Journal of Affective Disorders 288 (2021) 161–166

Contents lists available at ScienceDirect

Journal of Affective Disorders

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

Research paper

Using i-vectors from voice features to identify major depressive disorder

Yazheng Di [a,b], Jingying Wang [c], Weidong Li [d,*], Tingshao Zhu [a,b,**]

[a] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, China
[b] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China
[c] School of Optometry, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[d] Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Keywords:
Depression
Biological markers
Clinical trials
Computer/internet technology
Assessment/Diagnosis

ABSTRACT

Background: Machine-learning methods using acoustic features in the diagnosis of major depressive disorder (MDD) have insufficient evidence from large-scale samples and clinical trials. This study aimed to evaluate the effectiveness of the promising i-vector method on a large sample of women with recurrent MDD diagnosed clinically, examine its robustness, and provide an explicit acoustic explanation of the i-vectors.

Methods: We collected utterances edited from clinical interview speech records of 785 depressed and 1,023 healthy individuals. Then, we extracted Mel-frequency cepstral coefficient (MFCC) features and MFCC i-vectors from their utterances. To examine the effectiveness of i-vectors, we compared the performance of binary logistic regression between MFCC i-vectors and MFCC features and tested its robustness on different utterance durations. We also determined the correlation between MFCC features and MFCC i-vectors to analyze the acoustic meaning of i-vectors.

Results: The i-vectors improved 7% and 14% of area under the curve (AUC) for MFCC features using different utterances. When the duration is > 40 s, the classification results are stabilized. The i-vectors are consistently correlated to the maximum, minimum, and deviations of MFCC features (either positively or negatively).

Limitations: This study included only women.

Conclusions: The i-vectors can improve 14% of the AUC on a large-scale clinical sample. This system is robust to utterance duration > 40 s. This study provides a foundation for exploring the clinical application of voice features in the diagnosis of MDD.

1. Introduction

The deployment of objective assessments of psychiatric phenomenology would transform the ability to diagnose, treat and prevent major depressive disorder (MDD). MDD is very common, affecting almost one in ten people (Demyttenaere et al., 2004; Kessler et al., 2003) and recently recognized as the world's leading cause of disability(World Health Organization, 2017). Yet currently only about half of those suffering MDD are detected and offered treatment(Goldberg, 1995; Wells et al., 1989). One of the main obstacles preventing the effective use of existing therapies is the difficulty in the diagnosis of MDD. Diagnosis is still based on clinical interviews and mental status examination (Regier et al., 2013). Screening instruments are hindered by poor specificity and sensitivity, and there are no reliable biomarkers.

Further complicating the issue is that MDD is a syndromal diagnosis, leaving open the possibility that it consists of various separate conditions (Alexopoulos et al., 1997; Gustafsson et al., 2015; K. S. Kendler et al., 2001, 2006, 2013; Masters et al., 2015; Peterson et al., 2018), each with distinct outcomes and reaction to treat. One promising method to improve the objectivity of psychiatric assessment is the use of biological and behavioral indices, such as acoustic features(D. M. Low et al., 2020). Speech samples from clinical interviews are non-intrusive and easily accessible.

Recently, various machine-learning methods using acoustic features have been used in the diagnosis of MDD, but there is insufficient evidence from large-scale samples and clinical trials to prove their effectiveness. A review of articles assessing depression using voice features shows that only 38% of experiments analyzed subjects with a clinical diagnosis of MDD and the median sample size was 123 (D. M. Low et al., 2020). The widely used AVEC data sets (Valstar et al., 2016) for

* Corresponding author at: Shanghai Jiao tong University, 800 Dongchuan Road, Shanghai, 200240, China.
** Corresponding author at: CAS Key Laboratory of Behavioral Science, Institute of Psychology, 16 Lincui Road, Chaoyang District, Beijing 100101, China.
E-mail addresses: liwd@sjtu.edu.cn (W. Li), tszhu@psych.ac.cn (T. Zhu).

https://doi.org/10.1016/j.jad.2021.04.004
Received 23 February 2021; Received in revised form 27 March 2021; Accepted 2 April 2021
Available online 20 April 2021

depression employs a Patient Health Questionnaire-8 score rather than a clinical diagnosis of MDD. When using self-report measures, the goal of the study must be reoriented from predicting the diagnosis to predicting self-report questionnaire scores, which may not always be compatible with the clinical diagnosis. Large-scale trials are crucial because speaker and phonetic variabilities degrade the performance of depression detection systems(Nicholas Cummins et al., 2013), hindering the promotion and application of models that are tested on small samples.

The i-vector (Dehak et al., 2011) method provides a good paradigm for avoiding speaker and channel variability effects. The experimental results (Afshan et al., 2018; N. Cummins et al., 2014; Lopez-Otero et al., 2014; Nasir et al., 2016) demonstrated that the i-vector representation of speech for depression level estimations achieves state-of-the-art performance. However, these experiments did not adequately control the confounding variables such as gender, accent, and comorbidities, opening the possibility that their results arise from confounds in the data rather than hypothesized speech differences. The i-vector framework is utterance-level-based, which gives decisions on depression based on each utterance instead of on each person. An utterance is the smallest unit of speech. From a dialog we can obtain multiple utterances of one person. Lopez-Otero et al. (2014) found that previous information about a speaker's depressed mood dramatically improves system performance. However, some studies (Nicholas Cummins et al., 2011) assign utterance of the same person to the train and test sets.

Another problem with the utterance-based system is the highly arbitrary control over utterance duration when splitting speech segments into utterances. The differences in the richness of speech expression between depressed and healthy individuals(Alpert et al., 2001) likely have an impact on segment duration. The arbitrary editing of the audio can even erroneously improve the classification results. However, these studies did not report differences in utterance durations between cases and control groups, nor did they test the effect of speech duration on the classification results.

Here, using Mel-frequency cepstral coefficients (MFCC) i-vectors to predict MDD, we evaluate the effectiveness of this method on a large sample of women with recurrent MDD diagnosed clinically. We examine the robustness of utterance durations of this utterance-level method, and calculate the correlations of MFCC i-vectors and MFCC statistics, to provide a more explicit explanation of the i-vectors.

2. Materials and methods

2.1. Data collection

The recordings for this study were obtained as part of the collection of data for a large case-control study into the origins of major depression. This large multi-center study involving more than 20 hospitals across China, aimed to collect 24,000 cases of recurrent major depression and 24,000 controls. In this study, we report the results from the first 1808 subjects.

To obtain cases of recurrent MDD and ensure high-quality data, doctors were trained to use a computerized interview system. Interviews were recorded and the research team listens to at least two interviews from each interviewer to identify any errors in the way questions are asked and answers are interpreted. These recordings provided data for this study.

All patients with MDD were women who were aged between 30 and 60 and had two or more episodes of MDD meeting the DSM-IV criteria (Association, 1994), with the first episode occurring between 14 and 50 years of age. Patients were excluded if they had a history of bipolar disorder, drug or alcohol abuse, psychosis, or mental retardation. Control subjects were recruited from local communities. All four grandparents of both cases and controls are Han Chinese. Cases and controls were matched for location to reduce population structure effects and control the difference in accent between groups. To reduce the probability that patients with recurrent MDD would go on to develop bipolar disorder (an exclusion criterion), a minimum age was set at 30. Controls were screened by personal interview to ensure that they had no prior depressive episode and were interviewed to obtain data on environmental and other risk factors. They have a minimum age of 40 to reduce the risk of subsequent development of MDD.

The interview protocol acquired the following assessments for psychopathology: i) CIDI (WHO 1997) section of MDD expanded to include a "deep" assessment of the DSM-IV A criteria for MDD, symptoms of DSM-IV melancholia, Beck's cognitive triad (helplessness, hopefulness, and worthlessness), and irritability/anxiety; ii) CIDI section on dysthymia; iii) sections from interviews in the Virginia Adult Twin Study of Psychiatric and Substance Use Disorders (VATSPSUD) (K. S. Kendler and Prescott, 2007) for generalized anxiety disorder, panic, and five phobia subtypes (agoraphobia and social, situational, animal, and blood injury phobias); iv) brief assessments of premenstrual syndrome and postnatal depression (Cox et al., 1987; K. Kendler et al., 1992); v) smoking/nicotine dependence as assessed using the Fagerström Test for Nicotine Dependence (Heatherton et al., 1991) (alcohol and substance abuse was virtually absent in this study, so it was not assessed).

Four key environmental exposures known to be strongly associated with the risk of MDD were assessed in cases and controls: i) child sexual abuse; ii) parent-child relationships; iii) social support; and iv) stressful live events. Neuroticism was assessed using the full 23-item Eysenck personality questionnaire N scale. Family history of MDD was individually assessed in parents and full siblings using the Family History-Research Diagnostic Criteria. In each case, measures used are those developed, field-tested, and validated in the VATSPSUD studies (K. Kendler et al., 1992).

2.2. Data preprocessing

Participants' utterances were edited from recordings of the conversations between doctors and patients through the following steps: First, voice segments from the participants were selected and labeled by the questionnaire code. (Speech segments shorter than 1 s are eliminated (Ringeval et al., 2019).) Then, a participant's segments from the demographic questions are combined into one utterance named as Demo-utterance, while all segments of one participant are combined into one utterance named as All-utterance. We ignored the records that are not related to the questionnaires to ensure the homogeneity of the audio samples.

2.3. Data analysis

2.3.1. Acoustic features

2.3.1.1. MFCCs. The MFCC is a representation of the short-term power spectrum of a sound, which more closely approximates the human auditory system's response than the linearly-spaced frequency bands used in the normal cepstrum. Many studies have found the difference in MFCCs between depressed and healthy individuals (Pan et al., 2019; Wang et al., 2019). Moreover, it has been used as input features in machine-learning(L. A. Low et al., 2009) and deep-learning models (Afshan et al., 2018), which could be a baseline for our experimental results. In this study, MFCCs were extracted with a window size of 25 ms, a window shift of 10 ms, a pre-emphasis filter with a coefficient of 0.97, and a sinusoidal lifter with a coefficient of 22. A filter bank with 23 filters was used, and 12 coefficients were extracted. Utterances were downsampled to 8 kHz before feature extraction. We also used the first and second derivatives of MFCCs.

2.3.2. i-vector extraction

We followed the approach described by Dehak et al. (2011) to extract the i-vectors. To acquire frame-level features, the Universal Background Model (UBM) which represents the feature distribution of the acoustic

space, is adapted to a set of given speech frames, to estimate utterance-dependent Gaussian Mixture Models parameters. The adaptation technique (Kenny et al., 2005) assumes that all pertinent variability is captured by a low rank rectangular matrix T known as the total variability matrix. The i-vector extraction can be shown as follows:

$$M = m + Tv$$

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., and v is the i-vector, the representation of a speech utterance. In this study, we set the number of Gaussian mixtures as 256 and the i-vector dimension as 200.

2.3.3. Classification modeling

In this study, we trained a binary classifier on two classes: depressed and non-depressed. We used a binary logistic regression algorithm to train the classifier. The data were split into train and test sets by randomly assigning 70% of the speakers to the train set and 30% to the test set exclusively.

We obtained the classification results using only utterance durations as baseline and evaluated the performance using different utterances and features. We reported sensitivity, specificity (Parikh et al., 2008), receiver operating characteristic curve (ROC), and area under the curve (AUC).

To test the robustness of the classifier on different input utterance durations, we initially performed an independent *t*-test on durations of utterances between the case and control groups. If the difference is significant, we then split the All-utterance into smaller segments (10 s, 20 s, …, 70 s), and test the performance of the i-vector method on shorter utterances.

2.3.4. Pearson correlation

To understand the acoustic explanation of the MFCC i-vector, we calculated the Pearson correlation between MFCC statistics and MFCC i-vector. These statistics include maximum value, minimum value, arithmetic mean, and standard deviation. The MFCC statistics were calculated through the open-source openSMILE-3.0 toolbox (Eyben et al., 2013).

We initially calculated the pairwise correlation of all features to obtain r and p values. Since the MFCC statistics and i-vectors are both high-dimensional, we selected the pairs of features with $p < 0.05$ and $r > 0.3$ and plotted heat maps to represent the correlation (r value) of these selected features.

3. Results

3.1. Participants and utterance

There were 1808 participants (785 depression patients and 1023 healthy people). The case interview on average took 60 min and the control interview takes 25 min. This is because more questions about MDD are asked for case group to obtain cases with recurrent MDD and to ensure high quality data. Utterance durations of different group are shown in Table 1. For Demo-utterance, there is no significant difference in durations between two groups ($t = -0.83$, $p = 0.86$). For All-utterance, durations are significantly different between case group and control group ($t = 145.94$, $p = 0.00$).

Table 1
Duration of utterances.

Utterance	Group	N	Duration(s) Mean	Duration(s) SE
Demo-utterance	case	681	12.40	16.76
	control	873	13.23	125.10
All-utterance	case	784	249.16	333.29
	control	1024	103.22	153.68

3.2. Classification results

Table 2 shows the classification results using different features and utterance. Since there is a significant difference in All-utterance, we report the classification results using duration as input feature. Using Demo-utterance, durations show no better performance than chance. Using All-utterance, the AUC using duration as input feature is 0.64. Fig. 1 shows the ROC curve with or without i-vector framework. The i-vectors improved the AUC by 7% and 14% for MFCC features using Demo-utterance and All-utterance.

Table 3 shows the classification performance on different utterance durations. When the split utterance is from short to long, the sensitivities are always high, while the specificities and AUC gradually become higher. When duration is greater than 40 s, the classification results stabilize.

3.3. Pearson correlation coefficients between MFCC statistics and i-vectors

Almost all the MFCC statistics are significantly correlated with at least one dimension of MFCC i-vectors with $r > 0.3$. Significantly correlated MFCC i-vectors are concentrated from the dimension of 2 to 28, with $r > 0.3$. We coded the i-vectors as ivector001, ivector002, …, ivector200. The correlations of selected feature pairs are shown in Fig. 2–5. Note that the squares are not colored for non-significant correlated feature pairs ($p \geq 0.05$). Fig. 2 shows a mixture of warm and cold colors of the squares in each column. Figs. 3, 4 and 5 show that the colors of the squares in each column are similar.

4. Discussion

We evaluated the effectiveness of MFCC i-vectors in predicting depression in 1808 clinical samples. We tested the method's robustness to varying utterance durations and we explained the meaning of MFCC i-vectors by analyzing their correlation with MFCCs. The ROC (Fig. 1) shows that the i-vector method has better performance than the MFCC method both for Demo-utterance and All-utterance. The i-vectors improved the AUC by 7% and 14% for MFCC features using Demo-utterance and All-utterance, which is consistent with previous results (Afshan et al., 2018; Lopez-Otero et al., 2015, 2014; Nasir et al., 2016). The i-vector framework shows state-of-art performance in predicting depression in a large clinical sample.

The extraction of i-vectors from MFCCs is not a simple linear approach. Instead, it considers the global variability of MFCC features, including speaker and channel variability. That is, i-vectors are more sensitive to the variance of MFCCs rather than means, which is consistent with the results shown in Figs. 2–5. Fig. 2 shows a mixture of warm and cold colors of the squares in each column, indicating that the positive and negative correlations of MFCC mean with each i-vector are inconsistent. However, as shown in Figs. 3–5, the colors of the squares in each column are close to each other, which reflects the fact that the three MFCC statistics (maximum value, minimum value, and standard deviation) are consistently correlated with i-vectors (all positively or all negatively).

Our findings are consistent with recognized diagnostic vocal features

Table 2
Classification performance measures from five-fold cross validation.

Utterance	Feature	Sensitivity	Specificity	AUC
Demo-utterance	Duration	0.09	0.96	0.53
	MFCC	0.49	0.70	0.59
	MFCC i-vectors	0.66	0.70	0.66
All-utterance	Duration	0.40	0.89	0.64
	MFCC	0.56	0.77	0.66
	MFCC i-vectors	0.76	0.84	0.80

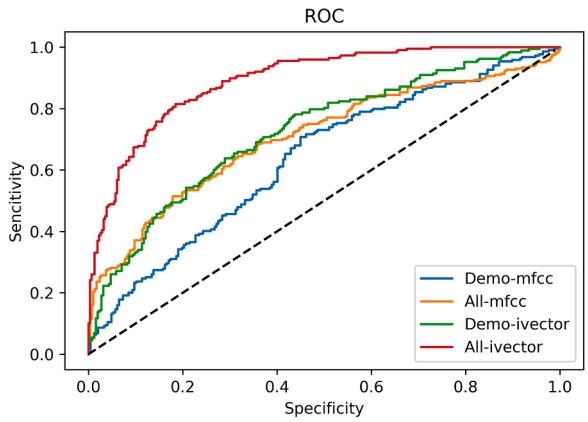

Fig. 1. Receiver operating characteristics (ROC) curve for different machine learning methods. 'Demo-' and 'All-'means using Demo-utterance and All-utterance respectively.

Table 3
Classification performance on different utterance durations.

Utterance duration (s)	Number of utterances Depressed	Healthy	Sensitivity	Specificity	AUC
10	19,133	10,046	0.88	0.52	0.70
20	9379	4760	0.90	0.60	0.74
30	6116	3000	0.91	0.63	0.77
40	4510	2146	0.91	0.64	0.78
50	3528	1599	0.93	0.64	0.78
60	2879	1266	0.93	0.64	0.78
70	2414	1033	0.93	0.64	0.78

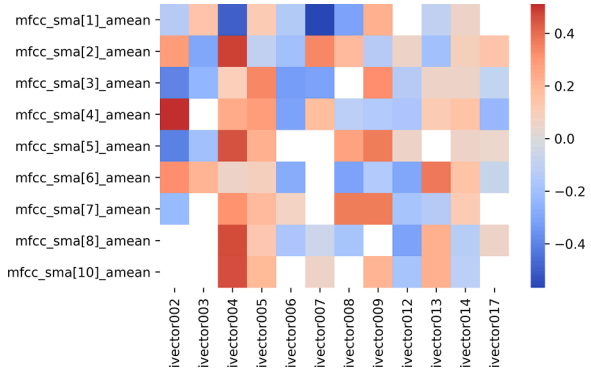

Fig. 2. Pearson Correlation Coefficients between MFCC arithmetic means and i-vectors. The suffix sma indicates that they were smoothed by a moving average filter with window length 3. The suffix de indicates that the current feature is a 1st order delta coefficient of the smoothed MFCC coefficients.

of MDD, which include slow speech, with increased pause time in talkback, reduced volume of speech, and reduction in both number and variation of pitch of speech (American Psychiatric Association, 2013) (p163). Cummins et al. (2013) found that an increase in depression severity was associated with a decrease in MFCC change over time, which is consistent with the description of depressed patients with hollow monotonous speech.

The durations of all-utterance are significantly different between case and control groups. Differences in utterance duration could arise from both depression and experimental design. Alpert et al. (2001) found the utterance duration of the depressed group in the free speech task is significantly shorter than that of the control group. Our clinical design was a semi-structured interview with a high degree of freedom in the interview process and a variable number of questions. Thus, we

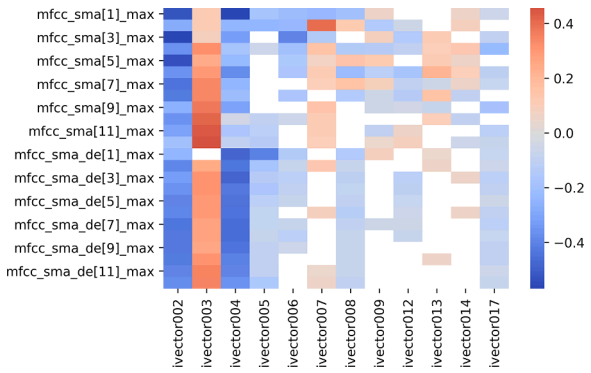

Fig. 3. Pearson Correlation Coefficients between MFCC maximum values and i-vectors. The suffix sma indicates that they were smoothed by a moving average filter with window length 3. The suffix de indicates that the current feature is a 1st order delta coefficient of the smoothed MFCC coefficients.

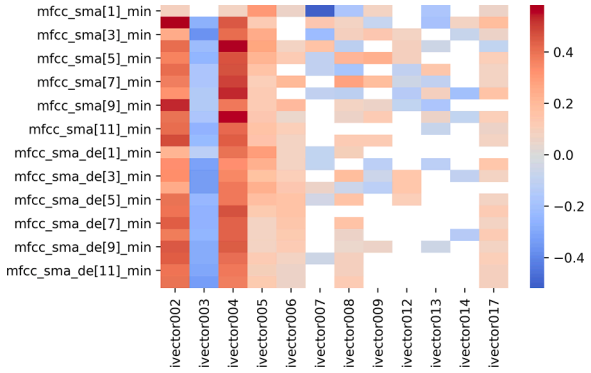

Fig. 4. Pearson Correlation Coefficients between MFCC minimum values and i-vectors. The suffix sma indicates that they were smoothed by a moving average filter with window length 3. The suffix de indicates that the current feature is a 1st order delta coefficient of the smoothed MFCC coefficients.

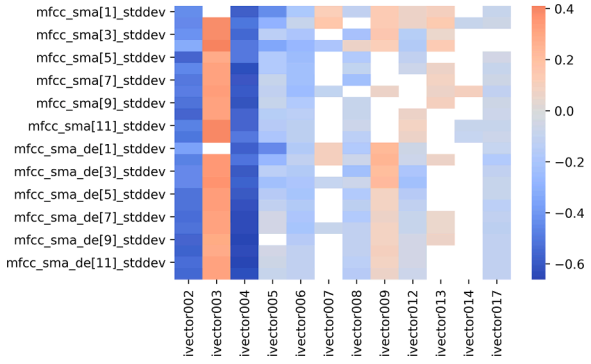

Fig. 5. Pearson Correlation Coefficients between MFCC standard deviations and i-vectors. The suffix sma indicates that they were smoothed by a moving average filter with window length 3. The suffix de indicates that the current feature is a 1st order delta coefficient of the smoothed MFCC coefficients.

reported the effect of utterance duration on classification. Using Demo-utterance, durations show no better performance than chance. Using All-utterance, the AUC using duration as input feature is 0.64. Table 3 shows the classification performance on different utterance durations. The sensitivities are always at a high level when the split utterance increases from 10 s to 70 s. The specificities and AUC gradually increases when utterance grows from 10 s to 40 s and are stabilized when duration is $>$ 40 s. To summarize, our classification system is robust to utterance duration $>$ 40 s.

There are also some limitations in this study. First, our study included only women, indicating that we should be more cautious in generalizing our findings to the male population. Voice features differ between men and women. Second, the i-vector system is based on utterance level. Further research should investigate how to combine more easily accessible information and fully use multiple utterances of the same person to achieve high performance person-level depression detection. There is more potentially useful information, such as demographic information and contextual information of a given utterance (i.e., what question does this utterance answer to or what emotion is activated in this utterance).

5. Conclusion

This study uses an i-vector framework for depression detection in a large clinical sample. Utterances are edited from records of clinical interview speech. The i-vectors improved 14% of the AUC for MFCC. This classification system is robust to utterance duration $>$ 40 s. This study examined the effectiveness and robustness of the i-vector framework in detecting MDD with the exclusion of common bias in previous studies. It provides a foundation for exploring the clinical application of voice features in diagnosing MDD.

Role of the funding source

Not applicable.

Author contributions

All authors contributed to conceptualization of the study. Y.D. conducted the formal analysis and wrote the original draft. J.W. helped with methodology and voice data curation. W.L. contributed to project administration and data curation. T.Z. supervised the statistical analysis and provided resources. All co-authors provided critical review and editing. All authors read and approved the final manuscript.

Data availability statement

Data used in this study is available by contacting the corresponding author to make arrangements.

Ethics statement

Ethical oversight of the study was provided by the Ethics Committee of Bio-x Center, Shanghai Jiao Tong University (M16033). All participants provided their written informed consent.

Declaration of Competing Interest

The authors have declared that no conflict interests exist.

Acknowledgments

The authors would like to thank the patients and staff at clinic where this study was conducted.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jad.2021.04.004.

References

Afshan, A., Guo, J., Park, S.J., Ravi, V., Flint, J., Alwan, A., 2018. Effectiveness of Voice Quality Features in Detecting Depression. Interspeech 2018, 1676–1680. https://doi.org/10.21437/Interspeech.2018-1399.

Alexopoulos, G.S., Meyers, B.S., Young, R.C., Campbell, S., Silbersweig, D., Charlson, M., 1997. Vascular Depression" Hypothesis. Arch. Gen. Psychiatry 54 (10), 915–922. https://doi.org/10.1001/archpsyc.1997.01830220033006.

Alpert, M., Pouget, E.R., Silva, R.R., 2001. Reflections of depression in acoustic measures of the patient's speech. J. Affect. Disord. 66 (1), 59–69. https://doi.org/10.1016/S0165-0327(00)00335-9.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Association, A.P., 1994. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association.

Cox, J.L., Holden, J.M., Sagovsky, R., 1987. Detection of postnatal depression: development of the 10-item Edinburgh Postnatal Depression Scale. Br. J. Psychiatry 150 (6), 782–786.

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: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 970–974. https://doi.org/10.1109/ICASSP.2014.6853741.

Cummins, Nicholas, Epps, J., Breakspear, M., & Goecke, R. (2011). *An Investigation of Depressed Speech Detection: features and Normalization*. 4.

Cummins, Nicholas, Joshi, J., Dhall, A., Sethu, V., Goecke, R., & Epps, J. (2013). *Diagnosis of depression by behavioural signals: a multimodal approach*. 11–20.

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-End Factor Analysis for Speaker Verification. IEEE Trans. Audio Speech Lang. Process. 19 (4), 788–798. https://doi.org/10.1109/tasl.2010.2064307.

Demyttenaere, K., Bruffaerts, R., Posada-Villa, J., Gasquet, I., Kovess, V., Lepine, J.P., Angermeyer, M.C., Bernert, S., De Girolamo, G., Morosini, P., 2004. Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. JAMA 291 (21), 2581–2590.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). *Recent developments in opensmile, the munich open-source multimedia feature extractor*. 835–838.

Goldberg, D., 1995. Epidemiology of mental disorders in primary care settings. Epidemiol. Rev. 17 (1), 182–190.

Gustafsson, H., Nordström, A., Nordström, P., 2015. Depression and subsequent risk of Parkinson disease: a nationwide cohort study. Neurology 84 (24), 2422–2429.

Heatherton, T.F., Kozlowski, L.T., Frecker, R.C., FAGERSTROM, K., 1991. The Fagerström test for nicotine dependence: a revision of the Fagerstrom Tolerance Questionnaire. Br. J. Addict. 86 (9), 1119–1127.

Kendler, K.S., Aggen, S.H., Neale, M.C., 2013. Evidence for multiple genetic factors underlying DSM-IV criteria for major depression. JAMA Psychiatry 70 (6), 599–607.

Kendler, K.S., Gardner, C., Neale, M., Prescott, C., 2001. Genetic risk factors for major depression in men and women: similar or different heritabilities and same or partly distinct genes? Psychol. Med. 31 (4), 605.

Kendler, K.S., Gatz, M., Gardner, C.O., Pedersen, N.L., 2006. A Swedish national twin study of lifetime major depression. Am. J. Psychiatry 163 (1), 109–114.

Kendler, K.S., Prescott, C.A., 2007. Genes, environment, and psychopathology: Understanding the Causes of Psychiatric and Substance Use Disorders. Guilford Press.

Kendler, K., Silberg, J., Neale, M., Kessler, R., Heath, A., Eaves, L., 1992. Genetic and environmental factors in the aetiology of menstrual, premenstrual and neurotic symptoms: a population-based twin study. Psychol. Med. 22 (1), 85–100.

Kenny, P., Boulianne, G., Dumouchel, P., 2005. Eigenvoice modeling with sparse training data. IEEE Trans. Speech Audio Process. 13 (3), 345–354. https://doi.org/10.1109/TSA.2004.840940.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K.R., Rush, A.J., Walters, E.E., Wang, P.S., 2003. The epidemiology of major depressive disorderresults from the National Comorbidity Survey Replication (NCS-R). JAMA 289 (23), 3095–3105. https://doi.org/10.1001/jama.289.23.3095.

Lopez-Otero, P., Dacia-Fernandez, L., Garcia-Mateo, C., 2014. A study of acoustic features for depression detection. In: 2nd International Workshop on Biometrics and Forensics, pp. 1–6. https://doi.org/10.1109/IWBF.2014.6914245.

Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., 2015. Assessing speaker independence on a speech-based depression level estimation system. Pattern Recognit. Lett. 68, 343–350. https://doi.org/10.1016/j.patrec.2015.05.017.

Low, D.M., Bentley, K.H., Ghosh, S.S., 2020. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig. Otolaryngol. 5 (1), 96–116. https://doi.org/10.1002/lio2.354.

Low, L.A., Maddage, N.C., Lech, M., Allen, N., 2009. Mel frequency cepstral feature and Gaussian Mixtures for modeling clinical depression in adolescents. In: 2009 8th IEEE International Conference on Cognitive Informatics, pp. 346–350. https://doi.org/10.1109/COGINF.2009.5250714.

Masters, M.C., Morris, J.C., Roe, C.M., 2015. Noncognitive" symptoms of early Alzheimer disease: a longitudinal analysis. Neurology 84 (6), 617–622.

Nasir, M., Jati, A., Shivakumar, P.G., Nallan Chakravarthula, S., & Georgiou, P. (2016). *Multimodal and multiresolution depression detection from speech and facial landmark features*. 43–50.

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.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., Thomas, R., 2008. Understanding and using sensitivity, specificity and predictive values. Indian J. Ophthalmol. 56 (1), 45–50. https://doi.org/10.4103/0301-4738.37595. PubMed.

Peterson, R.E., Cai, N., Dahl, A.W., Bigdeli, T.B., Edwards, A.C., Webb, B.T., Bacanu, S.-A., Zaitlen, N., Flint, J., Kendler, K.S., 2018. Molecular genetic analysis subdivided by adversity exposure suggests etiologic heterogeneity in major depression. Am. J. Psychiatry 175 (6), 545–554.

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. Am. J. Psychiatry 170 (1), 59–70.

Ringeval, F., Schuller, B., Valstar, M., Cummins, Ni., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-.M., Song, S., Liu, S., Zhao, Z., Mallol-Ragolta, A., Ren, Z., Soleymani, M., & Pantic, M. (2019). AVEC 2019 Workshop and Challenge: state-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. *ArXiv:1907.11510 [Cs, Stat]*. http://arxiv.org/abs/1907.11510.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., & Pantic, M. (2016). *Avec 2016: depression, mood, and emotion recognition workshop and challenge*. 3–10.

Wang, J., Zhang, L., Liu, T., Pan, W., Hu, B., Zhu, T., 2019. Acoustic differences between healthy and depressed people: a cross-situation study. BMC Psychiatry 19 (1), 300. https://doi.org/10.1186/s12888-019-2300-7.

Wells, K.B., Hays, R.D., Burnam, M.A., Rogers, W., Greenfield, S., Ware, J.E., 1989. Detection of depressive disorder for patients receiving prepaid or fee-for-service care: results from the Medical Outcomes Study. JAMA 262 (23), 3298–3302.

World Health Organization, 2017. Depression and Other Common Mental Disorders: Global Health Estimates. World Health Organization.