CANARY SPEECH

PEER-REVIEWED PUBLICATION

Depression Severity Detection Using Read Speech With A Divide-And-Conquer Approach

DEPRESSION SEVERITY DETECTION USING READ SPEECH WITH A DIVIDE-AND-CONQUER APPROACH

Namhee Kwon[1], Samuel Kim[1], Kristin Predeck[1,2], Jeff Adams[1,3], Henry O'Connell[1]

[1]Canary Speech LLC., UT., U.S.A.
[2]University of California at Davis, CA., U.S.A.
[3]Cobalt Speech LLC., MA., U.S.A.

{namhee, sam, kristin, jeff, henry}@canaryspeech.com

ABSTRACT

We propose a divide-and-conquer approach to detect de pression severity using speech. We divide speech features based on their attributes, i.e., acoustic, prosodic, and lan guage features, then fuse them in a modeling stage with fully connected deep neural networks. Experiments with 76 de pression 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.

Index Terms— Depression detection, speech analysis, mental health

1. INTRODUCTION

Depression is a mood disorder, relatively common yet seri ously affecting a person's life. Although there exist several in-clinic interview-based instruments such as Montgomery Asberg depression rating scale (MADRS) [1] and Hamilton Depression Rating Scale (HDRS) [2], self-reported measures such as 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 inves tigators.

We aim for a computer-to-human interaction tool to au tomatically detect depression and to measure the severity so that a person can use outside the clinic. 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 de tection of depression using voice has increased. Dresvyan skiy *et al.* used automatic speech recognition results in pre dicting PHQ and post-traumatic stress disorder (PTSD) [8], and Huang *et al.* have recently proposed a domain adapta tion algorithm using convolutional neural network (CNN) for binary classification based on PHQ scores [9]. Other related mental health conditions such as bipolar disorder [10, 11], schizophrenia [12], and anxiety [13] also showed promising results.

In this work, we propose a divide-and-conquer approach to detect depression severity. Although we can extract var ious 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 with applying different fusion methods (conquer).

The details on data is in Section 2, features are described in Section 3, and our suggested models are explained in Sec tion 4 with the experimental set up and result in Section 5, and finally followed by the conclusion in Section 6.

2. DATA

2.1. Data Collection

We have recruited 76 patients with depression (58 female and 18 male) and asked to perform voice recording sessions twice a week for a month. In each session, the patients recorded voice responses following 7 instructions shown in Table 1 through the Canary's mobile application.

22 participants stopped after one or two sessions, while the rest repeated for more weeks (3 to 14 sessions). Conse quently, we collected 727 sessions composed of 5001 audio responses.

In data collection procedures using an mobile application without a human administration, there is a legitimate concern around the risk of provoking severely depressed patients to sad or depressed emotions while it asks their feelings and thoughts. Therefore, our study focuses on detecting the level of depression from the 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.

Q1 Read the following passage (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 following passage (130 words)

Table 1: Instructions for speech collection session.

2.2. Labels

Each participating patient has scores from three different in struments: MADRS, MADRS-IVRS, and Snaith-Hamilton Pleasure Scale (SHAPS) [14]. The MADRS and MADRS IVRS are one of standard instruments that measure depres sion level, 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 [15]. 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 sad ness, inner tension, reduced sleep, concentration difficulties, lassitude, inability to feel, pessimistic thoughts, and suicidal thoughts.

The Snaith-Hamilton Pleasure Scale (SHAPS) [14] is a self-reported 14-item scale that measures anhedonia, i.e. the inability to experience the pleasure. The items cover social interaction, food and drink, sensory experience, and inter est/pastimes. The score range is from 0 to 14; score of 2 or less constitutes a normal score, while an abnormal score is defined as 3 or more.

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 stan dard [16]. When we use the binary classes (MADCLS) of moderate and severe instead of a finer-grained MADRS, it is equally balanced and the agreement between the MADCLS and the MADCLS-IVRS is 0.79. 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. It tells us that SHAPS is a quite different measure from MADRS and the anhedonia is one of the aspects of the depression.

3. FEATURES

Types of voice features are in 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 linguis tic features are to capture language-level patterns which may be influenced by the condition.

Acoustic features include various spectral characteris tics and voice quality features. They are extracted from 25 millisecond long frames sliding every 10 milliseconds. Spectral characteristics include spectral flux, spectral cen troid, spectral bandwidth, spectral contrast, spectral flatness, spectral rolloff, 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 pe riod perturbation quotient (PPQ5) jitter, and 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 pertubation quotient (APQ5) shimmer, 11-point amplitude perturbation quotient (APQ11) shimmer, and verage 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 comprises 19 statistical functions such as mean, me dian, skewness, kurtosis, quartile, percentile, and slope. The dimension of the acoustic features is 505.

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

$$y_i = \log\left(\frac{\varphi_i}{\varphi_1}\right),\ i \in \{2, 3, 4, \cdots, 9\} \quad (1)$$

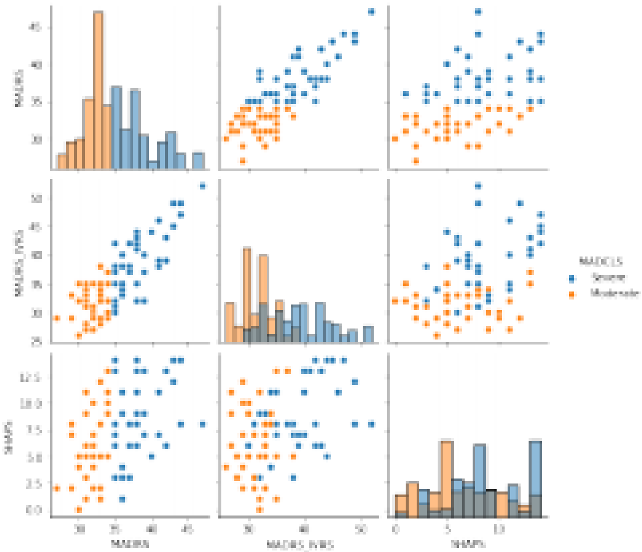


Fig. 1: Label distributions and correlations between labels.

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 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 informa tion of patients' response, we use an automatic speech recog nition (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 neu ral network (TDNN) architecture in Kaldi [17]. Since each speech signal has a given text for the patient to read, we com puted 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 spo ken 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.

4. MODEL

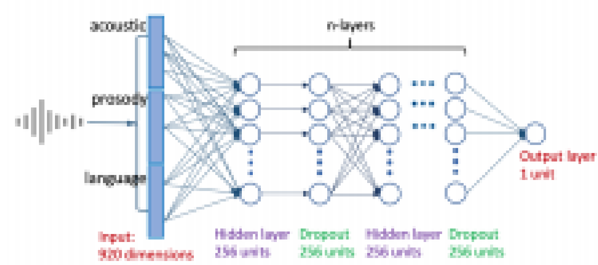
Using the features described in Section 3, we build a binary classification model for MADCLS, which is a binary label of MADRS into moderate and severe. For each feature set, we use a fully connected deep neural network (FC-DNN) with an empirically chosen number of hidden layers of 256 neu rons. Each 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, and 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 fea ture 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 concatenate or multiply the hidden layer output followed by another dense layer. For every model, we add a sigmoid layer as a binary classification output layer.

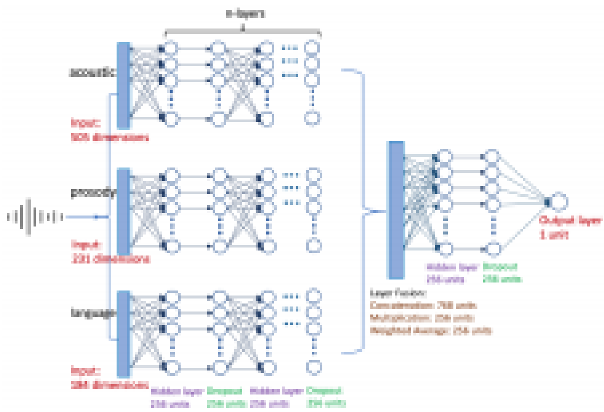
5. EXPERIMENT AND RESULT

We treat each session independently and build a classification model for a session composed of 7 voice recordings. We also predict a final label for each speaker by applying a majority voting using each session's predicted value.

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.



(a) Feature fusion



(b) Layer fusion

Fig. 2: Diagram of fusion methods.

Table 2 shows the accuracy of the model using different feature groups and fusing methods. The session-level accu racy reports the model performance for each assessment ses sion, and the speaker-level accuracy reports the performance from a majority voting for 76 subjects. The by-chance model accuracy is 0.5 and the self-assessed MADCLS-IVRS's accu racy compared to MADCLS is 0.79.

We compare the model using the opensmile toolkit [18] with various configurations [19] and the model using our pro posed features. The best-performing session-level model is the fusion model of all acoustic, prosody, and language fea tures using a weighted sum, whose accuracy is 0.69, but the final best model at a speaker level is the fusion model using a multiplication from acoustic and prosody features, that scored as 0.78. The speaker-level accuracy includes the cases when the speaker has only one session where it does not get any benefit of majority voting. The accuracy can reach up to 0.83 if we measure the accuracy only for the subjects who finished at least 3 sessions (54 subjects).

We also build regression models for MADRS and SHAPS using the same approach of divide-and-conquer. We apply the same group of features and fusion models as the classification model, and the result of the highest correlated models is re ported in Table 3. The correlation between the self-assessed MADRS score (MADRS-IVRS) and the actual MADRS score is 0.81. The correlation between MADRS-IVRS and

Features Speaker-level

	Session-lev el
	-
	-
eGeMAPS ComParE IS09	0.61 0.56 0.58
Acoustic Prosody Language	0.61 0.58 0.67
Acoustic + Prosody Prosody + Language Acoustic + Prosody + Language S	0.65 0.66 0.65

Acoustic Prosody Acoustic N Prosody Acoustic L Prosody Prosody S Language Prosody N Language Prosody L Language Acoustic S Prosody S Language Acoustic N Prosody N Language Acoustic L Prosody L Language	0.63 0.66 0.61 0.63 0.64 0.69 0.66 0.69 0.69

Chance level 0.50 MADCLS-IVRS 0.79 0.65

Opensmile

0.65 0.65 0.64 0.69 0.71 0.63 0.61
0.69 0.61 0.78 0.62 0.59 0.67 0.73
0.59 0.70 0.69

Individual Groups Feature Fusion

Layer Fusion

Table 2: MADCLS (Moderate vs. Severe) classification accuracy. In feature fusion, + represents concatenation of features. In Layer Fusion, S represents the concatenation of the layers, N represents the multiplication model, and L represents the weighted average model of the hidden layer outputs.

Regression Score Correlation

Features
MADRS-IVRS SS
Acoustic Prosody Language
MADRS-IVRS
Prosody only

MADRS	0.81	0.35 (p < 0.005)
SHAPS	0.38	0.47 (p < 0.001)

Table 3: Regression performance for MADRS and SHAPS.

SHAPS is 0.38, and this is reasonable because SHAPS covers only one aspect of depression, i.e. anhedonia, as we discussed earlier. Experimental results show that Pearson's correlation coefficient with the label is 0.35 and 0.47 for MADRS and SHAPS respectively.

6. 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 agree ment 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 ques tions 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 level of depression subjects.

7. REFERENCES

[1] Stuart A. Montgomery and Marie Asberg, "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 ab solute 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 Holtz man, To'Meisha S Edwards, James W Pennebaker, and Matthias R Mehl, "Depression, negative emotional ity, 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 Marki tantov, 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 de pression detection in natural environmental conditions using dilated CNNs," in *Proc. Interspeech 2020*, 2020.

[10] Katie Matton, Melvin G. McInnis, and Emily Mower Provost, "Into the Wild: Transitioning from Recogniz ing Mood in Clinical Interactions to Personal Conver sations for Individuals with Bipolar Disorder," in *Proc. Interspeech 2019*, 2019, pp. 1438–1442.

[11] Zakaria Aldeneh, Mimansa Jaiswal, Michael Picheny, Melvin G. McInnis, and Emily Mower Provost, "Iden tifying Mood Episodes Using Dialogue Features from Clinical Interviews," in

Proc. Interspeech 2019*, 2019, pp. 1926–1930.

[12] 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. Inter speech 2019*, 2019, pp. 3028–3032.

[13] 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.

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

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

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

[17] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Han nemann, 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.

[18] Florian Eyben, Felix Weninger, Florian Gross, and Bjorn 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.

[19] Florian Eyben, Klaus Scherer, Bjorn Schuller, Jo- ¨ han Sundberg, Elisabeth Andre, Carlos Busso, Lau rence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong, “The geneva minimal istic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE Transactions on Affec tive Computing*, vol. 7, pp. 1–1, 01 2015.