Detecting Postpartum Depression in Depressed People by Speech Features

Jingying Wang[1,2], Xiaoyun Sui[1], Bin Hu[3], Jonathan Flint[4(✉)], Shuotian Bai[5], Yuanbo Gao[2], Yang Zhou[1,2], and Tingshao Zhu[1(✉)]

[1] Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
{wangjingying, tszhu}@psych.ac.cn, oswicer@163.com
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Information Science and Engineering,
Lanzhou University, Gansu 730000, China
bh@lzu.edu.cn
[4] Department of Psychiatry and Biobehavioral Sciences,
UCLA David Geffen School of Medicine, Los Angeles, CA 90095, USA
jf@well.ox.ac.uk
[5] School of Information Engineering,
Hubei University of Economics, Wuhan 430205, China

Abstract. Postpartum depression (PPD) is a depressive disorder with peripartum onset, which brings heavy burden to individuals and their families. In this paper, we propose to detect PPD in depressed people via voices. We used openSMILE for feature extraction, selected Sequential Floating Forward Selection (SFFS) algorithm for feature selection, tried different settings of features, set 5-fold cross validation and applied Support Vector Machine (SVM) on Weka for training and testing different models. The best predictive performance among our models is 69%, which suggests that the speech features could be used as a potential behavioral indicator for identifying PPD in depression. We also found that a combined impact of features and content of questions contribute to the prediction. After dimension reduction, the average value of F-measure was increased 5.2%, and the precision of PPD was rose to 75%. Comparing with demographic questions, the features of emotional induction questions have better predictive effects.

Keywords: Postpartum depression · Depression · Speech features
Detecting · Classification

1 Introduction

Postpartum depression (PPD) is a kind of depressive disorder with peripartum onset, which can affect both genders after childbirth, and females usually suffer worse than males [1]. It is a heavy burden to not only patients themselves, but also their spouses, children and whole families [2]. The concept "PPD" was proposed by Pitt. B in 1968 [1], but there is still no agreement on its diagnostic criteria until now. PPD is one subtype of depressive disorder. It is liable to cause misdiagnosis between PPD and other subtypes of depressive disorder [3]. Accurate diagnosis is the critical for effective

Q. Zu and B. Hu (Eds.): HCC 2017, LNCS 10745, pp. 433–442, 2018.
https://doi.org/10.1007/978-3-319-74521-3_46

intervention. In the case of inconsistent diagnosis, it is necessary to develop new methods to aid PPD diagnosis.

Depressive disorder has a visible influence on patients' emotions [4]. The influence of emotion on people could reflect in their voices [5]. The study of Cannizzaro [6] shown that depressed people had less verbal production and variations comparing with healthy people. Sobin and Sackeim [7] summarized some specific speech features in depression, including slow speech speed, increased pause duration, and so on. As a subtype of depressive disorder, the influence of PPD on patients should reflect on their voices as well.

The diagnosis of PPD by using voice is feasible. First of all, speech-based diagnosis did not disturb patients too much. The procedure can be set during interviewing. Patients need not finish any complex or time-consuming tasks. Secondly, the check method is easier to hide. Behavioral features are easy to become untrue, because people are able to control their behaviors [7]. The method of speech check can avoid to directly contact with patients, which is benefit for data' ecological validity.

We proposed to detect PPD from depressed patients by using speech features. The purpose of this study is to investigate the effect of detecting PPD within depression via speech features under the state of natural experiment. The voices we used were originated from the conversation between patients and doctors recorded during interviewing. Patients' voices were separated from these recordings, and divided into two group: PPD and non-PPD. Speech features extracted by OpenSMILE, and 988 features were extracted in all. We used all speech features and features after dimension reduction to predict PPD, respectively. The predictive effects of speech features were evaluated in the light of three indexes: precision, recall and F-measure.

2 Related Work

Voice is one way of emotional expression. Speech features have been found to be able to identify different emotions. Nwe et al. [8] reported that classifying voice as different emotions based on HMM (hidden markov model) had a higher accuracy rate (average 7.7%) than artificial judging, the average rate was up to 78%. Wu et al. [9] used prosodic and spectral features to identify seven emotions, with the best precision as 91.6%. From the above studies, we know emotional can be predicted based on speech features. It motivates us to identify mental illness like PPD using speech features. There have been few studies about PPD patients' voices. We think those studies about phonetic changes of depression probably can be generalized to PPD.

The sounds of depressed patients have significant changes because of their illness [6]. The diagnostic speech features of depression in DSM-5 are described as slow, volume sank, variation of tone lessen, pause duration increase [4] (P163). Experiments revealed some specific vocal indicators in depression, such as speech speed slow down [10], increased pause duration and times [10, 11], shortened duration of utterance [12], longer initiative time latency [13]. Speech features in depression express with changes of F0 such as the decreasing of bandwidth, amplitude, energy [10, 14], shrunken F0 range (ΔF0) [14], weakened intensity [15], variation of frequency spectrum like shrunken second formant transition [12] and shrunken spectral energy distribution [16], and so on.

Reviewing recent findings, Cohen and Elvevåg [17] believed that computer-based assessments of natural language has the potential for measuring speech disturbances in people with severe mental illnesses. Some researchers attempted to predict depression via patients' speech. Mundt et al. [11] stated that the regression model consisted of F0, pause duration, speech speed, speech duration, etc. could predict depression, and the explanatory power of model to depression reaching 79.2%. Emerging evidence suggested that speech features have a strong performance in predicting depression, which obtained a RMSE of 10.17 well below the baseline of 14.12 [18]. The study of Cohn et al. shown that the accuracy in detecting depression was 79% for vocal prosody [19].

In our study, we choose the depressed voices collected from natural circumstance to improve ecological validity, differing from the controlled experimental environments in the previous studies. In clinical diagnosis, it is more crucial and harder to make a distinction between different mental illnesses than distinguish healthy people from psychiatric patients. To improve the differential diagnosis among depressive spectrum disorders, our detective aim is detecting PPD within depression.

3 Methods

3.1 Participants

In this study, patients' voices were secondary data which acquired from CONVERGE (China, Oxford and VCU Experimental Research on Genetic Epidemiology) project of MDD which recorded during interviewing. Our analyses were based on a total of 740 depressed patients recruited from 58 provincial mental health centers and psychiatric departments of general medical hospitals in 45 cities of 23 China provinces. All patients were female. They were excluded if they had bipolar disorder, intelligence deficiency or any type of psychosis. Patients were aged from 30 to 60, the mean age (standard deviation) of them was 44.4 (8.9). More details of this research include diagnosis and measures were described in [20].

3.2 Data Acquisition

Voices have been collected by recording pens during computer-based interviewing. All interviewers were professional medical staffs, and trained on how to carry out the interview by CONVERGE team for at least a week. The Interview (equal to the content of recordings) includes assessment of demographic, family history, life events, psychopathology (e.g. depression, anxiety, mania, psychosis, PPD) and psychosocial functioning. The interview lasted on average two hours. The answers of patients in interviews are the data what we want.

3.3 Data Preprocessing

The audios were recorded for auditing by trained editors who provided feedback on the quality of the interviews at the beginning, thus noise control had not been planned ahead. Since we only analyze the voices of patients', data preprocess is required before usage. There are three steps of data preprocessing before speech features extraction (see Fig. 1).

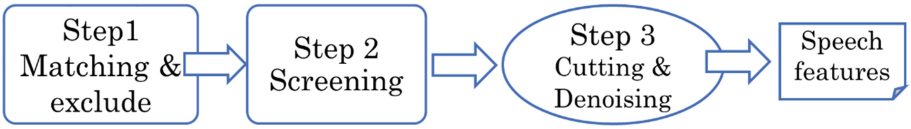

Fig. 1. The procedure of data preprocessing

The first step was matching and exclusion. First of all, we exported 11875 MP3 files from CONVERGE database. We need to match recordings with the results of psychological scales. We only left those patients who have both voice recording and questionnaire results. After matching, a total of 4243 patients were remained. The next step is exclusion. To make sure that enough recordings with enough length were used in the following steps, the short recordings should be excluded. On the basis of our experiences, we excluded those recordings which less than half an hour. Finally, a total of 3964 patients remained in this step.

The second step was screening. There were all kinds of noises in these recordings. We selected high-quality recordings to avoid the impact of noises on the predictive results. We divided all recordings into different levels according to the certain evaluation criteria (see Table 1). Finally, 774 recordings of level *A* were labelled, which were used for the further process.

Table 1. Evaluation criteria of voices in different levels

Level	Evaluation criteria
A	Background is quiet
B	Background has light noises
C	Noisy, hard to be cut
D	The voice of patient is not clear[a]

[a]Clear, means the voices are distinguishable, the content could be easily understood via hearing

In the last and most important step, our mission was cutting and denoising. We needed to separate the voices of patients from interviewers, and wiped out other noises. We recruited a few workers to engage in this part of work. The requirements of this work including: (1) the voice clips should be longer than 5 s; (2) the noises need to be cut include but not limited to ring, telephone ring, click, voices of other people, and so on. There were 740 patients remained after denoised, including 21459 voice clips. Each voice clip is equal to one answer of one patient.

3.4 Feature Extraction, Selection and Data Analysis

Open SMILE [21] was used to extract speech features. A total of 988 speech features were extracted. The procedure of feature extraction was as follows: firstly, in view of the above related work, 26 basic speech features were extracted from recordings,

including intensity, loudness, zero cross rate, voicing probability, F0, F0 envelope, eight linear spectral pair frequencies (LSP) and twelve Mel-frequency cepstral coefficients (MFCC). Secondly, to investigate the variability of voices, 26 features were turned into their first-order derivatives. Thirdly, we calculated 19 statistics of those features mentioned above, such as mean, standard deviation, range, etc. At last, we acquired 988 (= (26 + 26) × 19) speech features.

It is expected that there are some irrelevant and redundant data will weaken the prediction performance. Therefore, it is not a good idea to input all speech features for prediction. Speech features should be selected before prediction. For choosing relevant features and achieving dimension reduction of speech features, we used the Sequential Floating Forward Selection (SFFS) algorithm.

Data analyses mainly includes classification and correlation analysis. Patients were divided into two groups in the light of whether they have been diagnosed as PPD or not. The group labels were considered as golden standard in classification: patients with PPD were labeled 1, without PPD were labeled 0. As classification, we implemented SVM and 5-fold cross validation for training and testing different models. To figure out whether there are salient relationships between the independent variables "number of features" and "sample size" and the predictive effects, we used partial correlation analyses to test them. In addition, paired-sample t-test was used in trying out the impacts of dimension reduction and the content of question on the predictive effects.

4 Results

4.1 Prediction

The rate of PPD group and non-PPD group was kept 1:1 to ensure that sample size has no obvious impact on predictive results. In order to directly observe the effect of dimension reduction, we used all speech features and features after dimension reduction to predict PPD, respectively. The results are respectively shown in Tables 2 and 3. We list the results of top ten best-performing questions, and order by the sample size from small to large.

In Table 2, the classification was based on 988 speech features. The best predictive result of F-measure is 65%, which is the reply to one question of depression scale. Observation of row 4–9, we found that different questions with same sample size had different predictive powers, considering the speech features used in these ten questions were the same. In addition, the predictive effects of demographic answers were common lower than the other questions.

In Table 3, the number of features were dramatically reduced after dimension reduction. The best predictive result 69% of the selected features is *PSY.3*, which is the reply to one question of psychosis scale. The average value of F-measure was increased 5.2%. By looking into row 4–9, we found that the differences of predictive effects of different questions with same sample size decreased.

Table 2. Results of classification using 988 features

Sample size	Question	Precision	Recall	F-measure
80	DEP.E24.F	**0.65**	**0.65**	**0.65**
90	PSY.4	0.57	0.57	0.57
120	PSY.3	0.62	0.62	0.61
120	DEP.E29	0.58	0.57	0.57
130	GAD.D64.D	0.61	0.61	0.61
130	DEP.E26	0.53	0.53	0.53
170	D2.B	0.46	0.46	0.46
170	D4.A	0.49	0.49	0.49
190	D6.A	0.55	0.55	0.55
220	D6	0.53	0.53	0.53
Average		0.559	0.558	0.557

In the second column, these abbreviations before the first point represent the corresponding questionnaire, the content after the first point represents the corresponding item (means question) number in questionnaire (except scale demographics, the content after letter "D" is the item number). DEP, depression scale; PSY, psychosis scale; GAS, scale of general anxiety disorder; D, demographics.

Table 3. Results of classification after dimension reduction

Sample size	Number of features	Question	Precision	Recall	F-measure
80	12	DEP.E24.F	0.63	0.62	0.62
90	35	PSY.4	0.61	0.60	0.59
120	15	PSY.3	**0.69**	**0.69**	**0.69**
120	18	DEP.E29	0.65	0.64	0.64
130	28	GAD.D64.D	0.62	0.62	0.62
130	12	DEP.E26	0.62	0.62	0.62
170	13	D2.B	0.53	0.53	0.53
170	30	D4.A	0.54	0.54	0.53
190	9	D6.A	0.63	0.63	0.63
220	12	D6	0.63	0.62	0.62
Average			0.615	0.611	0.609

In the second column, these abbreviations before the first point represent the corresponding questionnaire, the content after the first point represents the corresponding item (means question) number in questionnaire (except scale demographics, the content after letter "D" is the item number). DEP, depression scale; PSY, psychosis scale; GAS, scale of general anxiety disorder; D, demographics.

PPD and non-PPD Confusion matrixes were shown in Table 4. The precision of PPD was markedly improved after dimension reduction, reaching 75%. In contrast, the precision of non-PPD had slightly decreased after dimension reduction.

Table 4. Confusion matrixes of the most effective questions

DEP.E24.F	0	1	Precision
0	27	13	67.5%
1	15	25	62.5%

a) 988 features (0, non-PPD; 1, PPD)

PSY.3	0	1	Precision
0	38	22	63.3%
1	15	45	75%

b) 15 features (0, non-PPD; 1, PPD)

4.2 Correlation Analysis and Significance Test

Do independent variables sample size and the number of features have significant relationships with predictive effects? To figure out it, partial correlation analyses were applied in consideration of the impact of the other factor. The analyzed result between the number of features and predictive indexes exhibited that there is no salient correlation between them, after controlling the sample size (precision: $r = -0.461$, $p > .1$; recall: $r = -0.422$, $p > .1$; F-measure: $r = -0.473$, $p > .1$). The correlation analysis between sample size and predictive indexes (see Fig. 2) shown that sample size had a significant negative correlation with three indexes which predicted by 988 features after controlling the impact of the number of features (precision: $r = -0.697$, $p < .05$; recall: $r = -0.691$, $p < .05$; F-measure: $r = -0.691$, $p < .05$). However, there is no significant correlation between sample size and three indexes which calculated via reduced features.

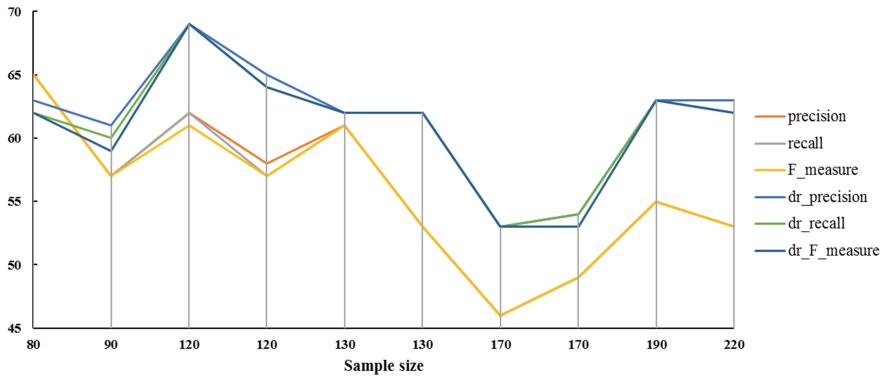

Fig. 2. The impact of sample size on predictive indexes (dr, dimension reduction)

To make it clear that if dimension reduction can evidently improve the predictive effect or not, we compared the predictive results of total features with reduced features by using paired-sample t test (Table 5 lists the means and standard deviations of

predictive indexes). The results indicated that the predictive results improved significantly after dimension reduction (precision: $t = -4.763$, $p < .01$; recall: $t = -4.31$, $p < .01$; F-measure: $t = -4.061$, $p < .01$). Further analysis, considering the impact of question, the contrast of three pairs questions with same sample size was ran by paired-sample t test. The results shown that the differences between different questions with same sample size had saliently shrank after dimension reduction (120 sample size: $t = -1$, $p = .42$; 170 sample size: $t = 7$, $p < .05$). The results of the sample size of 130 cannot test t value because their SD is zero. But their difference value's change is the largest after dimension reduction.

Table 5. The means and standard deviations of predictive indexes

	988 features			Reduced features		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Mean	55.90	55.89	55.79	61.50	61.10	60.90
SD	5.92	5.88	5.77	4.77	4.64	4.86

SD, standard deviation

5 Discussion

The purpose of this study is to detect PPD in depressed patients via speech features. We used openSMILE for feature extraction, selected SFFS for feature selection, tried different numbers of features, set 5-fold cross validation for strengthening generalizability of model and applied SVM in Weka for training and testing different models. The best performance of F-measure reaching 69%, comparing with the random predictive effect 50%, which suggests that voice could be used as a potential behavioral indicator to identify depression disorder' subtypes.

We speculate there may be important influences of *the number of features*, *sample size* and *the content of question* on the predictive effect. Our results indicated that the number of features has no significant relationship with the prediction, but the predictive effect is dramatically improved after dimension reduction. The number of features among different questions are different after dimension reduction, so we think the positive impact of dimension reduction on predictive results is a combined result of number of features and content of question. It is unexpected that there is a negative correlation between sample size and predictive effect. The probable cause is demographic questions lack of the ability of emotional induction, which results in the undistinguishable neutral emotion in all patients' voices. Just as it is shown in Fig. 2, questions D2.B and D4.A make significant contributions to obvious dents of curves.

Different question has different predictive effect. We can find some clues by the most effective questions in Tables 2 and 3. The most effective question is DEP.E24.F before dimension reduction. This question asked patients to recall the state in their severest depressive episode. The most effective question is PSY.3 after dimension reduction. This question asks, "*have you ever taken medicine for your nerves or the way you were feeling or acting?*". All patients are recurrent depressive sufferers, they

must have experiences on taking anti-depressants. Thus, it is a good question to induce emotion, because the experiences about psychotropic medicine probably were negative due to medicines' side effects. In summary, the effect of emotional induction of questions has an important influence on the predictive effect.

6 Conclusion

In this study, the best predictive performance of our speech-based models is F-measure 69%, which suggests that the speech features could be used as a potential behavioral indicator to identify PPD in depressed patients. A combined impact of features and question contribute to the improvement of predictive effect. After dimension reduction, the average value of F-measure was increased 5.2%, and the precision of PPD was rose to 75%. Compared with neutral demographic questions, the features of emotional induced questions have better predictive effects.

Acknowledgments. This work was supported by the National Basic Research Program of China (973 Program) (No. 2014CB744603), and Natural Science Foundation of Hubei Province (2016CFB208).

References

1. Pitt, B.: 'Atypical' depression following childbirth. Br. J. Psychiatry **114**(516), 1325–1335 (1968)
2. Burke, L.: The impact of maternal depression on familial relationships. Int. Rev. Psychiatry **15**(3), 243–255 (2003)
3. American College of Obstetricians and Gynecologists. Committee on Obstetric Practice, Committee opinion no. 453: Screening for depression during and after pregnancy. Obstet. Gynecol. **115**(2 Pt 1), 394–395 (2010)
4. Accounts Payable Association: Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Publishing (2013)
5. Kramer, E.: Elimination of verbal cues in judgments of emotion from voice. J. Abnorm. Soc. Psychol. **68**(4), 390–396 (1964)
6. Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., Snyder, P.J.: Voice acoustical measurement of the severity of major depression. Brain Cognit. **56**(1), 30–35 (2004)
7. Sobin, C., Sackeim, H.A.: Psychomotor symptoms of depression. Am. J. Psychiatry **154**(1), 4–17 (1997)
8. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden Markov models. Speech Commun. **41**(4), 603–623 (2003)
9. Wu, S., Falk, T.H., Chan, W.-Y.: Automatic speech emotion recognition using modulation spectral features. Speech Commun. **53**(5), 768–785 (2011)
10. Ellgring, H., Scherer, P.K.R.: Vocal indicators of mood change in depression. J. Nonverbal Behav. **20**(2), 83–110 (1996)
11. Mundt, J.C., Vogel, A.P., Feltner, D.E., Lenderking, W.R.: Vocal acoustic biomarkers of depression severity and treatment response. Biol. Psychiatry **72**(7), 580–587 (2012)

12. Flint, A.J., Black, S.E., Campbell-Taylor, I., Gailey, G.F., Levinton, C.: Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. J. Psychiatr. Res. **27**(3), 309–319 (1993)
13. Mandal, M.K., Srivastava, P., Singh, S.K.: Paralinguistic characteristics of speech in schizophrenics and depressives. J. Psychiatr. Res. **24**(2), 191–196 (1990)
14. Porritt, L.L., Zinser, M.C., Bachorowski, J.-A., Kaplan, P.S.: Depression diagnoses and fundamental frequency-based acoustic cues in maternal infant-directed speech. Lang. Learn. Dev. **10**(1), 51–67 (2014)
15. Cohen, A.S., Kim, Y., Najolia, G.M.: Psychiatric symptom versus neurocognitive correlates of diminished expressivity in schizophrenia and mood disorders. Schizophr. Res. **146**(1–3), 249–253 (2013)
16. Tolkmitt, F., Helfrich, H., Standke, R., Scherer, K.R.: Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. J. Commun. Disord. **15**(3), 209–222 (1982)
17. Cohen, A.S., Elvevåg, B.: Automated computerized analysis of speech in psychiatric disorders. Curr. Opin. Psychiatry **27**(3), 203–209 (2014)
18. Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., Epps, J.: Diagnosis of depression by behavioural signals: a multimodal approach. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, New York, NY, USA, pp. 11–20 (2013)
19. Cohn, J.F., et al.: Detecting depression from facial actions and vocal prosody. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–7 (2009)
20. Yang, F., et al.: Clinical features of and risk factors for major depression with history of postpartum episodes in Han Chinese women: a retrospective study. J. Affect. Disord. **183**, 339–346 (2015)
21. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia, New York, NY, USA, pp. 835–838 (2013)