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

John Gideon\textsuperscript{1}, Emily Mower Provost\textsuperscript{1}, and Melvin McInnis\textsuperscript{2}

Departments of: Computer Science and Engineering\textsuperscript{1} and Psychiatry\textsuperscript{2}, 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

Costly Consequences

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

Occurrence of Label Combinations

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

- 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>

Images: Samsung.com
## Acoustic Differences Between Models

<table>
<thead>
<tr>
<th>Galaxy S3 audio versus S5</th>
<th>Over 100 times as much Clipping</th>
<th>Over 6 times as loud (RMS)</th>
<th>3.9dB drop in estimated SNR</th>
</tr>
</thead>
</table>

S3

S5
Galaxy S3 audio versus S5:

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

Processing Pipeline – Preprocessing:

1. Declipping (RBAR)
2. Audio Normalization
3. Noise-Robust Segmentation
4. SVM Classification
5. Feature Normalization
6. 31 Call-Level Statistics
7. 7 Rhythm Features
8. Mood Prediction

Preprocessing:
- Feature Extraction
- Data Modeling

CHAI Lab
University of Michigan
Declipping Method

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

CBAR (*Harvilla and Stern, 2014*)

Harvilla and Stern. "Least Squares Signal Declipping For Robust Speech Recognition."
Declipping Method

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

- **RBAR**
  - Fast approximation to CBAR
  - Used in preprocessing pipeline

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

Harvilla and Stern. "Least Squares Signal Declipping For Robust Speech Recognition."
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 → SVM Classification → Mood Prediction

- Preprocessing
- Feature Extraction
- Data Modeling

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

- Preprocessing
- Feature Extraction
- Data Modeling

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

SVM Classification → Mood Prediction
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><strong>0.70 ± 0.17</strong>*</td>
<td>0.65 ± 0.15</td>
</tr>
<tr>
<td>Normalized By Subject</td>
<td><strong>0.67 ± 0.19</strong>*</td>
<td><strong>0.75 ± 0.14</strong>*</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><strong>0.72 ± 0.20</strong>*</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**

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>0.73 ± 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><strong>0.74 ± 0.24</strong></td>
<td><strong>0.77 ± 0.15</strong></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: $0.57 \pm 0.25 \rightarrow 0.72 \pm 0.20$
  – Depressed: $0.64 \pm 0.14 \rightarrow 0.75 \pm 0.14$

• No comprehensive solution

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

Questions?