Using Speech Emotion Recognition to Preclude Campus Bullying

Jianting Guo(✉) and Haiyan Yu

Harbin Institute of Technology at Weihai, Weihai, Shangdong, China
guojianting0616@163.com

Abstract. Campus bullying could have extremely adverse impact on pupils, leading to physical harm, mental disease, or even ultra behaviour like suicide. Hence, an accurate and efficient anti-bullying approach is badly needed. A campus bullying detection system based on speech emotion recognition is proposed in this paper to distinguish bullying situations from non-bullying situations. Initially, a Finland emotional speech database is divided into two parts, namely training-data and testing-data, from which MFCC (Mel Frequency Cepstrum Coefficient) parameters are garnered. Subsequently, ReliefF feature selection algorithm is applied to select the useful features to form a matrix. Then its dimensions is diminished with PCA (Principle Component Analysis) algorithm. Finally, KNN (K-Nearest Neighbor) algorithm is utilized to train the model. The final simulations show a recognition rate of 80.25%, verifying that this model is able to provide a useful tool for bullying detection.

Keywords: MFCC · PCA · KNN · Speech emotion recognition · Campus bullying

1 Introduction

It rarely surprises us that campus bullying has been a universal topic for the incredibly baneful effort it could bring to adolescence. However, most current anti-bullying approaches are either primitive (e.g. security patrol, surveillance cameras) or sketchy (e.g. ICE BlackBox). The main problem for the former is paucity of immediacy, while the latter requires behaviours which could be detected by the bullies. Although recently technics using movement sensors have been attached to this field. For example, Ye *et al.* [1] proposed a instance-based physical violence detection algorithm to prevent physical bullying, a great many bullying behaviours still manage to escape the detection.

The reason is simple. Neglecting the variety of bullying forms, most of these methodologies focus on physical bullying. Nevertheless, according to Olweus *et al.* [2], bullying action could be carried out by physical contact, by words and by other ways. Thus those speech bullying actions without any form of physical contact could easily escape most detections. Given the contemporary situation, this paper proposes a mechanism based on speech emotion recognition to

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 728–734, 2019.
https://doi.org/10.1007/978-3-030-32388-2_59

discover the emotions contained by speeches under both bullying situation and non-bullying situation and distinguish them from each other. Every smart phone with a microphone should be able to use this mechanism.

Additionally, Salmivalli *et al.* [3] claimed that apart from victims and bullies, other roles like outsiders are also involved in a bullying situation. Some of those outsiders could be too scared to stop the bullying or call for help. With this speech emotion recognition mechanism, they would be able to report the bullying situation without infuriating the bullies, which will furnish us with another immediate anti-bullying method.

The paper is constructed as follows: Sect. 2 gives out some previous researches on speech emotion recognition. The emotion-recognising-based bullying detection algorithm is proposed in Sect. 3. Section 4 shows the simulation results, while Sect. 5 draws conclusions.

2 Previous Research on Speech Emotion Recognition

Recently, emotion recognition in speech has been an extremely voguish field, attracting tremendous amounts of researchers to modify this technique. Thanks for the researches done by the previous researchers, a variety of classifiers using different speech features (e.g. frequency, energy, speaking rate...) and different databases (e.g. Berlin Emotional Database, BelFast Database...) appear one after another. Some of those classifiers are capable of distinguish 5–7 kinds of human emotion. Although in this paper our ultimate goal is to differentiate bullying situation and non-bullying situation, which means the ability to specify every kind of emotion is not necessary, it is still helpful to review a couple of spectacular classifiers.

Some of the commonest classifiers for speech emotion recognition are SVM (Support Vector Machine), HMM (Hidden Markov Model), K Star and so on. Iliou *et al.* [4] extracted features like pitch, energy and MFCCs (mel-frequency cepstral coefficients) and attached K Star classifier to distinguish 7 different emotions based on these features. Their final accuracy in speaker-independent framework reached 74%. Petrushin [5], however, ultilised neural network recogniser to classify 5 dissimilar emotion at distinct rate. His classifier's recognition ability (namely the accuracy) fluctuates depending on different emotions (e.g. 70–80%-anger, 35–55%-fear), which is very similar to humans' own capability of distinguishing emotions. To obtain a relatively high accuracy and specific recognition ability, these classifiers inevitably have a high sophistication. Besides, they also consume more hardware resources.

Since our purpose is to distinguish bullying situation and non-bullying situation based on emotion recognition, which does not demand for high specification, we will attempt to diminish the complexity with less features and less labels. Bicocchi *et al.* [6] pointed out that KNN algorithm could achieve a similar recognition accuracy with much less calculation, therefore, this paper will focus on KNN classifier.

3 Bullying Detection Algorithm Based on Speech Emotion Recognition

The database that we use contains voices from different campus scenarios performed by a school of Finnish students. These voices could be divided into 6 kinds—bullying voice, normal conversation, clap hands, laugh, cry and voice that shows fear. Given that we aim at distinguishing bullying and non-bullying situation, we categorize these different voices into two groups (e.g bulling situation: bullying voice, cry, voice that shows fear, non-bullying situation: normal conversation, clap hands, laugh.), after which we can get a training set and a testing set each containing approximately same number of emotional voices from both situations (training set: 42 voices from bulling situation and 41 voices from non-bullying situation. testing set: 41 voices from bullying situation and 41 voices from non-bullying situation). Similar quantity of training voices from both sides could efficiently prevent the classifier from having a bias towards one side.

According to Kim *et al.* [7], MFCC (Mel Frequency Cepstral Coefficient) is the most commonly used feature in distinguishing speeches, emotions, speakers and so on. Hence, the speech feature extraction in this paper will mainly focus on MFCC.

Initially, to improve the resolution of high-frequency part of the voice signal, we use first order FIR filter to do pre-emphasis to the signal's spectrum. The function of the first order FIR filter is:

$$H(z) = 1 - 0.9375z^{-1}. \tag{1}$$

Figure 1 shows the spectrum of a voice signal before pre-emphasis, while Fig. 2 demonstrates the spectrum of the same voice signal after pre-emphasis. As illustrated in the figures, the interval between two peaks in the spectrum after pre-emphasis is more obvious than that in the spectrum before pre-emphasis, which shows a relatively higher resolution.

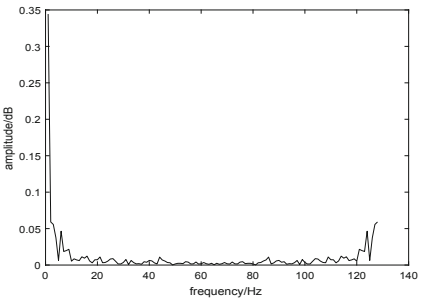

Fig. 1. Spectrum before pre-emphasis.

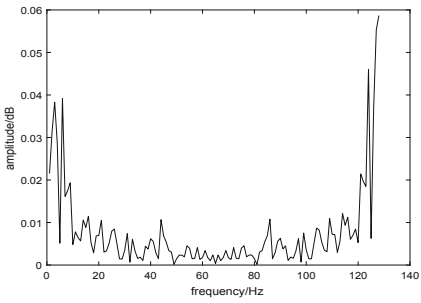

Fig. 2. Spectrum after pre-emphasis.

Subsequently, for the reason that voice signal is thought to be steady in 10 ms to 30 ms, we separate frame of all these voice signals with a Hamming window showed as following:

$$W(n,a) = (1-a) - cos[2\pi * n/(N-1)]. \tag{2}$$

where $a = 0.46$ and $N = 512$. To avoid data lost, the frame increment is set to be 256 which is half the length of Hamming window. Then MFCC parameters of every frame are calculated, after which first and second order differential MFCC parameters will also be reckoned. After the extraction of MFCC parameters, we calculate the arithmetic means of MFCCs and use them to form a matrix containing enormous characters of these voice signals. The formula used to compute the mean of MFCCs is:

$$x_{ave} = (x_1 + x_2 + ... + x_n)/n. \tag{3}$$

Considering that dealing with such an avalanche of features is sophisticated and unnecessary, we need to designate the useful ones from them. Thus, ReliefF feature selection algorithm is used to select features which is useful for the classifier. The ranks and weights of every feature is computed and only those features whose weight is greater than zero are kept while others are neglected. After the feature selection progress is done, 22 features are remained for each of the 83 samples in training set and 81 samples in testing set. Although the number of features is relatively small, they could still cost a lot of resources and be unnecessary. Consequently, PCA (Principle Component Analysis) is attached to diminish the dimension of features. The remaining 22 features of 83 elements in training set are used to form a matrix X with 83 rows and 22 columns. Then we do zero-mean to every row of matrix X, after which the covariance matrix C of X is calculated. Next, we find the eigenvalues of the covariance matrix C and the corresponding eigenvector r. Finally, the feature vectors are arranged in a matrix from top to bottom according to the corresponding feature value, and the first k rows are formed into a matrix P, which is the data after dimension reduction to k dimensions. The error and the percentage of information those features could deliver are computed using variances of features. We deem that keeping the features that could display 95% of primitive information would be adequate in this situation. Table 1 illustrates the relationship between the number of features and percentage of information they can deliver.

Table 1. Relationship between the number of features and the percentage of information they can deliver

The number of features	1	2	3	4	5	6	7
The percentage of information	47.45%	66.43%	79.50%	88.17%	93.76%	97.96%	99.09%

As can be seen from the table above, 6 features will be able to include 97.96% of the whole information in a signal voice. However, in order to make the classifier more accurate, we decide to keep 7 features, which could display 99.09% of all the information. Finally, we choose the KNN (K Nearest Neighbour) classifier to do the classification job due to its high accuracy and low complexity.

4 Classification and Analysis

According to Witten *et al.* [8], the class of testing set is predicted based on the nearest training instance in an instance-based learning situation. The KNN classifier we use in this article is also an instance-based classifier. Therefore, its classification result is mainly decided by how many kinds of labels the training set have and how many neighbours we set. Given that our mission is to distinguish two situations—bullying situation and non-bullying situation, we give 1 and 2 as two labels to the training set.

- 1 represents non-bullying situation.
- 2 represents bullying situation.

As for the value of K, we tried an array of different numbers and their corresponding accuracy with testing set varies. The accuracy calculation process is as follows.

Set $CTS(ClassifiedTestingSet)$ to be a vector that contains the classification result of testing set and $RTS(RealTestingSet)$ to be a vector containing the real value of testing set. In both vectors, 1 stands for non-bullying situation while 2 represents bullying situation. Considering that the two vectors have the same dimension, we can get their difference as:

$$D = CTS - RTS. \tag{4}$$

Set x to be the number of zeros in the D (Difference) vector, and y to be the length of the D vector. Then we have the formula for calculating the accuracy:

$$Accuracy = x/y. \tag{5}$$

Table 2 demonstrates some representative K values and their corresponding accuracy.

Table 2. Several representative K values and their corresponding accuracy

K	5	11	17	23	25	27
Accuracy	70.37%	76.54%	79.01%	79.01%	80.25%	79.01%

An increasing trend could be witnessed in accuracy with the rising of K value and the accuracy reaches its peak (80.25%) when K is 25, after which it begins to drop. So we set the ideal K value to be 25. The final accuracy is analogous to some of the recent speech emotion recognition algorithm. For example, Likitha *et al.* [9] also chose MFCC as the extracted feature in their paper and reached an efficiency of 80% for happy, sad and anger emotions. Nevertheless with ReliefF feature selection algorithm and PCA (Principle Component Analysis) algorithm which are not used in their algorithm, our method is able to gain a similar accuracy with less characters and less calculations.

Among all the incorrect classifications, most of them are from non-bullying situation, which means the classifier has a spectacular ability to detect bullying situation but its rate of misclassifying non-bullying situation is slightly high. The reason is not sophisticated. Bullying situation often includes shouting, crying, threatening and other fierce voices which are not commonly heard in normal situation. Their high specificity and low diversity make them relatively strong features, while normal situation often contains enormous different speech emotion—happy, excited, disappointed and so on, which makes features of normal situation weaker. Besides, some special situations in non-bullying situation like intense controversy contain some features that are very similar to features of bullying situation. Therefore, the classifier could sometimes mistake non-bullying situation for bullying situation.

5 Conclusion and Discussion

Campus bullying is universally acknowledged to be deleterious to students and most existing anti-bullying methods tend to focus on physical bullying and victims. In this paper, a instance-based speech emotion recognition algorithm is proposed to make great use of bystanders and detect language bullying which is easily neglected.

The database used in this paper consists of voice signals from different campus situations performed by a group of Finnish students. By analysing these voice signals, MFCC (Mel Frequency Cepstrum Coefficient) is extracted, after which ReliefF feature selection algorithm is attached to diminish the dimension of feature vector to 7 for reducing complexity. Then a two-label training set with approximately same number of factors from both bullying situation and non-bullying situation is used to train the KNN classifier and the classifier successfully reaches an accuracy of 80.25% with the testing sample. Considering the relatively high recognition accuracy, this approach can be attached to smart phones with microphones to detect and report campus bullying efficiently.

Aiming at developing a resource-friendly bullying detection mechanism, this paper utilizes only a limited number of features extracted from voice signals and 2 labels, which leads to a slightly high misclassification rate in non-bullying situation. In the future work, the author will focus on improving the specificity of classification using more emotional features in voice signals like pitch which is a character containing enormous information about emotional status.

References

1. Ye, L., Ferdinando, H., Seppanen, T., et al.: An instance-based physical violence detection algorithm for school bullying prevention. In: 2015 International Wireless Communications and Mobile Computing Conference (IWCMC). IEEE (2015)
2. Olweus, D.: Bullying At School: What We Know and What We Can Do. Wiley-Blackwell, New York (1993)

3. Salmivalli, C., Lagerspetz, K., Bjorkqvist, K., et al.: Bullying as a group process: particiant roles and their relations to social status within the group. Aggressive Behav. **1996**(22), 1–15 (2016)
4. Iliou, T., Paschalidis, G.: Using an automated speech emotion recognition technique to explore the impact of bullying on pupils social life. In: 2011 15th Panhellenic Conference on Informatics, Kastonia, pp. 18–22 (2011)
5. Petrushin, V.: Emotion recognition in speech signal: experimental study, development, and application. In: Proceedings of Sixth International Conference on Spoken Language Processing (ICSLP 2000), ISCA, vol. 2, pp. 222–225, October 2000
6. Bicocchi, N., Mamei, M., Zambonelli, F.: Detecting activities from body-worn accelerometers via instance-based algorithms. Pervasive Mobile Comput. J. **6**, 482–495 (2010)
7. Kim, S., Georgiou, P.G., Lee, S., Narayanan, S.: Real-time emotion detection system using speech: multi-modal fusion of different timescale features. In: IEEE 9th Workshop on Multimedia Signal Processing, Crete 2007, pp. 48–51 (2007)
8. Witten, I.H., Frank, E., Hall, M.A., et al.: Data mining: practical machine learning tools and techniques. ACM Sigmod Record **31**(1), 76–77 (2011)
9. Likitha, M.S., Gupta, S.R.R., Hasitha, K., Raju, A.U.: Speech based human emotion recognition using MFCC. In: 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, pp. 2257–2260 (2017)