PLOS ONE

RESEARCH ARTICLE

Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders

Wei Pan[1,2], **Jonathan Flint**[3], **Liat Shenhav**[4], **Tianli Liu**[5], **Mingming Liu**[1,2], **Bin Hu**[6], **Tingshao Zhu**[1]*

1 Institute of Psychology, Chinese Academy of Sciences, Beijing, China, **2** University of Chinese Academy of Sciences, Beijing, China, **3** Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, United States of America, **4** Department of Computer Science, University of California Los Angeles, Los Angeles, United States of America, **5** Institute of Population Research, Peking University, Beijing, China, **6** School of Information Science & Engineering, Lanzhou University, Lanzhou, China

* tszhu@psych.ac.cn

Citation: Pan W, Flint J, Shenhav L, Liu T, Liu M, Hu B, et al. (2019) Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders. PLoS ONE 14(6): e0218172. https://doi.org/10.1371/journal.pone.0218172

Editor: Zezhi Li, National Institutes of Health, UNITED STATES

Received: November 9, 2018

Accepted: May 28, 2019

Published: June 20, 2019

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors gratefully acknowledge the generous support from National Basic Research Program of China (2014CB744600). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

A large proportion of Depression Disorder patients do not receive an effective diagnosis, which makes it necessary to find a more objective assessment to facilitate a more rapid and accurate diagnosis of depression. Speech data is easy to acquire clinically, its association with depression has been studied, although the actual predictive effect of voice features has not been examined. Thus, we do not have a general understanding of the extent to which voice features contribute to the identification of depression. In this study, we investigated the significance of the association between voice features and depression using binary logistic regression, and the actual classification effect of voice features on depression was re-examined through classification modeling. Nearly 1000 Chinese females participated in this study. Several different datasets was included as test set. We found that 4 voice features (PC1, PC6, PC17, PC24, $P<0.05$, corrected) made significant contribution to depression, and that the contribution effect of the voice features alone reached 35.65% (*Nagelkerke's* R^2). In classification modeling, voice data based model has consistently higher predicting accuracy(F-measure) than the baseline model of demographic data when tested on different datasets, even across different emotion context. F-measure of voice features alone reached 81%, consistent with existing data. These results demonstrate that voice features are effective in predicting depression and indicate that more sophisticated models based on voice features can be built to help in clinical diagnosis.

Introduction

Depression disorder is the commonest psychiatric disorder (the lifetime prevalence reaching 16.2%) [1–7] and is now the leading cause of disability [8]. Yet despite its prevalence, many cases of depression remain unrecognized [9], depriving many of the possibility of receiving

effective treatment. Fewer than half of those eligible receive treatment and in many countries the figure is less than 10% [10]. One of the main obstacles in the way of treatment provision is the difficulty of recognizing and diagnosing depression. Diagnosis currently requires interview by a clinician, often over half an hour or more, a method that rarely exceeds an inter-rater reliability of 0.7 (kappa coefficient) [11]; in one large field study reliability was estimated to be as low as 0.25 [12]. According to a meta-analysis of 41 studies [13], community doctors had an accuracy of 47% in recognizing patients with depression recognition accuracy of depression by general practitioners is 47.3%. Methods to identify cases of depression that can be deployed at an appropriate scale are needed.

Researchers have attempted to find objective methods to increase the accuracy of depression diagnosis. Blood transcriptomic biomarkers and acoustic biomarkers, among other methods, have been investigated to help detect depression [14,15]. Among these, voice analysis has attracted increasing attention, being easy-to-acquire, non-invasive, and having objective advantages.

The voices of depression patients have recognizable characteristics, including slow speech rate, frequent pauses, little difference in speech features, and lack of cadence [16]. Voice features have been used in machine-learning methods to help diagnose depression. Multiple acoustic features have been investigated in empirical studies, including spectral, cepstral, prosodic, glottal, and Teager energy operator features [17–21] Potentially confounding factors may challenge the validity of such published studies, however.

The impact of confounding factors has been pervasively neglected in most of the current research on this topic. It has been suggested by many studies that demographic variables are significantly associated with depression. For example, occupations with higher status protect against depressive symptoms, and workers in specific occupational sectors and types report different levels of depressive symptoms [22]. What is more, demographic factors can also impact individual voice features, raising concerns [23]. For instance, the fundamental frequency (voice pitch) of females is usually higher than that of males [24]. Other studies have mentioned that demographic factors, such as age, gender, emotion, and the characteristics of the speaker can impact the classification results [25–27].

Several studies have attempted to explain the existence of nuisance factors and to control these to a certain degree. Sex-independent classifiers have been suggested, and have achieved better results [19,20,28,29]. However, none of these studies have examined the extent to which demographic information contributes to depression prediction. More importantly, the association between voice features and depression has not been evaluated to a significant level of confidence, and there has not been any reference to compare the predictive effect of voice features to. In addition, sample sizes in all studies are relatively small limiting power to detect subtle but possibly important classifiers [30–32], especially when there are some subcategories, such as gender.

In this study, the primary aims were: 1) to evaluate whether voice features can significantly contribute to the prediction of depression; 2) replicate the previous classification results for voice features on depression and examine the generalization ability of classification models; and 3) identify the effect of voice features and compare this to demographic variables. Once voice features have been shown to be useful in predicting depression, future research can confidently explore more robust classification models to aid in clinical depression diagnosis.

Materials and methods

Data collection

We have two different data sets. The first one is used for building models, named interview speech dataset. The second one is only for model test, which named 973 dataset. They are from different project with different research project.

Interview speech dataset. In this dataset, all subjects were interviewed using a computerized assessment system, during which participants' voice was recored. All participants were Chinese females with Han nationality and had three generations of Han nationality relatives. The depression patients (the case group) were all diagnosed by psychiatric specialists with DSM-5. All had two or more major depression episodes. Comorbidities, such as bipolar disorder and other mental deficiencies, were excluded. The healthy participants (the control group) had no experience of depression or any other mental illnesses. There were no blood relationships between the two groups. The research design is described in detail in a previous report [33, 34]. All participants were between 30 and 60 years old.

Only the voice datasets from the demographic questions numbered D2.A and D2.B were chosen for analysis because the others were either too short or the sample size was too small. Question D2.A is *When is your birthday*? The voice recordings of its answer varies from 0.29 to 33.09 s. Question D2.B is *How old are you*? The voice recordings of its answer ranges from 0.11 to 11.78 s. In this research, voice recordings of question D2.A was used to build logistic regression equations and classification models. While voice recordings of question D2.B was used as a test set for classification models mentioned above, as it has high homogeneity with D2.A dataset but not exactly the same—not all participants has both voice recordings for D2.A and D2.B. For convenience, these two datasets will be named as D2.A, D2.B.

It should be noted that each participant provided one voice sample, and all voice files were saved in a .wav/16 bit/16kHz format.

The study protocol was approved centrally by the Ethical Review Board of Oxford University (Oxford Tropical Research Ethics Committee) and the ethics committees of all participating hospitals in China. All participants provided their written informed consent.

973 dataset. This dataset was collected from the 973 project [35].

In this project, there was a priming of different emotions(positive, neutral, negative), under which participants' voice recordings were collected by different tasks(video watching, text reading, interview, question answering, picture describing). Participants from case group is diagnosed by clinicians with DSM-IV, their depression severity varies. For the control group, participants are both physically and mentally healthy people. Each participant experienced all three emotion priming three different emotion priming. 73 participants were included, 34 healthy individuals and 39 depressed patients. Their age also ranges between 30–60 years old. This dataset was included in this research to investigate whether models built by interview speech dataset can be used to predict voice recordings from different emotion context.

Data preprocessing

Interview speech dataset. After the data was cleaned (clearly audible, clips of doctors' voices cut), we chose sample for which complete demographic information (six demographic variables: age, accent, education, occupation, marital status, social class) was available. There were 1132 participants (584 depression patients and 548 healthy people) for question D2.A, and 904 participants (500 depression patients and 404 healthy people) for D2.B.

There are 6 different demographic variables in this dataset: age(30~60), occupation(1work; 2 wait for employment; 3 retired; 4 housewife; 5 others), education(1 uneducated; 2 primary school; 3 junior high school; 4 senior high school; 5 junior college; 6 senior college; 7 college and 8 above college), marital.status(1 married; 2 living apart; 3 divorced; 4 widowed; 5 single), social.class and accent. The accent areas were classified based on city and province information (Institute of Linguistics, CASS, 2012). There are 15 accent areas involved in our dataset: Beijing Mandarin, Zhongyuan Mandarin, Dongbei Mandarin, Jianghuai Mandarin, Lanyin Mandarin,

Jilu Mandarin, Jiaoliao Mandarin, Xinan Mandarin, Gan, Xiang, Jin, Wu, Min, Yue, and Tianjin. This variable was then translated into a dummy variable for further analysis.

At this point, each sample had one column headed with an ID number, 988 columns of voice feature data, six columns of demographic data, and one column containing the depression diagnosis. The dependent variable was depression and was divided into two classes: depressed (labeled 1) or not (labeled 0).

For feature extraction, based on existing findings, we chose 26 physical features that have been widely used in emotion recognition [36–38]: intensity, loudness, zero-crossing rate, voicing probability, fundamental frequency (F0), F0 envelope, eight line spectral pairs (LSP), and 12 mel-frequency cepstral coefficients (MFCC). The delta values, which reflect dynamic change in voice features, were then calculated for each of these 26 static features, and also for 19 statistical features, including maximum, minimum, range, mean, standard deviation, skewness, etc. By employing openSMILE [39], a total of 988 voice features were obtained [40] The voice data was then standardized.

973 dataset. For 973 dataset, all voice recordings were complete, no need to cut. And the background noise was well-controlled in this project. 988 voice features were extracted in the same way as interview data.

For 973 dataset, the demographic variables are age, education and occupation.

Data description

Most of the demographic variables were significantly different between case and control group in D2.A, D2.B and for 973 project. As some of the sample size for a category is too small, we performed all difference test based on the permutation test. To simplify the content and highlight the most important analysis in this research, we put all tables about data description as S1 Table in *Supporting Information*. As there are many biases between cases and controls in demographic variables, we need to match cases and controls to control the biases in demographic variables as much as we can.

Confounders matching

Propensity score matching (PSM) is a statistical matching technique that attempts to estimate the effect of a treatment, policy, or other intervention by accounting for the covariates that predict receiving the treatment. PSM attempts to reduce the bias due to confounding variables that could be found in an estimate of the treatment effect obtained from simply comparing outcomes among units that received the treatment versus those that did not [41]. We used matchIt package from R to match cases and controls on those confounders—demographic variables.

Data description after matching

After propensity score matching, most of the difference tests were changed to insignificant, or its z value, chi-square value decreased. But it should be noted that there are still many demographic variables significantly differed between groups. That is to say, our data is still biased to some level. Please see S1 Table in the *Supporting Information*.

Data analysis

Binary logistic regression. As our data is biased, we need to check the contribution of demographic variables and voice data separately. And compare model fitness when using voice data alone, demographic data alone, and two combined(voice data + demographic data)

predicting depression status to separate out the effect of voice data. Hence in the first stage, we employed statistical methods to explore whether voice data can significantly predict depression. We built three logistic regression models described above to investigate the contributing effect of both voice features and demographic variables when predicting depression. And we used ANOVA test to compare model fitness among these models. These analyses were based on dataset D2.A.

Principal component analysis (PCA) was initially conducted in order to reduce data dimension and avoid the multicollinearity problem. First, only demographic variables entered model. Second, only voice feature data entered model. Third, voice data and demographic data entered logistic model together.

Both the odds ratio (*OR*) and *Nagelkerke's* R^2 were employed as indicators of the contributing effects of the variables to depression. The *OR* [42] is the exponent based on the natural constant, *e*, which explains the extent to which the change in the independent variable causes the change in probability of the dependent variable. When the *OR* value is >1, the corresponding variable is the risk factor for depression; higher values imply a greater contributing effect to the dependent variable. *Nagelkerke's* R^2 [42] equals the adjusted R^2 in linear regression, which also means that the value of *Nagelkerke's* R^2 provides the amount of variance of the dependent variable explained by the explanatory variable. The stratified entry method allowed us to examine the contributions of important variables of interest gained after controlling the confounding variables.

Classification modeling. In the second stage, we tested the actual classification effect of voice features on depression using supervised learning methods. We built several classification models with identical data from question D2.A.

We split D2.A into training set and test set with ratio 7:3. All classification models were built based on the 70% training set of D2.A. To investigate predicting effect of voice data, we take classification models built on demographic variables as baseline to compare. To test model robustness, we tested the classification models on three different test sets. The first one is the rest 30% voice data of D2.A. The second one is the D2.B dataset. The third one is the 973 dataset. It should be noted that to investigate whether the models we built have robust predicting effect under different emotion context, we split 973 dataset as three test sets: data set under positive, neutral and negative emotion context separately.

The voice data were high dimensional, i.e. many variables were measured. To improve the generalizability of the models and to avoid the curse of dimensionality, feature selection was implemented, using random forest as the selection strategy with Boruta package in R [43].

The depression variable (labeled as 0 and 1) was included as a classification label. Classification models were built based on the training set, then the model performance was assessed on the test set. The classification models were built using random forest [44]. With each test set, three models were built, having only demographic data input, only voice data input, and both voice data and demographic data input, respectively, in order to compare the classification results under different situations. We used those identical models to select the same features under question D2.B and three 973 datasets, then classified them to estimate the generalization abilities of the models built on the training set of D2.A.

F measure [45], an evaluation indicator, was employed to assess the accuracy of the classification models. It considered both the precision) and the recall) of the classification models in order to compute the score. Precision is the number of correct positive results divided by the number of all positive results returned by the classifier. Recall is the number of all samples that should have been identified as positive, divided by the number of correct positive results returned by the classifier.

Results

Binary logistic regression

We examined how much demographic variables and voice features contributed to depression, using *Nagelkerke's* R^2 statistic, and estimated the effect of each variable using *OR*s. More importantly, we also compared model fitness among different variable input. The results showed that when only demographic data in model, it accounted for 10.87% (*Nagelkerke's* R^2) of the variance in the dependent variable depression. *OR*s for each demographic variable and their significance are shown in Table 1; age (*OR* = 0.95, *P*<*0.0001*), occupation (*OR* = 1.17, *P*<0.0001), and wu accent (*OR* = 2.90, *P* = 0.019) significantly predicted depression.

We applied PCA to the 988 voice features in order to reduce the data. We found that 137 principal components (PCs) captured 90% of the original data variance, with the PCs showing no significant correlation in a pairwise correlation analysis.

When only voice data entered logistic model, it accounted for 35.65% of variance in the dependent variable depression(Nagelkerke's R^2). ORs for each voice PCs and their significance are shown in Table 2. PC1 (OR = 0.58, P<0.0001), PC6 (OR = 1.57, P<0.001), PC17 (OR = 1.53, P<0.0001) and PC24 (OR = 1.45, P<0.05) significantly predicted depression.

Demographic data and voice data together explained 39.98% (*Nagelkerke's* R^2) of the risk for depression. This indicate that the unique contribution of voice features is 29.11%. The *OR*s for the PCs and demographic variables included in the model are shown in Table 3; the demographic variables occupation (*OR* = 1.21, *P*<0.05), and voice features PC1 (*OR* = 0.59, *P*<0.001), and PC17 (*OR* = 1.54, *P*<0.01) significantly predicted depression.

More importantly, we compared model fitness between different input. See Tables 4 and 5. In Table 4, we compare model fitness based on only voice data and demographic data + voice data. Results showed that model fitness improved when voice data entered the model($\varkappa^2$ = -241.11, P<0.001). In Table 5, model fitness between models on only voice data and only demographic data showed that model fitness of voice data is significantly better than demographic data($\varkappa^2$ = -200.97, P<0.001).

Classification modeling. Classification models were constructed to test whether voice features could be used to diagnose depression. We used feature selection, a method that improves a model's generalizability and avoids the curse of dimensionality. We identified 37 features (1 demographic variable and 36 voice features). These features are listed in S2 Table. Please check in the *Supporting Information*.

All classification models were built based on these 37 features. The training set was from the 70% of D2.A. An *F measure* was adopted to estimate to what extent models correctly classified depression patients and the controls.

Testing model performance on rest 30% of D2.A. When predictions were obtained for the test set of D2.A. Results showed, when only demographic data was used to classify

Table 1. Binary logistic regression model of demographic data.

	β	SE	*Wald*	*corrected P*	*OR*
age	-0.05	0.01	-4.83	2.1E-04***	0.95
occupation	0.15	0.04	4.07	7.4E-03**	1.17
wu	1.07	0.24	4.43	1.5E-03**	2.90

P<0.01**,
P<0.001***;
Bonferroni correction

https://doi.org/10.1371/journal.pone.0218172.t001

Table 2. Binary logistic regression model of voice data.

	β	SE	*Wald*	*corrected P*	*OR*
PC1	-0.5486592	0.1007107	-5.448	7.85E-06***	0.58
PC6	0.4536389	0.1117325	4.06	7.56E-03**	1.57
PC17	0.4251865	0.093759	4.535	8.87E-04***	1.53
PC24	0.3687606	0.0984286	3.746	2.76E-02*	1.45

$P<0.05^{*}$,
$P<0.01^{**}$,
$P<0.001^{***}$;
Bonferroni correction

https://doi.org/10.1371/journal.pone.0218172.t002

Table 3. Binary logistic regression model of demographic data and voice data.

	β	SE	*Wald*	corrected *P*	OR
occupation	0.19	0.05	3.82	0.021*	1.21
PC1	-0.53	0.12	-4.54	8.65E-04***	0.59
PC17	0.43	0.10	4.35	2.1E-03**	1.54

$P<0.05^{*}$,
$P<0.01^{**}$,
$P<0.001^{***}$;
Bonferroni correction

https://doi.org/10.1371/journal.pone.0218172.t003

Table 4. Compare model fitness between demo+voi and demo.

	resid.df	resid.dev	df	χ^2	*P*
demo+voi[a]	764	886.59		-241.11	9.848E-08***
demo	901	1127.7	-137		

[a] demo:demographic data; voi: voice data
$P<0.001^{***}$

https://doi.org/10.1371/journal.pone.0218172.t004

Table 5. Compare model fitness between demo+voi and demo.

	resid.df	resid.dev	df	χ^2	*P*
voi	780	926.73		-200.97	6.77E-06***
demo	901	1127.7	-121		

$P<0.001^{***}$

https://doi.org/10.1371/journal.pone.0218172.t005

Table 6. Classification results for the test set from D2.A.

	Accuracy	*Precision*	*Recall*	*F-measure*
demo	0.66	0.76	0.72	0.73
voi	0.72	0.73	0.91	**0.81**
demo+voi	0.73	0.75	0.87	0.81

https://doi.org/10.1371/journal.pone.0218172.t006

depression, the classification accuracy (*F measure*) was 73%. When only voice data was in the classification model, the accuracy was 81%. While the classification accuracy of demographic data and voice data is 81%. Compared to the classification accuracy of models based on demographic data alone, the classification accuracy of the models with voice data is higher. These results are shown in Table 6.

Testing model performance on D2.B. The number of participants overlap between D2.A and D2.B is 410. When predictions were obtained for the test set of D2.B. Results showed, when only demographic data was used to classify depression, the classification accuracy (*F measure*) was 75%. When only voice data was in the classification model, the accuracy was 80%. While the classification accuracy of demographic data and voice data is 81%. Compared to the classification accuracy of models based on demographic data alone, the classification accuracy of the models with voice data is consistently higher. These results are shown in Table 7.

Testing model performance on 973 dataset. To further test model robustness across different emotion context, we tested our model of 70% D2.A training set on three different emotion value relevant dataset under 973 data.

In the positive emotion situation, results showed, when only demographic data was used to classify depression, the classification accuracy (*F measure*) was 69%. When only voice data was in the classification model, the accuracy was 75%. While the classification accuracy of demographic data and voice data is 77%. Compared to the classification accuracy of models based on demographic data alone, the classification accuracy of the models with voice data is also consistently higher. These results are shown in Table 8.

In the neural emotion situation, results showed, when only demographic data was used to classify depression, the classification accuracy (*F measure*) was 75%. When only voice data was in the classification model, the accuracy was 80%. While the classification accuracy of demographic data and voice data is 81%. Compared to the classification accuracy of models based on demographic data alone, the classification accuracy of the models with voice data is higher, consistent to former results. These results are shown in Table 9.

In the negative emotion situation, results showed, when only demographic data was used to classify depression, the classification accuracy (*F measure*) was 69%. When only voice data was in the classification model, the accuracy was 76%. While the classification accuracy of demographic data and voice data is 78%. Compared to the classification accuracy of models based on demographic data alone, the classification accuracy of the models with voice data is also higher, consistent to former results. These results are shown in Table 10.

Discussion

In this study, the relationships among voice features, demographic variables, and depression were systematically examined. Our findings addressed the primary goals of the study finding that (1) voice features make a significant contribution to the prediction of depression; (2) In regression model, the model fitness of voice data alone is significantly better than the model fitness of demographic data alone; (3) Depression classification models built on voice features

Table 7. Classification results for the test set from D2.B.

	accuracy	*precision*	*recall*	*F-measure*
demo	0.67	0.78	0.73	0.75
voi	0.71	0.77	0.84	**0.8**
demo+voi	0.73	0.8	0.81	0.81

https://doi.org/10.1371/journal.pone.0218172.t007

Table 8. Classification results for the test set from 973 data under positive emotion context.

	accuracy	*precision*	*recall*	*F-measure*
demo	0.62	0.76	0.64	0.69
voi	0.63	0.69	0.82	**0.75**
demo+voi	0.66	0.7	0.85	0.77

https://doi.org/10.1371/journal.pone.0218172.t008

are effective with generalization ability; (4) Depression classification models built on voice features are robust across different emotion situations.

First, binary logistic regression analysis showed that voice features significantly contribute to the prediction of depression (voice features, PC1, PC6, PC1, and PC24, $P<0.05$). What is more, when voice data and demographic data was included, compared to the results of only demographic data in the model, the variance in depression explained by voice data was 29.11%, which represents the unique contribution of the voice features. What's more, the model fitness of voice data alone is significantly higher than that of demographic data, and when add voice data in the the demographic data based model, model fitness is also significantly improved.

This is the first time to systematically investigate whether voice features significantly predict depression, with a baseline of demographic variables. Several studies have investigated the correlation between voice features, such as prosodic and cepstral features, and depression [17,18,20]. Our results are the first to confirm that voice features can significantly predict depression, with considerable amount of contribution. There is no direct theory that specifies the mechanism of voice features in predicting depression; however, there is evidence to support our results.

On one hand, depression is associated with sustained activity in the brain areas responsible for coding emotional information [46,47]. On the other, plenty of studies have shown that voice parameters are affected by emotion. A review of the literature on human vocal emotion [48] noted that emotion influences voice in three main ways—voice quality, utterance timing, and utterance pitch contour. It has been suggested that basic emotions have a stable effect on voice features across different cultural backgrounds [48–50]. Prosodic features have been considered to be the most important factor in emotion recognition [51]. Spectrum features are also important in conveying emotion [52,53]. These findings show that depression patients may have stable emotional changes and corresponding speaking characteristics, which calls for further examination.

Second, the classification results of our study are in agreement: the classification accuracy of voice features is consistently higher than demographic data in each testing situation. More importantly, voice data can be used to predict depression under different emotion status, meaning depression detection using voice features is reliable and has its potential in clinical situation. This also indicate that voice features is a stable feature of depression, despite their emotion changes. The classification accuracy of voice features alone reached 81%(see Table 7). Classification models based on voice features alone have been widely examined. Classifying accuracy have been reported as between 60% and 90% in the machine learning field [17–

Table 9. Classification results for the test set from 973 data under neutral emotion context.

	accuracy	*precision*	*recall*	*F-measure*
demo	0.67	0.78	0.73	0.75
voi	0.71	0.77	0.84	**0.8**
demo+voi	0.73	0.8	0.81	0.81

https://doi.org/10.1371/journal.pone.0218172.t009

Table 10. Classification results for the test set from 973 data under negative emotion context.

	accuracy	*precision*	*recall*	*F-measure*
demo	0.62	0.76	0.64	0.69
voi	0.64	0.68	0.87	**0.76**
demo+voi	0.66	0.7	0.87	0.78

https://doi.org/10.1371/journal.pone.0218172.t010

21,28,29]. Here, the voice features following feature selection in the classification modeling were mainly calculated using some basic physical voice features: loudness, MFCC, LSP, voicing probability Both MFCCs and LSP are spectral features [54]; loudness, and voicing probability are prosodic features [55].

Spectral features, particularly MFCCs, were useful in classifying depression or not with an accuracy of 80% [18]. Spectral features are believed to reflect the relationship between changes in vocal tract shape and articulator movements [56]. These features have been observed to change in relation to the mental state of the speaker, relating to changes in muscle tension and control [16].

Prosodic features are the properties of syllables and larger units of speech, which contribute to linguistic functions such as intonation, tone, stress, and rhythm [57]. Previous research has shown consistent results that indicate some prosodic abnormality in the voice of depression patients. To illustrate, listeners were able to sense change in pitch, volume, speech rate, and pronunciation before and after treatment [58]. Prosody reflects various features of the speaker or the utterance, including: the emotional state of the speaker; the form of the utterance (statement, question, or command); the presence of irony or sarcasm; emphasis, contrast), and focus); and other elements of language that may not be encoded by grammar or by choice of vocabulary [59].

Taken together, our research has replicated previous results in which voice features were found to classify depression and has shown a stable generalizability when applied to new datasets, even under different emotion context. Though the length of voice recordings in our research are around10s, research on the same interview speech dataset has showed even 10 seconds length can reach ideal classification accuracy [60]. What's more, short utterance has been proved to be effective in speaker identification [61–64]. The consistently higher accuracy in this research also showed short voice recordings can reach ideal predicting accuracy.

In addition, demographic variables also significantly predicted depression (see Tables 1 and 3) and showed great predictive accuracy when classifying depression (see Tables 6,7, 8, 9 and 10). These results indicate that demographic information can improve the classification accuracy of depression in clinical applications, which should be noted by further research.

There are also some limitations in this research. First, our study only includes females, which means we should be more careful when generalize our conclusions to the male population, the intention of only females included is to keep both high sample homogeneity and large sample size, besides voice features between males and females differs a lot. Second, there exists bias on demographic variables, which affected the reliability to some degree. But we have to admit this is part of the nature of observational study, especially for epidemiological studies. And we tried our best to control the effect of confounders. This may implied that there exist some pattern between depression and demographic variables. And also remind us the necessarity to examine and control the effect of demographic variables.

Despite its shortcomings, this study took a pioneering step in examining the predictive effect of voice features with consideration of confounding factors. Demographic characteristics, such as gender, age, emotion, or personality of the speaker, etc., have been shown to be strong confounding factors for depression detection systems [18,25,26] Most of these studies

did not investigate the effect of these factors in a comprehensive way [19,20,28,29]. Morales et al. [23] mentioned the importance of examining and controlling the effect of confounding factors. Sex-independent models have been built [19]. Despite these studies, there has been no such work on evaluating the predicting effect of voice data with comparing baseline of potential confounding factors and how confounding factors affect the occurrence of depression.

Conclusion

Taken together, our findings suggest that voice features play an important role in predicting depression. In addition, demographic variables should be valued in future research. Our results contribute to our understanding of the actual effect of voice features on depression. This research provides a foundation to further explore more robust classification models, as well as to identify related voice features to build more robust models and exploit the clinical application value of voice features to the fullest.

Supporting information

S1 Table. Table containing data description information.
(DOCX)

S2 Table. Features selected for model building.
(DOCX)

S1 Appendix. Voice dataset for answer recordings to question D2.A.
(CSV)

S2 Appendix. Voice dataset for answer recordings to question D2.B.
(CSV)

S3 Appendix. Voice dataset for recordings in 973 dataset.
(CSV)

Author Contributions

Formal analysis: Wei Pan, Jonathan Flint, Liat Shenhav.

Funding acquisition: Tingshao Zhu.

Methodology: Wei Pan, Liat Shenhav, Mingming Liu, Tingshao Zhu.

Project administration: Tingshao Zhu.

Resources: Tingshao Zhu.

Supervision: Tianli Liu, Bin Hu, Tingshao Zhu.

Validation: Liat Shenhav.

Writing – original draft: Wei Pan.

Writing – review & editing: Wei Pan, Jonathan Flint, Tingshao Zhu.

References

1. Murray, C. J., Lopez, A. D., & World Health Organization. (1996). The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020: summary.

2. Melse J. M., Essink-Bot M. L., Kramers P. G., & Hoeymans N. (2000). A national burden of disease calculation: Dutch disability-adjusted life-years. Dutch Burden of Disease Group. American journal of public health, 90(8), 1241–1247. https://doi.org/10.2105/ajph.90.8.1241 PMID: 10937004
3. Michaud C. M., Murray C. J., & Bloom B. R. (2001). Burden of disease—implications for future research. Jama, 285(5), 535–539. https://doi.org/10.1001/jama.285.5.535 PMID: 11176854
4. Nierenberg A. A., Gray S. M., & Grandin L. D. (2001). Mood disorders and suicide. The Journal of clinical psychiatry.
5. Penninx B. W., Beekman A. T., Honig A., Deeg D. J., Schoevers R. A., van Eijk J. T., & van Tilburg W. (2001). Depression and cardiac mortality: results from a community-based longitudinal study. Archives of general psychiatry, 58(3), 221–227. https://doi.org/10.1001/archpsyc.58.3.221 PMID: 11231827
6. Alonso J., Angermeyer M. C., Bernert S., Bruffaerts R., Brugha T. S., . . . & Gasquet I. (2004). Disability and quality of life impact of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. Acta Psychiatrica Scandinavica, 109, 38–46. https://doi.org/10.1111/j.1600-0047.2004.00329.x
7. Üstün T. B., Ayuso-Mateos J. L., Chatterji S., Mathers C., & Murray C. J. (2004). Global burden of depressive disorders in the year 2000. The British journal of psychiatry, 184(5), 386–392. https://doi.org/10.1192/bjp.184.5.386
8. World Health Organization. (2017). Depression and other common mental disorders: global health estimates.
9. Goldberg D. (1995). Epidemiology of mental disorders in primary care settings. Epidemiologic reviews, 17(1), 182–190. https://doi.org/10.1093/oxfordjournals.epirev.a036174 PMID: 8521936
10. World Health Organization, 2018. Depression. Retrived from http://www.who.int/news-room/fact-sheets/detail/depression.
11. Spitzer R. L., Forman J. B., & Nee J. (1979). DSM-III field trials: I. Initial interrater diagnostic reliability. The American journal of psychiatry. http://dx.doi.org/10.1176/ajp.136.6.815
12. Regier D. A., Narrow W. E., Clarke D. E., Kraemer H. C., Kuramoto S. J., Kuhl E. A., & Kupfer D. J. (2013). DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. American journal of psychiatry, 170(1), 59–70. https://doi.org/10.1176/appi.ajp.2012.12070999 PMID: 23111466
13. Mitchell A. J., Vaze A., & Rao S. (2009). Clinical diagnosis of depression in primary care: a meta-analysis. The Lancet, 374(9690), 609–619. https://doi.org/10.1016/S0140-6736(09)60879-5
14. 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. Journal of neurolinguistics, 20(1), 50–64. https://doi.org/10.1016/j.jneuroling.2006.04.001 PMID: 21253440
15. Redei E. E., Andrus B. M., Kwasny M. J., Seok J., Cai X., Ho J., & Mohr D. C. (2014). Blood transcriptomic biomarkers in adult primary care patients with major depressive disorder undergoing cognitive behavioral therapy. Translational psychiatry, 4(9), e442. https://doi.org/10.1038/tp.2014.66
16. 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 transactions on Biomedical Engineering, 47 (7), 829–837. https://doi.org/10.1016/S0022-3956(99)00037-0 PMID: 10916253
17. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Gedeon, T., Breakspear, M., & Parker, G. (2013). A comparative study of different classifiers for detecting depression from spontaneous speech. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 8022–8026). IEEE. doi:10.1109/ICASSP.2013.6639227
18. Cummins, N., Epps, J., Breakspear, M., & Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. In Twelfth Annual Conference of the International Speech Communication Association.
19. Low L. S. A., Maddage N. C., Lech M., Sheeber L. B., & Allen N. B. (2011). Detection of clinical depression in adolescents' speech during family interactions. IEEE Transactions on Biomedical Engineering, 58(3), 574–586. https://doi.org/10.1109/TBME.2010.2091640 PMID: 21075715
20. Moore E. II, Clements M. A., Peifer J. W., & Weisser L. (2008). Critical analysis of the impact of glottal features in the classification of clinical depression in speech. IEEE transactions on biomedical engineering, 55(1), 96–107. https://doi.org/10.1109/TBME.2007.900562 PMID: 18232351
21. Scherer S., Stratou G., Gratch J., & Morency L. P. (2013). Investigating voice quality as a speaker-independent indicator of depression and PTSD. In Interspeech (pp. 847–851).
22. Christ S. L., Lee D. J., Fleming L. E., LeBlanc W. G., Arheart K. L., Chung-Bridges K., . . . & McCollister K. E. (2007). Employment and occupation effects on depressive symptoms in older Americans: does

working past age 65 protect against depression?. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 62(6), S399–S403. https://doi.org/10.1093/geronb/62.6.S399

23. Morales, M., Scherer, S., & Levitan, R. (2017). A Cross-modal Review of Indicators for Depression Detection Systems. In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality (1–12). doi: 10.18653/v1/W17-3101
24. Titze I. R. (1989). Physiologic and acoustic differences between male and female voices. The Journal of the Acoustical Society of America, 85(4), 1699–1707. https://doi.org/10.1121/1.397959 PMID: 2708686
25. 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 Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on (970–974). IEEE. doi: 10.1109/ICASSP.2014.6853741
26. 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 Communication, 71, 10–49. https://doi.org/10.1016/j.specom.2015.03.004
27. Sturim, D., Torres-Carrasquillo, P. A., Quatieri, T. F., Malyska, N., & McCree, A. (2011). Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In Twelfth Annual Conference of the International Speech Communication Association.
28. Scherer S., Stratou G., Lucas G., Mahmoud M., Boberg J., Gratch J., & Morency L. P. (2014). Automatic audiovisual behavior descriptors for psychological disorder analysis. Image and Vision Computing, 32(10), 648–658. https://doi.org/10.1016/j.imavis.2014.06.001
29. Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M. C., & Sahli, H. (2016). Decision tree based depression classification from audio video and language information. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (89–96). ACM. doi: 10.1145/2988257.2988269
30. Dobson C., Woller-Skar M. M., & Green J. (2017). An inquiry-based activity to show the importance of sample size and random sampling. Science Scope, 40(8), 76.
31. Sim J., Saunders B., Waterfield J., & Kingstone T. (2018). Can sample size in qualitative research be determined a priori?. International Journal of Social Research Methodology, 1–16. https://doi.org/10.1080/13645579.2018.1454643
32. Akobeng A. K. (2016). Understanding type I and type II errors, statistical power and sample size. Acta Paediatrica, 105(6), 605–609. https://doi.org/10.1111/apa.13384 PMID: 26935977
33. Yang F., Zhao H., Wang Z., Tao D., Xiao X., Niu Q., . . . & Li K. (2014). Age at onset of recurrent major depression in Han Chinese women–a replication study. Journal of affective disorders, 157, 72–79. https://doi.org/10.1016/j.jad.2014.01.004 PMID: 24581831
34. Yang F., Li Y., Xie D., Shao C., Ren J., Wu W., .. & Qiao D. (2011). Age at onset of major depressive disorder in Han Chinese women: relationship with clinical features and family history. Journal of affective disorders, 135(1–3), 89–94. https://doi.org/10.1016/j.jad.2011.06.056 PMID: 21782247
35. Liu, Z., Hu, B., Yan, L., Wang, T., Liu, F., Li, X., & Kang, H. (2015, September). Detection of depression in speech. In 2015 international conference on affective computing and intelligent interaction (ACII) (pp. 743–747). IEEE.
36. Wang, J., Sui, X., Hu, B., Flint, J., Bai, S., Gao, Y., .. & Zhu, T. (2017a). Detecting Postpartum Depression in Depressed People by Speech Features. In International Conference on Human Centered Computing (pp. 433–442). Springer, Cham. https://doi.org/10.1007/978-3-319-74521-3_46
37. Wang, J., Sui, X., Zhu, T., & Flint, J. (2017b). Identifying comorbidities from depressed people via voice analysis. In Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on (pp. 986–991). IEEE. doi: 10.1109/BIBM.2017.8217791
38. Weng, S., Chen, S., Yu, L., Wu, X., Cai, W., Liu, Z., .. & Li, M. (2015, December). The SYSU system for the interspeech 2015 automatic speaker verification spoofing and countermeasures challenge. In Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific (pp. 152–155). IEEE. doi: 10.1109/APSIPA.2015.7415492
39. Eyben F, Weninger F, Gross F, et al. Recent developments in opensmile, the munich open-source multimedia feature extractor[C]//Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013: 835–838
40. Sui XY.(2017) Depression Recognition with Audios Collected under Natural Environment. Postgraduate dissertation. Doctoral dissertation, Beijing. Graduate School of Chinese Academy of Sciences
41. Rosenbaum P. R., & Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55.
42. Bewick V., Cheek L., & Ball J. (2005). Statistics review 14: Logistic regression. Critical care, 9(1), 112. https://doi.org/10.1186/cc3045 PMID: 15693993

43. Miron B. Kursa, Witold R. Rudnicki (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), p. 1–13. URL: http://www.jstatsoft.org/v36/i11/

44. Iverson L. R., Prasad A. M., Matthews S. N., & Peters M. (2008). Estimating potential habitat for 134 eastern US tree species under six climate scenarios. Forest Ecology and Management, 254(3), 390–406. https://doi.org/10.1016/j.foreco.2007.07.023

45. Metz C. E. (1978, October). Basic principles of ROC analysis. In Seminars in nuclear medicine (Vol. 8, No. 4, pp. 283–298). WB Saunders. https://doi.org/10.1016/S0001-2998(78)80014-2 PMID: 112681

46. Davidson R. J., Pizzagalli D., Nitschke J. B., & Putnam K. (2002). Depression: perspectives from affective neuroscience. Annual review of psychology, 53(1), 545–574. https://doi.org/10.1146/annurev.psych.53.100901.135148 PMID: 11752496

47. Siegle G. J., Steinhauer S. R., Thase M. E., Stenger V. A., & Carter C. S. (2002). Can't shake that feeling: event-related fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals. Biological psychiatry, 51(9), 693–707. PMID: 11983183

48. Murray I. R., & Arnott J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America, 93(2), 1097–1108. https://doi.org/10.1121/1.405558 PMID: 8445120

49. Abelin, Å., & Allwood, J. (2000). Cross linguistic interpretation of emotional prosody. In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

50. Scherer, K. R. (2000). A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. In Sixth International Conference on Spoken Language Processing.

51. Bhatti, M. W., Wang, Y., & Guan, L. (2004, May). A neural network approach for human emotion recognition in speech. In Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on (Vol. 2, pp. II-181). IEEE. doi: 10.1109/ISCAS.2004.1329238

52. El Ayadi M., Kamel M. S., & Karray F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

53. Lieberman P., & Michaels S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. The Journal of the Acoustical Society of America, 34(7), 922–927. https://doi.org/10.1121/1.1918222

54. Toda T., Black A. W., & Tokuda K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Transactions on Audio, Speech, and Language Processing, 15 (8), 2222–2235. https://doi.org/10.1109/TASL.2007.907344

55. Vaissière J. (1983). Language-independent prosodic features. In Prosody: Models and measurements (pp. 53–66). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-69103-4_5

56. Coker C. H. (1976). A model of articulatory dynamics and control. Proceedings of the IEEE, 64(4), 452–460. https://doi.org/10.1109/PROC.1976.10154

57. Crystal, D. (1976). Prosodic systems and intonation in English(Vol. 1). CUP Archive.

58. Darby J. K., & Hollien H. (1977). Vocal and speech patterns of depressive patients. Folia Phoniatrica et Logopaedica, 29(4), 279–291. https://doi.org/10.1159/000264098

59. Frick R. W. (1985). Communicating emotion: The role of prosodic features. Psychological Bulletin, 97 (3), 412. http://dx.doi.org/10.1037/0033-2909.97.3.412

60. Afshan A., Guo J., Park S. J., Ravi V., Flint J., & Alwan A. (2018). Effectiveness of Voice Quality Features in Detecting Depression. In Proc. Interspeech (pp. 1676–1680).

61. Guo J., Xu N., Qian K., Shi Y., Xu K., Wu Y., & Alwan A. (2018). Deep neural network based i-vector mapping for speaker verification using short utterances. Speech Communication, 105, 92–102.

62. Guo J., Yang R., Arsikere H., & Alwan A. (2017). Robust speaker identification via fusion of subglottal resonances and cepstral features. the Journal of the Acoustical Society of America, 141(4), EL420–EL426. https://doi.org/10.1121/1.4979841 PMID: 28464674

63. Guo J., Nookala U. A., & Alwan A. (2017). CNN-Based Joint Mapping of Short and Long Utterance i-Vectors for Speaker Verification Using Short Utterances. In INTERSPEECH (pp. 3712–3716).

64. Guo J., Yeung G., Muralidharan D., Arsikere H., Afshan A., & Alwan A. (2016). Speaker Verification Using Short Utterances with DNN-Based Estimation of Subglottal Acoustic Features. In INTERSPEECH (pp. 2219–2222).