DISCRIMINATIVE ACOUSTIC EVENT RECOGNITION IN MULTIMEDIA RECORDINGS

Saurabh Khanwalkar, Guruprasad Saikumar, Amit Srivastava and Premkumar Natarajan

Raytheon BBN Technologies
10 Moulton St., Cambridge, MA, 02138
{skhanwal,gsaikuma,asrivast,pnataraj}@bbn.edu

ABSTRACT

In this paper, we describe an Acoustic Event Recognition (AER) system for locating events of interest in the audio stream of multimedia recordings. We focus on two non-speech acoustic events; bomb explosions and gunfire, which typically exist in surveillance videos and are of importance in monitoring and alerting applications. Recognition is performed using a discriminative approach based on Support Vector Machines (SVM). We compare the new approach to a baseline system that utilizes a Hidden Markov Model (HMM)-based classification approach. We performed experiments on a corpus of publicly available video files containing gunfire and explosion events. Our results show that the new discriminative approach, when configured to use a rich combination of acoustic features, achieves a high retrieval precision at a notable recall under noisy conditions. As compared to HMM-based system, we achieved 54% relative improvement in F-score for explosion recognition with 1.5% relative improvement in F-score for gunfire recognition.

1. INTRODUCTION

Video surveillance applications are becoming increasingly important both in private and public environments. The increase in the number of video cameras for surveillance and security purposes has rendered the manual detection of events of interest impractical and expensive. For this reason, research on automatic surveillance systems has recently received particular attention. A video can contain a wide variety of non-speech events including gunfire, explosion, screams and others. Automatic recognition of these events can be used for the alerting of hazardous situation at hand. In particular, the use of audio sensors in surveillance and monitoring applications has proved to be particularly useful for the detection of events like screams and gunfire [1], [2].

Audio-based surveillance stems from the field of automatic audio classification and matching. Traditional tasks in this area are speech/music segmentation and classification [3] and audio retrieval [4]. Previous approaches on the subject of acoustic monitoring include cases such as in [1] where a gunfire detection system is presented based on features derived from the time-frequency domain and a Gaussian Mixture Model (GMM) classifier. The authors use different Signal-to-Noise Ratio (SNR) during the training phase for achieving 10% and 5% false rejection and false detection rate, respectively. In [5], they report on building a parallel classification system based on GMM for the discrimination of ambient noise, scream and gunfire sounds. After a feature selection algorithm, they achieve 90% precision and an 8% false rejection rate.

The main objective of an AER system is to efficiently characterize the acoustic environment in terms of the hazardous and non-hazardous conditions while using a single microphone sensor; the goal of the system is to help/warn authorized personnel to take the appropriate action. In order for such an implementation to be useful and practical, it must offer high precision while keeping detection accuracy as high as possible under noisy conditions.

In this paper, we describe two different approaches to accurately recognize two types of acoustic non-speech events: *explosions* and *gunfire*. The first approach is a HMM-based system which uses perceptually-inspired speech features whereas the second approach is a discriminative SVM-based system that uses a larger set of traditional and novel acoustic features representing non-speech events like *explosion* and *gunfire*. We performed experiments with both these systems using publicly available surveillance data and evaluated the accuracy using widely adopted performance metrics.

2. HMM-BASED AER SYSTEM

The BBN Large Vocabulary Speech Recognition System consists of an HMM-based speech/non-speech detection component which is used for segmentation of input audio stream [6]. This component produces an acoustic event description of the input audio to segment the audio signal into regions of speech and non-speech. The component uses 14-dimensional Mel-frequency Cepstral Coefficients (MFCC), along with their derivatives as a 42-dimensional feature vector. It uses phonemes as detection units, where the speech class is modeled in terms of voiced, fricative or obstruent units and the non-speech class is modeled in terms of music, silence and noise phonetic units.

In order to use this component for AER, we adapted the system to model *explosion* and *gunfire* as additional non-speech classes along with the existing music, silence and noise phonetic classes. The explosion and gunfire models were trained using manually segmented and transcribed audio files for such non-speech events. These training transcripts were also used to re-train the language models so as

© EURASIP, 2011 - ISSN 2076-1465

to include these non-speech units in computation of bigram and trigram probabilities. Once the models were trained, Viterbi decoding was applied to recognize and produce transcriptions of explosion and gunfire events in the audio streams.

3. SVM-BASED AER SYSTEM

The HMM-based AER system uses speech perception inspired MFCC features for modeling speech as well as non-speech classes. These MFCC features are designed based on the speech spectral structure which is quite different from the non-speech acoustic events. For example, acoustic events like *explosions* and *gunfire* usually exist in the high frequency regions that are not fully resolved by the Mel-scale filter banks. To function accurately, such an AER system needs diverse and dynamic, low and high frequency acoustic features that represent such non-speech classes efficiently and help distinguish between *gunfire*, *explosion* and speech [7], [8].

In the context of a classifier design, a discriminative approach may be a better fit to recognize and separate these non-speech classes from speech. Discriminative models such as Support Vector Machines (SVM) are free of statistical and distributional assumptions and are easily scalable to high dimensional feature spaces. Despite their several significant advantages, generative models like HMM require far more labeled training samples than are usually available to model the target classes such as *explosions* and *gunfire* and are quite sensitive to distributional biases. Moreover, SVM natively support binary, categorical and continuous valued high-dimensional and sparse features which cannot be easily integrated into HMM-based system [9].

In this section, we describe a discriminative SVM-model based AER system which uses diverse low and high frequency acoustic features for modeling *explosion* and *gunfire* acoustic events.

3.1 Algorithm Flowchart

Figure 1 shows the procedural flowchart for SVM-based AER system that recognizes *explosion* and *gunfire* events in the input audio. The output of the system is a transcript with the timestamps of the recognized *explosions* and *gunfire* events. The algorithm consists of 3 major conceptual steps:

1. Discriminative AER feature extraction
2. Discriminative classification using SVM
3. Bottom-up segmental integration

Each of these steps is discussed in detail in the next sections.

3.2 Feature Extraction

In SVM-based AER system, we integrated the traditional acoustic features with some novel correlation-based features that give a measure of the periodicity and temporal energy [10]. We computed 6 diverse measures on each 60 ms audio frame at a frame rate of 30 ms:

1) *Zero-crossing rate measure* is defined as the weighted average of the number of times the speech signal changes sign within a frame.
2) *Energy measure* is the short-time energy obtained by squaring the windowed samples in the low frequency sub-band signal.
3) *Spectral flatness measure* is defined as the ratio of the geometric mean to the arithmetic mean of the power spectrum.
4) *Forward-backward autocorrelation change* is defined as the difference between the forward and backward prediction of autocorrelation coefficients computed on a frame [11].
5) *Inter-frame autocorrelation change* is defined as the difference between autocorrelation coefficients calculated from 1st half of the frame and those calculated from 2nd half of the same frame.
6) *Pitch frequency* is the frame-level estimated pitch values which may be absent for unvoiced and non-speech frames.

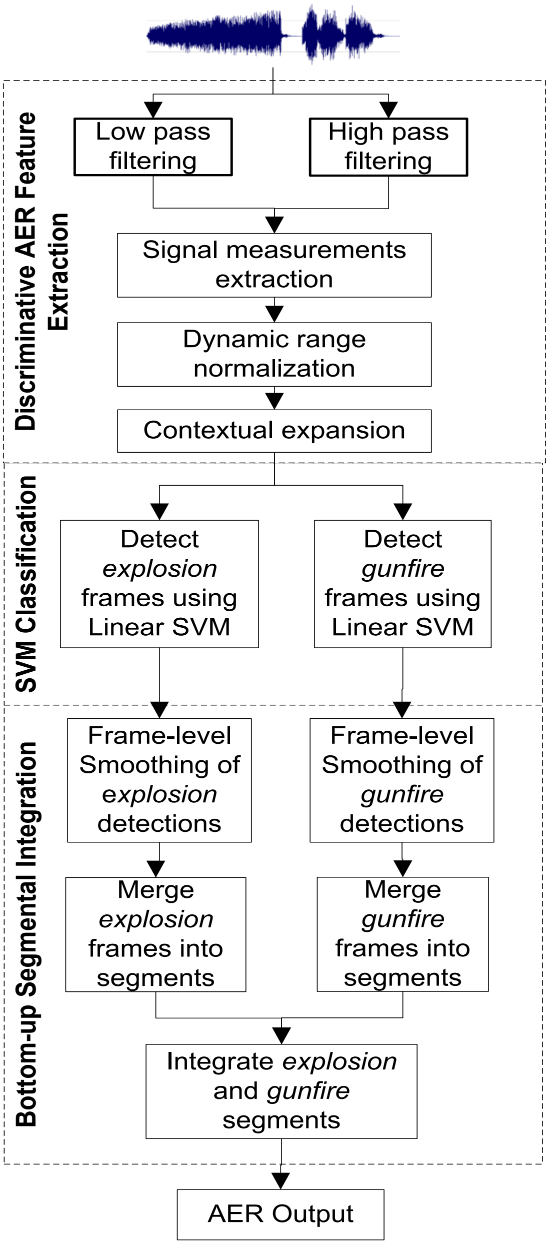
Discriminative AER Feature Extraction
Low pass filtering
High pass filtering
Signal measurements extraction
Dynamic range normalization
Contextual expansion
SVM Classification
Detect explosion frames using Linear SVM
Detect gunfire frames using Linear SVM
Bottom-up Segmental Integration
Frame-level Smoothing of explosion detections
Frame-level Smoothing of gunfire detections
Merge explosion frames into segments
Merge gunfire frames into segments
Integrate explosion and gunfire segments
AER Output

Figure 1 – SVM-based AER System Algorithm Flowchart

The acoustic events such as *explosions* and *gunfire* usually appear as loud high frequency bursts as compared to speech and music which are predominantly quasi-periodic harmonic events. In order to capture these high and low frequency patterns, we split the input audio utterance sampled at 16 KHz, into low frequency (LF) (0-4 KHz) and high frequency (HF) (4-8 KHz) sub-bands using low-pass and high-pass filter respectively. Except for zero-crossing rate and pitch measure, we computed all other measures separately on these 2 sub-band filtered signals. To characterize and compare the feature values across the 2 sub-band filtered signals and to extract the underlying dynamics; we computed absolute differences and ratios between the values and augmented the base feature vector. A total of 15 base features were then extracted per frame.

These base features were min-max normalized to bring their values into the [0, 1] range. This normalization was based on global minimum and maximum values for each feature dimension computed from the entire training set. To model the temporal variability of these spurious non-speech acoustic events, we expanded the base feature vector across frames by concatenating the center feature vector block with contextual feature vector blocks symmetrically on both sides. We choose the context window size to be large enough to reasonably capture the length of an acoustic event like *explosion* or *gunfire*. Based on preliminary experiments, we selected a context analysis window consisting of 5 neighboring frames before and 5 frames after the center frame for feature expansion. Then the final expanded feature dimension was equal to 15 x 11 = 165. This 165 dimensional feature vector per frame was used as a classification unit in SVM model, which is described in the next section.

3.3 SVM Classification

The linear SVM is a simple linear discriminative model that corresponds to the hyper-plane separating the positive class examples from the negative class examples as shown conceptually in Figure 2.

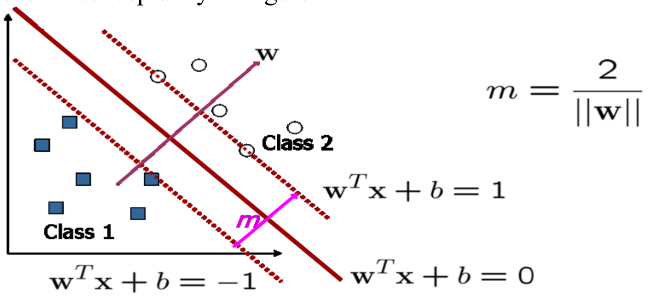
w
$m = \frac{2}{||\mathbf{w}||}$
Class 2
$\mathbf{w}^T\mathbf{x} + b = 1$
m
Class 1
$\mathbf{w}^T\mathbf{x} + b = -1$
$\mathbf{w}^T\mathbf{x} + b = 0$

Figure 2 – Large-margin discrimination for linearly separable problems

Let $\{x_1,\ldots$ be the set of n labeled training examples for a class and let $y_i \in \{1,-1\}$ be the class label of xi, which is a K-dimensional feature vector [12]. The optimal separating hyper-plane is represented as the weight vector w, found by solving the constrained optimization problem:

$$Minimize\ \frac{1}{2}\|w\|^2 \qquad (1)$$
$$subject\ to\ \ y_i\left(w^T x_i + b\right) \geq 1\ \ \forall i$$

The SVM classifier used in the AER system was based on the open-source publicly available SVM$^{\text{Light}}$ library [13]. In our AER system, we designed 2 one-versus-all SVM binary classifiers, (1) to detect *gunfire* and (2) to detect *explosions*. From the labeled *gunfire* and *explosion* examples, the linear SVM models were trained separately for both classes. Then, during decoding, incoming audio utterances were split into 60ms frames and each frame was classified as *gunfire*, *explosion* or everything else. The detection results from these binary decision classifiers were smoothened and merged via simple segmental integration methods.

3.4 Bottom-up Segmental Integration

We performed acoustic segmental integration step to merge the frame-level recognized units into segmental units. Based on the training data characteristics, we designed simple heuristic rules for frames-to-segment conversion. We used the following 2 heuristics: (1) an event must exceed a specified minimum duration of frames, and, (2) an event segment ends at the last in-class frame if a specified minimum number of consecutive out-of-class frames follow it. This step resulted into sets of *gunfire* and *explosion* segments which may or may not be overlapping in timestamps. Finally, these *explosion* and *gunfire* segments were merged into a single sequence of segments, resolving overlaps by using a simple heuristic of “longest segment wins”.

4. EXPERIMENTAL SETUP

To train the *explosion* and *gunfire* models, we selected videos publicly available on the web from liveleak.com and youtube.com.

Some keywords like bomb, IED, explosion, roadside, detonate, Iraq, army, etc were used to search for *explosion* and *gunfire* videos from these websites. The videos were originally collected in MP4, MPG, WMV and FLV formats and the audio track was then extracted and used for segmentation and acoustic event annotation. Overall, 36 files were selected for annotating *explosion* and *gunfire* acoustic events. The annotated data was split, with 90% of the data used to train the system and 10% held out as a test set. Table 1 shows the number of *gunfire* and *explosion* segments available in training and test data set.

	Training Set		**Test Set**	
Event	# Segments	Duration (sec)	# Segments	Duration (sec)
Explosion	137	154.2	14	17.1
Gunfire	961	1049.2	28	36.6
Other	4121	6094.4	345	718.0

Table 1 – Training and test data set characteristics

From the annotation, it was discovered that the training set for *explosion* class was less by one order of magnitude compared to the *gunfire* class. As compared to *explosion*, the corpus consisted of 895 seconds more *gunfire* training data. The overall training data duration for both *explosion* and *gunfire* including background speech was about 20 minutes. Similar to training data, we had twice the amount of *gunfire* as *explosion* in the test data set.

5. EVALUATION METRICS

The AER system accuracy was measured using AED-ACC metric (the first metric in CLEAR 2007 Acoustic Event Detection (AED) Evaluation [14]). The aim of this metric is to score detection of all instances of what is considered as a relevant acoustic event (AE). With this metric it is not important to reach a good temporal coincidence of the reference and system output timestamps of the AEs but to detect their instances. It is oriented to applications like real-time services for smart-rooms, audio-based surveillance, etc. AED-ACC is defined as the F-score (the harmonic mean between Precision and Recall):

$$\text{F-score} = \text{AED-ACC} = \frac{\left(1+\beta^2\right) * \text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}} \quad (2)$$

where

$$\text{Precision}=\frac{\#\text{correct system output AEs}}{\#\text{all system output AEs}} \quad (3)$$

$$\text{Recall}=\frac{\#\text{correct detected reference AEs}}{\#\text{all reference AEs}} \quad (4)$$

and β is the weighting factor that balances Precision and Recall.

For our system evaluation, the factor β was set to 1. A system output AE is considered correctly produced, if there exist at least one reference AE (with same label) whose temporal center is situated between the timestamps of the system output AE, or if the temporal center of the system output AE lies between the timestamps of at least one reference AE. A reference AE is considered correctly detected, if there exists at least one system output AE (with same label) whose temporal center is situated between the timestamps of the reference AE, or if the temporal center of the reference AE lies between the timestamps of at least one system output AE.

6. RESULTS

We first evaluated the HMM-based AER system which used only the MFCC features and Viterbi decoder. The decoder acoustic model and language model weights were optimized to yield the best performance on the test set. Table 2 below shows the Recall, Precision and AED-ACC (F-score) metrics for the HMM-based AER system. We report performance values for respective *explosion* and *gunfire* recognition. We also report the composite acoustic event recognition performance calculated from combined performance averaged over both *explosion* and *gunfire* AEs.

Event	Recall	Precision	F-score
Explosion	37.5	46.2	41.4
Gunfire	73.3	84.6	78.6
Composite	54.8	64.3	59.2

Table 2 – Performance of HMM AER System with MFCC speech features

The results indicate that HMM-based AER system gives much higher F-score for *gunfire* recognition as compared to *explosion* recognition. The balance between recall and precision values for *gunfire* recognition points out that the system generalized well for the *gunfire* class. We suspect that the poor performance in *explosion* recognition may be due to the insufficient annotated training data available for that class.

The second experiment was designed to evaluate the performance of SVM-based AER system with MFCC features used in the HMM system. We believe that the 42-dimensional MFCC feature vector does not capture the underlying non-linearity of *gunfire* and *explosion* events when used with linear SVM model. The objective of this experiment was to emphasize the importance of using high-dimensional and domain-specific features with the SVM model. Table 3 shows the performance of the system for this experiment. While the SVM model seems to underperform as compared HMM-based AER system when both use MFCC features, the native ability of SVM's to accept many more types of easy-to-extract features makes them more appealing in principle.

Event	Recall	Precision	F-score
Explosion	18.8	50.0	27.3
Gunfire	53.3	72.7	61.5
Composite	35.5	62.7	45.3

Table 3 – Performance of SVM AER System with MFCC speech features

The next experiment was designed to evaluate the performance of SVM-based AER system using the discriminative AER features described in section 3. Table 4 below shows the Recall, Precision and F-scores for this experiment.

Event	Recall	Precision	F-score
Explosion	56.3	75.0	64.3
Gunfire	66.7	93.0	77.7
Composite	61.5	83.8	70.8

Table 4 – Performance of SVM AER System with discriminative AER features

As compared to 42-dimensional MFCC features, using the expanded 165-dimensional AER features with SVM resulted in almost 135% relative F-score improvement in *explosion* recognition and 26% relative F-score improvement in *gunfire* recognition. As compared to HMM-based AER system, there is approximately 20% relative improvement in the composite F-score with the SVM-based AER system which uses high-dimensional AER features.

Although using MFCC features only with SVM did not yield improvements over baseline HMM-based system, we believe, that there is some discriminative information in them. This information could either be harnessed by using projection techniques such as concatenation of adjacent feature vectors (similar to AER feature expansion) or using non-linear kernels in SVM model. We decided to take advantage of the feature dimension flexibility of SVM by combining the 42-dimensional MFCC features with the 15-dimensional base AER features and applying the contextual window of 11 on the concatenated 57-dimensional extended feature set. The final expanded feature dimension was then equal to 57 x 11 = 627. This 627-dimensional feature vector per frame was used for classification in the SVM model. Table 5 below shows the results for this experiment.

Event	Recall	Precision	F-score
Explosion	57.8	70.0	63.7
Gunfire	74.7	85.7	79.8
Composite	65.5	77.8	71.1

Table 5 – Performance of SVM AER System with combination of MFCC and discriminative AER features

The results show that there is approximately 2.7% relative improvement in F-score for *gunfire* recognition at a modest 0.9% relative degradation in F-score for *explosion* recognition as compared to the SVM-based system with only AER features. Finally, with respect to HMM-based system (with MFCC features), we observed a composite relative improvement of 19.5% in recall and 21% relative improvement in precision, by using the SVM-based system with combined MFCC and AER features.

7. CONCLUSION

In this paper, we presented a discriminative SVM model-based approach to acoustic event recognition using publicly available surveillance video footage containing *explosions* and *gunfire*. The system performances were measured in terms of recall and precision using AED-ACC metric proposed in CLEAR 2007 evaluation.

The results presented in this paper show that the discriminative SVM-based AER system gives a high precision and high recall performance for both acoustic event classes as compared to the generative HMM model system. For any acoustic event recognition system, the underlying uncorrelated features play a very important role in class discrimination. We incorporated a combination of traditional and novel acoustic features that capture temporal and spectral characteristics of the non-speech acoustic events. Moreover, the discriminative distribution-free SVM model-based approach was easily applied to these large-dimensional feature vectors to model respective classes effectively.

We believe that additional tuning and unified feature normalization will lead to further improvements in retrieval precision and recall in the current SVM-based AER system. Additionally, using the current SVM model framework with non-linear kernels and additional feature expansions will give further improvements in accuracy. We plan to continue our investigations in those directions.

REFERENCES

[1] C. Clavel, T. Ehrette, G. Richard, "Events Detection for an Audio-Based Surveillance System", *IEEE International Conference on Multimedia and Expo*, 2005. ICME 2005, pages 1306–1309, 2005.

[2] J.L. Rouas, J. Louradour, S. Ambellouis, "Audio Events Detection in Public Transport Vehicle", *9th International IEEE Conference on Intelligent Transportation Systems*, 2006.

[3] L. Lu, H.J. Zhang, H. Jiang, "Content analysis for audio classification and segmentation", *IEEE Transactions on Speech and Audio Processing*, 0(7):504–516, 2002.

[4] T. Zhang, C.C.J. Kuo, "Hierarchical system for content based audio classification and retrieval", *Conference on Multimedia Storage and Archiving Systems III, SPIE*, 3527:398–409, 1998.

[5] L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi, A. Sarti, "Scream and gunfire detection in noisy environments," in *EURASIP*, Poland, Sept. 2007.

[6] R.M. Schwartz, Y. Chow, S. Roucos, M. Krasner, J. Makhoul, "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition", *ICASSP*, 1984, San Diago, CA, pp. 35.6.1-35.6.4.

[7] G. Zhoun, J. H. L. Hansen, J.F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress", *IEEE Transactions on Speech and Audio Processing*, pp. 201-216, March 2001.

[8] A. Harma, M.F. McKinney, J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *IEEE Conference on Multimedia and Expo*, 2005.

[9] A. Ganapathiraju, J. Hamaker, J. Picone, "Hybrid SVM/HMM architectures for speech recognition," in *Proc. of Speech Transcription Workshop*, May 2000.

[10] G. Peeters. "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," 2004.

[11] K. K. Paliwal, "A constrained forward-backward correlation prediction method for spectral estimation of noisy signals", *EURASIP, 1986.*

[12] Vladimir Vapnik, "*The Nature of Statistical Learning Theory",* Springer-Verlag, 1995.

[13] T. Joachims. (1999), "SVMLight: support vector machine", Cornell Univ., Ithaca, NY. http://svmlight.joachims.org/

[14] A. Temko, "Clear 2007 AED evaluation plan," http://isl.ira.uka.de/clear07, 2007.