Predicting Depression from Speech Recordings: A Machine Learning and Feature Selection Approach

Siamak K. Sorooshyari[1], Thomas Van Vleet[2], Alit Stark-Inbar[2,3], Heather E. Dawes[4], Deanna L. Wallace[4], Morgan B. Lee[4], Michael M. Merzenich[2], Edward F. Chang[4], Mor Nahum[5]

[1]Department of Integrative Biology, [3]Department of Psychology, University of California Berkeley, Berkeley, CA, USA
[2]Department of R&D, Posit Science Corporation, San Francisco, CA, USA
[4]Department of Neurological Surgery, UC San Francisco, San Francisco, CA, USA
[5]Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

Background

1) Importance of detecting depression:
- Mood disorders, such as major depressive disorder (MDD) afflict a significant portion of the population and are a costly public health issue.
- Characterization of day-to-day variation in symptoms of mood disorders are limited and difficult.

2) Predicting depression from speech:
- Changes in voice have been associated with mood states and MDD.
- Remotely administered voice capture tasks are cost-effective mood screener with tracking capability.
- Little consensus exists on the appropriate combinations of voice features required to reliably characterize mood.

3) Machine learning (ML) techniques:
- Analytically-justified and provide predictive capability.

Experimental Methods

1) Participants:
- N=49 ages 18-68 (23 females; mean age = 26.6 ± 11.8)
- Completed self-report and voice capture-based assessments using iPads.
- PHQ-9 was used to assess DSM-V symptoms of depression experienced in the two-weeks preceding administration in adults.

2) Mood categorization of participants:
- PHQ-9 threshold = 9 was used to differentiate depressed vs. non-depressed.
  - 37 non-depressed:
    - 23 with PHQ-9 scores of 0-4 (no/minimal depression)
    - 14 with PHQ-9 scores of 5-9 (mild depression)
  - 12 depressed
    - 7 with PHQ-9 scores of 10-14 (moderate depression)
    - 5 with PHQ-9 scores >14 (moderate/severe depression)

3) Tasks to capture speech recordings:
- Paragraph Reading task
- Story Teller task (spontaneous speech)

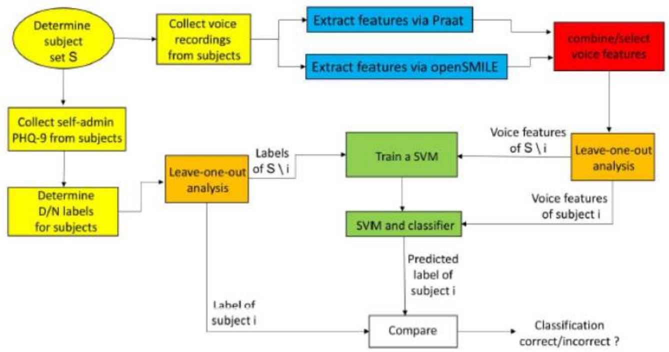
4) Evaluated phonetic, prosodic, and spectral features:
- Combine Praat features from three works: [1][2][3]

Feature Group	Feature Index (FI)	Feature Description	Designation
DeJong (D) [1]	1	# of syllables	nsyll
	2	# of pauses/silences	npause
	3	Duration of speech	dur
	4	Phonation time	phon time
	5	Speech rate (nsyll/dur)	speechrate
	6	Articulation rate (nsyll/phon time)	artic rate
	7	Average syllable duration	ASD
Kawahara (K) [2]	8	Mean of the per-syllable average intensities calculated across the wav file.	E[Avg Int]
	9	Mean of the per-syllable minimum intensities calculated across the wav file.	E[Min Int]
	10	Mean of the per-syllable maximum intensities calculated across the wav file.	E[Max Int]
	11	Average # of intervals in a .wav file	E[# of intervals]
Mielke (M) [3]	12-21	Mean of first 5 formants and their associated bandwidths computed across the wav file.	$E[F_i]$, $E[BW_i]$: i=1 to 5
	22-31	Standard deviation of first 5 formants and their associated bandwidths averaged across the wav file.	$SD[F_i]$, $SD[BW_i]$: i=1 to 5

- Use two openSMILE feature sets:
  - IS10 paraling.conf, 1582 features [4]
  - IS13 ComParE.conf, 6373 features [5]

Analysis and Findings

1) Overview of Voice Capture methodology:



2) Machine learning specifics:
- SVM with linear kernel, C=60.
- Leave one out (LOO) cross-validation analysis done across N=49 participants.
- Compute following metrics to assess predictive capability/accuracy:
  - $LOOC_i$: LOO classification accuracy with ith participant left-out
  - MLCA: mean LOO classification accuracy $MLCA = \frac{1}{N}\sum_{i=1}^{N} LOOC_i$
  - CPLCA: cross-participant LOO classification accuracy $\mathrm{CSLCA} = \frac{1}{N}\sum_{i=1}^{N} 1(LOOC_i > 0.5)$
  - FDR and MDR*: false discovery rate and missed diagnosis rate

3) Predictive capability with combinations of Praat features:

	D	K	M	D + K	K + M	D + M	D + K + M (full feature set)
# of features	7	4	20	11	24	27	31
MLCA	0.75	0.76	0.54	0.7	0.57	0.66	0.5
CPLCA-RU	0.76	0.76	0.57	0.76	0.65	0.73	0.55
CPLCA-RD	0.76	0.76	0.49	0.73	0.53	0.61	0.47
FDR	1.0 (1/1)	0.5 (1/2)	0.85 (11/13)	0.67 (2/3)	0.73 (8/11)	0.55 (6/11)	0.78 (14/18)
MDR	0.25 (12/48)	0.23 (11/47)	0.28 (10/36)	0.24 (11/46)	0.24 (9/38)	0.18 (7/38)	0.26 (8/31)

- Performance with 31 Praat features (D+K+M) is poor:
  MLCA = 0.5, CPLCA-RU = 0.55, CPLCA-RD = 0.47
- Voice intensity features (i.e. K) showed the best predictive capability:
  MLCA = CPLCA-RU = CPLCA-RD = 0.76
- Phonetic features (i.e. D) also performed well:
  MLCA = 0.75, CPLCA-RU = CPLCA-RD = 0.76
- Spectral features (i.e. M) showed poor predictive capability:
  MLCA = 0.54, CPLCA-RU = 0.57, CPLCA-RU = 0.49

4) Predictive capability with openSMILE features:

	oS IS10p	oS IS13cp
# of features	1582	6373
MLCA	0.61	0.75
CPLCA-RU	0.76	0.76
CPLCA-RD	0.53	0.73
FDR	0.5 (4/8)	0.5 (3/6)
MDR	0.2 (8/41)	0.21 (9/43)

- IS13 ComParE.conf features had better performance than IS10 paraling.conf features
  MLCA = 0.75 vs. 0.61, and CPLCA-RD = 0.73 vs. 0.53

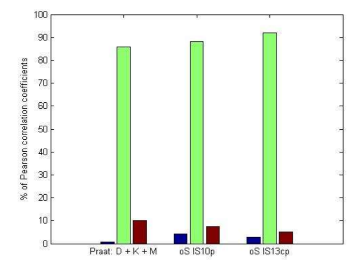
5) Predictive capability with Praat features and feature pruning:
- Remove one feature at a time and do LOO analysis.
- Repeat over all combinations of two features.
- Below results for feature pruning are best cases attained in terms of predictive accuracy

	D + K + M (full feature set)	D + K + M (1 feature pruned)	D + K + M (2 features pruned)
# of features	31	30	29
MLCA	0.5	0.67	0.72
CPLCA-RU	0.55	0.76	0.82
CPLCA-RD	0.47	0.65	0.76
FDR	0.78 (14/18)	0.67 (8/12)	0.42 (5/12)
MDR	0.26 (8/31)	0.22 (8/37)	0.14 (5/37)

- Best performance when pruning one feature $SD[F_2]$ (M-group)
  MLCA = 0.667, CPLCA-RD = 0.653
- Best performance when pruning two features E[Min Int] (K-group) and $E[BW_2]$ (M-group):
  MLCA = 0.72, CPLCA-RU = 0.82
- Noticeable improvements in performance when optimally pruning 1 and 2 features.

6) Compare predictive capability of Praat to openSMILE:
- Important to know which software and feature-group to consider.
- openSMILE IS13 ComParE.conf performs better than Praat features.
- Optimal pruning of 2 Praat features performs better than two openSMILE options.

7) Examine correlation structure among voice features across participants:
- Compute Pearson correlation coefficients, and quantize into 3 correlation levels.
- Large majority of features fall into the uncorrelated category.
- openSMILE IS13 ComParE.conf two openSMILE options.
- Correlation structure among features does not translate into classifier performance and predictive capability.



Conclusions

- Results provide encouraging evidence for remotely recorded speech as an effective means of predicting depression.
- Voice intensity and phonetic features yield better predictive capability than spectral features.
- Larger number of features does not necessarily result in superior classification.
- Feature selection and pruning the feature space is important prior to training ML algorithm.

References

[1]: DeJong, N.H. & Wempe T. Praat script to detect syllable nuclei and measure speech rate automatically. Behav. Res. Methods., 2009.
[2]: Kawahara, S. get_intensity_min_max.praat. Available from http://user.keio.ac.jp/~kawahara/resource.html, 2010.
[3]: Mielke, J. get_formants.praat. Available from https://phon.wordpress.ncsu.edu/lab-manual/scripts/praat-scripting, 2019.
[4]: Schuller, B. et al. The INTERSPEECH 2010 paralinguistic challenge. Interspeech, 2010.
[5]: Schuller, B. et al. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. Interspeech, 2013.