CLASSIFICATION OF HUMAN SOUNDS USING SUPPORT VECTOR MACHINE WITH PSYCHOACOUSTIC DATA

Shahzad Ahmed, Hyun In Jo, Jin Yong Jeon

Department of Architecture engineering, Hanyang University, 04763 Seoul, Korea

Email: jyjeon@hanyang.ac.kr

This paper presents the classification of human sounds based on support vector machine (SVM) using psychoacoustic data. A scream classification model, with sounds of speech and screams indicating different acoustical characteristics, was investigated. Temporal changes were observed by evaluating the physical characteristics of waveforms and spectrograms with psychoacoustic parameters, including loudness and sharpness. Mel frequency cepstral coefficients were used to identify the spectral energy distribution of screams. Further, a Mel filter bank and frequency band filter were used to extract the high spectral energy, and differentiate between the lower and higher energy spectra. The classification accuracy was improved by combining the SVM with the psychoacoustic parameters of scream sound.

Keywords: Human sound classification, Support vector machine, Mel frequency cepstral coefficients, Psychoacoustics.

1. Introduction

Acoustic signals classification is widely used to detect hazardous situations, commonly called as an acoustic surveillance system. Although video surveillance approaches for safety purpose has been widely implemented with considerable accuracy [1], but the detection of abnormal situations based on acoustic signals has a low-cost computation, compared with video or image processing [2]. In comparison with speech, scream sound is generated by a sudden air explosion. This sound has different spectral and temporal characteristics, in comparison with other ambient sounds [3]. Previous studies considered audio signal classification doubly, as a signal processing problem [4] and as a pattern detection problem [5]. Currently, robust and efficient algorithms are based on machine learning approaches driven by extracted features.

Classification based on features extracted by mel frequency cepstral coefficients (MFCC) is widely used for scream detection and also for detection and classification of other events, such as cough sound [6]. Several combinations of feature extraction and classifiers exist: the MFCC with a Gaussian mixture model (GMM) classifier [7], the MFCC with a hidden Markov model (HMM) [1, 8], and the support vector machine (SVM) classifier [9]. Further, MFCC based cough classification algorithm was also presented by [6]. All of these classifiers directly used audio signals, without considering acoustic parameters (loudness & sharpness) and spectrogram analyses.

The proposed method comprises a machine learning approach with feature extraction, and the psychoacoustic evaluation of scream and speech sounds. Furthermore, the method consists of database creation for different screams and speech, followed by feature extraction and classification. The paper is organized as follows: Section 2 presents the spectral and temporal comparison between
speech and scream sounds, and includes psychoacoustic parameters of scream sound. Section 3 provides the details of feature extraction. Section 4 presents the classification and experimentation. The summary is then presented in Section 5.

2. Scream sound characteristics

2.1 Temporal and spectral characteristics

Speech is comprised mostly of low frequency components, and in comparison to scream sound, it has a higher concentration at the frequency ranging from 0–4 kHz. Conversely, the frequency spectrum of scream sound is in the range of 0–12 kHz frequency. Figure 2 demonstrates the spectral and temporal variations occurring in speech and scream sounds. The lower portion of the figure represents the audio signal in the time domain, and the upper portion shows the time vs frequency spectrogram representation. The color intensity represents the power of each frequency component according to the specific time. Higher the intensity, higher is the value of frequency components. Figure 2(a) presents the spectrogram of the female scream sound. The pattern of scream contains distinctive harmonics that are not visible in the continuous speech signal.

Figure 2: Scream sound and its spectrogram (a) speech and its spectrogram (b)
2.2 Psychoacoustical parameters

Zwiker’s psychoacoustic parameters for different scream sounds were analyzed. Figure 3 shows the acoustical parameters (loudness and sharpness) of three different kind of screams chosen randomly, and these values are quantified in Table 1. There is a similar pattern for the different screams visible in Fig. 3. Furthermore, this pattern is investigated using an SVM classifier.

![Figure 3: Loudness of different screams (a), sharpness of different screams (b)](image)

<table>
<thead>
<tr>
<th>Mean</th>
<th>Scream</th>
<th>Scream 2</th>
<th>Scream 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loudness</td>
<td>29.1</td>
<td>24.5</td>
<td>23.4</td>
</tr>
<tr>
<td>Sharpness</td>
<td>1.91</td>
<td>1.91</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Table 1: SQ parameters: Loudness, Sharpness

3. Features extraction

3.1 Data acquisition

While selecting the training dataset, it should be considered that the selected data must contain within it the features required for the overall classification process. Moreover, training data should also contain variations within it, as this contributes to the robustness of the process. The database comprised of speech and scream sounds was created using different online resources. Female screams were gathered from different age groups, to include tonality differences within the data set.

3.2 Features vector

Our framework uses a modified MFCC technique to extract features from audio signals. Moreover, MFCC uses a mel scale that is based on human hearing, i.e., humans are more sensitive to low frequencies and can distinguish even small changes occurring at low frequencies. In contrast to that, sensitivity to high frequencies is much lower. The equation for converting the standard frequency into the mel scale is:

\[ M(f) = 1125 \ln \left(1 + \frac{f}{700}\right) \]  

The time domain of the scream and speech signal \( S(t) \) is converted into sub frames that are 25 ms in duration, to obtain \( S_i(t) \), where \( i \) represents the frame index. The fast Fourier transform (FFT) is then computed for each frame, and is denoted as \( S_i(k) \):
\[
S_i(k) = \sum_{n=0}^{N} s_i(n) h(n - \frac{p^\ast kn}{N})
\] (2)

Where \(h(n)\) is the frame shape (e.g. hamming) having length \(N\). In this framework, a rectangular window is used. The periodogram power spectral estimate of \(S_i(k)\) computed from Equation (2) is then calculated as:

\[
\mathcal{P}_i(k) = \frac{1}{N} |S_i(k)|^2
\] (3)

As can be seen above, Equation (3) is multiplied by the mel-filter bank comprised of 26 triangular filters. The output now consists of 26 values, each corresponding to an individual filter bank. This method utilizes only 13 coefficients out of these 26 as feature vectors. The log square of these 26 values is passed through the discrete cosine transform (DCT) to generate the final features vector. Figure 4 describes the overall method used to create the features vector.

All the database sounds of scream and speech are passed through the MFCC algorithm, and a database of features called a ‘feature set’ is constructed. The data is pre-labelled as speech and scream sounds. Furthermore, the proposed method utilizes 13 MFCC coefficients.

4. Support Vector Machine (SVM) Classifier

The proposed method uses a supervised, statistical machine learning based technique: SVM, based on binary linear classification. Further, SVM was proposed by [9] and has been extensively used to solve certain practical problems related to detection and classification. This study uses the SVM classifier to generate a multiclass recognition scenario driven by MFCC coefficients. The system is trained using SVM, and it is expected that the test data will meet the requirement mentioned in Equation (4), where the complexity of the model is described by using a fitted model. The error that occurred during test phase should be less than the sum of the complexity and the training error:

\[
\text{Test Error} \leq \text{Training Error} + \text{Complexity of Models}
\] (4)

In this study, more than 50 audio events were used for classification, including speech and scream sounds. To verify the robustness, an SVM classifier was tested with and without noise.

5. Summary

A method to classify human sounds, driven by MFCC feature extraction, is proposed. To demonstrate the algorithm, the classification of scream and speech is implemented. The study considers scream detection and classification as a pattern recognition (machine learning) problem, and not as a traditional signal processing problem. Moreover, the acoustic parameters of different scream sounds contain patterns that can be further investigated.
REFERENCES


