Vered Aharonson*, Verushen Coopoo, Craig S. Carlson, and Michiel Postema

Speech biomarkers for automated depression level detection

https://doi.org/10.1515/cdbme-2025-0172

Abstract: This study investigates the contribution of speech audio and speech verbal content in the automated detection of depression levels. Recordings from the Distress Analysis Interview Corpus Wizard-of-Oz dataset and the depression severity labels of the recordings were used to extract acoustic features. A transcription of the recordings was used to extract textual features. The acoustic set included prosodic, cepstral, and glottal feature categories. The textual features consisted of semantic and syntactic categories. Mutual information feature selection, followed by a random forest classifier identified the set of features which optimised the depression level classification. The optimised binary classification of depression from non-depressed yielded an accuracy of 0.89 and an F1 score of 0.87. A classification of the five depression levels yielded an accuracy of 0.79 and an F1 score of 0.72. The ratio of importance scores of acoustic to textual of the speech acoustic features was greater than 3:1. Our method thus provided acoustic and textual indicators in depressed speech. These might increase the acceptability of automated depression detection by healthcare professionals. Our initial findings indicate a select set of features that can improve the effectiveness of automated depression detection and monitoring tools.

Keywords: speech-based screening, glottal features, semantic and syntactic textual features, prosody.

***Corresponding author: Vered Aharonson,** School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, 1 Jan Smutslaan, Braamfontein 2050, ZA and Medical School, University of Nicosia, 93 Agiou Nikolaou Street, Engomi 2408, CY. email: vered.aharonson@wits.ac.za
Verushen Coopoo, School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, Braamfontein, ZA
Craig S. Carlson, Department of Electrical Engineering and Automation, Aalto University, Espoo, FI; Department of Biomedical Technology, Faculty of Medicine and Health Technology, Tampere University, Tampere, FI and School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, Braamfontein, ZA.
Michiel Postema, Department of Biomedical Technology, Faculty of Medicine and Health Technology, Tampere University, Tampere, FI and School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, Braamfontein, ZA.

1 Introduction

An estimated 5% of the world population is suffering from depression, a large portion of those undiagnosed [1]. Speech is a central tool in depression diagnosis and therapy and might be employed in automated screening and monitoring technologies in everyday digital devices such as mobile phone and computers [2]. To be both effective and acceptable by healthcare professionals, the tools should convey which speech biomarkers of depression, in its different severity levels, are assessed by these tools [3].

A vast number of studies examined automated speech-based detection of depression and a classification of depression severity. A dataset commonly shared by these investigations is the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WoZ). This dataset contains speech recordings of 189 interviews conducted by a virtual psychologist controlled by a human, and their text transcriptions [4]. Each speaker has a verified depression severity as defined by the validated patient health questionnaire (PHQ-8) rating scale [5]. Previous DAIC-WoZ studies employed speech processing to examine acoustic features in speech of depressed individuals or natural language processing to examine verbal features. Recent studies included both types of features for examination [1, 6–9]. The performance in depression classification ranged between 75 % to 88 %. Discrimination between depressed and non-depressed was detected with 94 % accuracy [10–13]. Most of these studies, however, employed deep machine learning that lacked an explanation which features and types are the best biomarkers of depression [2, 7, 14]. Moreover, to enable deep learning on the relatively small and unbalanced cohort of DAIC-WoZ, the recorded speech was segmented into subsegments of less than a second prior to augmentation. These methods, although considered necessary and commonly used in machine learning, might further obscure longer temporal patterns of natural speech, on a word or sentence level. A processing of natural time units in straightforward machine learning tools may provide both insights on the manifestation of physiological, emotional and cognitive processes in depressed speech. These insights then might be used to better tune and enhance digital tools for depression diagnosis and monitoring.

In this study we examined the DAIC-WoZ recordings on a sentence-level, extracted acoustic and textual features based

on literature observational psychiatric characteristics of depressed voice, and employed Mutual Information and Random Forest importance scores to highlight which subsets of features most contribute to depression-severity classification.

2 Materials and methods

For this study, dedicated software was written in Python (version 3.11) [15], with a MATLAB® (The MathWorks, Inc., Natick, MA, USA) plugin for Collaborative Voice Analysis Repository (COVAREP) features extraction.

Voice data were downloaded from DAIC-WoZ [4]. The distribution of depression severity in the 189-participants cohort is portrayed in Table 1.

The interviewer questions and comments were removed from the audio and textual data. Each response of the participants was saved in an audio `wav` file and a text file. Single-word sentences such as "yes" and "hello" were discarded. This step yielded a set of 60 sentences on average per speaker of mean duration of 22 seconds. The feature set included five feature categories, summarised in Table 2, which includes examples for each category, literature-based indication of their perceptual observations in depression, and the number of features taken from each category. The acoustic categories include prosodic, cepstral and glottal, while textual categories include syntactic and semantic. Acoustic features were extracted from the recordings using the COVAREP plugin. Textual features were extracted from the stored text files using the Python libraries `nltk`, `string` and `scacy`.

Feature selection employed the Mutual Information. This straightforward method captured non-linear relationships between features and was not limited to discrete features [16]. The output number of features was determined as the top fifty percentile of Mutual Information scores. Two Random Forest classifiers were trained on the selected feature set. These classifiers incorporated a built-in feature permutation importance metric and thus constituted a second selection specifically tuned to the classification task [17]. A class weighting was applied in the classifier implementation to mitigate the effect of imbalance between the classes. The first classifier was binary and discriminated the No-depression class from the Depression classes. The second classifier performed a 5-class classification for the five severity levels.

The classifiers provided importance scores for each feature which were aggregated per feature category. The first aggregation grouped the two main categories of features: acoustic and textual. The second classifier grouped the five feature types. The classifiers performance was evaluated using F1 and accuracy metrics that allowed for comparison with previous results.

Tab. 1: Depression severity distribution of participants in DAIC-WoZ corpus.

Depression severity rating	# participants
No depression	86
Mild	46
Moderate	30
Moderate severe	20
Severe	7

3 Results

The number of features in each of the categories in the original set and following the Mutual Information selection is presented in Table 3. The numbers are complemented by percentages in parenthesis. The final column presents the selection percentage out of the original set.

Only 44 % and 42 % from the cepstral and semantic feature subsets, the largest feature categories in the original set, were selected in the first step. The next two, in terms of original categories size, were the prosodic and glottal. Both increased their proportion in the selected set. The smallest subset, syntactic features, was least favoured in the Mutual Information selection, and only 37 % of these features were selected.

The accuracy and F1-score for the 5-class classification were 0.79 and 0.72, respectively. The binary classifier yielded an accuracy and F1-score of 0.89 and 0.87, respectively. The feature importance scores aggregation across the five categories and across the two main categories, acoustic and textual, are shown in Table 4.

The importance scores for the binary depression detection show a higher aggregated importance for the acoustic features compared to the textual ones, of 81 % to 19 %. The classification of the five depression-severity classes reveals a smaller ratio of acoustic to textual importance ratio of 75 % to 25 %.

4 Discussion and conclusions

Our methodology employed machine learning in a transparent way and helped understand which categories of features are most influential for the discrimination of depression severity levels. The application of Mutual Information selection prior to the classification enabled a tracking of the feature selection process and improved transparency, as did The random forest classifier by providing feature importance ranking.

Tab. 2: Acoustic and textual feature categories related to depression.

Feature category	Examples	Indication in depression	# features
Prosodic	Intonation, speech rate, pause duration, loudness, jitter, shimmer.	Reduced speech energy, pitch, and speech rate – perceived as monotonous, with lower intensity, slower speed, and reduced pitch range.	40
Cepstral	Mel-frequency cepstral coefficients, cepstral peak prominence, spectral tilt.	Muscle tension, breath control, and vocal stability changes, indicated in weaker cepstral coefficients and spectral flattening.	54
Glottal	Amplitude quotients, harmonics ratio, harmonic richness factor, maxima dispersion quotient, and R glottal shape parameter.	Changes in vocal fold behaviour due to emotional and psychological states, perceived as increased breathiness, weakness and inconsistent glottal cycles.	36
Syntactic	Number and length of tokens, characters and syllables, parts of speech, disfluency rate.	Sentence structure, grammatical complexity, and fluency change indicated by shorter sentences, simpler grammar, less conjunctions, more pauses, false starts, and repetition.	19
Semantic	Words to vectors embeddings, words sentiment and emotion scores, cognitive processing words, pronoun usage, lexical richness indices.	Increased self-focus and negative emotion, indicated by more “I, me” pronouns, fewer diverse words, and more negative words.	5

Tab. 3: Number of features from each acoustic and textual feature category in the initial set and the Mutual Information selected set.

Feature category	Initial number	Mutual Information selected number	Percent selected
Prosodic	40 (0.20 %)	29 (0.28 %)	0.70
Cepstral	54 (0.26 %)	19 (0.23 %)	0.44
Glottal	36 (0.18 %)	25 (0.20 %)	0.56
Syntactic	19 (0.09 %)	6 (0.07 %)	0.37
Semantic	55 (0.27 %)	23 (0.22 %)	0.42

Tab. 4: Importance of the acoustic and textual feature categories in the five-class and binary random forest classifications, as well as their aggregated totals.

Feature category	5-class importance		Binary importance	
Prosodic	0.29		0.34	
Cepstral	0.23	0.75	0.22	0.81
Glottal	0.23		0.25	
Syntactic	0.09	0.25	0.05	0.19
Semantic	0.16		0.14	

The proportion of the acoustic feature in the initial feature set was 64 %. The combined two-tier feature selection of mutual information and feature importance in optimising the random forest classifier increased its proportion to 75–25 for five-class and 81–19 in the binary classification. Within the acoustic categories, the prosodic features increased their proportion in the Mutual Information selection and hold a larger importance for both classifiers optimisation. This finding corroborate the importance of prosody, which was reported in many emotion recognition and mental state voice-based recognition studies. This increased importance, however is higher for the binary classification compared to the five-severity classes. This may imply that in discriminating more severe depression levels, other features become more important. Notably, in the five-classes case, both textual feature categories, semantic and syntactic have higher prominence compared to the binary classification. The glottal features were indicated as important in both feature selection tiers. These were least used in former studies although their existence in depressed speech was perceived by clinicians (cf. Table 2). These features, however, have distinct overlap with prosodic features that were frequently shown as important in emotional speech recognition in general and depressed speech in particular [6, 7, 9–12, 14]. The accuracy of the binary classifier is comparable to or higher than previous studies that used the DAIC dataset for depression detection [6–9, 12, 14]. These studies, however, used deep networks and could not provide feature importance insights.

Importantly, our analysis refrained from using preprocessing techniques that could alter subtle properties in the raw data. Previous studies employed audio filtering, divided the recording to small segments to enlarge the number samples for deep learning, and used augmentation techniques either to enlarge the samples or to balance classes imbalance in data [6, 7, 9]. Our methods did not apply noise-reduction filters, did not segment the participants response, which then remained as whole sentence time unit, and did not apply augmentation. Naturally this prohibited the use of deep learning, which needs a much larger data size. Even for shallow learning, as our ran-

dom forest classification, a feature space of 204 hand-crafted features was too large and hence a feature selection was applied. The effects of not using filters or augmentation need to be further examined. Listening to the recordings and examining the DAIC-WoZ recording protocol [4] yielded that recordings were relatively clean of noise. To quantitatively measure and validate noise reduction filters, a comparison of filter usage effects on classification performance is needed. Similarly, the effect of augmentation and up/down sampling techniques need to be assessed. Importantly, these need to be assessed for real-world implementation of speech-based depression recognition, where clean recording environments and balanced classes of depression severity could not be assumed [2].

This study is preliminary, but reinforces the importance of diverse acoustical and textual feature analysis in depression. It highlights the potential of speech-based depression analysis in accessible and affordable automated tools.

Author Statement

This work has been supported by the Erasmus+ Mobility Programme for Staff Mobility For Teaching and by the Academy of Finland, Grant Number 340026. Authors state no conflict of interest. The research related to the use of human dataset complied with all the relevant national regulations and institutional policies and has been approved by the IRB committee at the University of the Witwatersrand, Johannesburg with clearance number M220330.

References

[1] Gianfredi V, Blandi L, Cacitti S, Minelli M, Signorelli C, Amerio A, Odone A. Depression and objectively measured physical activity: a systematic review and meta-analysis Int J Environ Res Public Health 2020;17:3738.

[2] Shin J, Bae SM. Use of voice features from smartphones for monitoring depressive disorders: scoping review. Digit Health 2024;10:20552076241261920.

[3] Malik S, Singh R, Arora G, Dangol A, Goyal S. Biomarkers of major depressive disorder: knowing is half the battle. Clin Psychopharmacol Neurosci 2021;19,12–25.

[4] Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) 2014:3123–8.

[5] Kroenke K, Spitzer RL, Williams JBW, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010:32:345–59.

[6] Aharonson V, de Nooy A, Bulkin S, Sessel G. Automated classification of depression severity using speech: a comparison of two machine learning architectures. Proceedings of the 2020 IEEE International Conference on Healthcare Informatics (ICHI) 2020:1–4.

[7] Muzammel M, Salam H, Othmani A. End-to-end multimodal clinical depression recognition using deep neural networks: a comparative analysis. Comput Methods Programs Biomed Update 2021;211:106433.

[8] Lam G, Huang D, Lin W. Context-aware deep learning for multi-modal depression detection. Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019:3946–50.

[9] Lau C, Chan WY, Zhu XD. Improving depression assessment with multi-task learning from speech and text information. Proceedings of the 55th Asilomar Conference on Signals, Systems, and Computers 2021:449–53.

[10] Scibelli F, Roffo G, Tayarani M, Bartoli L, De Mattia G, Esposito A, Vinciarelli A. Depression speaks: automatic discrimination between depressed and non-depressed speakers based on nonverbal speech features. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018:6842–6.

[11] Aloshban N, Esposito A, Vinciarelli A. What you say or how you say it? Depression detection through joint modeling of linguistic and acoustic aspects of speech. Cogn Comput 2021;14:1585–98.

[12] Tao F, Ge X, Ma W, Esposito A, Vinciarelli A. Multi-local attention for speech-based depression detection. Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023:1–5.

[13] Jiang H, Hu B, Liu Z, Wang G, Zhang L, Li X, Kang H. Detecting depression using an ensemble logistic regression model based on multiple speech features. Comput Math Methods Med 2018;2018:6508319.

[14] Liu L, Liu L, Wafa HA, Tydeman F, Xie W, Yanzhong WY. Diagnostic accuracy of deep learning using speech samples in depression: a systematic review and meta-analysis. J Am Med Inform Assoc 2024;31:2394–404.

[15] Srinath KR. Python: the fastest growing programming language. Int J Res Eng Technol 2017;4:354–7.

[16] Tourassi GD, Frederick ED, Markey MK, Floyd CE. Application of the mutual information criterion for feature selection in computer-aided diagnosis. Med Phys 2001;28:2394–402.

[17] Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics 2008;9:307.