Journal of
Health & Medical Informatics

Long et al., J Health Med Informat 2017, 8:5
DOI: 10.4172/2157-7420.1000293

Research Article Open Access

Detecting Depression in Speech: A Multi-classifier System with Ensemble Pruning on Kappa-Error Diagram

Hailiang Long#, Xia Wu#, Zhenghao Guo, Jianhong Liu and Bin Hu*

College of Information Science and Technology, Beijing Normal University, Beijing, 100875, China
#These authors are contributed equally.

Abstract

Depression is a severe mental health disorder with high societal costs. Despite its high prevalence, its diagnostic rate is very low. To assist clinicians to better diagnose depression, researchers in recent years have been looking at the problem of automatic detection of depression from speech signals. In this study, a novel multi-classifier system for depression detection in speech was developed and tested. We collected speech data in different ways, and we examined the discriminative power of different speech types (such as reading, interview, picture description, and video description). Considering that different speech types may elicit different levels of cognitive effort and provide complementary information for the classification of depression, we can utilize various speech data sets to gain a better result for depression recognition. All individual learners formed a pool of classifiers, and some individual learners with a high diversity and accuracy in the pool were selected. In the process, the kappa-error diagram helped us make decisions. Finally, a multi-classifier system with a parallel topology was built, and each individual learner in this system used different speech data types and speech features. In our experiment, a sample of 74 subjects (37 depressed patients and 37 healthy controls) was tested and a leave-one-out cross-validation scheme was used. The experiment result showed that this new approach had a higher accuracy (89.19%) than that of single classifier methods (the best is 72.97%). In addition, we also found that the overall recognition rate using interview speech was higher than those employing picture description, video description, and reading speech. Furthermore, neutral speech showed better performance than positive and negative speech.

Keywords: Depression detection; Speech signal; Multi-classifier system; Kappa-Error Diagram

Introduction

Depression is a common mood disorder that is characterized by sadness, loss of interest, feelings of guilt or low self-worth, and poor concentration [1]. It can be long-lasting or recurrent, and can have a significant impact on individuals and their families, and even on society as a whole [2]. However, effective depression treatment is limited by current diagnostic methods, which depend excessively on the experience of the clinicians and on the cooperation of the patient, risking a range of objective biases [3]. Therefore, it is particularly important to look for new objective and convenient methods that assist clinicians in their diagnosis [4]. Affective sensing technology provides the potential possibility of objective depression recognition, and many researchers have aimed at the characterization of depression using physiological and behavioral signals (e.g., facial expression, body gesture, speech, eye movement, etc.) [5-7]. Among these signals, speech signals can be collected easily by a cheap, non-invasive, and portable instrument [8]. Hence, many researchers have focused on the use of speech signals for depression recognition [9,10].

Speech, which can serve as one of the most convenient means of communication, contains not only semantic information but also emotional characteristics and the state of the speaker [11]. The voice of depressed individuals reflects the perception of qualities such as monotony, slur, and less fluctuation [12]. Some researchers have supported the feasibility and validity of depression detection in speech signals [13,14], and many studies have focused on the correlation between depression and speech parameters [15,16]. Darby et al. found that listeners could perceive a change in the pitch, loudness, and speaking rate of depressed patients [17]. Hollien et al. suggested that depressed patients use different speaking patterns and highlight some potential characteristics: reduced speaking intensity, reduced pitch range, slower speech, reduced intonation, and a lack of linguistic stress [18]. Low et al. found that the first three formant frequencies and bandwidths, grouped together, had significant differences between their depressed and control patients ($p<0.05$) [19]. In addition, jitter and shimmer features have also been analyzed in the detection of depression, where jitter was found to be higher and shimmer lower [20,21]. Some papers on depression have reported a relative shift in energy from lower to higher frequency bands [16,22]. Furthermore, Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Coefficient (LPC), and Line Spectrum Pair (LSP) are all some of the most popular spectral features used in detecting depression in speech [23,24].

With the development of machine learning and affective sensing technology, many approaches on the automatic detection of depression using speech signals have been investigated lately [25,26]. A wide range of features have been explored for the automatic classification of depressed speech. Shankayi et al. collected speech data by reading emotional and scientific texts, and investigated the influences of some speech features (e.g., pitch, energy, formants, some glottal features, etc.) using a Support Vector Machine (SVM); they found that the scientific text speech worked better than the emotional speech [27]. Alghowinem et al. analyzed voiced, unvoiced, and mixed speech on interview speech data sets [28] and suggested that both mixed and unvoiced speech were useful in detecting depression. In their series of studies,

***Corresponding author:** Bin Hu, College of Information Science and Technology, Beijing Normal University, Beijing, 100875, China, Tel: +4155898; E-mail: bh@bnu.edu.cn

Received November 23, 2017; **Accepted** November 28, 2017; **Published** November 30, 2017

Citation: Long H, Wu X, Guo Z, Liu J, Hu B (2017) Detecting Depression in Speech: A Multi-classifier System with Ensemble Pruning on Kappa-Error Diagram. J Health Med Informat 8: 293. doi: 10.4172/2157-7420.1000293

these authors used some features such as energy, intensity, loudness, jitter, shimmer, HNR, MFCC, etc. In recent years, several studies have attempted to identify which speech types and emotions provide the most reliable recognition of depression. Some researchers investigated the performance of interview speech and reading speech data sets, and they verified that the overall recognition rate using spontaneous speech was higher than that using reading speech [29]. Liu et al. employed the methods of reading, interview, and picture description to establish a Chinese language database for depression recognition [30]. They tested the performance of three widely used classifiers (SVM, naïve Bayes, and random forest) on a large voice feature data set and further testified to the effectiveness of evaluating depression severity using speech [13]. Different speech types and emotions may elicit different levels of cognitive effort and various emotional effects, producing changes in speech acoustics that affect the depression recognition [31]. By considering that various speech data sets collected in different ways can provide complementary information for the classification of depression, we can utilize various speech data sets to gain a better result for depression recognition than that using only one speech data set.

In this paper, we propose a new multi-classifier system that combines different speech types and emotions. Multi-classifier systems have created considerable excitement in the machine learning community because of the potential to greatly increase the classification accuracy [32-34]. The system performs information fusion of classification decisions, overcoming the limitations of traditional approaches based on single classifiers. The design of the whole system is divided into three steps: construction of the classifier pool, selection of individual learners, and combination of classifiers. Finally, our system is structured in a parallel topology consisting of several SVMs. Each individual learner (SVM) uses different speech data types and voice features. This gathers the strengths of the individual learners, obtaining enhanced performance by their combination. Another contribution of this paper is that we also investigated the impact of different speech types and emotions in depression detection. First, the study investigated the discriminative power of four speech types-reading, interview, picture description, and video description-for the recognition of depression using an SVM. Second, our research determined the different classification results of three speeches emotions-positive, neutral, and negative.

In the remainder of the paper, Section II describes the speech database that was collected for this study. Section III introduces the new depression prediction method. Following this, we showed our experiment results in Section IV. And the discussion of the experiments can be found in Section V.

Speech Data Collection

The procedure of speech data collection includes four parts: reading, interview, picture description, and video description. Each part can be divided into three groups in terms of its induced emotion: positive, negative, and neutral emotion. Details of the experiment are as follows:

- **Reading:** This part consisted of three short articles. The number of words in each article was consistent, and all of them are commonly used in Chinese. These materials described different scenes that expressed a positive emotion, neutral emotion, and negative emotion respectively. These text materials contained the corresponding emotional words, which came from the Chinese Affective Words System (CAWS) [35].

- **Interview:** The interview part contained nine questions: three positives, three neutrals, and three negatives. The question topics came from some depression scales, such as the Self-Rating Depression Scale (SDS) [36], Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [37], Hamilton Depression Scale (HAMD) [38], etc. The following were sample questions: If you have a vacation, please describe your travel plans (positive emotions). How do you evaluate yourself (neutral emotions)? Is there anything that makes you feel sorry or remorseful (negative emotions)?

- **Picture description:** This part comprised six pictures. There were three facial expressions and three scene pictures that expressed a positive, neutral, and negative emotion. These pictures were selected from the Chinese Facial Affective Picture System (CFAPS) [39] and Chinese Affective Picture System (CAPS) [40]. After observing the picture, the participants were asked to describe the contents of the picture and what they thought of them.

- **Video description:** Three types of video clips were selected, and the duration of each of them was 1 min. These video clips were all from the Chinese Affective Video System (CAVS) [41]. A comedy clip, which was excerpted from an animation series entitled Larva, expresses positive emotions. Neutral emotional video clips were excerpted from a documentary entitled Space Millennium, and the negative ones were excerpted from a film entitled Bodyguards and Assassins.

During the course of our experiments, voice data were collected in a quiet, soundproof room without any other interference. The ambient noise of the experiment was less than 60 dB. The voice signals were collected with 44.1 kHz sampling rate and 24-bit sampling bits and saved in WAV format. To be included in the study, the participants must be 18-55 years old, could be male or female, and must be a Chinese speaker. Before the experiment, each participant was asked to fill in some questionnaire that contains information such as age, gender, educational level, health condition, etc. The Beck Depression Inventory (BDI) scores of the participants were tested too, which were used to help us distinguish the categories of the participants.

All subjects were divided into two groups: depressive patients (DP) and healthy persons (HP). Depressive patients were persons who were all diagnosed with depression by a psychiatrist and whose BDI scores were higher than 14. Healthy persons were all persons who have had no history of depression or other mental disorder in the past or in the present and whose BDI scores were lower than 14. Careful editing and inspection of each sample ensured that only high-quality recordings without noise and unwanted interference were selected. After the completion of the selection process, a subset of 37 control and 37 depressed subjects were chosen. The basic information of the subjects is summarized in Table 1. We gave a gender balance between the DP group and the HP group. The actual class labels were set to be equal to +1 for the DP and-1 for the HP.

Each of the 74 subjects presented four different speech recordings, which were reading, interview, picture description, and video

	Healthy subjects (-1)		**Depressive Subjects (+1)**	
Parameter	**Male**	**Female**	**Male**	**Female**
Number	19	18	19	18
Age (years)	AVGa: 41.8	AVGa: 31.9	AVGa: 34.7	AVGa: 36.7
	Devb: 10.2	Devb: 12.9	Devb: 9.2	Devb: 12.8
BDI Score	AVGa: 4.5	AVGa: 5.2	AVGa: 27.1	AVGa: 30.1
	Devb: 4.0	Devb: 6.2	Devb: 9.3	Devb: 9.4

Table 1: Basic information of subjects.

description. In addition, each type of speech recording was grouped into three categories according to the emotions used in the task: positive, neutral, and negative. All in all, there were 12 speech data recordings for each subject: positive interview (INT_POS), neutral interview (INT_NEU), negative interview (INT_NEG), positive read (RED_POS), neutral read (RED_NEU), negative read (RED_NEG), positive picture description (PIC_POS), neutral picture description (PIC_NEU), negative picture description (PIC_NEG), positive video description (VID_POS), neutral video description (VID_NEU), and negative video description (VID_NEG) (Table 1).

Method

Figure 1 illustrates the overall framework of the basic classification system commonly used in depression recognition [11,42]. The procedure includes two main stages: training stage and classification stage. The training stage uses a set of data with known classes to build a classification model. After the pre-processing, feature parameters characterizing the speech acoustics of each class are calculated. Finally, a classification model (SVM) for distinguishing between depressive patients and healthy persons is built. At the decision-making stage, the procedures also include pre-processing and feature calculation, and the SVM model is used to provide the final prediction result. In our system, these basic steps are necessary too. However, we adopted an ensemble learning approach to construct a classifier, and the overall structure is shown in Figure 2. In this section, the detailed construction method of the multi-classifier system is introduced.

Construction of the individual learners' pool

Multi-classifier systems (MCS) focus on the combination of some individual learners from heterogeneous or homogeneous modeling backgrounds to give the final decision [43]. The construction of the pool of classifiers is the first step in creating a multi-classifier system. In our experiments, the pool of classifiers contained a large number of homogeneous classifiers. These individual learners were SVMs, and each learner used different speech data types and speech features to train the model. As shown in Figure 1, every single learning process involves pre-processing, feature extraction, and classification steps. Next, these steps are described separately.

Before the feature extraction, some pre-processing steps are needed. Pre-processing mainly includes filtering, framing, windowing, and sometimes endpoint detection for some particular feature extraction. In our experiment, the frame size was set to 25 ms at a shift of 10 ms and used a Hamming window.

Voice features can be divided into acoustic and linguistic features [44]. In this experiment, only acoustic features were extracted since we wanted to analyse the characteristics of depressed speech regardless of the language used. We computed some low-level descriptors on a frame-by-frame basis: short time energy (log Energy), Intensity, Loudness, Zero-Crossing Rate (ZCR), F0, jitter, shimmer, formants, and Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Coefficient (LPC), Spectral Centroid, Line Spectrum Pair (LSP), Perceptual Linear Predictive Coefficients (PLP), etc. To obtain a fixed number of features per item, we computed some functional characterizing the statistical and temporal properties of these low-level descriptors. These features have been shown to have a favourable effect on depression classification, and they can be calculated easily by the open SMILE software [45]. All of these features are shown in Table 2.

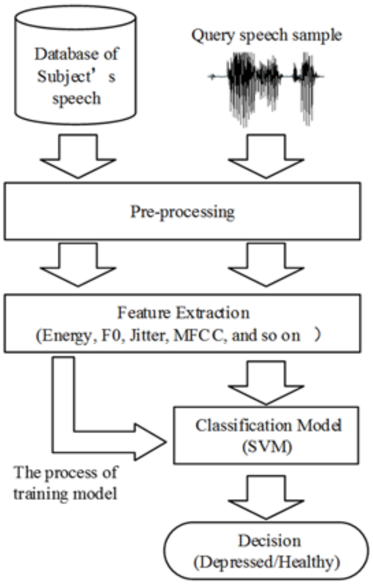
Database of Subject's speech
Query speech sample
Pre-processing
Feature Extraction (Energy, F0, Jitter, MFCC, and so on)
Classification Model (SVM)
The process of training model
Decision (Depressed/Healthy)

Figure 1: Overview of the basic classification system for depression re cognition using speech signal.

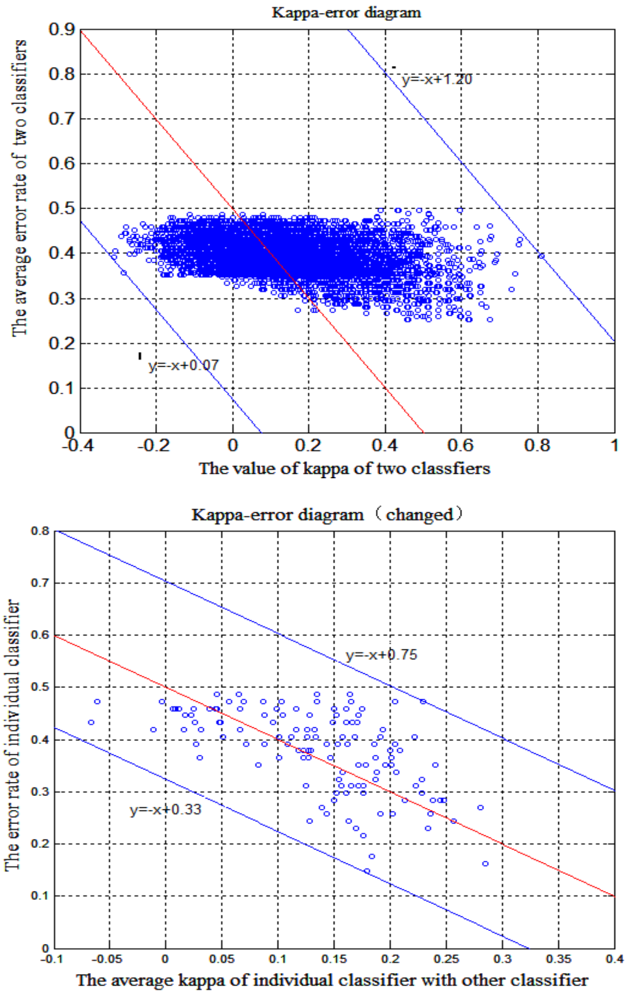

Figure 2: Kappa-error diagram of 15876 classifier pairs (top) and the new kappa-error diagram of 126 classifiers (bottom).

Moreover, the calculation method of the relevant features can be found in L. Rabiner and R. Schafer theory [46].

SVM is considered the current state-of-the-art classifier and has been effectively used in modeling speech information for depression recognition [11]. Alghowinem et al. [42] compared four classifiers: Gaussian mixture model (GMM), SVM, Hierarchical Fuzzy Signature (HFS), and Multilayer Perceptron Neural Network (MLP). They concluded that SVM performed the best. SVM is a supervised learning model that looks for an optimal hyper plane as a decision function. In other words, it hopes that the samples of the separate categories are divided by a clear gap that is as wide as possible. New samples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall [47,48]. In our experiment, we used the LibSVM toolkit [49] to build the model. The Radial Basis Function (RBF) kernel function was chosen, and the cost and gamma parameters were optimized via a wide range grid search. For dimensionality reduction, principal component analysis (PCA) was used before building the model. PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [50,51]. The percentage of variance retained was set to 95%, which was used to determine the number of principal components to retain (Table 2).

There were 12 speech data sets and 16 speech features that were used. Each individual learner used different speech data types and speech features to train the model. Therefore, 192 individual learners were built. These learners formed a large pool of classifiers.

Selection of individual learners

Multi-classifier systems are being used to achieve the best possible classification by increasing the efficiency and accuracy of the classification. In a situation where a number of classifiers are available, the simplest approach would be to select all classifiers and use them for the classification task.

This approach, although simple and easy to implement, does not guarantee a good performance. It is highly probable that worse performing classifiers might only add to the complexity of the problem and could even provide lower results than those of the worst classifier. Thus, we needed to adopt some strategy for selecting the part of the classifiers for the multi-classifier system.

Low-level descriptors (LLDs)	Functions applied to LLDs
LogEnergy	
Formant (3)	
Intensity	
Jitter	
LogHNR	maximum, minimum, range,
Loudness	mean, four quantiles and the
LPC (8)	distance between them, variance,
LSP (8)	standard deviation, skewness,
MFCC (12)	kurtosis, coefficient of the liner
Pitch	regression and the error
PLP (5)	
Shimmer	
Spectral Centroid	
Spectral Entropy	
Spectral Flux	
ZCR	

Table 2: Low level descriptors of the acoustic features and their statistic functions.

Dietterich et al. [52] indicated that the use of an ensemble of classifiers could achieve better recognition rates than those obtained through a singular classifier when: (1) the recognition rate of each individual learner of the multi-classifier is greater than 0.5 (i.e., not random classifiers) and (2) the errors made by each individual learner are uncorrelated (i.e., have a high diversity). In other words, multi-classifier systems will only work when it is possible to build individual learners that have a high accuracy and have sufficient independence between them [53-55]. Therefore, we needed to select learners with a high recognition rate and a high diversity from 192 individual learners. At first, we selected the individual learners that had an accuracy of better than 0.5. Then, we continued to make a choice on the remaining classifiers (126 individual learners).

Our aim was to choose learners with a high recognition rate and a high diversity from the pool of individual learners. We used classification errors to measure the recognition rates and the kappa statistic to measure the diversity. Smaller values of kappa (k) indicate a high diversity, k=0 indicates independent classifiers and k=1 indicates identical classifiers. The value of k is usually nonnegative; it is negative only if the probability of agreement between the two classifiers is even below chance [56]. Then, we used a kappa-error diagram to relate the two concepts. The kappa-error diagram, which was proposed by Margineantu and Dietterich [57], visualizes individual accuracy and diversity in a 2D plot (Figure 2). It has been used to decide which ensemble members can be pruned without much harm to the overall performance [58]. An ensemble of L classifiers is shown on the kappa-error diagram as a scatterplot of L (L – 1)/2 points, each corresponding to a pair of classifiers. The x coordinate of the pair is the value of kappa for the two classifiers. The y coordinate of the pair is the average of their error rates. The smaller the error, the higher the recognition rate, and the larger value of kappa, the higher the diversity. Thus, points that are closer to the bottom left corner of the diagram are preferable (high diversity and low error).

Consider the problem of separating the set of training data (x_1, y_1), (x_2, y_2), … (x_m, y_m) into two classes, where xi ∈ R^N is a feature vector and y_i ∈{-1, +1} is class label. Suppose the contingency table of two classifiers (Ci and Cj) is:

	Cj correct (DP,+1)	Cj wrong (HP,-1)
Ci correct (DP,+1)	a	b
Ci wrong (HP,-1)	c	d

Where the table entries are the number of points jointly classified as indicated, and a+b+c+d=N.

The averaged individual error for the pair of classifiers is:

$$e = \frac{1}{2}\left(\frac{c+d}{N} + \frac{b+d}{N}\right) = \frac{b+c+2d}{2N} \tag{1}$$

Kappa is constructed as

$$p_1 = \frac{a+d}{N} \tag{2}$$

Where p_1 is the observed agreement, i.e., the probability that the two classifiers will be both correct or both incorrect when classifying a randomly chosen data point, and p_2 is the agreement by chance, i.e., the probability that the two classifiers will agree by chance on a randomly chosen data point. The two quantities are calculated as

$$p_1 = \frac{a+d}{N} \tag{3}$$

$$p_2=\frac{(a+b)(a+c)+(b+d)(c+d)}{N^2} \qquad (4)$$

The top part of Figure 2 shows an example of a kappa-error diagram for our experiment. There are 15,876 classifier pairs (126 × 126) in this figure. The points that are closer to the bottom left corner of the diagram have a high diversity and a low error (i.e., high accuracy). Therefore, we can use a straight line with a slope of −1 (i.e., a 45° falling line, y = −x + b) to separate the good and the bad points, and points at the bottom left of the line are preferable. As shown in Figure 2, all points are contained in two blue lines, and the red line gradually moves from line y = −x + 0.07 to line y = −x +1.20 in parallel. We choose all the points at the bottom left of the red line to build the multi-classifier system. As the red line moves, the performance of the multi-classifier system gets better. The red line stops moving when the performance of the multi-classifier system is not increasing. At this time, the selected classifiers are the best combination. When the red line moves to line y = −x + 1.20, it should stop too. At this time, all classifiers are selected to build multi-classifiers. The slope of the line has a great influence on the result of the selection. The reason why we choose a straight line with a slope of −1 is that both the error rate and the kappa value are considered equally. When the straight line is parallel to the x-axis, we only consider the effect of the y (i.e., error rate). When the straight line is parallel to the y-axis, we only consider the effect of the x (i.e., kappa).

However, this method has a particularly serious problem. The point in the kappa-error diagram represents a pair of classifiers. After the straight line moves a very small distance, all the classifiers have been included. It does not make sense for the straight line to continue moving. As shown in Figure 3, all the classifiers are already included in the multi-classifier system when the value of b adds to 0.3. Therefore, we made a slight improvement to the kappa-error diagram. Each point of the new kappa-error diagram corresponds to an individual classifier. The x coordinate of the point is the average of kappa for one individual classifier with all the other classifiers:

$$k_i=\sum_{j=1}^{N-1}k_{ij} \qquad (5)$$

where k_{ij} is the value of kappa of classifiers Ci and Cj. The y coordinate of the point is the value of the error rate for one individual classifier:

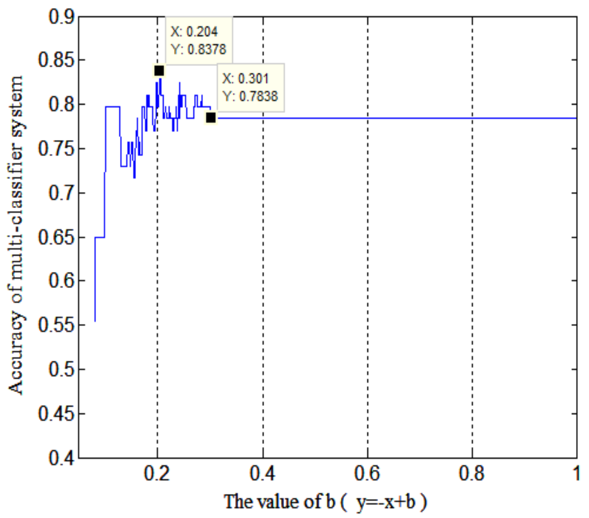

Figure 3: The relationship between b (y = - x + b) and the accuracy of multi-classifier (For traditional kappa-error diagram, i.e. the top part of Figure 2).

$$e=\frac{c+d}{N}\,or\,\frac{b+d}{N} \qquad (6)$$

The bottom part of Figure 2 shows a situation about the new kappa-error diagram for our experiment. The selection process is the same as described before, but the two border lines are y = −x + 0.33 and y = −x +0.75. This small improvement not only retains the advantages of the traditional kappa-error diagram but also solves the previous problems. Moreover, this is a significant reduction in computing time. Most importantly, it improves the accuracy of the multi-classifier system more than before, as shown in Section IV.

Combination of classifiers

After selecting those beneficial individual learners, we will consider how to interconnect these classifiers. The overwhelming majority of multi-classifier systems reported in the literature are structured in a parallel topology [59]. In this architecture, the final decision of the combined classifier output is made on the basis of the outputs of the individual learners obtained independently. As shown in Figure 4, speech samples were classified as DP and HP using a parallel configuration of some single classifiers. Each individual classifier included a single SVM classifier using different types of speech data sets to train the model, and every individual classifier used a single category of features (e.g., LPC, MFCC, energy, etc.)

Fuser design is another consideration for multi-classifier system [60,61]. As a common fusion method, majority voting was adopted in our system. In our experiment, the class estimates (Figure 4) given by the system were calculated by parameter r (x_i) given as:

$$r(x_i)=\sum_{i=1}^{n}\tilde{y}_k(x_i) \qquad (7)$$

Where $y_k(x_i)$is the estimated results by the kth classifier and the final classification decision was then made according to the sign of $r(x_i)$. When the symbol of $r(x_i)$ was positive, the final result was judged to be a DP. However, when the symbol was negative, the final result was judged to be an HP.

Experiments and Results

Before building a multi-classifier system, we trained a number of individual learners. These learners used different speech data types and speech features to train the model. Analysing the results of these classifiers helped us to understand the impact of speech types and

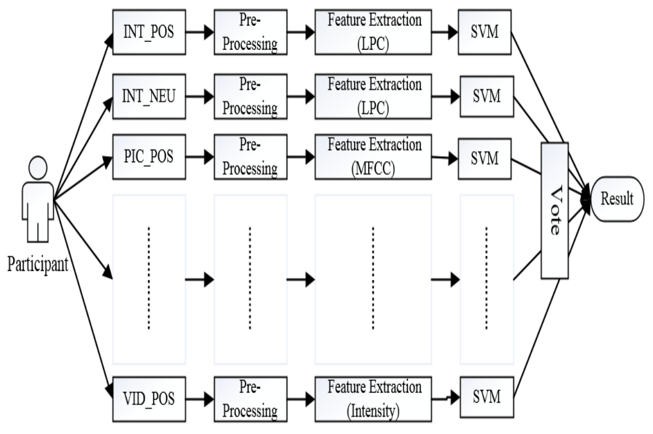

Figure 4: A multi-classifier system using different speech types and emotions.

emotions in depression classification. After that, we tested the new multi-classifier system, and the results obtained from the new system were compared with single classifier techniques using a single type of feature category. In this study, the correct classifications of DP and HP were measured in terms of accuracy. To mitigate the effect of the limited amount of data, we used a leave-one-out cross-validation, without any overlap between the training and the testing data.

Results of the individual learners

The accuracy of each classifier classification is shown in Table 3. Most of individual classifiers had an accuracy of 50%-70%. The highest accuracy of these classifiers was 72.97%, and in this classifier, neutral interview speech was used and the speech feature that we selected was MFCC. As shown in Figure 5, for the average case, the overall recognition rate using interview speech was higher than that using picture description, video description, and reading speech. In addition, the results also indicated that neutral speech worked better than positive and negative speech. Specifically, the neutral interview speech type gave the best result.

The performances of various features on different data sets were different. As shown in Table 3, energy, LPC, LSP, and MFCC were the common features that gave a high accuracy in the interview, picture description, video description, and reading speech, whereas F0, jitter, shimmer, formants, and PLP coefficients were the worst. For interview speech, energy, intensity, jitter, loudness, LPC, LSP, MFCC, spectral flux, and ZCR gave a better result. For picture description speech, energy, jitter, LPC, LSP, MFCC, and spectral flux gave a better result. For video description speech, energy, LPC, LSP, MFCC, and spectral flux gave a better result. For reading speech, most features did not produce good results (Figure 6).

Results of the multi-classifier system

Using the method described in Section III, we tested a new multi-classifier method. At first, we created a large number of individual learners, and each learner used different speech types and different speech features. Then, we selected individual learners with a high diversity and a high recognition rate. During this process, the improved kappa-error diagram was used to make the choice. Finally, we combined the classifiers into a multi-classifier system with parallel topologies (Figure 5).

As previously stated, diversity and accuracy are two of the guiding measures of the design process for multi-classifier systems [62,63].

	INT_NEG	INT_NEU	INT_POS	PIC_NEG	PIC_NEU	PIC_POS	RED_NEG	RED_NEU	RED_POS	VID_NEG	VID_NEU	VID_POS
LogEnergy	61	65	62	55.41	58.1	46	51	57	54	55	57	66
Formant	45	54	47	48.65	48.7	42	49	31	49	59	53	34
Intensity	50	62	61	44.59	60.8	45	59	57	51	55	41	64
Jitter	45	58	54	56.76	54.1	61	41	46	23	30	42	39
LogHNR	30	42	35	40.54	51.4	41	47	50	38	28	53	30
Loudness	53	69	57	51.35	59.5	54	53	45	59	42	62	59
LPC	65	69	66	66.21	71.6	61	45	47	49	62	69	64
LSP	72	70	69	67.57	70.3	69	57	51	62	69	69	72
MFCC	66	73	68	71.62	66.2	70	53	64	59	70	72	69
Pitch	50	58	46	33.78	29.7	36	49	54	50	38	24	41
PLP	64	41	45	31.08	54.1	47	47	51	46	36	23	35
Shimmer	58	27	28	43.24	25.7	45	41	43	28	47	39	22
Spectral Centroid	54	55	47	48.65	33.8	31	47	54	53	53	61	51
Spectral Entropy	42	47	46	48.65	58.1	54	50	46	51	50	57	53
Spectral Flux	51	59	53	62.16	59.5	54	54	58	57	53	64	61
ZCR	49	54	53	31.08	50	45	49	51	46	26	65	53
Average	53	57	52	50.08	53.2	50	49	50	48	48	53	51

Table 3: Accuracy (%) for acoustic features classification for different speech types and emotions.

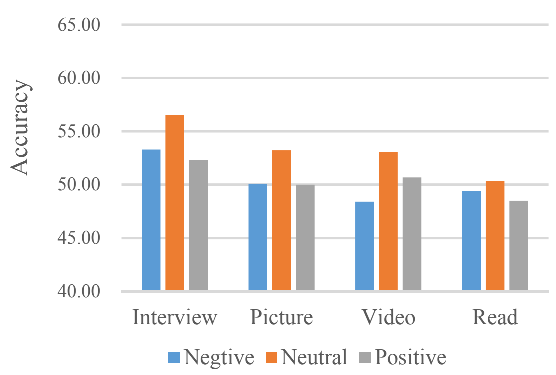

Figure 5: The average accuracy of different speech types.

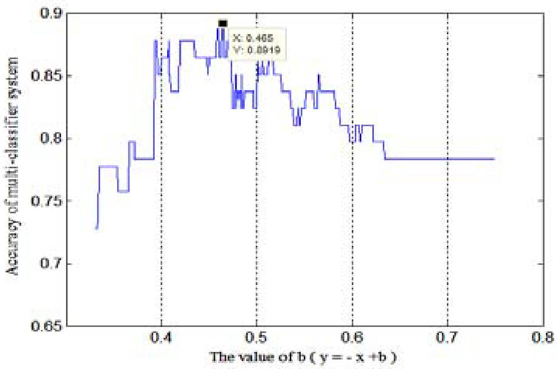

Figure 6: The relationship between b (y = - x + b) and the accuracy of multi-classifier (For new kappa-error diagram, i.e. the bottom part of Figure 2).

To increase diversity, we can use a number of methods [64]. First, every individual learner used different types of speech data to train the classification model, leading to a diversity of input data. Second, the speech features used by each learner were different, increasing the diversity of the attribute. Finally, although all learners used the SVM, the parameters of each learner were different. To increase the accuracy, we used the RBF kernel function to build the SVM, and the cost and gamma parameters were optimized via a wide range grid search. This method can improve the accuracy of every individual classifier in this system and the diversity between them.

As shown in Figure 5, the prediction accuracy of the multi- classifier system was 89.19%, indicating that this system gave better results than those of the single classifier methods (whose best result was 72.97%). The accuracy was highest when the value of b was 0.465, and the number of classifiers remaining in the pool after a pruning procedure was 11. Figure 3 shows the result of the unmodified kappa-error diagram, and the best result was 83.78% when the value of b was 0.204. The number of classifiers remaining in the pool after a pruning procedure was 86. Our ameliorations to the kappa-error diagram resulted in an overall improvement in accuracy and fewer individual learners for the multi-classifier system. In addition, as can be seen from Figure 3, if all the classifiers were integrated, the final result was just 78.38%.

Discussion and Conclusion

In our experiments, we researched two things. First, we investigated the performance of different speech data sets used for the identification of depression. Second, we combined a variety of different data sets to design a new multi-classifier system.

We found that using interview speech gave better results than that using picture description, video description and reading speech. It indicated that interview speech contains more relevant information about the subjects' general characteristics including their affective state. Interview speech can be considered as spontaneous speech, which can better express our feelings and emotions, leading to good results. Picture and video description can also be considered as spontaneous speech, but their performance was worse than that of interview speech. We speculate that most of the interview questions referred to the subjects themselves such that it was easy for them to get into an emotional state. In addition, we also found that neutral speech performed better than positive and negative speech. Previous studies have reported that depression is characterized by sadness, loss of interest or pleasure, and feelings of guilt or low self-worth [65]. This leads to the tendency in which depressive patients are more likely to generate negative emotions when they are expressing their feelings, but it is difficult for them to generate positive emotions. Hence, in the face of neutral experimental materials, depressive patients were more likely to generate negative emotions, whereas healthy persons were more likely to generate positive emotions, resulting in larger deviations between depressive patients and healthy persons. The deviations caused by positive materials were a little worse than those caused by neutral experimental materials. Nevertheless, in the face of negative experimental materials, depressive patients and healthy persons all tended to generate negative emotions, resulting in minor deviations.

Although a single speech data set provided weak prediction results, it was possible that it could provide important information for a multi-classifier system. The multi-classifier system was structured in a parallel topology, and every learner used an SVM to distinguish between depressive patients and healthy persons on different speech data sets and different speech features. This system gathers the strengths of the individual classifiers, obtaining enhanced performance by their combination. Finally, the prediction accuracy of the multi-classifier system was 89.19%, which was higher than those of single classifier methods.

Finally, some of the problems in the experiment cannot be ignored. It should be noted that the amount of data used here was relatively small. Therefore, in our further study, we intend to collect more data, and future studies will be able to report on a larger data set.

Acknowledgement

#Hailiang Long and Dr. Xia Wu contributed equally to this work. This work was supported in part by Natural Science Foundation of Beijing Municipality under Grant Z171100000117005, the general Program of National Natural Science Foundation of China under Grant 61571047 and the Fundamental Research Funds for the Central Universities [2017EYT36].

References

1. Aaron TBMD, AAPD Brad (2014) Depression: Causes and Treatment. Clin Geriatr Med 14: 765-786.
2. Willner P (2016) The chronic mild stress (CMS) model of depression: History, evaluation and usage. Neurobiol Stress 6: 78.
3. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS (2007) Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J Neurolinguist 20: 50-64.
4. Hu B, Wan J, Dennis M, Chen HH, Li L, et al. (2010) Ontology-based ubiquitous monitoring and treatment against depression. Wirel Commun Mob Com 10: 1303-1319.
5. Cummins N, Joshi J, Dhall A, Sethu V, Goecke R, et al. (2013) Diagnosis of depression by behavioural signals: a multimodal approach. ACM 1: 11-20.
6. Girard JM, Cohn JF, Mahoor MH, Mavadati SM, Hammal Z, et al. (2014) Nonverbal Social Withdrawal in Depression: Evidence from manual and automatic analysis. Image Vis Comput 32: 641-647.
7. Scherer S, Stratou G, Mahmoud M, Boberg J (2014) Automatic behavior descriptors for psychological disorder analysis. Image Vis Comput 32: 648-658.
8. Tan TES, Dai JA, Lan SS (2017) Speech analysis and depression. Informat Proces Assoc Summ Conf 1: 1-4.
9. Cummins N, Epps J, Breakspear M, Goecke R (2011) An Investigation of Depressed Speech Detection: Features and Normalization.
10. Hashim NW, Wilkes M, Salomon R, Meggs J, France DJ (2017) Evaluation of Voice Acoustics as Predictors of Clinical Depression Scores. J Voice 31: 26.
11. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, et al. (2015) A review of depression and suicide risk assessment using speech analysis. Speech Commun 71: 10-49.
12. Horwitz R, Quatieri TF, Helfer BS, Yu B, Williamson JR, et al. (2013) On the relative importance of vocal source, system, and prosody in human depression. IEEE Int Conf 1: 1-6.
13. Liu Z, Hu B, Liu F, Kang H, Li X, et al. (2016) Evaluation of Depression Severity in Speech. Springer 1: 312-321.
14. Mundt JC, Vogel AP, Feltner DE, Lenderking WR (2012) Vocal acoustic biomarkers of depression severity and treatment response. Biol Psychiatry 72: 580-587.
15. Nilsonne A, Sundberg J, Ternström S, Askenfelt A (1998) Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. J Acoust Soc Am 83: 716.
16. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM (2000) Acoustical properties of speech as indicators of depression and suicidal risk. IEEE 47: 829-837.
17. Darby JK, Hollien H (1977) Vocal and speech patterns of depressive patients. Folia Phoniatr Logo 29: 279-291.
18. Hollien H (2010) Vocal indicators of psychological stress. Ann N Y Acad Sci 347: 47-72.
19. Low LS, Maddage NC, Lech M, Sheeber LB, Allen NB (2011) Detection of Clinical Depression in Adolescents' Speech During Family Interactions. IEEE Trans Biomed Eng 58: 574-586.

20. Scherer KR (1987) Vocal assessment of affective disorders. Lawrenc Erlbau Assoc 1: 57-82.

21. Nunes A, Coimbra RL, Teixeira A (2010) Voice quality of European Portuguese emotional speech in Computational Processing of the Portuguese Language. Int Conf 1: 142-151.

22. Ozdas A, Shiavi RG, Silverman SE, Silverman MK (2004) Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. IEEE T Bio-Med Eng 51: 1530-1540.

23. Flint AJ, Black SE, Campbell-Taylor I, Gailey GF, Levinton C (1993) Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. J Psychiatr Res 27: 309-319.

24. Elliot Moore II, Clements MA, Peifer JW, Weisser L (2008) Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech. IEEE Trans Biomed Eng 55: 96.

25. Ooi KEB, Lech M, Allen NB (2013) Multichannel Weighted Speech Classification System for Prediction of Major Depression in Adolescents. IEEE T Bio Med Eng 60: 497-506.

26. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J (2015) Analysis of acoustic space variability in speech affected by depression. Speech Commun 75: 27-49.

27. Shankayi R, Vali M, Salimi M, Malekshahi M (2013) Identifying depressed from healthy cases using speech processing. Biom Eng 1: 242-245.

28. Goecke R, Wagner M, Epps J, Parker G, Breakspear M (2013) Characterising Depressed Speech for Classification. Int Speec Communicat Assoc 1: 3-5.

29. Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M (2013) Detecting Depression: A Comparison between Spontaneous and Read Speech. IEEE Int Conf Acoust Speech Sig Proces 1: 7547-7551.

30. Liu Z, Hu B, Yan L, Wang T, Liu F, et al. (2015) Detection of depression in speech. IEEE 1: 743-747.

31. Jiang H, Hu B, Liu Z, Yan L, Wang T, Liu F, et al. (2017) Investigation of different speech types and emotions for detecting depression using different classifiers. Researchgate 90: 39-46.

32. Dietterich T (2000) Ensemble methods in machine learning. SPRINGERLINK 1: 1-15.

33. Breiman L (1996) Bagging predictors. Springer 24: 123-140.

34. Dietterich T (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Springer 40: 139-158.

35. Shu-Jing XU, Yin HF (2008) Initial Establishment of the Chinese Affective Words Categorize System used in Research of Emotional Disorder. 22:770-774.

36. Zung WW (1965) A self-rating depression scale. Americ Psycho Assoc 12: 63-70.

37. A. P. Association (2013) Diagnostic and Statistical Manual of Mental Disorders: American Psychiatric Association.

38. Hamilton M (1960) A rating scale for depression. BMJ 23: 56.

39. Xu G, Huang YX, Yan W, Luo YJ (2011) Revision of the Chinese Facial Affective Picture System. RESEARCHGATE.

40. Lu B, Hui M, Yuxia H. (2005) The Development of Native Chinese Affective Picture System-A pretest in 46 College Students. RESEARCHGATE 19: 719-722.

41. Xu P, Yuxia H, Luo Y (2010) Establishment and assessment of native Chinese affective video system. Chinese Mental Health J 24: 551-554.

42. Alghowinem S, Goecke R, Wagner M, Epps J (2013) A comparative study of different classifiers for detecting depression from spontaneous speech. IEEE 1: 8022-8026.

43. Ranawana R, Palade V (2006) Multi-Classifier Systems: Review and a roadmap for developers.

44. Polzehl T, Schmitt A, Metze F, Wagner M. (2011) Anger recognition in speech using acoustic and linguistic cues. Speech Communication 53: 1198-1209.

45. Eyben F, Wollmer M, Schuller B (2010) Open smile: the munich versatile and fast open-source audio feature extractor. Researchgate 1: 1459-1462.

46. Rabiner L, Schafer R (2011) Theory and applications of digital speech processing.

47. Cristianini N, Shawe-Taylor J (2000) An introduction to support Vector Machines: and other kernel-based learning methods.

48. Williams CKI (2005) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. IEEE 16: 781-781.

49. Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. Research Gate 2: 27.

50. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Science Direct 2: 37-52.

51. Vidal R, Ma Y, Sastry SS (2016) Principal Component Analysis. Springer.

52. Dietterich TG (1988) Machine-learning research: Four current directions. Research Gate 1: 97-136.

53. Sharkey AJC, Sharkey NE (1997) Combining diverse neural nets. Cambrid Uni Pres 12: 231-247.

54. T.G. Dietterich, Machine-learning research: Four current directions, The AI Magazine, pp: 97-136, 1998.

55. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles. Springer link 5: 181-207.

56. Fleiss JL (1981) Statistical Methods for Rates and Proportions. Research gate.

57. Margineantu DD, Dietterich TG (1997) Pruning Adaptive Boosting. Research gate 1: 378-387.

58. Rokach L (2009) Collective-Agreement-Based Pruning of Ensembles. Science Direct 53: 1015- 1026.

59. Kuncheva L (2004) Combining Pattern Classifiers: Methods and Algorithms.

60. Ho TK, Hull JJ, Srihari SN (2002) Decision combination in multiple classifier systems. IEEE 16: 66-75.

61. Roli F, Giacinto G (2007) Design of Multiple Classifier Systems by clustering of classifiers. IEEE 22: 25-33.

62. Wozniak M, Grana M, Corchado E (2014) A survey of multiple classifier systems as hybrid systems. Science Direct 16: 3-17.

63. Roli F, Giacinto G, Vernazza G (2001) Methods for Designing Multiple Classifier Systems.

64. Kuncheva LI (2005) Diversity in multiple classifier systems. Sciencedirect 6: 3-4.

65. Selbaek G, Borza T (2017) Depression in Later Life.