AUDIO BASED EVENT DETECTION FOR MULTIMEDIA SURVEILLANCE

Pradeep K. Atrey[†], *Namunu C. Maddage*[*] *and Mohan S. Kankanhalli*[†]

[†]School of Computing, National University of Singapore
[*]Institute for Infocomm Research
Republic of Singapore

ABSTRACT

With the increasing use of audio sensors in surveillance and monitoring applications, event detection using audio streams has emerged as an important research problem. This paper presents a hierarchical approach for audio based event detection for surveillance. The proposed approach first classifies a given audio frame into vocal and nonvocal events, and then performs further classification into normal and excited events. We model the events using a Gaussian Mixture Model and optimize the parameters for four different audio features ZCR, LPC, LPCC and LFCC. Experiments have been performed to evaluate the effectiveness of the features for detecting various normal and the excited state human activities. The results show that the proposed top-down event detection approach works significantly better than the single level approach.

1. INTRODUCTION

In addition to the traditional video cameras, the use of audio sensors in surveillance and monitoring applications is becoming increasingly important [1]. Audio is useful especially in situations when other sensors such as video fails to reliably detect the events. For example, when the object is occluded or is in the dark, the audio sensors can be more appropriate in detecting the presence of object(s) assuming that the existence of the objects makes some sound. There are many events which can be effectively detected using the audio sensors when compared to using other sensors, e.g. human shouting/crying, door knocking and talking etc. The audio sensors can also be used to capture footstep sound of walking and running even in the dark when the video sensors usually fail to detect the human motion. In such cases, both audio and video sensors can be used to detect the events, overall confidence goes up. Audio is a cheaper sensor as well.

Audio based surveillance has been studied earlier for detecting various types of acoustic events such as human's coughing in the office environment [1], impulsive sounds like gunshot detection [2], glass breaks, explosions or door alarms [3]. In this paper, we focus on detecting a set of events such as human's crying, shouting, knocking, talking, walking and running using the audio streams. This work is a part of a multimedia surveillance system [4] which we are currently building. The system utilizes various heterogenous sensors including video and audio. This paper reports the results of event detection using only a single microphone data.

Our work is different from the previous works in the following aspects. First, we adopt a more sophisticated multi-level classification approach which works better than single-level approaches. Second, we provide extensive experimental evidence and analysis to evaluate the effectiveness of various features for detecting different kinds of events.

The proposed method adopts a hierarchical classification approach to assign a label to an event in a given "audio frame". We define an audio frame to be a fixed size audio segment which is extracted from the continuous audio stream. The various time-domain features - Zero-Crossing Rate (ZCR), Linear Predictor Coefficient (LPC), Linear Predictive Cepstral Coefficient (LPCC); and the frequency domain feature - Log Frequency Cepstral Coefficient (LFCC) are used. A Gaussian Mixture Model (GMM) classifier is employed to classify an input audio frame, at the top level - into foreground or background, at second level - into vocal or nonvocal, and at third level - into excited events (e.g. shout/cry, door knock, running footsteps) or normal events (talk, walking footsteps).

2. AUDIO EVENT DETECTION SYSTEM

The system consists of two phases - offline training (or event modelling) and online testing (event detection) as shown in figure 1. We describe its various components as follows.

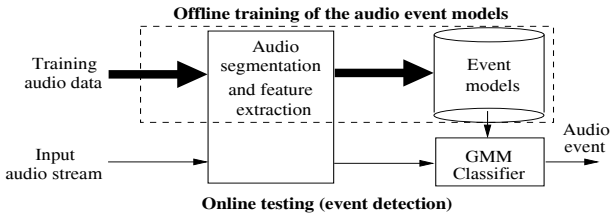

Fig. 1. Audio event detection system

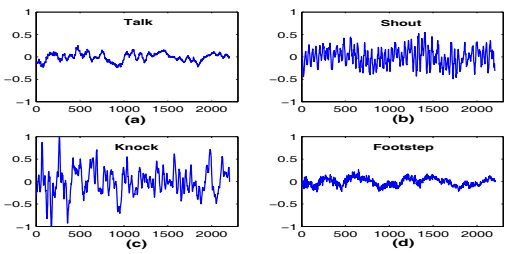

Fig. 2. The 50 ms sample audio frames for different events

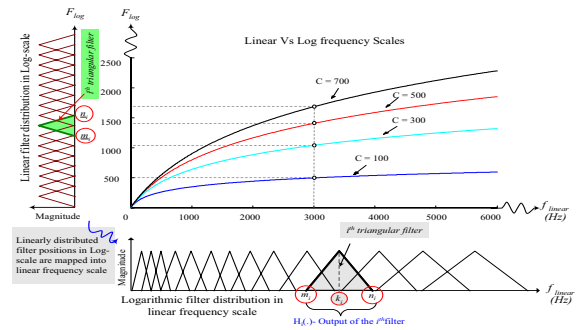

Fig. 3. Log scale filter distribution in Log and Linear scale

2.1. Audio segmentation

The training audio data is first segmented into the audio frames of 50 ms each. The size 50 ms is chosen by experimentally observing the minimum length of audio frame which can capture events such as footstep etc. We recorded audio for around two hours in the real environment (office corridor) and collected a large number of audio samples for each of the event - talk, shout, knock and footsteps (walking and running). The example of audio frames for these events are shown in figure 2.

2.2. Feature extraction

2.2.1. Zero Crossing Rate

The Zero Crossing Rate (ZCR) measures the number of times in the given time interval that the signal amplitude passes through a value of zero moving from negative to positive and vice versa. This feature helps in distinguishing the excited events from the normal events. We compute the mean ZCR value for each audio frame.

2.2.2. Linear Predictor Coefficient

The Linear Predictor Coefficient (LPC) have been widely used in speech processing community. The LPCs are filter coefficients described in all pole model which approximates the characteristics of speech production system. Therefore, LPCs are sensitive to the vocal sounds. This motivated us to use LPC for the detection of vocal and nonvocal events in surveillance scenarios. We used LPC algorithm from Matlab toolbox. The technical details of computation of LPCs are well documented in the literature.

2.2.3. Linear Predictive Cepstral Coefficient

The Linear Predictive Cepstral Coefficients are derived from LPCs. The LPCCs are more robust against sudden signal changes or the noise because these coefficients are derived from the impulse response of speech model. Therefore, we explore their use in detecting vocal and nonvocal audio events to see how they perform compared to LPCs.

2.2.4. Log Frequency Cepstral Coefficient

The Log Frequency Cepstral Coefficients (LFCCs) are computed by using logarithmic filter bank in frequency domain. The position of filters are calculated as follows. First, we transform the frequencies in the linear scale (F_{linear}) into log scale (F_{log}) using equation 1.

$$F_{log} = \frac{Clog_{10}(1 + \frac{F_{linear}}{C})}{log_{10}2} \tag{1}$$

where C is the frequency scaling factor. Then, the filters are linearly positioned in the log frequency scale and these positions are transformed back to linear frequency scale (figure 3) [5]. As C increases, more filters are positioned in the lower frequencies and vice versa.

The output $Y(i)$ of the ith filter is computed as -

$$Y(i) = \sum_{j=m_i}^{n_i} log_{10}[S(j)]H_i(j) \tag{2}$$

where $S(.)$ is the signal spectrum, $H_i(.)$ is the ith filter, and m_i and m_i are boundaries of the ith filter. The equation (3) describes the computation of nth LFCC.

$$C(n) = \frac{2}{n}\sum_{i=1}^{N_{cb}} Y(i)cos(k_i\frac{2\pi}{N}n) \tag{3}$$

k_i is the center frequency of the ith filter (figure 3), and N and N_{cb} are number of frequency sample points and number of filters, respectively.

2.3. Event modeling and detection

We consider four activities - talk, shout, knock and footsteps (walking and running). We adopt a hierarchical (top-down) approach to model these events using a mixture of Gaussians (GMM). The top-down event modelling approach works better compared to the single-level multi-class modelling approach which is shown in the experiment section.

As shown in figure 4, at the top level (0), each input audio frame is classified as the foreground or the background.

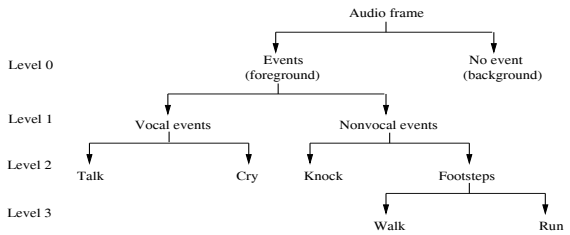

Fig. 4. A hierarchical approach for event detection

Table 1. Features and parameters used in the experiment

Features	Number of coefficients (n)	Scaling factor (C)	Number of filters (N_{cb})
ZCR	1	-	-
LPC	5 to 40	-	-
LPCC	5 to 40	-	-
LFCC	5 to 40	50 to 400	5 to 30

The background is the environment noise which represents 'no event' and is ignored. The foreground that represents the events, are further categorized into two classes - vocal and nonvocal (level 1). At the next level (2), both vocal and nonvocal events are further classified into normal and the excited events. Finally, at the last level (3), the footsteps sequences are classified as walking or running based on the frequency of their occurrence in a specified time interval.

3. EXPERIMENTAL RESULTS

The experiments have been performed with the following two basic objectives - first, to evaluate the effectiveness of various features for the specific events, and second, to evaluate the performance of the proposed hierarchical classification approach over the single-level approach.

Each event is modeled using over 10 minutes of audio data. For testing, we used manually annotated 2 hours of audio data stream which consists of around 10 minutes of each event. The parameter optimization of the audio features used for event detection is done on the same test data.

The parameter optimization of the features for event modelling is performed by varying the feature parameters (n, C and N_{cb}) for each of 10 Gaussian Mixture Models. Table 1 shows the parameter values for different features used in the experiment. The classification accuracy (shown in %) is used as the performance measure of our method.

3.1. Foreground/background detection

As described in section 2, we first segment the incoming audio stream into 50ms frames. Then, in both the hierarchical and single-level approaches, we use GMM classifier to identify the foreground frames. Table 2 shows the effectiveness of the features for foreground/background detection. We found that the LPC with n = 5 coefficients using a single Gaussian performs better than the other features.

Table 2. Accuracy (%) of background/foreground detection

Number of GMM	ZCR (n = 1)	LPC (n = 5)	LPCC (n = 5)	LFCC (n = 5, C = 200, N_{cb} = 20)
1	85	**96**	78	82

3.2. Event detection and feature evaluation

We employed 1 to 10 GMMs at multiple levels of classification hierarchy and made the following observations -

- For vocal-nonvocal classes, LFCC with $n = 10$ coefficients, $C = 200$ scaling factor and $N_{cb} = 25$ filters is found most appropriate feature. The details of the best classification accuracies obtained by using the different features and parameters are shown in Table 3. The optimized value [89,90(10, 200, 25)] shown in the Table indicates that the 'Vocal' and 'Nonvocal' events have been 89% and 90% times correctly detected using LFCC feature with 10 coefficients, 200 scaling factor and 25 filters. Note that the optimized parameters are considered to be the one which provide a higher accuracy for both the classes.

Table 3. Classification accuracies (%) for vocal and nonvocal events using hierarchical approach

Number of GMM	ZCR (n = 1)	LPC(n)	LPCC(n)	LFCC(n,C,N_{cb})
	Vocal,Nonvocal	Vocal,Nonvocal	Vocal,Nonvocal	Vocal,Nonvocal
1	72,58	83,76(15)	97,65(5)	**89,90(10, 200, 25)**
2	78,66	89,76(5)	88,63(5)	89,90(10, 200, 25)
3	81,66	83,81(5)	88,63(5)	89,90(15, 250, 20)
4	81,65	83,84(10)	86,65(5)	80,89(5, 200, 10)
5	81,65	83,84(10)	86,65(5)	89,90(10, 200, 25)
6	81,65	89,76(5)	86,65(5)	88,90(10, 200, 5)
7	81,66	83,81(5)	84,64(5)	89,90(10, 200, 25)
8	81,66	83,81(5)	82,60(5)	80,89(5, 200, 10)
9	81,66	83,76(15)	84,63(5)	89,90(10, 200, 25)
10	81,66	83,76(15)	85,63(5)	80,89(5, 200, 10)

- The best classification accuracies obtained for talk-shout classes are shown in Table 4. For these two types of events, the results clearly shows that the LPC with 25 coefficients is the better feature when used with 10 GMMs.

- For door knock and footstep events, ZCR with 8 GMMs and LFCC with 5 coefficients, 150 scaling factor and 30 filters also perform decently (Table 4). However, since ZCR is a less computationally-expensive feature than the LFCC, the ZCR seems to be the better choice.

- Combining features did not work well. We observed that it even reduced the classification accuracies.

Table 4. Classification accuracies (%) for Talk/shout and Knock/Footsteps events using hierarchical approach

Number of GMM	ZCR (n = 1)	LPC(n)	LPCC(n)	LFCC(n,C,N_{cb})	ZCR (n = 1)	LPC(n)	LPCC(n)	LFCC(n,C,N_{cb})
	Talk,Shout	Talk,Shout	Talk,Shout	Talk,Shout	Knock,Foot	Knock,Foot	Knock,Foot	Knock,Foot
1	59, 84	29, 84(5)	24, 63(20)	41, 89 (10 50 5)	90,62	40, 90(5)	45, 88(5)	70,74(5,150,30)
2	41, 100	41, 84(10)	35, 89(5)	35, 84 (10 50 15)	75,71	35, 88(5)	25, 93(5)	70,74(5,150,30)
3	53, 84	47, 63(10)	35, 79(5)	35, 84 (10 50 15)	75,69	35, 100(10)	50, 83(5)	70,74(15,150,30)
4	53, 84	41, 84(10)	35, 79(5)	41, 89 (10 50 5)	75,71	35, 100(10)	40, 95(15)	70,74(10,150,25)
5	53, 89	47, 63(25)	35, 79(5)	38, 85 (20 250 10)	75,71	35, 100(15)	60, 86(5)	70,74(10,150,30)
6	53, 89	47, 74(40)	35, 79(5)	41, 89 (10 50 5)	75,71	35, 98(5)	40, 98(15)	70,74(10,150,30)
7	53, 89	59, 63(25)	35, 79(5)	38, 85 (20 250 10)	75,71	35, 98(15)	40, 98(15)	70,74(10,150,25)
8	53, 89	53, 74(40)	35, 100(20)	38, 85 (20 250 10)	**75,74**	35, 98(15)	40, 98(15)	70,74(5,150,30)
9	53, 89	59, 63(25)	35, 79(5)	41, 89 (10 50 5)	75,74	35, 100(15)	40, 95(15)	70,74(10,150,30)
10	53,89	**65, 68(25)**	35, 79(5)	35, 84 (10 50 15)	75,71	35, 98(15)	40, 95(15)	70,74(5,150,30)

Table 5. Classification accuracies (%) for all four events using single-level approach

Number of GMM	ZCR (n = 1)	LPC(n)	LPCC(n)	LFCC(n, C, N_{cb})
	Talk, Shout, Knock, Footstep	Talk, Shout, Knock, Footstep	Talk, Shout, Knock, Footstep	Talk, Shout, Knock, Footstep
1	23, 80, 25, 60	23, 35, 26, 50(5)	23, 47, 20, 28 (5)	46, 80, 72, 43 (5, 100, 10)
2	10, 98, 45, 67	26, 50, 30, 55(35)	35, 54, 32, 41 (5)	42, 82, 72, 41 (5, 100, 10)
3	11, 80, 42, 65	36, 60, 41, 52(5)	35, 54, 32, 41 (5)	39, 82, 72, 36 (15, 200, 10)
4	12, 83, 44, 65	33, 62, 36, 52(5)	35, 54, 31, 38 (5)	42, 84, 67, 46 (5, 100, 10)
5	11, 87, 42, 60	35, 64, 32, 50(5)	35, 45, 32, 38 (10)	42, 82, 72, 41 (5, 100, 10)
6	10, 90, 46, 65	36, 67, 40, 55(5)	38, 57, 35, 42 (5)	42, 82, 72, 41 (5, 100, 10)
7	15, 84, 42, 65	36, 60, 41, 52(5)	35, 52, 30, 38 (10)	42, 84, 67, 46 (5, 100, 10)
8	23, 90, 42, 64	36, 56, 41, 53(10)	35, 45, 32, 38 (10)	42, 78, 67, 43 (15, 250, 15)
9	22, 89, 32, 65	36, 64, 41, 55(10)	35, 45, 32, 38 (10)	42, 82, 72, 41 (5, 100, 10)
10	12, 87, 25, 62	34, 56, 42, 53(5)	34, 54, 34, 41 (5)	42, 82, 72, 41 (5, 100, 10)

- We compared our results with the single-level multi-class approach by running 1 to 10 GMMs on the same data. We observed that single-level approach fails to provide a good accuracy for all the events (Table 5). It performs good for one class, but bad for the other. In contrast, our hierarchical method works better as can be seen from the reported results.

- The distinction between 'walking' and 'running' events is made based on the occurrence of number of footsteps in a specified time interval (2 seconds in our case). We used a single Gaussian classifier that provided 76% and 80% classification accuracy for walking and running footsteps, respectively.

4. CONCLUSIONS

The experimental results and analysis show that LPC performs well for the segmentation of background/foreground; and also, as expected, distinguishes (normal) talk and (excited) shouting more accurately. LFCC performs better in demarcating between the vocal and nonvocal events. Also, LFCC as well as ZCR are good for classifying between door knock and footstep events. The results also show that hierarchical classification performs significantly better than the single-level approach. Future work will be to extend it to the microphone arrays to obtain better robustness.

5. REFERENCES

[1] Aki Harma, Martin F. McKinney, and Janto Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE International Conference on Multimedia and Expo*, Amsterdam, July 2005.

[2] C. Clavel, T. Ehrette, and G. Richard, "Event detection for an audio-based surveillance system," in *IEEE International Conference on Multimedia and Expo*, Amsterdam, July 2005.

[3] Alain Dufaux, Laurent Bezacier, Michael Ansorge, and Fausto Pellandini, "Automatic sound detection and recognition for noisy environment," in *European Signal Processing Conference*, Finland, September 2000, pp. 1033–1036.

[4] Pradeep K. Atrey, Mohan S. Kankanhalli, and Ramesh Jain, "Timeline-based information assimilation in multimedia surveillance and monitoring systems," in *ACM International Workshop on Video Surveillance and Sensor Networks*, Singapore, November 2005, pp. 103–112.

[5] Namunu C. Maddage, *Content based music structure analysis*, Ph.D. thesis, School of Computing, National University of Singapore, 2006.