The PRIORI Emotion Dataset:
Linking Mood to Emotion Detected In-the-Wild

Soheil Khorramp, Mimansa Jaiswal, John Gideon, Melvin McInnis, Emily Mower Provost

Departments of: Computer Science and Engineering and Psychiatry, University of Michigan
{khorrrams, mimansa, gideonjn, mmcinnis, emilykmp}@umich.edu

Abstract

Bipolar Disorder is a chronic psychiatric illness characterized by pathological mood swings associated with severe disruptions in emotion regulation. Clinical monitoring of mood is key to the care of these dynamic and incapacitating mood states. Frequent and detailed monitoring improves clinical sensitivity to detect mood state changes, but typically requires costly and limited resources. Speech characteristics change during both depressed and manic states, suggesting automatic methods applied to the speech signal can be effectively used to monitor mood state changes. However, speech is modulated by many factors, which renders mood state prediction challenging. We hypothesize that emotion can be used as an intermediary step to improve mood state prediction. This paper presents critical steps in developing this pipeline, including (1) a new in the wild emotion dataset, the PRIORI Emotion Dataset, collected from everyday smartphone conversational speech recordings, (2) activation/valence emotion recognition baselines on this dataset (PCC of 0.71 and 0.41, respectively), and (3) significant correlation between predicted emotion and mood state for individuals with bipolar disorder. This provides evidence and a working baseline for the use of emotion as a meta-feature for mood state monitoring.

Index Terms: Emotion Dataset, Emotion in the Wild, Emotion Recognition, Mood Prediction

1. Introduction

Bipolar disorder (BD) is a severe, chronic mental illness that typically begins in early adulthood and is characterized by periodic and pathological mood changes ranging from extreme lows (depression) to extreme highs (mania) [1]. It is found in 1% of the world population, with a core clinical expression pattern related to emotion, energy, and psychomotor activity [2]. These core clinical signs and symptoms are monitored to gauge the health and progress of the individual in treatment [1]. The dynamic nature of BD demands efficient clinical monitoring to detect mood changes in sufficient time to treat or mitigate their severity. Intense clinical monitoring is effective but unrealistic due to cost and the availability of skilled health care providers [3]. Automatic passive mood monitoring addresses the need for ongoing monitoring in a cost efficient manner to predict the course and outcome of chronic human disease such as BD.

Current mood recognition systems are focused on mapping between speech to mood directly, which is challenging due to the complexity of the speech signal. We hypothesize that emotion can simplify mood prediction by acting as an intermediary between speech (rapidly varying) and mood (slowly varying). Further, one of the hallmark symptoms of BD is emotion dysregulation, suggesting that the tracking of emotion changes will provide important insights into an individual’s mood variation. In this paper, we define emotion in terms of valence (positive vs. negative) and activation (calm vs. excited), both of which are observable from expressed behaviors such as speech.

Strategies for mobile monitoring for mental health have mostly relied upon self-reported diagnosis of a disorder on a device or social media to identify features indicative of the disorder [4], be it anxiety [5], BD [6], depression or anorexia [7]. However, these interactive self-reports are often incomplete or misleading [8]. An alternative approach is to directly recognize mood from observed behavior. Vanello et al. [9] and Faurholt et al. [10] investigated how speech features could be used to characterize mood states. Muaremi et al. in [11] used statistics of phone calls, such as duration and frequency, to predict mood episodes. Our own work has demonstrated that properties such as speaking rate are also effective for detecting mood [12, 13]. However, a recurring theme in these studies is the challenge associated with detecting mood directly from speech, due in part to the highly varying nature of the speech signal. We hypothesize that we will be able to improve mood detection by using an intermediary (emotion) that has more slowly varying properties.

The relationship between emotion and mood has been gaining attention. Stasak et al. investigated the utility of using emotion to detect depressed speech [14], using the AVEC 2014 dataset [15]. However, these data were collected in a controlled environment, potentially limiting their use “in the wild”. Carrillo et al. identified a relationship between emotional intensity and mood in the context of BD [16]. However, they relied upon transcribed interviews, rather than on acoustics directly.

The work presented in this paper leverages the PRIORI dataset, a longitudinal dataset of natural speech patterns from individuals with BD [17], which is itself a component of a longitudinal study of BD [18]. Data were collected from individuals with BD for six to twelve months using smartphones with a secure app that recorded their side of all phone conversations. They were assessed weekly for depressive and manic symptoms using standardized scales by a study clinician [12]. We analyze a subset of this dataset, referred to as the PRIORI Emotion Dataset, which is annotated with labels of valence and activation. This provides an opportunity to associate natural expressions of emotion with changes in mood.

In this paper, we address: (1) the predictability of emotion in natural smartphone conversations and (2) the relationship between mood and natural expressions of emotion. We describe the collection of the full PRIORI dataset, as well as the creation of the PRIORI Emotion Dataset. We establish natural speech emotion classification baselines on this dataset, which achieve a Pearson correlation coefficient (PCC) of 0.71 and 0.41 for detecting activation and valence, respectively. Finally, we demonstrate that there is a significant positive correlation between heightened mood and both activation and valence. Critically, we note that these emotion patterns are inherently subject-dependent, highlighting the importance of attuning to individual variability when designing mental health monitoring support.
### Table 1: Distribution of mood states. Shown are the binned HAMD and YMRS ranges, the number of assessments, and the mean and standard deviation of subject observations.

<table>
<thead>
<tr>
<th>Mood</th>
<th>HAMD</th>
<th>YMRS</th>
<th>Number</th>
<th># Per Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euthymic</td>
<td>≤ 6</td>
<td>≤ 6</td>
<td>70</td>
<td>5.8 ± 3.4</td>
</tr>
<tr>
<td>Manic</td>
<td>&lt;10</td>
<td>≤ 10</td>
<td>27</td>
<td>2.7 ± 1.9</td>
</tr>
<tr>
<td>Depressed</td>
<td>≥10</td>
<td>&lt;10</td>
<td>120</td>
<td>10.0 ± 6.4</td>
</tr>
<tr>
<td>Excluded</td>
<td>Else</td>
<td>Else</td>
<td>96</td>
<td>8.7 ± 6.7</td>
</tr>
</tbody>
</table>

2. PRIORI Dataset

The PRIORI dataset is composed of one-sided natural conversations recorded during daily smartphone usage (Samsung Galaxy S3, S4, S5) [17, 12, 13]. The participants include 51 BD and nine healthy controls. The inclusion criteria were: BD type I or II, no medical or neurological disease, and no active history of substance abuse. All study participants were provided with a smartphone and were asked to use the smartphone as their primary device. The app runs silently in the background, recording the speech with 8 kHz sampling frequency, and then uploads the recordings to our servers for analysis. Participants were enrolled for an average of 32 ± 16 weeks. The collection includes 52,931 calls and over 4,014 hours of speech.

Participants were clinically evaluated in weekly assessment calls to assess the level of depression (Hamilton Depression Scale (HAMD) [19]) and mania (Young Mania Rating Scale (YMRS) [20]). All other recordings are referred to as personal calls. We assign mood labels to all assessment calls based on the HAMD and YMRS scales. We bin the continuous ratings into categories euthymic, manic, depressed, and excluded based on thresholds applied to the HAMD and YMRS scores. A call is labelled *euthymic* if it has a score of six or less on both the HAMD and YMRS scales; *manic* if the score is ten or greater on the YMRS and less than ten on the HAMD; and *depressed* if the score is ten or greater on the HAMD and less than ten on the YMRS. All other assessments are excluded from this paper’s experiments (see Table 1).

### 3. PRIORI Emotion Annotation

This work explores the relationship between mood state and emotion expression, which necessitates access to a labeled corpus over which emotion can be detected and emotion classification algorithms can be validated. However, there are no natural smartphone conversational speech datasets annotated in this manner. We addressed this limitation by generating the PRIORI Emotion Dataset, a subset of the larger PRIORI dataset. The PRIORI Emotion Dataset contains manual valence/activation annotations of both assessment and personal calls. We use a dimensional labeling strategy in this work [21], motivated by the concept of *core affect* [22]. This construct provides a de-contextualized manner of considering emotion expression.

We selected a subset of twelve PRIORI subjects, based on three factors: (1) BD diagnosis: our goal is to examine the link between emotion and bipolar mood (future work will focus on the healthy controls). (2) Used Samsung S5: This provided microphone consistency, lack of which was identified as a challenge identified in our prior work [12]. (3) Provided informed consent for annotation of personal calls, which provides the basis for ground-truth emotion labels. This subset includes a total of 11,337 calls, encompassing over 928 hours of natural conversational speech.

The emotion labeling process has four steps: (1) segmentation, (2) segment selection, (3) segment inspection, and (4) segment annotation. Each is explained in the following sections.

3.1. Segmentation

We first filter the set of calls to exclude all recordings longer than one hour. This restriction is due to the large memory requirements and processing time associated with these data. We then perform speech activity detection (SAD), using the COMBO-SAD algorithm introduced by Sadjadi and Hansen [23]. We form contiguous segments following the methodology used in our prior work [12]. The resulting segments contain continuous speech with no intermediate silence. This procedure provides 167,339 segments with the average length of 6.32 ± 5.89 seconds.

3.2. Segment Selection

We identified a subset of 17,237 segments for manual annotation. Our first filter was for segment length, to increase the likelihood that segments contained sufficient data to assess, but were not so long that the emotion would vary over the course of the segment. Next, we sampled from both personal calls and assessment calls to ensure a diversity of examples. Finally, segments were chosen whose mood could be inferred, prioritizing segments that were more likely to correspond to euthymic, manic, or depressed mood states. The challenge is that the only mood labels available are those defined by the weekly assessment calls. Therefore, we sampled personal calls as a function of proximity in time to assessment calls, preferring those that occurred closer to the assessments. Specifically, our selection process was as follows: (1) filter on segment length, excluding segments shorter than three seconds and longer than 30 seconds, (2) for each assessment call select up to ten random segments. For each subject’s personal calls, select 1,200 segments randomly considering the weight of $\max(4 - d, 1)$, where $d$ is the number of days between the call and its future assessment day. Calls on the day of assessment receive a weight of four (these calls are most closely linked to the HAMD/YMRS score). Other calls receive a weight that reduces linearly up to 3 days before assessment day, giving further away calls a weight of one. This procedure selects 2,837 and 14,400 segments from assessment and personal calls respectively.

3.3. Segment Inspection

Each segment was manually inspected before the annotation process and we removed those that were not deemed appropriate for the annotation task. We removed segments when: (1) background noise dominates speech signal, (2) the segment contains less than two seconds of speech, (3) the subject is not talking to the phone (i.e. talking to someone else in the room), (4) when the emotion clearly varies over the course of the segment, and (5) when the segment contains identifiable information such as name, address, phone number, etc. After this step, the dataset contains 2,209 assessment and 11,402 personal segments (in total 13,611 segments, 25.20 hours of speech).

3.4. Segment Annotation

We annotated the activation and valence of the selected speech segments using a 9-level discrete representation (1: very low, 5: neutral, 9: very high), using the established pictorial manikins method [21]. There were 11 annotators (7 female, 4 male) aged between 21 and 34. All annotators were students of the University of Michigan and native English speakers.

We conducted a training session for each annotator, including a training video and manuscript, to introduce the annotation software and provide annotation examples. In the training ses-
1. Although challenging, we asked annotators to only consider the acoustic characteristics of the recordings, not the content of them. They were asked to avoid letting speech content “color” their activation and valence labels.

2. We asked that annotators consider the subject-specificity of emotion expression. When approaching a new subject, annotators were asked to spend some time listening to a few segments without assigning a rating in order to get a better sense of what that person’s baseline sounds like.

We further supported the assessment of subject-dependent emotion patterns by providing individual context for each participant. Annotators rated segments over a single participant at a time. The software randomly selected a patient and randomly presented all segments of that patient to the annotator before moving on to the next patient’s segments.

We collected between two and six labels for each segment (3.83 ± 1.31 labels per segment). Figure 1 shows the distribution of the number of annotations for each segment. See Figure 1 for a distribution of the activation and valence labels defined by the annotators. We found that the activation and valence values are significantly correlated with a PCC of 0.46 (p < 0.01).

4. Methods

We describe two emotion prediction systems used in this work. The first system is a deep feed-forward neural network (FFNN) that operates on the eGeMAPS feature set [24]. The second system is a convolutional neural network (CNN) classifier that operates on log-Mel-frequency bank (log-MFB) features [25].

4.1. Acoustic Features

eGeMAPS – The eGeMAPS feature set is carefully designed to standardize acoustic features that are effective in affective computing. It is an 88-dimensional feature vector that includes features relating to energy, excitation, spectral, cepstral, and dynamic information. We extract the eGeMAPS features using the openSMILE toolkit with default parameters [26].

Log-MFB – Previous research has demonstrated that log mel-frequency bank (log-MFB) spectral features outperform other temporal frame-level acoustic features such as MFCCs [25]. We extract 40-dimensional log-MFB features with 25ms frame length and 10ms frame shift using the Kaldi toolkit [27]. We perform global ε-normalization on all features.

Table 2: The results of the baseline emotion systems. A bolded result indicates significantly better performance than the other method, using a repeated cross-validation test with p < 0.01.

### 4.2. Models

**Deep FFNN** – This network contains a stack of fully-connected dense layers with tanh activation functions followed by an output layer with a linear activation function. We predict activation and valence values independently.

**Conv-Pool** – We implemented Conv-Pool network, proposed in [28], due to its high emotion recognition accuracy on the IEMOCAP [29] and MSP-IMPROV [30] datasets. The Conv-Pool network contains three major components: (1) a stack of convolutional layers; (2) a global pooling over time layer; and (3) a stack of dense layers. The convolutional layers create a sequence of feature maps that determine emotionally salient regions within variable length utterances. The global pooling layer automatically extracts a set of call-level statistics. Aldeneh et al. [28] found that a max-pooling layer is effective for emotion recognition. Finally, the stack of dense layers predicts the labels from the call-level features. We used ReLU and linear activation functions for intermediate and output layers.

5. Emotion Detection Baselines

We normalized the labels by subtracting the rating midpoint of 5 and scaling to the range of [−1, 1]. Let x be a rating, the normalization was performed through $\frac{x - 5}{2}$. Our preliminary analyses showed that this transformation helped the networks to learn the bias and standard deviation of the labels more quickly. All annotations were averaged to produce a single segment rating.

We assess the performance of the methods using repeated cross-validation method introduced in [31]. We repeat each experiment for five total runs, where a run is defined as six randomly selected folds. In each run, the folds are shuffled by randomly assigning two subjects to each of the six folds. We then use round-robin cross-validation. At each step, one fold (two subjects) is assigned to testing, one is used for tuning parameters and early stopping, and the rest are used for training. This round-robin procedure generates one test measure per fold, resulting in six measures. Over the course of the five runs, a matrix of 6-by-5 test measures is output. We report the mean over all experiments as the experiment mean. The experiment standard deviation is the mean standard deviation within runs. Significance was determined using a repeated cross-validation paired t-test with six degrees of freedom, as shown in [31].

We implemented both networks using Keras [32] with TensorFlow backend [33]. We optimized RMSE during training through the Adam optimizer [32] with a fixed learning rate of 0.0001. All weights were initialized using the Xavier uniform algorithm [32] and all bias parameters were set to zero. We set epoch size to 64. To train FFNN, we performed cross-validation to tune the number of dense layers (2,4,8) and the number of channels (200,400,800). To train Conv-Pool network, we set the number of initial convolutional layers and final dense layers equal and validated them over a set of (2,3). The number
of channels (200,400) and the length of the convolution kernels (4,8) were also validated. We trained FFNN and Conv-Pool networks for 100 and 15 epochs respectively and selected the best epoch based on the validation performance.

We use three popular metrics to compare the resulting networks: (1) PCC, (2) concordance correlation coefficient (CCC), and (3) root mean square error (RMSE). Using Conv-Pool, we achieved a PCC of 0.712 and 0.405 for activation and valence respectively, and found that Conv-Pool has significantly better performance than FFNN using all measures except RMSE (Table 2). This supports previous work that demonstrated the importance of the temporal structure of speech to characterize emotion [28]. We hypothesize that the lack of improvement for RMSE is due to the fact that RMSE places more weight on selecting the correct bias of the ratings. Subject-dependent or speaker-adapted models may provide improve RMSE. Finally, as shown in previous emotion datasets, activation is easier to predict from speech than valence [34].

6. Mood Analysis

In this section, the link between BD mood states and predicted emotion is tested. To facilitate this analysis, we use our Conv-Pool models to predict emotion on the 10,563 assessment call segments. We use the 30 different models from the repeated cross-validation as an ensemble and take the mean output.

We normalized the predicted emotion labels using subject-dependent euthymic z-normalization. Our preliminary analyses demonstrated the importance of considering how a subject varies about his/her own baseline, which is defined as his/her euthymic periods. We calculate the mean and standard deviation of the valence/activation ratings over all calls associated with euthymic mood states. We then normalize each segment based on these values, reducing the effect of subject biases.

Table 3 shows the mean activation and valence ratings, calculated over the segments in each of the different mood states. Table 4 shows the PCC between each of the dimensional emotion ratings (activation and valence) and each of the mood ratings (YMRS and HAMD). We note that:

- In the majority of subjects, the mean of both emotion ratings during manic states is more positive and activated compared to the corresponding within-subject ratings during depressed states. Significance was determined by a t-test with p < 0.01. This provides evidence that emotion behavior may be effective for predicting mood states.
- For almost all subjects, activation and valence are significantly positively correlated with YMRS and significantly negatively correlated with HAMD (p < 0.01). This supports the hypothesis that heightened mood states come with heightened emotions.
- Even after normalizing each subject by his/her euthymic segments, the distribution of emotion ratings between subjects is significantly different (using a one-way ANOVA with p < 0.01). We also used a Tukey-Kramer posthoc test of the 66 possible pairwise subject comparisons. Activation was found to be significantly different in 51 cases and valence was found to be significantly different in 48 cases (p < 0.01).
- Our experiments did not show a correlation between the within-call variance of emotion ratings and mood states.

7. Conclusion

In this work, we present the PRIORI Emotion Dataset - a natural dataset of emotional speech passively-recorded from patients with BD. The dataset is unique in that it has a high proportion of emotional segments of speech and is the only in the wild telephonic dataset, annotated for emotion. The dataset contains more than 23 hours of speech (13,617 segments) with the average of 3.83 labels per segment. We train a CNN model and show that it is possible to accurately extract activation ratings from unstructured speech. We achieve a PCC of 0.712 and 0.405 for activation and valence, respectively. We perform exploratory analysis to show how these predicted emotion labels, normalized using subject’s euthymic baseline, correlate to YMRS and HAMD values. We find that mean of both emotion ratings during manic states is significantly higher than the mean of the corresponding ratings during depressed states.

The annotation process is ongoing. Future work includes implementing subject-dependent models for emotion recognition so that we can directly predict the ratings instead of normalizing them post-hoc. This would help in building subject-dependent mood recognition.

8. Acknowledgements

Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. This work was supported by the National Science Foundation (CAREER-1651740). NIMH R34MH100404, the Heinz C Prechter Bipolar Research Fund and the Richard Tam Foundation at the University of Michigan.
9. References


