INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Towards Smart-cars that can Listen: Abnormal Acoustic Event Detection on the Road

Mahesh Kumar Nandwana, Taufiq Hasan

Research and Technology Center (RTC), Robert Bosch LLC
4005 Miranda Ave, Palo Alto, CA, USA

Mahesh.Nandwana@gmail.com, Taufiq.Hasan@us.bosch.com

Abstract

Even with the recent technological advancements in smart-cars, safety is still a major challenge in autonomous driving. State-of-the-art self-driving vehicles mostly rely on visual, ultrasonic and radar sensors to assess the surroundings and make decisions. However, in certain driving scenarios, the best modality for context awareness is environmental sound. In this study, we propose an acoustic event recognition framework for detecting abnormal audio events on the road. We consider five classes of audio events, namely, ambulance siren, railroad crossing bell, tire screech, car honk, and glass break. We explore various generative and discriminative back-end classifiers, utilizing Gaussian Mixture Models (GMM), GMM mean supervectors and the I-vector framework. Evaluation results using the proposed strategy validate the effectiveness of the proposed system.

Index Terms: Acoustic event recognition, Gaussian mixture model, I-vector, autonomous driving, audio classification.

1. Introduction

In recent years, there has been an increasing interest in self-driving car technology from industry researchers across the globe. While this technology has advanced rapidly, there are still concerns regarding safety. Current self-driving cars rely heavily upon visual, ultrasonic and radar sensors to understand the environment and make decisions. However, in some cases, recognizing the environmental sound can be an important indication for a potentially hazardous situation on the streets. One can think of multiple events that may be of interest to the smart-car, including siren from an ambulance, or a tire screeching sound nearby. Until now, the audio signal has mostly been used inside the car for speech recognition. In this work, we propose the use of audio sensors for detection of abnormal events on the road.

There are several advantages of audio-based detection compared to other sensors. Firstly, the microphone is a relatively low-cost sensor. Secondly, audio can be useful in situations where other sensors may fail, for example, in the case of darkness, fog or other low visibility conditions. Thirdly, with modern cars already equipped with embedded speech recognition engines, additional computational processing for event detection may be achieved relatively easily.

In the past, researchers have used acoustic event detection for various purposes such as scream and gunshot detection [1], non-speech sound detection [2, 3], and in acoustic surveillance [4]. Early work on non-speech audio event detection considered hidden Markov model (HMM) and support vector machines (SVM) [5]. Much previous work deals with the application of surveillance in general [6, 7, 8]. A major challenge faced by the research community in this domain is the lack of a standardized dataset. Recently, the RWCP (Real-world computing partnership) Sound Scene Database in Real Acoustical Environments (RWCP-SSD) [1] dataset has been used by researchers for audio event detection [9, 10, 11]. Although this corpus provides a reasonable number of event classes, it only provides short impulsive sound events. Understandably, techniques that are effective on this corpus may not be generalized on other datasets.

In this paper, we aim to detect unusual audio events which can occur in the surroundings of a car while driving. We consider five different events, namely, siren, railroad crossing bell, tire screech, car honk, and glass break. We use standard cepstral features in the front-end. For modeling and classification, we consider a number of techniques based on Gaussian mixture models (GMM), inspired by techniques developed in the speaker recognition community [12]. These systems include GMM and universal background model (UBM) framework, Gaussian supervector method, and the i-vector framework [13]. To the best of our knowledge, this is the first work in the area of acoustic event detection for environment monitoring outside the car.

This paper is organized as follows. Sec. 2 describes the corpus collection used for this study. Sec. 3, describes the acoustic features and various back-end classifiers used for this task. In Sec. 4, the experimental setup and results are discussed and summarized. Finally, concluding remarks are presented in Sec. 5.

2. Corpora

A number of audio events can occur on the road while driving. We, human drivers, routinely consider these audio events along with visual cues for driving decisions, or alertness. It may be emphasized that, even though the considered events are proposed with self-driving cars in mind, these can also be helpful to warn the driver in standard cars. This can be effective while the driver is listening to loud music while driving. We consider the following five events in this work.

- **Siren:** This is one of the most common and important events that can occur while driving. There are different types of sirens for a police vehicle, fire truck and ambulance. For this work, we have considered all types of sirens in this category. This signal is of periodic nature, repeating a specific set of ascending and descending tones.

[1] http://research.nii.ac.jp/src/en/RWCP-SSD.html

http://dx.doi.org/10.21437/Interspeech.2016-1366

- **Glass Breaks:** This event occurs either in the case of theft/burglary or in an accident.
- **Horn:** Horn is an inseparable component of any automobile. Horn is used to warn others of the vehicle's approach or presence, or to call attention to a potential hazard.
- **Railroad Crossing:** Railroad crossing signs usually convey the message of "stop, look and listen". However, in the case of low visibility conditions, the distinct bell sound at a railroad crossing can be an effective cue for an approaching train.
- **Screech:** Car screech is a high pitch sound which can be heard if another vehicle has a skidding tire. This is clearly a warning sign to the nearby drivers of a slippery road or a potential accident.

At present, there is no publicly available audio dataset for the above audio events. Accordingly, for this study, we collected audio data from the web repositories, sound libraries and YouTube. All the recordings were converted to single channel audio segments with an 8kHz sampling rate. The number of audio segments collected from each event, their total duration and how they are utilized for training and test for our experiments, are described in Table 1.

Table 1: Description of collected audio event corpus for the five classes. Audio segments have a duration of about $1 \sim 4$ seconds. Number of training and test segments are shown.

Event Name	Total duration (min)	# Segments train	# Segments test
Siren	13	100	130
Glass Break	8	100	105
Horn	7	100	103
Railroad	9	100	130
Screech	6	100	102
Total	43	500	570

3. Features and Classifiers

3.1. Acoustic features

In this work, we utilize the well-known Mel-frequency cepstral coefficients (MFCC) [14]. 39 dimensional features are extracted from overlapping frames of 25ms duration. 13 static coefficients are computed including C_0, and the velocity (Δ) and acceleration ($\Delta + \Delta$) coefficients are appended. Unlike in speech processing, we do not remove the silent, or low energy frames.

3.2. The GMM-UBM framework

This framework was originally proposed for speaker verification [15]. At first, a GMM based universal background model (UBM) is trained using all of the training data from five different classes. This serves as a generic audio event model for these five classes. In the next step, maximum a-posteriori (MAP) adaptation is performed on the mean vectors of the GMM-UBM using the training data of each class.

For a given set of acoustic features $\mathcal{X} = \{\mathbf{x}_n | n \in 1 \cdots T\}$, a GMM-UBM model λ_0 with M Gaussian components is represented as:

$$f(\mathbf{x}_n|\lambda_0) = \sum_{g=1}^{M} \pi_g \mathcal{N}(\mathbf{x}_n|\mu_g, \mathbf{\Sigma}_g), \tag{1}$$

where, π_g, μ_g and $\mathbf{\Sigma}_g$ indicate the weight, mean vector, and covariance matrix of the g-th mixture component. The UBM, λ_0, is independent of an audio class since it is trained on all of the classes using the expectation maximization (EM) [16] algorithm.

In order to adapt this model towards a specific audio event class, we utilize the methods in [15]. First, we define $\gamma_n(g) = p(g|\mathbf{x}_n, \lambda_0)$ as the posterior probability of the g-th Gaussian component given a feature vector. For a set of features $\mathcal{X}_c$ that belong to class c, we then compute the zero and first order Baum-Welch statistics as:

$$N_c(g) = \sum_{n=1}^{T} \gamma_n(g) \tag{2}$$

$$\mathbf{F}_c(g) = \sum_{n=1}^{T} \gamma_n(g)\mathbf{x}_n \tag{3}$$

Using these parameters, the posterior mean given the data vectors $\mathcal{X}_c$ is given by:

$$E_g[\mathbf{x}_n|\mathcal{X}_c] = \frac{\mathbf{F}_c(g)}{N_c(g)} \tag{4}$$

Utilizing the MAP adaptation for mean vectors alone [17, 15], the updated mean vectors of the GMM are given by:

$$\hat{\mu}_g = \alpha_g E_g[\mathbf{x}_n|\mathcal{X}] + (1 - \alpha_g)\mu_g \tag{5}$$

where, the parameter α_g controls how the adapted GMM parameter will be affected by the observed audio event data from class c. It is computed as:

$$\alpha_g = \frac{N_c(g)}{N_c(g) + r} \tag{6}$$

where $r > 0$ is known as the relevance factor. Thus, using the adapted mean vectors $\hat{\mu}_g$, new GMM models λ_c are generated for each audio class c, while the weight and covariance matrices are kept the same as the UBM λ_0. During the evaluation phase, the likelihood of the test feature vectors are computed across the GMM models from each audio class, and the maximum scoring class is selected as the predicted class. In this work, we use $M = 16$ mixture components and relevance factor, $r = 16$.

3.3. The GMM-supervector system

This method generates a super-vector by concatenating the adapted GMM mean vectors extracted from different audio segments using MAP adaptation [18]. These super-vectors are then used as features for other classifiers. However, since the supervector dimensions are large, some form of dimensionality reduction is usually used for efficient processing.

In this work, we first train the UBM as in the previous section. Next, we compute the adapted mean vectors from each training audio segment using (5). GMM mean super-vector (SV) $\mathbf{m}_s$ from audio segment s, is then computed by concatenating the adapted mean vectors obtained from each mixture components, as follows:

$$\mathbf{m}_s = [\hat{\mu}_1^T \; \hat{\mu}_2^T \cdots \; \hat{\mu}_M^T]^T. \tag{7}$$

Here $\hat{\mu}_g$ are the adapted mean vectors. These SVs are extracted from each of the training segments from the five classes. With 16 mixture components and 39 dimensional acoustic features, we have the SV dimension of 624. Next, we perform mean and variance normalization of each component of the SV to have zero mean and unit variance. Afterwards, we divide each SV by it's own l^2 norm for length normalization [19]. Using these normalized training SVs, we train a probabilistic principal component analysis (PPCA) model [20] and project the SV on the first 100 principal components for dimensionality reduction. These 100 dimensional vectors are then used to learn a linear discriminant analysis (LDA) projection matrix using the labels of the five classes. LDA reduces the dimension of these vectors further to 4. All the normalization parameters and projections are learned from the training data and applied on the test data during evaluation. Finally, for classification, we measure the cosine distance (CD) between the training and evaluation vectors. CD score obtained for each test audio segment across the training segments are averaged to obtain the classification score. The highest scoring class is assigned as the predicted class by the system.

3.4. The i-vector system

This framework is similar to the super-vector framework discussed previously. Originally proposed in [13], this method utilizes a variant of the PPCA algorithm to reduce the dimensionality of the GMM super-vectors. Essentially, this method assumes that the variability of the different acoustic events lies in a lower dimensional subspace defined by a total variability space. Accordingly, GMM super-vector extracted from an audio segment is represented by the following factor analysis model [13]:

$$\mathbf{m} = \mathbf{m}_0 + \mathbf{T}\mathbf{w}. \tag{8}$$

Here, $\mathbf{m}_0$ is the UBM super-vector, $\mathbf{T}$ is the total variability matrix, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are the total factors. The posterior mean of the hidden variable $\mathbf{w}$ is known as the i-vector, or identity vector. The $\mathbf{T}$ matrix is estimated using an EM algorithm [21].

In our proposed framework, we train the $\mathbf{T}$ matrix using all the training data. The i-vector dimension is set at 200. The i-vectors are normalized to have unit length [19], and their dimension is reduced to 4 using an LDA projection matrix as in the super-vector system. The final classification is again performed using the cosine similarity measure.

4. Experiments and Results

The collected audio event data described in Sec. 2 is utilized for the evaluation. 100 audio segments from each class is selected for training and the remaining ones are retained for testing. The performance of each individual system is evaluated using overall classification accuracy and one-vs-rest accuracy for each class. The results are shown in Table 2.

From the results, we observe that the GMM-SV system provides the best performance across different classes, with an overall accuracy of 87.54%. The performance difference between I-vector and GMM-UBM system is not significant. This is somewhat surprising since the GMM-UBM system uses a generative model, whereas the I-vector system utilizes a discriminative projection via LDA. This is possibly due to lack of sufficient data for training the $\mathbf{T}$ matrix for i-vector extraction. Typically, a very large speech corpus is used to train this hyper-parameter for speaker recognition applications. However, for audio event detection, we were not able to find a relevant dataset to train this matrix.

Table 2: Performance comparison of different systems for acoustic event detection on the road. Performance reported with respect to overall and one-vs-rest accuracy for five classes.

Systems	%Accuracy		
	GMM-UBM	I-vector	GMM-SV
Siren	73.08	78.46	75.38
Glass	96.19	94.29	97.14
Horn	86.41	94.17	98.06
Railroad	72.31	52.31	90.77
Screech	53.92	70.59	78.43
Overall	76.14	76.84	87.54

5. Conclusion

In this paper, we have proposed an audio event classification framework for detecting abnormal sounds on the road. We have considered five event classes, namely, siren, glass break, car horn, railroad crossing, and tire screech. Different classifiers based on Gaussian mixture models have been examined and evaluated on an in-house audio dataset accumulated from the web. Experimental results demonstrated that the GMM super-vector based framework with a discriminative back-end classifier performs well on this task.

6. References

[1] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *IEEE AVSS 2007*. IEEE, 2007, pp. 21–26.

[2] M. K. Nandwana, H. Bořil, and J. H. L. Hansen, "A new front-end for classification of non-speech sounds: A study on human whistle," in *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 1982–1986.

[3] M. Nandwana, A. Ziaei, and J. H. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," in *Proc. IEEE ICASSP*, Brisbane, Australia, April 2015.

[4] P. Transfeld, S. Receveur, and T. Fingscheidt, "An acoustic event detection framework and evaluation metric for surveillance in cars," in *Proc. Interspeech*, Dresden, Germany, 2015.

[5] J. P. Elo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, "Non-speech audio event detection," in *IEEE ICASSP*. IEEE, 2009, pp. 1973–1976.

[6] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, "Audio based event detection for multimedia surveillance," in *IEEE ICASSP*, vol. 5. IEEE, 2006, pp. V–V.

[7] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *IEEE ICME*. IEEE, 2005, pp. 1306–1309.

[8] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted hmms for unusual event detection," in *IEEE CVPR*, vol. 1. IEEE, 2005, pp. 611–618.

[9] N. Cho and E.-K. Kim, "Enhanced voice activity detection using acoustic event detection and classification," *IEEE Trans. on Consumer Electronics*, vol. 57, no. 1, pp. 196–202, 2011.

[10] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *IEEE ICASSP*. IEEE, 2015, pp. 559–563.

[11] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 3, pp. 540–552, 2015.

[12] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 99, pp. 788 – 798, May 2010.

[14] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[15] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19 – 41, 2000.

[16] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[17] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[18] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE ICASSP*, May 2006, pp. 97–100.

[19] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, Florence, Italy, Oct. 2011, pp. 249 – 252.

[20] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. of the Royal Stat. Soc.: Series B*, vol. 61, no. 3, pp. 611–622, 1999.

[21] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.