The Discriminatory Power of Vocal Features in Detecting Mental Illnesses Under Complex Context

Wei Pan
Central China Normal University
Liat Shenhav
University of California, Los Angeles
Amber Afshan
University of California, Los Angeles
Abeer Alwan
University of California, Los Angeles
Jonathan Flint
University of California, Los Angeles
Tianli Liu
Peking University
Bin Hu
Lanzhou University
Tingshao Zhu (tszhu@psych.ac.cn)
Chinese Academy of Sciences

Research Article

Keywords: depression, healthy controls, schizophrenia, bipolar disorder, i-vectors, logistic regression

Posted Date: March 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-341817/v1

Abstract

Background: Vocal features have been proposed as a way to identify depression by distinguishing depression from healthy controls, but while there have been some claims for success, the degree to which changes in vocal features are specific to depression has not been systematically studied. In particular, it is not clear whether vocal features are characteristic of mental ill health in general, rather than characteristic of different psychiatric diagnoses. We examined the performance of vocal features in recognizing three diseases (depression, schizophrenia and bipolar disorder) in comparison with healthy controls and in pairwise comparison with each disease in turn.

Methods: We sampled 32 bipolar disorder patients; 106 depression patients, 114 healthy controls and 20 schizophrenia patients. We extracted i-vectors from MFCCs features, and built logistic regression models with ridge regularization and 5 fold cross validation on the training set, then applied models to the test set.

Results: Our results showed that AUC score for classifying depression and bipolar disorder is 0.5 (F-score = 0.44). For other comparisons, AUC scores range from 0.75 to 0.92 (F-score ranges from 0.73 to 0.91). The performance (AUC) of depression and bipolar disorder classification model is significantly worse than the performance of bipolar disorder and schizophrenia classification model (corrected $P < 0.05$). We found no significant difference in pairwise ROC difference tests among the remaining classification tasks.

Conclusions: Vocal features have robust discriminatory power not only in classifying depression and health, but also in pairwise classification among different mental illness.

Introduction

The identification and diagnosis of depression still relies on a clinical interview, which is often slow and unreliable: inter-rater reliability rarely exceeds of 0.7 (kappa coefficient)(Spitzer, Forman, & Nee, 1979). Furthermore, about half of cases go unrecognized: in a meta-analysis of 41 studies recognition accuracy of depression by general practitioners was 47.3%(Mitchell, Vaze, & Rao, 2009). Accurate, and fast ways to identify cases of depression would have major clinical benefits.

Novel applications of computational methods are making some inroads into this problem. A review of 14 studies comparing diagnostic accuracies between deep learning models and health-care professionals showed that both sensitivity and specificity of deep learning models are higher than that of health-care professionals(X. Liu et al., 2019). In the last decade there has been interest in the ability of using speech to identify depression, and other mental illness(Alberto, Arndis, Vibeke, Riccardo, & Parola, 2019; Cummins et al., 2015; Faurholt-Jepsen et al., 2016; Maxhuni et al., 2016; Rapcan et al., 2010).

Most research in depression detection has focused on examining the ability of vocal features to classify depression and health with machine learning models to investigate whether voice can be used as

auxiliary tool assisting clinical diagnosis(Afshan et al., 2018; He & Cao, 2018; Jiang et al., 2018; Pan et al., 2019; Rohani, Faurholt-Jepsen, Kessing, & Bardram, 2018; Taguchi et al., 2018). Investigating the ability of voice features to diagnose depression is complicated by the presence of other mental illness. Various speech features have been shown to be indicative of depression. Mel-frequency cepstrum coefficients (MFCCs) have been shown to reflect perception relevant information, which are most commonly used in speech recognition and mental illness detection(Zhu, Kim, Proctor, Narayanan, & Nayak, 2013). Investigation of MFCCs by Cummins et al.(Cummins, Epps, Sethu, & Krajewski, 2014; Cummins et al., 2011) and Joshi et al.(Joshi et al., 2013) found that the classification results were statistically significant for detecting depression.

Identity vector (i-vector), using total variability framework, is state-of-the-art technique for speaker-verification. The total variability framework provides an effective way to capture speaker variability and channel variability in a low dimensional subspace. i-vectors are very informative in capturing the cepstral variability, and can predict depression with high and classification accuracy. Research found that the i-vector based model outperformed the baseline model set by KL-means supervectors (Cummins et al., 2013). Nasir et al.(Nasir, Jati, Shivakumar, Chakravarthula, & Georgiou, 2016) used i-vectors to investigate a number of audio and video features for classification, and reported a highest accuracy with i-vector modelling based on MFCC features. Another research showed an improvement of 40% for predictive accuracy (F-score) with the i-vector method(Afshan et al., 2018).

Voice features characteristic of other psychiatric diseases have also been reported. A review and meta-analysis of voice in schizophrenia identified weak atypicalities in pitch variability related to flat affect, and stronger atypicalities in the proportion of spoken time, speech rate, and pauses related to the alogia and flat affect(Alberto et al., 2019). Rapcan et al.(Rapcan et al., 2010) extracted temporal, energy and vocal pitch features from recordings, results showed that when classifying schizophrenia patients from healthy ones, the classification accuracy reached 79.4%. For bipolar disorder, researchers using voice features to predict their different emotion states (i.e. depressive state, manic state, mixed state). Classification results showed the AUC score reached 0.89 (Faurholt-Jepsen et al., 2016; Maxhuni et al., 2016).

In this research, we aimed at illustrating the classification ability of voice features under different scenarios. There are three types of binary classification tasks in total: 1) the ability of voice features on general classification, i.e. classifying any disease (depression, bipolar disorder, schizophrenia) vs. healthy-controls; 2) the ability of voice features on classifying specific mental disorder vs. healthy-controls; 3) the ability of voice features on pairwise among different mental disorders. Among these, type 1 was set as baseline level to show whether the dataset we used achieves results comparable to existing research, and as a reference of how models performing under other types in this research. Type 2 and type 3 was set to show to what extent voice features can differentiate case group and control group.

Methods

Dataset

All participants are Chinese aged between 18 ~ 59 years old. There are four categories according to diagnosis: health group (58 males; 57 females), depression group (53 males; 70 females), bipolar disorder group (16 males; 21 females) and schizophrenia group(10 males; 10 females). All patients were assessed with the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)(American Psychiatric Association. (1994) by psychiatrists. Depressed patients were diagnosed as having depression only, with no other mental disorders or medical tasks. Exclusion criteria for all patients are: serious physical illness, pregnant and lactating, alcohol and other substance abuse in one year. Participants didn't take any medication in two weeks before the experiment.

For each participant, there are 21 voice recordings collected from 21 speech tasks. Ambient noise level of the lab was less than 60 dB. Audio length ranged between 10s ~ 2min30s. The database is in Mandarin. All the audio recordings were collected with a sampling rate of 16kHz and saved as .wav format. Recordings of high sound quality were selected. Participants provided written informed consent before the experiment. This project is conducted from August 25, 2015 to September 29, 2017, approved by the Institutional Review Board (IRB) of Institute of Psychology, Chinese Academy of Sciences. All methods performed in this research were in accordance with the relevant guidelines and regulations.

Data processing

Preprocessing

There are 7 classification tasks in total: any disease vs. healthy controls; depression vs. healthy controls; bipolar disorder vs. healthy controls; schizophrenia vs. healthy controls; depression vs. bipolar disorder; depression vs. schizophrenia; bipolar disorder vs. schizophrenia (AH, DH, BH, SH, DB, DS, BS). In each task, either mental illness group or depression group is set as case group, the other one is set as control group.

First, we split the data into training set (70%) and test set (30%), we calculated the sample size needed for the test set. To avoid distribution bias, data was matched between groups for each binary classification task. We randomly sampled the exact number for each task. For training set, we also did random sampling in the rest of the dataset for each category. As the data set is small, we conducted difference test based on permutation test. Age, gender and education variables were investigated to see if the data is balanced on the training set.

It should be noted that for cases and controls of each classification task, there is a further matching. In total, we did the matching for two times: 1) matching gender within categories; 2) matching cases and controls within classification tasks. As a result, the number for cases and controls in each task is balanced. For example, there are 20 schizophrenia patients and 20 healthy ones (gender balanced) in the model of classifying health and schizophrenia.

Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients (MFCCs) were extracted with a window size of 25 ms, a window shift of 10 ms, a pre-emphasis filter with coefficient 0.97, and a sinusoidal lifter with coefficient 22(Afshan et al., 2018). A filter bank with 23 filters was used and 13 coefficients were extracted. Utterances were downsampled to 8 kHz before feature extraction. We also used the first and second derivatives of MFCCs.

I-vector extraction

After MFCCs feature extraction, a UBM of 64 mixtures was trained for each feature set. Followed by total variability matrix calculation, and used it to extract i-vectors of dimension 20.

The i-vector extraction formula can be represented as follows:

$$M = m + Tv \quad (1)$$

where m is the mean super-vector of the UBM. M is the mean centered super-vector of the speech utterance derived using the 0th and 1st order Baum-Welch statistics. v is the i-vector, the representation of a speech utterance(Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2011).

20 i-vectors were generated for each participant. All i-vector features were quantile normalized. Then the mean value of 21 voice tasks for each participant was calculated for classification. Then the training data was used for building logistic regression model.

Logistic Regression

Using i-vectors, we performed logistic regression on the training set. Glmnet method(Friedman, Hastie, & Tibshirani, 2010) was conducted to achieve logistic regression with 5-fold cross validation. And ridge regularization was performed. Then we applied models on test sets in each classification task.

Model Building

To compare the classification ability of voice features under different situation, we built logistic models for 7 binary classification tasks. First, we examined the classification ability of voice features in distinguishing between health and mental disorders in general. Second, we examined model performance in classifying health and specific mental disorders separately. After this, we investigate model performance in specific mental illness classification context (depression vs. bipolar disorder; depression vs. schizophrenia; bipolar disorder vs. schizophrenia).

Results

Descriptive statistics

Demographic information

Demographic information was investigated to see if there was any distribution bias between groups. There were three demographic variables: age, gender, education.

Descriptive statistics for within model difference test

Age difference was tested and showed in Table 1. Results showed that except for the classification task schizophrenia vs. health, cases and controls were matched for the relevant variables. We further conducted propensity score matching(Dehejia & Wahba, 2002) for the schizophrenia vs. health task. Case group and control group in this task is then balanced, see Table 2.

After matching, in total, there are 32 bipolar disorder patients; 106 depression patients, 114 healthy ones and 20 schizophrenia patients (equal number of male and female in each category).

Table 1
t tests of age difference for each classification task

	Type	*M ± SD*	*t*	*P*
AH	**any disease**	32.35 ± 9.82	-1.42	0.16
	healthy controls	34.48 ± 10.63		
DH	**depression**	33.76 ± 10.23	-0.4	0.69
	healthy controls	34.43 ± 10.48		
BH	**bipolar disorder**	29.68 ± 8.99	-1.85	0.07
	healthy controls	34.86 ± 9.03		
SH	**schizophrenia**	29.07 ± 7.42	-2.11	0.03
	healthy controls	37.14 ± 7.42		
DB	**depression**	34.18 ± 9.67	1.57	0.12
	bipolar disorder	29.68 ± 8.99		
DS	**depression**	33.86 ± 9.49	1.45	0.15
	schizophrenia	29.07 ± 7.42		
BS	**bipolar disorder**	30.00 ± 7.75	0.33	0.77
	schizophrenia	29.07 ± 7.42		

Table 2
One-way test after propensity score matching for SH task

	Type	*M ± SD*	*t*	*P*
S_H	**schizophrenia**	28.79 ± 8.15	-0.74	0.48
	healthy controls	31.14 ± 8.74		

Chi-square test with permutation test was conducted on both gender variable and education variable. Results showed that there is no significant difference either on gender or education across different tasks. See Table 3 and Table 4.

Table 3
Chi-square tests on gender for each classification task

	Type	Gender		χ^2	*P*
		Male	**Female**		
AH	**any disease**	55	55	0	1
	healthy controls	40	40		
DH	**depression**	37	37	0	1
	healthy controls	37	37		
BH	**bipolar disorder**	11	11	0	1
	healthy controls	11	11		
SH	**schizophrenia**	7	7	0	1
	healthy controls	7	7		
DB	**depression**	11	11	0	1
	bipolar disorder	11	11		
DS	**depression**	7	7	0	1
	schizophrenia	7	7		
BS	**bipolar disorder**	7	7	0	1
	schizophrenia	7	7		

Table 4
Chi-square tests on education for each classification task

	Type	Education		χ^2	*P*
		High school level and below	**Undergraduate and above**		
AH	**any disease**	46	64	0.18	0.76
	healthy controls	31	49		
DH	**depression**	29	45	0.03	1
	healthy controls	28	46		
BH	**bipolar disorder**	9	13	0	1
	healthy controls	9	13		
SH	**schizophrenia**	6	8	1.29	0.45
	healthy controls	9	5		
DB	**depression**	10	12	0.09	1
	bipolar disorder	9	13		
DS	**depression**	5	9	1.29	0.45
	schizophrenia	8	6		
BS	**bipolar disorder**	7	7	0.14	1
	schizophrenia	8	6		

Classification

Model performance obtained for different tasks are summarized in Table 5. For general classification ability of voice features, when classifying health vs. non-health group (AH), F-score is 0.82, AUC is 0.79. When classifying different mental diseases from the healthy ones, results showed that: for DH task, F-score = 0.78, AUC = 0.77; for BH task, F-score = 0.80, AUC = 0.80; for SH task, F-score = 0.73, AUC = 0.75. To further examine the distinguishing ability of voice features on pairwise classification among three mental diseases, DB, DS and BS tasks were performed. Results showed: for DB classification task, F-score = 0.44, AUC = 0.50; for DS task, F-score = 0.83, AUC = 0.83; for BS task, F-score = 0.91, AUC = 0.92.

Table 5
Results on 7 classification tasks with the i-vector framework.

Tasks	*Sensitivity*	*Specificity*	*Accuracy*	*Precision*	*Recall*	*F-score*	*AUC*
AH	0.81	0.76	0.79	0.83	0.81	0.82	0.79
DH	0.81	0.72	0.77	0.74	0.81	0.78	0.77
BH	0.80	0.80	0.80	0.80	0.80	0.80	0.80
SH	0.67	0.83	0.75	0.80	0.67	0.73	0.75
DB	0.40	0.60	0.50	0.50	0.4	0.44	0.50
DS	0.83	0.83	0.83	0.83	0.83	0.83	0.83
BS	0.83	1.00	0.92	1.00	0.83	0.91	0.92

We also compared model performance for all tasks. After Bonferroni correction, only the performance (AUC) of depression and bipolar disorder classification model is significantly worse than the performance of bipolar disorder and schizophrenia classification model ($P < 0.05$), see Table 6.

Table 6
ROC difference tests for model comparison

	DeLong's test	*df*	*P*	*Corrected P*
AH_DH	0.32	134.67	0.75	4.5
AH_BH	-0.11	29.02	0.91	5.46
AH_SH	0.27	13.78	0.79	4.74
AH_DB	2.32	25.54	0.03	0.18
AH_DS	-0.35	14.66	0.73	4.38
AH_BS	-1.34	18.71	0.2	1.2
DH_BH	-0.32	32.18	0.75	4.5
DH_SH	0.11	14.7	0.92	5.52
DH_DB	2.09	27.64	0.046	0.276
DH_DS	-0.52	15.88	0.61	3.66
DH_BS	-1.53	21.29	0.14	0.84
BH_SH	0.3	21.48	0.76	4.56
BH_DB	2.01	36.64	0.052	0.312
BH_DS	-0.22	23.92	0.83	4.98
BH_BS	-0.93	29.35	0.36	2.16
SH_DB	1.41	25.27	0.17	1.02
SH_DS	-0.47	21.63	0.65	3.9
SH_BS	-1.05	18.37	0.31	1.86
DB_DS	-2.02	27.56	0.053	0.318
DB_BS	-2.93	29.93	0.006	0.036*
DS_BS	-0.58	19.8	0.57	3.42

* $P < 0.05$

Discussion

We investigated the classification ability of voice features in different scenarios. Results showed that voice features can assist clinical diagnosis.

First, after matching, descriptive statistics investigation indicates no distribution bias between case group and control group in each task. By achieving this we ruled out the potential confounding factors(Pan et al., 2019) for voice predicting depression. Using MFCCs features, we built several logistic regression models for different classification scenarios with the i-vector method.

Type 1 includes AH and DH models to examine the baseline classification ability voice features have. Given the results, when classifying health and non-health, the F-score is 0.82, AUC score is 0.79. This showed to what extent voice features can distinguish mental illness from healthy controls.

Type 2 consists of DH, BH and SH models to investigate the ability of voice features in distinguishing specific mental illness from healthy ones. For DH, BH and SH tasks, the F-score ranged from 0.73 to 0.80, the AUC score ranged from 0.75 to 0.80. There already exist lots of research classifying depression and health using voice features. Our results about DH are consistent with the findings of others(Afshan et al., 2018; Alghowinem et al., 2013; Cummins et al., 2015; Horwitz et al., 2013; Quatieri & Malyska, 2012; Pan et al., 2019; Sidorov & Minker, 2014), so that we can take this task performance as another baseline. The DH task results illustrated the effectiveness of our dataset. ROC difference test showed there is no significant difference in the pairwise comparison among AH and the three mental illnesses vs. healthy tasks.

Type 3 comprises DB, DS, BS models to further show the performance of voice features on pairwise classification among the three mental illnesses. Model performance of both DS (F-score = 0.83; AUC = 0.83) and BS (F-score = 0.91; AUC = 0.92) are ideal. In fact, the F-score and AUC score for BS task are both the highest across all 7 tasks. But model performance of DB is the worst (F-score = 0.44) among all tasks, and the AUC score 0.5 means voice features doesn't contribute to the discrimination between depression and bipolar disorder.

Further pairwise ROC test showed there is no significant difference of model performance among AH, DH, BH, SH, DS and BS tasks. But for DB and BS comparison, the model performance of DB is significantly worse than that of BS model. See Table 6.

To our knowledge, this is the first research examining the discriminatory power of voice features concerning depression and other mental illnesses with similar depression symptoms. It showed that voice features can be applied not only on classifying depression and health, but also on detecting other mental illnesses, with considerable amount of predicting accuracy. This research fully described the distinguishing abilities of voice features under different classification scenarios.

The results are promising, which might be because the i-vectors are able to catch the mental illness relevant information. To extract i-vectors, first it needs to put both case group and control group together to learn the shared information between the two group, and then this shared part is removed from the voice data to get the i-vectors, which means that i-vectors capture mental illness relevant voice information. i-vector based system has been proved to be effective in both short and long utterances from 10s to 5 min(Guo, Nookala, & Alwan, 2017; Guo et al., 2018, 2016; Guo, Yang, Arsikere, & Alwan,

2017). And clearly, it also captures different voice information for different mental illnesses when classifying between mental illnesses. Different mental illness has different symptom in voice. For example, vocal features observed to change with a depression patient's mental condition and emotional state is motivated by perception of monotony, hoarseness, breathiness, glottalization, and slur in the voice of a depressed subject(France & Shiavi, 2000; Low, Maddage, Lech, Sheeber, & Allen, 2010; Moore, Clements, Peifer, & Weisser, 2003; Mundt, Snyder, Cannizzaro, Chappie, & Geralts, 2007; Ozdas, Shiavi, Silverman, Silverman, & Wilkes, 2004; Trevino, Quatieri, & Malyska, 2011). The voice of schizophrenia patients can be described as poverty of speech, increased pauses, distinctive tone and intensity of voice, which has been associated with core negative symptoms such as diminished emotional expression, lack of vocal intonation and alogia, having difficulty in controlling voice to express affective and emotional contents in proper social text(Alpert, Rosenberg, Pouget, & Shaw, 2000; Cohen, Mitchell, Docherty, & Horan, 2016; Cohen, Najolia, Kim, & Dinzeo, 2012; Galynker, Cohen, & Cai, 2000; Guo, Xu, et al., 2018; Hoekert, Kahn, Pijnenborg, & Aleman, 2007; Millan, Fone, Steckler, & Horan, 2014; Trémeau et al., 2005).

This research failed to classify depression and bipolar disorder. Unipolar depression and bipolar depression are quite similar. Bipolar disorder is combined with depression phase and also manic phase, that's why it's so hard to distinguish between the two them. Research about bipolar disorder usually tracks user phone call recordings to extract voice features and detecting different emotion states(depression state, manic state)(Faurholt-Jepsen et al., 2016; Maxhuni et al., 2016). We still should working on tracking emotion states to help distinguish depression from bipolar disorder or finding some other ways.

Conclusion

This research fully examined the discriminatory power of voice features under different clinical scenarios for mental illness diagnosis. Robust results illustrate the clinical value for voice features in complex clinical diagnosis for depression. In future work, tracking illness episode for depression and other mental illnesses and larger sample size might get higher and more accurate results.

Declarations

Ethics approval and consent to participate

All patients provided informed consent for their participation in this study after the procedure had been fully explained to them, and the study protocol was approved by the ethical board of the Institute of Psychology, Chinese Academy of Science.

Consent for publication

Not applicable.

Availability of data and materials

Datasets of this study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by National Basic Research Program of China (2014CB744600) and the Key Research Program of the Chinese Academy of Sciences (ZDRW-XH-2019-4). They only provide funding.

Authors' Contributions

In this article, W.P. was in charge of overall data analysis and drafting the paper. L.S. helped on data analysis. A.Af. and A.Al. helped extracting MFCC features and i-vectors. J.F. helped with the analysis and revising the manuscript. T.L. and B.H. were in charge of collecting and organizing the data. T.Z. was in charge of the whole data collecting, analyzing, and paper drafting procedures.

Acknowledgements

We thank generous support from National Basic Research Program of China (2014CB744600) and the Key Research Program of the Chinese Academy of Sciences (ZDRW-XH-2019-4).We are also very grateful to everyone for their time and effort during data collection.

References

1. Afshan, A., Guo, J., Park, S. J., Ravi, V., Flint, J., & Alwan, A. (2018). Effectiveness of Voice Quality Features in Detecting Depression. https://doi.org/10.21437/Interspeech.2018-1399
2. Alberto, P., Arndis, S., Vibeke, B., & Riccardo, F. (2019). Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis.*bioRxiv*, 583815.
3. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., & Parker, G. (2013). Detecting depression: A comparison between spontaneous and read speech. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 7547–7551. https://doi.org/10.1109/ICASSP.2013.6639130
4. Alpert, M., Rosenberg, S. D., Pouget, E. R., & Shaw, R. J. (2000). Prosody and lexical accuracy in flat affect schizophrenia. *Psychiatry Research*, *97*(2–3), 107–118. https://doi.org/10.1016/S0165-1781(00)00231-6
5. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-IV). American Psychiatric Pub, 1994
6. Cohen, A. S., Mitchell, K. R., Docherty, N. M., & Horan, W. P. (2016). Vocal expression in schizophrenia: Less than meets the ear. *Journal of Abnormal Psychology*, *125*(2), 299–309. https://doi.org/10.1037/abn0000136

7. Cohen, A. S., Najolia, G. M., Kim, Y., & Dinzeo, T. J. (2012). On the boundaries of blunt affect/alogia across severe mental illness: Implications for Research Domain Criteria. *Schizophrenia Research*, *140*, 41–45. https://doi.org/10.1016/j.schres.2012.07.001
8. 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. *In2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)(pp. 970-974). IEEE.*
9. 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
10. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(4), 788–798. https://doi.org/10.1109/TASL.2010.2064307
11. Dehejia, R. H., & Wahba, S. (2002, February). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, Vol. 84, pp. 151–161. https://doi.org/10.1162/003465302317331982
12. Faurholt-Jepsen, M., Busk, J., Frost, M., Vinberg, M., Christensen, E. M., Winther, O., … Kessing, L. V. (2016). Voice analysis as an objective state marker in bipolar disorder. *Nature Publishing Group*, *6*, 856. https://doi.org/10.1038/tp.2016.123
13. France, D. J., & Shiavi, R. G. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, *47*(7), 829–837. https://doi.org/10.1109/10.846676
14. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01
15. Galynker, I. I., Cohen, L. J., & Cai, J. (2000). Negative symptoms in patients with major depressive disorder: a preliminary report. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *13*(3), 171–176. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10910087
16. 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*. https://doi.org/10.21437/Interspeech.2017-430
17. Guo, J., Xu, N., Chen, X., Shi, Y., Xu, K., & Alwan, A. (2018). *Filter sampling and combination CNN (FSC-CNN): a compact CNN model for small-footprint ASR acoustic modeling using raw waveforms*. https://doi.org/10.21437/Interspeech.2018-1370
18. 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* (Vol. 00).
19. Guo, J., Yang, R., Arsikere, H., & Alwan, A. (2017). *Robust speaker identification via fusion of subglottal resonances and cepstral features*. https://doi.org/10.1121/1.4979841

20. 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*. https://doi.org/10.21437/Interspeech.2016-282
21. He, L., & Cao, C. (2018). Automated depression analysis using convolutional neural networks from speech. *Journal of Biomedical Informatics*, *83*, 103–111. https://doi.org/10.1016/j.jbi.2018.05.007
22. Hoekert, M., Kahn, R. S., Pijnenborg, M., & Aleman, A. (2007, November). Impaired recognition and expression of emotional prosody in schizophrenia: Review and meta-analysis. *Schizophrenia Research*, Vol. 96, pp. 135–145. https://doi.org/10.1016/j.schres.2007.07.023
23. 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. *2013 IEEE International Conference on Body Sensor Networks, BSN 2013*. https://doi.org/10.1109/BSN.2013.6575522
24. 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.
25. Quatieri, T. F., & Malyska, N. (2012). Vocal-source biomarkers for depression: A link to psychomotor activity*. In Thirteenth Annual Conference of the International Speech Communication Association.*
26. Jiang, H., Hu, B., Liu, Z., Wang, G., Zhang, L., Li, X., & Kang, H. (2018). Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech Features. *Computational and Mathematical Methods in Medicine*, *2018*, 6508319. https://doi.org/10.1155/2018/6508319
27. Joshi, J., Goecke, R., Alghowinem, S., Dhall, A., Wagner, M., Epps, J., … Breakspear, M. (2013). Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces*, *7*(3), 217–228. https://doi.org/10.1007/s12193-013-0123-2
28. Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., … Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, *1*(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2
29. Low, L. S. A., Maddage, N. C., Lech, M., Sheeber, L., & Allen, N. (2010). Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 5154–5157. https://doi.org/10.1109/ICASSP.2010.5495018
30. Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., & Morales, E. F. (2016). Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing*. https://doi.org/10.1016/j.pmcj.2016.01.008
31. Millan, M. J., Fone, K., Steckler, T., & Horan, W. P. (2014). Negative symptoms of schizophrenia: Clinical characteristics, pathophysiological substrates, experimental models and prospects for improved treatment. *European Neuropsychopharmacology*, *24*(5), 645–692. https://doi.org/10.1016/j.euroneuro.2014.03.008

32. 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

33. Moore, E., Clements, M., Peifer, J., & Weisser, L. (2003). Analysis of Prosodic Variation in Speech for Clinical Depression. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, *3*, 2925–2928. https://doi.org/10.1109/iembs.2003.1280531

34. 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

35. Nasir, M., Jati, A., Shivakumar, P. G., Chakravarthula, S. N., & Georgiou, P. (2016). Multimodal and multiresolution depression detection from speech and facial landmark features. *AVEC 2016 - Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Co-Located with ACM Multimedia 2016*, 43–50. https://doi.org/10.1145/2988257.2988261

36. 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 Transactions on Biomedical Engineering*, *51*(9), 1530–1540. https://doi.org/10.1109/TBME.2004.827544

37. Pan, W., Flint, J., Shenhav, L., Liu, T., Liu, M., Hu, B., & Zhu, T. (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

38. Rapcan, V., D'Arcy, S., Yeap, S., Afzal, N., Thakore, J., & Reilly, R. B. (2010). Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. *Medical Engineering and Physics*, *32*(9), 1074–1079. https://doi.org/10.1016/j.medengphy.2010.07.013

39. Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2018, August 1). Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: Systematic review. *JMIR MHealth and UHealth*, Vol. 6. https://doi.org/10.2196/mhealth.9691

40. Sidorov, M., & Minker, W. (2014). Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach. *AVEC 2014 - Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Workshop of MM 2014*, 81–86. https://doi.org/10.1145/2661806.2661816

41. Spitzer, R. L., Forman, J. B. W., & Nee, J. (1979). DSM-III field trials: I. Initial interrater diagnostic reliability. *American Journal of Psychiatry*, *136*(6), 815–817. https://doi.org/10.1176/ajp.136.6.815

42. Taguchi, T., Tachikawa, H., Nemoto, K., Suzuki, M., Nagano, T., Tachibana, R., … Arai, T. (2018). Major depressive disorder discrimination using vocal acoustic features. *Journal of Affective Disorders*, *225*, 214–220. https://doi.org/10.1016/j.jad.2017.08.038

43. Trémeau, F., Malaspina, D., Duval, F., Corrêa, H., Hager-Budny, M., Coin-Bariou, L., … Gorman, J. M. (2005). Facial expressiveness in patients with schizophrenia compared to depressed patients and

nonpatient comparison subjects. *American Journal of Psychiatry*, *162*(1), 92–101. https://doi.org/10.1176/appi.ajp.162.1.92

44. Trevino, A. C., Quatieri, T. F., & Malyska, N. (2011). Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, *2011*(1), 42. https://doi.org/10.1186/1687-6180-2011-42
45. Zhu, Y., Kim, Y. C., Proctor, M. I., Narayanan, S. S., & Nayak, K. S. (2013). Dynamic 3-D visualization of vocal tract shaping during speech. *IEEE Transactions on Medical Imaging*, *32*(5), 838–848. https://doi.org/10.1109/TMI.2012.2230017