2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Depression Severity Detection Using Read Speech with a Divide-and-Conquer Approach

Namhee Kwon and Samuel Kim

Canary Speech, LLC., Provo, Utah, U.S.A.
{namhee, sam}@canaryspeech.com

Abstract— **We propose a divide-and-conquer approach to detect depression severity using speech. We divide speech features based on their attributes, i.e., acoustic, prosodic, and language features, then fuse them in a modeling stage with fully connected deep neural networks. Experiments with 76 clinically depressed patients (38 severe and 38 moderate in terms of Montgomery-Asberg Depression Rating Scale (MADRS)), we obtain 78% accuracy while patients' self-reporting scores can classify their own status with 79% accuracy.**

I. Introduction

Depression is a mood disorder, relatively common yet seriously affecting a person's life. Traditionally depression disorder diagnosis depends on an in-clinic interview with a patient using the instruments such as Montgomery-Asberg Depression Rating Scale (MADRS) [1] and Hamilton Depression Rating Scale (HDRS) [2]. As an alternative, self-reported measures such as the Patient Health Questionnaire (PHQ) [3] and MADRS-IVRS [4] have been developed to make the process easier; the subjects are requested to answer the self-assessment scale for the questionnaires instead of being interviewed by investigators.

We aim at a non-invasive tool to automatically measure depression severity so that a person can use it outside the clinic without worrying about the sensitive questions potentially affecting the patient's mood. In this regard, we use voice as researchers found that there are language patterns in the depression patients' word usage and speech changes in their pitch, tone, pauses, etc. [5], [6], [7].

Along with advances in speech analysis techniques and machine learning algorithms, the interest in the automatic detection of depression using voice has increased. Dresvyanskiy *et al.* used automatic speech recognition results in predicting PHQ scores and post-traumatic stress disorder (PTSD) [8], and Huang *et al.* have recently proposed a domain adaptation algorithm using a convolutional neural network (CNN) for binary classification based on PHQ scores [9]. Afshan *et al.* [10] successfully applied the i-vector of acoustic features to detect depression from the interview corpus of Chinese female patients. Alghowinem *et al.* [11] showed that spontaneous speech was a better tool in recognizing depression, yet acoustic features such as jitter, shimmer, energy were robust in both read speech and spontaneous speech. Diverse studies were done in speech features [12], [13], [14], [15], [16], feature selection [17], and model architecture design [18] regarding depression.

Other related mental health conditions such as bipolar disorder [19], [20], schizophrenia [21], and anxiety [22] also showed promising results.

In this work, we propose a divide-and-conquer approach to detect depression severity. Although we can extract various features from speech signals, they may represent some attributes of different layers in speech production procedures. Speech production is inherently multifaceted, and how it is modulated by the speaker's health and emotional status is not completely discovered yet. Therefore, we categorize speech features into groups according to their attributes (divide) and build models based on groups by applying different fusion methods (conquer).

The details on data are in Section II, features are described in Section III, and our suggested models are explained in Section IV with the experimental setup and result in Section V, and finally followed by the conclusion in Section VI.

II. Data

A. Data Collection

We have recruited 76 patients diagnosed with depression (58 female and 18 male) and asked voice responses following the instructions shown in Table I. Each patient repeated the same assessment session composed of 7 questions twice a week. The assessment was performed through Canary's mobile application on each patient's mobile phone, where a voice response was collected in a 16-bit wav file with a 16kHz sampling rate. The number of repeated sessions per patient varies from 1 to 14 sessions.

Since our research goal is to build a mobile application without human administration, there is a legitimate concern around the risk of provoking sad or depressed emotions in severely depressed patients while asking their feelings and thoughts. Therefore, our study focuses on detecting the level of depression from speech without asking any emotional or personal questions. We only use read speech and cognitive responses and study if we can detect depression disorder from how they speak rather than what they speak.

B. Labels

Each participating patient was scored using three different instruments: MADRS, MADRS-IVRS, and Snaith-Hamilton Pleasure Scale (SHAPS) [23]. The MADRS and

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

MADRS-IVRS are one of the standard instruments that measure depression levels, as described earlier; MADRS is an investigator-administered score based on the conversation in a clinic, while MADRS-IVRS is a self-reported score over the phone using IVR [24]. They are rated from 0 to 60 where normal is 0 to 6, mild is 7 to 19, moderate is 20 to 34, and severe is 35 to 60. The scale is composed of apparent sadness, reported sadness, inner tension, reduced sleep, concentration difficulties, lassitude, inability to feel, pessimistic thoughts, and suicidal thoughts.

The Snaith-Hamilton Pleasure Scale (SHAPS) [23] is a self-reported 14-item scale that measures anhedonia, i.e. the inability to experience a pleasure. The items cover social interaction, food and drink, sensory experience, and interest/pastimes. The score ranges from 0 to 14; a higher score indicates greater anhedonia.

Figure 1 shows the distributions of the scores, i.e., MADRS, MADRS-IVRS, and SHAPS. As shown in the figure, MADRS is distributed from 27 to 47, meaning that our data collection includes only a moderate or severe level of depression. The correlation between the MADRS and the MADRS-IVRS is 0.81, which is aligned with the field standard [25]. When we use the binary classes (MADCLS) of moderate and severe instead of a finer-grained MADRS, the agreement between the MADCLS and the MADCLS-IVRS is 0.79.

Q1	Read the 10 short sentences (65 words)
Q2	Read a list of words backwards (45 words)
Q3	Read a list of numbers forward and backward (15 numbers are given)
Q4	Say months forward and backward
Q5	Count from 1 to 20, Say A to Z
Q6	Repeat PA-TA-KA for 5 times
Q7	Read the “Grandfather Passage” (130 words)

TABLE I: Instructions for speech assessment session.

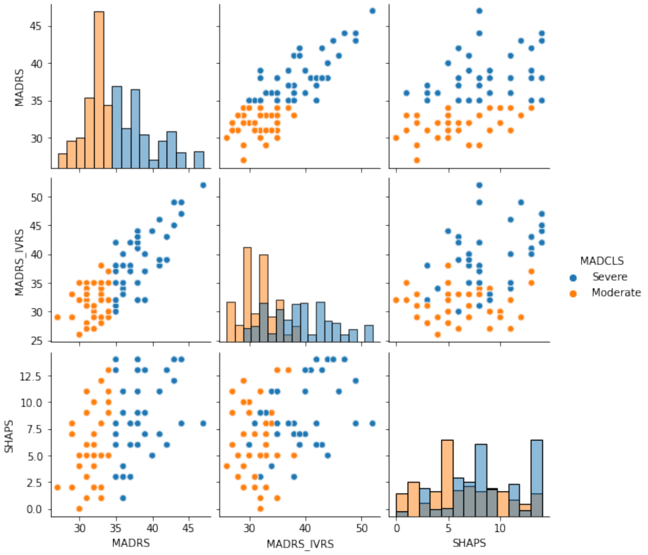


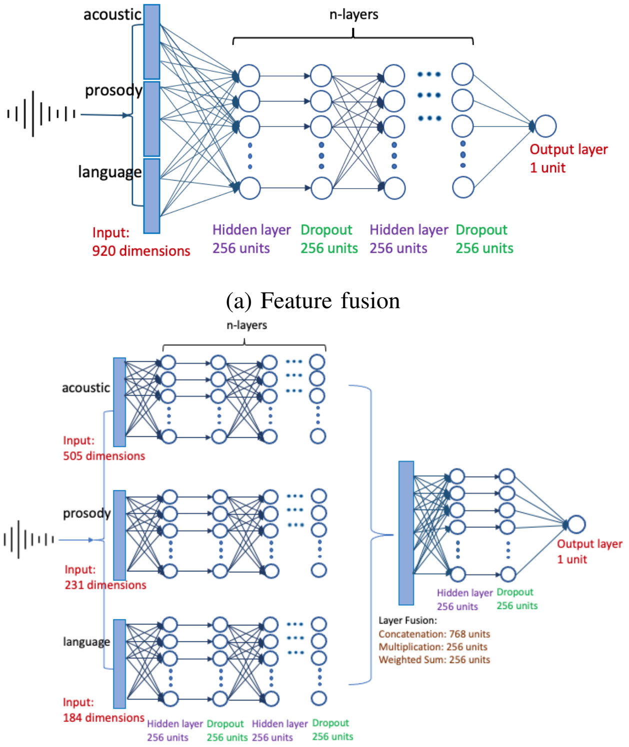
Fig. 1: Label distributions and correlations between labels.

On the other hand, SHAPS shows the full range of available scores from 0 to 14 and the correlation between MADRS and SHAPS is 0.45 because anhedonia is one of the symptoms of depression rather than a direct depression scale.

III. Features

Types of voice features are divided into three categories: acoustic, prosodic, and linguistic features. We consider frame-level signal characteristics as acoustic features, while variations in pitch, loudness, and tempo as prosodic features. The linguistic features capture the language-level patterns which may be influenced by the condition.

Acoustic features include various spectral characteristics and voice quality features. They are extracted from 25 millisecond long frames sliding every 10 milliseconds. Spectral characteristics include spectral flux, spectral centroid, spectral bandwidth, spectral contrast, spectral flatness, spectral roll-off, mel-frequency cepstral coefficients (MFCC), while voice quality features include harmonics-to-noise ratio (HNR), various jitter measures (local jitter, local absolute jitter, relative average perturbation (RAP) jitter, five-point period perturbation quotient (PPQ5) jitter, and the average absolute difference between consecutive differences (DDP) jitter) and various shimmer measures (local shimmer, local shimmer in dB, three-point amplitude perturbation quotient (APQ3) shimmer, five-point amplitude perturbation quotient (APQ5) shimmer, 11-point amplitude perturbation quotient



(a) Feature fusion

(b) Layer fusion

Fig. 2: Diagram of fusion methods.

Features		Session-level	Speaker-level
Chance level		-	0.50
MADCLS-IVRS		-	0.79
Opensmile	eGeMAPS	0.61	0.65
	ComParE	0.55	0.57
	IS09	0.58	0.65
Individual Groups	Acoustic	0.58	0.64
	Prosody	0.57	0.68
	Language	0.58	0.59
Feature Fusion	Acoustic + Prosody	0.65	0.63
	Prosody + Language	0.66	0.61
	Acoustic + Prosody + Language	0.65	0.69
Layer Fusion	Acoustic $\bigcup$ Prosody	0.62	0.61
	Acoustic $\bigotimes$ Prosody	0.66	**0.78**
	Acoustic $\bigoplus$ Prosody	0.62	0.63
	Prosody $\bigcup$ Language	0.63	0.58
	Prosody $\bigotimes$ Language	0.64	0.67
	Prosody $\bigoplus$ Language	**0.69**	0.73
	Acoustic $\bigcup$ Prosody $\bigcup$ Language	0.66	0.60
	Acoustic $\bigotimes$ Prosody $\bigotimes$ Language	**0.69**	0.71
	Acoustic $\bigoplus$ Prosody $\bigoplus$ Language	**0.69**	0.69

TABLE II: MADCLS (Moderate vs. Severe) classification accuracy (unweighted average recall). In feature fusion, $+$ represents concatenation of features. In Layer Fusion, $\bigcup$ represents the concatenation of the layers, $\bigotimes$ represents the multiplication model, and $\bigoplus$ represents the weighted sum model of the hidden layer outputs.

(APQ11) shimmer, and the average absolute difference between consecutive differences (DDP) shimmer). After extracting these frame-level features for a given speech signal, we compute various statistics of individual features to represent the signal. The statistics comprise 19 statistical functions such as mean, median, skewness, kurtosis, quartile, percentile, and slope. The dimension of the acoustic features is 505.

Prosody features include normalized deciles of the fundamental frequency (f0) and energy, and speech rate. The normalized deciles are calculated by normalizing deciles of f0 and energy values from a given speech signal with respect to its first decile to illustrate how they vary.

$$\gamma_i = \log(\frac{\phi_i}{\phi_1}), i \in \{2, 3, 4, \cdots, 9\} \tag{1}$$

where ϕ_i indicates i-th decile. We also compute the same for the maximum and minimum values. For speech rate, we analyze the rhythm of energy pattern to estimate the number of syllables, number of pauses, speech duration, phonation time, speech rate, articulation rate, and average speaking duration (ASD). The dimension of the prosody features is 231.

Since language features are based on the lexical information present in the patients' response, we use an automatic speech recognition (ASR) system. In particular, we use Canary's general English model which is trained on publicly available datasets like Tedlium and Librispeech using the time-delayed neural network (TDNN) architecture in Kaldi [26]. Since each speech signal has a given text for the patient to read, we computed ASR errors such as insertion, deletion, and substitution to evaluate how it is articulated. We also extract average word duration, average vowel duration, filler (ah, hmm, eh, uh, etc.) ratio, and word repetition ratio over the total number of spoken words. For the word order questions from Q2 through Q5, we measure the total correct word order ratio, the longest correct word order ratio, and the unexpected word ratio. The dimension of the language features is 184.

IV. Model

Using the features described in Section III, we build a binary classification model for MADCLS, which is a binary label of MADRS into moderate and severe. We use a fully connected deep neural network (FC-DNN) with an empirically chosen number of hidden layers of 256 neurons. The number of stacked layers for each model is also chosen empirically; 4 layers for acoustic and prosodic features and 1 layer for language features. Each layer is defined with an activation function of ReLU (Rectified Linear Unit) using l2 regularization and 50% dropout to avoid overfitting.

As the feature sets are *divided* in a way that they are grouped by their attributes, we *conquer* by fusing them in various ways. In particular, feature fusion and layer fusion are applied and compared as illustrated in Figure 2. The feature fusion is done by concatenating the feature vectors for a high-dimensional feature vector and then building a fully connected dense model. For the layer fusion, we build a fully connected dense model for each group of features and then apply and compare 3 different fusing methods on the hidden layer output (concatenation, multiplication, or weighted sum) followed by another dense layer. For every model, we add a sigmoid layer as a binary classification output layer.

V. Experiment and Result

We treat each speaker's multiple assessment sessions independently and build a binary classification model on MADCLS at a session-level. For each assessment session composed of the voice responses to 7 questions, we generate the acoustic, prosodic, and language features and feed them

MADCLS	Subjects			Sessions		
	Male	Female	Total	Male	Female	Total
Moderate	8	30	38	71	324	395
Severe	10	28	38	100	252	352
Total	18	58	76	171	576	747

TABLE III: Number of participants per label.

into the different model structures as described in Figure 2. We also predict a speaker-level label by applying a majority voting using each session's predicted value. The number of sessions and speakers per class is shown in Table III.

We perform 6-fold cross-validation; we split the data into 6 folds and iteratively use one fold as a test set and the rest as a training set. Each fold is subject-independent in the sense that different folds do not share data from the same subject.

Table II shows the model performance from the cross-validation experiment on MADCLS. We report the unweighted average recall of the models using different feature groups and fusing methods. The session-level performance shows the unweighted accuracy of the model on 747 sessions, and the speaker-level performance shows the unweighted accuracy from a majority voting for 76 subjects. As the speaker's label is balanced, the by-chance model accuracy is 0.5, and the self-assessed MADCLS-IVRS's accuracy compared to the investigator-administered MADCLS is 0.79.

We compare the model using the openSMILE toolkit [27] with various configurations [28] and the model using our proposed features. openSMILE is an open-source toolkit widely used in emotion recognition from an audio signal. The best-performing session-level accuracy is 0.69 for the fusion model using the acoustic, prosody, and language features, and the final best model at a speaker level is the multiplication layer-fusion model using the acoustic and prosody features, which scored at 0.78.

It shows that our proposed layer-fusion model predicts the depression severity at a higher accuracy than the models using each feature set separately or the feature fusion models using a concatenated high-dimensional vector. Each feature group captures the different aspects of the speech and each model's confusion arises from different reasons, so the benefit of the majority voting for the speaker-level classification differs for each model. Further, the number of sessions per speaker is inconsistent; there are speakers who finished only one session. If we measure the accuracy only for the subjects who finished at least 3 sessions (54 subjects), the accuracy can reach up to 0.83.

VI. Conclusion

We have described our divide-and-conquer approach using acoustic, prosodic, and language features in a fusion model toward depression severity detection. Considering the agreement between the investigator-administered and self-assessed depression severity is 79%, our model using only read speech reaching 78% is very encouraging. There are interesting questions such as which audio responses are more informative and how many sessions are required for a reliable evaluation. We leave these questions for future work as we deal with limited training data and inconsistent user behaviors. We also plan to extend our study to a wider range of subjects to include normal or mild levels of depression severity.

References

[1] Stuart A. Montgomery and Marie Åsberg, "A new depression scale designed to be sensitive to change," *British Journal of Psychiatry*, vol. 134, no. 4, pp. 382–389, 1979.

[2] M. Hamilton, "A rating scale for depression," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 23, pp. 56–62, 1960.

[3] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams, "The phq-9," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[4] James C. Mundt, David J. Katzelnick, Sidney H. Kennedy, Beata S. Eisfeld, Beverley B. Bouffard, and John H. Greist, "Validation of an ivrs version of the madrs," *Journal of Psychiatric Research*, vol. 40, no. 3, pp. 243 – 246, 2006.

[5] Hollien H Darby J, K, "Vocal and speech patterns of depressive patients," *Folia Phoniatr Logop*, vol. 29, pp. 279–291, 1977.

[6] Mohammed Al-Mosaiwi and Tom Johnstone, "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clinical Psychological Science*, vol. 6, no. 4, pp. 529–542, 2018, PMID: 30886766.

[7] Allison M Tackman, David A Sbarra, Angela L Carey, M Brent Donnellan, Andrea B Horn, Nicholas S Holtzman, To'Meisha S Edwards, James W Pennebaker, and Matthias R Mehl, "Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis.," *Journal of personality and social psychology*, vol. 116, no. 5, pp. 817, 2019.

[8] Denis Dresvyanskiy, Danila Mamontov, Maxim Markitantov, and Albert Ali Salah, "Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics," in *AVEC 2019*, 2019.

[9] Zhaocheng Huang, Julien Epps, Dale Joachim, Brian Stasak, James R Williamson, and Thomas F Quatieri, "Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated CNNs," in *Proc. Interspeech 2020*, 2020.

[10] Amber Afshan, Jinxi Guo, Soo Jin Park, Vijay Ravi, Jonathan Flint, and Abeer Alwan, "Effectiveness of voice quality features in detecting depression," *Interspeech 2018.*

[11] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear, "Characterising depressed speech for classification," 01 2013.

[12] Wei Pan, Jonathan Flint, Liat Shenhav, Tianli Liu, Mingming Liu, Bin Hu, and Tingshao Zhu, "Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders," *PLOS ONE*, vol. 14, no. 6, pp. 1–14, 2019.

[13] Nik Wahidah Hashim, Mitch Wilkes, Ronald Salomon, Jared Meggs, and Daniel France, "Evaluation of voice acoustics as predictors of clinical depression scores," *Journal of Voice*, vol. 31, 07 2016.

[14] Zhenyu Liu, Bin Hu, Lihua Yan, Tianyang Wang, Fei Liu, Xiaoyu Li, and Huanyu Kang, "Detection of depression in speech," 09 2015, pp. 743–747.

[15] Mao Yamamoto, Akihiro Takamiya, Kyosuke Sawada, Michitaka Yoshimura, Momoko Kitazawa, Kuo Ching Liang, Takanori Fujita, Masaru Mimura, and Taishiro Kishimoto, "Using speech recognition technology to investigate the association between timing-related speech features and depression severity," *PLoS One*, vol. 15, no. 9 September, Sept. 2020.

[16] Meysam Asgari, Izhak Shafran, and Lisa Sheeber, "Inferring clinical depression from speech and spoken utterances," *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, vol. 2014, 11 2014.

[17] Heysem Kaya, Florian Eyben, Albert Salah, and Björn Schuller, "Cca based feature selection with application to continuous depression recognition from acoustic speech features," 05 2014.

[18] Lang He and Cui Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103–111, 2018.

[19] Katie Matton, Melvin G. McInnis, and Emily Mower Provost, “Into the Wild: Transitioning from Recognizing Mood in Clinical Interactions to Personal Conversations for Individuals with Bipolar Disorder,” in *Proc. Interspeech 2019*, 2019, pp. 1438–1442.

[20] Zakaria Aldeneh, Mimansa Jaiswal, Michael Picheny, Melvin G. McInnis, and Emily Mower Provost, “Identifying Mood Episodes Using Dialogue Features from Clinical Interviews,” in *Proc. Interspeech 2019*, 2019, pp. 1926–1930.

[21] Mary Pietrowicz, Carla Agurto, Raquel Norel, Elif Eyigoz, Guillermo Cecchi, Zarina R. Bilgrami, and Cheryl Corcoran, “A New Approach for Automating Analysis of Responses on Verbal Fluency Tests from Subjects At-Risk for Schizophrenia,” in *Proc. Interspeech 2019*, 2019, pp. 3028–3032.

[22] Samuel Kim, Namhee Kwon, Henry O’Connell, Nathan Fisk, Scott Ferguson, and Mark Bartlett, “How are you? Estimation of anxiety, sleep quality, and mood using computational voice analysis,” in *IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2020, pp. 5369–5373.

[23] R. P. Snaith, M. Hamilton, S. Morley, A. Humayan, D. Hargreaves, and P. Trigwell, “A scale for the assessment of hedonic tone the snaith–hamilton pleasure scale,” *British Journal of Psychiatry*, vol. 167, no. 1, pp. 99–103, 1995.

[24] Ken Kobak, John Greist, James Jefferson, and David Katzelnick, “Computer-administered clinical rating scales a review,” *Psychopharmacology*, vol. 127, pp. 291–301, 10 1996.

[25] Mundt JC, Katzelnick DJ, Kennedy SH, Eisfeld BS, Bouffard BB, and Greist JH., “Validation of an ivrs version of the madrs,” *Journal of Psychiatric Research*, , no. 3, pp. 243–246, 2006.

[26] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, “The kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society.

[27] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” in *Proceedings of the 21st ACM International Conference on Multimedia*, New York, NY, USA, 2013, MM ’13, pp. 835–838, ACM.

[28] Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, pp. 1–1, 01 2015.