A Comparison of KNN Based Classifiers for Detecting Emotion from Mandarin Speech

以最近鄰居分類法為基礎的分類器在中文語音情緒辨識表現之比較

Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, and Yuan-Hao Chang
Tatung University （大同大學資訊工程研究所）
tlpao@ttu.edu.tw, d8906005@ms2.ttu.edu.tw, d9306002@ms2.ttu.edu.tw

Abstract

Humans communicate through speech, movement, hand gestures and facial expressions. We express our emotions in speech by the words that we use and intonation of the voice. Whereas research about automated recognition of emotions in facial expressions is now very rich, research dealing with the speech modality has only been active for very few years and is almost for English. In this paper, we presented a comparison of three KNN based classification algorithms for detecting emotion from Mandarin speech. The results show that the proposed weighted D-KNN outperforms the other two classification techniques: 13.1% improvement for traditional KNN and 7.4% improvement for M-KNN. The highest recognition rate (79.31%) is obtained with weighted D-KNN using Fibonacci series.

Keywords：KNN, Emotion Detection, Weighted D-KNN

摘要

人類透過說話、動作、手勢及臉部表情等方式來進行溝通。在說話方式上，我們可以藉由說話的用字遣詞及聲音的語調變化來表現出說話當時的情緒。有鑑於自動臉部表情情緒辨識的研究已經非常豐富，而自動語音情緒辨識的研究近幾年才比較活躍，不過卻幾乎都以英文為主。在這篇論文中，我們比較三個以最近鄰居分類法為基礎的演算法在中文語音情緒辨識上的表現。實驗結果呈現出我們所提出的權重式最近鄰居分類法，勝過其他兩種一樣以最近鄰居分類法為基礎的演算法：較傳統式的最近鄰居分類法改善了 13.1%，而較改良式最近鄰居分類法改善了 7.4%。在使用以費氏級數為權重序列的權重式最近鄰居分類法時，可得到最好的辨識率(79.31%)。

關鍵詞：最近鄰居分類法，情緒辨識，權重式 D-KNN

1、Introduction

Humans interact with one another in several ways such as speech, eye contact, body language, and so on. Among them, speech communication is the most common in human-to-human interaction. Speech signal is a rich source of information and convey more than spoken words. The additional information conveyed in speech includes gender information, age, accent, speaker's

identity, health, prosody and emotion [5]. For example, a listener can recognize different emotion of a speaker if the latter speaks the same sentences in different mood.

Recognizing emotions from speech has gained increased attention recently and can be applied to develop many applications. Emotion recognition classifies speech into categories, which are related to the psychological state of the user. The usual emotion categories are anger, fear, sad, happy, disgust and surprise. The term “basic emotions” is widely used without implying that these emotions can be mixed to produce others [4].

Possible applications of recognizing emotions from speech include emotion recognition game, emotion recognition software for call center and robots [8]. Emotion recognition game allows a user to compete against the computer or another person to see who can better recognize emotion in recorded speech. One potential practical application of the game is to help autistic people in developing better emotional skills at recognizing emotion in speech. Emotion recognition software for call center has the ability to detect the emotional state in telephone conversions. If anger is recognized from speech signal, the system sets higher priority to the voice message and assigns the proper person to response the message. Robots can do a lot of work but they just do simple actions for what a man order. If we want a robot to be a more appropriate supporter, the robot must have the ability to understand its owner’s thought or emotion. Besides, other applications may include interactive movie, intelligent toys, situated computer-assisted speech training system and supported medical instruments.

This paper is organized as follows. In Section 2, some related works are described. In Section 3, we will describe emotion recognition system with KNN based classifiers in more detail. Experimental results are reported in Section 4. Finally, conclusions are given in Section 5.

2、Related Works

The speech signal contains different kind of information. From the automatic emotion recognition task point of view, it is useful to think about speech signal as a sequence of features that characterize both the emotions as well as the speech. In this section, some related works are described.

2.1 Emotional Speech Corpus

The performance of an emotion classifier relies heavily on the quality of emotional speech data and the similarity of it to real world samples. As mentioned in [6], there are three different categories of emotional speech: acted speech, elicited speech, and spontaneous speech.

In acted speech recording, actors are invited to record utterances, where each utterance needs to be spoken with multiple emotions. The method is adopted by most researches because it can get large amount of data in a short time and the data is undistorted. For general use, we should invite speakers with different age, gender, even with different social or educational background if possible. And if we hope the emotion in the data to be more obvious, we could invite professional actors.

In elicited speech recording, the Wizard-of-Oz (WOZ) is used. The WOZ means using a program that interacts with the speaker and drives him into a specific emotion situation and then records his voice. This method needs a good program that can induce the participator to say something in our expected emotion state. So how to design such a program may not be easy.

In spontaneous speech recording, the real-world utterances that express emotions are recorded. Although data got from this method has the best naturalness, it is the most difficult because we need to follow the speaker. When he or she is in some emotion state, his voice is recorded immediately. This method will face many problems. For examples, we must hide our recording

device in order to make the speaker without any pressure to present his real emotion. Furthermore, we also cannot assure the environment is quiet. Generally speaking, the method is generally infeasible.

2.2 Classifiers

The problem of detecting emotion can be formulated as that of assigning an emotion category to each utterance. Two main types of information sources can be used to identify the speaker's emotion: the word content of the utterance and acoustic features such as variance of the first formant. In the following, we describe the various classification methods that have been taken into consideration for emotional speech recognition [3].

The simplest classification algorithm is K Nearest Neighbor (KNN) and it is based on the assumption that the examples residing closer in the instance space have same class values. Thus, while classifying an unclassified example, the effects of the k nearest neighbors of the example are considered. It yields accurate results in most of the cases.

Gaussian Mixture Models (GMMs) provide a good approximation of the originally observed feature probability density functions by a mixture of weighted Gaussians. The mixture coefficients were computed by use of an Expectation Maximization algorithm. Each emotion is modeled in one GMM. The decision is made for the maximum likelihood model.

Neural Nets are a standard procedure in pattern classification. They are renowned for their non-linear transfer functions, their self-contained feature weighting capabilities and discriminative training. Considering the sparsely available emotion training material their good performance on small training sets compared to GMMs seems advantageous.

A great interest in Support Vector Machines (SVM) in classification can be observed recently. They tend to show a high generalization capability due to their structural risk minimization oriented training. Non-linear problems are solved by a trans-formation of the input feature vectors into a generally higher dimensional feature space by a mapping function where linear separation is possible. Maximum discrimination is obtained by an optimal placement of the separation plane between the borders of two classes. The plane is spanned by the support vectors leading to a reduction of references. A number of approaches to solve multi-class problems exist.

3、Emotion Recognition System with KNN based Classifier

Figure 1 shows the block diagram of the KNN based emotion recognition system. To do the emotion recognition we have to create and evaluate a corpus of emotional

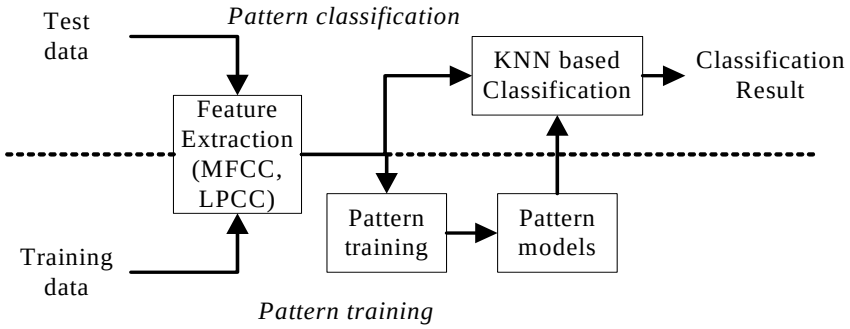

Figure 1. Block diagram of KNN based emotion recognition system

data. In our research, five emotions are investigated: anger, happiness, sadness, boredom, and neutral. We invite 18 males and 16 females to simulate five emotions. A prompting text with 20 different sentences is designed. The length of each sentence is from one word to six words. The sentences are meaningful so speakers could easily simulate them with emotions. During the recordings process, speakers are asked to try their best to simulate each emotion. And speakers can simulate one sentence many times until they are satisfied what they simulated. Finally, we obtained 3,400 emotional speech sentences.

It is very important to use speech with unambiguous emotional content for further analysis. This can be guaranteed by a listening test [2], in which listeners evaluate the emotional content of a recorded sentence. Moreover, we can understand the performance of human in emotion recognition.

Table 1 shows the human performance confusion matrix. The rows and the columns represent simulated and evaluated categories, respectively. For example, first row says that 89.56% of utterances that were portrayed as angry were evaluated as angry, 4.29% as happy, 0.88% as sad, 0.77% as bored, 3.25% as neutral, and 0.99% if none of above.

For further analysis, we only need the speech data that can be recognized by most human. So we divide speech data into different dataset by their recognition accuracy. We will refer to these data sets as D80, D90, D100, which stand for recognition accuracy of at least 80%, 90%, and 100%, respectively, as listed in Table 2.

A critical problem of all recognition systems is the selection of the feature set to use. Various features relating to pitch, to energy, to durations, to tunes, to spectral, to intensity… have been studied. In our previous experiment, we estimated the following: formants (F1, F2 and F3), Linear Predictive Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), first derivative of MFCC (dMFCC), second derivative of MFCC (ddMFCC), Log Frequency Power Coefficients (LFPC), Perceptual Linear Prediction (PLP) and RelAtive SpecTrAl PLP (Rasta-PLP). Due to the highly redundant information, a forward feature selection (FFS) or backward feature selection (BFS) should be carried out to extract only the most representative features. In FFS, LPCC is the most representative

Table 1. Human Performance Confusion Matrix (%)

	Angry	Happy	Sad	Bored	Neutral	Others
Angry	89.56	4.29	0.88	0.77	3.52	0.99
Happy	6.67	73.22	3.28	2.36	13.56	0.92
Sad	2.94	1.00	82.76	9.29	3.29	0.71
Bored	1.26	0.44	8.62	75.16	13.65	0.88
Neutral	1.69	0.91	1.56	12.27	83.51	0.06

Table2. Datasets

Data set	D80	D90	D100
Size (number of sentences)	570	473	283

feature. In BFS, MFCC is the most representative feature. Finally, we combine MFCC and LPCC as the feature set used in emotion recognition system.

When we get the features from the training data and the test data, we can calculate the distance between them to classify the test data. There are various techniques for classification such as KNN, GMM, Neural Network and SVM. In our system, the classification was performed using KNN based classifiers. They are traditional KNN, Modified-KNN (M-KNN) and proposed weighted D-KNN.

Being simple, elegant and straightforward, many researchers often adopt KNN as a classifier for their applications today [1]. It is an intuitive method that classifies unlabeled data based on their similarities with data in the training set. When a new sample data *x* arrives, KNN finds the *k* neighbors nearest to the unlabeled data from the training space based on some suitable distance measure. In our case, the Euclidean distance is used. Now let the *k* prototypes nearest to *x* be $N_k(x)$ and *c*(*z*) be the class label of *z*. Then the subset of nearest neighbors within class $j \in \{1,..., \text{number of classes } l\}$ is

$$N_k^j(x) = \{y \in N_k(x) : c(y) = j\} \quad (1)$$

Finally, the classification result $j^* \in \{1,...,l\}$ is defined as a majority vote:

$$j^* = \arg\max_{j=1,...,l} |N_k^j(x)| \quad (2)$$

Modified-KNN is a technique based on the KNN [7]. It calculates the *k* nearest neighbor's distances d^i to the new sample data in each class. The classification result $j^* \in \{1,...,l\}$ is obtained by the following equation.

$$j^* = \arg\min_{j=1,...,l} \sum_{i=1}^{k} d_j^i \quad (3)$$

In this paper, we propose a weighted D-KNN to improve the performance of M-KNN. The purpose of weighting is to find a vector of real-valued weights that would optimize classification accuracy of some classification or recognition system by assigning low weights to less relevant features that provide little information for classification and higher weights to features that provide more reliable information. In M-KNN case, among the k nearest neighbors in each class, the one have the smallest distance value d^1 is the most important. The classification result $j^* \in \{1,...,l\}$ is defined as:

$$j^* = \arg\min_{j=1,...,l} \sum_{i=1}^{k} w_i d_j^i \quad (4)$$

In addition, we set a constraint $w_1 \geq w_2 \geq \cdots \geq w_k$ to conform the idea of weighting.

4、Experimental Results

The basis of the experiments is the Mandarin emotional speech corpus as presented in Section 3. In this paper, the D80 dataset is used and contains 570 utterances recorded in 8-bit PCM with a sampling frequency of 8k Hz including 151 angry, 83 bored, 96 happy, 116 neutral, and 124 sad utterances.

To compare the three KNN based emotion detection classifiers described in the previous section, all experiments were conducted using the MATLAB software, and all results are based on the leave-one-out (LOO) cross-validation. The extracted acoustic features include MFCC and LPCC.

Table 3 is the experimental result of the traditional KNN classifier with k=10. The average recognition rate is 68.90%. The results show that the extracted features with KNN can distinguish angry/happy from bored emotion.

Table 3. Experimental result (%) of traditional KNN (*k*=10)

	Angry	Happy	Sad	Bored	Neutral
Angry	84.106	4.6358	3.9735	0	3.3113
Happy	19.792	45.833	7.2917	0	12.5
Sad	4.8387	3.2258	66.935	4.0323	12.097
Bored	0	0	6.0241	79.518	4.8193
Neutral	0	2.5862	7.7586	10.345	68.103

Table 4. Experimental result (%) of M-KNN (*k*=10)

	Angry	Happy	Sad	Bored	Neutral
Angry	88.079	4.6358	2.649	0	4.6358
Happy	28.125	51.042	4.1667	1.0417	15.625
Sad	6.4516	3.2258	73.387	5.6452	11.29
Bored	0	0	12.048	83.133	4.8193
Neutral	0	3.4483	12.069	12.931	71.552

Table 5. Comparison of weighted D-KNN classifiers

Weighting scheme	Accuracy (%)
k→1	75.39
The power of 2	78.86
Fibonacci series	79.31

Table 6. Experimental result (%) of weighted D-KNN (*k*=10, weighting=10→1)

	Angry	Happy	Sad	Bored	Neutral
Angry	88.742	3.9735	2.649	0	4.6358
Happy	22.917	54.167	6.25	0	16.667
Sad	4.0323	1.6129	79.032	6.4516	8.871
Bored	0	0	9.6386	84.337	6.0241
Neutral	0.8620	4.3103	12.931	11.207	70.69

Table 7. Experimental result (%) of weighted D-KNN (*k*=10, weighting= power of 2)

	Angry	Happy	Sad	Bored	Neutral
Angry	90.066	5.298	1.9868	0	2.649
Happy	19.792	61.458	4.1667	0	14.583
Sad	3.2258	2.4194	82.258	2.4194	9.6774
Bored	0	2.4096	6.0241	85.542	6.0241
Neutral	0.862	6.0345	7.7586	10.345	75

Table 8. Experimental result (%) of weighted D-KNN (*k*=10, weighting=Fibonacci series)

	Angry	Happy	Sad	Bored	Neutral
Angry	90.728	4.6358	1.3245	0	3.3113
Happy	18.75	62.5	3.125	0	15.625
Sad	4.0323	2.4194	82.258	2.4194	8.871
Bored	0	1.2048	8.4337	84.337	6.0241
Neutral	0.862	5.1724	6.8966	10.345	76.724

Table 4 shows the recognition rate of the M-KNN classifier with k=10. The average recognition rate is 73.44%. M-KNN yields better results than traditional KNN; an improvement of the classification accuracy by 6.1%.

The experimental results of the weighted D-KNN with different weighting series are summarizes in Table 5. Their corresponding confusion matrices are given from Table 6 to Table 8. The results show that different weighting schemes have different ability and property. The best accuracy is obtained with Fibonacci series scheme.

5、Conclusions

Accurate detection of emotion from speech has clear benefits for the design of natural human-machine speech interfaces or for the extraction of useful information from large quantities of speech data. The task consists of assigning an emotion category to a speech utterance.

In this paper, we presented a comparison of three KNN based classification algorithms for detecting emotion from Mandarin speech. The results show that the proposed weighted D-KNN outperforms the two other classification techniques: 13.1% improvement for traditional KNN and 7.4% improvement for M-KNN. Besides, various weighting schemes result in different results. The highest recognition rate (79.31%) is obtained with weighted D-KNN using Fibonacci series. Like human performance, we can see that the most easily recognizable category is anger and the poorest recognizable category is happiness.

In the future, it is necessary to collect more acted or spontaneous speech sentences, in terms of both speakers and listeners. Furthermore, it might be useful to measure the confidence of the decision after performing classification. Based on confidence threshold, classification result might be classified as reliable or not. Unreliable tests can be for example further processed by human. Besides, how to optimize the weights in weighted D-KNN to improve the recognition rate in emotion recognition system is still a challenge for our future work.

六、Acknowledgement

The authors would like to thank the National Science Council (NSC) for financial supporting this research under NSC project NO: NSC 93-2213-E-036-023

6、References

[1] C. Thiel, Multiple Classifier Fusion Incorporating Certainty Factors, Universität Ulm, Fa-kultät für Informatik, 2004

[2] Inger Samso Engberg, Anya Varnich Hansen, “Documentation of the Danish Emotional Speech Database”, Department of Communication Technology Institute of Electronic Sys-tems, Aalborg University, Sep. 1996

[3] Izhak Shafran and Mehryar Mohri, "A comparison of classifiers for detecting emotion from speech", Proc. of IEEE Int'l Conference on Acoustic Signal and Speech Processing (ICASSP), Philadelphia, PA, Mar 19-23, 2005

[4] P. Eckman, “An argument for basic emotions”, Cognition and Emotion vol. 6, pp. 169-200, 1992.

[5] Rabiner L.R. and Juang B.H. Fundamentals of Speech Recognition. Prentice-Hall, Engle-wood Cliffs, NJ, 1993.

[6] Raquel Tato, Rocio Santos, Ralf Kompe, “Emotional Space Improves Emotion Recognition”, Man Machine Interface Lab, Advance Technology Center Struttgart, Sony International (Europe) GmbH.

[7] Tsang-Long Pao , Yu-Te Chen, Jhih-Jheng Lu, and Jun-Heng Yeh, "The

Construction and Testing of a Mandarin Emotional Speech Database", Proceeding ROCLING XVI, Sep. 2004

[8] V. Petrushin, “Emotion Recognition in Speech Signal: Experimental Study, Development, and Application”, in. Proc. of International Conference on Spoken Language Processing, ICSLP 2000