A REAL-TIME UNUSUAL VOICE DETECTOR BASED ON NURSING AT HOME

MIN-QUAN JING, CHAO-CHUN WANG, LING-HWEI CHEN

Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan 300, ROC
E-MAIL: mqjing@msn.com, lupin@debut.cis.nctu.edu.tw, lhchen@cc.nctu.edu.tw

Abstract:

In this paper, we will propose a method to detect an unusual voice for nursing system. Based on the healthy condition of a person, we define four kinds of unusual voices including cough, groan, wheeze and cry for help. When the person nursed sends out the unusual voices, we judge that his health condition have a doubt, and need someone to pay attention. In order to detect the unusual voices, we extract five features on audio waveform, including the number of segmented parts, duration of waveform, mean of volume, zero crossing rate and correlation. Experimental results show that the detection rate is 94%~97% for these four kinds of unusual voices. In false alarm, there are only 0.08% of wrong rates.

Keywords:

Nursing system; Cough; groan; Wheeze; Cry for help; Zero crossing and correlation

1. Introduction

While the physically handicapped are receiving increasing concern in modern social welfare, it is necessary to develop a home nursing system to reduce the social costs and labor costs. Most home nursing systems track user health condition by means of video or electronic devices. Huang *et al.* [1] employ video to track if users get the attention of medical staff for their alleged health problems with voluntary warnings signified by special actions, such as pointing out their index finger. Such an approach is questionable. First, whether or not a person having a physical problem and needing immediate medical attention can correctly make a help sign is an open case. Second, users under video surveillance have no privacy at all. Huang and Wu *et al.* [2] installed a device on user to detect the electrocardiograph (ECG) in order to determine the health condition and send back to a medical center. In this case, users are in a passive situation and are given no choice in voluntarily sending critical voice signals. In current state, voice detection techniques can be used on nursing system. Zwicker [3] points out that an ordinary human voice falls within a lower frequency range. Based on their work, audio signals at a lower frequency can be considered as speeches. Saunders [4] employs the normalized zero-crossing rate (ZCR) as the feature to distinguish speeches from non-speeches. Abu-El-Quran and Goubran [5] identify speeches and non-speeches by calculating the pitch of audio signals with the pitch ratio. However, these two-way classification methods are inadequate in practice. The multi-way classification methods are receiving increasing concern. Wyse and Smoliar [6] classify ordinary audio signals into music, speech and other. Kimber and Wilcox [7] further classify audio signals into speech, silence, laughter and non-speech, marking a step forward in the two-way classification. Lastly, Lin and Chen [8] classify audio signals into five classes: pure speech, music, solo (unaccompanied) human singing, background music speech, and background noise speech. Though these techniques can effectively classify audio signals, they fail to further classify pure speech in greater detail. In this paper, we develop a voice-detection-based nursing system which classify pure speeches into severe coughing, groaning, wheezing and help-asking sound, and sounds that do not belong to the these four classes of signals. Experimental results show that the delectability of the proposed method is up to 94-97%; while the false alarm rate is only 0.08%.

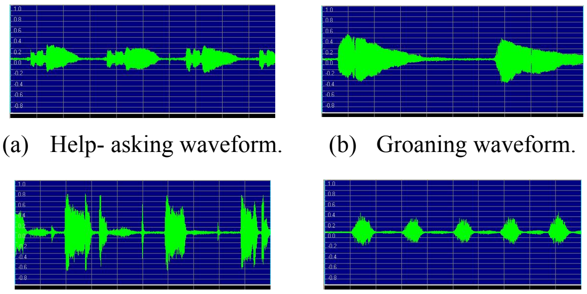
The remaining of the paper is described as follows. In Section 2, the feature vectors of the four critical voices (help-asking, groaning, coughing, and wheezing) and the extraction method are defined. Section 3 presents an algorithm to classify critical voice signals with these features. In Section 4, some experimental results are given to demonstrate the effectiveness of the proposed method. Section 5 makes a conclusion.

2. Feature Extraction and Selection

Four critical voice signals are defined as audio keywords: help-asking sound, groaning, severe coughing

978-1-4244-3703-0/09/$25.00 ©2009 IEEE

and wheezing. A total of 92 samples of these four types of critical voice signals, each 23 samples, were collected to build a database for locating their distinctive features. The sampling criteria of audio signals are as follows: length five seconds, sampling rate 16KHz, resolution 16 bits, mono channel digital signal. After collecting the samples, audio segments were extracted based on waveform, and possible features were located and analyzed with statistics in order to select the appropriate features. The waveforms of the critical voice signals are shown in Fig. 1. The following five items were selected as the potential features: segments, mean volume, segment duration, spectrogram correlations, and normalized ZCR.



(a) Help- asking waveform. (b) Groaning waveform.

(c) Coughing waveform. (d) Wheezing waveform.

Figure. 1. The waveforms of the critical voice signals.

2.1. Segmentation

As shown in Fig. 1, user critical voice signals have a distinctive cycle, except coughing. And they are quite distinctive from the sounds usually picked up around users, e.g. conversation and TV sound. Therefore, in order to effectively detect the cycle of critical voice signals, individual continuous audio intervals, i.e. voiced segments, were extracted by means of pre-processing and segmentation, and every voiced segment was analyzed. First is the smoothing of waveform. Based on each sampling point x on the waveform, a smooth window is applied for 16 points on the left and right, respectively. Eq. (2-1) is adopted the task.

$$\overline{f(x)} = sign(f(x)) \times \frac{1}{33}\sum_{i=-16}^{16}\left|f(x+i)\right|, \qquad (2\text{-}1)$$

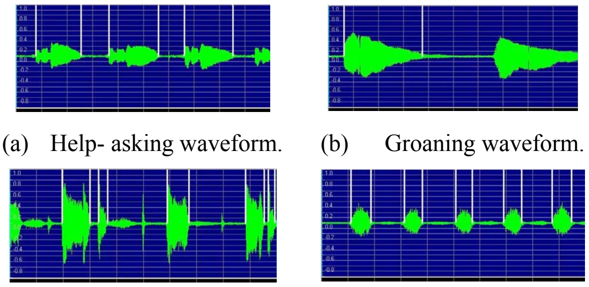
where $f(x)$ is the original volume signal, and $\overline{f(x)}$ is the volume signal after smoothing.

Second is noise removing of smoothed signals $\overline{f(x)}$. We use Eq. (2-2) to handle this.

$$\overline{\overline{f(x)}} = \begin{cases} 0 & if \left|\overline{f(x)}\right| \le 1500 \text{ or } \left|\overline{f(x)}\right| \le (L/6) \\ \overline{f(x)} & \text{otherwise} \end{cases}, \qquad (2\text{-}2)$$

where $\overline{\overline{f(x)}}$ is the signal after noise removing and $L = \max(|\overline{f(x)}|)$.

For sudden and transient noise is detected occasionally, it is assumed that a voiced segment lasts for 1/16 second, and ranges smaller than this length and with a high volume are considered as sudden noise and will be deleted. After smoothing and noise removing, the voiced segments and voiceless segments in the five-second audio signal are clearly distinguished. Fig. 2 shows the audio signals after segmentation. The start and end points of each voiced segment are indicated with a white line.



(a) Help- asking waveform. (b) Groaning waveform.

(c) Coughing waveform. (d) Wheezing waveform.

Figure2. The waveforms of the critical voice signals after Smoothing and Noise Removing.

2.2. Feature Analysis

As shown in the waveform of the four critical voice signals after segmentation, every voiced segment within the help-asking, groaning and wheezing is very similar and has a cycle; i.e. the pattern of voiced segments is regular. As shown in the waveform of coughing, it is clear that the mean volume of voiced segments is higher than that of other critical voice signals. By applying the commonly used normalized ZCR, the five potential features obtained are defined as follows:

- Number of segments (*N*): the number of voiced segments within a five-second audio signal.
- Mean volume (*M*): the mean of the absolute volume value of all voiced segments calculated as shown in Eq. (2-3):

$$M = \frac{1}{\sum_{n=1}^{N}\left(E(n)-S(n)\right)}\sum_{n=1}^{N}\sum_{x=S(n)}^{E(n)}\left|f(x)\right|, \qquad (2\text{-}3)$$

where N is the number of segments within a five-second audio signal; *S(n)* and *E(n)* the start and end points of the *n*-th segment expressed in sampling point; and *f(x)* the volume value.

- Duration (D_n): the duration of each voiced segment calculated as shown in Eq. (2-4):

$$D_n = E(n) - S(n) \quad , \qquad (2\text{-}4)$$

where n is n-th segment expressed in sampling point.

- Correlations (*C*): the correlations between the form of voiced segments and distinctive form of various sounds. How to obtain the distinctive form of various sounds and calculate their correlations are described below.
- Normalized *ZCR*: the frequency of the up and down vibrations passing through the zero point of all segmented voiced segments divided by the total duration of all voiced segments as shown in Eq. (2-5):

$$ZCR = \frac{ZCRN}{\sum_{n=1}^{N} \left(E(n) - S(n)\right)} \quad , \qquad (2\text{-}5)$$

where ZCRN is the frequency of the up and down vibrations passing through the zero point of all segmented voiced segments within a five-second audio signal.

2.3. Feature Statistics and Selection

Table 1 shows the statistics results on the samples according to the features. We can see that the mean volume and normalized ZCR of coughing is significantly different from that of other sounds. Therefore, these two items will be the distinctive features of coughing.

Table 1. Feature Statistics

	N	Mean Volume	Duration D_n	Normalized ZCR
Help-asking Sound	3~4	1000 ~1250	11800 ~14900	0.08 ~0.12
Groaning	1~2	800 ~2200	17800 ~29200	0.11 ~0.115
Coughing	3~11	5500 ~7500	1200 ~18400	0.14 ~0.19
Wheezing	5~6	400 ~1000	3500 ~6500	0.21 ~0.24

Next, the number of segments (N) and segment duration (D_n) are selected as the distinctive features of help-asking sound, groaning and wheezing. We explain the determination of the thresholds and the threshold ranges in Tables 2-6.

Table 2. Mean, Variant and Feature Threshold of the Mean Volume of Coughing.

	Mean Volume	
	$Mean_M$	$Variance_M$
Coughing	$\dfrac{\sum_{f=1}^{23} M_f}{23}$	$\left(\dfrac{\sum_{f=1}^{23} (M_f - Mean_M)^2}{23}\right)^{1/2}$

Table 3. Mean, Variant and Feature Threshold Range of the Normalized ZCR of Coughing.

	Normalized ZCR	
	$Mean_{ZCR}$	$Variance_{ZCR}$
Coughing	$\dfrac{\sum_{f=1}^{23} ZCR_f}{23}$	$\left(\dfrac{\sum_{f=1}^{23} (ZCR_f - Mean_{ZCR})^2}{23}\right)^{1/2}$
Threshold Range	$Mean_{ZCR} - Variance_{ZCR} \le TR_{ZCR} \le Mean_{ZCR} + Variance_{ZCR}$	

Table 4. Mean, Variant and Feature Threshold Range of the Duration of Help-asking Sound.

	Duration	
	$Mean_{HD}$	$Variance_{HD}$
Help-asking Sound	$\dfrac{\sum_{f=1}^{DN} HD_f}{DN}$	$\left(\dfrac{\sum_{f=1}^{DN} (HD_f - Mean_{HD})^2}{DN}\right)^{1/2}$
Threshold Range	$Mean_{HD} - Variance_{HD} \le TR_{HD} \le Mean_{HD} + Variance_{HD}$	

Table 5. Mean, Variant and Feature Threshold Range of the Duration of Groaning.

	Duration	
	$Mean_{GD}$	$Variance_{GD}$
Groaning	$\dfrac{\sum_{f=1}^{DN} GD_f}{DN}$	$\left(\dfrac{\sum_{f=1}^{DN} (GD_f - Mean_{GD})^2}{DN}\right)^{1/2}$
Threshold Range	$Mean_{GD} - Variance_{GD} \le TR_{GD} \le Mean_{GD} + Variance_{GD}$	

Table 6. Mean, Variant and Feature Threshold Range of the Duration of Wheezing.

	Duration	
	$Mean_{WD}$	$Variance_{WD}$
Wheezing	$\dfrac{\sum_{f=1}^{DN} WD_f}{DN}$	$\left(\dfrac{\sum_{f=1}^{DN}(WD_f - Mean_{WD})^2}{DN}\right)^{1/2}$
Threshold Range	$Mean_{WD} - Variance_{WD} \le TR_{WD} \le Mean_{WD} + Variance_{WD}$	

Table 2 shows the calculation of the mean and variance the mean volume of coughing and the determination of its threshold. In the table, f refers to the index of coughing files in the database, and M_f is the mean volume of the f -th file. $Threshld_M$ is the threshold of the mean volume of coughing. In the testing phase, after receiving any five-second audio signal with the mean volume greater than the threshold, the system will determine it as a feature of coughing. Table 3 shows the calculation of the mean and variance the normalized ZCR of coughing and the determination of its threshold range. In the table, f refers to the index of coughing files in the database, and ZCR_f is the normalized ZCR of the f -th file. TR_{ZCR} is the threshold range of normalized ZCR. In the testing phase, after receiving any five-second audio signal with the normalized ZCR within this threshold range, the system will determine it as a feature of coughing. Table 4 shows the calculation of the mean and variance the duration of help-asking sound and the determination of its threshold range. In the table, f refers to the index of voiced segment of help-asking sound in the database, HD_f is the duration of the f -th voiced segment, and DN is the total number of voiced segments. TR_{HD} is the threshold range of the duration of help-asking sound. In the testing phase, after receiving any five-second audio signal containing a voiced segment whose duration falls within this threshold range, the system will determine it as a feature of asking for help. Table 5 shows the calculation of the mean and variance the duration of groaning and the determination of its threshold range. In the table, f refers to the index of voiced segment of groaning in the database, GD_f is the duration of the f -th voiced segment, and DN is the total number of voiced segments. TR_{GD} is the threshold range of the duration of groaning. In the testing phase, after receiving any five-second audio signal containing a voiced segment whose duration falls within this threshold range, the system will determine it as a feature of groaning. Table 6 shows the calculation of the mean and variance the duration of wheezing and the determination of its threshold range. In the table, f refers to the index of voiced segment of groaning in the database, WD_f is the duration of the f -th voiced segment, and DN is the total number of voiced segments. TR_{WD} is the threshold range of the duration of wheezing. In the testing phase, after receiving any five-second audio signal containing a voiced segment whose duration falls within this threshold range, the system will determine it as a feature of wheezing. Table 7 shows the actual values of the threshold range of features calculated with the equations specified in Tables 2-6. Based on this table, we can test a voice signal and the detail is introduced in the next section.

Table 7. Feature Threshold Range Statistics.

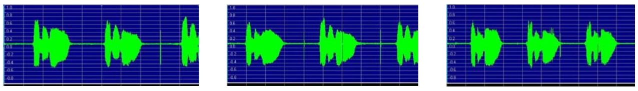
	N	Mean Volume	Duration D_n	Normalized ZCR
Help-asking Sound	3~4		12256 ~14892	
Groaning	1~2		18863 ~28903	
Coughing		5498		0.152~0.189
Wheezing	5~6		3755 ~6292	

3. Unusual Voice Detection

As the pattern of the voiced segments of help-asking sound, groaning and wheezing is comparatively regular, the spectrogram correlations are applied to the reference for determining these three critical voice signals. The pattern of these three control voiced segments can be obtained through training data and then stored in the system as the criteria for comparing with the segmented voiced segments.

3.1. Feature Statistics and Selection

In order to accelerate the computational efficiency, we re-sample every segmented voiced segment. Let $RSN(i)$ is configured as the number of sampling point of each class. In this paper, the values of $RSN(i)$ are 13, 23 and 5 for $i = 1,2,3$ (1: help-asking sound; 2: groaning; and 3: wheezing). $D_n(i)$ is the duration of the n -th voiced segment in the waveform of respective classes. This way, the following equation is introduced $R(i) = \frac{D_n(i)}{RSN(i)}$; where $i = 1, 2, 3$ (1: help-asking sound; 2: groaning; and 3: wheezing). In practice, $R(i)$ intervals on the waveform are merged into one point, and the mean of the absolute values of $R(i)$ interval is the value of the merged point. As a result, the help-asking sound, groaning and wheezing are expressed in 13, 23 and 5 sampling points respectively. Observation shows that there are 3 subclasses in the similarity of the spectrogram of segmented voiced segments of help-asking sound as shown in the first voiced segment in Fig. 3. The same result is also found in groaning, which is also further classified into 3 subclasses as shown in the first voiced segment in Fig. 4.

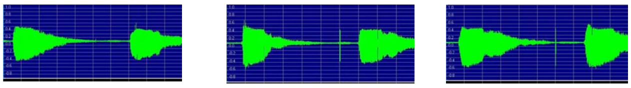


(a) Waveform of Subclass 1 of Help-asking.

(b) Waveform of Subclass 2 of Help-asking.

(c) Waveform of Subclass 3 of Help-asking.

Figure3. The waveforms of subclasses of Help-asking Sound.



(a) Waveform of Subclass 1 of Groaning.

(b) Waveform of Subclass 2 of Groaning.

(c) Waveform of Subclass 3 of Groaning.

Figure4. The waveforms of subclasses of Groaning Sound.

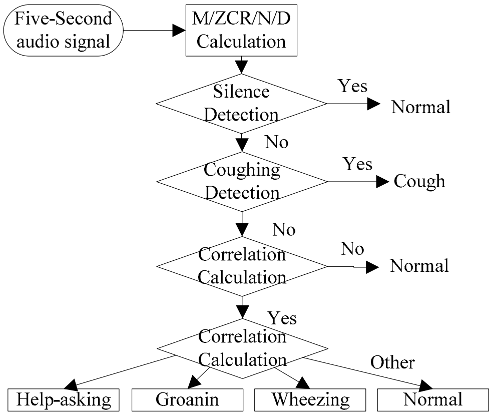
The voiced segments of wheezing are the most stable and require no further classification. The mean voiced segment of each subclass is selected as the control voiced segment for the system to perform comparisons. Details of the calculation of the mean voiced segment are as follows.

All voiced segments in a subclass: $X^1 X^2 \cdots X^k \cdots X^n$; where n is the number of segments within the voiced interval. $X^k = (x_1{}^k, x^k{}_2 \cdots x_j^k \cdots x^k{}_{RSN_{(i)}})$, $1 \le k \le n$, where $x_j{}^k$ is the re-sampling point and i is the class. Mean voiced segmet $X = (x_1, x_2 \cdots x_l \cdots x_{RSN(i)})$, $1 \le l \le RSN(i)$, where $x_l = MEAN(x_l{}^1, x_l^2, \cdots x_l^n)$. When running the program, the system will re-sample all segmented voiced segments into $RSN(i)$ number of sampling points to perform the correlation calculation of the mean voiced segment of the i-th class. The correlation calculation is shown in Eq. (3-1), where X is the mean voiced segment, and Y the re-sampled voiced segment segmented by the system. Then, the class of the segmented voiced segment, i.e. help-asking sound, groaning or wheezing, is determined based on the greatest value.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{3-1}$$

3.2. Determination Criteria and System Process of Various Sounds

Concluding the above analysis results (Table 7), and the results of calculation of the correlations between the mean voiced segments and samples of various classes in the database, the distinctive features and optimal threshold range of various critical voice signals are determined. The main judgment process of the system is illustrated below.



- Silence Detection: *N*=0; when there is no voiced segment segmented in a five-second audio signal after processing, it is determined as silence.
- Coughing Detection: (*M*>5500) and (0.15<*ZCR*<0.2); when the mean volume and normalized ZCR of a five-second audio signal after processing fall within the threshold range, it is determined as coughing.
- Correlation Calculation: is the calculation of the correlations between the voiced segments segmented from a five-second audio signal and the control mean voiced segments of help-asking sound, groaning and wheezing in the system. The system will continue to compare the following criteria if one of the help-asking voiced segments is greater than 0.7; one of the groaning voiced segments is greater than 0.9; or one of the wheezing voiced segments is greater than 0.8.
- Help-asking Sound Detection: (*N*=3 or 4) and (*Dn*=12000~15000); when a five-second audio signal meets the correlation calculation requirements of help-asking sound, and the number of segments and duration of one of the voiced segments falls within the

threshold range, it is determined as a help-asking sound.

- Groaning Detection: (*N*=1 or 2) and (*Dn*=17000~29000); when a five-second audio signal meets the correlation calculation requirements of groaning, and the number of segments and duration of one of the voiced segments falls within the threshold range, it is determined as a groaning.
- Wheezing Detection: (N=5 or 6) and (Dn=3500~6500); when a five-second audio signal meets the correlation calculation requirements of wheezing, and the number of segments and duration of one of the voiced segments falls within the threshold range, it is determined as wheezing.

4. Experimental Results

In this paper, we develop a real-time unusual voice detector system under Intel Pentium IV 2.8Hz platform with Bluetooth microphone and the receiver. The audio format is mono channel audio samples of 16 KHz sampling rate with 16 bits resolution. In order to show the effectiveness of the proposed method, the system was operated for 150 minutes. The length of data is five-second, and they were processed in alternative seconds; therefore, data of 4 seconds are overlapped. A total of 8996 entries of data (150*60-4) were processed. Sounds in the experiment included talking sound, phone ringing, TV sound and ambient sound. There were help-asking sound, groaning, coughing and wheezing, each 150 times during the system operation of 150 minutes. Table 8 shows the detection results. The correct detections are 145, 144, 143, and 142 respectively; the number of normal situation detections is 8396, including 7 false alarms. Coughing, 4 times, is the most common false alarms.

Table 8. The Detection Results.

Class (Q'ty)	Help-asking sound	Groaning	Coughing	Wheezing	Normal
Help-asking sound (150)	145 (96.7%)	0 (0%)	0 (0%)	0 (0%)	5 (3.3%)
Groaning (150)	0 (0%)	144 (96%)	0 (0%)	0 (0%)	6 (4%)
Coughing (150)	0 (0%)	0 (0%)	143 (95.3%)	0 (0%)	7 (4.7%)
Wheezing (150)	0 (0%)	0 (0%)	0 (0%)	142 (94.7%)	8 (5.3%)
Normal (8396)	0 (0%)	1 (0.01%)	4 (0.05%)	2 (0.02%)	8389 (99.92%)

5. Conclusions

This paper proposed a real-time usual voice detection based home nursing system. Results of the experiment indicate that the system detectability of help-asking sound, groaning, severe coughing and wheezing is up to 94-97%; while the false alarm rate is only 0.08%. When compared with the previous video-based nursing system, the proposed system can protect the privacy of users. Improvements may be achieved by employing new de-noise methods in preprocessing and delicately designed the database.

Acknowledgements

This work is supported in part by National Science Council (NSC 97-2221-E-009-137-.).

References

[1] C. L. Huang and W. Y. Huang, “Sign language recognition using model-based tracking and 3-D Hopfield neural network,” Machine Vision and Application, vol.10, no. 5, pp. 292-301, 1998.

[2] T. Y. Huang, M. L. Wu, T. Y. Chiang, Y. D. Lin and H. W. Chung, “Design and implementation of a Fascimile electrocardiographic system for convenient remote monitoring,“ National Taiwan University Engineer Magazine, vol.88, pp. 43-50, 2003.

[3] E. Zwicker and H. Fastl, “Psychoacoustics: Facts and Model,” Springer, 1990.

[4] J. Saunders, “Real-time discrimination of broadcast speech/music,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing, vol. 2, Atlanta, GA, pp. 993-996, May 1996.

[5] A. Abu-El-Quran, and R. Goubran, “Pitch-based feature extraction for audio classification,“ in Proc. of the 2nd IEEE International Workshop on Haptic, Audio and Visual Environments and Their Applications, Ottawa, Canada, pp.43-47, Sep. 2003.

[6] L. Wyse and S. Smoliar, “Toward content-based audio indexing and retrieval and a new speaker discrimination technique,” in Proc. ICJAI’95, Singapore, Dec. 1995.

[7] D. Kimber and L. D. Wilcox, “Acoustic segmentation for audio browsers,” in Proc. Interface Conf., Sydney, Australia, July 1996.

[8] R. S. Lin and L. H. Chen, “A new approach for classification of generic audio data,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 19, no. 1, pp.63-78, 2005.