Tradeoff Between Quality And Quantity Of Raters To Characterize Expressive Speech

Alec Burmania, Mohammed Abdelwahab, and Carlos Busso

Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science
Labels from expressive speech

- Emotional databases rely on labels for classification
  - Usually obtained via perceptual evaluations

- Lab Setting
  + Allows researcher close control over subjects
  - Expensive
  - Small demographic distribution
  - Smaller corpus size

- Crowdsourcing
  + Can solve some of the above issues
  + Widely tested and used in perceptual evaluations
  - Raises issues with rater reliability
Labels from expressive speech

- How do we balance quality and quantity in perceptual evaluations?
  - How many labels is enough?
  - Crowdsourcing makes these decisions important

<table>
<thead>
<tr>
<th>Many Evaluators &amp; Low Quality</th>
<th>Few Evaluators &amp; High Quality</th>
</tr>
</thead>
</table>

- How does this affect classification?

- Interprets reliability as a function of quality and quantity
- We use kappa as our metric ($\kappa$) and raters ($n$)

**Effective Reliability**

$$\text{Effective Reliability} = \frac{n\kappa}{1+(n-1)\kappa}$$

<table>
<thead>
<tr>
<th>n raters</th>
<th>0.42</th>
<th>0.45</th>
<th>0.48</th>
<th>0.51</th>
<th>0.54</th>
<th>0.57</th>
<th>0.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>78</td>
<td>80</td>
<td>82</td>
<td>84</td>
<td>85</td>
<td>87</td>
<td>88</td>
</tr>
<tr>
<td>10</td>
<td>88</td>
<td>89</td>
<td>90</td>
<td>91</td>
<td>92</td>
<td>93</td>
<td>94</td>
</tr>
<tr>
<td>15</td>
<td>92</td>
<td>92</td>
<td>93</td>
<td>94</td>
<td>95</td>
<td>95</td>
<td>96</td>
</tr>
<tr>
<td>20</td>
<td><strong>94</strong></td>
<td>94</td>
<td>95</td>
<td>95</td>
<td>96</td>
<td>96</td>
<td>97</td>
</tr>
</tbody>
</table>

MSP-IMPROV Corpus

- Recordings of 12 subjects improvising scenes in pairs (>9 hours, 8,438 turns) [2]
- Actors are assigned context for a scene that they are supposed to act out
- Collected for corpus of fixed lexical content but different emotions

Data Sets
- Target – Recorded Sentences with fixed lexical content (648)
- Improvisation – Scene to produce target
- Interaction – Interactions between scenes

MSP-IMPROV Corpus

How can I not?

Anger
Lazy friend asks you to skip class

Happiness
Accepting job offer

Sadness
Taking extra help when you are failing classes

Neutral
Using coupon at store
Perceptual Evaluation

- **Idea:** Can we verify if a worker is spamming even while lacking ground truth labels for most of the corpus?

- **We will focus on a five class problem (Angry, Sad, Neutral, Happy, Other)**

**Collect Reference Set** (Gold Standard)

**Interleave Reference Set with Data** (Online Quality Assessment)

**Collect reference set**

**Trace performance in real time**

---

Metric: Angular Agreement

- Assign categories (angry, sad, happy neutral, other) as a 5D space (v).
- We calculate the LOWO inter-evaluator agreement as below:

\[
 Agreement(\theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{V(i) \cdot \hat{V}_i}{\|V(i)\| \|\hat{V}_i\|}
\]

- Assume the rater we are evaluating chooses angry:
- We then recalculate the agreement as above and find the difference:

\[
 \Delta \theta = \theta_t - \theta_s
\]
Average Difference of Gold Standard
Performance Averaged over first two sets
First Group of Evaluators Removed
This is still an issue!
Offline Filtering Process

- Because we have the quality at each of the checkpoints, we can filter results that fall below a certain threshold.
- This gives us target sets with an average of number of evaluations >20.
- Thus we can filter to have sets with different inter-evaluator agreement.
- We choose Angular agreement as our metric (useful for minority emotions).

We can control this to produce sets of varying quality.
Secondary Post-processing threshold ($\Delta \theta$)

$\Delta \theta = 25^\circ$
\[ \Delta \theta = 5^\circ \]
### Rater Quality

** Constant sample size **

<table>
<thead>
<tr>
<th>Δθ</th>
<th>5 Raters</th>
<th>10 Raters</th>
<th>15 Raters</th>
<th>20 Raters</th>
<th>25 Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># sent</td>
<td>κ</td>
<td># sent</td>
<td>κ</td>
<td># sent</td>
</tr>
<tr>
<td>5</td>
<td>638</td>
<td>0.572</td>
<td>525</td>
<td>0.558</td>
<td>246</td>
</tr>
<tr>
<td>10</td>
<td>643</td>
<td>0.532</td>
<td>615</td>
<td>0.522</td>
<td>466</td>
</tr>
<tr>
<td>15</td>
<td>648</td>
<td>0.501</td>
<td>643</td>
<td>0.495</td>
<td>570</td>
</tr>
<tr>
<td>20</td>
<td>648</td>
<td>0.469</td>
<td>648</td>
<td>0.471</td>
<td>619</td>
</tr>
<tr>
<td>25</td>
<td>648</td>
<td>0.452</td>
<td>648</td>
<td>0.450</td>
<td>643</td>
</tr>
<tr>
<td>30</td>
<td>648</td>
<td>0.438</td>
<td>648</td>
<td>0.433</td>
<td>648</td>
</tr>
<tr>
<td>35</td>
<td>648</td>
<td>0.425</td>
<td>648</td>
<td>0.433</td>
<td>648</td>
</tr>
<tr>
<td>40</td>
<td>648</td>
<td>0.420</td>
<td>648</td>
<td>0.427</td>
<td>648</td>
</tr>
<tr>
<td>90</td>
<td>648</td>
<td>0.422</td>
<td>648</td>
<td>0.419</td>
<td>648</td>
</tr>
</tbody>
</table>

- **Increasing agreement due to filter**
- **Decreasing samples meeting size criteria**
Experimental Setup

- Let's choose 4 scenarios which trade off quality and quantity, assess their effective reliabilities and classification performance
  - **Case 1**: High Quality, Low Quantity
    - 5 degree filter, and 5 Raters ($\kappa = 0.572$)
  - **Case 2**: Moderate Quality, Moderate Quantity
    - 25 Degree Filter, 15 raters ($\kappa = 0.450$)
  - **Case 3**: Low Quality, Low Quantity
    - No Filter, 5 Raters ($\kappa = 0.422$)
  - **Case 4**: Low Quality, High Quantity
    - No Filter, 20 Raters ($\kappa = 0.419$)
Classification

- Five Class Problem (Angry, Sad, Neutral, Happy, Other)
  - Excluded turns w/o majority vote agreement
  - Acoustic Features IS 2013 - OPENSIMILE

Feature Extraction
D = 6373

CAE Feature Selection
D = 1000

Forward Feature Selection
D = 50

SVM Classifier

6F-SI Cross Validation

Quality

Quantity

C1

C2

C3

C4
# Results

## Common Turns in all Cases

<table>
<thead>
<tr>
<th>Case</th>
<th># Turns</th>
<th>Acc. (%)</th>
<th>Pre. (%)</th>
<th>Rec. (%)</th>
<th>F-score(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>514</td>
<td>47.39</td>
<td>46.53</td>
<td>47.39</td>
<td>46.96</td>
</tr>
<tr>
<td>Case 2</td>
<td>514</td>
<td>48.23</td>
<td>47.42</td>
<td>48.23</td>
<td>47.82</td>
</tr>
<tr>
<td>Case 3</td>
<td>514</td>
<td>47.07</td>
<td>46.62</td>
<td>47.07</td>
<td>46.84</td>
</tr>
<tr>
<td>Case 4</td>
<td>514</td>
<td>47.88</td>
<td>47.17</td>
<td>47.88</td>
<td>47.52</td>
</tr>
</tbody>
</table>

## EF Reliability, Reliability Rank, F-Score Rank

<table>
<thead>
<tr>
<th>Case</th>
<th>EF Reliability</th>
<th>Reliability Rank</th>
<th>F-Score Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>87</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Case 2</td>
<td>92</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Case 3</td>
<td>78</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Case 4</td>
<td>94</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Discussion

- Relatively small differences appear in labels (<10%)
  - “Wisdom of the crowd” seems to be useful for emotion

- Cost
  - Accuracy desired may be a function of cost
    - Is it worth 4x cost for minor improvement?
    - What is the cost of quality?

<table>
<thead>
<tr>
<th>Case</th>
<th>Cost</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>26</td>
<td>32</td>
</tr>
<tr>
<td>Case 2</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>Case 3</td>
<td>-</td>
<td>36</td>
</tr>
<tr>
<td>Case 4</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
What does this mean?

- We can establish a rough crowdsourcing framework for emotion.

- Test collection for reliability

- Establish reliability target and cost target

- Data Collection

Repeat as needed
Questions?

Interested in the MSP-IMPROV database? Come visit us at msp.utdallas.edu and click “Resources”
References

