Speech Communication 90 (2017) 39–46

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Investigation of different speech types and emotions for detecting depression using different classifiers

Haihua Jiang[a], Bin Hu[a,*], Zhenyu Liu[b], Lihua Yan[b], Tianyang Wang[b], Fei Liu[b], Huanyu Kang[b], Xiaoyu Li[b]

[a] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[b] Ubiquitous Awareness and Intelligent Solutions Lab, Lanzhou University, Lanzhou 730000, China

ARTICLE INFO

Article history:
Received 22 October 2016
Revised 17 January 2017
Accepted 11 April 2017
Available online 26 April 2017

Keywords:
Acoustic features
Depression
Classifiers
Speech types
Speech emotions

ABSTRACT

Depression is one of the most common mental disorders. Early intervention is very important for reducing the burden of the disease, but current methods of diagnosis remain limited. Previously, acoustic features of speech have been identified as possible cues for depression, but there has been little research to link depression with speech types and emotions. This study investigated acoustic correlates of depression in a sample of 170 subjects (85 depressed patients and 85 healthy controls). We examined the discriminative power of three different types of speech (interview, picture description, and reading) and three speech emotions (positive, neutral, and negative) using different classifiers, with male and female subjects modeled separately. We observed that picture description speech rendered significantly better ($p < 0.05$) classification results than other speech types for males, and interview speech performed significantly better ($p < 0.05$) than other speech types for females. Based on speech types and emotions, a new computational methodology for detecting depression (STEDD) was developed and tested. This new approach showed a high accuracy level of 80.30% for males and 75.96% for females, with a desirable sensitivity/specificity ratio of 75.00%/85.29% for males and 77.36%/74.51% for females. These results are encouraging for detecting depression, and provide guidance for future research.

1. Introduction

Depression is one of the most common mental disorders. Globally, an estimated 350 million people of all ages suffer from depression (World Health Organization, 2016), which can cause the affected person to function poorly at work, school, and within the family. The lifetime risk for depression is reported to be at least 15% (Kessler et al., 2003). At its worst, depression can lead to suicide (Hawton et al., 2013). Early intervention aimed at preventing the onset of clinical depression can provide a very important means for reducing the burden of the disease. However, currently the range of diagnostic tools for identifying depression is quite limited. Assessment methods rely almost exclusively on patient self-reporting and clinical opinion (Mundt et al., 2007), risking a variety of subjective biases. Consequently, it is particularly important to look for new objective measures that assist clinicians in their diagnosis and monitoring of clinical depression.

The emotional state of a person suffering from a depressive disorder affects the acoustic qualities of his/her speech (Cummins et al., 2015). Therefore, depression could be detected through an analysis of perceived changes in the acoustical properties of speech. The link between acoustic parameters in speech signals and depression has been researched extensively. The speech behavior produced by depressed patients has been shown to vary as a result of the negativity of conversational content (Vanger et al., 1992) and the cognitive effort required (Calev et al., 1989). Different speech types and emotions may elicit different levels of cognitive effort, or induce various emotional effects, which can produce changes in speech acoustics that affect the classification of depression. However, there has been little research exploring the correlation between depression, speech types, and speech emotions. There is also a lack of objective tools for clinical analysis of depression based on speech.

The purpose of our work was to investigate the impact of speech types and emotions in depression classification, and provide an effective measure for detecting depression. First, the study investigated the discriminative power of three speech types — interview, picture description, and reading — for the recognition of depression using three popular classifiers: K nearest neighbors (KNN), Gaussian mixture model (GMM), and Support vector machine (SVM). Second, our research determined the different

* Corresponding author.
E-mail address: bh@bjut.edu.cn (B. Hu).

http://dx.doi.org/10.1016/j.specom.2017.04.001
0167-6393/

classification results of three speech emotions — positive, neutral and negative — using these different classifiers. Finally, based on speech types and emotions, we proposed a new computational methodology for detecting depression.

The remaining parts of this paper are organized as follows. Section 2 contains a brief review of existing methods. Section 3 describes the speech database that was collected for this study. Section 4 describes the methodology of the study. The experiments and results are described in Section 5, followed by the conclusions in Section 6.

2. Previous work

Depressed speech has been characterized consistently by clinicians as dull, monotone, and lifeless (Sobin and Sackeim, 1997). Darby and Hollien (1977) conducted a pilot study of severely depressed patients, and found that listeners could perceive noticeable differences in prosodic characteristics of depressed speech. A number of recent studies have demonstrated that acoustic speech analysis can be used efficiently to recognize symptoms of depression.

A wide range of features have been explored for automatic depressed speech classification. Moore et al. (2008), Low et al. (2011), and Ooi et al. (2013, 2014) investigated the suitability of forming a classification system from combinations of prosodic, spectral, and glottal features. Alghowinem et al. (2013a), Valstar et al. (2014), and Low et al. (2010) summarized and compared Low-Level descriptors and statistical features of depression classification. Investigation of mel-frequency cepstrum coefficients (MFCC) by Cummins et al. (2011, 2014) and Joshi et al. (2013) found that the classification results were statistically significant for detecting depression. Ozdas et al. (2004) and Quatieri and Malyska (2012) found that depressed patients exhibited higher energy in the upper frequency bands of the glottal spectrum.

The two most popular modeling and classification techniques used in the literature include SVM and GMM. Ooi et al. (2013) presented a multi-feature approach using GMM classifiers, and they reported a binary classification of 73% (Sens. 0.79, Spec. 0.67). Low et al. (2011) used a 2-class gender-independent GMM classifier and reported classification accuracies ranging from 50% to 75%. Cummins et al. (2011) used a GMM back-end and reported a classification of 79%. Alghowinem et al. (2013b) compared four classifiers: GMM, SVM, Hierarchical Fuzzy Signature (HFS), and Multilayer Perceptron Neural Network (MLP). They concluded that GMM and SVM performed better. Helfer et al. (2013) reported the stronger performance of SVM over GMM when classifying the severity of depression. Cohn et al. (2009) used a gender independent SVM classifier and reported an accuracy of 79% (Sens. 0.88, Spec. 0.64). To the best of our knowledge, the method using KNN for depressed speech classification has not been described in the literature. However, KNN has been used effectively in speech emotion modeling (Pao et al., 2008; He et al., 2011; Muthusamy et al., 2015).

Several papers have attempted to identify which speech types and emotions provide the most reliable recognition of depression. Low et al. (2011) compared three interactions: event-planning interaction (EPI), problem-solving interaction (PSI), and family consensus interaction (FCI). They concluded that PSI provides consistently higher results. Alghowinem et al. (2013a) found that using spontaneous speech gave a more accurate result than using read speech for most features. Gupta et al. (2014) and Sidorov and Minker (2014) found that using freeform data attained superior performance to using read passage data. Alghowinem et al. (2012) found that talking about positive emotions in an interview resulted in increased correct recognition of depression. Goeleven et al. (2006) reported that depressed patients showed a specific failure to impair inhibitions relating to negative information. Gollan et al. (2008) and Leyman et al. (2007) found that depressed individuals exhibited enhanced memory and attention for negative expressions, and they interpreted neutral faces more negatively than controls. However, in most of the previous research, males and females were modeled together, and just one classifier was used in each study without comparisons.

Early studies of acoustic correlates in speech were usually limited to small databases, using very few participants and short audio recordings. For instance, Moore et al. (2008) and Mantri et al. (2013) interviewed 33 subjects (15 depressive patients, 18 healthy controls) speaking American English. Alghowinem et al. (2012) recruited 40 depressed subjects and 40 healthy controls speaking English. Low et al. (2011) studied 139 adolescents (68 depressed, 71 healthy) speaking English. A depression corpus composed of 84 subjects speaking German was used by Valstar et al. (2014), Mitra and Shriberg (2014), Lopez-Otero et al. (2015), and Williamson et al. (2014). Mundt et al. (2007) and Horwitz et al. (2013) each recruited 35 patients (20 females and 15 males) to participate. It should be noted that majority of the participants in most of this previous research spoke Western languages. Therefore, it is necessary that further research should validate the proposed measures with larger sample sizes and a greater variety of languages.

3. Speech database

In our research, depressed patients and healthy controls, both male and female, were included as subjects. They ranged in age from 18 to 55 years old, were native Chinese speakers, and had at least a primary school education (Liu et al., 2015). Each subject was asked to complete a pre-assessment booklet that included general information, such as healthy history, as well as demographic information, such as age, gender, education level, and type of employment. Next, each participant was assessed by psychiatrists using the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) (American Psychiatric Association, 1994) rules for diagnosis. All the participants were asked to finish the Patient Health Questionnaire-9 (PHQ-9) (Kroencke et al., 2001). Once the screening process was completed, subjects were divided into two groups according to the PHQ-9 scores: healthy controls (PHQ-9$\geqslant$5), and depressed patients (PHQ-9$\geq$5). Depressed patients were diagnosed with pure depression, and had no other mental disorders or medical conditions. Healthy controls were selected who had no history of mental illness and who matched the depressed subjects broadly in terms of demographics.

During the course of our experiments, it was necessary to keep the ambient noise level of the lab to less than 60 dB The audio signals were recorded with a 44.1 KHz sampling rate and 24-bit sampling depth. All recording data were saved in uncompressed WAV format. The experimental paradigm contained three parts, including a reading task, an interview with the subjects, and a picture description task. The reading task contained a short story and three groups of words with positive, neutral, and negative emotions. The story was named, “The North Wind and the Sun,” from the booklet, “The Principles of the International Phonetic Association” (France et al., 2000). The interview task contained 18 questions, which were divided into three groups according to their emotion valence: 6 positive, 6 neutral, and 6 negative. The question topics came from DSM-IV, Self-rating Depression Scale (SDS), and Hamilton Depression Scale (HAMD). The following are sample questions: What is the best gift you have ever received, and how did you feel (positive emotions)? If you have a vacation coming up, please describe your travel plans (positive emotions). How do you evaluate yourself (neutral emotions)? Please describe one of your friends, including that person’s age, job, character, and hobbies (neutral emotions). What would you like to do when you are un-

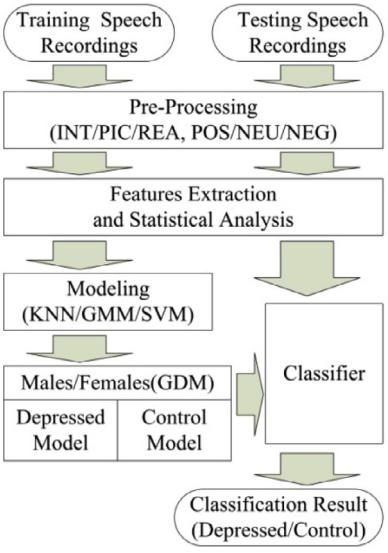

Fig. 1. Block diagram for modeling speech of depressed and control subjects.

able to fall asleep (negative emotions)? What circumstances could actually cause you to be desperate (negative emotions)? The picture description task included four pictures in all. Three pictures, expressing positive, neutral, and negative faces, were selected from the Chinese Facial Affective Picture System (CFAPS). The last picture with a "crying woman" came from the Thematic Apperception Test (TAT) (Hönig et al., 2014). In this task, subjects were told to describe these four pictures freely. More details about the experiment for data collecting can be found in Liu et al. (2015).

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, the speech database used in the experiments consisted of recordings of 85 depressed subjects (53 females and 32 males) and 85 controls (51 females and 34 males). Each subject's speech was divided into 29 recordings according to different subtasks. In other words, there were 4930 speech recordings in this study.

4. Methodology

The proposed framework for the modeling and classification of the depressed and control participants' speech is illustrated in Fig. 1. In the following sections of this paper, the pre-processing, features extraction and modeling techniques are described.

4.1. Pre-processing

Each of the 170 participants was represented by 29 speech recordings. These 29 recordings were split into three partitions according to different speech types: interview (INT, including 18 speech recordings), picture description (PIC, including 4 recordings), and reading (REA, including 7 recordings). In addition, we grouped these recordings into three categories according to the emotions used in the task: positive (POS, including 6 INTs, 2 REAs, and 1 PIC), neutral (NEU, including 6 INTs, 3 REAs, and 1 PIC), and negative (NEG, including 6 INTs, 2 REAs, and 2 PICs). The total durations of INT, PIC, and REA were 52,427 s, 16,203 s, and 21,425 s, respectively. The total durations of POS, NEU, and NEG were 23,961 s, 32,998 s, and 33,096 s, respectively. The preprocessing was performed on a frame-by-frame basis, with a frame length of 25 ms and 50% overlap between frames.

Table 1
Description of the acoustic features based on 38 LLDs and their first derivate and 21 feature statistics functions.

Descriptors	Functions
PCM loudness	maxPos, minPos, mean,
MFCC[0–14]	stddev, skewness, kurtosis,
Log Mel-frequency band[0–7]	quartile 1/2/3
LSP frequency[0–7]	quartile range (2–1)/(3–2)/(3–1)
F0 envelop	lin.regression coeff.1/2
Voicing probability	lin.regression error Q/A
F0final	percentile 1/99
jitterLocal, jitterDDP	percentile range (99–1)
shimmerLocal	up-level time 75/90

4.2. Features extraction

Acoustic features can be categorized into two branches: low-level descriptors (LLD), which are extracted frame-by-frame, and statistical functions, which are statistical measurements over the low-level features. In this work, we employed the publicly available open-source software openSMILE (Eyben et al., 2010) to extract several low level voice features and functional features from the pure subject speech. The feature set consisted of 1582 features that resulted from a base of 34 LLDs with 34 corresponding delta coefficients appended, and 21 functions applied to each of these 68 LLD contours (1428 features). In addition, 19 functions were applied to the 4 pitch-based LLDs and their four delta coefficient contours (152 features), where 19 functions were selected from the 21 functions mentioned by removing the minimum value and the range functions. Finally the number of pitch onsets (pseudo syllables) and the total duration of the input were appended (2 features). Table 1 gives an overview of the low-level descriptors and associated feature statistics functions. The details of each item can be seen in Schuller et al. (2010).

Most of these features have been verified to be useful for depression classification (Moore et al., 2008; Low et al., 2011; Ooi et al., 2013; Alghowinem et al., 2013a; Valstar et al., 2014). These features were then normalized to a range of [0, 1]. Principal component analysis (PCA) was applied to reduce feature space dimensionality.

4.3. Modeling and classification

Considering the development of gender differences in depressive symptoms (Nolenhoeksema and Girgus, 1994), there are two classification techniques: gender-dependent modeling (GDM) and gender-independent modeling (GIM). Low et al. (2011) found that GDM performed better than GIM. In our research, we employed GDM, in which males and females were modeled separately. To examine whether the classification accuracy was biased with respect to classifiers, three different popular methods were compared in this work.

KNN is an algorithm that stores all available cases and classifies new cases based on a similarity measure. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its K nearest neighbors. In this research, Euclidean distance was calculated for the testing samples and their neighbors. The value of K was selected by 5-fold cross-validation on the training samples.

GMM is defined as the weighted sum of multiple Gaussian components that represent a density of a particular random variable. Mathematical formulation of GMM is given by:

$$p(x|\Theta) = \sum_{j=1}^{M} w_j p(x|j) \quad (1)$$

where x is the feature vector, M is the maximum number of Gaussian components, and Θ represents the Gaussian mixture model parameters that include mean vector (μ), co-variance Matrix (Σ), and weight (w) that satisfies $\sum_{i=1}^{M} w_i = 1$. $p(x|j)$ is the density of Gaussian component j given by:

$$p(x|j) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_j|^{\frac{1}{2}}} e^{\frac{-1}{2}(x-\mu_j)^T \sum_i^{-1}(x-\mu_j)} \tag{2}$$

where D is the dimension of x. In the implementation, the expectation-maximization (EM) algorithm was used for estimating the Gaussian mixture model parameters Θ of each Gaussian component. For computational efficiency, diagonal covariance matrices were used in the Gaussian component instead of full covariance.

SVM aims to construct an optimal hyperplane that has the largest distance to the nearest training-data point of two classes while minimizing the training error. It can be represented as the following optimization problem:

$$\text{minimize}: \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \tag{3}$$

$$\text{subject to}: y_i(w \cdot x_i + b) + \xi_i \geq 1,$$
$$\xi_i \geq 0, i = 1, 2, \ldots, N \tag{4}$$

where w is the vector normal to the hyperplane, b is a scalar bias, and C is a constant that penalizes the training errors and controls the tradeoff between margin maximization and error minimization. We use a radial basis function (RBF) $\exp(-\gamma\|u-v\|^2)$ as SVM's kernel function, where u and v are feature vectors. In this study, searching for the most appropriate (γ, C) pair was performed through a grid search using 5-fold cross validation on the training dataset using the LIBSVM toolbox (Chang and Lin, 2011).

4.4. New methodology for the classification of depression

In order to increase classification accuracy, a combination of classifiers can be applied. In this study, each participant was represented by 29 speech recordings. Alternatively, we could have developed 29 individual classifiers, and then, when each of them voted, a class label prediction would have been returned by the ensemble based on the collection of votes. However, this method needed to train too many classifiers. Furthermore, the speech recordings of INT, PIC, REA, POS, NEU, and NEG were unbalanced. If the discriminative power of each of these different speech types and emotions was significantly different from the others, the classification results could be unstable. In order to overcome these challenges, a new ensemble methodology based on speech types and emotions was developed for detecting depression (STEDD). A weighted decision fusion process was adopted in this methodology. Fig. 2 provides an overview of the proposed methodology.

At the first stage, as shown in Fig. 2(a), the training speech recordings a_m ($m = 1, \ldots, 29$) of subjects x_k ($k = 1, \ldots, N$) were grouped into INT(S_{1j}, $j = 1, \ldots, 18$), PIC(S_{2j}, $j = 1, \ldots, 4$), REA(S_{3j}, $j = 1, \ldots, 7$), POS(S_{4j}, $j = 1, \ldots, 9$), NEU(S_{5j}, $j = 1, \ldots, 10$), and NEG(S_{6j}, $j = 1, \ldots, 10$). Six base classifiers were trained using the speech of S_i ($i = 1, \ldots, 6$). Each of the classifiers C_i ($i = 1, \ldots, 6$) was developed using the framework shown in Fig. 1. The class estimate given by C_i was $y_i(x_k)$, and a majority vote was utilized to calculate the values of $y_i(x_k)$. $y_i(x_k)$ was set to equal $+1$ for the depressed subjects and -1 for the healthy controls. The value of the weight coefficient W_i of each C_i was calculated as follows:

$$g_i(x_k) = \frac{1}{N}\sum_{k=1}^{N}(y_i(x_k) - f_i(x_k))^2 \tag{5}$$

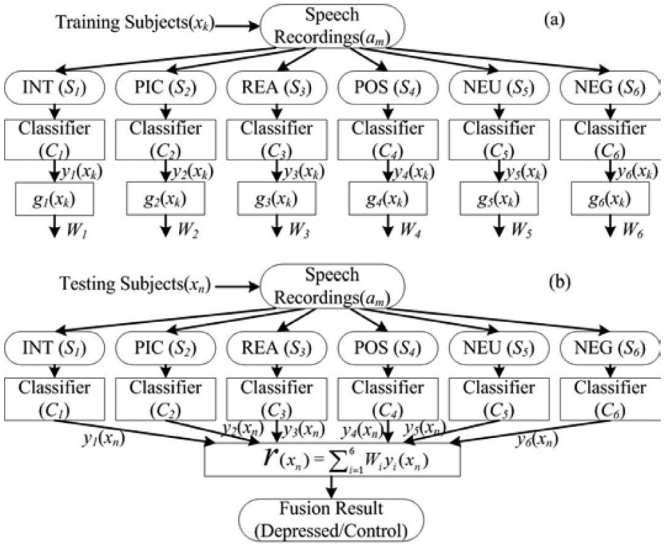

Fig. 2. Overview of the proposed methodology: (a) is the training process and determines the weight values; (b) is the testing process.

$$W_i = \frac{\frac{1}{g_i(x_k)}}{\sum_{i=1}^{6}\left(\frac{1}{g_i(x_k)}\right)} \tag{6}$$

$$\sum_{i=1}^{6} W_i = 1 \tag{7}$$

where N is the total number of the training samples, and $f_i(x_k)$ is the actual class of subject(x_k). $f_i(x_k)$ was set to equal $+1$ for the depressed subjects and -1 for the healthy controls. The function $g_i(x_k)$ defined in (5) represents a total average squared error between the actual classes and the classes estimated by C_i. This process of weight calculation assigned higher weight values to those C_i's that provided better classification results, and smaller values to those that provided lesser performance. The additional advantage given by this approach was that the relative performance of different speech types and emotions can be observed.

After finding the values of the weight coefficients, the performance of this new STEDD methodology for classification was tested. First, as shown in Fig. 2(b), testing speech samples of subjects x_n, were classified by each of the six classifiers. Each classifier produced its own class estimates $y_i(x_n)$. Next, the class estimates given by the classifiers were combined into a weighted score parameter $r(x_k)$ given as follows:

$$r(x_n) = \sum_{i=1}^{6} W_i y_i(x_n) \tag{8}$$

Then the final classification decision was made based on the sign of $r(x_n)$. If the value of $r(x_n)$ was greater than 0, subject x_n was classified as depressed; otherwise, x_n was classified as a control.

For the purpose of comparison, a methodology using an unweighted decision fusion process (UDD) was also tested, in which the values of W_i ($i = 1, \ldots, 6$) were set to be the same, and other processes were the same as STEDD as well.

5. Experiments and results

Experiments with the framework outlined in Section 4 were carried out using the database described in Section 3. In this study, the correct classifications of depressed patients and health controls were measured in terms of sensitivity, specificity, and accu-

Table 2
Classification results from the speech of INT, PIC, and REA for males.

Classifiers	Sensitivity %		Specificity %		Accuracy %	
+Types	mean	st.dev.	mean	st.dev.	mean	st.dev.
KNN+INT	60.07	10.18	63.73	13.37	61.95	8.05
KNN+PIC	63.28	8.94	67.65	8.58	65.53	7.90
KNN+REA	42.86	12.58	73.11	6.18	58.44	6.47
GMM+INT	61.81	9.41	59.15	9.19	60.44	7.92
GMM+PIC	64.06	5.63	70.59	7.50	67.42	6.29
GMM+REA	57.14	12.58	60.08	6.83	58.66	7.19
SVM+INT	61.28	8.79	69.93	8.80	65.74	5.49
SVM+PIC	74.22	3.41	67.65	4.65	70.83	3.45
SVM+REA	49.55	8.75	75.63	7.97	62.99	5.11

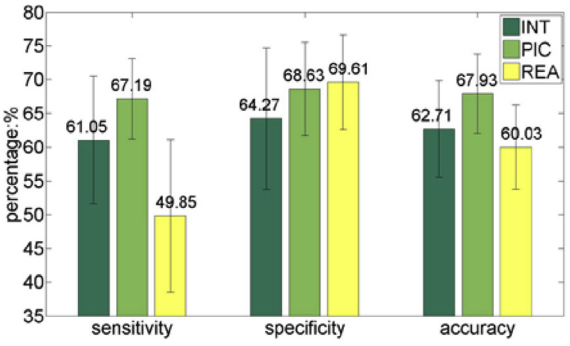

Fig. 3. The mean and st.dev. of sensitivity, specificity, and accuracy of the INTs, PICs, and REAs for males.

Table 3
Classification results from the speech of POS, NEU, and NEG for males.

Classifiers	Sensitivity %		Specificity %		Accuracy %	
+Emotions	mean	st.dev.	mean	st.dev.	mean	st.dev.
KNN+POS	58.33	13.18	68.95	11.44	63.80	9.38
KNN+NEU	55.00	14.87	64.12	14.99	59.70	8.41
KNN+NEG	55.94	10.87	66.76	8.32	61.52	5.12
GMM+POS	60.42	6.07	64.38	8.48	62.46	6.50
GMM+NEU	63.44	13.48	60.88	11.16	62.12	10.34
GMM+NEG	59.06	8.55	57.94	6.45	58.49	5.60
SVM+POS	61.46	9.32	68.95	6.95	65.32	5.22
SVM+NEU	56.88	12.33	73.82	11.35	65.61	6.32
SVM+NEG	62.50	10.46	70.00	5.39	66.36	5.28

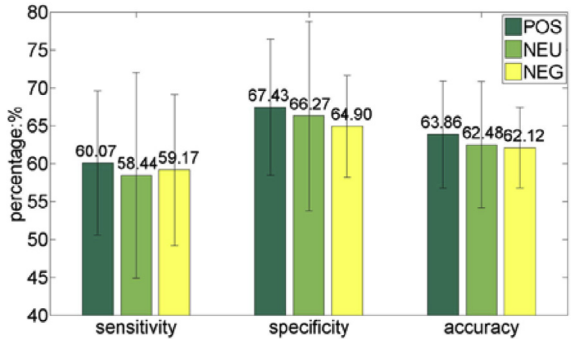

Fig. 4. The mean and st.dev. of sensitivity, specificity, and accuracy of the POSs, NEUs, and NEGs for males.

racy. The depressed patients were considered as the positive cases, and the healthy controls were considered as the negative cases.

When assessing the performance, a well-performing system would have high values for all of these three parameters. However, if a compromise had to be made, it was desirable to achieve the highest overall accuracy by obtaining an optimal sensitivity to specificity ratio (ideally >1). In this study, we employed a 10-fold cross-validation and speaker-independent split of train and test data. The mean and standard deviation (st.dev.) of sensitivity, specificity, and accuracy were calculated in order to establish an effective and robust classification method. The one-way analysis of variance (ANOVA) followed by the least significant difference (LSD) test was carried out to determine whether the differences in the discriminative power of different speech types and emotions were statistically significant. The significance level was defined as $p<0.05$.

5.1. Experiment for males using KNN, GMM and SVM (EXP1)

Table 2 shows the classification results from the speech of INT, PIC, and REA for males using KNN, GMM, and SVM classifiers. Fig. 3 shows the mean and st.dev. of sensitivity, specificity, and accuracy of the INTs (KNN+INT, GMM+INT and SVM+INT), PICs (KNN+PIC, GMM+PIC and SVM+PIC), and REAs (KNN+REA, GMM+REA and SVM+REA) for males. ANOVA and LSD tests were conducted on paired speech types over the 10-fold cross-validation results using KNN, GMM, and SVM classifiers. The specificity of INT, PIC, and REA were similar ($p>0.05$). The sensitivity and accuracy were significantly different among the three speech types ($p=0.001$ and $p=0.015$). After the LSD test, from the accuracy and sensitivity in Fig. 3, it can be observed that PIC improved accuracy over INT and REA (67.93% vs. 62.71% vs. 60.03%, $p=0.029$ and $p=0.004$). It can also be observed that the sensitivity of REA was worse than the sensitivity of INT and PIC (49.85% vs. 61.05% vs. 67.19%, $p=0.001$ and $p=0.001$). The classification results of other paired speech types were similar ($p>0.05$). Moreover, the PICs showed a decrease in standard deviation of sensitivity over the INTs and REAs of 3.47% and 5.31%, and resulted in a 1.27% and 0.38% decrease in standard deviation of accuracy over the INTs and REAs, respectively.

Based on these results, it was found that PIC performed best among these three speech types, while REA performed worst. This finding might have resulted because PIC evoked situations more likely to elicit greater cognitive effort and produced more pronounced changes in speech acoustics for identifying depression for males. Both INT and PIC can be considered as spontaneous speech, making this result consistent with the findings of Alghowinem et al. (2013a) and Gupta et al. (2014) in demonstrating that spontaneous speech gave better results than reading. However, different from previous work, PIC was separated out in this study. In future research, in order to get better performance, it would be valuable to collect more PIC data for males.

Table 3 shows the classification results from the speech of POS, NEU, and NEG for males using KNN, GMM, and SVM classifiers. Fig. 4 shows the mean and st.dev. of sensitivity, specificity and accuracy of the POSs, NEUs, and NEGs for males. From Table 3, it was observed that KNN+POS outperformed than KNN+NEU and KNN+NEG in sensitivity, specificity and accuracy; GMM+NEU performed better than GMM+POS and GMM+NEG in sensitivity; and SVM+NEG yielded good results in sensitivity and accuracy. No one could outperform others consistently between the POSs, NEUs, and NEGs. Furthermore, it was observed that POS provided very similar results as compared with NEU and NEG in Fig. 4. After ANOVA and LSD tests were conducted on paired speech emotions over the 10-fold cross-validation results, we found that the classification results (accuracy, sensitivity, and specificity) were similar among the three speech emotions ($p>0.05$). Based on these findings, it was more likely that POS, NEU, and NEG had almost the same classification results for males. On the other hand, there were no statistically significant differences while expressing different emotions between depressed and healthy males. This result contrasts

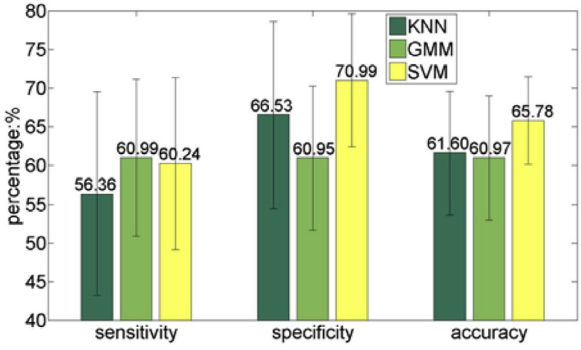

Fig. 5. KNN, GMM, and SVM classification results for males.

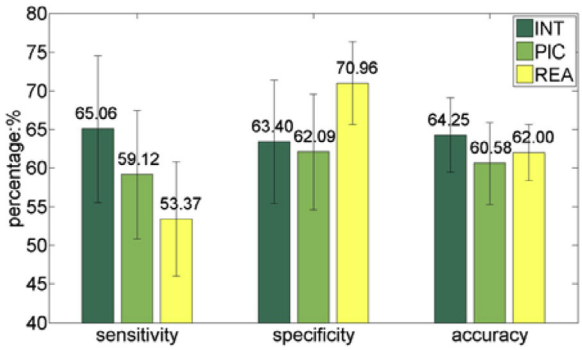

Fig. 6. The mean and st.dev. of sensitivity, specificity, and accuracy of the INTs, PICs, and REAs for females.

Table 4
Classification results from the speech of INT, PIC, and REA for females.

Classifiers	Sensitivity %		Specificity %		Accuracy %	
+Types	mean	st.dev.	mean	st.dev.	mean	st.dev.
KNN+INT	62.79	9.24	60.68	9.16	61.75	4.88
KNN+PIC	53.30	11.96	67.65	4.49	60.33	7.76
KNN+REA	44.74	7.53	79.27	2.94	61.68	4.25
GMM+INT	66.35	10.53	60.89	8.90	63.68	4.86
GMM+PIC	63.68	8.16	56.38	7.25	60.10	3.50
GMM+REA	62.00	6.41	58.54	5.69	60.30	4.38
SVM+INT	66.04	8.76	68.63	5.99	67.31	4.76
SVM+PIC	60.38	4.81	62.25	10.68	61.30	4.68
SVM+REA	53.37	8.18	75.07	7.39	64.01	2.17

Table 5
Classification results from the speech of POS, NEU, and NEG for females.

Classifiers	Sensitivity %		Specificity %		Accuracy %	
+Emotions	mean	st.dev.	mean	st.dev.	mean	st.dev.
KNN+POS	54.72	8.20	68.19	6.05	61.32	3.07
KNN+NEU	60.19	13.41	66.86	10.61	63.46	5.18
KNN+NEG	56.23	12.87	63.53	13.56	59.81	6.18
GMM+POS	63.73	9.57	57.52	8.77	60.68	3.22
GMM+NEU	68.30	9.91	57.45	8.18	62.98	5.98
GMM+NEG	62.64	8.17	63.92	5.56	63.27	4.49
SVM+POS	58.07	6.88	67.32	7.79	62.61	4.31
SVM+NEU	64.72	8.93	70.59	5.26	67.60	3.80
SVM+NEG	63.40	11.45	69.80	10.23	66.54	4.82

with the findings of two previous researchers. Alghowinem et al. (2012) observed that when subjects talked about positive emotions in the interview, the result was higher correct recognition of depression, and Goeleven et al. (2006) noted that depressed patients showed a specific failure to impair inhibitions relating to negative information. The disparity of these prior results with ours may be attributed to the fact that the tasks used in the previous research were not the same as ours, and different classifiers were employed. Moreover, different from previous research, in this study, we adopted and compared three classifiers, and we modeled males and females separately.

The KNN, GMM, and SVM classification results from speech of all types and emotions for males are shown in Fig. 5. The statistical significance analysis was also conducted on paired classifiers over the 10-fold cross-validation results. For males, the sensitivity of KNN, GMM, and SVM were similar ($p > 0.05$). The specificity and accuracy were significantly different among the three classifiers ($p = 0.002$ and $p = 0.032$). From the specificity and accuracy in Fig. 5, SVM resulted in an accuracy improvement over KNN and GMM (65.78% vs. 61.60% vs. 60.97%, $p = 0.035$ and $p = 0.016$). The specificity of GMM was worse than the specificity of SVM and KNN (60.95% vs. 70.99% vs. 66.53%, $p = 0.001$ and $p = 0.042$). The classification results of other paired classifiers were similar ($p > 0.05$). Moreover, SVM showed a decrease in standard deviation of specificity over GMM and KNN of 0.71% and 3.49%, and also resulted in a 2.34% and 2.31% decrease in standard deviation of accuracy over GMM and KNN, respectively. The results of the comparison of classifiers showed that SVM outperformed the other classifiers and GMM performed worst for males in this study.

5.2. Experiment for females using KNN,GMM and SVM (EXP2)

Table 4 shows the classification results from the speech of INT, PIC, and REA for females using KNN, GMM, and SVM classifiers. Fig. 6 shows the mean and st.dev. of sensitivity, specificity, and accuracy of the INTs, PICs, and REAs for females. ANOVA and LSD tests were conducted on paired speech types. For females, the sensitivity, specificity, and accuracy were significantly different among the three speech types ($p = 0.001$, $p = 0.006$, and $p = 0.047$). It can be observed in Fig. 6 that INT improved sensitivity over REA (65.06% vs. 53.37%, $p = 0.001$), and gave a classification accuracy increase compared to PIC (64.25% vs. 60.58%, $p = 0.030$). The specificity of REA was better than the specificity of PIC and INT (70.96% vs. 62.09% vs. 63.40%, $p = 0.012$ and $p = 0.003$). The classification results of other paired speech types were similar ($p > 0.05$). It was found that INT performed better than PIC for females, which contrasted with the result in EXP1. This finding might be attributed to INT producing more changes in speech acoustics than PIC when detecting depression for females. It can be observed that REA showed a sensitivity/specificity ratio of 53.37%/70.96%, and INT showed a sensitivity/specificity ratio of 65.06%/63.40%. Meanwhile, the accuracy was found to be similar between INT and REA ($p > 0.05$). As the objective of our experiment was to identify more depressed patients (higher sensitivity) rather than to screen out healthy controls (higher specificity), INT also provided better performance than REA for females in this study.

Table 5 shows classification results from the speech of POS, NEU, and NEG for females using KNN, GMM, and SVM classifiers. Fig. 7 shows the mean and st.dev. of sensitivity, specificity and accuracy of the POSs, NEUs, and NEGs for females. From Table 5, it can be observed that the NEUs performed better than the POSs and NEGs. It can be observed in Fig. 7 that NEU improved sensitivity over POS and NEG, and increased classification accuracy compared to POS and NEG. However, ANOVA and LSD were also conducted on paired speech emotions. It was found that the classification results (accuracy, sensitivity, and specificity) were similar among the three speech emotions ($p > 0.05$). Based on these findings, it was more likely that POS, NEU, and NEG had almost the same classification results for females, which was consistent with the findings for males in EXP1.

The KNN, GMM, and SVM classification results from speech of all types and emotions for females are shown in Fig. 8. The statistical significance analysis was also conducted on paired classifiers.

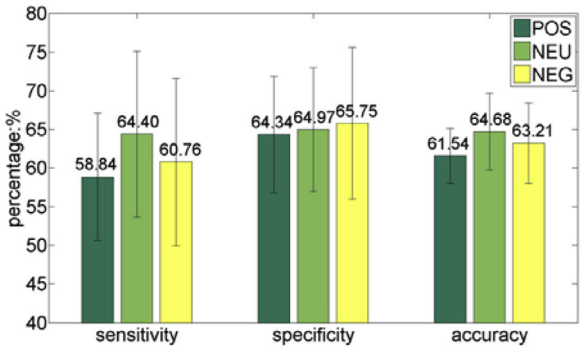

Fig. 7. The mean and st.dev. of sensitivity, specificity, and accuracy of the POSs, NEUs, and NEGs for females.

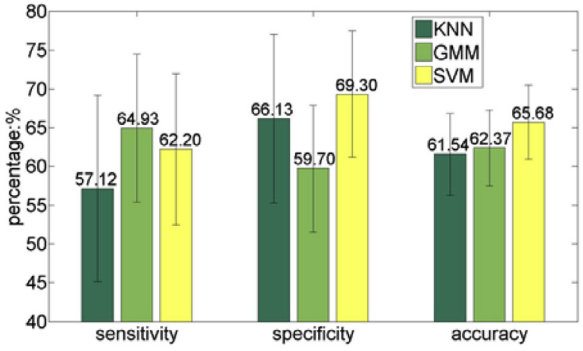

Fig. 8. KNN, GMM, and SVM classification results for females.

Table 6
Weight values for STEDD.

Gender	INT	PIC	REA	POS	NEU	NEG
Male	0.1474	0.1751	0.1273	0.1751	0.1751	0.2001
Female	0.1937	0.1326	0.1292	0.1481	0.2099	0.1866

The sensitivity, specificity and accuracy were significantly different among the three classifiers ($p = 0.023$, $p = 0.001$, and $p = 0.006$). From the specificity and accuracy in this figure, SVM resulted in an accuracy improvement over KNN and GMM, respectively (65.68% vs. 61.54% vs. 62.37%, $p = 0.003$ and $p = 0.015$). The specificity of GMM was worse than the specificity of SVM and KNN (59.70% vs. 69.30% vs. 66.13%, $p = 0.001$ and $p = 0.010$). The sensitivity was found to be significantly different between GMM and KNN (64.93% vs. 57.12%, $p = 0.007$). The classification results of other paired classifiers were similar ($p > 0.05$). The results of the comparison of classifiers showed that SVM was more effective than the other classifiers, which was consistent with the findings in EXP1.

5.3. Experiment for the proposed methodology (EXP3)

An experiment with the framework outlined in Section 4 (see Fig. 2) was carried out. In EXP1 and EXP2, SVM showed the best overall performance. Therefore, SVM was employed for modeling in this phase of our experiments. At the first stage, each base classifier provided an individual classification result. At the second stage, the outputs from the six base classifiers were combined into the final decision by calculating a weighted sum of the intermediate decisions generated by each base classifier. The weights were calculated using (5)–(7).

Table 6 shows the resulting weights averaged over 10-fold cross-validation. The weights in this table reflect contributions of different speech types and emotions into the formation of the final classification decision based on the weighted score parameter given in (8). In the case of speech types, the order of weights (from the highest to the lowest) was PIC, INT, and REA for males. This indicated that the features from PIC showed the highest correlation with the depression classification for males, which was consistent with the findings in EXP1. For females, the order of weights was INT, PIC, and REA. This indicated that the features from INT were more highly correlated with the depression classification for females, which was consistent with the findings in EXP2. In the case of speech emotions, NEG had the greatest weight for males, while POS and REA had the same weight. For females, the order of weights was NEU, NEG, and POS. The differences between the weights of speech emotions were smaller than the differences between the weights of speech types.

Table 7
Classification results of UDD and STEDD.

Classifier	Gender	Sensitivity %	Specificity %	Accuracy %
UDD	Male	75.00	79.41	77.27
	Female	73.58	64.71	69.23
STEDD	Male	75.00	85.29	80.30
	Female	77.36	74.51	75.96

The final classification results are presented in Table 7. From this table, it was observed that STEDD provided very promising results, and provided better performance than UDD in terms of classification accuracy, sensitivity, and specificity both for males and females. This can be explained by the fact that STEDD provided higher weights compared to the base classifiers that provided better classification results. It also can be observed that the classification results of STEDD were better than all of the results in EXP1 and EXP2. This can be attributed to the general ability of an ensemble of classifiers to provide better performance than a single learner. Therefore, it can be concluded that STEDD was effective for detecting depression.

6. Conclusion

In this study, first we compared the classification results using speech of different types. It was observed that using picture description speech provided significantly better ($p < 0.05$) classification results for males than using interview and reading speech. It was found that using interview speech gave significantly better ($p < 0.05$) classification results for females than using picture description and reading speech. It was also noted that classification using speech associated with positive, neutral, or negative emotions had similar performances ($p > 0.05$) for both males and females. Based on this research, the collection of more data for males and females should be targeted further in future research. Compared to the classifiers GMM and KNN, SVM showed the highest classification result and the best stability for both males and females.

In the second part of our research, we presented a new ensemble methodology for the classification of depression. The novel aspects of this methodology were that males and females were modeled separately, and different weights were provided for different speech types and emotions according to their respective contributions in detecting depression. This new approach provided very promising results, showing a high accuracy level of 80.30% for males and 75.96% for females, plus a desirable sensitivity/specificity ratio of 75.00%/85.29% for males and 77.36%/74.51% for females.

Although the presented results show encouraging trends, a potential limitation of this study is that depressed speech may contain other features that are connected with different speech types and emotions. Future research will involve the collection of more

data, as well as improvements for the feature extraction and selection strategy.

Acknowledgment

This work was supported by the National Basic Research Program of China (973 Program) (No.2014CB744600).

References

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G., 2012. From joyous to clinically depressed: mood detection using spontaneous speech. In: Proceedings of FLAIRS-25. Marco Island, Florida, pp. 141–146.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G., 2013a. Detecting depression: a comparison between spontaneous and read speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada, pp. 7547–7551.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Gedeon, T., Breakspear, M., 2013b. A comparative study of different classifiers for detecting depression from spontaneous speech. In: Proceedings of ICASSP. Vancouver, Canada, pp. 8022–8026.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, 4th ed. American Psychiatric Association, Washington, DC.

Calev, A., Nigal, D., Chazan, S., 1989. Retrieval from semantic memory using meaningful and meaningless constructs by depressed, stable bipolar and manic patients. British J. Clin. Psychol. 28, 67–73.

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 1–27.

Cohn, J.F., Kruez, T.S., Matthews, I., Yang, Y., Nguyen, M.H., Padilla, M.T., Zhou, F., Torre, F.D., 2009. Detecting depression from facial actions and vocal prosody. In: Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. IEEE, pp. 1–7.

Cummins, N., Epps, J., Breakspear, M., Goecke, R., 2011. An investigation of depressed speech detection: features and normalization. In: Proceedings of Interspeech. ISCA, Florence, Italy, pp. 2997–3000.

Cummins, N., Epps, J., Sethu, V., Krajewski, J., 2014. Variability compensation in small data: oversampled extraction of i-vectors for the classification of depressed speech. In: Proceedings of ICASSP. Florence, Italy. IEEE, pp. 970–974.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F., 2015. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 71, 10–49.

Darby, J.K., Hollien, H., 1977. Vocal and speech patterns of depressivepatients. Folia Phoniatrica 29, 279–291.

Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile-The Munich versatile and fast open-source audio feature extractor. In: Proceedings of the ACM Multimedia International Conference(MM). Firenze, Italy, pp. 1459–1462.

France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, M., 2000. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Bio Eng. 47, 829–837.

Goeleven, E., Raedt, R.D., Baert, S., Koster, E.H.W., 2006. Deficient inhibition of emotional information in depression. J. Affect. Disorders 93, 149–157.

Gollan, J.K., Pane, H.T., McCloskey, M.S., Coccaro, E.F., 2008. Identifying differences in biased affective information processing in major depression. Psychiatry Res. 159, 18–24.

Gupta, R., Malandrakis, N., Xiao, B., Guha, T., Segbroeck, M.V., Black, M., Potamianos, A., Narayanan, S., 2014. Multimodal prediction of affective dimensions and depression in human–computer interactions. In: Proceedings of AVEC '14. Orlando, Florida, USA. ACM, pp. 33–40.

Hawton, K., Casanas i Comabella, C., Haw, C., Saunders, K., 2013. Risk factors for suicide in individuals with depression: a systematic review. J. Affect. Disorders 147, 17–28.

He, L., Lech, M., Maddage, N.C., Allen, N.B., 2011. Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. Biomed. Sig. Process. Control 6, 139–146.

Helfer, B.S., Quatieri, T.F., Williamson, J.R., Mehta, D.D., Horwitz, R., Yu, B., 2013. Classification of depression state based on articulatory precision. In: Proceedings of Interspeech. Lyon, France. ISCA, pp. 2172–2176.

Horwitz, R., Quatieri, T.F., Helfer, B.S., Yu, B., Williamson, J.R., Mundt, J., 2013. On the relative importance of vocal source, system, and prosody in human depression. In: Proceedings of the IEEE International Conference on Body Sensor Networks. Cambridge, MA, USA, pp. 1–6.

Hönig, F., Batliner, A., Nöth, E., Schnieder, S., Krajewski, J., 2014. Automatic modelling of depressed speech: relevant features and relevance of gender. In: Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association. Singapore, pp. 1248–1252.

Joshi, J., Goecke, R., Alghowinem, S., Dhall, A., Wagner, M., Epps, J., Parker, G., Breakspear, M., 2013. Multimodal assistive technologies for depression diagnosis and monitoring. J. Multimodal User Interf. 7, 217–228.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K.R., Rush, A.J., Walters, E.E., Wang, P.S., 2003. The epidemiology of major depressive disorder: results from the national comorbidity survey replication (NCS-R). Jama J. Am. Med. Assoc. 289, 3095–3105.

Kroencke, K., Spitzer, R., Williams, J., 2001. The phq-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16, 606–613.

Leyman, L., Raedt, R.D., Schacht, R., Koster, E.H., 2007. Attentional biases for angry faces in unipolar depression. Psychol. Med. 37, 393–402.

Liu, Z.Y., Hu, B., Yan, L.H., Wang, T.Y., Liu, F., Li, X.Y., Kang, H.Y., 2015. Detection of depression in speech. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII). Xian, China, pp. 743–747.

Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., 2015. Assessing speaker independence on a speech-based depression level estimation system. Pattern Recog. Lett. 68, 343–350.

Low, L.A., Maddage, N.C., Lech, M., Sheeber, L.B., Allen, N.B., 2010. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In: Proceedings of ICASSP. Dallas, Texas. IEEE, pp. 5154–5157.

Low, L.A., Maddage, N.C., Lech, M., Sheeber, L.B., Allen, N.B., 2011. Detection of clinical depression in adolescents speech during family interactions. IEEE Trans. Bio Eng. 58, 574–586.

Mantri, S., Agrawal, P., Dorle, S.S., Patil, D., Wadhai, V.M., 2013. Clinical depression analysis using speech features. In: Proceedings of the 6th International Conference on Emerging Trends in Engineering and Technology (ICETET). Nagpur, India, pp. 111–112.

Mitra, V., Shriberg, E., 2014. Effects of feature type, learning algorithm and speaking style for depression detection from speech. In: Proceedings of ICASSP. Florence, Italy. IEEE, pp. 4774–4778.

Moore, E., Clements, M., Peifer, J.W., Weisser, L., 2008. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. IEEE Trans. Biomed. Eng. 55, 96–107.

Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., Geralts, D.S., 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response(IVR) technology. J. Neurolinguistics 20, 50–64.

Muthusamy, H., Polat, K., Yaacob, S., 2015. Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals. Math. Prob. Eng. 2015, 1–13.

Nolenhoeksema, S., Girgus, J.S., 1994. The emergence of gender differences in depression during adolescence. Psychol. Bull. 115, 424–443.

Ooi, K.E.B., Lech, M., Allen, N.B., 2013. Multichannel weighted speech classification system for prediction of major depression in adolescents. IEEE Trans. Bio Eng. 60, 497–506.

Ooi, K.E.B., Lech, M., Allen, N.B., 2014. Prediction of major depression in adolescents using an optimized multi-channel weighted speech classification system. Biomed. Signal Process. Control 14, 228–239.

Ozdas, A., Shiavi, R.G., Silverman, S.E., Silverman, M.K., Wilkes, D.M., 2004. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. IEEE Trans. Bio Eng. 51, 1530–1540.

Pao, T.L., Chen, Y.T., Yeh, J.H., 2008. Emotion recognition and evaluation from Mandarin speech signals. Int. J. Innov. Comput. Inf. Control 4, 1695–1709.

Quatieri, T.F., Malyska, N., 2012. Vocal-source biomarkers for depression: a link to psychomotor activity. In: Proceedings of Interspeech. Portland, USA. ICSA, pp. 1059–1062.

Schuller, B., Steidl, S., Batliner, A., 2010. The INTERSPEECH 2010 paralinguistic challenge. In: Proceedings of Interspeech. Makuhari, Japan, pp. 2794–2797.

Sidorov, M., Minker, W., 2014. Emotion recognition and depression diagnosis by acoustic and visual features: a multimodal approach. In: Proceedings of AVEC '14. Orlando, Florida, USA. ACM, pp. 81–86.

Sobin, C., Sackeim, H.A., 1997. Psychomotor symptoms of depression. Am. J. Psychiatry 154, 4–17.

Vanger, P., Summerfield, A.B., Rosen, B.K., Watson, J.P., 1992. Effects of communication content on speech behavior of depressives. Comprehensive Psychiatry 33, 39–41.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M., 2014. 3D dimensional affect and depression recognition challenge. In: Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14). Orlando, Florida, USA. ACM, pp. 3–10.

Williamson, J., Quatieri, T., Helfer, B., Ciccarelli, G., Mehta, D.D., 2014. Vocal and facial biomarkers of depression based on motor incoordination and timing. In: Proceedings of AVEC '14. Orlando, Florida, USA. ACM, pp. 65–72.

World Health Organization, 2016. Depression Fact Sheet Reviewed April 2016 http://www.who.int/mediacentre/factsheets/fs369/en/.