AN ABNORMAL SOUND DETECTION AND CLASSIFICATION SYSTEM FOR SURVEILLANCE APPLICATIONS

Cheung-Fat Chan[1] and Eric W.M. Yu[2]

1. Department of Electronic Engineering, City University of Hong Kong
83, Tat Chee Avenue Kowloon, HONG KONG
Email: itcfchan@cityu.edu.hk
2. Hong Kong Applied Science and Technology Research Institute (ASTRI)
Hong Kong Science Park, Shatin, HONG KONG

ABSTRACT

A detection and classification system for sound surveillance is presented. A human/non-human voice classifier is firstly applied to separate the input sound into human voice sound or non-human emergency sound. It utilizes a sliding window Hidden Markov Model (HMM) with trained background, human voice and non-human sound templates. In case of human voice, a scream/non-scream classification is performed to detect screaming in an abnormal situation such as screaming for help during bank robbery. In case of non-human sound, an emergency sound classifier capable of identifying abnormal sounds such as gun shot, glass breaking, and explosion, is employed. HMM is used in both scream/non-scream classification and emergency sound classification but with different sound feature sets. In this research, a number of useful sound features are developed for various classification tasks. The system is evaluated under various SNR conditions and low error rates are reported.

1. INTRODUCTION

Acoustic surveillance has advantages over video surveillance in some special situations such as in darkness and in confined area where camera privacy is a public concern [1]. Previous approaches on developing acoustic monitoring system for automatic surveillance include cases such as gunshot detection system based on features derived from the time-frequency domain and Gaussian Mixture Model (GMM) classifier [2]. A two-stage approach is reported in [3] where the first stage is to classify sound events into typical and atypical cases and subsequent processing is done on atypical events such as gun shot, screaming and explosion. An audio surveillance system for typical office environment is reported in [4]. This system employs a background noise model to continuously update for event detection while both supervised and k-means data clustering are observed. In this paper, we look at sound surveillance as a problem to firstly classify sound events into human and non-human sounds, and apply different strategies to do subsequent processing of human screaming or non-human emergency sounds.

2. METHODOLOGY

A block diagram of the proposed sound detection/classification system is shown in Fig. 1. The proposed sound detection and classification system composes of a feature extraction module, a human/non-human voice classifier/detector, a human screaming classifier and a non-human emergency sound classifier.

2.1 Acoustic Feature Extraction

It is very important to derive suitable sound features for the identification of various human and non-human sounds. There are many acoustic features available in literature; they can be roughly grouped into temporal, spectral, parametric and harmonic features [5]. However, it is computational expensive to utilize all these features. In this work, a number of useful features suitable for the targeted acoustic surveillance applications are developed and they will be described in the following sub-sections. Note that a 16 bit A/D converter with 16 kHz sampling is used to digitize the sound. The digitized signal is processed on a frame-by-frame basis with a frame size and frame shift of 512 and 64 samples, respectively.

2.1.1 Weighted Average Delta Energy (Δ_E)

$$\Delta_E(n) = 1.5\sqrt{\frac{1}{M(0.4+v_n)}\left(\sum_{m=0}^{M-1}\delta_{n-m}\right)} \quad (1)$$

where $M = 40$, $\delta_n = |u_n - v_n|$, and

$u_n = 0.8u_{n-1} + 0.2\bar{x}_n$ with $\bar{x}_n = \frac{3.2}{N}\sum_{m=0}^{N-1}|x_{n+m}|$, where x_n is speech time sample, and

$$v_n = \begin{cases} 0.98v_{n-1} + 0.015u_n + 0.005u_{n-1} & u_n > 0.8v_{n-1} \text{ and } u_n > 5u_{\min} \\ 0.8v_{n-1} + 0.15u_n + 0.05u_{n-1} & u_n \le 0.8v_{n-1} \text{ or } u_n \le 5u_{\min} \end{cases}$$

with $u_{\min} = \min[u_n],\ \forall n$ within a frame of N samples.

The advantage of using this delta energy feature is that it is relatively small and smooth for various noise background levels, but relatively large at abnormal acoustic events.

 ISSN 2076-1465

2.1.2 LPC Spectrum Flatness (F_{LPC})

$$F_{LPC}(n) = 0.95F_{LPC}(n-1) + 0.05F_n \quad (2)$$

where $F_n = \frac{2}{L-M}\sum_{k=0}^{L-M-1}\left[\frac{\frac{1}{M}\sum_{m=0}^{M-1}\left|S_{k+m} - \bar{S}_k\right|}{\bar{S}_k} - 0.8\right]$

with $L = 128$, $M = 20$, $\bar{S}_k = \frac{1}{M}\sum_{m=0}^{M-1} S_{k+m}$, and S_k is the current frame LPC magnitude spectrum obtained as $S_k = \left|\frac{\rho}{A(z)}\right|_{z=e^{j\pi k/L}} \quad 0 \le k < L$, where ρ is the LPC gain and $A(z)$ is an order-10 LPC polynomial computed from a frame of 512 audio samples using Hamming window.

2.1.3 FFT Spectrum Flatness (F_{FFT})

$$F_{FFT}(n) = 0.95F_{FFT}(n-1) + 0.05D_n \quad (3)$$

where $D_n = \left|\frac{\sum_{k=0}^{W-1}\left|X_k - \bar{X}_k\right|}{\sum_{k=0}^{W-1} X_k} - 1\right|$

with $\bar{X}_k = 0.8\bar{X}_{k-1} + 0.2\sum_{m=0}^{M-1} X_{k-m+M/2}$, $W = 256$, $M = 20$, and X_k being the FFT magnitude spectrum of x_n

2.1.4 Zero Crossing Rate (R_{ZC})

$$R_{ZC}(n) = 0.95R_{ZC}(n-1) + \frac{0.2}{N}C_{ZC}(n) \quad (4)$$

where $C_{ZC}(n)$ being the count of zero crossing within the current block of size $N = 64$. The count is increased by one if a crossing between +0.01 and -0.01 of the DC removed signal is detected.

2.1.5 Harmonicity (H)

Frequency domain pitch analysis method is used to compute harmonicity [6]. An FFT of size 512 is used to compute the Fourier spectrum for pitch analysis. The analysis is based on minimizing the matching error between the input spectrum and an artificial harmonic spectrum given a pitch value as search variable. Due to the extreme pitch variation of people screaming, the pitch search range is very large and a fixed window of size 512 is not sufficient to capture the pitch variation of various speakers during screaming. A multi-resolution approach for pitch analysis is used. A long buffer of 1024 samples is employed. The middle part of the 512 samples is used for normal range frequency-domain pitch analysis covering the pitch frequency variation from 200Hz to 1.067 kHz. For low pitch speakers, the 1024 sample frame is firstly down-sampled by 2 to have a 512 sample frame for pitch analysis covering the pitch frequency range from 150Hz to 533Hz. For very high pitch speakers, the middle part of 256 samples is firstly up-sampled by 2 to have a 512 sample frame for pitch analysis covering the pitch frequency range from 500Hz to 2.133 kHz. Each pitch detector returns a pitch value. If the signal under investigation is unvoiced, a pitch value of 0 is returned indicating a non-harmonic signal. A counting process over a 40-frame window is used to compute the harmonicity. If any one of the returned pitch values is non-zero, the count is increased by one. Harmonicity is defined as the ratio of the count over 40.

Note that feature 1 to feature 5 are basically sufficient for human/non-human sound classification because these features are quite distinct between human and non-human sounds, for example; human screaming voice sounds have strong harmonicity and LPC spectral flatness while non-human emergency sounds do not. Moreover, non-human emergency sounds have large delta energy and zero crossing rate while human screaming voice sounds do not. For non-human sound classification, two additional features are also developed. These two features are:

2.1.6 Mid-Level Crossing Rate (R_{MC})

$$R_{MC}(n) = 0.95R_{MC}(n-1) + \frac{0.2}{N}C_{MC}(n) \quad (5)$$

where $C_{MC}(n)$ being the count of mid-level crossing within the current block(n) of size $N = 64$. The count is increased by one if a crossing between +0.01 and -0.01 of the positive or negative mid-level of the signal is detected. The mid-level crossing line is defined as the 60% of the maximum value of the current speech block(n).

2.1.7 Peak and Valley Count Rate (R_{PV})

$$R_{PV}(n) = 0.95R_{PV}(n-1) + \frac{0.2}{N}C_{PV}(n) \quad (6)$$

where $C_{PV}(n)$ being the count of number of strong peaks or valleys within the current block(n) of size $N = 64$. A strong peak is identified when a sample is at least 5 times higher than its two adjacent samples. A strong valley is identified when a sample is at least 5 times lower than its two adjacent samples. The count is increased by one if such a peak or valley is detected.

2.2 Human/Non-Human Sound Classification

The human/non-human voice classifier is built based on HMM. The HMM used is an 8-state parallel left-right model with three incoming and outgoing connections in each state. The model has three groups of tied output probabilities. The output probabilities are drawn from a finite set of 256 values, each corresponds to a codeword symbol in a VQ codebook. As a sliding window with width of 100 frames is applied to

evaluate the probabilities from the HMMs for each frame instant, computational complexity is the major concern and discrete density HMM is a desirable choice. For practical implementation, a state can be configured to have 3 transitions from previous states and each transition associates with a transition probability and a group of output probabilities. Three HMMs are trained; one for the voiced harmonic-type of speech signal, one for the non-human sounds which are mostly non-harmonic, e.g., gunshot, glass-breaking, etc. Features 1 to 5 described previously are used for human/non-human sound classification. The third HMM is trained for the background sound for both clean and noisy backgrounds. All groups of training features are hand-marked by visual inspection. They consist of over 400 feature vector sequences for the human voiced case, over 300 feature vector sequences for the non-human sounds, and over 100 vector sequences for the background sounds. All these features were sampled and computed using the screaming and non-human sound databases specially developed by us for audio surveillance application. The sound database consists of human voice (both screamed and non-screamed) from 44 speakers and several types of abnormal sounds including the emergency sounds (gun shot, glass breaking, and explosion) and non-emergency sounds (car braking, sawing, thunder, and car passing, etc). All sound samples are mixed with ambient noise recorded in rental apartment and bank environment under various SNR settings. During evaluation, the unknown feature vector sequence with a width of 100 frames is obtained by sliding through the input signal and fed to the HMM for evaluation against the three trained models. After evaluation, the probability value for each model is obtained:
$P_V(n) = \text{HMM}(M_V, Y_n)$, $P_N(n) = \text{HMM}(M_N, Y_n)$, and $P_B(n) = \text{HMM}(M_B, Y_n)$, where M_V, M_N, and M_B are the trained HMMs for human voice, non-human sound and background signal, respectively, Y_n is the unknown input feature vector sequence for frame *n*. These probability values will then be used by the Decision and Duration Labeling Module for determining the start point and end point of the classified type of signal.

Decision Rules and Duration Marking and Labeling

Given the three probability values in each frame, i.e., P_V, P_N and P_B (note that the frame index *n* is dropped for convenient here), a decision is made to determine the following states of the input signal; HUMAN VOICE, NON-HUMAN SOUND, BACKGROUND and UNSURE. The decision rules are as follows:

Define some sensitivity threshold constants
HHB Human-voice to background margin
HHN Human-voice to non-human sound margin
NNB Non-human sound to background margin
NNH Non-human sound to human voice margin
BBN Background to non-human sound margin
BBH Background to human voice margin

The decision logic is:
if ((P_V - P_B)>HHB) && ((P_V - P_N)>HHN))
    state = HUMAN_VOICE
else
    if ((P_N - P_B)>NNB) && ((P_N - P_V)>NNH))
        state = NON_HUMAN_SOUND
    else
        if ((P_B - P_V)>BBH) && ((P_B - P_N)>BBN))
            state = BACKGROUND
// exception
if ((P_N >1.3 P_V) && ((P_V - P_B)>0.7) && ((P_N - P_B)>0.7))
    state = NON_HUMAN_SOUND
else
    state = UNSURE ;

These sensitivity threshold constants are derived from the training sound database under various SNR conditions. Basically, three sets of thresholds are derived; high sensitivity for SNR<10dB, mid sensitivity for SNR between 10 and 20dB, and low sensitivity for SNR>20dB. A sensitivity set can be selected by the users during surveillance after the system is deployed. The voice state duration marking is based on the true/false transitions of the corresponding state with screening to avoid glitches of short durations and is implemented using a finite state machine. Only short glitches of UNSURE state are allowed between two voice states of the same type. After the marking of start and end positions, the signal within the duration is assigned to either the HUMAN VOICE or the NON-HUMAN SOUND class. The decision is based on the average delta energy and average probabilities of the corresponding states:
$\xi = \frac{1}{N_E - N_S}\sum_{n=N_s}^{N_E}\Delta_E(n)$, $\rho_V = \frac{1}{N_E - N_S}\sum_{n=N_s}^{N_E}P_V(n)$ and $\rho_N = \frac{1}{N_E - N_S}\sum_{n=N_s}^{N_E}P_N(n)$ where N_S and N_E are the start point and end point of the marked duration, respectively.

Given the sensitivity thresholds as:
ETH average delta energy threshold
PVTH average probability threshold for human voice
PNTH average probability threshold for non-human sound

The decision logic is as follows:
For the case of marked human voice duration
If ((ξ >ETH) && (ρ_V >PVTH))
    sound_classified = HUMAN_VOICE
else
    sound_classified = NONE

For the case of marked non-human sound duration
If ((ξ >ETH) && (ρ_N >PNTH))
    sound_classified = NON_HUMAN_VOICE
else
    sound_classified = NONE

Again, these threshold constants are derived with three associated SNR ranges. Fig. 2 demonstrates the successful classifications of a human voice and a non-human sound, respectively, by the proposed classifier. Here the time waveform (above), the spectrograms (middle) and the sound features (below) are shown with the start point and end point marked.

2.3 Human Scream/Non-Scream Classification

Generally, it is very difficult to define the features for human scream. Some people tend to have extreme screaming with highly stretched pitch and raised high frequency components. However, some tend to have very soft screaming which is almost indistinguishable from normal speech. In this research, we apply HMM with classification feature based on likelihood ratio measure of an order-16 LPC system. We divide the training database into two sets of voices manually labeled as screamed and non-screamed by expert listeners. The screamed set consists of pitch-stretched voices such as "呀" ("ah" in English), "救命" ("help") in Mandarin. The non-screamed set composes of normal conversation speech recorded in office environment. Two 8-state parallel left-right HMMs; one for screamed and one for non-screamed are trained using forward-backward algorithm.

2.4 Non-Human Sound Classification

Our system is designed to discriminate between sets of emergency and non-emergency sounds. Seven types of non-human sounds were considered. The emergency sounds consist of gunshot, glass breaking, and explosion while the non-emergency sounds consist of car braking, sawing, thunder, and car passing. In addition to the delta energy Δ_E and zero crossing rate R_{ZC}, mid-level crossing rate R_{MC} and peak-valley count R_{PV} are also used. Hamonicity and spectral flatness are not used because they are not effective for non-human sounds. Again, the classifier is HMM-based. An 8-state parallel left-right HMM for each class of abnormal sounds is trained for this purpose.

3. EVALUATION RESULTS

A total of 6 minutes of human sounds are used for evaluation; within them 3 minutes are labeled as scream sounds. Non-scream sounds are conversation or word utterances in normal conditions. For the non-human sounds, a total of 10 minutes sound recording is used. About 5 minutes of the non-human sounds are emergency sounds (e.g., gunshot, glass break, and explosion sounds). All other non-human sounds are non-emergency sounds (e.g., car braking, sawing, thunder, and car passing). The performance of the proposed system is tested under various noise levels. To evaluate the system in a more realistic scenario, the samples of noise were recorded from the typical operating environments of surveillance systems, including the rental apartment and the bank and were mixed with the sounds under test. The sound detection/classification errors of the proposed system are summarized in Table 1 and Table 2.

Table 1 Classification Error of the Human/Non-Human Sound Classifier

SNR	Error Rate (%)
30dB	5.5
20dB	8.3
10dB	13.5
5dB	18.7

Table 2 Detection Errors of the Human Scream Detector and the Emergency Sound Detector

SNR	Equal Error Rate (%)	
	Human Scream Detection	Emergency Sound Detection
30dB	10.7	9.8
20dB	13.3	13.2
10dB	15.0	16.3
5dB	18.0	19.8

As the missing of a critical event can be costly, sound surveillance system is usually designed to tolerate more false alarms than false rejections. Table 3 is an example that shows the detection results when the decision threshold is adjusted to trade false alarm rate for miss rate under various noise levels. In a typical operating environment where the range of SNR levels varies from 30dB to 10dB, the average classification error rate of the human/non-human sound classifier is 9.1%. The average miss rates of the scream detector and emergency sound detector in Table 3 are 8.4% and 8.6%, respectively.

Table 3 Performance of Example Detectors in Typical Surveillance Application

SNR	Human Scream Detection		Emergency Sound Detection	
	False Alarm Rate (%)	Miss Rate (%)	False Alarm Rate (%)	Miss Rate (%)
30dB	18.4	6.6	16.3	7.6
20dB	19.5	8.3	19.2	8.2
10dB	22.1	10.2	20.5	10.1
5dB	23.6	13.9	25.8	14.2

4. CONCLUSION

An abnormal sound detection and classification is reported. A number of useful sound features are developed for the classification tasks. A sliding window HMM with trained background, human voice and non-human sound templates is applied for the classification of human/non-human sounds with an error rate of 5.5% at 30 dB SNR. This classifier provides the first stage screening of sounds for subsequent detections of human scream and non-human emergency sounds which achieves good performances with miss rates of 6.6% and 7.6%, respectively, at 30 dB SNR.

REFERENCES

[1]. 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.

[2]. L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi, and A. Sarti, "Scream and gunshot detection in noisy environments," in *EUSIPCO*, Poznan, Poland, Sept 2007.

[3]. Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis, "On Acoustic Surveillance of Hazardous Situations," in ICASSP, pp165-168, Taiwan, April, 2009.

[4]. A. Harma, M.F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *in Proc of IEEE ICME*, pp.634-637, 2005.

[5]. G. Peeters, A large set of audio features for sound description (similarity and classification) in the cuidado project, CUIDADO Project Report, 2003.

[6]. C.-F. Chan and E.W.M. Yu, "Improving pitch estimation for efficient multiband excitation coding of speech," *IEE Electronics Letters,* vol. 32, no. 10, pp. 870-872, May 1996.

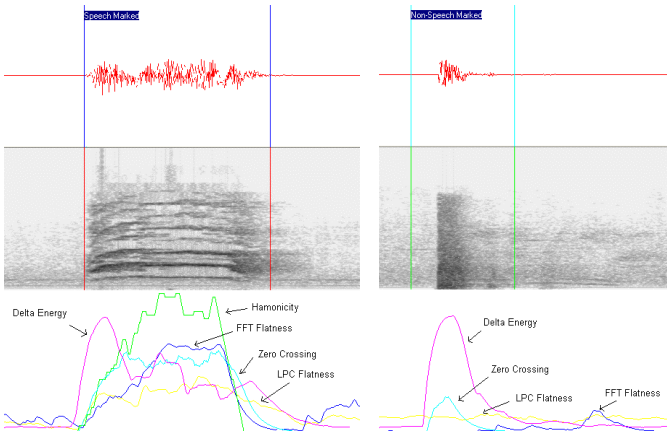

Fig. 2 Demonstration of Successful Classification of: (a) Human Voice – an "Ah" sound and (b) Non-Human Sound – a "Gun Shot" Sound

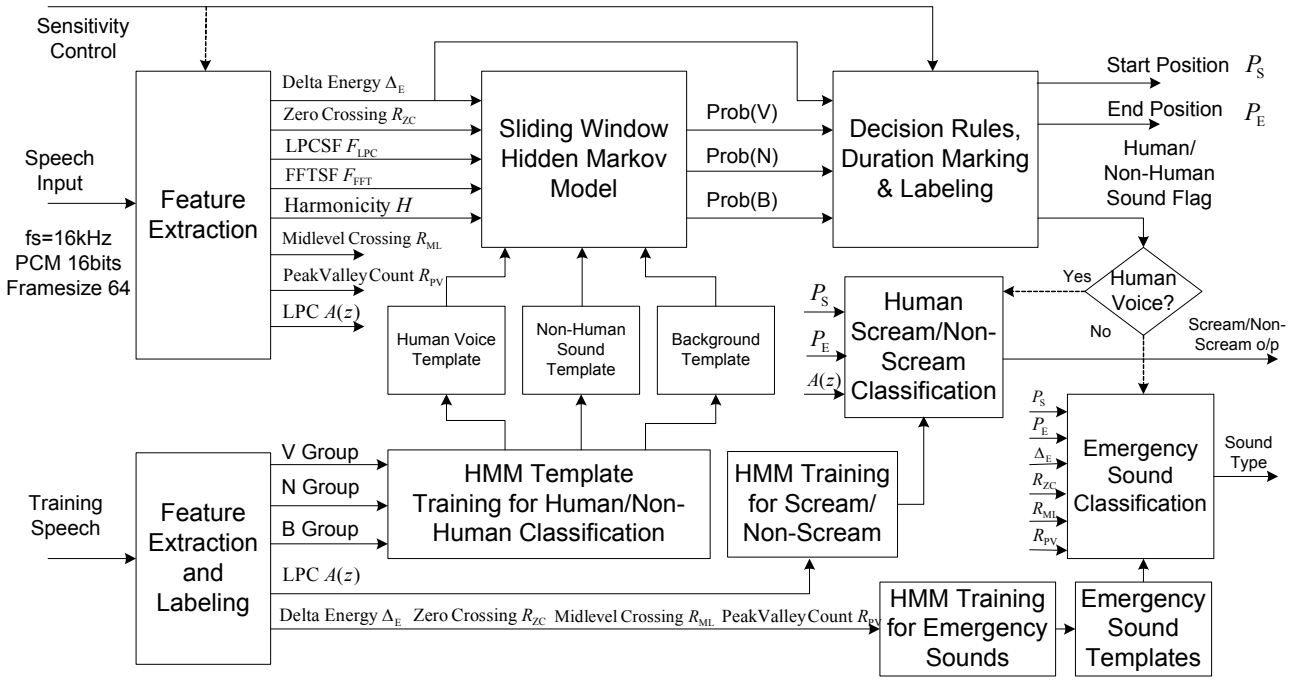

Fig. 1 The Proposed Sound Detection and Classification System