Human-Centered Computing: Using Speech to Understand Behavior

Emily Mower Provost
Computer Science and Engineering
University of Michigan
Automated agents provide support, entertainment, and interaction
Embracing Complexity

Environments

Lexical Content

Speech

Individual Differences

Emotion
CHAI Lab Research Directions

• Audio-visual emotion modeling:
  – Perception modeling
  – Expression modeling
  – Methods: deep learning, multitask learning, time series modeling, knowledge-driven

• Assistive technology:
  – Speech assessment for individuals with aphasia
  – Mood state tracking for individuals with bipolar disorder
  – [Early states] Estimating suicidality
  – [Early states] Speech assessment: Huntington’s Disease
CHAI Lab Research Directions

• Audio-visual emotion modeling:
  – Perception modeling
  – Expression modeling
  – Methods: deep learning, multitask learning, time series modeling, knowledge-driven

• Assistive technology:
  – Speech assessment for individuals with aphasia
  – Mood state tracking for individuals with bipolar disorder
  – [Early states] Estimating suicidality
  – [Early states] Speech assessment: Huntington’s Disease
CHAI Lab Research Directions

• Audio-visual emotion modeling:
  – Perception modeling
  – Expression modeling
    – Methods: deep learning, multitask learning, time series modeling, knowledge-driven

• Assistive technology:
  – Speech assessment for individuals with aphasia
  – Mood state tracking for individuals with bipolar disorder
  – [Early states] Estimating suicidality
  – [Early states] Speech assessment: Huntington’s Disease

chai.eecs.umich.edu
Focus on Behaviors

• Goal: detect human behavior from speech
  • Emotion: valence (positivity), activation (energy), categories
  • Mood: depression, suicidality
  • Diagnosis: Huntington Disease, aphasia

extract speech signal

speech signal

extract features + model

activation/valence patterns
Why is this area so important?

ALGORITHMS ➔ IMPACT
Motivation

• Bipolar Disorder (BP)
  – A leading cause of disability worldwide
  – Common, chronic, and severe psychiatric illness
  – Characterized by swings into mania and depression
  – Devastating personal, social, vocational consequences

• Current Treatment
  – Pharmaceutically
  – Periodic follow-up visits for monitoring
  – Reactively post manic/depressive episodes

Costly / Majority Unnecessary
Devastating Consequences
Wellness Monitoring

Depression

Mania

Baseline

chi.eecs.umich.edu
Wellness Monitoring

Over-arching question:
How can we automatically identify an individual’s early warning signs?
Wellness Monitoring

Engineering question:
How can we augment algorithm design with clinical knowledge?

This work:
What if we focus on emotion?
PRIORI Roadmap

Clinical Mood Assessment
- 2014: Initial paper
- 2016: Speech rhythm
- 2016: Personalization
- 2018: Emotion

Natural Speech
- Important change:
  - Inclusion of speech recognition
  - Innovations in robust and generalizable emotion recognition
PRIORI Roadmap

Clinical Mood Assessment

- 2014: Initial paper
- 2016: Speech rhythm
- 2016: Personalization
- 2018: Emotion

Important change:
- Inclusion of speech recognition
- Innovations in robust and generalizable emotion recognition

Natural Speech

- Language
- Anomaly Detection
- Emotion in-the-wild
PRIORI Roadmap

Clinical Mood Assessment

2014
Initial paper

2016
Speech rhythm

2016
Personalization

2018
Emotion

Important change:
- Inclusion of speech recognition
- Innovations in robust and generalizable emotion recognition

Natural Speech

Language [Interspeech 2019]

Anomaly Detection [In prep.]

Emotion in-the-wild
PRedicting Individual Outcomes for Rapid Intervention
Types of PRIORI Calls

- Personal calls:
  - Calls made as someone goes about his/her day
  - Natural speech
Types of PRIORI Calls

• Personal calls:
  – Calls made as someone goes about his/her day
  – Natural speech

• Assessment calls:
  – Clinical interactions over the phone
  – Young Mania Rating Scale (YMRS)
  – Hamilton Depression Scale (HamD)
Types of PRIORI Calls

- **Personal calls:**
  - Calls made as someone goes about his/her day
  - Natural speech

- **Assessment calls:**
  - Clinical interactions
  - Young Mania Rating Scale (YMRS)
  - Hamilton Depression Scale (HamD)

Personal calls grouped by assessment call
The PRIORI dataset

- PRIORI:
  - Longitudinal study of bipolar disorder
  - Collect and analyze mood data for individuals with BP
  - Develop a mood recognition systems

- Participants
  - Patients: BP I and II (51)
  - Healthy controls (9)
  - Dataset size: over 50K calls, over 4K hours of speech
Goal

- Collect and present a new dataset!
- Determine the efficacy of emotion techniques for recognizing mood

Insight

- Emotion and mood both modulate the speech signal

Approach

- Extract common emotion features
- Classify using common emotion recognition techniques

Findings

- There do seem to be differences!

Speech Rhythm

- **Goal**: Determine whether a clinician would designate a person in a mood episode using the rhythm of speech in a clinical interaction

- **Insight**: When manic, speech rate increases, when depressed, it decreases

- **Approach**: 
  - Create a robust pre-processing pipeline
  - Classify mood episode

- **Findings**: 
  - Rhythm can be used to estimate mood
  - It is critical to control for extraneous factors!

---

Methods

Preprocessing

Feature Extraction

Classification

Audio Signal → Device Compensation → Segmentation → Estimate Rhythm → SVM Classification → Mood

RBAR Declipping → Single-task

Combo-SAD → Multi-task
Methods and Results

<table>
<thead>
<tr>
<th>Measure: AUC</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mania</td>
<td>0.57 ± 0.25</td>
</tr>
<tr>
<td>Depression</td>
<td>0.64 ± 0.14</td>
</tr>
</tbody>
</table>
Methods and Results

- **Audio Signal**
  - Device Compensation
  - Segmentation
  - Estimate Rhythm
  - ST SVM Classification

**Measure: AUC**

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>RBAR Declipping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mania</td>
<td>0.57 ± 0.25</td>
<td>0.70 ± 0.17*</td>
</tr>
<tr>
<td>Depression</td>
<td>0.64 ± 0.14</td>
<td>0.65 ± 0.15</td>
</tr>
</tbody>
</table>

* Paired t-test over subjects, p < 0.05
Methods and Results

Audio Signal → Device Compensation → Estimate Rhythm → SVM Classification → Mood

<table>
<thead>
<tr>
<th>Measure: AUC</th>
<th>Baseline</th>
<th>RBAR Declipping</th>
<th>Ignoring Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mania</td>
<td>0.57 ± 0.25</td>
<td>0.70 ± 0.17*</td>
<td>0.74 ± 0.24*</td>
</tr>
<tr>
<td>Depression</td>
<td>0.64 ± 0.14</td>
<td>0.65 ± 0.15</td>
<td>0.77 ± 0.15*</td>
</tr>
</tbody>
</table>

* Paired t-test over subjects, p < 0.05
Methods and Results

- **Measure: AUC**
- **Baseline**
- **RBAR Declipping**
- **Multitask Learning**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Baseline</th>
<th>RBAR Declipping</th>
<th>Multitask Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mania</td>
<td>0.57 ± 0.25</td>
<td>0.70 ± 0.17*</td>
<td>0.72 ± 0.20*</td>
</tr>
<tr>
<td>Depression</td>
<td>0.64 ± 0.14</td>
<td>0.65 ± 0.15</td>
<td>0.71 ± 0.15</td>
</tr>
</tbody>
</table>

* Paired t-test over subjects, p < 0.05
Methods and Results

<table>
<thead>
<tr>
<th>Measure: AUC</th>
<th>Baseline</th>
<th>RBAR Declipping</th>
<th>Subject Normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mania</td>
<td>0.57 ± 0.25</td>
<td>0.70 ± 0.17*</td>
<td>0.67 ± 0.19*</td>
</tr>
<tr>
<td>Depression</td>
<td>0.64 ± 0.14</td>
<td>0.65 ± 0.15</td>
<td>0.75 ± 0.14*</td>
</tr>
</tbody>
</table>

* Paired t-test over subjects, p < 0.05
Personalization

Goal • Improve the prediction of depression

Insight • Individuals are unique and so is their expression of mood

Approach • Speaker verification techniques (i-vectors)

Findings • We can improve depression prediction over speech rhythm features alone

Personalization

Feature Fusion

- Rhythm
- i-vector
- Concatenate Feature Sets
- Hybrid SVM
- Prediction

Decision Fusion

- Rhythm
- i-vector
- Population-General SVM
- Subject-Specific SVM
- Constant / Soft / Hard Decision
- Prediction

Timeline:

- 2014: Initial paper
- 2016: Speech rhythm
- 2016: Personalization
- 2018: Emotion
- Language
- Anomaly Detection
- Emotion in-the-wild
Personalization

Feature Fusion

- Rhythm
- i-vector

<table>
<thead>
<tr>
<th>Concatenate Feature Sets</th>
<th>Population-General SVM</th>
<th>Subject-Specific SVM</th>
<th>Hybrid SVM</th>
<th>Prediction</th>
</tr>
</thead>
</table>

Decision Fusion

- Rhythm
- i-vector

<table>
<thead>
<tr>
<th>Constant / Soft / Hard Decision</th>
<th>Prediction</th>
</tr>
</thead>
</table>

Results

<table>
<thead>
<tr>
<th>System Characteristics</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population-general</td>
<td>0.69 ± 0.15</td>
</tr>
<tr>
<td>Subject-specific</td>
<td>0.70 ± 0.18</td>
</tr>
<tr>
<td>Feature Fusion</td>
<td><strong>0.76 ± 0.13</strong>*</td>
</tr>
<tr>
<td>Constant Decision Fusion</td>
<td>0.74 ± 0.16</td>
</tr>
<tr>
<td>Soft Decision Fusion</td>
<td><strong>0.78 ± 0.12</strong>*</td>
</tr>
<tr>
<td>Hard Decision Fusion</td>
<td>0.76 ± 0.13</td>
</tr>
</tbody>
</table>
Emotion

Goal
• Move to personal calls!

Insight
• Mood is slowly varying, can we improve prediction by focusing on factors more directly expressed in speech?

Approach
• Annotate the data for emotion!
• Transcribe the data!

Findings
• We can accurately predict emotion from natural speech
• In clinical interactions, emotion patterns change with symptom severity

Identifying an intermediary step

- Mood prediction is challenging:
  - Not directly observable
  - Long time scale

- Emotion can simplify mood prediction:
  - Primary BP symptom: emotion dysregulation, utility in classification*
  - Time course: emotion variation between speech and mood

Reference:

Valence and activation annotation:
- 9-point Likert scale
- 11 annotators (7 female, 4 male), between 21 and 34, native speakers of English

Annotators were asked to consider two important points:
- Only the acoustic characteristics, not the content
- Subject-specificity of emotion expression
Emotion Distributions

*Note: categorical labels for demonstration purposes only.

chai.eecs.umich.edu
Emotion Recognition Experimental Setup

• Normalize ground truth labels:
  – Subtracting the rating midpoint of 5
  – Scaling to the range of $[-1, 1]$

• **Subject-independent** cross-validation
  – Experiments repeated for five total runs (six randomly selected folds)
  – Each run: randomly assign two subjects to each fold.
  – Round-robin cross-validation
  – Generates one test measure per fold, resulting in six measures.
  – Output: matrix of 6-by-5 test measures

• Parameter selection: max CCC over validation set
Features and Models

- **Baseline** system
  - 88-dimensional eGeMAPS features
  - Features globally normalized
  - Feed-forward neural network, tanh activation function, linear output

- **Alternative** system
  - 40-dimensional MFB features
  - Features globally normalized
  - Conv-pool network (convolutional layers, global max pooling, dense layers)
  - ReLU and linear activation functions for intermediate and output
Emotion Results

- Conv-Pool > FFNN (PCC, CCC)

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Metric</th>
<th>eGeMAPS FFN</th>
<th>MFBs Conv-Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activation</td>
<td>PCC</td>
<td>0.642 ± 0.076</td>
<td>0.712 ± 0.077</td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td>0.593 ± 0.071</td>
<td>0.660 ± 0.090</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.207 ± 0.012</td>
<td>0.201 ± 0.028</td>
</tr>
<tr>
<td>Valence</td>
<td>PCC</td>
<td>0.271 ± 0.053</td>
<td>0.405 ± 0.062</td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td>0.191 ± 0.031</td>
<td>0.326 ± 0.052</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.199 ± 0.015</td>
<td>0.194 ± 0.016</td>
</tr>
</tbody>
</table>

Bold: p<0.01, paired t-test

chai.eecs.umich.edu
Emotion Results

- Conv-Pool > FFNN (PCC, CCC)
- Activation more accurately recognized

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Metric</th>
<th>eGeMAPS FFN</th>
<th>MFBs Conv-Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activation</td>
<td>PCC</td>
<td>0.642 ± 0.076</td>
<td><strong>0.712 ± 0.077</strong></td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td>0.593 ± 0.071</td>
<td><strong>0.660 ± 0.090</strong></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.207 ± 0.012</td>
<td>0.201 ± 0.028</td>
</tr>
<tr>
<td>Valence</td>
<td>PCC</td>
<td>0.271 ± 0.053</td>
<td><strong>0.405 ± 0.062</strong></td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td>0.191 ± 0.031</td>
<td><strong>0.326 ± 0.052</strong></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.199 ± 0.015</td>
<td>0.194 ± 0.016</td>
</tr>
</tbody>
</table>

Bold: p<0.01, paired t-test
Mood Dataset

- **Goal**: Analyze link between mood and predicted emotion
Experimental Setup

• **Goal**: Analyze link between mood and predicted emotion

• Considerations:
  – Importance of considering how a subject varies about his/her own baseline (euthymic periods)
  – Normalize depressed, manic segments by subject (euthymic segments)

• Approach:
  – Apply conv-pool models to predict emotion
  – Use ensemble over the cross-validation models
  – Analyze over all 10,563 assessment call segments (10,563)
What is the link between mood and emotion?

• Ways to measure:
  – Observe clinical interactions
  – Relate emotion to mood symptom severity (classes or continuous)

• Finding: valence/activation significantly higher in manic vs. depressed episodes

Valence: positive vs. negative
Activation: calm vs. excited
What is the link between mood and emotion?

• Ways to measure:
  – Observe clinical interactions
  – Relate emotion to mood symptom severity (classes or continuous)

• Finding: valence/activation are significantly correlated with mood severity

Valence: positive vs. negative
Activation: calm vs. excited
Comparing Emotion Distributions

• Comparing distributions of valence/activation across subjects

• Comparisons:
  – Over all subjects: one-way ANOVA with p < 0.01
  – Pairwise comparisons: Tukey-Kramer posthoc test (66 pairs)

• Findings:
  – Activation: overall difference, significantly different in 51 cases
  – Valence: overall difference, significantly different in 48 cases
Embracing Complexity

Environments

Lexical Content

Speech

Individual Differences

Emotion

chai.eecs.umich.edu
Research Question

Emotion is a big data problem!

But, what is the best method for transferring paralinguistic information and datasets with different conditions to emotion?

Reference:

chai.eecs.umich.edu
Domain Generalization

- Goal: creates a *middle-ground* representation for *unseen data*
- Removes factors particular to individual datasets
Domain Generalization – Autoencoders

Denoising Autoencoder (DAE)  \hspace{1cm}  Adversarial Autoencoder (AAE)

Variational Autoencoder (VAE)  \hspace{1cm}  Adversarial Variational Bayes (AVB)

Eskimez et al. 2018
Domain Generalization – DANNs

- Domain Adversarial Neural Networks
- **Encode** a middle representation
- **Discriminative**: Classify emotion and domain from middle layer
- **Adversarial**: Backpropagate the reverse gradient of domain
- “**Unlearns**” domain
- No clear target – challenges with converging

Ajakan et al. 2014; Abdelwahab et al. 2018
What if we could still be discriminative?
What if we could still be discriminative?

What if we had a clear target?
Datasets

<table>
<thead>
<tr>
<th></th>
<th>IEMOCAP</th>
<th>MSP-IMPROV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subjects (Male/Female)</strong></td>
<td>10 (5/5)</td>
<td>12 (6/6)</td>
</tr>
<tr>
<td><strong>Environment</strong></td>
<td>Laboratory</td>
<td>Laboratory</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td><strong>Sample Rate</strong></td>
<td>16 kHz</td>
<td>44.1 kHz</td>
</tr>
<tr>
<td><strong>Total Utterances</strong></td>
<td>10039</td>
<td>8438</td>
</tr>
</tbody>
</table>
## Labels

<table>
<thead>
<tr>
<th></th>
<th>IEMOCAP</th>
<th>MSP-IMPROV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Utterances</td>
<td>10039</td>
<td>8438</td>
</tr>
<tr>
<td>Likert Scale</td>
<td>1-5</td>
<td>1-5</td>
</tr>
<tr>
<td>Class Boundaries</td>
<td>1-2, 3, 4-5</td>
<td>1-2, 3, 4-5</td>
</tr>
<tr>
<td>Mean (Std.) Activation</td>
<td>3.08 (0.90)</td>
<td>2.57 (1.10)</td>
</tr>
<tr>
<td>Utt. Without Ties</td>
<td>4814</td>
<td>7290</td>
</tr>
<tr>
<td>Mean (Std.) Valence</td>
<td>2.79 (0.99)</td>
<td>3.02 (1.06)</td>
</tr>
<tr>
<td>Utt. Without Ties</td>
<td>6816</td>
<td>7852</td>
</tr>
</tbody>
</table>
### Labels

<table>
<thead>
<tr>
<th></th>
<th>IEMOCAP</th>
<th>MSP-IMPROV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Utterances</td>
<td>10039</td>
<td>8438</td>
</tr>
<tr>
<td>Likert Scale</td>
<td>1-5</td>
<td>1-5</td>
</tr>
<tr>
<td>Class Boundaries</td>
<td>1-2, 3, 4-5</td>
<td>1-2, 3, 4-5</td>
</tr>
<tr>
<td>Mean (Std.) Activation</td>
<td>3.08 (0.90)</td>
<td>2.57 (1.10)</td>
</tr>
<tr>
<td>Utt. Without Ties</td>
<td>4814</td>
<td>7290</td>
</tr>
<tr>
<td>Mean (Std.) Valence</td>
<td>2.79 (0.99)</td>
<td>3.02 (1.06)</td>
</tr>
<tr>
<td>Utt. Without Ties</td>
<td>6816</td>
<td>7852</td>
</tr>
</tbody>
</table>
## Labels

<table>
<thead>
<tr>
<th></th>
<th>IEMOCAP</th>
<th>MSP-IMPROV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Utterances</td>
<td>10039</td>
<td>8438</td>
</tr>
<tr>
<td>Likert Scale</td>
<td>1-5</td>
<td>1-5</td>
</tr>
<tr>
<td>Class Boundaries</td>
<td>1-2, 3, 4-5</td>
<td>1-2, 3, 4-5</td>
</tr>
<tr>
<td>Mean (Std.) Activation</td>
<td>3.08 (0.90)</td>
<td>2.57 (1.10)</td>
</tr>
<tr>
<td>Utt. Without Ties</td>
<td>4814</td>
<td>7290</td>
</tr>
<tr>
<td>Mean (Std.) Valence</td>
<td>2.79 (0.99)</td>
<td>3.02 (1.06)</td>
</tr>
<tr>
<td>Utt. Without Ties</td>
<td>6816</td>
<td>7852</td>
</tr>
</tbody>
</table>
Method Overview

Audio

Generate Representation

Emotion Classification

Domain Critic

chai.eecs.umich.edu
Baseline: CNN

**CNN**: Convolutional Neural Network trained on all labeled data;

**SP**: Specialist CNN trained on just target labeled data (if available)
ADDoG: Adversarial Discriminative Domain Gen.
MAADDoG: Multiclass ADDoG
Experimental Overview

• Four datasets:
  – IEMOCAP (16 kHz)
  – MSP-Improv (44.1 kHz)
  – PRIORI Emotion (8 kHz)

• Features: Mel Filterbanks (40d, length zero-padded to longest in batch)

• Task: cross-domain valence recognition (three-class)

• Setups:
  – Train on one lab dataset, test on another (IEMOCAP/MSP-Improv)
  – Train on one lab dataset, test on PRIORI Emotion
  – Train on two lab datasets, test on PRIORI Emotion
Experiment 1 – Cross Dataset

<table>
<thead>
<tr>
<th>Train Dataset</th>
<th>Test Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Labeled IEMOCAP or MSP</td>
<td>None Labeled Other Lab Dataset</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>MSP-Improv to IEMOCAP</th>
<th>IEMOCAP to MSP-Improv</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>0.439 ± 0.022 UAR</td>
<td>0.432 ± 0.012 UAR</td>
</tr>
<tr>
<td>ADDoG</td>
<td>0.474 ± 0.009 UAR*</td>
<td>0.444 ± 0.007 UAR*</td>
</tr>
</tbody>
</table>

*Denotes results significantly better than CNN (paired t-test, p=0.05)
Experiment 1 – Increasing Target Labels

Train
IEMOCAP or MSP

All Labeled

Test
Other Lab Dataset

Some Labeled

TAR: Split in Half

Fold 1
Use for Test

Fold 2
Use 200 for Train/Val.

TAR: Split in Half

Fold 1
Use 200 for Train/Val.

Fold 2
Use for Test

SRC
Use for Train

SRC
Use for Train
Experiment 1 – Increasing Target Labels

All Labeled

Train
IEMOCAP or MSP

Test
Other Lab Dataset

Some Labeled

MSP-Improv to IEMOCAP

IEMOCAP to MSP-Improv

Dots denote results significantly different than ADDoG (paired t-test, p=0.05)
Experiment 1 – Increasing Target Labels

- Train:
  - All Labeled
  - IEMOCAP or MSP

- Test:
  - Some Labeled
  - Other Lab Dataset

MSP-Improv to IEMOCAP

IEMOCAP to MSP-Improv

Dots denote results significantly different than ADDoG (paired t-test, p=0.05)
Experiment 1 – Increasing Target Labels

Train
- IEMOCAP or MSP

Test
- Other Lab Dataset

All Labeled

Some Labeled

MSP-Improv to IEMOCAP

IEMOCAP to MSP-Improv

Dots denote results significantly different than ADDoG (paired t-test, p=0.05)
Experiment 2 – To In-the-Wild Data

Train IEMOCAP or MSP

Test PRIORI Emotion

IEMOCAP to PRIORI Emotion

MSP-Improv to PRIORI Emotion

Dots denote results significantly different than ADDoG (paired t-test, p=0.05)
Experiment 2 – To In-the-Wild Data

All Labeled

Train
IEMOCAP
or MSP

Test
PRIORI Emotion

Some Labeled

IEMOCAP to
PRIORI Emotion

MSP-Improv to
PRIORI Emotion

Dots denote results significantly different than ADDoG (paired t-test, p=0.05)
Experiment 3 – To In-the-Wild Data

Train
IEMOCAP or MSP

Test
PRIORI Emotion

All
Labeled

Some
Labeled

IEMOCAP and MSP-Improv to PRIORI Emotion

Dots denote results significantly different than MADDoG (paired t-test, p=0.05)
Experiment 3 – To In-the-Wild Data

Train
IEMOCAP and MSP

Test
PRIORI Emotion

IEMOCAP and MSP-Improv to PRIORI Emotion

Dots denote results significantly different than MADDoG (paired t-test, p=0.05)
Experiment 3 – To In-the-Wild Data

What we learn:
We can’t train a model on outside datasets and expect them to just work

IEMOCAP and MSP-Improv to PRIORI Emotion

Dots denote results significantly different than MADDoG (paired t-test, p=0.05)

Train IEMOCAP and MSP

Test PRIORI Emotion

All Labeled

Some Labeled
Experiment 3 – To In-the-Wild Data

Train
IEMOCAP and MSP

Test
PRIORI Emotion

All Labeled

Some Labeled

IEMOCAP and MSP-Improv to PRIORI Emotion

Dots denote results significantly different than MADDoG (paired t-test, p=0.05)
Experiment 3 – To In-the-Wild Data

Train
IEMOCAP and MSP

Test
PRIORI Emotion

All Labeled
Some Labeled

Where we can go:
We can use these models to derive emotion features in other domains [Interspeech 2019]

IEMOCAP and MSP-Improv to PRIORI Emotion

Dots denote results significantly different than MADDoG (paired t-test, p=0.05)
Conclusions

- ADDoG and MADDoG **consistently converge**
  - Clear target at each step (other dataset)
  - “Meet in the middle” approach
- Effective at detecting emotion in **smartphone calls**
Remaining challenge:
We still aren’t sure about the representation itself!
Emotion Recognition Representation

What if the representation held emotional meaning?

What if points close in emotion were close to each other?

Deep Metric Learning (DML)

• **Goal**: learn an embedding space where pairwise distance corresponds to label similarity
Deep Metric Learning (DML)

- **Goal**: learn an embedding space where pairwise distance corresponds to label similarity

![Diagram showing a range from Calm to Excited with labels for Negative and Positive]
Deep Metric Learning (DML)

- **Goal**: learn an embedding space where pairwise distance corresponds to label similarity
Deep Metric Learning (DML)

• **Goal**: learn an embedding space where pairwise distance corresponds to label similarity

Triplet Loss
[Weinberger and Saul 2009; Chechik et al. 2010; Hoffer and Ailon 2015; Schroff et al. 2015]
Deep Metric Learning (DML)

- **Goal**: learn an embedding space where pairwise distance corresponds to label similarity

Triplet Loss
[Weinberger and Saul 2009; Chechik et al. 2010; Hoffer and Ailon 2015; Schroff et al. 2015]
Deep Metric Learning (DML)

- **Goal**: learn an embedding space where pairwise distance corresponds to label similarity

![Diagram showing triplet loss](chai.eecs.umich.edu)

**Triplet Loss**
[Weinberger and Saul 2009; Chechik et al. 2010; Hoffer and Ailon 2015; Schroff et al. 2015]
Deep Metric Learning (DML)

• **Goal**: learn an embedding space where pairwise distance corresponds to label similarity

Variability is signal, not just noise
Hard labels are too limiting.

- Disagreement in evaluation is extremely common
$f$-Similarity Preservation Loss

**Goal**: learn an embedding space where feature similarity = label similarity
**Goal**: learn an embedding space where feature similarity = label similarity
**Goal**: learn an embedding space where feature similarity = label similarity
Goal: learn an embedding space where feature similarity = label similarity
New Representations

Enforce emotional meaning!

Extract features

Convolutional layer

MFB features

16 width

128 kernels

128-dim vector

Global maxpooling

Fully-connected layer (FC1)

Fully-connected layer (FC2)

Softmax

Classification loss

f-SPL

40 dim

108x337

Enforce emotional meaning!

MFB features

Extract features

Convolutional layer

Global maxpooling

Fully-connected layer (FC1)

Fully-connected layer (FC2)

Softmax

Classification loss

f-SPL

40 dim

108x337
Performance on heldout data

• $f$-SPL less susceptible to overfitting

• Statistically significantly higher performance compared to cross-entropy loss
Embedding with emotional meaning

Baseline
Embracing Complexity

Environments

Lexical Content

Speech

Individual Differences

Emotion

chai.eecs.umich.edu
Thanks!

Questions?

The Heinz C. Prechter Bipolar Research Fund at the University of Michigan Depression Center

chai.eecs.umich.edu