INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Into the Wild: Transitioning from Recognizing Mood in Clinical Interactions to Personal Conversations for Individuals with Bipolar Disorder

Katie Matton[1], *Melvin G McInnis*[2], *Emily Mower Provost*[1]

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, Michigan, USA
[2]Psychiatry, University of Michigan, Ann Arbor, Michigan, USA

{katiemat, mmcinnis, emilykmp}@umich.edu

Abstract

Bipolar Disorder, a mood disorder with recurrent mania and depression, requires ongoing monitoring and specialty management. Current monitoring strategies are clinically-based, engaging highly specialized medical professionals who are becoming increasingly scarce. Automatic speech-based monitoring via smartphones has the potential to augment clinical monitoring by providing inexpensive and unobtrusive measurements of a patient's daily life. The success of such an approach is contingent on the ability to successfully utilize "in-the-wild" data. However, most existing work on automatic mood detection uses datasets collected in clinical or laboratory settings. This study presents experiments in automatically detecting depression severity in individuals with Bipolar Disorder using data derived from clinical interviews and from personal conversations. We find that mood assessment is more accurate using data collected from clinical interactions, in part because of their highly structured nature. We demonstrate that although the features that are most effective in clinical interactions do not extend well to personal conversational data, we can identify alternative features relevant in personal conversational speech to detect mood symptom severity. Our results highlight the challenges unique to working with "in-the-wild" data, providing insight into the degree to which the predictive ability of speech features is preserved outside of a clinical interview.

Index Terms: Bipolar Disorder, mood prediction, computational paralinguistics, mobile health

1. Introduction

Bipolar Disorder (BD) is a severe and lifelong mental illness characterized by pathological mood transitions into episodes of mania and depression. Current patient monitoring strategies depend on the availability of specialized medical professionals to conduct regular in-person health assessments. However, we face a global shortage of mental health workers, rendering this form of clinical support inaccessible to a majority of those who need it [1]. Even when available, clinical assessment is limited by a reliance on retrospective patient accounts of mood symptoms. Thus, there is a glaring need for more effective mood monitoring methods to supplement existing clinical approaches.

Smartphone-based monitoring is a potential solution [2]. Smartphones can passively collect behavioral data relevant to mood, and smartphone ownership is becoming increasingly ubiquitous [3], meaning such an approach could help to make mental health care more widely accessible. Among the data generated by smartphones, speech is a promising medium for BD mood detection given its clinically validated connection to mood symptoms [4, 5]. An extensive body of work explores the automatic detection of mood states using speech captured in clinical or laboratory settings (see [6] for details). A smaller, more recent collection of work suggests that mood can also be detected from natural interactions captured via smartphones [7, 8, 9]. However, the differences between mood expression inside and outside clinical environments remain underexplored.

We investigate differences between contexts, inside and outside clinical environments, using two types of speech features: speaker timing patterns and language usage. Existing work has shown that timing features related to conversation dynamics, such as number of speaking turns and average speaking length, are useful in classifying BD mood states [9]. Other work has found that depression is associated with a decrease in speech rate and an increase in pause time [10, 11]. Investigations into changes in language use associated with depression have primarily been limited to analyses of social media data, writing, and clinical interactions. Changes observed include an increased usage of both negatively valanced language and first person singular pronouns [12, 13, 14]. These changes, among others, are commonly captured using the Linguistic Inquiry and Word Count (LIWC) tool [15] or n-gram features [16, 17].

In this paper, we present the first study of how language and speech timing patterns change as a function of both depression symptom severity and the context of the interaction. We present models that can detect depression severity from both clinical interactions and personal conversations and show that there are substantial differences in the features important to each task. We demonstrate that features with little value in a natural conversation setting can emerge as maximally important in a clinical setting as a result of their ability to quantify interview structure. Out of the features analyzed, we find that measures of emotion generalize best across the two settings and are especially salient to mood when captured from personal interactions.

2. PRIORI Dataset

The PRIORI dataset is a longitudinal collection of conversational speech data collected from 51 individuals with BD and 9 healthy controls [8]. The inclusion criteria for the study was a diagnosis of BD type I or II. Exclusion criteria were the presence of a co-morbid illness or a history of substance abuse. Participants were enrolled for six to twelve months, and during this time, they used a smartphone with a secure recording application installed. The app recorded their side of all incoming and outgoing calls and then securely uploaded the data to a HIPAA-compliant server. The dataset contains 52,931 calls amounting to 4,000+ hours of speech. There are two types of calls, assessment and personal, which we will describe in more detail.

Assessment calls: Study participants engaged in weekly interviews with clinicians using the smartphones provided. During these calls, a clinician evaluated the participant's mood using

http://dx.doi.org/10.21437/Interspeech.2019-2698

Table 1: *Distribution of mood in all data and in test data.*

	# Assessment Calls		# Personal Calls	
	Total	Test	Total	Test
Depressed	155	100	127	87
Euthymic	131	67	88	50

the Hamilton Depression Scale (HamD) and the Young Mania Rating Scale (YMRS) [4, 5]. The HamD and YMRS scores are the ground truth prediction targets in our experiments.

Personal calls: All other calls in the dataset are examples of "in-the-wild" data and consist of participants interacting naturally with their personal contacts. Since the HamD and YMRS scales are used to assess mood with respect to the previous week, the assessment ratings can be applied to all personal calls made in the week leading up to the assessment. However, we hypothesize that the ratings are most strongly influenced by the mood symptoms exhibited by the participant on the day of the interview from which they were obtained. We restrict the set of personal calls used to those that occurred on the same day as an assessment to render it more likely that the assessment label is relevant to the personal calls. We group all personal calls for each subject on the day of an assessment into a single observation by concatenating their transcripts. From here on, we refer to an assessment call and to grouped personal calls as a "call".

Data selection: The size of the dataset we use is substantially reduced compared to the full PRIORI dataset. In addition to the restriction on personal calls previously described, we apply several other selection criteria. We use only calls: (1) recorded on Samsung S5 devices due to the lower recording quality of the Samsung S3 and S4 devices, also used in this study, which negatively affected Automatic Speech Recognition (ASR) performance (see [18]), (2) with enough speech for feature extraction (see Section 3.1), (3) with a depressed or euthymic (asymptomatic) label, and (4) in which the speaker has BD. Items 3 and 4 will be discussed in the following paragraph. Table 1 shows a description of the data after applying this criteria.

Outcomes of interest: The conventional approach to mood tracking is to predict mood classes [7, 9, 18, 19]. However, monitoring mood fluctuations requires fine-grained measurement. Thus, we treat the problem as a regression task and predict mood symptom severity. We focus on depression prediction because there are relatively few examples of mania in the dataset. Further, we only use calls in which the speaker was evaluated to be euthymic (HamD $<$ 6, YMRS $<$ 6) or depressed (HamD $>$ 10, YMRS $<$ 6). This restriction has been applied in previous work [18, 19] and allows us to focus on cases of clear symptoms, making it easier to interpret the effects of context. We exclude observations from healthy controls because our goal is to track within-subject mood changes rather than to distinguish between subjects with and without BD. In order to better isolate the effects of context, we use the same set of subjects for both our assessment and personal call experiments.

Train-test split: We test in a leave-one-subject-out (LOSO) manner, ensuring that there is no overlap between the train and test subjects in each fold. We require that subjects used for testing have at least two euthymic and two depressed observations in both the assessment and personal call data (12 subjects). There are 23 subjects that do not fit this criteria but which have at least one observation in both types of data. For each test fold, we train on the subset of the set of 12 subjects that excludes the current test subject (11 subjects) and the set of 23 subjects.

3. Methods

3.1. Data pre-processing

We segment the calls into regions of continuous speech using the COMBO-SAD algorithm [20] and the approaches outlined in our previous work [18]. We transcribe the segments using an ASR model, implemented in Kaldi using the 'nnet2' recipe [21]. The ASR model obtained an average word error rate of 39.7% per segment when tested on the manually transcribed subset of the PRIORI dataset (around 25 hours of speech) [22]. We use only calls with at least 100 words and 5 speech segments, as preliminary evidence suggested features extracted on shorter calls were either extracted incorrectly or not meaningful.

3.2. Feature extraction

We identify a set of speech features that are motivated by clinical studies of BD symptoms as well as existing work in automatic mood state detection. Some of these features are computed by applying a set of statistics, which are consistently: mean, median, standard deviation, min, and max.

Speech intelligibility: ASR has higher confidence for well enunciated speech. Previous work has found a link between depression severity and a disruption in articulatory precision [23]. We capture this phenomenon by extracting statistics from the segment-level ASR confidence measures to produce call-level features. We also extract the proportion of out-of-vocabulary (OOV) words in each call.

Non-verbal expressions: Non-verbal behavior provides information regarding emotion and psychological well-being. We compute counts of the instances of laughter and noise detected by the ASR model, normalized by total word count.

Linguistic style: We present three feature sets that capture different attributes of linguistic style.

Syntax: We use the LIWC dictionary [15], used in previous work in depression classification [14, 17], to compute normalized counts of: (1) Part of Speech (POS) categories (e.g. first person pronouns, adverbs) (2) verb tenses (e.g. past, present), (3) swear words, (4) non-fluencies (e.g "hmm", "um"), and (5) fillers (e.g. "you know"). We supplement the 18 POS measures included in LIWC with 5 additional POS categories derived using the Natural Language Toolkit (NLTK) POS tagger and with 13 POS ratio features (e.g. adjective:verbs).

Speech complexity and verbosity: We compute statistics from the number of words and syllables present in each speech segment. We also use mean word length and the fraction of long words (6+ characters) as features.

Speech graphs: Mota et al. created speech graphs to measure thought disorder [24] and psychosis [25]. Thought disturbances, such as rumination, are present in individuals with depression [26, 27], and we hypothesize that the graph measures can be used to capture them. We represent calls graphically using each unique word as a node. We insert an edge for every pair of words uttered consecutively within the same speech segment. We transform each call into three graphs: (1) uses the words directly, (2) uses the lemmatized form of each word, and (3) represents each word as its associated POS. We use 12 graph attributes as features, including average degree, density, diameter, the size of connected components, and loop, node, and edge counts (see [24] for details). We also include a version of each feature that is normalized by total word count.

Semantic content: We capture speech content using two different approaches. We use LIWC to measure the presence of psychologically meaningful categories, such as emotion (e.g. anger, anxiety), biological processes (e.g. body, health), and personal concerns (e.g. work, death). Between these and the LIWC linguistic style features, we use all categories from the LIWC 2007 dictionary. In addition, we use Term Frequency-Inverse Document Frequency (TF-IDF) features to encode the use of specific words. We collect a vocabulary using all calls in the PRIORI dataset with 50+ words that were recorded on Samsung S5 devices (22,939 calls). We take all unigrams present in 10+ calls and all bigrams present in 50+ calls, resulting in a set of of 20,738 n-grams. When computing TF-IDF values, we normalize the term frequency by total word count.

Speaker timing: We use Kaldi to produce aligned word and phone timing annotations for each call. We extract the following features for words, phones, and pauses: (1) statistics over the durations of all instances (e.g. mean word duration), (2) statistics over the per second timing within all segments (e.g. mean words per second across segments), (3) total count (e.g. number of words), and (4) overall per second timing (e.g. words per second). We also extract total call duration, total participant speaking duration, ratio of participant speaking duration to total duration, total pause duration, ratio of pause duration to total duration, segment count, segments per minute, count of short utterances (lasting less than 1-second), and short utterances per minute, some of which were motivated by [9].

3.3. Data modeling

We use linear regression due to its high interpretability and effectiveness at modeling datasets of limited size. We assess performance with LOSO cross-validation across the 12 test subjects (see Section 2). For each fold, we first use the training data to eliminate all features that do not have statistically significant Pearson Correlation Coefficients (PCC) with the ground truth depression ratings. We use a p-value of .01 as a cutoff for all features aside from the TF-IDF features. The number of TF-IDF features (20,738) is much larger than the total number of all other features (217). We apply a stricter p-value cutoff of .0001 for the TF-IDF features to create a balance between the feature sets and to compensate for the increased chance of spurious correlations that arises from performing a large number of statistical tests. After this step, we order the remaining features by magnitude of correlation. We perform a second (nested) stage of LOSO cross-validation, this time over the training speakers with sufficient mood examples (11 subjects), to select the best number of features with respect to this ranked list. We evaluate performance using the PCC between the predicted and target scores. Our goal, as defined by our clinical team, is to detect individual-specific mood irregularities rather than to precisely replicate clinical interview ratings. This makes being able to determine the relative symptom severity of an individual at two points in time more important than optimizing based on absolute error.

4. Results

We obtain a PCC of .64 for assessment calls with a standard deviation of .12 across subjects and a PCC of .32 for personal calls with a standard deviation of .25 across subjects. These results demonstrate that we can detect depression severity from both structured assessments and "in-the-wild" personal phone calls. Our performance is better on assessment calls than on personal calls, and in our remaining analysis, we examine the factors that contribute to this performance disparity.

Table 2: *Feature ablation results: mean and standard deviation of the PCC between predicted and target depression ratings.*

Features	Assessment	Personal
Speech Intelligibility	-	.15 ± .40
Non-Verbal Expressions	-	-
Ling. Style	.30 ± .38	-
Speaker Timing	.63 ± .15	-
LIWC (emotion only)	.10 ± .32	.25 ± .42
TF-IDF	.46 ± .22	-

We perform a feature ablation study, training models using each feature category (see Section 3.2) individually. We apply the same feature set reduction steps described previously (see Section 3.3), keeping TF-IDF features correlated with $p < .0001$ and all other features with $p < .01$. If the resulting feature set is empty, we do not present a correlation score ('–' in Table 2). There is a marked dissimilarity between the feature sets effective at detecting depression severity in the assessment and personal data. While the linguistic style, speaker timing, and TF-IDF feature sets have high predictive value for the assessment calls, they are not significantly correlated with depression for the personal call data. We also observe that the LIWC emotion and speaker intelligibilty features have increased usefulness for personal calls. These two feature types may be connected; previous work has shown that speech recognition performance degrades when the speech is emotional [28, 29, 30].

We examine the utility of individual features by computing their correlation with the clinically derived depression severity ratings (both are call-level). We exclude TF-IDF features from this analysis to limit the number of statistical tests applied. We adopt the Bonferroni correction when assessing significance values and use $\alpha = .05/217 = 2.3e^{-4}$. The most strongly correlated features for the assessment and personal call data are presented in Table 3. There are 21 features that are statistically significant for the clinical data and 11 for the personal data, with notably little overlap between the two sets. Only two features, negative emotion and anxiety, are significant for both data types.

5. Discussion

We identify key patterns from our results that have meaningful implications for the development of automatic mood detection systems using both clinical and "in-the-wild" data.

5.1. Benefits of structure in clinical interactions

Features that directly or indirectly measure call duration, such as total duration, segment count, and short utterance count, have a strong positive correlation with depression severity for assessment calls, but are not statistically significant for personal calls (Table 3). The structure of the HamD interview provides an explanation for this: many of the interview questions, such as "have you experienced feelings of guilt?" and "have you had difficulty falling asleep?", naturally elicit longer responses from individuals who have experienced the symptoms in question. In inspecting the assessment calls, we also find that clinicians tend to ask more follow-up questions to highly symptomatic individuals. These findings provide insight into why a model that uses only speaker timing features is effective at detecting mood for assessment calls, but not for personal calls (Table 2).

Assessment calls are explicitly focused on mood symptoms, which makes mood detection based on word choice easier. As seen in Table 2, a model trained using only TF-IDF features ob-

Table 3: *Correlation with depression severity of top ten features for assessment and personal calls (showing correlation for both assessment and personal calls) excluding TF-IDF measures. Significance is asserted (*) for $\alpha = 2.3e^{-4}$ (Bonferroni correction).*

Best Assessment Features	Assessment	Personal	**Best Personal Features**	Assessment	Personal
noun	.46*	.23	negative emotion (LIWC)	.25*	.37*
total duration	.42*	.11	laughter	$-.04$	.32*
ave degree pos graph norm	$-.41$*	$-.15$	ASR confidence median	$-.07$	$-.32$*
segment count	.37*	.07	anger (LIWC)	.04	.31*
determiner	$-.30$*	$-.20$	ASR confidence mean	$-.08$	$-.31$*
assent (LIWC)	.29*	.23	anxiety (LIWC)	.23*	.30*
lemma graph density	$-.28$*	$-.01$	death (LIWC)	.12	.30*
naive graph density	$-.27$*	$-.01$	noise	.04	.29*
leisure (LIWC)	$-.26$*	$-.11$	ASR confidence std	.09	.28*
short utterance count	.25*	.05	pauses per second	.01	.25*

tains a PCC of .46 for the assessment calls, but fails to work for personal calls (we also tested a less restrictive p-value cutoff of .1 to produce a non-empty feature set and obtained a PCC of $-.02$). We examine the TF-IDF features selected by the model on the assessment calls and present the features selected by all training folds in Table 4. We find that most of the features selected, such as "yes", "normal", and "really bad", appear to be direct responses to questions about mood symptom severity. As shown in Table 3, we also observe that assessment calls have higher correlation scores than personal calls for assent (which provides a measure of "yes" responses) and leisure (which relates to the topic of outside activities on the HamD test). Again, we see that the narrow topic focus of the assessment interviews increases the utility of certain language features. These results highlight the increased difficulty of finding language patterns that generalize across subjects when using "in-the-wild" data, especially for datasets of limited size.

We examine how performance changes when we explicitly remove features that capture the structure of the clinical interview. First, we exclude all features that directly or indirectly measure call length: all duration measures, unnormalized counts, and most of the graph features (many are correlated with call duration with a PCC $> .6$). The results are displayed in Table 5. With this restriction, the performance of the model on the assessment calls drops from a PCC of .64 to a PCC .51. We also test the exclusion of the TF-IDF features and find that performance drops to a PCC of .56. Without call length related features or TF-IDF features, the correlation obtained (PCC $= .31$) is comparable to that obtained on the personal calls. When we apply these same restrictions to the personal calls, the performance remains relatively constant, indicating that these features lack significance as measures of natural interactions.

5.2. Emotional openness in natural interactions

We see that features that measure emotional distress, such as negative emotion, anger, and anxiety, have increased correlation scores for the personal calls, compared to the assessment calls (Table 3). Surprisingly, we also observe a high correlation between laughter and depression. However, upon investigating, we found that our ASR model, which was not trained to detect crying, often outputs either "laughter" or "noise" when crying occurs in a speech segment. Therefore, the high correlation of these two features with depression is likely explained by an increase in instances of crying. These results, which are unique to the personal calls, suggest that individuals display more openness in their emotional expressions during interactions with personal contacts than they do in a clinical interview setting. Our findings provide evidence that this greater range and intensity of emotional expression is useful for mood detection.

Table 4: *Features selected by TF-IDF-only model on assessment calls: mean and standard deviation of model coefficients.*

Feature	β	Feature	β
yes	$2.3 \pm .49$	people	$.84 \pm .16$
good	$-1.14 \pm .35$	bad	$.61 \pm .18$
normal	$-1.12 \pm .28$	hand	$.60 \pm .21$
yeah	$.93 \pm .14$	nope	$-.56 \pm .15$
really bad	$.90 \pm .10$	every day	$.42 \pm .33$

6. Conclusion

In this paper, we demonstrate that the predictive ability of speech features varies depending on the nature of the interactions from which they are derived. We present evidence that the structure of clinical interviews simplifies the task of mood detection; features that capture aspects of interview structure are effective at detecting mood from the assessments calls, but are less useful when applied to the "in-the-wild" personal call data. However, we still successfully detect mood from the personal calls using an alternative set of features that are made more salient by the increased emotional openness present in natural interactions. In doing so, our work demonstrates the potential for smartphone-based monitoring of BD mood symptoms. In future work, we plan to fine-tune our ASR model with the manually transcribed subset of our dataset to examine the effect of increased ASR accuracy on system performance. We also will explore how our findings translate to larger datasets and to the detection of manic symptom severity.

7. Acknowledgements

This work was supported by the National Science Foundation (CAREER-1651740), NIMH R34MH100404, the National Institute of Mental Health (R01MH108610, R34MH100404), the Heinz C Prechter Bipolar Research Fund and the Richard Tam Foundation at the University of Michigan. The data collection effort was reviewed and approved by the Institutional Review Board of the University of Michigan, HUM00052163.

Table 5: *Ablation results for features associated with the structure of a clinical interview: mean and standard deviation of the PCC between predicted and target depression ratings.*

Features	Assessment	Personal
All	$.64 \pm .12$	$.32 \pm .25$
All - Call Length	$.51 \pm .22$	$.31 \pm .26$
All - TF-IDF	$.56 \pm .16$	$.34 \pm .27$
All - Call Length - TF-IDF	$.31 \pm .23$	$.33 \pm .27$

8. References

[1] T. Butryn, L. Bryant, C. Marchionni, F. Sholevar *et al.*, “The shortage of psychiatrists and other mental health providers: causes, current state, and potential solutions,” *International Journal of Academic Medicine*, vol. 3, no. 1, p. 5, 2017.

[2] M. Faurholt-Jepsen, M. Bauer, and L. V. Kessing, “Smartphone-based objective monitoring in bipolar disorder: status and considerations,” *International journal of bipolar disorders*, vol. 6, no. 1, p. 6, 2018.

[3] E. J. Topol, S. R. Steinhubl, and A. Torkamani, “Digital medical tools and sensors,” *JAMA*, vol. 313, no. 4, pp. 353–354, 2015.

[4] M. Hamilton, “The hamilton rating scale for depression,” in *Assessment of depression*. Springer, 1986, pp. 143–152.

[5] R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer, “A rating scale for mania: reliability, validity and sensitivity,” *The British Journal of Psychiatry*, vol. 133, no. 5, pp. 429–435, 1978.

[6] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Communication*, vol. 71, pp. 10–49, 2015.

[7] M. Faurholt-Jepsen, J. Busk, M. Frost, M. Vinberg, E. M. Christensen, O. Winther, J. E. Bardram, and L. V. Kessing, “Voice analysis as an objective state marker in bipolar disorder,” *Translational psychiatry*, vol. 6, no. 7, p. e856, 2016.

[8] Z. N. Karam, E. M. Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, and M. G. Mcinnis, “Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4858–4862.

[9] A. Muaremi, F. Gravenhorst, A. Grünerbl, B. Arnrich, and G. Tröster, “Assessing bipolar episodes using speech cues derived from phone calls,” in *International Symposium on Pervasive Computing Paradigms for Mental Health*. Springer, 2014, pp. 103–114.

[10] Z. Liu, H. Kang, L. Feng, and L. Zhang, “Speech pause time: A potential biomarker for depression detection,” in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 2020–2025.

[11] J. D. Teasdale, S. J. Fogarty, and J. M. G. Williams, “Speech rate as a measure of short-term variation in depression,” *British Journal of Social and Clinical Psychology*, vol. 19, no. 3, pp. 271–278, 1980.

[12] S. Rude, E.-M. Gortner, and J. Pennebaker, “Language use of depressed and depression-vulnerable college students,” *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[13] N. S. Holtzman *et al.*, “A meta-analysis of correlations between depression and first person singular pronoun use,” *Journal of Research in Personality*, vol. 68, pp. 63–68, 2017.

[14] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, “Predicting depression via social media,” in *Seventh international AAAI conference on weblogs and social media*, 2013.

[15] J. Pennebaker, C. Chung, M. Ireland, A. Gonzales, and R. J Booth, “The development and psychometric properties of liwc2007,” 2007.

[16] M. De Choudhury, S. Counts, and E. Horvitz, “Social media as a measurement tool of depression in populations,” in *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 2013, pp. 47–56.

[17] G. Coppersmith, M. Dredze, and C. Harman, “Quantifying mental health signals in twitter,” in *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 2014, pp. 51–60.

[18] J. Gideon, E. M. Provost, and M. G. McInnis, “Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder,” *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2359–2363, 2016.

[19] S. Khorram, J. Gideon, M. G. McInnis, and E. M. Provost, “Recognition of depression in bipolar disorder: Leveraging cohort and person-specific knowledge.” in *INTERSPEECH*, 2016, pp. 1215–1219.

[20] S. O. Sadjadi and J. H. Hansen, “Unsupervised speech activity detection using voicing measures and perceptual spectral flux,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” IEEE Signal Processing Society, Tech. Rep., 2011.

[22] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E. M. Provost, “The priori emotion dataset: Linking mood to emotion detected in-the-wild,” *arXiv preprint arXiv:1806.10658*, 2018.

[23] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, “Classification of depression state based on articulatory precision.” in *Interspeech*, 2013, pp. 2172–2176.

[24] N. B. Mota, N. A. Vasconcelos, N. Lemos, A. C. Pieretti, O. Kinouchi, G. A. Cecchi, M. Copelli, and S. Ribeiro, “Speech graphs provide a quantitative measure of thought disorder in psychosis,” *PloS one*, vol. 7, no. 4, p. e34928, 2012.

[25] F. Carrillo, N. Mota, M. Copelli, S. Ribeiro, M. Sigman, G. Cecchi, and D. F. Slezak, “Automated speech analysis for psychosis evaluation,” in *Machine Learning and Interpretation in Neuroimaging*. Springer, 2013, pp. 31–39.

[26] L. C. Foland-Ross, J. P. Hamilton, J. Joormann, M. G. Berman, J. Jonides, and I. H. Gotlib, “The neural basis of difficulties disengaging from negative irrelevant material in major depression,” *Psychological science*, vol. 24, no. 3, pp. 334–344, 2013.

[27] I. Demeyer, E. De Lissnyder, E. H. Koster, and R. De Raedt, “Rumination mediates the relationship between impaired cognitive control for emotional information and depressive symptoms: A prospective study in remitted depressed adults,” *Behaviour research and therapy*, vol. 50, no. 5, pp. 292–297, 2012.

[28] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, “Asr for emotional speech: clarifying the issues and enhancing performance,” *Neural Networks*, vol. 18, no. 4, pp. 437–444, 2005.

[29] M. Sheikhan, D. Gharavian, and F. Ashoftedel, “Using dtw neural–based mfcc warping to improve emotional speech recognition,” *Neural Computing and Applications*, vol. 21, no. 7, pp. 1765–1773, 2012.

[30] M. Bashirpour and M. Geravanchizadeh, “Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, p. 9, 2018.