Detection of Abnormal Sound Using Multi-stage GMM for Surveillance Microphone

Akinori Ito, Akihito Aiba, Masashi Ito and Shozo Makino
Graduate School of Engineering, Tohoku University
6-6-05 Aramaki aza Aoba, Aoba-ku, Sendai 980-8579 Japan
{*aito,akihito,itojin,makino*}*@makino.ecei.tohoku.ac.jp*

Abstract

We developed a system that detects abnormal sound from sound signal observed by a surveillance microphone. Our system learns the "normal sound" from observation of the microphone, and then detects sounds never observed before as "abnormal sounds." To this end, we developed a technique that uses multiple GMMs for modeling different levels of sound events efficiently. We also consider how to determine thresholds of GMM switching and event detection. As a result, we obtained almost same detection performance using the percentile method to the manually optimized GMMs. Besides, we exploited the segment-based feature, which gave the best result among all methods.

1. Introduction

Closed-circuit TV (CCTV) systems for surveillance are widely used nowadays in various places such as shopping center or office. In most of current CCTV systems, captured images are just recorded, and they are verified after an incident happens. Such systems can be deterrents to crimes; however, they cannot help victims when incidents happen unless monitored by a human operator. To realize immediate detection of incidents without human operators, video-based methods for automatic incident detection have been developed so far [2, 7]. In spite of much efforts of developing such image-based detection systems, it is still difficult to detect incident using only images accurately [1].

Recently, audio-based incident detection has been proposed. The audio-based method can compensate the video-based one; the target need not to be captured by the camera, and it requires less computation compared with the video-based method. Kim *et al.* developed a system that detects incident using sound [4]. Their system observes sounds using a microphone array, estimates the direction of the sound, and moves camera toward the sound. Similar system was proposed by Kawamoto *et al.* [3]. Their systems focused on how to estimate the direction of the sounds.

In addition to the estimation of sound direction, it is also important and difficult issue how to detect the only sounds concerning to incidents while suppressing false alarms. To make the situation even more difficult, it is unrealistic to prepare all samples of sounds related to incidents beforehand. If the systems treat specific sounds such as human fall sounds[8], explosive sounds and gunshot sounds[6], those systems will only be useful in those situations. To realize a system that can be used in more general situations, we need to develop a method for detecting "unknown" sounds in a situation where the system is installed.

In this paper, we propose a new method for detecting abnormal sounds without using any samples of the sounds to be detected. Our method creates a model of "normal" sounds using the multi-stage Gaussian mixture models for modeling "rare" events contained in normal sounds. We compared several methods of determining number of GMMs and thresholds for efficient training of the model.

2. Abnormal sound detection based on Multi-stage GMM

2.1. Basic framework

Figure 1 shows an overview of the proposed system. In our system, we assume that the sounds to be detected are those which rarely occur in a normal situation. Based on this assumption, the system first observes sounds in the normal situation, and trains a statistical model of the normal sounds. After training the model, the system continues to observe sounds, and calculates likelihood of the sound. If the value of likelihood of a sound is lower than a predefined threshold, that sound is determined as an abnormal sound.

Note that the definition of "normal" depends on the situation at which the system is installed, including place, time in a day, day of the week, and the season. Therefore, the sys-

978-0-7695-3744-3/09 $25.00 © 2009 IEEE
DOI 10.1109/IAS.2009.160

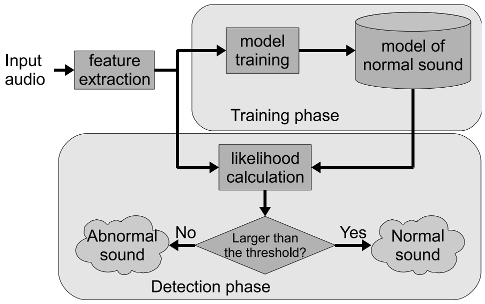

Figure 1. System overview

tem must update the model frequently to adapt the change of the situation. However, in the experiment presented in this paper, we assume that the acoustic situation is stable throughout the experiment.

2.2. Multi-stage GMM

Modeling of the sounds is based on a Gaussian mixture model (GMM). A GMM is mixture of multivariate Gaussian distribution, each of which expresses some kind of sound.

However, from a preliminary experiment, one problem arose by using single GMM for modeling normal sounds. Sounds in a normal situation can be classified into the following two classes:

- *Background sound:* Stable sounds such as air conditioner noise or computer fan noise. Most of the observed sounds are involved in this class.
- *Salient sound:* Sounds occurred by events such as talking voice, door open and close etc. Although this kind of sounds is salient, ratio of this class in the entire sounds is very small.

The preliminary experiment suggested that the most false alarms were caused by sounds in the salient sound class. Therefore, it is important to create a detailed model that focuses on the sounds in the salient sound class. However, if we train a GMM with the maximum likelihood criteria, the salient sounds are just ignored because the vast majority of the sounds are background sound. Increasing the number of mixture of the GMM does not help, because most of the mixture component is trained for modeling the sounds in the background class, which is the best way of increasing the entire likelihood. Therefore, we need another way of modeling the salient sounds using a criterion other than the ML criteria.

Generally speaking, the discriminative training is effective in this case. However, we cannot employ the discriminative training in this framework because we do not have negative examples in the training phase, which is indispensable for conducting the discriminative training.

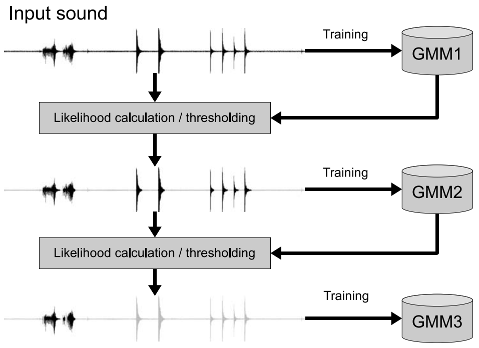
Input sound
Training
GMM1
Likelihood calculation / thresholding
Training
GMM2
Likelihood calculation / thresholding
Training
GMM3

Figure 2. Training procedure of the multi-stage GMM

Therefore, we propose a new model that uses multiple GMMs. We call this method “the multi-stage GMM.” The basic idea of the multi-stage GMM is to model the majority of the training sample using the first GMM, and then model the rest of the samples using the second GMM. The overview of the training procedure is shown in Figure 2.

The detail of the training procedure is as follows. First, GMM1 is trained using the entire training samples. Next, likelihood value of each of the training samples is calculated using the GMM1. Then the samples that have smaller likelihood values than a threshold are chosen for the next training.

In the next stage, the selected samples are used as training samples to train GMM2. Then the samples with lower likelihood are chosen using a threshold for training GMM3. This training process continues until a sufficient number of GMMs are trained.

The trained multi-stage GMM classifier $\mathcal{M}$ is a list of pairs of a GMM and threshold value.

$$\mathcal{M} = (\langle \mathcal{G}_1, \theta_1 \rangle, \ldots, \langle \mathcal{G}_S, \theta_S \rangle) \tag{1}$$

Here, $\mathcal{G}_k$ is the k-th GMM where

$$p(x|\mathcal{G}_k) = \sum_{i=1}^{M_k} \lambda_{k,i} N(x; \mu_{k,i}, \Sigma_{k,i}), \tag{2}$$

θ_k is the k-th threshold, and S is the number of GMMs.

The detection of abnormal sound is conducted as follows. Let $D_a(x;\mathcal{M})$ be a discrimination function that returns 1 when the sound x is determined as an abnormal sound. Now, $D_a(x;\mathcal{M})$ is defined as follows.

$$D_a(x;\mathcal{M}) = D_1(x;\mathcal{M}) \tag{3}$$

$$D_k(x;\mathcal{M}) = \begin{cases} 0 & \text{if} \quad p(x|\mathcal{G}_k) > \theta_k \\ 1 & \text{else if} \quad k = S \\ D_{k+1}(x;\mathcal{M}) & \text{otherwise} \end{cases} \tag{4}$$

Table 1. Experimental conditions

Training data	Environmental sound in our lab, 24 hours (from 2007/10/01 10:00)
Test data (normal)	Environmental sound in our lab, 24 hours (from 2007/10/02 10:00)
Test data (abnormal)	7 kinds of sounds (buzzer, bell, fire cracker, glass clash, male scream, female scream)×4
Sampling freq.	16kHz
Analysis window	25ms length / 10ms shift
Window function	Hamming
Feature (dimension)	MFCC(16)+power(1)+ΔMFCC(16) +Δpower(1)
Number of Gaussian mixture	2 – 256, selected *a posteriori*

Table 2. The experimental result

# stages	Recall	Precision	F-measure
1	0.464	0.765	0.578
2	0.536	0.789	0.638
3	0.536	0.938	0.682

The meaning of this procedure is that we first calculate likelihood of the input sound x using the first GMM $\mathcal{G}_1$. If the likelihood is lower than the threshold, we try to determine if the sound is abnormal using the next GMM. If the likelihood by the final GMM is lower than the threshold, the input sound is determined as an abnormal sound.

2.3. Experiment

We conducted an experiment for investigating the effectiveness of the multi-stage GMM. In this example, all thresholds and the number of mixture for each GMM were determined *a posteriori*. Experimental conditions are shown in Table 1.

Table 2 shows the experimental result. In this result, all thresholds were chosen so that the maximum F-measure was obtained. From this result, we can conclude that the multi-stage GMM can improve the accuracy of detection of abnormal sounds, as long as appropriate thresholds are used.

3. Automatic determination of optimum threshold

Although the proposed method could improve the accuracy of abnormal sound detection, the thresholds were selected manually. This requires the abnormal sound data, which is actually impossible. Therefore, we need to develop a method of determining thresholds without referring the abnormal sound data.

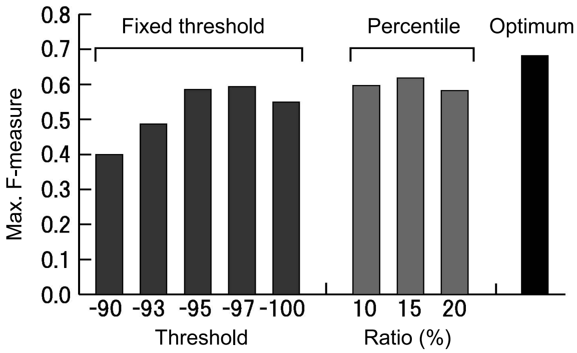

Figure 3. Experimental result

To this end, we investigated the following two methods of threshold determination:

1. *Fixed threshold method.* In this method, only one threshold is used for all GMMs. Instead, we increase the number of GMMs in the training phase until all training samples are determined as normal data.

2. *Percentile method.* In this method, a pre-defined fraction of samples with lower likelihood are chosen for the training data of the next GMM. The stopping criteria of GMM generation are based on difference of the minimum and maximum likelihood, or number of frames for training.

In this experiment, we used five thresholds (-90 to -100) for the fixed threshold method and three ratios (10 to 20 %) for the percentile method. The other experimental conditions were same as that in the previous experiment. Figure 3 shows the experimental result. As shown in the figure, we could improve the F-measure by optimizing only one parameter by both methods. However, the best F-measures by the proposed methods were still lower than that by optimizing all thresholds manually.

The best results by the three threshold determination methods are shown in Table 3. The percentile methods gave higher F-measure than the fixed threshold method. The determined number of stages was 3 for the investigated methods.

4. Incorporating segment model

Although a GMM does not consider the order of feature vectors, the order of feature vectors does matter because timbre of a salient sound changes temporally. Therefore, we investigated using segments of feature vectors [5]. This

Table 3. Comparison of threshold determination methods

Threshold	Recall	Precision	F-measure	#stages
Fixed threshold	0.679	0.528	0.594	3
Percentile	0.607	0.630	0.618	3
Optimum	0.536	0.938	0.682	3

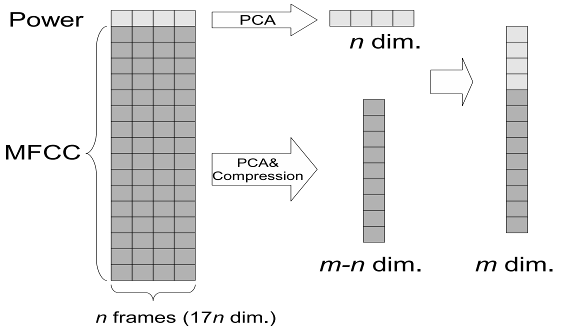

Figure 4. Segment-based feature calculation

method uses several contiguous feature vectors as one feature vector, which involves temporal change of the original feature vectors. As a segment-based feature vector has high-dimension, the principal component analysis (PCA) is applied for reducing the dimensionality of feature vectors.

Calculation of a segment-based feature vector is depicted in Figure 4. The contiguous n frames are used for calculating one vector. Powers and MFCCs are treated separately; The n contiguous powers are orthogonalized using PCA without compression. The MFCCs are orthogonalized and compressed using PCA. Finally, the powers and the compressed MFCCs are combined for making one feature vector.

We carried out an experiment for investigating the performance of segment-based feature. In this experiment, we used 2 to 8 contiguous frames for calculating a feature vector, and the total dimension of the compressed feature vector was changed from 13 to 25. The thresholds are determined by the percentile method (ratio=15%). Figure 5 shows the results. The best result was obtained when using 25-dimension feature vector calculated from two frames. The F-measure of the best condition was 0.696, which was better than the result by manually-optimized thresholds shown in Table 3.

5. Conclusion

We proposed a novel method for detecting abnormal sounds with no samples of abnormal sounds. Our method is based on multi-stage GMM, which efficiently models the rare events contained in the training samples. We investigated two methods for threshold determination, and the percentile method gave the best result. Besides, we examined the segment-based feature. Using the proposed feature, we obtained 0.696 F-measure.

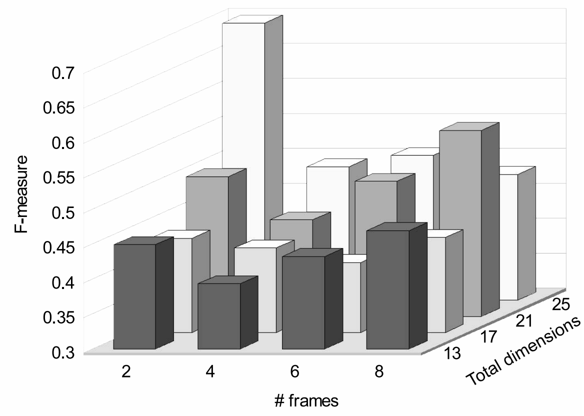

Figure 5. Experimental results of segment-based feature

References

[1] J. Bijhold, A. Ruifrok, M. Jessen, Z. Geradts, S. Ehrhardt, and I. Alberink. Forensic audio and visual evidence 2004-2007: A review. In *15th INTERPOL Forensic Science Symposium*, 2007.

[2] K. feng Wang, X. Jia, and S. Tang. A survey of vision-based automatic incident detection technology. In *Int. Conf. on Vehicular Electronics and Safety*, pages 290–295, 2005.

[3] M. Kawamoto, F. Asano, and K. Kurumatani. A security monitoring system of detecting unusual sounds and hazardous situations by sound environment measurement using microphone arrays. In *IPSJ SIG Tech. Rep., 2008-UBI-019*, volume 2008, pages 19–26, 2008 (in Japanese).

[4] Y. Kim, S. W. Lee, D. H. Lee, J. Kim, and M. W. Lee. Sound detection as an aid to increase detectability of CCTV in surveillance system. In *Usability and Internationalization, Part II, HCII 2007, LNCS 4560,*, pages 382–389, 2007.

[5] S. Nakagawa and K. Yamamoto. Speech recognition by hidden Markov model using segmental statistics. *IEICE Trans. D-II*, J79-D-II(12):2032–2038, 1996 (in Japanese).

[6] S. Ntalampiras, I. Potamitis, and N. Fakotakis. On acoustic surveillance of hazardous situations. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP2009)*, pages 165–168, 2009.

[7] S. A. Velastin, B. A. Boghossian, and A. Lazzarato. Detection of potentially dangerous situations involving crowds using image processing. In *Intelligent Industrial Automation*, 1999.

[8] X. Zhuang, J. Huang, G. Potamianos, and M. Hasegawa-Johnson. Acoustic fall detection using gaussian mixture models and GMM supervectors. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP2009)*, pages 69–72, 2009.