CROWDSOURCING EMOTIONAL SPEECH

Jennifer Smith, Andreas Tsiartas, Valerie Wagner,
Elizabeth Shriberg, Nikoletta Bassiou

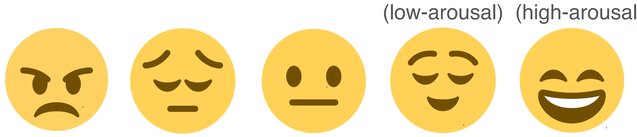
SRI International®

Introduction

Introduction

- Crowdsourced both data collection & annotation
- Subjects record themselves multi-condition data
- Annotate for emotion in both tone of voice and content
- Five emotions:

angry | sad | neutral | happy (low-arousal) | happy (high-arousal)



Collection Method

Design process

- Collected data from 20 subjects using prompts that varied in:
  - Number of emotional recordings
  - Length of emotional recordings
  - Descriptive words for emotions
  - Number of emotions per batch
  - With and without audio examples
- Expert annotators evaluated recordings for quality of emotion in the tone of voice and content
- Iterated and used expert annotations and feedback from subjects to tune the variables

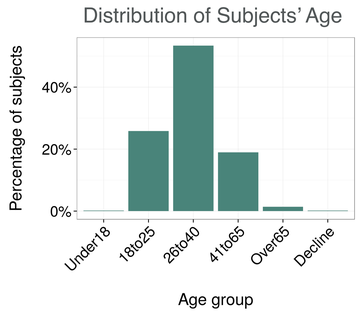
Final Prompt for Eliciting Emotional Speech

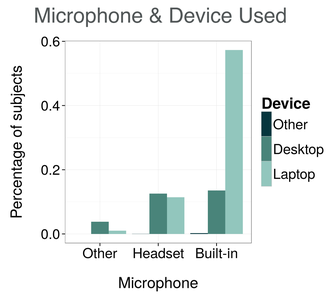
- Ten recordings:
  - Eight unscripted emotion recordings
  - Two non-emotional scripted recordings
- Subject prompted to use past emotional experiences
- Unscripted: imagine yourself in that moment and express your emotion as if speaking with a trusted friend or family member
- Express “full blown” emotion
- Embody the emotion (e.g. sit up straight and smile if happy or frown if sad)
- Listen to three varied examples of emotional speech before recording

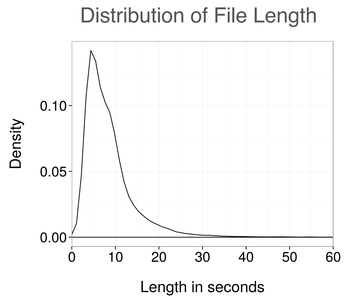
The Corpus

Descriptive Statistics

- 187 hours of data
- 110,068 audio recordings
- 2,965 subjects
- Gender: 57.1% female, 42.6% male, 0.3% declined to answer
- 39.7% of subjects contributed data to only one emotion, 16% contributed data for all five emotions







Annotation Method

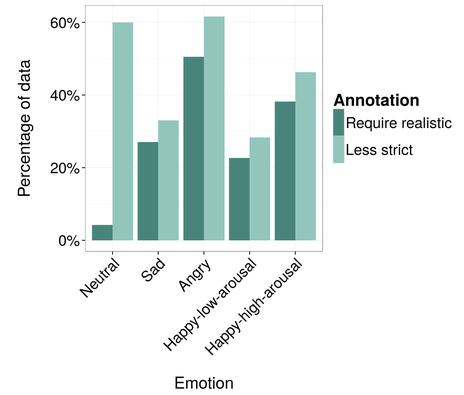
Annotation Questions

- Audio quality
  - No speech
  - Voices or sounds in background
  - Distortion, skipping
- Separately categorize the emotion in the tone of voice from the content of the words
- Rate the emotional quality:
  - Emotion is faint
  - Emotion is very exaggerated or the person is goofing around
  - Emotion sounds very realistic
  - None of these

Results & Discussion

Annotation Results

- Annotated 5,168 recordings (5% of the data)
- Audio quality issues present in 2% of the data
- Emotion in the tone of voice & content matched for 46% of the data
- Emotion rated “very realistic” for 29% of the data
- Requiring data sound “very realistic” removes about 8-10% of data for all emotions except neutral, for which it removes 60% of data



Initial Evaluation of Annotation

- Classification experiment using SenSay™ platform
  - Single-layer Neural Network
  - 2 sets: (1) annotated subset; (2) equally sized subset drawn randomly from data
  - Features include spectral, prosodic, articulatory, noise-robust, etc.
  - 4-way classification: Neutral removed due to lack of data
  - Balanced classes
- 5.3% absolute (13.3% relative) improvement in accuracy

Lessons Learned

- Crowdsourcing is fast and inexpensive way to collect and annotate a large corpus
- Multiple short utterances best for eliciting “full-blown” emotion
- Ten recordings / batch maximizes the number of samples while minimizing emotional-quality loss due to fatigue/boredom
- Collecting emotions separately minimizes cognitive load of switching emotions
- Requiring subjects listen to three varied audio examples increased emotional-quality without narrowing emotional expression
- Difficult to recruit long-term annotators – better to account for annotator error in post-processing
- Special attention should be paid to eliciting “neutral” data