Mood State Prediction
From Speech Of Varying Acoustic Quality
For Individuals With Bipolar Disorder

John Gideon¹, Emily Mower Provost¹, and Melvin McInnis²

Departments of: Computer Science and Engineering¹ and Psychiatry², University of Michigan
Overview

**Bipolar disorder**
Pathological mood-state swings of mania and depression
A leading cause of disability – 4% of Americans affected

**Current Treatment**
Periodic follow-up visits for monitoring
Reactively after manic/depressive episodes

**Clinical Need**
To passively detect & predict mood and health state changes in order to intervene and prevent episodes

---

National Institute of Mental Health, “Bipolar Disorder in Adults.”
Kessler et al., “Lifetime Prevalence and Age-of-onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication.”
Angst et al., “Long-term Outcome and Mortality of Treated Versus Untreated Bipolar and Depressed Patients: A Preliminary Report.”
Problem Statement

- **Speech** patterns shown to **reflect mood** in clinic
  - Controlled environments
  - Single type of **recording device**
- Real world recordings
  - Variations in **background noise**
  - Variations in **microphone quality**

Speech recorded in the **real world** has **large variations in quality** making a **distributed** mobile health system using speech **infeasible** without controlling for these differences.

Hamilton, “Hamilton Depression Scale.”
Young et al., “A Rating Scale For Mania: Reliability, Validity And Sensitivity.”
UM PRIORI Acoustic Database

- **Participants**: 37 subjects enrolled for 6-12 months
- **Total Data**: 2,400 hours across 30,000 calls
- **Ground Truth**: 780 Recorded weekly phone-based clinical assessments (About 15 minutes each)
  - Structured clinical interview
  - Rated on mania and depression severity
    - Young Mania Rating Scale (**YMRS**)
    - Hamilton Rating Scale for Depression (**HAMD**)
  - 23 assessments transcribed for validating segmentation
  - Only used assessment calls in this analysis

Feelings of guilt? Insomnia? Anxiety? Weight loss?

Assessment Call Audio → ? → Assessment Mood

Hamilton, “Hamilton Depression Scale.”
Young et al., “A Rating Scale For Mania: Reliability, Validity And Sensitivity.”
Mood Label Assignment

- **Euthymic (30%)**
- **Depressed (28%)**
- **Manic (12%)**

Occurrence of Label Combinations

- YMRS Mania Score
- HAMD Depression Score
## Models of Phones

<table>
<thead>
<tr>
<th>Samsung Galaxy S3</th>
<th>Samsung Galaxy S5</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="" alt="Samsung Galaxy S3" /></td>
<td><img src="" alt="Samsung Galaxy S5" /></td>
</tr>
<tr>
<td>18 Participants</td>
<td>17 Participants</td>
</tr>
<tr>
<td>456 Assessments</td>
<td>287 Assessments</td>
</tr>
</tbody>
</table>
Acoustic Differences Between Models

| Galaxy S3 audio versus S5 | Over 100 times as much Clipping | Over 6 times as loud (RMS) | 3.9dB drop in estimated SNR |

![Waveform Comparison](image)
Processing Pipeline – Preprocessing

Galaxy S3 audio versus S5

- Over 100 times as much Clipping
- Over 6 times as loud (RMS)
- 3.9dB drop in estimated SNR

Audio Signal → Declipping (RBAR) → Audio Normalization → Noise-Robust Segmentation → SVM Classification → Mood Prediction

- Preprocessing
- Feature Extraction
- Data Modeling

7 Rhythm Features → 31 Call-Level Statistics → Feature Normalization
Declipping Method

- **CBAR**
  - Extrapolates clipped regions
  - Minimizes pointiness (acceleration)

CBAR *(Harvilla and Stern, 2014)*
Declipping Method

• **CBAR**
  – Extrapolates clipped regions
  – Minimizes pointiness (acceleration)

• **RBAR**
  – Fast approximation to CBAR
  – Used in preprocessing pipeline

CBAR *(Harvilla and Stern, 2014)*

Before

After RBAR
Noise-Robust Segmentation

5 Sources of Speech Activity (Sadjadi and Hansen, 2013)

- Harmonicity
- Clarity
- Prediction Gain
- Periodicity
- Perceptual Spectral Flux

Combine with PCA keeping largest $\lambda$

25ms Hanning Window

Normalize by 5th Percentile and Std.

Segmentation Signal

Noise-Robust Segmentation (Cont.)

- Validation used to determine segments
  - Exceeds a **threshold of 1.8**
  - Minimum silence of 0.7 seconds
- Only include segments longer than two seconds
  - **Subsegment** into two seconds with one second overlap
  - Necessary for feature extraction

Audio:

Segments: 1.5 sec.  4.25 seconds

Subsegments: 2 sec.  2 sec.  2 seconds
Processing Pipeline – Feature Extraction

Audio Signal → Declipping (RBAR) → Audio Normalization → Noise-Robust Segmentation

7 Rhythm Features → 31 Call-Level Statistics → Feature Normalization

SVM Classification → Mood Prediction

Preprocessing
Feature Extraction
Data Modeling
Rhythm Features

• Both mania and depression have rhythm related symptoms
  – **Mania:** Speech is more frequent, quicker, and louder
  – **Depression:** Slowing of speech and difficulty articulating

• Uses constant **two second segments**
  – Extract audio envelope
  – Extract seven statistics of syllable vs supra-syllable rhythm
  – Calculate **31 statistics** over segments for call-level features

• Normalize either **globally** or by **subject**

Tilsen and Arvaniti. "Speech Rhythm Analysis With Decomposition Of The Amplitude Envelope: Characterizing Rhythmic Patterns Within And Across Languages."
Processing Pipeline – Data Modeling

Audio Signal → Declipping (RBAR) → Audio Normalization → Noise-Robust Segmentation → SVM Classification → Mood Prediction

- Preprocessing
- Feature Extraction
- Data Modeling

7 Rhythm Features → 31 Call-Level Statistics → Feature Normalization
Data Partitioning

• Binary cases considered
  – Euthymic vs. manic
  – Euthymic vs. depressed
• Used participant-independent testing
• Participants have at least six calls
  – At least two euthymic
  – At least two manic and/or depressed

<table>
<thead>
<tr>
<th>Model</th>
<th># Subjects for Mania Test</th>
<th># Subjects for Depressed Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>S5</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Both</td>
<td>15</td>
<td>18</td>
</tr>
</tbody>
</table>
Validation, Training, and Testing

- Use **participant-independent validation**
  - Calculate **weighted information gain** and rank features
- Certain experiments use a **Multi-Task SVM**
  - Phone device (S3/S5) is second task
  - Weight kernel function based on device
- Performance measure: **Area Under the Receiver Operating Characteristic Curve (AUC / AUROC)**
Results – Declipping, Normalization, and Multitask

<table>
<thead>
<tr>
<th>Pipeline Test</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.57 ± 0.25</td>
<td>0.64 ± 0.14</td>
</tr>
<tr>
<td>Declipped Using RBAR</td>
<td>*<em>0.70 ± 0.17</em></td>
<td>0.65 ± 0.15</td>
</tr>
<tr>
<td>Normalized By Subject</td>
<td>*<em>0.67 ± 0.19</em></td>
<td>*<em>0.75 ± 0.14</em></td>
</tr>
<tr>
<td>Multi-Task Using Baseline Preprocessing</td>
<td>0.68 ± 0.23*</td>
<td>0.66 ± 0.18</td>
</tr>
<tr>
<td>Multi-Task Using Best Preprocessing</td>
<td>*<em>0.72 ± 0.20</em></td>
<td>0.71 ± 0.15</td>
</tr>
</tbody>
</table>

- **Significantly improved manic performance**
  - S5: Significantly more clipping in manic vs. depressed calls
  - Hypothesis: Individuals speak more loudly in a manic state

- **Normalization by subject** significantly improves both

*Denotes results significantly better than baseline (paired t-test, p=0.05)
## Results – No Speech Segmentation

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.52 ± 0.22</td>
<td>0.66 ± 0.17</td>
</tr>
<tr>
<td>S5</td>
<td>0.78 ± 0.31</td>
<td>0.62 ± 0.09</td>
</tr>
<tr>
<td>Both</td>
<td>0.57 ± 0.25</td>
<td>0.64 ± 0.14</td>
</tr>
</tbody>
</table>

Baseline

- **Remove speech segmentation**
  - Divide all audio into two second segments with one second overlap
  - Silence is included in features
- **Accuracy significantly improves**
  - Hypothesis: Rhythm features *indirectly capturing information* about the assessment interview
  - Requirement: **Accurate segmentation to avoid misleading results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.73 ± 0.22</td>
<td>0.74 ± 0.10</td>
</tr>
<tr>
<td>S5</td>
<td>0.79 ± 0.37</td>
<td>0.80 ± 0.21</td>
</tr>
<tr>
<td>Both</td>
<td>*<em>0.74 ± 0.24</em></td>
<td>*<em>0.77 ± 0.15</em></td>
</tr>
</tbody>
</table>

*Denotes results significantly better than baseline (paired t-test, p=0.05)
Conclusion

• Results demonstrate ability to counter variations in recording device quality
  – Differences include clipping, loudness, and noise
  – Combination of preprocessing, feature extraction, and data modeling

• Significantly better than baseline
  – Manic: 0.57 ± 0.25 → 0.72 ± 0.20
  – Depressed: 0.64 ± 0.14 → 0.75 ± 0.14

• No comprehensive solution

• Techniques could also be used to increase subject comparability when performing analysis on personal calls
Thank you for listening!

Questions?
Speech for Mood Monitoring

• Computational Analysis of Speech
  – **Emotion Recognition:** Mower 2011, Schuller 2009
  – **Major Depression:** Mundt 2007, Cohn 2009, Trevino 2011, Quatieri 2012, Helfer 2013, Cummins 2013
  – **PTSD:** Sluis 2011, Broek 2011, Tsumatori 2011
  – **Autism:** Hoque 2009, Van Santen 2010, Bone 2012, Chaspari 2013

• Challenges to adoption of remote monitoring
  – Collected in lab or disruptive phone calls
  – Clinical setting: prompted speech, fixed text
Rhythm Features

• Uses constant **2 second segments**
  – Constant to ensure changes in features due to rhythm, not segment size
  – Provides enough syllables without too much variation

• Perform preprocessing to extract audio envelope (Tilsen, 2013)

• Find power spectra
  – High frequency
    • Syllables
  – Low frequency
    • Supra-syllables

\[ \text{SBR}_{3.25} = 0.17 \quad \text{centroid} = 4.8 \text{Hz} \]

\[ \text{SBR}_{3.25} = 0.93 \quad \text{centroid} = 3.8 \text{Hz} \]
Rhythm Features (Cont.)

- **Empirical mode decomposition**
  - Extracts the intrinsic mode functions (IMFs)
- Calculate **ratio of power** between IMF$_1$ and IMF$_2$
- Determine **instantaneous frequency** over the first two IMFs
  - Time derivative of phase
  - Calculate mean and std.
- Calculate **31 statistics** over segments for call-level features
  - Total Features: $31 \times 7 = 217$ **total features**
- Normalize either **globally** or by **subject**
Results – Declipping

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.52±0.22</td>
<td>0.66±0.17</td>
</tr>
<tr>
<td>S5</td>
<td>0.78±0.31</td>
<td>0.62±0.09</td>
</tr>
<tr>
<td>Both</td>
<td>0.57±0.25</td>
<td>0.64±0.14</td>
</tr>
</tbody>
</table>

Baseline

• **Galaxy S5s perform better** than S3s when considering mania
  – Higher quality recordings
  – Subject population could also be more homogeneous
• **Significantly improved manic performance**
  – Significantly more clipping in manic calls than depressed calls from the S5
  – We hypothesize this is due to individuals speaking louder in a manic state

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.68±0.16</td>
<td>0.62±0.14</td>
</tr>
<tr>
<td>S5</td>
<td>0.79±0.21</td>
<td>0.69±0.18</td>
</tr>
<tr>
<td>Both</td>
<td><strong>0.70±0.17</strong></td>
<td>0.65±0.15</td>
</tr>
</tbody>
</table>

Declipped Using RBAR

*Denotes results significantly better than baseline (paired t-test, p=0.05)
## Results – No Speech Segmentation

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.52±0.22</td>
<td>0.66±0.17</td>
</tr>
<tr>
<td>S5</td>
<td>0.78±0.31</td>
<td>0.62±0.09</td>
</tr>
<tr>
<td>Both</td>
<td>0.57±0.25</td>
<td>0.64±0.14</td>
</tr>
</tbody>
</table>

### Baseline

- **Segments were no longer found** using previous algorithm
  - All audio divided into 2 second segments with 1 second overlap
  - Results in much silence being captured
- Performs the **best of all tests**
  - Hypothesize this is actually caused by rhythm features *indirectly capturing information* about the assessment interview
  - Shows need for **accurate segmentation to avoid misleading results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.73±0.22</td>
<td>0.74±0.10</td>
</tr>
<tr>
<td>S5</td>
<td>0.79±0.37</td>
<td>0.80±0.21</td>
</tr>
<tr>
<td>Both</td>
<td>0.74±0.24*</td>
<td>0.77±0.15*</td>
</tr>
</tbody>
</table>

*Denotes results significantly better than baseline (paired t-test, p=0.05)
## Results – Normalization By Subject

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.52±0.22</td>
<td>0.66±0.17</td>
</tr>
<tr>
<td>S5</td>
<td>0.78±0.31</td>
<td>0.62±0.09</td>
</tr>
<tr>
<td>Both</td>
<td>0.57±0.25</td>
<td>0.64±0.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.66±0.15</td>
<td>0.73±0.15</td>
</tr>
<tr>
<td>S5</td>
<td>0.71±0.35</td>
<td>0.78±0.10</td>
</tr>
<tr>
<td>Both</td>
<td>0.67±0.19*</td>
<td>0.75±0.14*</td>
</tr>
</tbody>
</table>

**Baseline**

**Normalized By Subject**

- **Significant improvement for both mood tests**
- Previously shown to be able to correct for variations in feature distributions between speakers
  - Method also has ability to correct for phone models

*Denotes results significantly better than baseline (paired t-test, p=0.05)
### Results – Multi-Task Learning

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.52±0.22</td>
<td>0.66±0.17</td>
</tr>
<tr>
<td>S5</td>
<td>0.78±0.31</td>
<td>0.62±0.09</td>
</tr>
<tr>
<td>Both</td>
<td>0.57±0.25</td>
<td>0.64±0.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.67±0.20</td>
<td>0.67±0.21</td>
</tr>
<tr>
<td>S5</td>
<td>0.72±0.41</td>
<td>0.65±0.11</td>
</tr>
<tr>
<td>Both</td>
<td><strong>0.68±0.23</strong>*</td>
<td>0.66±0.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Manic AUC</th>
<th>Depressed AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.71±0.19</td>
<td>0.66±0.14</td>
</tr>
<tr>
<td>S5</td>
<td>0.78±0.23</td>
<td>0.79±0.13</td>
</tr>
<tr>
<td>Both</td>
<td><strong>0.72±0.20</strong>*</td>
<td>0.71±0.15</td>
</tr>
</tbody>
</table>

- **Significantly improves manic** test performance without any preprocessing modifications
- We hypothesize depressed tests are less affected due to being more comparable before preprocessing
- Best manic performance when using all techniques

*Denotes results significantly better than baseline (paired t-test, p=0.05)