Scream Detection for Home Applications

Weimin Huang, Tuan Kiang Chiew, Haizhou Li, Tian Shiang Kok, Jit Biswas
Institute for Inofcomm Researach, Singapore 138632
{wmhuang, chiewtk, hli, tskok, biswas}@i2r.a-star.edu.sg

Abstract— **Audio signal is an important clue for the situation awareness. It provides complementary information for video signal. For home care, elder care, and security application, screaming is one of the events people (family member, care giver, and security guard) are especially interested in. We present here an approach to scream detection, using both analytic and statistical features for the classification. In audio features, sound energy is a useful feature to detect scream like audio. We adopt the log energy to detect the energy continuity of the audio to represent the screaming which is often lasting longer than many other sounds. Further, a robust high pitch detection based on the autocorrelation is presented to extract the highest pitch for each frame, followed by a pitch analysis for a time window containing multiple frames. Finally, to validate the scream like sound, a SVM based classifier is applied with the feature vector generated from the MFCCs across a window of frames. Experiments of screaming detection is discussed with promising results shown. The algorithm is ported and run in a Linux based set top box connected by a microphone array to capture the audio for live scream detection.**

Keywords-component; screaming detection; pitch detection; SVM learning; monitorig, home applications; elderare; security

I. INTRODUCTION

Audio-visual monitoring plays important roles in situation awareness [1,2]. In home and many other applications, audio is some times the dominant or sole accessible signal indicating what happens [3-7], such as screaming, yelling, crying, gunshot, explosion, and door opening/closing etc. Of these audio events, screaming/yelling is one of the specifically concerned signals in home care and eldercare for the security and safety reason.

Early works related to audio activity/background recognition start on the auditory scene analysis that recognizes an environment using audio information [8,9]. Besides continuous effort on auditory scene recognition [10], some researchers worked on the efficient indexing and retrieval of audio data given a large amount of music, speech and other sound clips available today, in order for human to browse the data easily [11,12,19], where audio classification of speech and music genre is one of the main focuses [13,14].

Recently, more researchers are attracted to the study of detection of unique audio event from sounds [4,6,15-18] for different applications, of which gunshot and scream were studied by energy surge and audio feature analysis. Other works [20-22] have studied the stress detection from human speech while some researchers worked out some approaches to general sound classification [7,14,23].

In this paper, we are focusing on the detection of one specific audio event, i.e. screaming, which usually represents urgent or serious cases for home care and elderly care. One example is the measurement of verbal aggression of dementia patients by detection of the duration and frequency of screaming. In this work, we propose a new method for human screaming detection, based on the perceptual analysis of the scream sound. An energy analysis is first introduced to detect abnormal sound, followed by a pitch extraction to pick up the 'scream' like audio. An example based learning classifier (Support Vector Machine) is adopted to further validate the two audio classes (scream/non-scream), based on the spectrum features on a sliding window of frames.

A. Related works

Most related works to scream detection are presented in [4-7,18,19]. In [18], the system detects and locates gunshot from distance, the features used to detect the shock wave (usually generated by gunshot or other supersonic projectiles) are the features to detect the N shape wave, such as pulse width, peak waveform amplitude, ratio of the positive peak to the negative peak of the N of the gunshot pulse, and slope. Periodicity of wave is detected as one feature of gunshot. The dynamic synapse neural networks are then used to classify gunshots (12 guns from other sound as noise) with input from wavelet decomposition of the wave. Generally, [18] relies on the gunshot features, which will not be usable directly to scream detection. Rabaoui et al [7] use one class SVM (support vector machine) to character each class of sound and compare the distance to each class to decide the class of sound. It exams a big set of common audio features such as Zero-Crossing Rate, Spectral Centroid, Spectral Roll-off point, MFCCs, LPCCs, PLP, Wavelet features. According to their experiments, the 1st and 2nd derivative is not much helpful to the performance. It is contrary to the observation in [6] for gunshot feature and our observation in scream features. In [6] it uses Gaussian Mixture Model (GMM) for classification of scream, gunshot and other sounds. A feature selection process is being used to select the 'optimal' features. However it is controversial for the features in MFCC derivatives discarded in [6] are found useful in our study for scream detection. In [19], simple energy feature of audio is used to detect the action scene based on the assumption that there will be high energy in the audio of the action scene. However, in live monitoring system, the energy

978-1-4244-5046-6/10/$26.00 ©2010 IEEE

power of sound will decay with the distance increased between the sensor and the sound source [16]. GMM and HMM are applied in [4] to classify explosion, gunshot and screaming, where MFCC and MPEG7 features such as Spectrum Flatness, Waveform and Fundamental Frequency. Similar to [6], GMM method is applied to model the low level MFCC features for abnormal sound detection [5].

B. *Contribution and organization*

Based on the perceptual features, a new approach to scream detection is designed in this paper. It is based on the energy envelop proposed to locate the continuous sound, the ration of highest pitch of consecutive frames extracted using the autocorrelation measure the 'high' pitch of screaming sound and the compact representation of MFCC in a sliding window for SVM learning.

The paper is organized as follows. Section II describes the brief approaches of the related works. Section III presents the features we are using. Section IV proposes the classifier and system diagram for screaming detection. The experiments and system setup are shown in Section V, followed by the conclusion and future works in Section VI.

II. FEATURE DESCRIPTION

There are many features used for audio analysis. Perceptually we studied some of them to represent the property of scream feature. One is the energy feature, which represents the sound power. However as indicated in [16], the energy power itself is not reliable because the signal decays with distance increase between the sound source and the microphone. Although the sound power may not be high enough as a reliable feature, compared to many other sounds such as normal speech, knocking door, individual laughing etc, scream usually presents as a sound segment with continuous and relative high energy. This feature enables us to distinguish scream from many non-scream sounds.

For convenience, in a live scream detection system we can fix the sampling frequency of audio signal, here we set the sampling rate at $F_s=11k$ per second. With the given F_s, we use different frame length to computing the audio features in the following sections. For energy calculation, around 20ms audio is used as a frame f_e which is 256 samples, x_i, $i=1,\ldots,256$. Experiments show that it is able to detect the short stop reliably between two lauder sounds or the breath between the voice segments. The log energy of a 20ms frame is computed as

$$E = 10\log(\sum_{i=1,\ldots 256} x_i^2) \quad (1)$$

Let a segment of audio signal S(t) is composed of overlapped frames, …, $f_e(k-1)$, $f_e(k)$. There are 128 samples overlap between any two consecutive frames $f_e(k-1)$ and $f_e(k)$. To detect the segment $S(t)$ of a continuous sound until time t, we use the energy measurement, checking if the energy is larger than a threshold T within the segment:

$$S(t) = \{f_e(k) \mid E(k)>T, \text{ any } k\} \quad (2)$$

The duration of $S(t)$ is written as $|S(t)|$. In the experiment setting, we observed that most of the screams last more than 0.5 second, which can be adjusted accordingly later. Based on the probability of duration of the screaming in the audio corpus, and considering the beginning and ending of the scream, we set $|S(t)|>0.3$ which will keep 99% of screaming clips detected using the duration only. If it is too short, many of speech could be detected as scream due to the similarity of the sound 'ah'. If it is too long, many scream segments will be filtered out.

An example of the energy envelope (continuity) is shown in the Figure 1 below. The figure on top is the sound wave consisted of different type of sound, such as speech, scream, (glass) breaking, explosion etc. The figure at the bottom is the

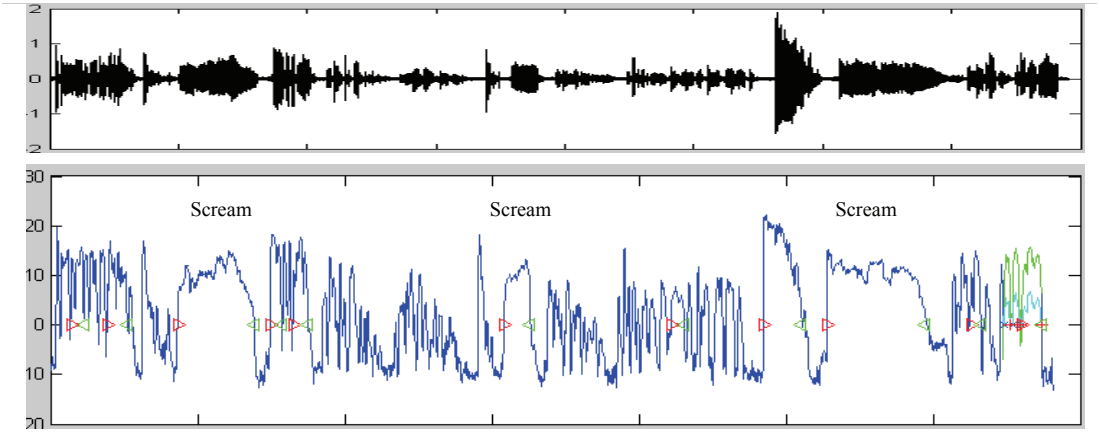

Figure 1 A sound wav and the energy, energy envelop

energy and the envelope detection marked by a pair of triangles, starting from a red triangle and ending at a green one. The time span between the two triangles is $|S(t)|$. In the sound shown, there are three scream segments that are labeled out.

In a scream sound the pitch is usually higher than the normal speech. However, there are both low and high frequency in some of the normal speech or other non-voice sounds. By examining the scream sound, we can see that the high frequency portion of sound is usually falling into certain range with relative high energy if it is a screaming. Based on the observation, we use the autocorrelation technique to extract the high pitch in a sound with high energy. Autocorrelation R of a segment of signal x_i, $i=0,...,N-1$, is

$$R(t) = \sum_i (x_i x_{i+t}) \qquad (3)$$

We are interested in the high frequency part in the signal, which is one of the features of scream. To minimize the noise, we select $N=1024$, $t=1,...,256$. For screaming feature, the frequency is usually higher than *500*. Thus $t=1,...,256$ is enough to capture the feature. For signal length N, if it is too small, the result is not reliable due to the noise presented. If it is too high it may lose local sound feature. By detecting the peaks in the $R(t)$, we can find the high frequency $P(A)$ of the signal. Some typical examples of autocorrelation R is shown in Figure 2. By the natural of the autocorrelation, we are searching for the first high pitch after the first valley in the wave. To remove noise, a smoothing filter is applied to $R(t)$ before computing of the peak.

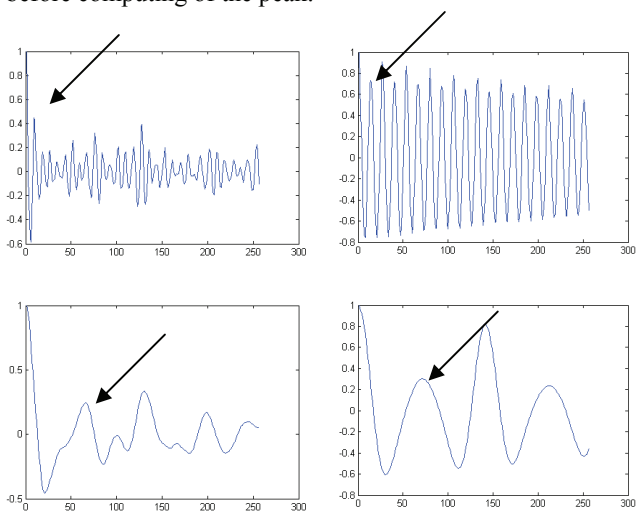

Figure 2 Autocorrelation R of segments of sound. Top row are coming from two scream segments, bottom row coming from two other sound segments. The arrows indicate the detected highest pitch.

One of the challenges in pitch detection using auto-correlation is to set the threshold to detect the real pitch. Fortunately we can use the training data to help setting a threshold *thre*. For every $N=1024$ segment of screaming data, the magnitude of the highest pitch is labeled. Take *1024* samples as a frame for pitch detection, with *512* frames overlapped window, we can extract multiple pitches in one screaming clip. For the *thre*, it can be selected to detect 95% of the pitches in all the scream frames.

The peaks and valleys are first detected by locating local minima or maxima of the normalized wave $R = R/R(0)$.

$$R_1(t) = R(t)-R(t+1)$$

$$R_2(t) = R(t-1)-R(t)$$

$$Peaks = \{t|\ R_1(t)>0\ \&\ R_2(t)>0\}$$

$$Valleys = \{t|\ R_1(t)<0\ \&\ R_2(t)<0\}$$

Then the threshold *thre* is applied to the Peaks and Valleys,

$$P_1 = \{Peaks|\ R_1(Peaks)>thre\}$$

$$V_1 = \{Valleys|\ R_1(Valleys)<-thre\},\ V = min(R(V_1))$$

The first peak in P_1 after the lowest valley V is set as the high pitch in the segment of signal,

$$Pitch\_high = P_1(i),\ i=min(k),\ so\ P_1(k)>V \qquad (4)$$

Some of the detection results are shown in Figure 2.

The distribution of high pitch of screams is in most cases in the range of *(4, 18)*, corresponding to 650Hz to 2200Hz. A statistical analysis of the distribution of the high pitches in a long segment of signal is therefore applied. For the past m pitches $Pitch\_high(n)$, $n=0,1,...,m$, we calculate the ratio of the high pitch similar to a scream to that of low pitch (more like speech) and noise, so we have a high pitch ratio for a segment of audio

$$Ratio(t) = |\{n,\ 4<Pitch\_high(n)<18\}|/(m+1) \qquad (5)$$

The ratio $Ratio(t)$ is one of the features used for detection of screams.

With the detection of high pitch and detection of long segment of high energy, it is still not enough to detect screams accurately. Many false alarms can be reported for sounds with the similar feature, although it can remove most of the sound of human speech and other low frequency sound (for example the explosion). To enhance the detection, we need to introduce features that characterize the sound more precisely.

From many audio features, Mel Frequency Cepstral Coefficients (MFCCs) show the good characterization capability. MFCCs are widely used in speech recognition for they are similar to the perception to human ears. Further the 1st and 2nd derivatives are also used to represent the changes between neighboring frames. Comparing the scream and non-scream audio, we can observe that the MFCCs show quite strong difference for each segment of the sound.

We have tested the MFCCs using single frame of 256 samples for the classification. However the short segment does not show the good performance to classify screams from other sounds. One reason could be that the temporal dynamics are

not covered within a short frame. To characterize the temporal feature, we expand to use 20 frames to extract dynamic MFCCs as the audio feature. For each frame, we have 36 MFCCs (12 MFCC without energy, plus 1st and 2nd derivative of the 12 MFCCs). Here each frame is 256 samples (~20ms). With an overlap of 128 samples, we obtain the next frame. Thus 20 frames signal is about 0.25s.

III. CLASSIFIER ON FEATURES IN SLIDING WINDOW

The diagram of online scream detection system is shown in Figure 3. The input audio signal is sampled at 11k Hz. A overlapped 1024-sample frame is input in the buffer for processing. By analysis of the extracted features characterizing scream, we can detect such sound in real time, only with the delay of a 512-sample frame, which is about 40ms

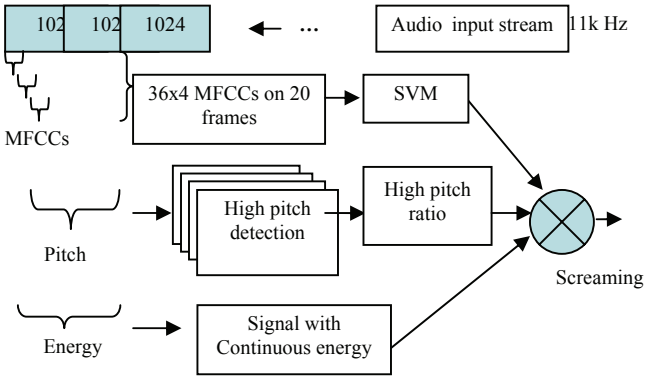

Figure 3. System diagram of live scream detection

The 36x20 MFCCs can be used for training of a SVM. One of the issues is the training time given the size of feature vectors. It also causes the problem of over-fitting in case the number of training samples is not bigger enough. We also extracted and tested the statistics of the MFCCs across 20 frames by obtaining the mean, minimum, maximum and standard deviation of each coefficient in MFCCs. In this way, the features are further reduced to 36x4. The new representation of the compact features is shown in Figure 4.

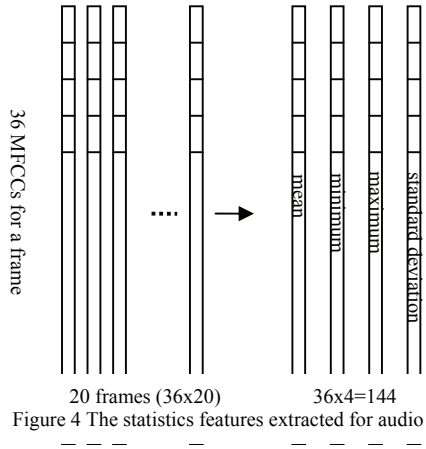

Figure 4 The statistics features extracted for audio signal.

In the training phase, we labeled the scream-segment and non-segments (speech, laugh, applause, cry etc) in a data set and extracted the same amount of MFCC features for scream and non-scream in these segments for SVM training. The experiments show that the classifier using 36x20 features has a better performance for the training set than using 36x4 compact features. However the classifier using 36x4 features performs better in the testing set.

A SVM classifier is represented as

$$f(x) = \sum w_i \phi(x_i, x) + C \qquad (6)$$

Where x is the input MFCC-vector of 144 dimensional over a short sequence of audio, x_i are the support vectors (from training samples), w_i are the weights learned by Support Vector Machine toolbox [24]. A scream is possible if $f(x)>T$, $T=0$. For real application, the threshold T can be tuned with some samples.

Further, similar to the pitch detection, we can apply the statistical analysis to the MFCC-SVM result over a period of time, by detection of more than one positive output within the potential scream segments as the detection of scream.

With the energy distribution, pitch detection and SVM output of MFCCs, we can detect a scream from a segment of audio by applying a GMM classifier to the parameters vector

$$\{|S(t)|,\ ratio(t),\ mean< f(x>,\ x \in S(t)\}. \qquad (7)$$

IV. EXPERIMENTS

A. Experiment setting

To validate the results, we collected the audio corpus from our live recording, the internet and some from movies. The training set of scream contains 26 different scream clips, duration from 1.0s to 10.88s. The testing set contains 56 clips, duration from 0.34s to 9.4s. Within each clips there can be more than one segment of screams. Other non-screaming sounds, including speech (female, male), laugh, cry, applause, clap, knock, glass break, etc, totally have 49 clips for training, from 0.5s to 51.99s. Test set for non-screaming audio has 271 clips for speech, cry, break, applause, knock, laugh, etc., duration from 0.31s to 51.23s.

For simplicity, the performance is measured based on the clips. If there is a detection of scream at any time in a non-scream clip, we call that one false positive (FP: false alarm). If no scream is detected at any time in a scream clip, we call that a false negative (FN: miss detection). The performance is indicated by the FP rate (FPR) and FN rate (FNR):

$$FPR = \frac{number\_of\_FPs}{number\_of\_non\_scream\_clips}$$

$$FNR = \frac{number\_of\_FNs}{number\_of\_scream\_clips}$$

The system has been implemented in Residential Gateway (RG), which is a Linux system (Pentium Mobile CPU~2.0G, 400M front bus, Linux-UBUNTU6.06). The audio sampling rate is 11k Hz and it can complete the detection of a 5s audio clip in less than 1s. A microphone array is used to capture the audio which has a noise removal function can be turned on to further enhance the signal.

B. Results and discussion

To find out the impact of noise on the performance, we conducted four different experiments, trained using the training set with or without the noise removal filtering, and tested using the testing set with or without the noise removal filtering. The noise removal filter is implemented in software to remove the high frequency noise.

The table below summarizes the results based on the above settings.

TABLE I. SCREAM DETECTION RESULTS

Results	Trained without noise removal	
	Tested without noise removal	Tested with noise removal
FPR	12.18%	5.54%
FNR	8.93%	26.79%
	Trained with noise removal	
	Tested without noise removal	Tested with noise removal
FPR	23.62%	11.44%
FNR	3.57%	8.93%

From the results we can see that the screaming detection approach can do quite well if trained properly in the similar environment. With some noise, the system tends to detect more of scream as well as more false alarms. The examination of the data shows that in quite lot cases, the noise if amplified does sounds like screaming in high pitch sound. The main difference is the energy level between the real scream and noise, which we did not use in our method.

The ROC curve of the method is shown in Figure 6, when the signal is trained and tested both with a noise removal filtering.

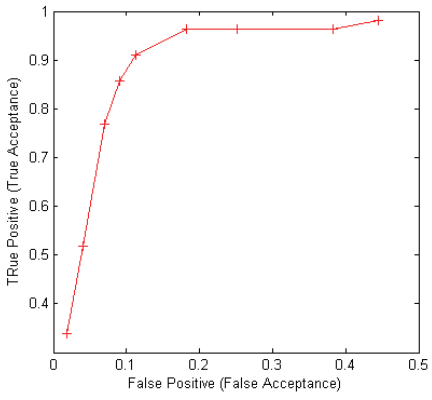

Figure 6 ROC curve of scream detection

Investigating the data, we observed that in a 'scream', there are many segments that sound like amplified noise perceptually, which caused the false negative. By further checking other data, we found that many of the false detections happen in the crying and laugh clips, which actually sound very similar to screams if just listening to these segments. By taking out of these clips from the testing, we can get a result of FAR=7.66% in the case of using noise removal in both training and test phases. The FPR can be reduced in other cases too if we take out of more scream-like sounds from the non-scream clips. In the scream clips, one is distorted due to the recording problem. Removing it from the set, the FNR is reduced to 7.27%.

V. CONCLUSION AND FUTURE WORKS

The scream detection is useful in many surveillance applications, such as home monitoring, public security and safety surveillance. We have proposed a method and system for real time scream detection, by combination of continuity of log energy detection, high pitch analysis and compact MFCCs across frames using SVM. The experiment and promising results are shown for home environment applications. It can be used as a real time alert system as well as a data indexing and retrieval system for audio event searching. One of the alternative uses is for agitation measurement of dementia patient.

The experiments showed that the training based method is useful even there is strong background noise. One example is the screaming in the laughing and applause clips. The method successfully detected some of the short 'screams' in the segments. The future works can be focusing on enlarging the database to test the efficiency of screaming detection, as well as to incorporate the background sounds (speech, and other environment sounds) together with screaming to develop a system for screaming detection, which is more challenging yet more useful for home applications. Another interesting work is to apply a long duration audio analysis for situation awareness, such as detection of laughing, clapping and crying, by characterizing the audio dynamics in a much longer period.

VI. ACKNOWLEDGEMENT

This work is partially supported by EU project ASTRALS (FP6-IST-0028097).

REFERENCES

[1] W. Zajdel, J. D. Krijnders, T. Andringa, D. M. Gavrila, "CASSANDRA: Audio-video Sensor Fusion for Aggression Detection", AVSS 2007, pp.200-205

[2] P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli. Audio Based Event Detection for Multimedia Surveillance. ICASSP 2006

[3] A. F. Smeaton, M. McHugh, Towards event detection in an audio-based sensor network, Proceedings of the third ACM international workshop on Video surveillance & sensor networks, 2005, pp.87-94

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

[5] R. Radhakrishnan, A. Divakaran, "Systematic Acquisition of Audio Classes for Elevator Surveillance", SPIE Image and Video Communications and Processing, Vol. 5685, March 2005, pp. 64-71

[6] L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi, A. Sarti, "Scream And Gunshot Detection In Noisy Environments", EURASIP European Signal Processing Conference, September, Poznan, Poland, 2007

[7] A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri and N. Elouze, "Improved One-Class SVM Classifier for Sounds Classification", *IEEE AVSS*, London, Sept 2007

[8] V. Peltonen, J. Eronen, M. Parviainen and A. Klapuri, "Recognition of Everyday Auditory Scenes: Potentials, Latencies and Cues," The AES 110th Convention, Amsterdam, The Netherlands, 2001

[9] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, T. Sorsa, "Computational auditory scene recognition," IEEE International Conference on Audio, Speech and Signal Processing, Florida, 2002

[10] Ing-Jr Ding, "Events Detection for Audio Based Surveillance by Variable-Sized Decision Windows Using Fuzzy Logic Control," Tamkang Journal of Science and Engineering, Vol. 12, No. 3, 2009, pp. 299-308

[11] J. Foote. "Content-based retrieval of music and audio," In C. C. J. Kuo et al., editors, Multimedia Storage and Archiving Systems II, Proc. SPIE, volume 3229, pp.138-147, 1997

[12] C. Yang, "MACS: Music Audio Characteristic Sequence Indexing for Similarity Retrieval," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, 2001

[13] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," IEEE Trans on Speech and Audio Processing, Vol. 10, No. 5, pp.293-302, July 2002

[14] L. Lu, S. Z. Li and H.-J. Zhang, "Content-Based Audio Segmentation Using Support Vector Machine," IEEE ICME 2001, Tokyo, Japan, Aug, 2001

[15] Olajec J., Jarina R., Kuba M., "GA-Based feature extraction for clapping sound detection," IEEE NEUREL 2006, September 2006, pp. 21-25.

[16] Hiroaki Nanjo, Takanobu Nishiura, Hiroshi Kawano, "Acoustic-Based Security System: Towards Robust Understanding of Emergency Shout," IAS'09, vol. 1, pp.725-728, 2009 Fifth International Conference on Information Assurance and Security, 2009

[17] C. Clavel, T. Ehrette, and G. Richard. Events Detection for an Audio-Based Surveillance System. ICME'05, 2005. pp.1306–1309

[18] Theodore W. Berger, Real time acoustic event location and classification system with camera display, US Patent 7203132 Issued on April 10, 2007

[19] Yun Zhai, Zeeshan Rasheed, Mubarak Shah, "A Framework for Semantic Classification of Scenes Using Finite State Machines," CIVR 2004, LNCS3115, pp.279-288

[20] Xi Li, et al, "Stress and Emotion Classification using Jitter and Shimmer Features," International Conference on Acoustics Speech and Signal Processing 2007 (ICASSP07), Hawii, 2007

[21] Guojun Zhou, John H.L. Hansen, and James Kaiser, "Classification of speech under stress based on features derived from the nonlinear Teager energy operator," ICASSP '98, pp.549-552, 1998.

[22] R. Fernandez and R.W. Picard, "Modeling Drivers' Speech under Stress," Speech Comm., 40:145-159, 2003.

[23] Hugo Meinedo, João Paulo da Silva Neto, "Audio Segmentation, Classification and Clustering in a Broadcast News Task," In International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2003, April 2003

[24] Cawley, G. C., "MATLAB Support Vector Machine Toolbox v0.50 Beta, http://theoval.sys.uea.ac.uk/~gcc/svm/toolbox," University of East Anglia, School of Information Systems, Norwich, Norfolk, U.K. NR4 7TJ", 2000