Scream sound detection based on SVM and GMM

Sukhwan Chung and Yongjoo Chung

***Abstract*—In recent days, one of the critical social problems is that violent crimes are frequent in places like public toilets and streets. Due to this problem, the importance of surveillance systems for the safety of pedestrians is increasing gradually. As conventional visual surveillance systems have some limitations, many attempts have been made to support the system by adding audio-based functionality. In this study, we try to detect scream sounds based on SVM and GMM, respectively, and compare the performance of the two methods through various experiments. From the experimental results, SVM could obtain 0.559% False Acceptance Rate, which means that there is a very low possibility of incorrectly deciding non-scream sound as a scream sound. In contrary, the GMM method could achieve 12.03% of False Rejection Rate, which implies that GMM has relatively good sensitivity to the scream sound compared with SVM. From these results, we could conclude that both GMM and SVM have a distinctive merit from each other and the plausibility of further performance improvement by combining the two approaches is also observed.**

***Keywords*—scream detection, surveillance system, GMM, SVM.**

I. Introduction

In recent days, at places like public toilets and public parking lots as well as private houses and apartments, the rate of crime occurrences is increasing considerably [1], which makes environment security of our major concern. Surveillance systems commonly used now employ infrared sensors to detect intruders from outside or record videos as in CCTVs installed in certain places. In particular, CCTVs aid the security enforcement of the police and help prevent crimes but they suffer from environmental restrictions like vision angle and dimming light and have fatal demerit that they can't identify the crime situation in real time. To overcome this problem, there have been efforts to apply audio-based detection to the existing video surveillance system [2][3][4].

GMM(Gaussian Mixture Model) is one of the popular methods in audio detection as well as in speech recognition. In [5], GMMs for the door/trunk closing and car accident sound are trained respectively and they are used to detect the sound from door/trunk closing so that the black box in the car do not operate falsely by confusing it with car accident sound. In other applications, GMM has shown successful results in detecting specific sounds such as scream, shout and gun fire [6][7][8]. In [7], to detect scream and gun fire in noisy conditions, a parallel type recognizer is proposed by using the GMM for each sound.

Recently, researches using SVMs (Support Vector Machines) have been popularly used in audio detection [9][10][11]. SVMs are known to have equal or superior performance in generalization compared with other classifiers. Rather than using frame-level feature vectors, they use as inputs, the means and variances of feature vectors computed periodically. In [9], the means, variances, maximum and minimum values of the 36-dimentonal MFCC(Mel-frequency cepstral coefficients) computed in every 0.25 seconds are used as the input of the SVM to show superior performance. In addition, pitch and energy values of the signal is combined with the output of the SVM to detect the existence of scream sound.

As mentioned previously, GMM and SVM are used independently in many researches to detect various audio signals but there are few research results to compare the two methods using the same audio data. But it is needless to say that we need to compare them to implement more efficient classifier for scream detection. In this study, to obtain reliable classification results, we gathered various audio sound signals compared with other researches for scream detection. Also, the input feature vectors and architecture of the classifier is varied to make the comparison in more detail.

The paper is organized as follows. In section II, feature extraction methods and classifiers used for scream detection are introduced. In section III, we show and compare various experimental results. Finally, in section IV, we make conclusion and discuss further studies.

II. Feature Extraction and Classifiers

A. Feature Extraction

The waveform of sound signal can't be used as input of classifiers due to the irregularities in its characteristics. So some kind of values which can well explain the characteristics of the sound signal and traditionally, we use features like ZCR(zero crossing rate), pitch and correlation for audio signal detection [7][9]. In this study, we use MFCC features which have shown to be quite efficient with noise-robustness in speech recognition [6].

In Fig. 1, we show the process of MFCC feature extraction process in a block diagram. Audio signal sampled at 16 KHz is processed in frames each of which has 25ms duration and the interval between frames is 10ms. The audio signal is pre-emphasized as in (1) to emphasize high-frequency components of the signal and then hamming-window is used

Manuscript received April. 12, 2017. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by Ministry of Education(No. 2015R1D1A1A01059925).

Sukhwan Chung is with the Keimyung University, Daegu, South Korea (e-mail: kbsong11@ naver.com).

Yongjoo Chung is with the Keimyung University, Daegu, South Korea. (e-mail: yjjung@kmu.ac.kr).

before applying FFT(Fast Fourier Transform) to the signal.

$$s(n) = s(n) - 0.9s(n-1). \quad (1)$$

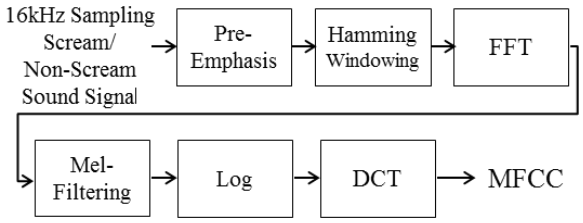

Fig. 1. MFCC feature extraction process

Mel-scale filter bank output is computed from the result of FFT and log of the filter bank output is transformed into 13-dimentional MFCC via DCT(Discrete Cosine Transform).

We determined the input vectors of the classifiers to give the best performance on recognition experiments. 12-dimentional MFCCs excluding c0 are used as the input vectors of the GMM while 36-dimentional MFCCs from the 12-dimentional MFCCs including both delta and acceleration coefficients are used as the input vectors of the SVM.

In this study, we used AFE(Advanced Front-End) which is defined as the standard for MFCC feature extraction by ETSI(European Telecommunication Standards Institute) [12]. AFE is expected to be robust against background noise occurring in real environments since it contains efficient algorithms to suppress noise signals.

B. Gaussian Mixture Model

GMM is the sum of weighted Gaussian probability density functions and has been popularly used in speech recognition to model the acoustic characteristic of speech signals in MFCC domains. Training GMM in this study is done as shown in Fig. 2. After feature extraction, vector quantization is done for scream and non-scream data to find the GMM parameters for each of them. The GMM parameters consist of the weight ω_m, mean vector μ_m and covariance matrix Σ_m, $\{m = 1,2,\cdots,M\}$ of the Gaussian probability density functions comprising the GMM where M is the number of the number of mixture components.

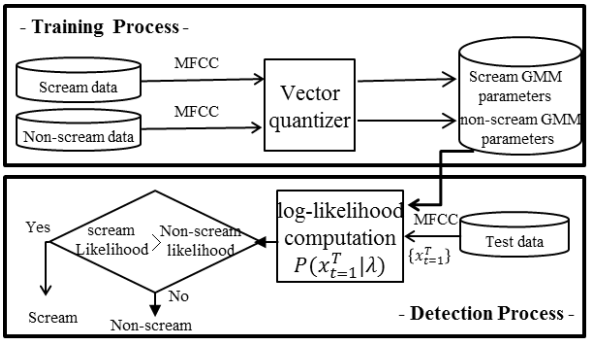

Fig. 2. GMM training and detection process

Given feature vectors of length T, $\{o_t\}_{t=1}^T$ the log-likelihood for each feature vector is computed as follows.

$$b(o_t) = \log\left(\sum_{m=1}^{M} \omega_m N(o_t \mid \mu_m, \Sigma_m)\right) \quad (2)$$

$$N(o_t \mid \mu_m, \Sigma_m) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_m|}} e^{-\frac{1}{2}(o_t-\mu_m)\Sigma_m^{-1}(o_t-\mu_m)'} \quad (3)$$

Here, $N(o_t \mid \mu_m, \Sigma_m)$ is the Gaussian probability density function with mean vector μ_m and diagonal covariance matrix Σ_m. For the whole feature vectors of length T, the log-likelihood is computed by adding the log-likelihood of each feature vector by assuming that the feature vectors are independent from each other.

C. Support Vector Machine

SVM is a non-probabilistic binary classifier which tries to maximize the distance margin between two classes [13]. In this study, the input for the SVM is obtained by averaging the 36 dimensional MFCC feature vectors for 20 frames.

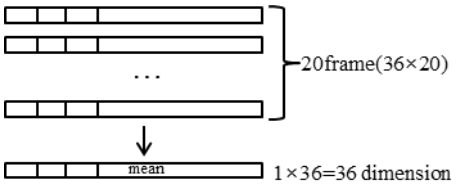

Fig. 3. MFCC input vector for SVM

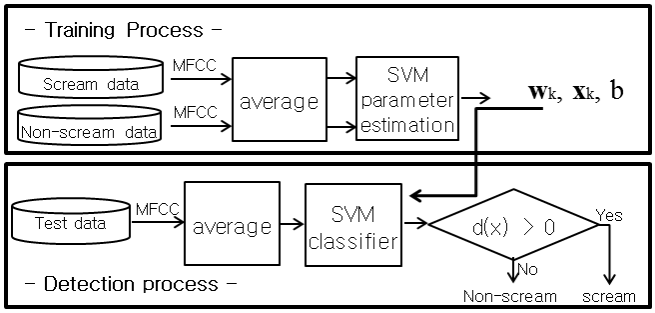

Fig. 4. Training and detection process of SVM

In Fig. 4, we show the training and detection process for the SVM in this study. During training, scream and non-scream data is inputted as in Fig. 3 to estimate the SVM parameters, w_k, x_k, b. During detection process, the output $d(x)$ of the SVM classifier is evaluated to determine whether the input data is scream or not.

$$d(x) = \sum_{k=1}^{K} w_k G(x_k, x) + b \quad (4)$$

Here, K is the number of support vectors. w_k, x_k, b are weights, support vectors and bias obtained during training process.

If the value of $d(x)$ is positive, we determine the input as scream sound and if negative, it is determined as non-scream sound.

https://doi.org/10.17758/EAP.AE0517126

III. EXPERIMENTAL RESULTS

A. Database

For the experiments in this study, we used data gathered from internet [14]. The data was recorded in clean conditions. The whole data can be divided into scream data set and non-scream data set. The scream data set consists of 63 files with durations from 1 seconds to 12 seconds. The non-scream data set consists of 213 files with durations from 1 seconds to 225 seconds. For the experiments, the ratio of data amount between training and testing is set 3:1. For the reliability of the experimental results, the whole data is divided into 4 parts and Jack knife method is used for the experiments.

B. Experimental Results

The decision on the test data is done in every 600 ms using the labelling information. This is based on the idea that duration of the scream sound in real environments would be at least 600 ms. If the decision interval is too short, sounds like cough would be classified as scream as the sound signal is too short for us to confirm that it is not scream. In contrary, if the decision interval is too long, there is a possibility that we may miss short scream sound. To measure the performance of the classifiers, we used two popular metrics used in this field, FAR(False Acceptance Rate) and FRR(False Rejection Rate). FAR is the ratio of non-scream data which is misclassified as scream sound and FRR is the ratio of scream data which is misclassified as non-scream sound.

$$FAR = \frac{\#\ of\ decisions\ mis-classified\ as\ scream}{\#\ of\ total\ decisions\ for\ non-scream\ data}$$

$$FRR = \frac{\#\ of\ decisions\ classified\ as\ non-scream}{\#\ of\ total\ decisions\ for\ scream\ data}$$

In Table 1, we show the results of GMM classifier when 12-dimensional MFCCs are used as feature vectors. We can see the results depend greatly on the mixture components of the GMM. FRR has the best result of 19.41% when the number of mixture components is 50. FAR improves as the number of mixture components increases and it has the best result of 9.1% when the number of mixture components is 76. The reason for the performance improvement of FAR with the number of mixture components is that the various sound signal in the non-scream data is modelled better as the number mixture components is increased. In contrary, the improvement of FRR with the number of mixture components is limited as the acoustic characteristic of the scream sound signal is relatively stationary.

TABLE I: EXPERIMENTAL RESULTS USING GMM CLASSIFIER

Mixture number	FRR(%)	FAR(%)
1	21.14	18.64
10	22.86	15.14
20	19.74	14.17
30	21.14	13.3
40	19.74	12.78
50	19.41	11.85
55	20.29	11.27
60	21.43	10.67
70	22.29	10.02
76	23.43	9.10

In Table II, we show the comparison between GMM and SVM classifiers. For the GMM, the best results from Table 1 when the number of mixture components is 50(FRR) and 76(FAR) are shown in Table 2. From Table 2, we can see that SVM shows better performance in FAR(2.54% vs. 9.1%) while GMM is better than SVM in FRR(38.38% vs. 19.41%).

TABLE II: PERFORMANCE COMPARISON BETWEEN SVM AND GMM

	FRR(%)	FAR(%)
SVM	38.38	2.54
GMM	19.41	9.10

The scream files used in the previous experiments included background noises and breathing sound in addition to real scream sound. Thus, for the strict performance comparison, we eliminated those parts from the scream files and recognition experiments were done again and the results are shown in Table 3 and 4.

TABLE III: EXPERIMENTAL RESULTS USING GMM CLASSIFIER WHEN NOISES AND BREATHING SOUND SIGNAL IS REMOVED FROM SCREAM FILES.

Number of Mixture	FRR(%)	FAR(%)
1	12.03	15.43
3	26.55	11.39
5	32.78	9.15
7	34.43	7.39
10	34.85	6.01
20	38.58	5.63
30	35.68	4.76
40	36.51	4.10
50	39.83	3.58
60	43.15	2.98
70	43.15	2.61
76	44.39	2.35

Compared with Table I, the results in Table III show much better performance in FAR but worse performance in FRR. As the characteristics of signals in scream file is very stationary due to the removal of noises and breathing sounds, the increased number of mixture components in scream GMM seems to badly affect FAR.

https://doi.org/10.17758/EAP.AE0517126

TABLE IV: PERFORMANCE COMPARISON BETWEEN GMM AND SVM WHEN NOISES AND BREATHING SOUND SIGNAL IS REMOVED FROM SCREAMING FILES

	FRR(%)	FAR(%)
SVM	36.93	0.55
GMM	12.03	2.35

In Table IV, we compare the performance between GMM and SVM when noises and breathing sounds are removed from screaming files. For the GMM, the best results from Table 3 when the number of mixture components is 1(FRR) and 76(FAR) are shown in Table 4. Similarly as in Table 2, we can see that better performance is attainable by using GMM classifier. However, we can see that SVM classifier has very good performance in FAR attaining 0.55% in which only a few non-scream sounds like breaking windows and cat crying are misclassified as scream sound.

IV. CONCLUSIONS

Surveillance systems relying on visual information do not show satisfying results due to restrictions in real environments. To overcome this problem, various audio detection methods have been proposed recently.

In this study, we compared the performance of SVM and GMM classifiers which are representative methods in audio detection. From the experiments, we could find that the two methods show contrary recognition results. GMM is superior in FRR while SVM is better than GMM in FAR.

We think that the contrary results can be utilized appropriately in commercial products. For example, when we need to know exactly the moment of screaming as in crime investigation, we should not miss the screaming event and so GMM is advantageous as it has better FRR performance. But in real time surveillance system, SVM which has better FAR is advantageous as the system reliability is of major concern.

For further studies, we aim to implement a commercial product utilizing the methods developed in this paper. Also, a deep neural network based detection architecture will be studied for better performance than GMM and SVM. Finally, since SVM and GMM show different characteristics in detection results, we may find a method to combine them and improve performance in scream sound detection.

REFERENCES

[1] W. Huang, T. K. Chiew, H. Li, T. S. Kok and J. Biswas, "*Scream detection for home applications*", Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on, No. 399, pp. 2115-2120, June. 2010.

[2] M. K. Nandwana, A. Ziaei and J. H. L. Hansen, "*Robust Unsupervised Detection of Human Screams In Noisy Acoustic Environments*", Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pp. 161-165, April. 2015.

[3] J. Pohjalainen, P. Alku and T. Kinnunen, "*Shout detection in noise*", Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 4968-4971, May 2011.

[4] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci and A. Sarti, "*Scream and Gunshot Detection in Noisy Environments*", Signal Processing Conference, 2007 15th European, pp. 1216-1220, Sep 2007.

[5] B. Lei, M. W. Mak, "*Sound-event partitioning and feature normalization for robust sound-event detection*", Digital Signal Processing (DSP), 2014 19th International Conference on, pp. 389-394, Aug. 2014.

[6] S. G. Oh, J. U. Lee, H. S. Lee, Y. W. Chung and D. H. Park, "Abnormal Sound Detection and Identification in Surveillance System", Journal of KIISE, Vol. 39, No. 2, pp. 144-152, Feb. 2012.

[7] J. H. Park, J. Y. Lim, J. Y. Yang, J. M. Kyung and M. S. Hahn, "False Positive Movie Clip Decision in Black-box Using Car Door-Closing Sound Classification", IEIE, Vol. 37, No. 1, pp. 761-763, June. 2014.

[8] J. H. Seo, H. I. Lee and S. P. Lee, "A Design of a Scream Detecting Engine for Surveillance Systems", KIEE, Vol. 63, No. 11, pp. 1559-1563, Nov. 2014.

[9] S. Ntalampiras, I. Potamitis, N. Fakotakis, "*On acoustic surveillance of hazardous situations*", Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, pp. 165-168, April. 2009.

[10] Support Vector Machines for Binary Classification, http://kr.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html?requestedDomain=www.mathworks.com. [Accessed: Sep. 02, 2016]

[11] Korean National Police Agency, http://www.police.go.kr/portal/main/contents.do?menuNo=200197 [Accessed: Sep. 05, 2016]

[12] Sound Effects Download, http://www.soundsnap.com/ [Accessed: Sep. 05, 2016]