Using Crowdsourcing in the Rating of Emotional Speech Assets

Sarah Jane Delany
*Technological University Dublin*, sarahjane.delany@tudublin.ie

Alexey Tarasov
*Technological University Dublin*, tarasovsaleksejs@gmail.com

John Snel
*Technological University Dublin*

Charlie Cullen
*Technological University Dublin*, charlie.cullen@tudublin.ie

Follow this and additional works at: https://arrow.tudublin.ie/scschcomoth

Part of the Computer Sciences Commons

**Recommended Citation**


This work is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 3.0 License](https://creativecommons.org/licenses/by-nc-sa/3.0/)
Using Crowdsourcing in the Rating of Emotional Speech Assets

Sarah Jane Delany
Alexey Tarasov
Charlie Cullen
John Snel
Agenda

• Introduction to the domain
  – Automatic Recognition of Emotion from Speech
• Process of emotion recognition
  – Main challenges
• Labelling of corpora by using crowdsourcing
  – Main challenges and our directions
Applications of Emotion Recognition

• Monitoring call center operators [Gupta2007]
• ICT services for hospitalized children [Yilmazyildiz2006]
• E-learning [D’mello2007]
• Recognition of driver’s sleepiness [Grimm2007]
• Smart homes [Kostoulas2008]
• Software usability testing [Schultz2007]
• Surveillance [Clavel2006]
Training Emotion Recognition Classifiers

Data Acquisition

Labelling

Feature Extraction

Build Classifier

Model

speech utterances

happy

active

neutral

Data Acquisition

Labelling

Feature Extraction

Build Classifier

Model
Classification

Data Acquisition

Labelling

Feature Extraction

Build Classifier

Model

neutral
Classification

Data Acquisition

Labelling

Feature Extraction

Build Classifier

Model

Challenges
## Data Acquisition

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acted speech</td>
<td>Actors are asked to depict certain emotion</td>
</tr>
<tr>
<td>Natural speech</td>
<td>Recordings of real-life situations are used [call-centers, talk-shows etc.]</td>
</tr>
<tr>
<td>Induced/elicted speech</td>
<td>Emotions are induced in a controlled environment</td>
</tr>
</tbody>
</table>
Control vs. naturalness
Control vs. naturalness

Naturalness

Control

Acted
Control vs. naturalness

- Control
- Naturalness
- Natural
- Acted
Control vs. naturalness

Naturalness

Control

Natural

Induced

Acted
Control vs. naturalness

- Naturalness
- Control

- Induced
  - MIPs
  - Mood Induction Procedures
- Acted
Mood induction procedures

- Subjects are placed in controlled environment
- Wizard-of-Oz changes it to induce emotions

<table>
<thead>
<tr>
<th>AIBO robot</th>
<th>[Steidl2009]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer games</td>
<td>[Vaughan2010]</td>
</tr>
<tr>
<td>Driving simulator</td>
<td>[Jones2005]</td>
</tr>
</tbody>
</table>
Labelling: Categories vs. Dimensions

- Boundaries are vague
- Different meanings for different people
- Culture-dependent

![Diagram showing dimensions of active and passive, positive and negative with categories of feelings including Anger, Disgust, Fear, Happiness, Sadness, and Surprise.](image)
Discretization of dimensions

Please choose the activation level:

- Passive
- Slightly Passive
- Average
- Slightly Active
- Active

Please choose the evaluation level:

- Negative
- Slightly Negative
- Neutral
- Slightly Positive
- Positive

[Grimm2007a]
Feature Extraction

- Features
  - Acoustic
  - Lexical
  - Contextual
- Prosodic
- Spectral
Most widely used classifiers

- k Nearest Neighbours
- C4.5 Decision Trees
- Support Vector Machines
- Artificial Neural Networks
- Naïve Bayes
- Ensembles
How corpora are labelled now?

• Small number of experts
  – Not cheap
  – Not fast
  – Emotion is subjective
Crowdsourcing

“The act of taking a task traditionally performed by a designated agent and outsourcing it to an undefined, generally large group of people in the form of an open call”  [Jeff Howe]
Crowdsourcing

• June 2006 Wired magazine article by Jeff Howe

...the power of many...
**Mechanical Turk is a marketplace for work.**
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

**86,131 HITs** available. [View them now.](https://www.mturk.com/mturk/)

---

### Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. [Find HITs now.](https://www.mturk.com/mturk/)

**As a Mechanical Turk Worker you:**
- Can work from home
- Choose your own work hours
- Get paid for doing good work

---

### Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. [Register Now](https://www.mturk.com/mturk/)

**As a Mechanical Turk Requester you:**
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

---

[Find HITs Now](https://www.mturk.com/mturk/)

[Get Started](https://www.mturk.com/mturk/)
How to Play

1. You and a partner see the same image.

2. Each of you must guess what words your partner is typing.
Crowdsourcing

• Triggered a shift in the way labels or ratings are obtained in variety of domains:
  – natural language tasks [Snow2008]
  – computer vision [Sorokin2008, von Ahn2004]
  – sentiment analysis [Hsueh2008, Brew2010]
  – machine translation [Ambati2010]
Labelling process

1.wav
2.wav
3.wav
4.wav
Labelling process

1.wav

2.wav

3.wav

4.wav
Labelling process

1.wav

2.wav

3.wav

4.wav
Labelling process

1.wav

2.wav

3.wav

4.wav
Labelling process

1.wav

2.wav

3.wav

4.wav
Labelling process

1.wav
2.wav
3.wav
4.wav
Labelling process

1.wav

2.wav

3.wav

4.wav
Challenges

Our objective: to label large collections of speech assets using crowdsourcing but identifying and using good raters

• How to determine the sequence of assets?
• How to estimate the reliability of raters?
• How to use multiple ratings received?
Determining the sequence of assets

• Active learning is a typical approach
  – [Yan2011] (logistic regression only)
  – [Ambati2010] (applicable only to the natural language processing)
Estimating Rater Reliability

• Everyone rates everything [Raykar2010, Whitehill2009]

  – typically for binary classification task
  – prior information about raters needed
Using Ratings

• Majority vote or weighted average

• Learning algorithms, e.g. EM [Raykar2010]
  – everyone has to rate everything
Listen and Rate

June 21, 2011, 1:02 pm

Logged in as: tarasovs.aleksejs@gmail.com. You have rated in total 30 assets. In this session you have rated 2 assets and listened to 2 assets.

Step 01

Please listen to the audio file and rate it accordingly.

(Important note: Be aware that if you use the 'Refresh', 'Back' or 'Forward' buttons in the browser you will be presented with a new speech clip. The same applies if you go to the 'Instructions' page. Please rate accordingly.)

Step 02

Please choose the activation level:

- Passive
- Slightly Passive
- Average
- Slightly Active
- Active
Preliminary Experiment

• Using the fully labelled VAM corpus
• Simulating the labelling process
• Three approaches to labeller selection
  – Random selection
  – Best overall selection
  – Best so far selection
Preliminary results
Next steps

• Using active learning for the presentation of assets with SVM and SVR
• Perform dynamic assessment of raters’ reliabilities without having any prior information about them
• Label our corpus using these techniques
Thank you for your attention!

Any questions?
References

- Brew. Using crowdsourcing and active learning to track sentiment in online media.