RESEARCH REVIEW 2019

Emotion Recognition from Voice in the Wild

Oren Wright

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM19-1106

Toward Machine-Emotional Intelligence

Current

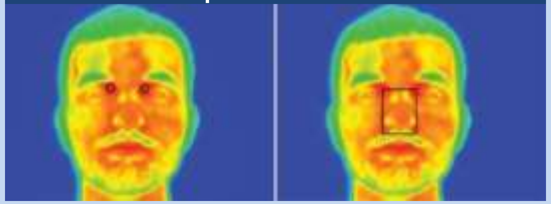
Heart Rate

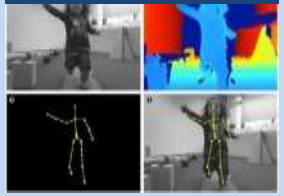

Facial Microexpressions


Voice


Proposed and Future

Gaze


Respiration Rate


Gait


Pupil Dilation


Perspiration


Passive Biometrics at the SEI

Real-Time Heartrate Extraction (2016)

Facial Micro-Expression Analysis (2017)

Used with permission of the Poker Channel:
youtube.com/user/sergeypoker/

Voice Forensics at CMU Language Technologies Institute

U.S. Coast Guard photo by Eric D. Woodall

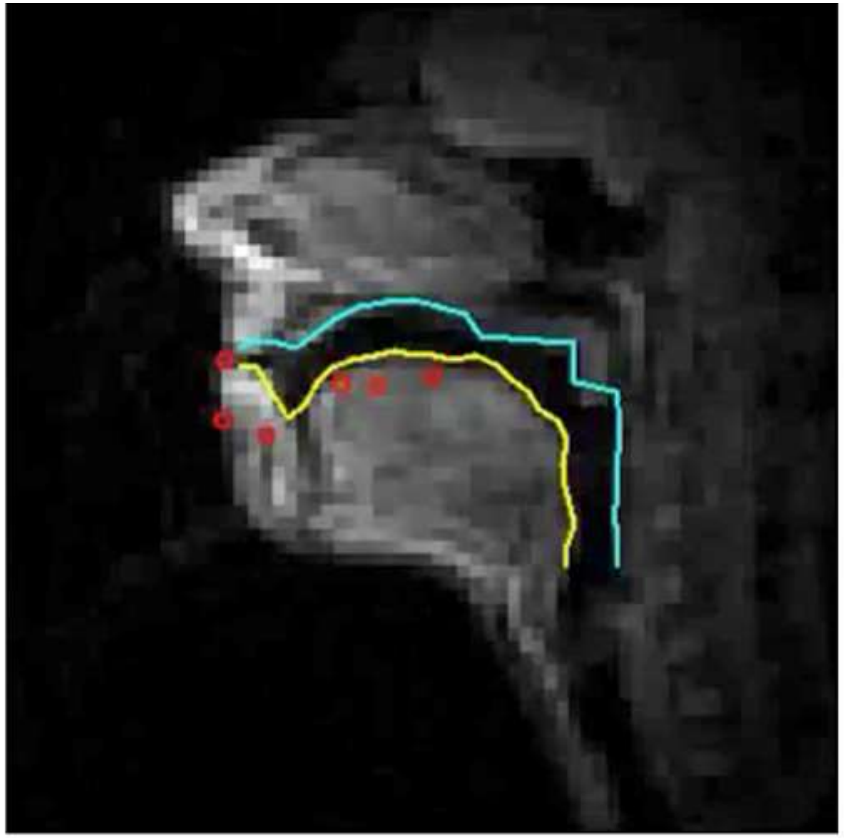
Profiling Hoax Callers

R. Singh, et al., 2016

This person is:

- White
- Brought up in the U.S.A.
- Approx. 175 cm tall
- Approx. 75 kg
- Approx. 40 years old
- Not in any trouble
- Not on a boat
- In a warehouse of some kind
- Using homemade equipment
- Sitting on a metal chair upon a concrete floor

Emotion Recognition from Voice



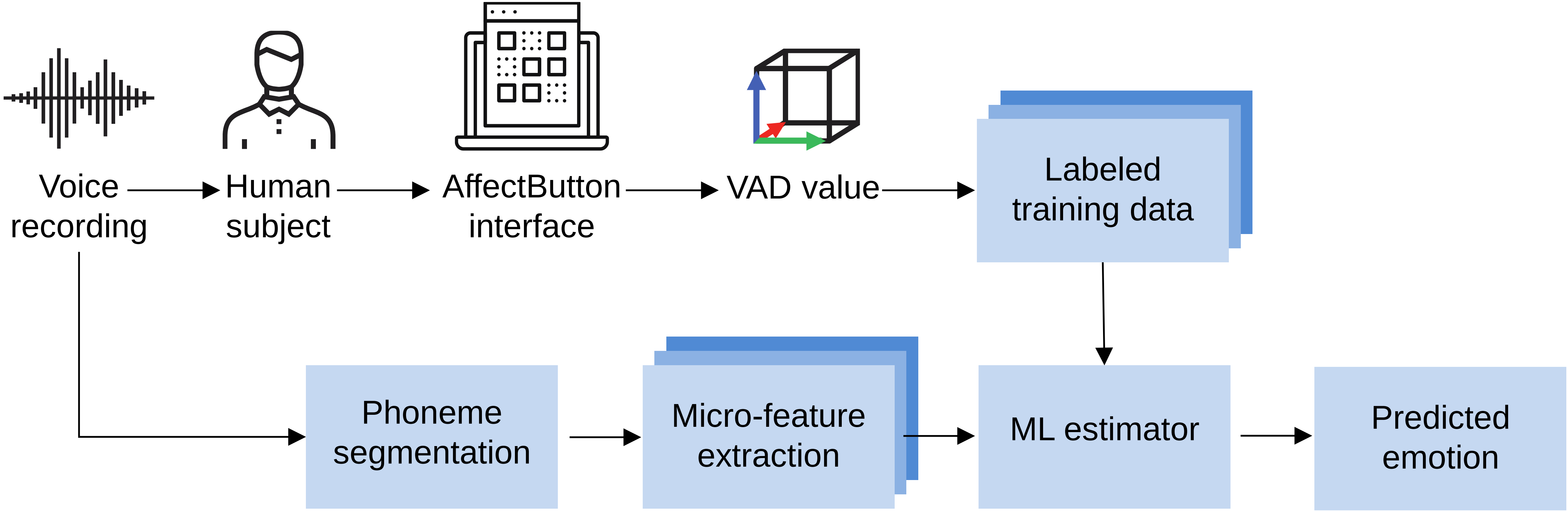
- Voice is a complex process that presents bio-markers
- Bio-marker analysis enabled by **micro-articulometry**
- Made possible by 30 years of automatic speech recognition technology at CMU

"cat" → /k/æ/t/ → Micro-feature extraction

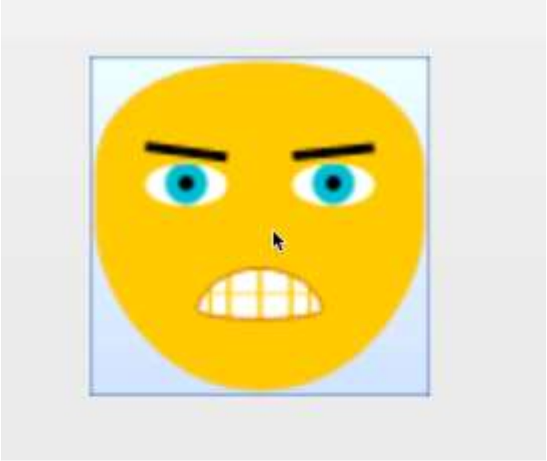
Mission Applications

- Security checkpoints and encounters
- Interrogations
- Intelligence profiling
- Media analysis and exploitation
- Detection of stress, PTSD
- Human-machine teaming

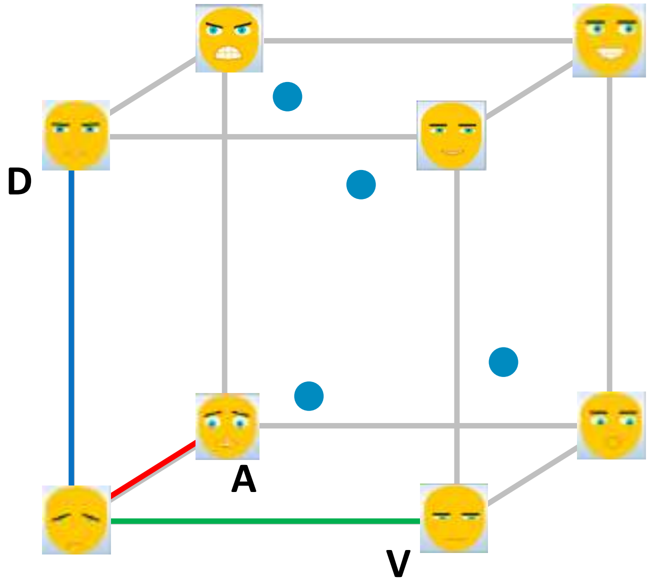
Emotion Recognition from Voice



Database Construction



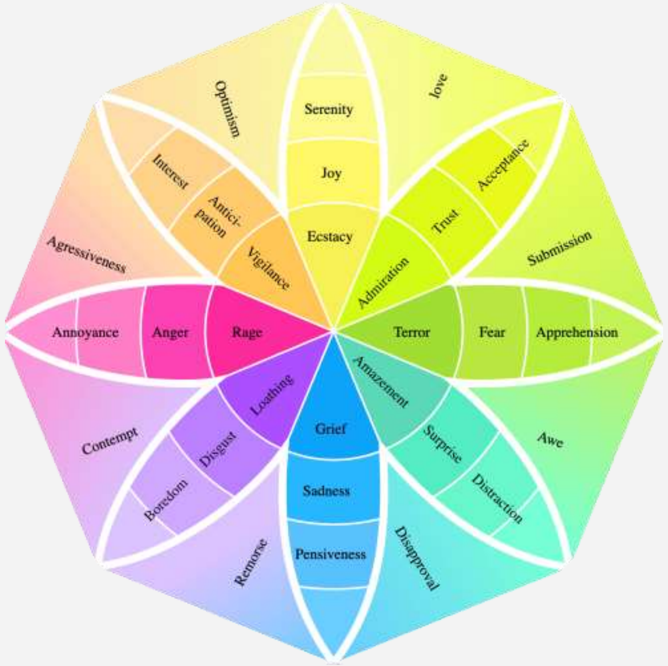
Broekens & Brinkman, 2009



Mehrabian & Russell, 1974

CMU-SER Database

- **Largest ever in-the-wild speech emotion database**
  - Over **29,000** annotated audio clips, totaling over **54** hours of voice recordings
  - Over **324,000** unique annotations
- **Open source tools**
  - Voice processing and exemplar creation
  - Crowdsourcing platform with Amazon Mechanical Turk

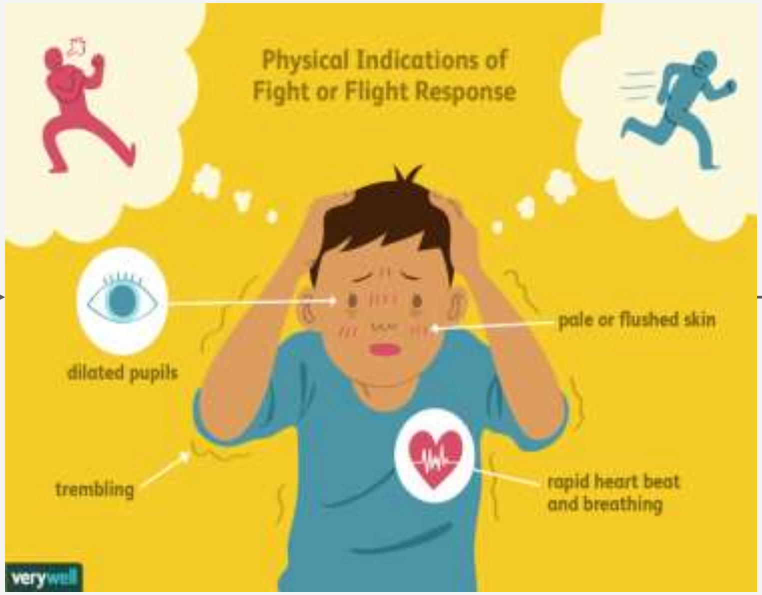


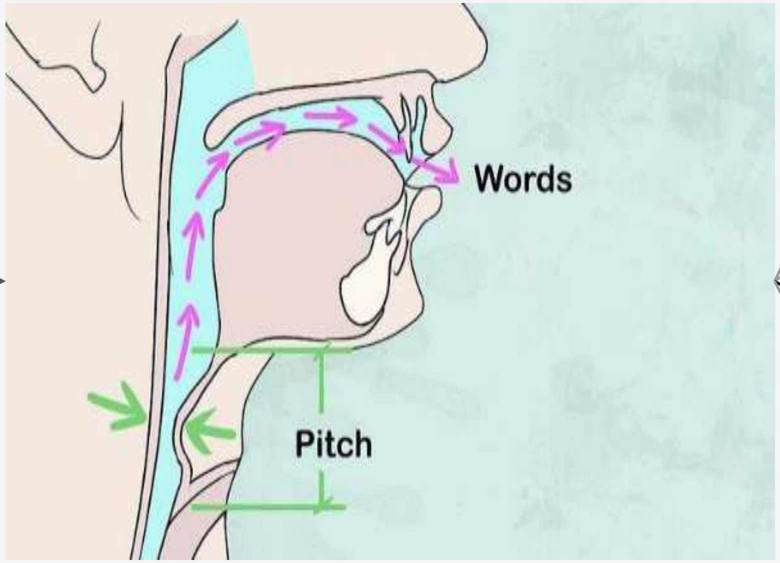
AI for Super-Intelligent Hearing:

Deducing Human Emotion Status from Voice

Rita Singh
Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Motivation: Voice and Emotions





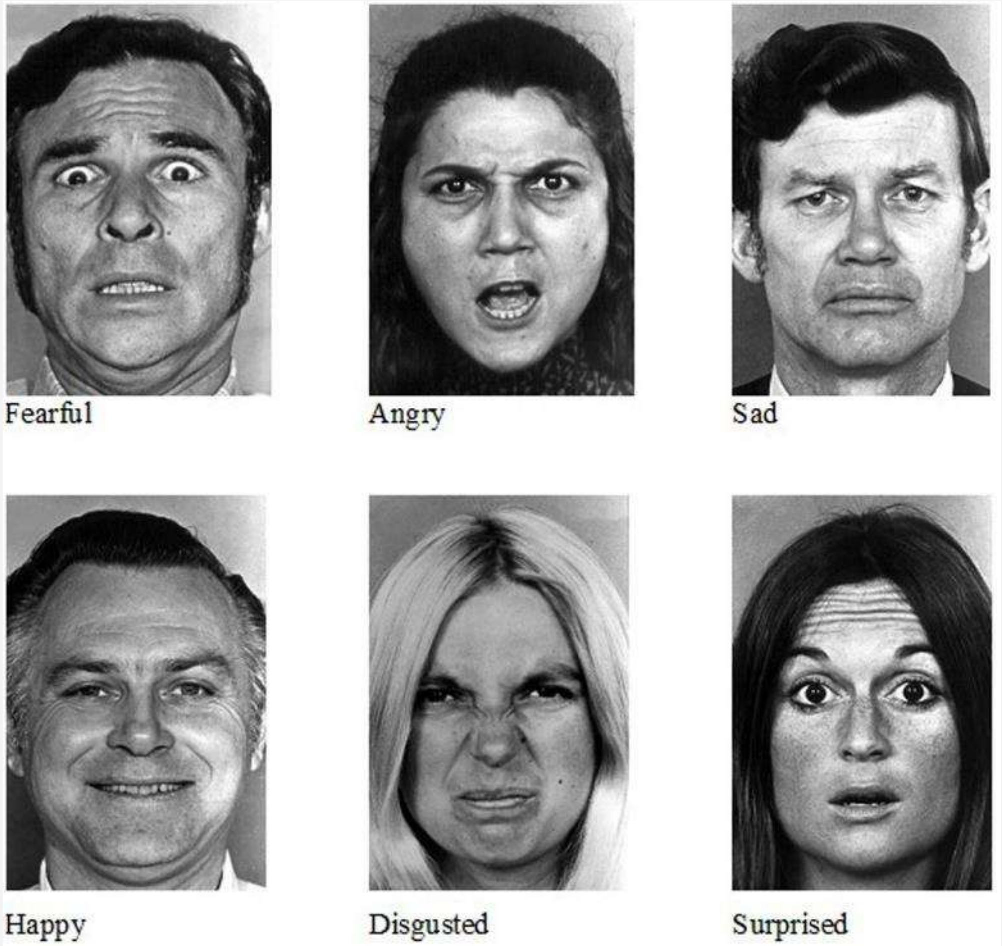
HAPPY

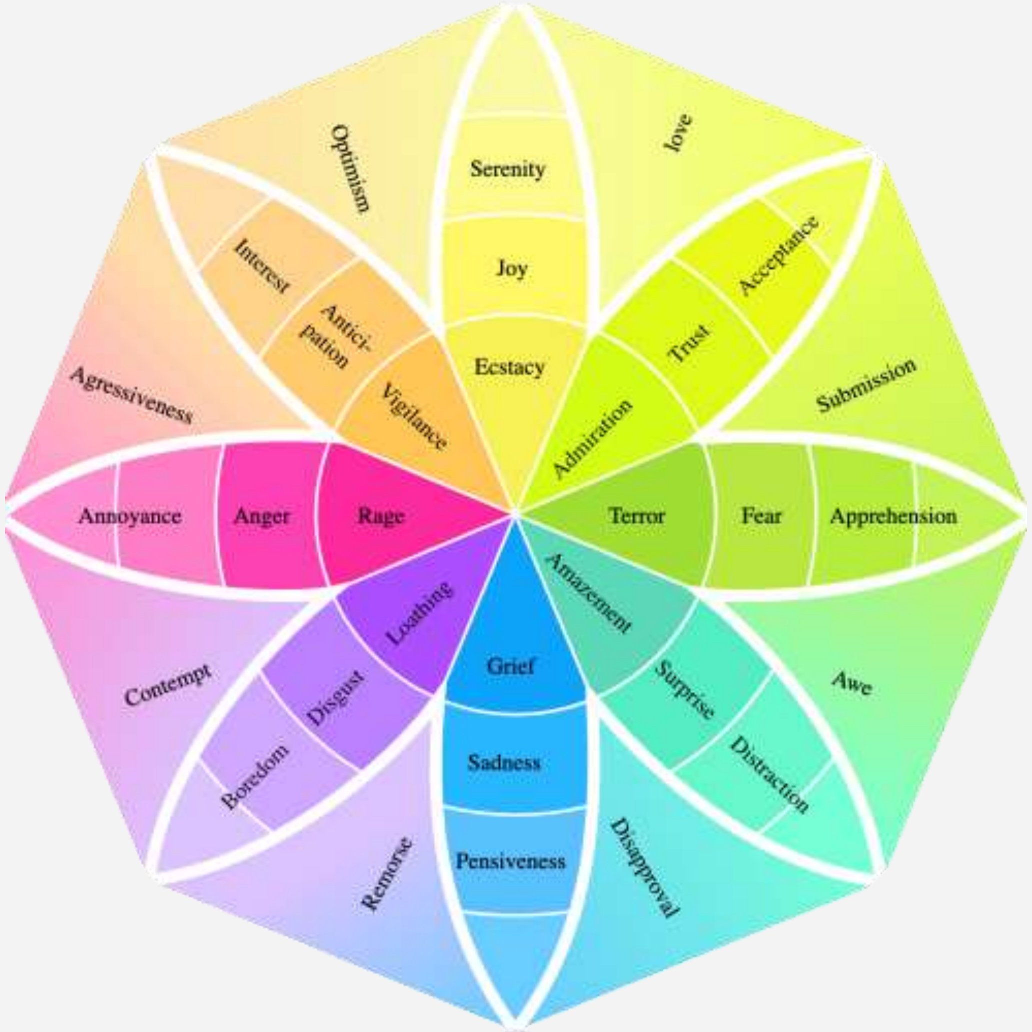
SAD

FEAR

NEUTRAL

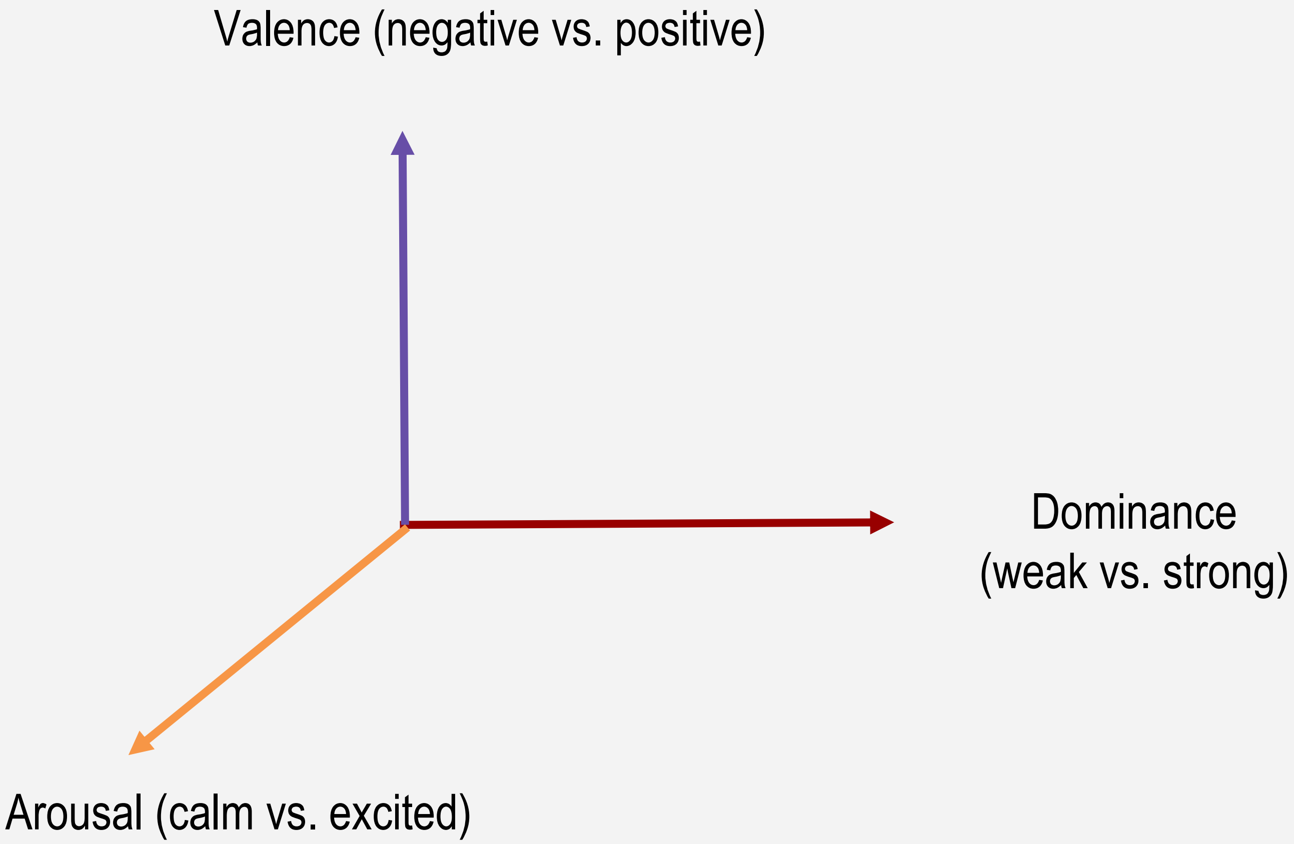
Motivation: Discrete Emotions



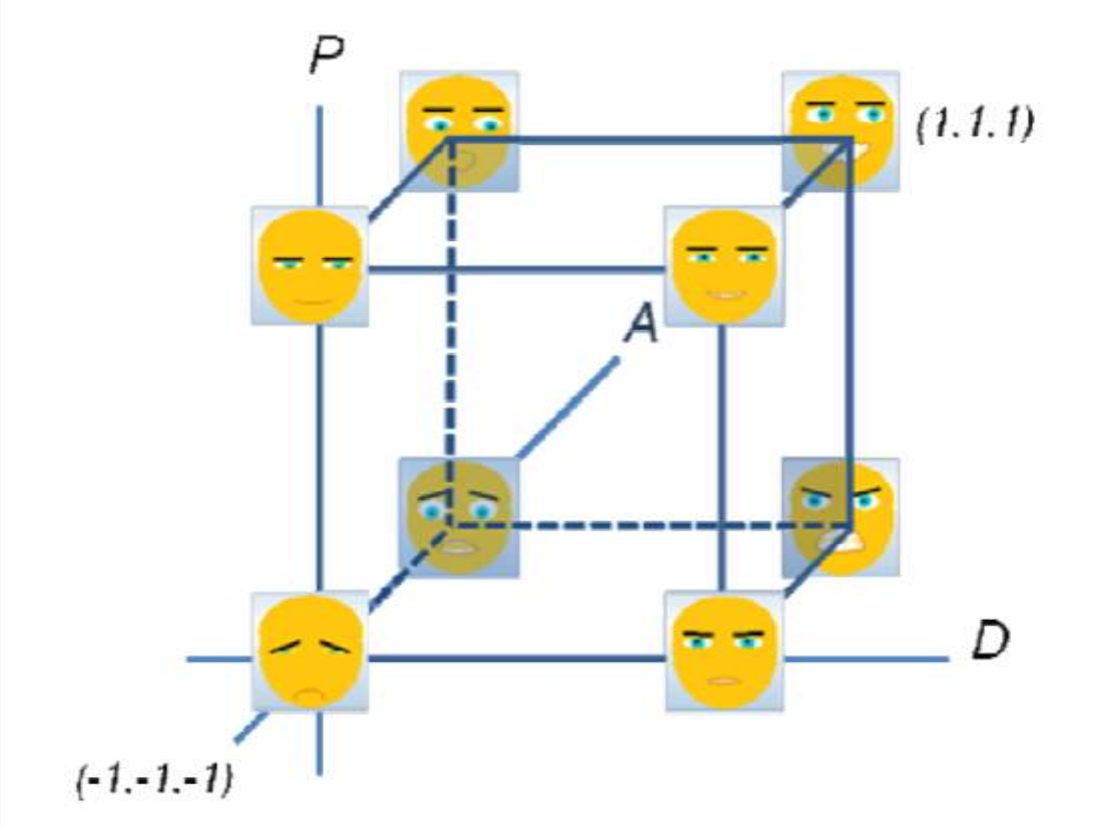


Motivation: Continuous Emotions

- Discrete emotions are limited.
- People exhibit complex affective states.
- Complexity of emotions is difficult to capture via discrete emotions.
- Emotion Primitives: Describing emotions on a continuum (valence, arousal, dominance)

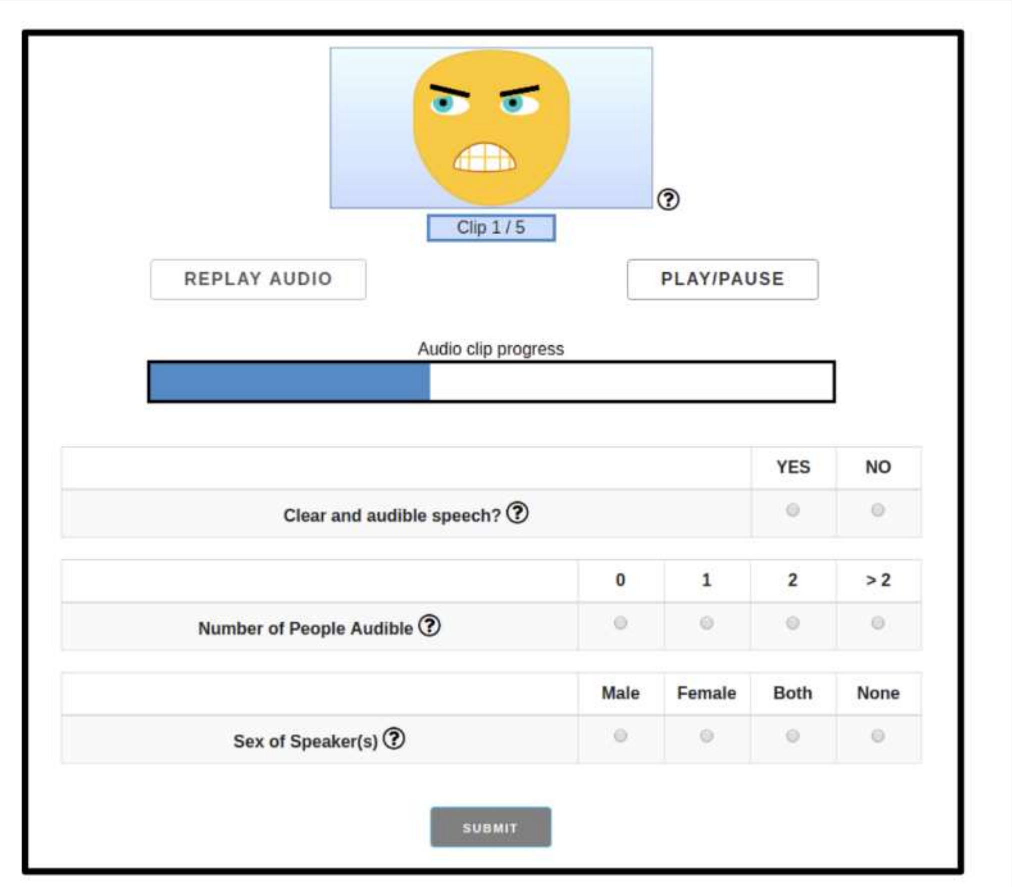


Methodology: Affect Button

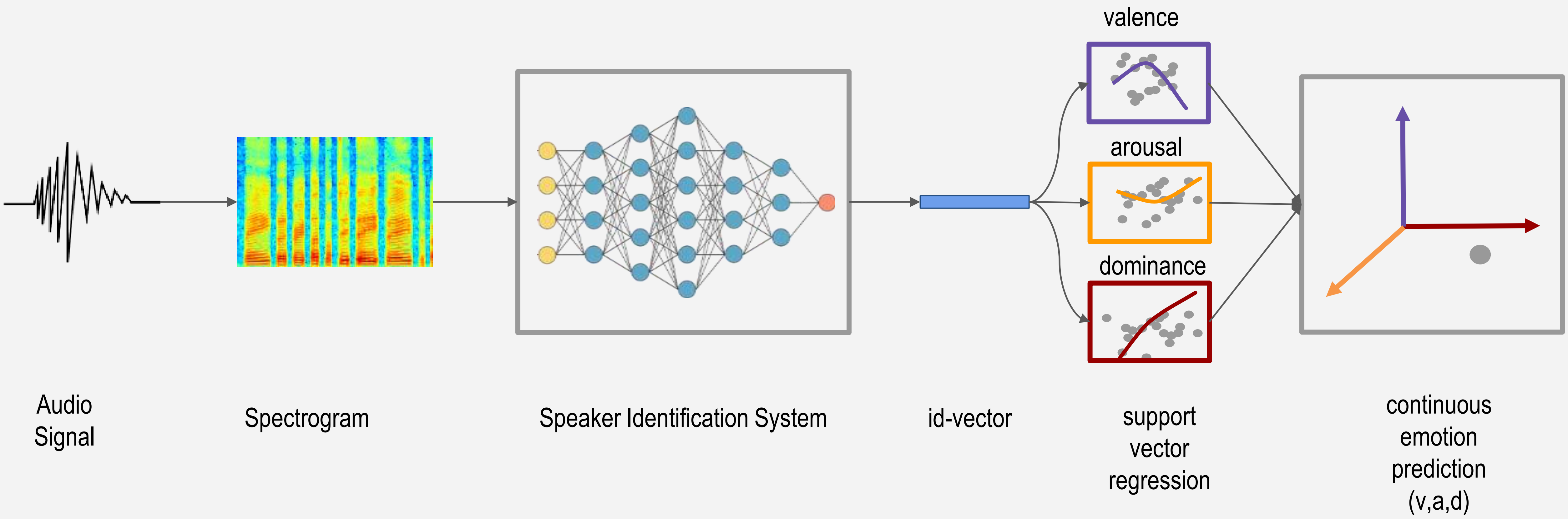


Methodology: Data Collection & Crowdsourcing

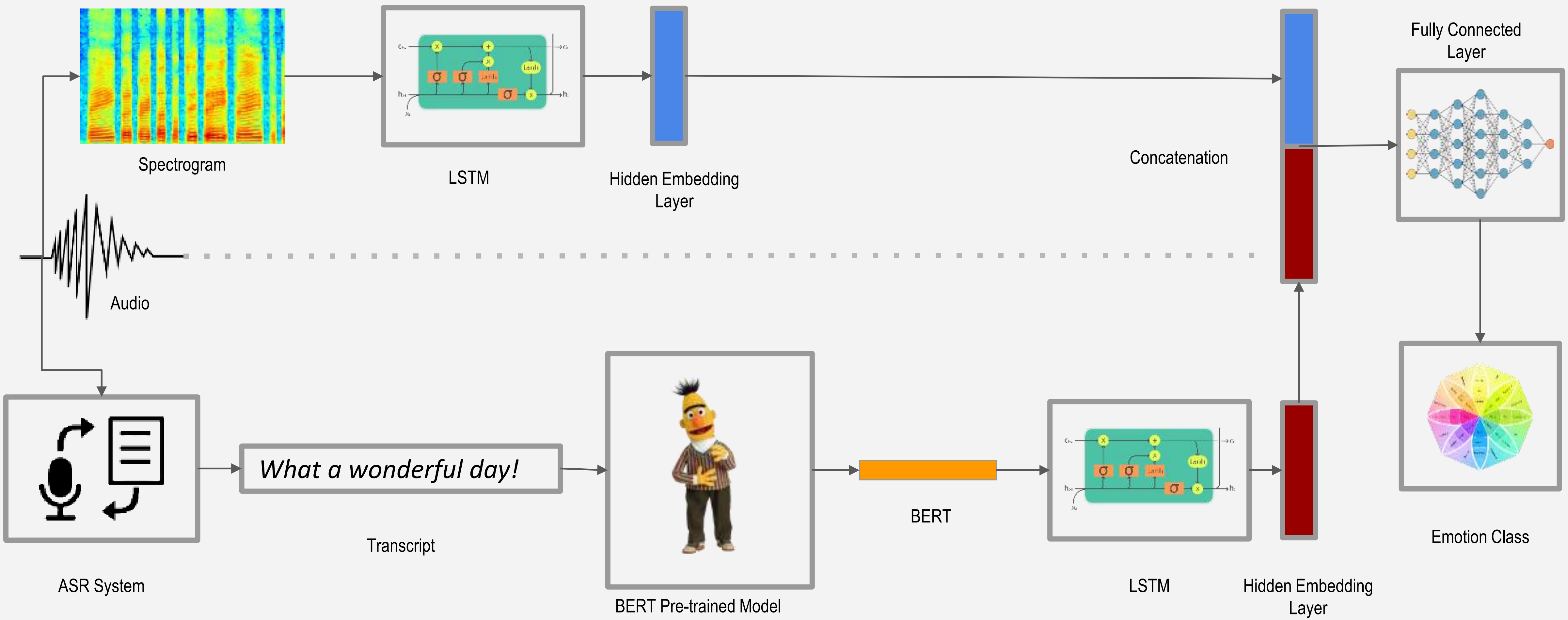
- Data selection (internet archive, news channels, etc.)
- Data cleaning (noisy utterances, music, manual annotations, etc.)
- Pre-hoc quality control
- Challenges of crowdsourcing:
  - Re-sampling
  - Filtering



Methodology: Model I (speaker-specific embeddings)



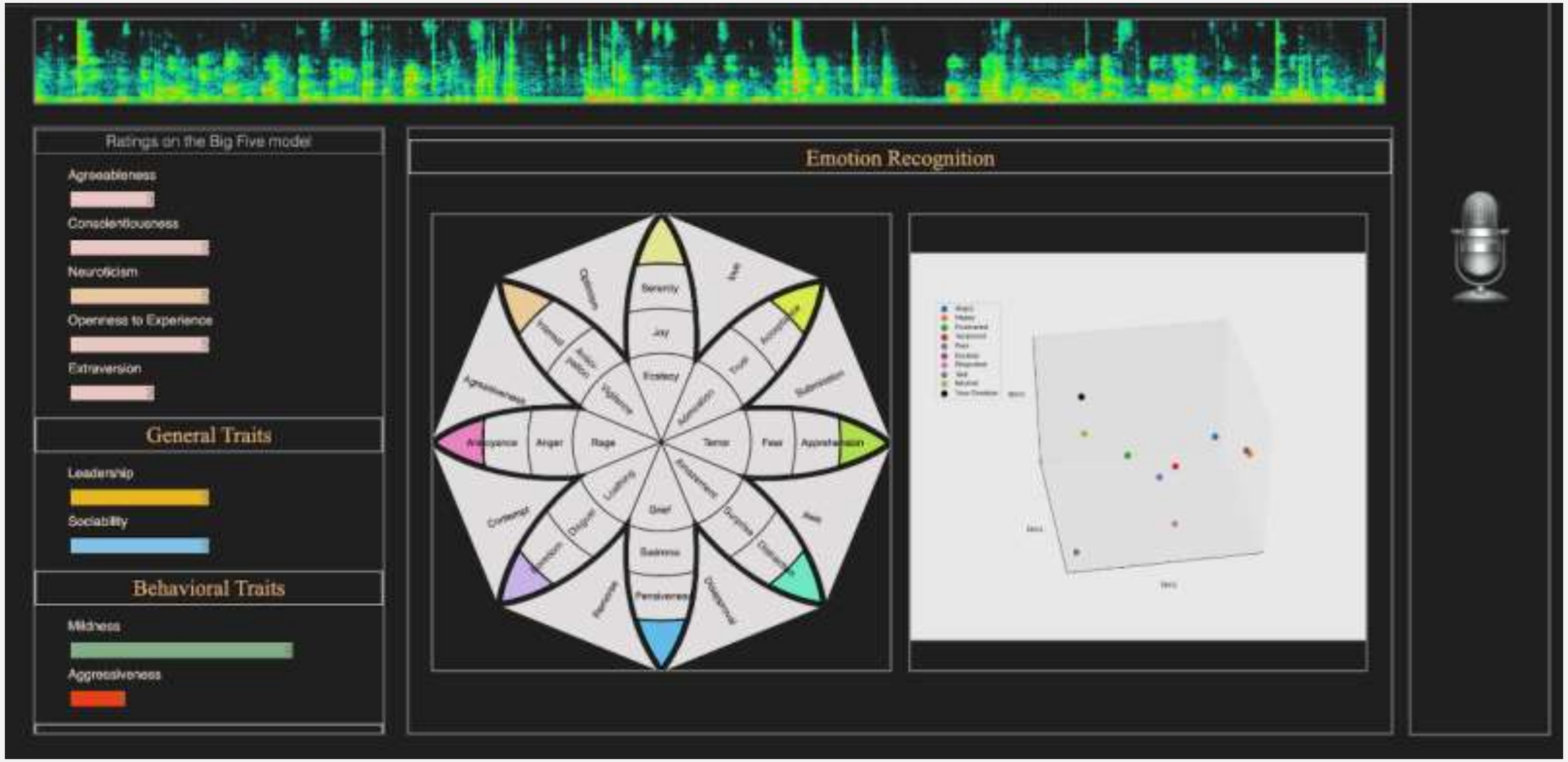
Methodology: Model II (multimodal embedding network)



Results

speaker specific embedding network (data: CMU-SER, task: regression)		
	MAE	RMSE
Baseline	0.26	0.31
Proposed Model	**0.16**	**0.20**

multimodal embedding network (data: IEMOCAP, task: classification)	
	Accuracy
Baseline	30.8%
Proposed Model	**53%**



Snippet of a lecture by Christopher Manning marked as "neutral" in terms of emotion by our demo

Perception

- Expression versus Perception
- Are emotions expressed through speech universal in how they are perceived?
- Studying gender differences in perception of emotion
- Challenges of crowdsourced data in relation to statistical analysis

Detecting gender differences in perception of emotion in crowdsourced data

Shahan Ali Memon
Hira Dhamyal
Carnegie Mellon University
Pittsburgh, USA
{samemon,hyd}@cs.cmu.edu

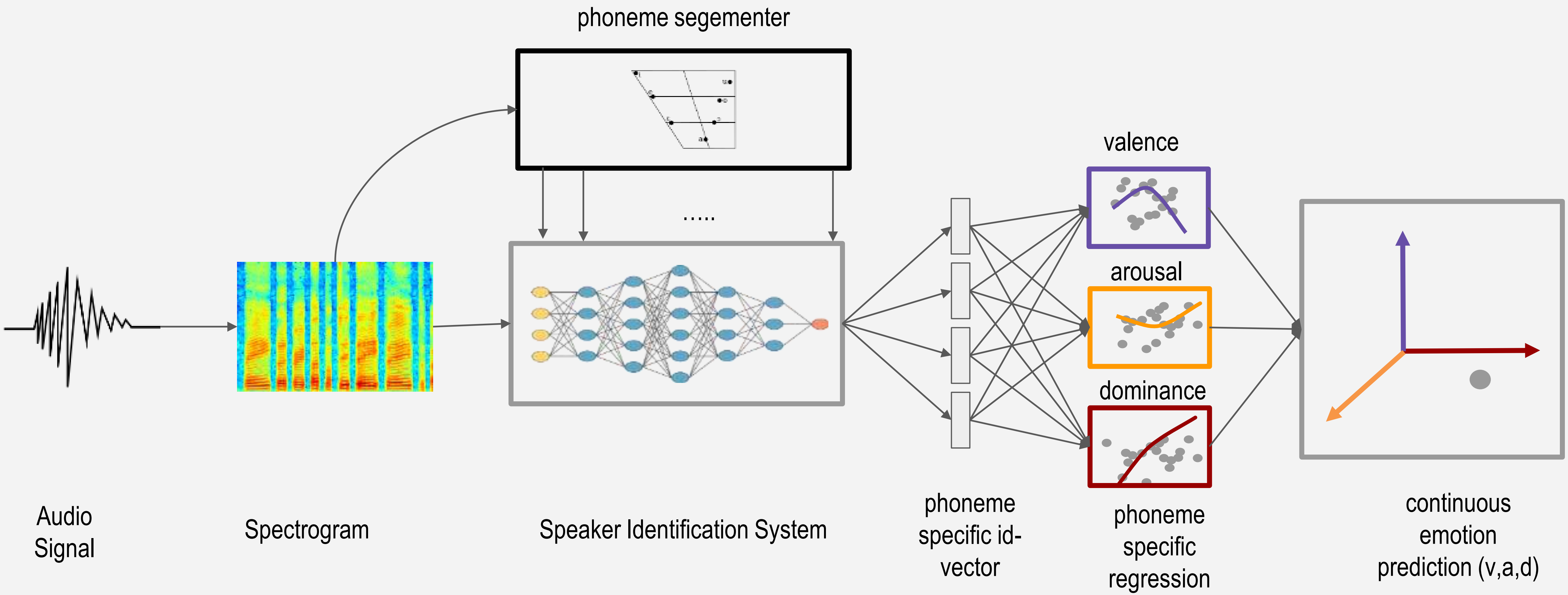
Oren Wright
Daniel Justice
Vijaykumar Palat
William Boler
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, USA
{owright,dljustice,vpalat,wmboler}@sei.cmu.edu

Bhiksha Raj
Rita Singh
Carnegie Mellon University
Pittsburgh, USA
{bhiksha,rsingh}@cs.cmu.edu

ABSTRACT
Do men and women perceive emotions differently? Popular convictions place women as more emotionally perceptive than men. Empirical findings, however, remain inconclusive. Most prior studies focus on visual modalities. In addition, almost all of the studies are limited to experiments within controlled environments. Generalizability and scalability of these studies has not been sufficiently established. In this paper, we study the differences in perception of emotion between genders from speech data in the wild, annotated through crowdsourcing. While we limit ourselves to a single modality (i.e. speech), our framework is applicable to studies of emotion perception from all such loosely annotated data in general. Our paper addresses multiple serious challenges related to making statistically viable conclusions from crowdsourced data. Overall, the contributions of this paper are two fold: a reliable novel framework for perceptual studies from crowdsourced data; and the demonstration of statistically significant differences in

and interpret emotions [30]. It is important to emphasize that perception, while correlated to emotional *expression*, is different from it. This distinction can be best explained by the modified version of Brunswik's lens model proposed in [37]. There are three main stages in this model: the encoding, the transmission, and the decoding of emotions. Encoding is the process where individual conveys their internal state by modifying their communicative channel. Decoding is the process where another individual makes an inference about the state of the first individual. The cues that are encoded and the cues that are decoded may differ based on the noise in the transmission. When studying expression, the focus is on how the emotions were encoded, and the primary subject of study is the encoder. On the other hand, perception deals with how the emotions were interpreted or decoded, and, hence, the focus is on the decoder. To create emotionally intelligent machines that can interact with humans, understanding how humans perceive emotions is a crucial first step. Not only

Future Work: Phonetic Embeddings



Questions?

Rita Singh (LTI, ECE)
(rsingh@cs.cmu.edu)

The Amazing Team

Hira Dhamyal (LTI)
Shahan Ali Memon (LTI)
Bhiksha Raj (LTI, ECE)
Richard Stern (ECE, LTI)
Oren Wright (SEI)
Daniel Justice (SEI)
Vijaykumar Palat (SEI)
William Boller (SEI)