Restricted Boltzmann Machines* (RBM) [13]. In the current work, we build on this rich basis set; and extend applicability of unsupervised learning using RBMs to the problem of anomaly detection in audio recordings.

Operating on this feature analysis often comes a robust backend classifier whose role is to capture variability across different instances of the sound class. Unsupervised classifiers like Support vector machines (SVM) and Gaussian mixture models (GMM) have proved to be very efficient in modeling the mean statistics of analytical audio features in tasks of scream, laughter and gunshot detection [14, 15]. These models do provide well defined average representation of isolated events but fail to capture the information contained in the temporal dynamics of these events. In contrast, HMM based models are capable of capturing such temporal trajectories [15]. However, because of their markovian constraint, they become inefficient in modeling the long term temporal dependencies across events essential to obtain a global context of an acoustic scene. Recent work started using more representationally powerful generative models based on dis-

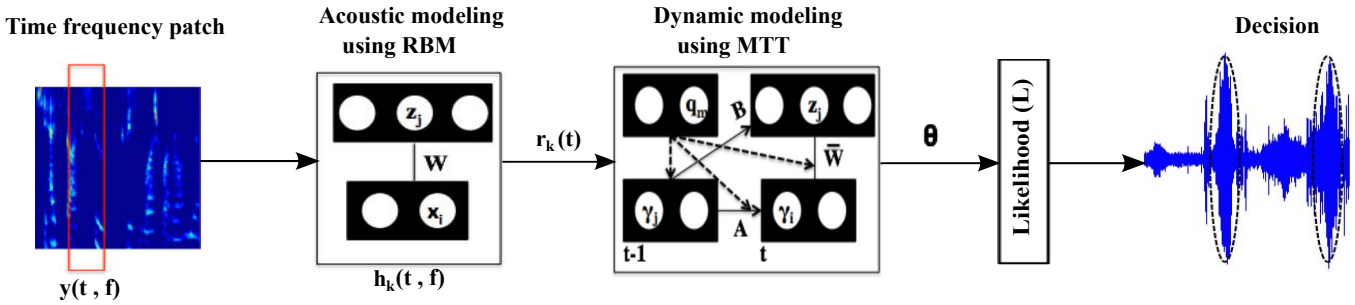

Fig. 1. *Block diagram of MTT based abnormal sound event detection*

tributed hidden states, such as Conditional RBMs [16] to learn representation of temporal dynamics from data rather than explicitly modeling them under hard wired assumptions. In the current study, we develop a hybrid RBM-CRBM scheme for modeling normal acoustic behavior in a subway station. An "event" such as normal conversation among riders is typically comprised of multiple sub-events like speech, laugh, cheerful banter etc., each having its own set of spectral and temporal dynamics. In order to capture these different modes of temporal dynamics as well as their interactions and transitions across each other, we propose a *mixture of dynamic trajectories* that can decompose the global temporal space of a normal event into multiple trajectories, each of which belongs to a semantically different sub-event. We develop an integrated framework of learning the localized spectro-temporal attributes in an unsupervised fashion as well as capturing their different modes of temporal trajectories by using a set of mixtures of temporal trajectories (MTTs). The framework flags as '*abnormal*' events that don't fall within the span of learned trajectories.

The organization of this paper is as follows: Section 2 provides a detailed description of the proposed methodology using a hybrid RBM-MTT framework. Section 3 outlies the experimental setup and event detection results, while section 4 provides conclusion and discussion of the results.

2. METHOD

Our proposed framework for abnormal sound event detection comprises 3 main processing blocks; acoustic modeling using RBM, dynamic modeling using MTT and finally using these models for abnormal sound event detection as shown in Figure 1. The proposed system operates on time frequency representation of acoustic signal. A time-frequency auditory spectrogram $y(t, f)$ is extracted from each audio file based on a model of peripheral processing in the mammalian auditory system [17]. The spectrogram representation $y(t, f)$ is sampled with frame size of 10 ms. 10 consecutive frames are then grouped together to form a one dimensional vector x in a process of shingling [18]. A dataset of n sampled patches given by $X = x^1, x^2, \ldots, x^n$ is formed, where $x^{(i)} \in R^N$ and N = 1280 in our case.

2.1. Acoustic modeling using RBM

We use Sparse restricted Boltzmann machine (RBM) as the unsupervised learning algorithm to discover features from the unlabeled dataset X. Sparse RBMs are undirected graphical models with K binary hidden variables [19]. We train the first layer RBM representations comprised of 400 hidden units using the contrastive divergence (CD) approximation with same type of hyper-parameters and sparsity penalty as used in [20]. The training produces the weights W_k for $k = 1, 2, \ldots, 400$ which are a representation of localized spectro-temporal attributes. In order to get a representation similar to localized 2D filters, we transform these one dimensional weights W_k into $h_k(t, f)$ where t = 10 and f = 128. We apply these 2D filters over the time-frequency patch $y(t, f)$ extracted from the labeled dataset of normal conversations to obtain filter responses $\mathbf{r}_k(t)$ given by:

$$\mathbf{r}_k(t) = \sum_f \int y_l(\tau, f) h(t - \tau, f) d\tau \qquad (1)$$

Filter responses $\mathbf{r}_k(t)$ are used as our feature representation for the next processing block.

2.2. Dynamic modeling using MTT

Next, a mixture of CRBMs (mCRBM) [21] is proposed as a *dynamical* mixture model to decompose the global temporal space of a normal event into multiple trajectories, where each such trajectory belongs to a particular sub-event. A *dynamical* mixture model can be created by introducing a mixture component variable, $\boldsymbol{q}$, with M possible states [21]. The dynamical model is defined by a joint distribution:

$$p(\boldsymbol{\gamma}_t, \boldsymbol{z}_t, \boldsymbol{q}_t | \boldsymbol{\gamma}_N) = exp(-E(\boldsymbol{\gamma}_t, \boldsymbol{z}_t, \boldsymbol{q}_t | \boldsymbol{\gamma}_N))/Z(\boldsymbol{\gamma}_N) \qquad (2)$$

where $\boldsymbol{\gamma}_t$ is real valued representation of current filter response, $\boldsymbol{z}_t$ is a collection of binary hidden units such that $z \in (0, 1)$, and $\boldsymbol{\gamma}_N$ contains the history of past N filter responses to provide a way for capturing the long term temporal

dependencies across the responses. The energy function $\boldsymbol{E}$ is given by:

$$E(\boldsymbol{\gamma}_t, \boldsymbol{z}_t, \boldsymbol{q}_t|\boldsymbol{\gamma}_N) = \frac{1}{2}\sum_i(\boldsymbol{\gamma}_{it} - \hat{c}_{it})^2 - \sum_j \boldsymbol{z}_{jt}\hat{d}_{jt} - \sum_m \boldsymbol{q}_{mt}\sum_{i,j}\bar{W}_{ij}\boldsymbol{\gamma}_{it}\boldsymbol{z}_{jt} \quad (3)$$

where $\bar{W}$ captures the interactions between the filter responses and hidden variables and the dynamical terms $\hat{c}_{it}$ and $\hat{d}_{jt}$ are linear functions of previous N filter responses $\boldsymbol{\gamma}_N$, given by:

$$\hat{c}_{it} = \sum_m \boldsymbol{q}_{mt}\left(C_{im} + \sum_l A_{ilm}\boldsymbol{\gamma}_{lN}\right)$$
$$\hat{d}_{jt} = \sum_m \boldsymbol{q}_{mt}\left(D_{jm} + \sum_l B_{ilm}\boldsymbol{\gamma}_{lN}\right) \quad (4)$$

where C and D are static biases and A and B are autoregressive model parameters. The parameter set $\theta = (\bar{W}, A, B, C, D)$ of mCRBM are learned using contrastive divergence (CD) approximation. We refer the reader to [21] for details of learning mCRBM by CD. This learned parameter set θ becomes our representation of mixture of temporal trajectories (MTT) models. We use M=10 assuming a mixture of 10 components can span the entire temporal trajectory space of a single event and use 200 hidden units in our mCRBM architecture.

2.3. Abnormal Sound Event Detection

In the detection stage, we use the measure of log-likelihood score of a given test frame under our learned MTT model to decide whether the frame under consideration belongs to an abnormal event or normal conversation [21]. A test audio signal is processed through the learned RBM weights to obtain feature representation $\mathbf{r}_k(t)$ as per equation 1. On applying the parameter set θ over $\mathbf{r}_k(t)$, we obtain a log likelihood score L given by:

$$L = log(p(\mathbf{r}_t|\mathbf{r}_N;\theta)) = log(\sum_{\mathbf{z}_t,\mathbf{q}_t} p(\mathbf{r}_t, \mathbf{z}_t, \mathbf{q}_t|\mathbf{r}_N)) \quad (5)$$

We compare this likelihood score L with a threshold value obtained from development set and we label the frame as 'normal' if $L >$ threshold or 'abnormal' if $L <$ threshold.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Data

We prepare an unlabeled training dataset by randomly mixing the recordings from both TIMIT [22] and BBC sound effects library [23] to train our first layer RBM bases. BBC sound effects library contain classes like Ambience, Animals, Office, Transportation and Musical etc. Because of such heterogeneity across the scenes, RBM weights are not biased towards one particular kind of scene. The dataset used for abnormal sound events detection contains recordings of audio events in a metro station [24]; the duration of each file ranging from 1 minute to about 6 minutes. We resample each recording in the dataset to 8 KHz and preprocess them through a pre-emphasis filter with coefficients $[1 - 0.97]$ in order to boost the high frequencies. The recordings contain events like normal speech, laughter, cheerful banter etc. annotated as *normal conversation*. The frames belonging to normal events are split randomly into 80 % for training the MTT models and rest 20 % as development and test set. The recordings also contain events like train passing by, shout, scream, fights, aggressive behavior etc. which we consider as 'abnormal' in our analysis and include them in the test set for detection.

3.2. System variants

The performance of an abnormal sound event detection system depends on how good our model representation is. The key aspect of our model representation is based on a set of mixtures of temporal trajectories capturing the interactions and transitions across multiple events in a complex acoustic scene. In order to quantify its importance and effect on system performance, we contrast our proposed system against 3 system variants based on similar generative framework and backbone architecture but with variabilities in mixture components and trajectory representations. In one case, we train our MTT model using $M = 1$ to see how the performance of the detection system changes when a single mixture component is used to model different modes of temporal trajectories existing within a single event. Secondly, in order to quantify the importance of temporal trajectories based representation, we build a detection system by replacing mCRBM block with a regular RBM that models only the localized spectro-temporal modulations without any information of long term temporal dependencies. Our final system is based on learning first layer RBM bases only from normal conversation and use these learned bases for detecting the abnormal events.

3.3. Results and Analysis

Figure 2 shows the ROC for each of the detection systems by including/excluding the MTT stage as well as varying the number of mixtures capturing the temporal trajectories. The figure shows that our proposed system using MTT model with $M = 10$ performs the best in terms of true positive rate. When MTT is replaced by a RBM layer in the framework, we see that the detection performance of the system degrades because of incapability of RBM based representation in capturing the long term temporal dependencies. Single layer RBMs

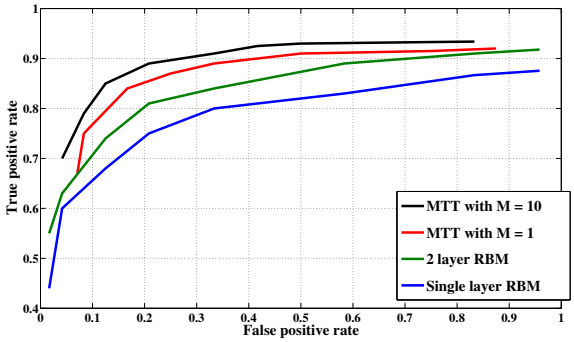

Fig. 2. *ROC curves for 4 systems regarding to detection of abnormal sound events*

trained only on normal conversations gives the worst performance in terms of true positives; the main reason being the first layer RBM bases trained on a small set of data are not able to capture a good representation of localized spectro-temporal attributes. As a result of this poor characterization, we get a lot more false negatives for this system compared to other systems. For MTT model with $M = 1$, the observation gets interesting. We see that its detection performance is better than RBM based systems, thus illustrating the importance of long term temporal dependencies over short term temporal structure for better characterization of sound events. However, the true positive rate for this system decreases when compared to MTT model with $M = 10$. This observation is mainly accounted for by the fact that due to presence of different modes of temporal trajectories within an event of normal conversation, MTT with $M = 1$ fails to span the entire temporal trajectory space of such a broader class. As a result, when an event like *laughter* occurs in a continuous audio, the system detects it as an abnormal event even though it is labeled as *normal conversation*.

To provide more insight into the idea of MTTs capturing different modes of temporal trajectories, we apply our MTT model to a sample recording of normal conversation among riders in a subway station. At several points during the conversation, other than normal speech, there are instances of *laugh*, *excitement* etc. which are non stationary events having their own set of dynamics. In our experimental analysis, we find that frames belonging to the instances of *laugh* and *excitement* are assigned to components 1, 4 and 7 with an average probability of 0.9572; while component 9 captures the temporal trajectories of normal speech in the conversation with an average probability of 0.9851. This probabilistic assignment of frames to different components of MTT confirms our intuition that MTT with desired number of components is able to segment an event with different modes of temporal trajectories into statistically salient sub-events.

We further test the robustness of the proposed system by adding noise from NOISEX-92 database [25] to the test set

SNR (dB)	F-measure (%) for 4 detection systems			
	MTT (M=10)	MTT (M=1)	2 layer RBM	Single layer RBM
Clean	**93.11**	89.12	86.55	78.77
20 dB	**92.03**	85.41	79.66	71.82
10 dB	**88.85**	80.15	72.99	64.88
0 dB	**65.77**	59.87	51.66	43.77
-5 dB	**50.76**	43.28	34.88	25.75
-10 dB	**42.36**	34.88	25.77	10.99

Table 1. Abnormal sound events detection results for 4 systems at different SNR levels

at different SNR levels of 20, 10, 0, -5 and -10 dB. The performance of the systems are measured in terms of percentage F-measure. We see from Table 1 that MTT (M=10) based detection system not only outperforms the other three system variants in clean scenario but exhibits robustness in presence of noise as well. When noise level increases, the detection performance of our MTT based system degrades at a much lower rate compared to the other three system variants. We also observe that for upto 10 dB SNR, our proposed system gives a very satisfactory performance in detecting the abnormal sound events. Another interesting point to note from Table 1 is that even MTT models with $M = 1$ performs better than RBM based models for all noise cases. This clearly shows the importance of incorporating the information of temporal trajectories along with localized spectro-temporal attributes in the model representation of sound events for a robust characterization..

4. CONCLUSION

In this work, we develop a hybrid RBM-MTT framework for abnormal sound event detection in subway station by using a joint representation of localized spectro-temporal attributes with mixtures of temporal trajectories. Such a joint representation is very effective in capturing the intricate details and commonalities across a broader sound class spanned by multiple events. We show that MTT as a *dynamical* mixture model spans the complete temporal trajectory space of a complex acoustic scene by decomposing it into multiple trajectories, each of which belongs to a particular sub-event. In abnormal sound event detection task, the detection accuracy improves by an absolute 7 % over RBM class of models when information of different modes of temporal dynamics is incorporated in model representation of sound objects via our proposed MTT. We also find that our MTT based representation augments the detection system with high degree of noise robustness at low SNR levels, thus illustrating the fact that the joint representation provides a much robust characterization of broader sound classes.

5. REFERENCES

[1] V.K. Mittal and B. Yegnanarayana, "Production features for detection of shouted speech," in *Consumer Communications and Networking Conference (CCNC), 2013 IEEE*, Jan 2013, pp. 106–111.

[2] M.K. Nandwana, A. Ziaei, and J.H.L. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," in *Proceedings of ICASSP'15*, April 2015, pp. 161–165.

[3] M. Knox and Nikki Mirghafori, "Automatic laughter detection using neural networks," in *Interspeech*, Antwerp, Belgium, August 2007.

[4] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proceedings of ICASSP'09*, April 2009, pp. 165–168.

[5] X.Zhuang, X.Zhou, T.Huang, and M.Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," in *Proceedings of ICASSP'08*, 2008.

[6] O. Kalini, S. Sundaram, and S. Narayanan, "Saliency driven unstructured acoustic scene classification using latent perceptual indexing," in *Proceedings of MMSP'09, Rio de Janiero, Brazil*, October 5-7, 2009.

[7] W. Nogueira, G. Roma, and P. Herrera, "Sound scene identification based on MFCC, binaural features and a support vector machine (SVM) classifier," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[8] S. Nemala and M. Elhilali, "Relevant spectro-temporal modulations for robust speech and nonspeech classification," 2010.

[9] C.V. Cotton and D.P.W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, Oct 2011, pp. 69–72.

[10] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proceedings of Eurospeech*, 2003, pp. 2573–2576.

[11] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems 22*, pp. 1096–1104. Curran Associates, Inc., 2009.

[12] Geoffrey E. Hinton, "Learning multiple layers of representation," *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2015/09/18.

[13] Weimin Huang, Tuan Kiang Chiew, Haizhou Li, Tian Shiang Kok, and J. Biswas, "Scream detection for home applications," in *Proceedings of ICIEA'10*, June 2010, pp. 2115–2120.

[14] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proceedings of AVSS '07*, Washington, DC, USA, 2007, pp. 21–26.

[15] Graham W Taylor, Geoffrey E Hinton, and Sam T Roweis, "Modeling human motion using binary latent variables," in *Advances in neural information processing systems*, 2006, pp. 1345–1352.

[16] 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.

[17] J. Salamon and J.P. Bello, "Unsupervised feature learning for urban sound classification," in *Proceedings of ICASSP'15*, April 2015, pp. 171–175.

[18] Marc'aurelio Ranzato, Christopher S. Poultney, Sumit Chopra, and Yann Lecun, "Efficient Learning of Sparse Representations with an Energy-Based Model," in *Neural Information Processing Systems*, 2006, pp. 1137–1144.

[19] K Cho, T Raiko, and A Ilin, "Enhanced gradient for training restricted boltzmann machines," *Neural Computation*, vol. 25, no. 3, pp. 805–831, March 2013.

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

[21] William M. Fisher, George R. Doddington, and Kathleen M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," in *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[22] "The bbc sound effects library original series," *http://www.soundideas.com*, May 2006.

[23] W. Zajdel, J.D. Krijnders, T. Andringa, and D.M. Gavrila, "Cassandra: audio-video sensor fusion for aggression detection," in *Proceedings of AVSS '07*, Sept 2007, pp. 200–205.

[24] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," *Tech.Rep., Speech Research Unit, Defense Research Agency, Malvern, U.K.*, 1992.