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

Identifying Mood Episodes Using Dialogue Features from Clinical Interviews

Zakaria Aldeneh[1], *Mimansa Jaiswal*[1], *Michael Picheny*[3], *Melvin McInnis*[2], *Emily Mower Provost*[1]

[1]Department of Computer Science and Engineering, University of Michigan, USA
[2]Department of Psychiatry, University of Michigan, USA
[3]IBM T.J. Watson Research Center, USA

{aldeneh, mimansa, mmcinnis, emilykmp}@umich.edu, picheny@us.ibm.com

Abstract

Bipolar disorder, a severe chronic mental illness characterized by pathological mood swings from depression to mania, requires ongoing symptom severity tracking to both guide and measure treatments that are critical for maintaining long-term health. Mental health professionals assess symptom severity through semi-structured clinical interviews. During these interviews, they observe their patients' spoken behaviors, including both what the patients say and how they say it. In this work, we move beyond acoustic and lexical information, investigating how higher-level interactive patterns also change during mood episodes. We then perform a secondary analysis, asking if these interactive patterns, measured through dialogue features, can be used in conjunction with acoustic features to automatically recognize mood episodes. Our results show that it is beneficial to consider dialogue features when analyzing and building automated systems for predicting and monitoring mood.

Index Terms: Spoken dialogue, Mood Modeling, Bipolar Disorder, Depression, Mania

1. Introduction

Bipolar disorder (BP) is a mental disorder that affects more than 20 million people worldwide [1]. Individuals who suffer from BP experience mood episodes that range from mania (elevated mood) to depression (lowered mood). These mood episodes have negative consequences on an individual's life and work. The current standard for monitoring symptom severity is through regular clinical interviews. In these interviews, a mental health professional asks a series of questions that address different aspects of a patient's life and mood. The success of these interviews relies on how a clinician interprets both the verbal and non-verbal cues of a patient's response to the open-ended interview questions, making the interviews subjective in nature. In addition, regular clinical appointments can be both costly and time consuming. Recent advances in affective computing have made it possible to develop automated methods to help make the current standard not only more objective, but also more accessible to a wider range of people. Automated approaches can also help quantify patient's behavior and provide clinicians with actionable data. This work focuses on leveraging interaction patterns from clinical interviews for automatically recognizing mood episodes.

Previous research showed that patient mood episodes affect the acoustics of both the patients and the clinicians in clinical interviews. For instance, variations in pitch, speaking rate, and intensity in a patient's speech were shown to correlate with symptom severity [2–8]. Given the dyadic nature of the clinical interviews, changes in speech patterns of the patient affect speech patterns of the clinician through the phenomenon of entrainment [9–13]. These observed changes in speech patterns of both the patient and the clinician in the dialogues affect the dynamics of the interaction [14]. In other words, we expect patients' mood episodes to affect not only patient and clinician acoustics, but also the turn-taking behaviour in clinical dialogues. The goal of this work is to study the effect of mood changes on interaction dynamics in clinical interviews. We show that the extracted high-level dialogue features can be used to augment prosodic features to improve the performance of automatically detecting depression severity in clinical interviews. Our results show that mood measurably affects interaction patterns, and that it can be beneficial to take dialogue features into account when building automated agents for conducting clinical interviews.

2. Related Work

Although the use of acoustic bio-markers has not yet been embraced in clinical practice, there has been extensive research done on the topic. We refer the reader to a paper by Cummins et al. [15] for a more comprehensive literature review.

Prosody. Prosodic features capture variations in rhythm, stress, and intonation of speech. They represent information present in syllables and larger units of speech (i.e., not at the phone level). A study by Hashim et al. [16] showed that acoustic measures that capture timing properties of speech are predictive of depression scores. Specifically, the authors showed that features that capture transitions between different voicing properties can be effective for predicting clinical depression scores. Many studies have shown the effectiveness of pitch and rhythm features for detecting mood state [2, 5, 7, 8, 17].

Interaction. Modulations in patients' speech patterns affect clinicians' speech patterns during clinical interviews [2]. Acoustic features extracted from the clinician during clinical interviews have been shown to correlate with the mood symptom severity of patients. For instance, Scherer et al. [17] showed that acoustic features extracted from interviewers varied with patients' depression severity. The authors also found that some entrainment measures had an inverse relationship with depression severity. Dibeklioglu et al. [18] and Yu et al. [3] showed that conversational features (e.g., onset time and utterance durations) can be effective for detecting mood symptom severity in patients.

In contrast to previous work, the novelty of our work is three-fold: (1) we introduce a set of dialogue features to aid in the prediction of mood symptom severity; (2) we analyze dialogue features using a linear mixed effect model to study how mood episodes affect interaction patterns; (3) we show that explicitly adding high-level dialogue features to acoustic-based systems can improve the performance of automatic mood symptom severity prediction.

http://dx.doi.org/10.21437/Interspeech.2019-1878

3. The Data

The PRIORI (Predicting Individual Outcomes for Rapid Intervention) dataset is a longitudinal dataset composed of cellphone recordings collected as part of a large-scale effort to study how the properties of speech change with mood symptom severity [7]. The participants in the study include individuals who are diagnosed with type-I or type-II BP. Participants were provided with a smartphone equipped with specialized secure recording software that they used as their primary cellphone during the duration of the study (maximum duration of 13-months). The software records only their side of the conversation for both incoming and outgoing calls. It encrypts the speech in real-time and then uploads the data to a HIPAA compliant server for offline processing.

The data include two types of calls: assessment calls and personal calls. Assessment calls are weekly calls between a participant and a clinician in which the clinician evaluates the participant's mood using the the Hamilton Depression Scale (HAMD) [19] and the Young Mania Rating Scale (YMRS) [20] as part of a clinical interview. YMRS is a rating scale used for assessing mania severity while HAMD is a rating scale used for assessing depression severity. In our dataset, both YMRS and HAMD scores range from 0 (least symptomatic) to 35 (most symptomatic). Whereas personal calls were only collected from patient-side cellphone microphones, assessment calls were collected from two sources: patient-side cellphone microphones, and clinician-side landline telephone recorder microphone. This gives us the data needed for studying clinician-patient interactions.

The assessment call component of the PRIORI dataset contains over 380 hours of speech from over 1,280 assessment calls. Following the work of Gideon et al. [8], we define three mood episodes based on the HAMD and YMRS scores: euthymic state (YMRS ≤ 6 & HAMD ≤ 6), depressed state (YMRS ≤ 6 & HAMD ≥ 10), and manic state (YMRS ≥ 10 & HAMD ≤ 6). We exclude all calls outside of these ranges. Our final dataset includes 155 hours of speech from 641 calls (317 euthymic, 268 depressed, 56 manic) by 47 unique speakers (34 females) and 9 unique clinicians (5 females).

3.1. Extracting Speaker Turns

The patient-side cellphone recordings of the assessment calls contain single channel streams with the patient's speech signal. The clinician-side landline telephone recordings of the assessment calls contain single channel streams with both the clinician's and the patient's speech signals. The goal is to obtain the start and end times of each speech segment in the conversation. To do so, we cross-correlate the two signals in the frequency domain and use the maximum value as the alignment offset. Once the two signals are aligned, we run COMBO-SAD, a voice activity detection (VAD) algorithm by Sadjadi and Hansen [21], to extract speech segments from the two aligned signals. The VAD output from the cellphone recordings gives the patient turns, while the VAD output from the landline recordings gives the merged patient and clinician turns. Regions where the two VAD signals overlapped were assigned "patient speech". Regions where the two VAD signals did not overlap were assigned "clinician speech". To form speaker turns, we merge speech activity from a single speaker whenever there is silence that is $<$500 milliseconds separating speech segments. We pick this value as conventional spoken dialog systems use a silence threshold of around 500 milliseconds to determine utterance end-points [22].

4. Features

4.1. Dialogue Features

We extract a set of high-level dialogue features to quantify the patient-clinician interactions motivated by the fact that patients experiencing different mood episodes display less expressive interactive styles, increased talkativeness, racing thoughts, and inflated self-esteem [14, 24]. All of the dialogue features that we study in this work are time-based features and can be easily extracted using a conventional VAD. This makes the extracted dialogue features more robust to noisy conditions in the recordings when compared to features that are extracted directly from the acoustic signal.

Floor control ratio. This feature measures the relative amount of time an individual spends speaking to the total amount of speech in a conversation. Floor control has been studied in the entrainment, turn-taking, and dialogue literature [25, 26]. This feature can quantify dominance, brevity, and relative duration of the patient's response to the interview questions.

Turn hold offset. This feature measures the duration of pauses that are less than half a second within turns from the same speaker. Turn hold offset is a well-studied feature in the turn-taking literature [27, 28]. Previous work showed that depressed individuals tend to have longer pauses in their speech [16].

Number of consecutive turns. This feature measures the tendency of a speaker to hold the floor in a conversation. In other words, this feature measures the tendency for a speaker to include long pauses ($>$500 milliseconds) between his or her sentences.

Number of turn switches per minute. The current data segmentation approach makes measuring durations (or frequencies) of overlapping speech difficult. Previous research, however, showed that the number of turn switches is correlated with the number of interrupts and overlaps in a conversation [29, 30]. We use this feature as a proxy for the amount of overlapping speech that occurs in a clinical interview.

Turn switch offsets. This feature measures the latency between turn transitions. Previous work showed that different dialogue contexts have different turn switch latencies [31]. Clinically, previous work demonstrated that depressed individuals take longer to respond to clinicians' questions [3].

Turn lengths. This feature measures the duration of each turn by a speaker. Previous research showed that variants of this feature were effective for detecting depression [3].

We summarize dialogue features by taking the mean and standard deviation across each conversation. This results in 20 features representing the interactions in each clinical interview.

4.2. Rhythm Features

Previous work showed that speech rhythm features are effective for predicting mood states [2, 6, 8]. We follow the approach mentioned in [8] and extract seven rhythm features using the algorithm proposed by Tilsen and Arvaniti [32]. These rhythm features capture power distribution, rate, and rhythm stability metrics. To obtain call-level features, we calculate statistics over the seven rhythm features, including: mean, standard deviation, kurtosis, skewness, max, min and their normalized locations, linear regression slope, intercept, and error. This results in a total of 70 features representing each assessment call.

Table 1: *List of investigated dialogue features with coefficients and standard errors from LMEMs. The main effect coefficients indicate changes from euthymia to depression (or from euthymia to mania). We report p-values obtained from likelihood ratio test against a null model with no mood effect.* ***Bolded*** *estimates indicate significance after correcting the FDR at $\alpha = 0.05$ [23]. We use '–' to denote estimates for features that showed obvious deviations from homoscedasticity or normality after visually inspecting the residual plots. Unless noted otherwise, all values for time-based features are reported in milliseconds. All ratios are reported as percentages (%). p-value codes: '***' <0.001; '**' <0.01; '*' <0.05; '.' <0.1*

	Depression		Mania	
Feature	**Estimate**	***p*-value**	**Estimate**	***p*-value**
call duration in minutes[†]	**0.578±0.052**	***	**0.468±0.073**	N/A
number of turn-switches per min.	-0.567±0.319	.	0.004±0.526	
patient features				
floor control ratio	**2.657±1.173**	*	**8.276±1.981**	***
hold offset (mean)	-0.263±77.601		-215.175±143.812	
hold offset (SD)	–		–	
number of continuous turns (mean)	**0.072±0.020**	***	0.066±0.034	.
number of continuous turns (SD)	**0.104±0.041**	*	0.140±0.067	*
switch offset (mean)	8.434±26.940		-48.812±41.987	
switch offset (SD)	**63.812±20.925**	**	49.700±34.688	
turn lengths (mean)	75.175±52.825		**313.163±84.050**	***
turn lengths (SD)	12.875±73.000		**299.200±117.013**	*
clinician features				
hold offset (mean)	**128.150±30.200**	***	-16.799±44.070	
hold offset (SD)	**200.262±39.575**	***	-33.312±55.038	
number of continuous turns (mean)	**0.054±0.022**	*	-0.027±0.034	
number of continuous turns (SD)	0.052±0.033		-0.039±0.052	
switch offset (mean)	6.474±33.411		-90.675±50.237	.
switch offset (SD)	55.075±81.487		–	
turn lengths (mean)	**-97.525±41.600**	*	**-215.300±70.100**	**
turn lengths (SD)	-20.137±38.688		**-175.863±61.812**	**

[†] Produced significant ($p < 0.0005$) interaction effect with patient gender control variable in manic episodes. The estimate for males in the manic state was $+0.631 \pm 0.161$ minutes.

5. Analyzing the Dialogue Features

We run a series of linear mixed effects models (LMEMs), using the *lme4* [33] package in R [34], to analyze the effect of mood on turn-taking in clinical dialogues. We only consider clinical interviews with female clinicians in this analysis (~82% of the interviews). This obviates the need for considering three-way interactions between the predictors (i.e., $mood \times gender_{patient} \times gender_{clinician}$). We set each of the dialogue features as a response variable in our linear models. Depending on the task, we set the binary mood state $\{euthymic, depressed\}$ or $\{euthymic, manic\}$ as a fixed effect test variable. We set the gender of the patient as a random effect control variable. Finally, we set random intercepts for the patients and the clinicians. We use likelihood ratio tests to test for statistical significance. For each dialogue feature, we test a full model (with the mood fixed effect) against a null model (without the mood fixed effect). In the case of a significant interaction between mood state and patient gender, we report the p-values for the interaction effect since those from the main effect are not interpretable [35].

Our final data that we use for this analysis contain 525 clinical interviews with a total duration of around 130 hours from 46 unique patients (33 females) and 5 unique clinicians (all females).

5.1. Results and Discussion

We report the results of the LMEM analysis in Table 1.

5.1.1. Depression

We find that call duration goes up by an average of 0.578 ± 0.052 minutes when patients are in a depressed episode (compared to a ehuthymic episode). Call duration did not produce a significant interaction effects with the gender effect, showing that the increase in duration is consistent across male and female patients.

Patient Features. Floor control of patients significantly goes up by $2.657\% \pm 1.173\%$ when patients are depressed. We also find that both the mean and variability of the number of continuous turns go up by an average of 0.072 ± 0.020 and 0.104 ± 0.041 respectively. Additionally, the variability in turn switch offset for the patients goes up by 63.812 ± 20.925 milliseconds when patients are depressed. None of the patient features demonstrated statistically significant interactions with gender. The results suggest that depressed patients are more likely to insert longer pauses (>500 milliseconds) while speaking. Additionally, the results suggest that depressed patients exhibit higher variability in the time they take to respond to questions by clinicians.

Clinician Features. We find that both the mean and variability of turn hold offsets in clinicians go up by 128.150 ± 30.200 milliseconds and 200.262 ± 39.575 milliseconds, respectively. We find that the number of continuous turns for clinicians goes up by an average of 0.054 ± 0.022. Finally, we find that the average turn length of clinicians goes down by 97.525 ± 41.600 milliseconds. There were no significant interaction effects. These results suggest that although clinicians

insert longer silences between their turns while interviewing depressed patients, they tend to speak for a slightly shorter time.

5.1.2. Mania

We find that call duration goes up by an average of 0.468 ± 0.073 minutes when patients are in a manic episode (compared to a ehuthymic episode). Call duration produced significant interaction with gender ($p < 0.0005$), indicating that the increase in call duration is mainly driven by male patients ($+0.631 \pm 0.161$ minutes).

Patient Features. We find that floor control of patients significantly goes up by an average of $8.276\% \pm 1.981\%$ when patients are manic. We find that both the mean and variability of turn lengths in patients go up by 313.163 ± 84.050 milliseconds and 299.200 ± 117.013 milliseconds, respectively. None of the patient features produced significant interactions with their gender, meaning that both male and female patients speak longer relative to clinicians when patients are manic.

Clinician Features. We find that both the mean and variation in clinicians' turn lengths go down by 215.300 ± 70.100 and 175.863 ± 61.812 milliseconds, respectively. This finding was consistent for clinicians interviewing both male and female patients.

6. Predicting Mood Episodes

The previous section demonstrated that symptom severity has a significant impact on the dynamics of interaction as captured by our dialogue features. In this section, we assess whether mood symptom severity prediction can be improved by integrating dialogue features with standard prosodic features. We define two classification tasks, similar to [8]: (1) discriminate between episodes of euthymia and depression and (2) discriminate between episodes of euthymia and mania. We only include patients in this analysis if they have at least two euthymic calls and two manic/depressed assessment calls.

We study the efficacy of dialogue features using three classifiers: logistic regression, support vector machines (SVM), and deep neural networks (DNN). We train each classifier with rhythm features, dialogue features, and their combination via early fusion. We follow a leave-one-speaker-out evaluation scheme, and report the average area under the receiver operating characteristic curve (AUROC) across all test speakers. For each test speaker, we run a five-fold cross-validation over the training speakers to pick optimal hyper-parameters. We build and train our classifiers using the Scikit-learn library [36]. We optimize for the following hyper-parameters: Logistic regression: $\{C$: [0.001, 0.01, . . . , 1000]$\}$; SVM: $\{kernel$: $[rbf]$, C: [0.001, 0.01, . . . , 1000], γ: [0.0001, 0.001, . . . , 100]$\}$; DNN: {activation: $[relu]$, number of layers: [2, 3], layer width: [32, 64], batch size: [64], learning rate: [0.001], }. We train the DNNs with the log-loss function for a total of 10 epochs using the ADAM optimizer. To reduce variance due to random initialization, we train DNNs with 10 different random seeds and report the average of the runs. We scale the input features using the maximum values from the training speakers for each test fold before feeding the features into the classifiers.

For every test speaker, we run feature selection on the training speakers using likelihood ratio tests to determine whether individual features are significantly affected by mood. We retain features if the resulting p-value is less than 0.05. We found that the total call duration was highly predictive of clinical outcomes. As a result, we do not include it as a feature in our analysis to focus our study on dialogue features that capture local interaction dynamics of the interviews.

Table 2: *Detecting mania/depression from clinical calls using three classifiers. Results shown are average AUROCs across all test speakers. Early fusion was used to combine rhythm and dialogue features.*

Classifier	Features	AUROC Depressed	AUROC Manic
LR	rhythm	0.739	0.658
	dialogue	0.634	0.606
	both	0.761	0.641
SVM	rhythm	0.724	0.650
	dialogue	0.618	0.547
	both	0.764	0.630
DNN	rhythm	0.729	0.676
	dialogue	0.617	0.618
	both	0.761	0.651

6.0.1. Results and Discussion

We summarize the results for the classification tasks in Table 2. Consistent with previous work [8], our results show that rhythm features are effective for detecting both depression and mania. When using rhythm features alone, we obtain a maximum AUROC of 0.739 when predicting depression using a logistic regression classifier and a maximum AUROC of 0.676 when predicting mania using a DNN classifier.

When using dialogue features alone, we obtain a maximum AUROC of 0.634 when predicting depression using a logistic regression classifier and a maximum AUROC of 0.618 when predicting mania using a DNN classifier.

Next, we study how combining the two feature sets, via early fusion, affects the performance of predicting depression and mania. When fusing rhythm and dialogue features, we obtain a maximum AUROC of 0.764 when predicting depression using a SVM classifier and a maximum AUROC of 0.651 when predicting mania using a DNN classifier. Augmenting rhythm features with dialogue features resulted in improved AUROCs for all three classifiers when detecting depression, suggesting that interaction patterns are complementary to speech rhythm patterns. This pattern was not true for detecting mania, however. For detecting mania, we found that none of the classifiers was able to make use of the additional dialogue features to get improved AUROCs over using rhythm features alone. This could be due to the fact that the relatively small number of manic episodes in our dataset makes it hard for the trained classifiers to generalize to unseen test speakers.

7. Conclusion

In this work we showed that high-level dialogue features can be used to quantify interaction dynamics in clinical interviews, highlighting how changes in mood episodes can significantly affect the values of the features. Additionally, we showed that dialogue features can be used to augment prosodic features to improve automatic detection of depression severity in clinical interviews. For future work, we plan consider building gender-specific models for mood prediction, as those have shown promise in emotion recognition tasks [37].

8. References

[1] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters, "Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication," *Archives of general psychiatry*, vol. 62, no. 6, pp. 593–602, 2005.

[2] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, 2013.

[3] Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell, "Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs," in *Semdial 2013 DialDam: Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*, 2013, pp. 160–169.

[4] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech*, 2018, pp. 1716–1720.

[5] 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.

[6] 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.

[7] 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.

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

[9] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[10] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[11] K. G. Niederhoffer and J. W. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 337–360, 2002.

[12] M. Nasir, B. Baucom, S. Narayanan, and P. Georgiou, "Towards an unsupervised entrainment distance in conversational speech using deep neural networks," *arXiv preprint arXiv:1804.08782*, 2018.

[13] B. Xiao, Z. E. Imel, D. C. Atkins, P. G. Georgiou, and S. S. Narayanan, "Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] J. A. Hall, J. A. Harrigan, and R. Rosenthal, "Nonverbal behavior in clinicianpatient interaction," *Applied and Preventive Psychology*, vol. 4, no. 1, pp. 21–37, 1995.

[15] 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.

[16] N. W. Hashim, M. Wilkes, R. Salomon, J. Meggs, and D. J. France, "Evaluation of voice acoustics as predictors of clinical depression scores," *Journal of Voice*, vol. 31, no. 2, pp. 256–e1, 2017.

[17] S. Scherer, Z. Hammal, Y. Yang, L.-P. Morency, and J. F. Cohn, "Dyadic behavior analysis in depression severity assessment interviews," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 112–119.

[18] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. F. Cohn, "Multimodal detection of depression in clinical interviews," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 307–310.

[19] M. Hamilton, "Development of a rating scale for primary depressive illness," *British journal of social and clinical psychology*, vol. 6, no. 4, pp. 278–296, 1967.

[20] 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.

[21] 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.

[22] H. Arsikere, E. Shriberg, and U. Ozertem, "Enhanced end-of-turn detection for speech to a personal assistant," in *2015 AAAI Spring symposium series*, 2015.

[23] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[24] F. Cassidy, E. Murry, K. Forest, and B. J. Carroll, "Signs and symptoms of mania in pure and mixed episodes," *Journal of affective disorders*, vol. 50, no. 2-3, pp. 187–201, 1998.

[25] T. Meshorer and P. A. Heeman, "Using past speaker behavior to better predict turn transitions." in *Interspeech*, 2016, pp. 2900–2904.

[26] Š. Beňuš, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[27] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in psychology*, vol. 6, p. 731, 2015.

[28] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[29] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[30] S. H. Yella and H. Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1688–1700, 2014.

[31] P. A. Heeman and R. Lunsford, "Turn-taking offsets and dialogue context." in *Interspeech*, 2017, pp. 1671–1675.

[32] S. Tilsen and A. Arvaniti, "Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 628–639, 2013.

[33] D. Bates, D. Sarkar, M. D. Bates, and L. Matrix, "The lme4 package," *R package version*, vol. 2, no. 1, p. 74, 2007.

[34] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[35] J. Zar, "Biostatistical analysis 4th ed," *New Jersey*, 1999.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[37] B. Zhang, E. M. Provost, and G. Essi, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5805–5809.