MCU-Based Isolated Appealing Words Detecting Method with AI Techniques

Liang Ye[1,2,3(✉)], Yue Li[3,4], Wenjing Dong[1], Tapio Seppänen[5], and Esko Alasaarela[2]

[1] Department of Information and Communication Engineering, Harbin Institute of Technology, Harbin 150080, China
yeliang@hit.edu.cn
[2] Health and Wellness Measurement Research Group, OPEM Unit, University of Oulu, 90014 Oulu, Finland
[3] Key Laboratory of Police Wireless Digital Communication, Ministry of Public Security, Harbin 150080, China
[4] Electrical Engineering School, Heilongjiang University, Harbin 150080, China
[5] Physiological Signal Analysis Team, University of Oulu, 90014 Oulu, Finland

Abstract. Bullying in campus has attracted more and more attention in recent years. By analyzing typical campus bullying events, it can be found that the victims often use the words "help" and some other appealing or begging words, that is to say, by using the artificial intelligence of speech recognition, we can find the occurrence of campus bullying events in time, and take measures to avoid further harm. The main purpose of this study is to help the guardians discover the occurrence of campus bullying in time by real-time monitoring of the keywords of campus bullying, and take corresponding measures in the first time to minimize the harm of campus bullying. On the basis of Sunplus MCU and speech recognition technology, by using the MFCC acoustic features and an efficient DTW classifier, we were able to realize the detection of common vocabulary of campus bullying for the specific human voice. After repeated experiments, and finally combining the voice signal processing functions of Sunplus MCU, the recognition procedure of specific isolated words was completed. On the basis of realizing the isolated word detection of specific human voice, we got an average accuracy of 99% of appealing words for the dedicated speaker and the misrecognition rate of other words and other speakers was very low.

Keywords: Appealing words detection · Speech recognition · MCU · AI

1 Introduction

In recent years, more and more bullying events have been reported in middle school campus or primary school campus. Slight campus bullying includes curse and push, and serious campus bullying includes beat and abuse. A survey in China reported that, over 40% of students had suffered from various campus bullying [1], which showed a different campus life from people's mind in which campus should be a safe place. In USA, a survey by *USA Today* reported that about 50% of the surveyed high school students had bullied others in the past one year, whereas 47% of them said that they had

S. Han et al. (Eds.): AICON 2019, LNICST 287, pp. 300–308, 2019.
https://doi.org/10.1007/978-3-030-22971-9_26

been bullied, ridiculed, or mocked. 44% of the surveyed boys and 50% of the girls said that they had ever been victims of campus bullying [2]. Obviously, campus bullying has become a very common and serious problem in all societies. Campus bullying seriously endangers the normal study and life of the victims, and more seriously, it will affect the establishment of their world outlook and outlook on life in the growing stage. However, after being hurt by violence, the bullied often do not dare to give timely feedback to teachers or parents because of fear, self-esteem and other reasons. On one hand, the bullied children cannot be comforted and protected; on the other hand, the bullies are not timely educated and monitored, and eventually the phenomenon of bullying on campus becomes more and more serious.

Fortunately, campus bullying can be monitored by many indicators with artificial intelligence (AI) [3], and the acoustic characteristics of voice are one of them. With the rapid development of speech recognition technology in recent years, it has become a hot research field to monitor campus bullying events through speech keywords. Different types of campus bullying incidents are more or less accompanied by verbal bullying, and equally, in such circumstances, there are also appealing words or begging words from the victims. Therefore, through real-time monitoring of specific common words of campus bullying, teachers and parents can be informed of the occurrence of campus bullying incidents in the first time, so as to take timely measures to minimize the impact of campus bullying. In addition, due to the development of large-scale integrated circuit technology, these complex speech recognition systems can also be made into dedicated chips. Speech keyword recognition technology and embedded systems are combined, which is convenient to students who are not allowed to carry mobile phones. Hand rings, watches and other portable devices play a monitoring role for teenagers.

The remainder of this paper is organized as follows: Sect. 2 talked about some related work; Sect. 3 describes the proposed campus bullying word detecting method; Sect. 4 displays the experiment result; and Sect. 5 finally draws a conclusion.

2 Related Work

Nordic countries were the first to study campus bullying, but most of these studies were from the perspectives of pedagogy, analysis of students' psychology, and giving students the right teaching. However, many students who suffer from bullying dare not report their own experiences to their parents, so it seems that the prevention and control of campus bullying from the perspective of education alone is weak. With the popularity of smartphones, some researchers have developed campus bullying prevention programs based on smartphones, such as Stop Bullies, Campus Safety, ICE BlackBox, TipOff, Back Off Bully and so on. "Stop Bullies": When bullying occurs, the user presses a key on the mobile phone, and the mobile phone can send live video, photos or text messages along with the user's GPS information to the designated recipient. Receiver can find the user's location and take corresponding measures according to the site information to stop bullying. TipOff: Users can upload bullying or crime scene data (such as photos) recorded on their mobile phones to a secure server, and only the administrator of that server can view the evidence. Other bullying detection technologies work similarly. These methods are passive and need to be triggered manually by the

user. In the process of bullying, it is very difficult for the victim, and even invites further aggression from the perpetrator. Bystanders may be afraid of being retaliated by the perpetrator, and not dare to operate mobile phones to alarm in the process.

Therefore, there should be an automatic method to detect campus bullying in an artificial intelligence way [4], and speech recognition is a possible one. In recent years, with the development of computer technology and microelectronics technology, the research of speech recognition technology in the AI field has moved from laboratory to application field, and has achieved breakthrough results. Many enterprises in developed countries, such as BIM, APPLE, AT&T, Microsoft and other well-known companies in the United States, Japan and Korea, have done a lot of research in the field of speech recognition technology. The technology of Speaker-Independent and continuous speech recognition is becoming more and more mature, and the research of speaker-specific recognition has also made some achievements in embedded applications. China has also invested a lot of energy in speech recognition research. The Chinese Academy of Sciences, Tsinghua University, Northeast University, Beijing University of Technology, Shanghai Jiaotong University, Huazhong University of Science and Technology, and Harbin Institute of Technology are all engaged in the research and development of speech recognition.

In speech feature recognition of isolated words, linear prediction coefficient has the lowest complexity, and the combination of linear prediction coefficient and piecewise linear matching method is one of the mainstream methods of speech keyword recognition in China. Linear prediction coefficients are the basic features of speech. In order to reduce the complexity of the algorithm, we can use fixed coefficients with constant prediction coefficients for a long time. In order to improve the recognition accuracy, we can use the adaptive prediction in which each frame recalculates the prediction coefficients and predicts the average energy of the remaining signals. In addition, there are single-level prediction, multi-level prediction and other ways [5, 6]. In this paper, a fixed coefficient prediction scheme is adopted. Linear predictive coding with fixed coefficients can analyze speech signals by estimating the formant of speech signals, eliminating the role of formant in speech signals, and estimating the strength and frequency of retained speech signals. In the training process, the digital signals describing the strength and frequency of speech signals, common peaks and residual signals are stored in a Microcontroller Unit (MCU) and ready to be called at any time.

In speech recognition, the piecewise linear dynamic time matching recognition method based on time series eigenvalue difference has the least computation and is suitable for short-term speech recognition. The basic idea of this method is to find out the relative quantities of phonological features (consonants, vowels, transitional tones, *etc*.) of speech signals for distance comparison. The specific method is to find out the difference of frame features according to the time sequence, and then divide the difference of phonological features by the total difference of phonological features to get the relative accumulated difference of features. In this way, although the pronunciation speed is different, the relative cumulative difference of phonological features is basically unchanged [7]. Through the analysis of speech data, it is found that although the spectrum of the end segment changes dramatically, the semantics of the end segment are few, and it has little effect on distinguishing speech. The feature of the end segment is deleted in this recognition method.

3 Isolated Word Recognition for Specific Voice

Considering the practical application scene, it is the victim who carries the bullying detecting device, so the voice of the victim is clearer than that of the bullies. Therefore, this paper focuses on the detection of appealing words and begging words by the victims.

The isolated word recognition system consists of the following steps: data sampling, data pre-processing, feature extraction, feature selection, classifier training, and classifier testing (or practical use).

3.1 Data Sampling and Pre-processing

Students in most primary schools and middle schools in China are not allowed to take mobile phones at school, so this paper considers applying the campus bullying detecting system in a MCU. This paper chooses the Sunplus for this purpose because it has the following advantages:

(1) Small size, high integration, good reliability and easy expansion.
(2) Strong interrupt handling ability. The Sunplus MCU interrupt system supports 14 interrupt vectors and more than 10 interrupt sources, which is suitable for real-time applications.
(3) ROM, static RAM and multi-functional I/O ports with high addressing capability.
(4) The instruction system with strong function and high efficiency has compact format and fast execution.
(5) Low energy consumption.

The Sumplus MCU provides a microphone input with automatic gain control (AGC). Voice data were gathered by the embedded microphone. Because voice is affected by oronasal radiation, the high frequency part of voice attenuates more seriously than the low frequency part. Therefore, pre-emphasis is essential in the pre-processing procedure. Since voice is short-term stationary stochastic processes, faming is needed to cut long-term voice into short-term segments. In order to avoid spectrum leakage, the frames should be windowed. Normally the Hamming window is used because its sidelobe attenuation is large. In a complete utterance, there are always blank segments, so voice activity detection (VAD) is used to detect the valid part in a speech in speech recognition. Usually, VAD is judged by short-time energy and zero crossing ratio (ZCR).

3.2 Acoustic Features

After pre-processing, acoustic features can be extracted from the speech. Commonly used features include pitch, fomant, Linear Predictive Cepstral Coefficient (LPCC), Mel Frequency Cepstrum Coefficient (MFCC), *etc*. MFCC has been proved to be a good set of acoustic features for speech recognition as well as speech emotion recognition.

MFCC is based on human auditory model. Mel is pitch unit. Pitch is a subjective psychological quantity, and it is the sense of human auditory system to sound frequency [8]. Through years of research on human ear auditory system, scholars have found that the sensitivity of human ear auditory system to different frequencies of

signals is different. Generally speaking, the treble is easily concealed by the bass, and vice versa. In frequency domain, the critical bandwidth of high frequency speech signal masking is larger than that of low frequency speech signal masking. Therefore, band-pass filters can be arranged densely to sparsely according to the critical bandwidth from low frequency to high frequency, and the input speech signal can be filtered. The energy of the speech signal after the filter is used as the characteristic parameter of the speech signal. Because the modified feature parameters take into account the particularity of human ear auditory system and make use of some research achievements in the field of human ear auditory system research, the feature parameters are more in line with human ear auditory characteristics. The corresponding relationship between Mel frequency and actual frequency is shown as,

$$Mel(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right) \tag{1}$$

Transform the time domain signals into frequency domain signals by FFT (Fast Fourier Transform), and calculate the energy of the frequency domain signals after passing the Mel filters as,

$$S(i,m) = \sum_{k=0}^{N-1} E(i,k) H_m(k), \qquad 0 \le m \le M \tag{2}$$

where $E(i, k)$ is the energy spectrum before the filters and $H_m(k)$ is the frequency response of the filter and calculated as,

$$H_m(k) = \begin{cases} 0 & (k < f(m-1)) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & (f(m-1) \le k \le f(m)) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & (f(m) \le k \le f(m+1)) \\ 0 & (k > f(m+1)) \end{cases} \tag{3}$$

Then MFCC can be got by DCT (Discrete Cosine Transform),

$$mfcc(i,n) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} \log[S(i,m)] \cos\left(\frac{\pi n(2m-1)}{2M}\right) \tag{4}$$

3.3 Classifier Design

There are normally many kinds of classifiers that can be used for speech recognition. Considering that the classification is to be applied on a MCU in which the resources are limited, the authors chose an efficient DTW (Dynamic Time Warp) classifier with small computational cost.

Different voice signals produced by different people have different modes. Even the same person will produce different speeds and other changes in speech characteristic

parameters at different times due to different methods of voice production. Thus, speeches can be recognized by means of template matching.

Assume that the test and reference templates are represented by T and R, respectively. The smaller the distance $D[T, R]$ between them is, the higher the similarity is. In order to calculate the distortion distance, the distance between the corresponding frames in T and R should be calculated. Let n and m be arbitrarily selected frame numbers in T and R, respectively. $D[T(n), R(m)]$ denotes the distance between the two frame feature vectors. Distance function is executed by the distance measure actually adopted, and Euclidean distance is usually used in DWT algorithm. If $N = M$, it can be calculated directly. Otherwise, $T(n)$ should be aligned with $R(m)$. Alignment is mainly based on dynamic time warping. Each frame number $n = 1{:}N$ of the test template is marked on the horizontal axis in a two-dimensional rectangular coordinate system, and the frame number $m = 1{:}M$ of the reference template is marked on the vertical axis. By drawing some vertical and horizontal lines from the integer coordinates representing the frame number, a grid can be formed. Each intersection point (n, m) in the grid represents the intersection point between a frame in the test mode and a frame in the training mode.

The DTW algorithm is to find a path through a number of grid points in this grid, which is the frame number of distance calculation in the test and reference templates. The path is not optional. First of all, the pronunciation speed of any kind of voice may change, but the order of each part cannot change. Therefore, the chosen path must start from the lower left corner and end at the upper right corner.

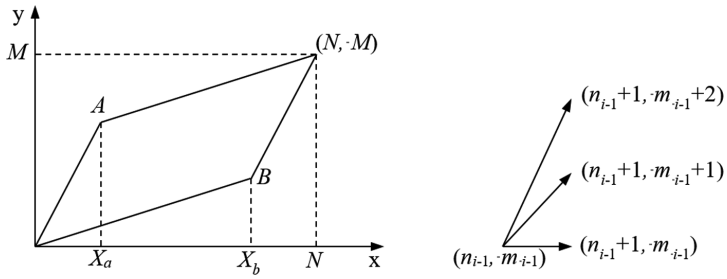

Fig. 1. Efficient DTW searching path

However, because the slope of bending is limited in the matching process, many lattices cannot actually be reached. Therefore, it is not necessary to calculate the matching distance of the lattice points outside the diamond. In addition, it is not necessary to save all the frame matching distance matrices and cumulative distance matrices, because only three grids of the previous column are used for matching calculation at each grid point in each column, making full use of these two characteristics can reduce the computational load and storage space requirements. This procedure is given in Fig. 1.

If the actual dynamic bending is divided into three sections, namely $(1, X_a)$, $(X_a+ 1, X_b)$, (X_b, N), where $X_a = 1/(6M - 3N))$, and $X_b = 2/(6N - 3M)$, one can get the restrictive conditions as,

$$2M - N \geq 3 \tag{5}$$

$$2N - M \geq 3 \tag{6}$$

When each frame on the x-axis no longer needs to be compared with each frame on the y-axis but only with the frames on the y-axis $(y_{\min}, y_{\max})$, the calculations of $y_{\min}$ and $y_{\max}$ are given as follows,

$$y_{\min} = \begin{cases} 0.5x, & 0 \leq x \leq X_b \\ 2x + (M - 2N), & X_b \leq x \leq N \end{cases} \tag{7}$$

$$y_{\max} = \begin{cases} 2x, & 0 \leq x \leq X_a \\ \frac{1}{2x} + (M - \frac{1}{2N}), & X_a \leq x \leq N \end{cases} \tag{8}$$

4 Experiments and Results

The algorithms mentioned above were implemented on a Sunplus MCU, and the system flow chart is given in Fig. 2.

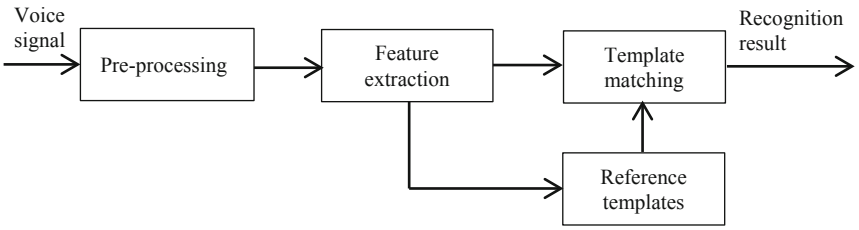

Fig. 2. Speech recognition system on a Sunplus MCU

The Sunplus MCU provided a microphone with 8kNz sampling rate. The voice signal then passed a low-pass filter and A/D converter. Pre-emphasis was used to enhance the high frequency part. The components of MFCC extraction and DTW classifier were programed on the MCU.

Totally 4 girls and 2 boys were invited to say the appealing or begging words, and each of them said 1000 words, including 500 appealing or begging words and 500 other words. The recognition result is given in Table 1. The aim of this experiment was to recognize the appealing or begging words (known as keywords in Table 1) of the dedicated speaker (Girl 1 in this experiment) but ignore other words of that speaker or appealing or begging words of the other speakers.

Table 1. Recognition results of different speakers

Speaker	Keywords	Recognized	Non-keywords	Recognized
Girl 1	500	497	500	0
Girl 2	500	3	500	0
Girl 3	500	5	500	1
Girl 4	500	0	500	0
Boy 1	500	0	500	0
Boy 2	500	0	500	0

It can be seen from Table 1 that for the dedicated speaker, the keywords can be recognized with an average accuracy of 99.4%, none of the non-keywords were misrecognized. For the other speakers, only 9 out of 5000 words were misrecognized. It shows that the implemented system can recognize the appealing or begging words of the dedicated speaker with a high accuracy.

5 Conclusions

This paper implemented an isolated appealing words detecting method on a Sunplus MCU. The voice of the speakers was gathered by a microphone embedded on the MCU and passed a low-pass filter and A/D converter. Then pre-processing was performed to enhance the speech signal and MFCC features were extracted. An efficient DTW algorithm acted as the classifier. In the experiments, the keywords of the dedicated speaker could be recognized with an average accuracy of 99.4%, while only very few of the other words of the dedicated speaker and speeches of the other speaker were misrecognized.

Acknowledgement. This work was supported by the National Natural Science Foundation of China under Grant No. 61602127, the Basic scientific research project of Heilongjiang Province under Grant No. KJCXZD201704, and the Key Laboratory of Police Wireless Digital Communication, Ministry of Public Security under Grant No. 2018JYWXTX01. The authors would like to thank those people who have helped with these experiments.

References

1. Dawei, W., Hongmei, Y.: The alarming sound of campus alarm bells. People's Public Secur. **10**, 19–21 (2004)
2. Sharon, J.: Bullying survey: most teens have hit someone out of anger. USA Today **26** (2010)
3. Lu, W., Gong, Y., Liu, X., et al.: Collaborative energy and information transfer in green wireless sensor networks for smart cities. IEEE Trans. Industr. Inf. **14**(4), 1585–1593 (2017)
4. Ye, L., Wang, P., Wang, L., et al.: A combined motion-audio school bullying detection algorithm. Int. J. Pattern Recogn. Artif. Intell. **32**(12), 1850046 (2018)
5. Liu, J., Zhang, W.: Research progress on key technologies of low resource speech recognition. J. Data Acquisit. Process. **32**(2), 205–220 (2017)

6. Zhang, P., Ji, Z., Hou, W., et al.: Design and optimization of a low resource speech recognition system. J. Tsinghua Univ. (Sci. Technol.) **57**(2), 147–152 (2017)
7. Wang, J., Zhang, J., Lu, W., et al.: Automatic speech recognition with robot noise. J. Tsinghua Univ. (Sci. Technol.) **57**(2), 153–157 (2017)
8. Remesa, U., Lópeza, A.R., Juvela, L., et al.: Comparing human and automatic speech recognition in a perceptual restoration experiment. Comput. Speech Lang. **34**, 450–457 (2016)