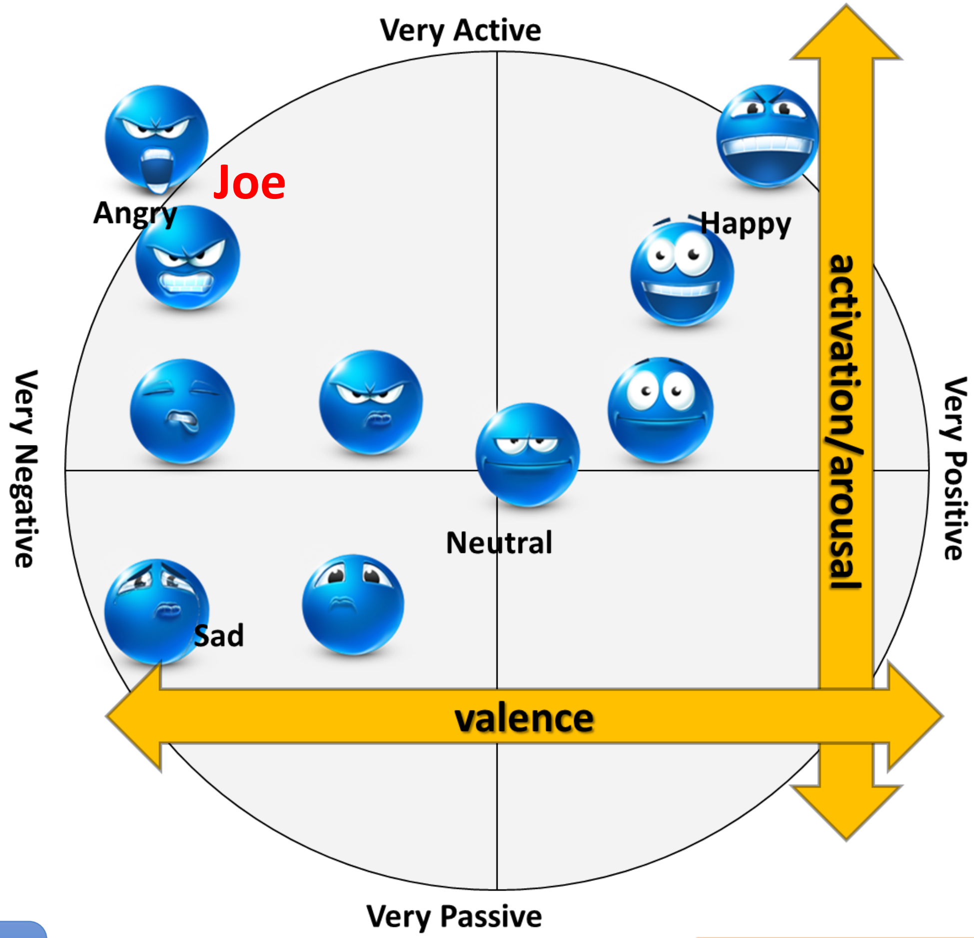
Exploring the Intersection Between Speaker Verification and Emotion Recognition

Michelle Bancroft, Reza Lotfian, John Hansen, and Carlos Busso

Background

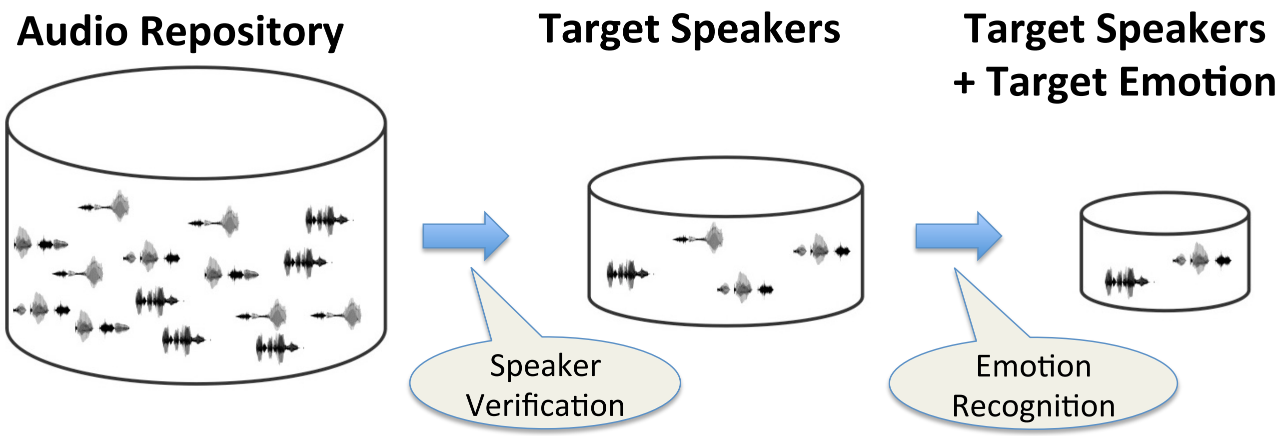
- **Recognizing emotional speech is an important research problem**
  - Security and defense
  - Quality control in customer service
  - Human computer interaction
- **Goal: Create effective methods for retrieving emotional data from known speakers**
  - Arousal
  - Valence
  - Dominance
- **Relevant problem for forensic analysis**

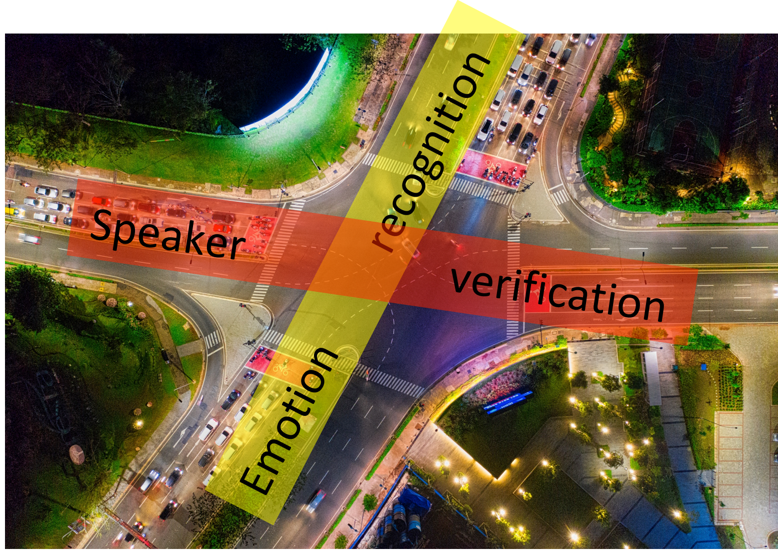


Retrieve angry sentences from "Joe"!

Proposed Work

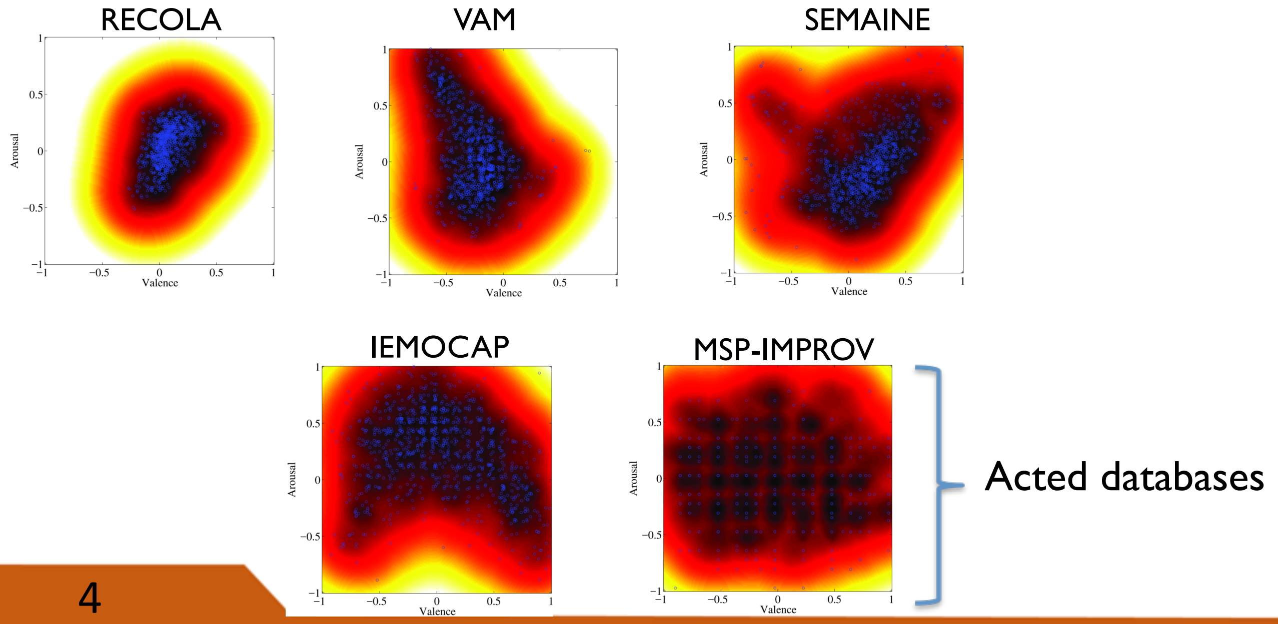
- **Combining emotional recognition with speaker verification tasks**
  - Use a speaker verification system to identify new sentences from target speakers
  - Use emotion recognition model to predict arousal, valence and dominance for sentences



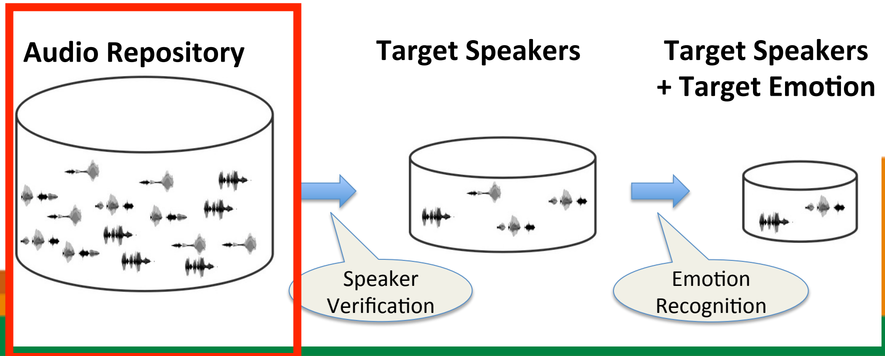


Infrastructure for the study

- Lack of naturalness
- Limited in size
- Limited number of speakers
- Unbalanced emotional content

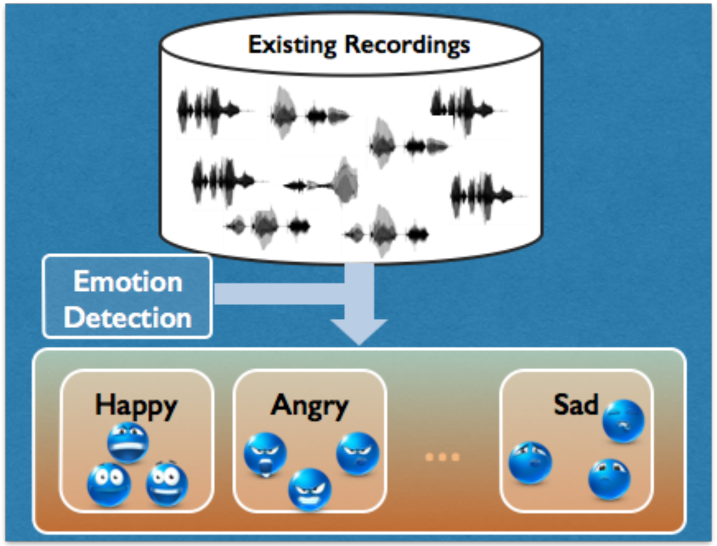


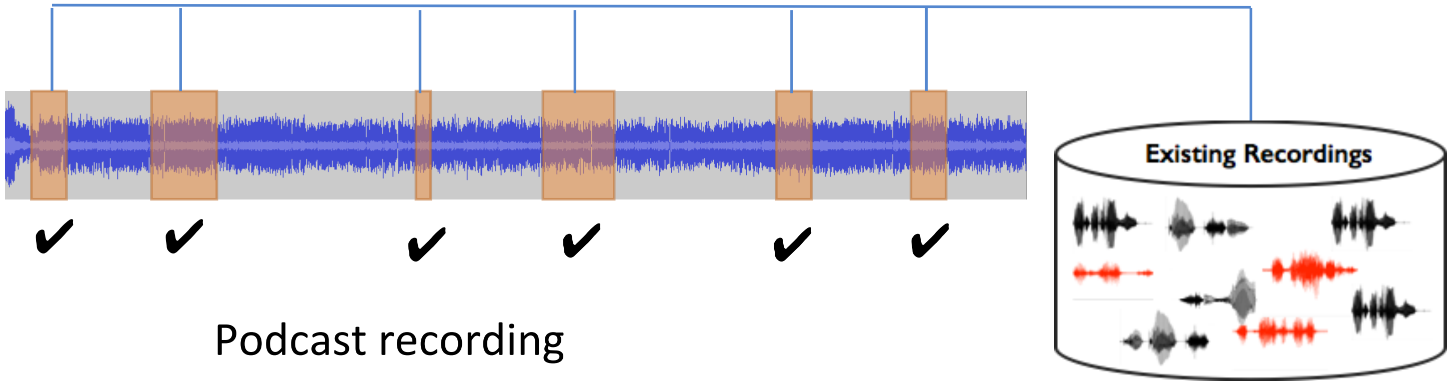
	Corpus Size	# Spkr.	Type	Lang.
IEMOCAP	12h26m	10	acted	English
MSP-IMPROV	9h35m	12	acted	English
CREMA-D	7,442 samples	91	acted	English
Chen Bimodal	9,900 samples	100	acted	English
Emo-DB	22m	10	acted	German
GEMEP	1,260 samples	10	acted	-
VAM-Audio	48m	47	spont.	German
TUM AVIC	10h23m	21	spont.	English
SEMAINE	6h21m	20	spont.	English
FAU-AIBO	9h12m	51	spont.	German
RECOLA	2h50m	46	spont.	French



The MSP-Podcast Database

- Use existing podcast recordings
- Divide into speaker turns
- Emotion retrieval to balance the emotional content
- Annotate using crowdsourcing framework

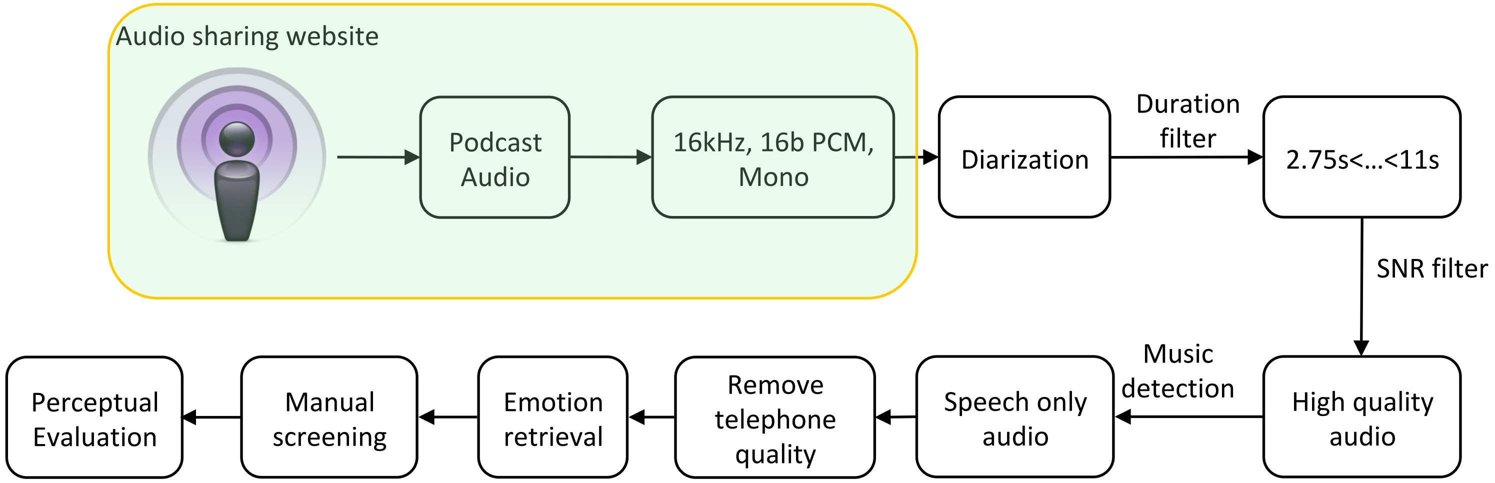




Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," IEEE Transactions on Affective Computing, vol. To appear, 2018.

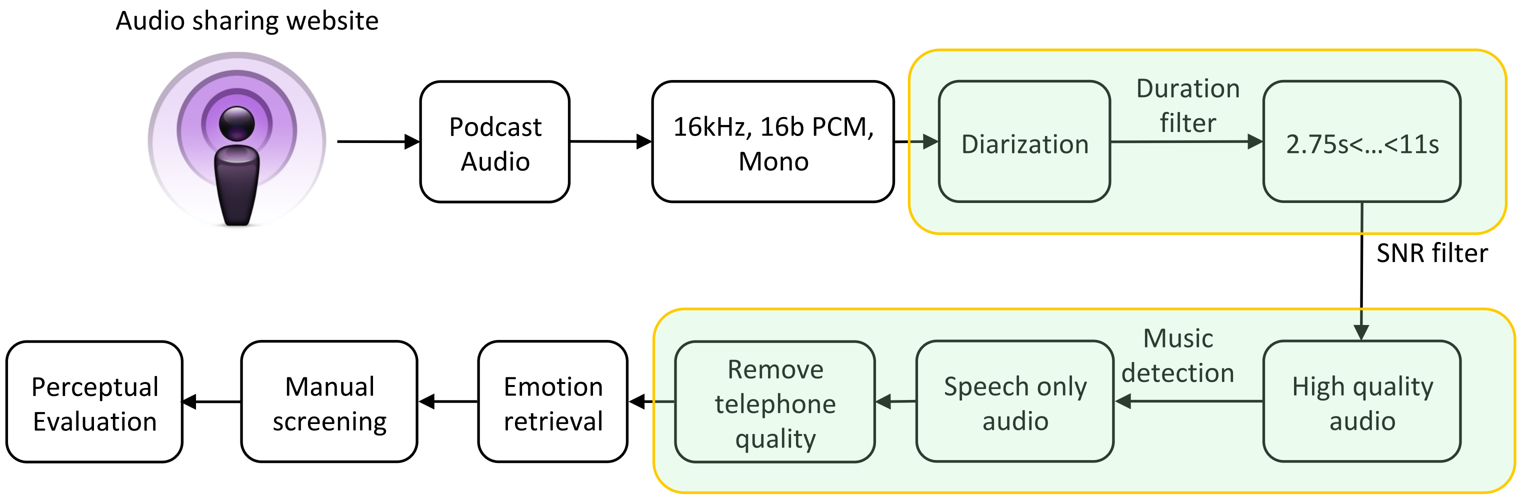
MSP-Podcast Corpus

- **Collection of audio recordings (Podcasts)**
  - Naturalness and the diversity of emotions
  - Creative Commons copyright licenses
  - Interviews, talk shows, news, discussions, education, storytelling, comedy, science, technology, politics, economics, business, arts, culture, sports



MSP-Podcast Corpus

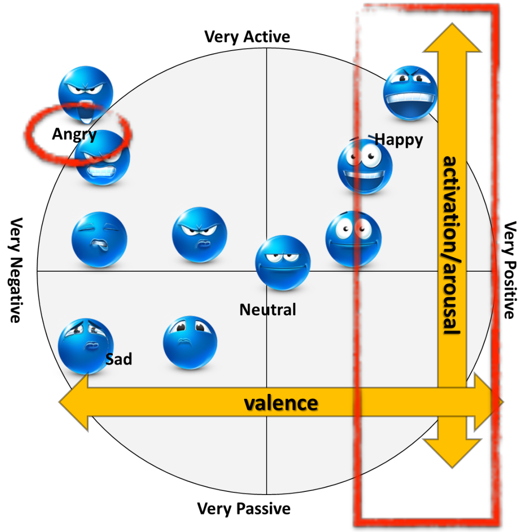
- **Automatic speaker diarization**
  - Single speaker segments
  - High SNR, no music, no phone quality
- **Duration:**
  - Longer than 2.75sec: Long enough for annotators + extract reliable features
  - Shorter than 11sec: Emotion content not changing significantly

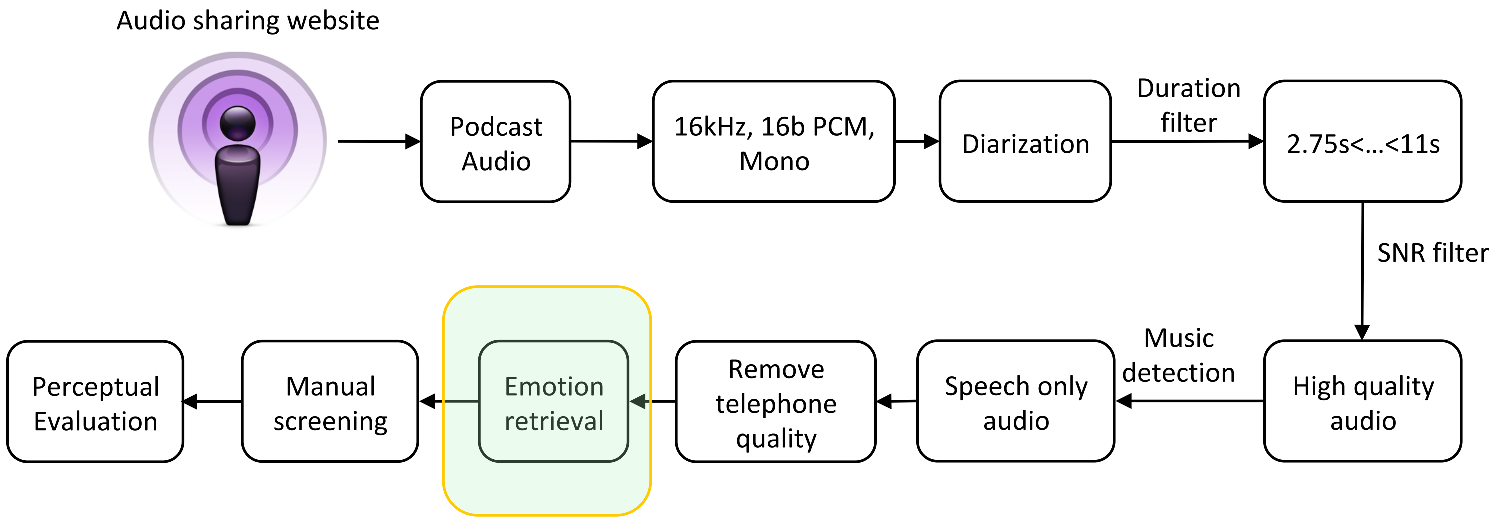


"The speaker attribution intelligent service," http://diarizationservice3.cloudapp.net/

MSP-Podcast Corpus

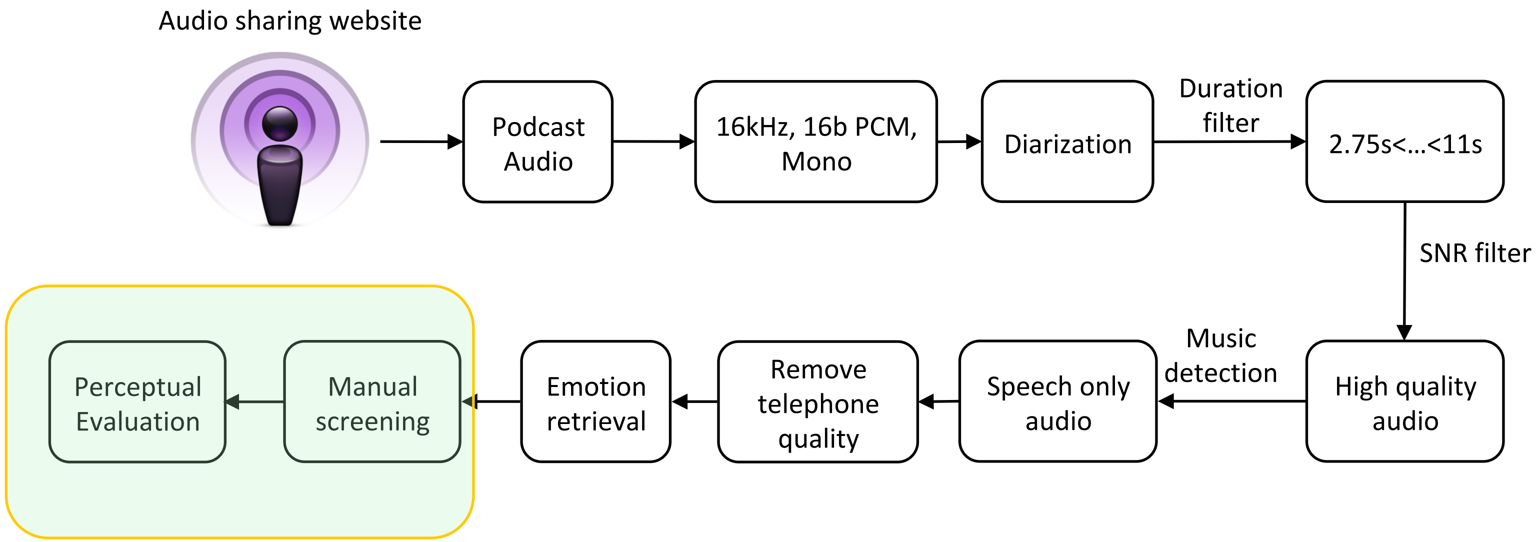
- **Retrieve samples that convey desired emotion**
  - Developing and optimizing different machine learning framework using existing databases
  - Balance the emotional content



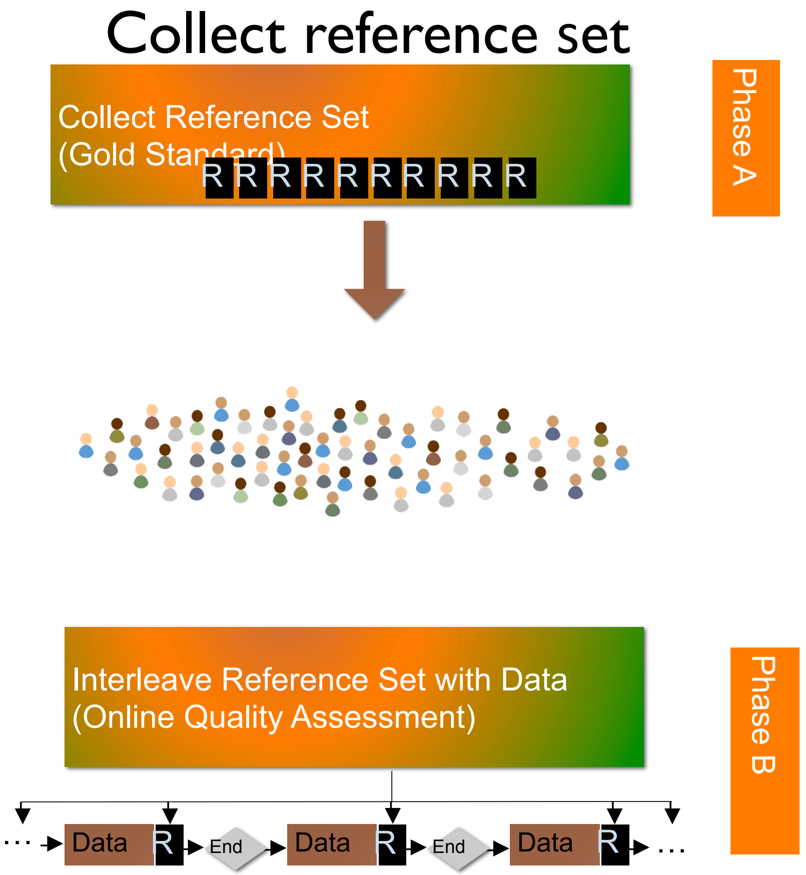


MSP-Podcast Corpus

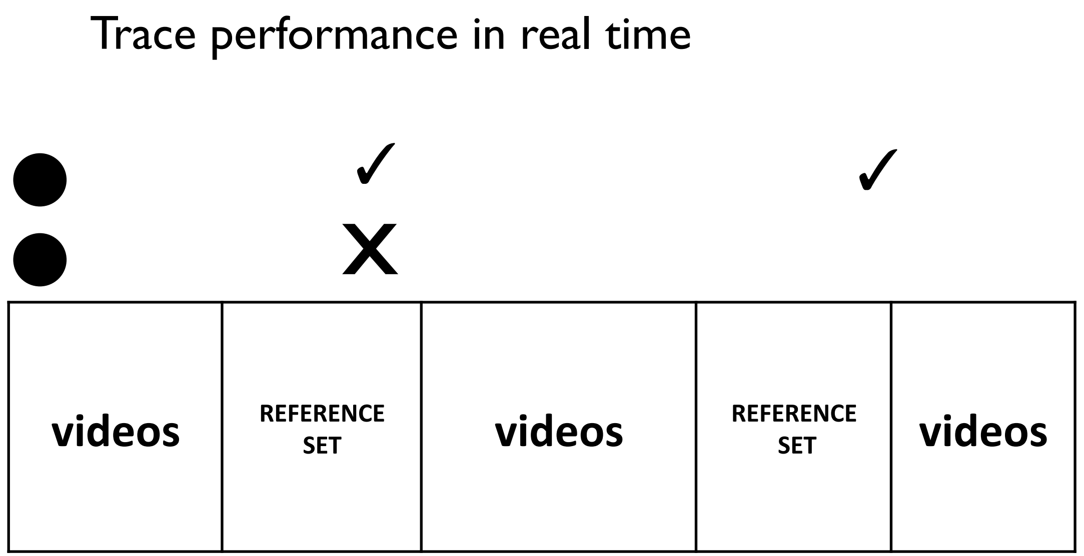
- **Perceptual evaluation**
  - Subjective annotation is costly
  - screening only retrieved samples before uploading for annotations



Perceptual Evaluation



- Use Amazon Mechanical Turk Crowdsourcing
- Verify if a worker is spamming in real time

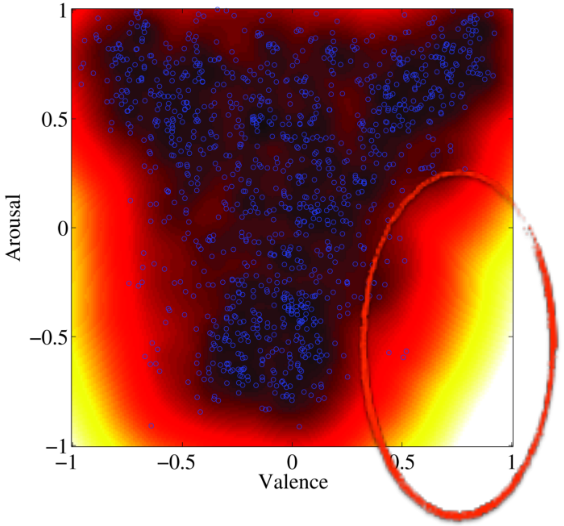


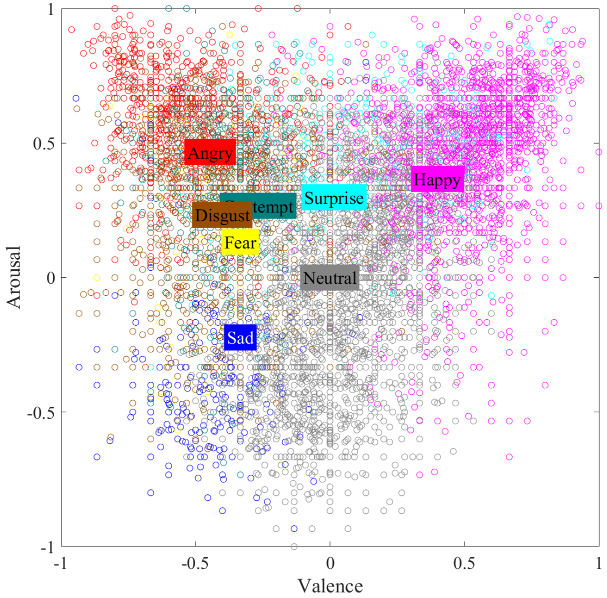
Alec Burmania, Srinivas Parthasarathy, and Carlos Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," IEEE Transactions on Affective Computing, vol. 7, no. 4, pp. 374-388, October-December 2016.

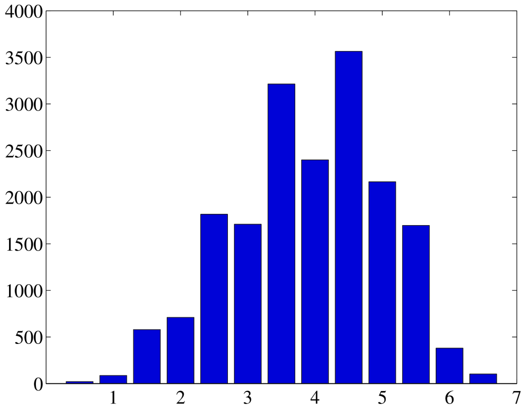
MSP-Podcast corpus version 1.0

With emotion labels:
20,032 sentences
(38h, 57m)

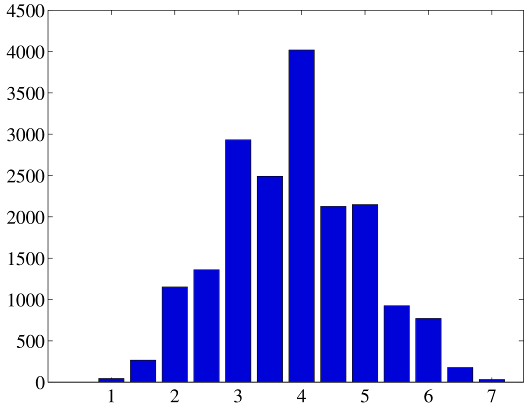
Segmented turns
152,975 sentences over 1,000 podcasts



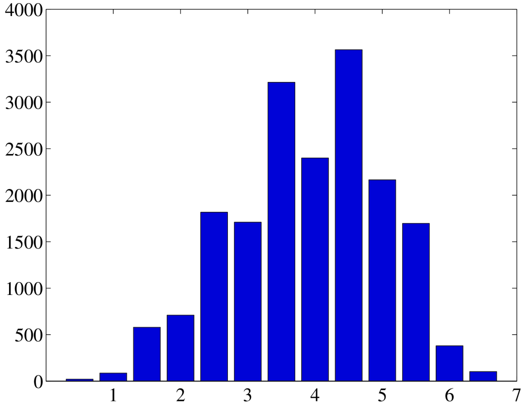




Arousal



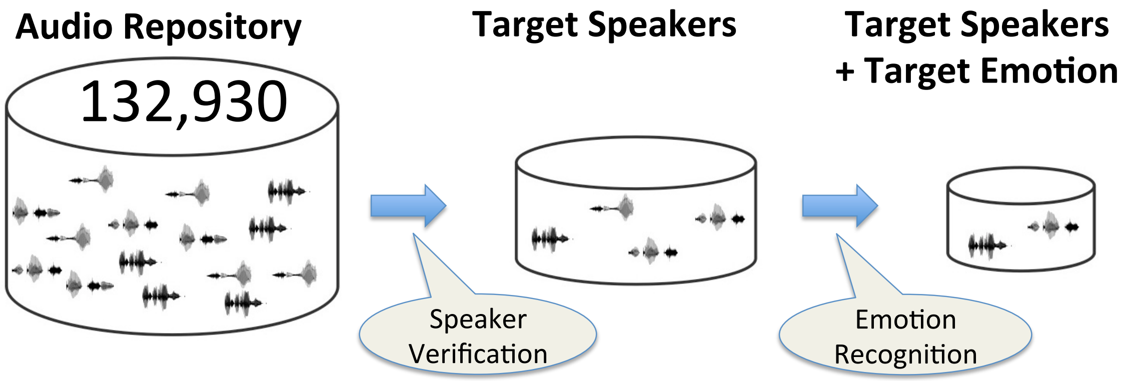
Valence



Dominance

The MSP-Podcast Corpus

- **Manual annotations of data with emotional labels**
  - 16,015 out of 20,032 speaking turns
- **Target speakers**
  - 146 speakers with at least 150s
- **Audio repository: segments without emotional labels**
  - 132,930 speaking turns
  - They include speech from target speakers



Ideal infrastructure with labeled data with both emotion and speaker information, but also a large unlabeled speech repository for the retrieval task

Speaker Verification

- **Speaker verification toolkit**
- **i-Vector Modeling**

$$\underbrace{M}_{\substack{\text{GMM supervector}\\ \text{After MAP adaptation}}} = \underbrace{m}_{\substack{\text{Universal}\\ \text{mean vector}}} + \underbrace{T}_{\substack{\text{Total variability}\\ \text{matrix}}}\overbrace{\mathbf{x}}^{\text{i-vector}}$$

- **Mean normalized Probabilistic linear discriminant analysis (PLDA)**

Model for speaker *j*

$$\bar{x}_j = \frac{1}{D}\sum_{d=1}^{D} x^{(d)}$$

- **The log-likelihood ratio (LLR)**
  - The LLR computes the ratio between two alternative hypothesis:

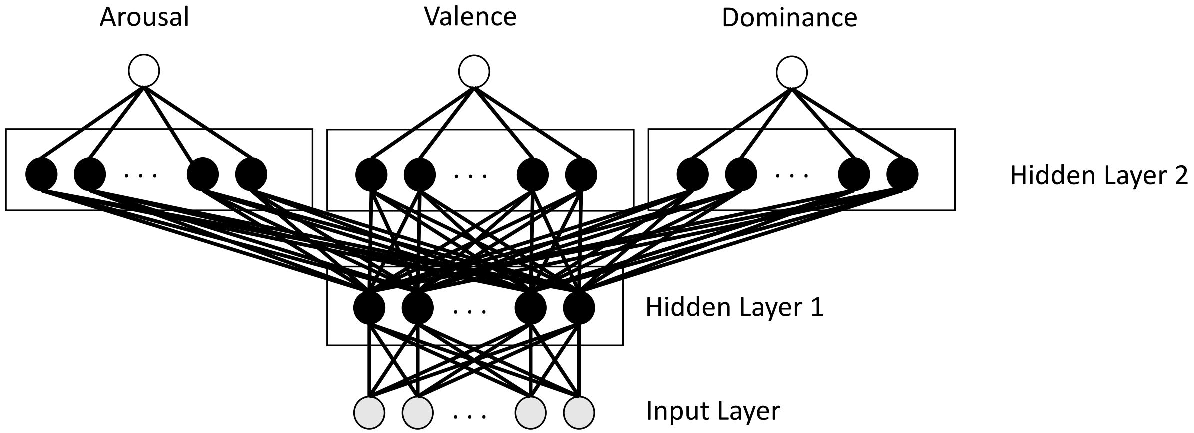
H_1 : $\mathbf{x_1}$ and $\mathbf{x_2}$ same speaker

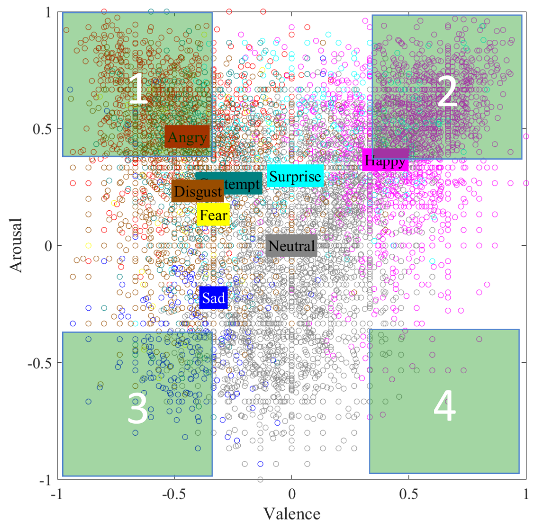
H_2 : $\mathbf{x_1}$ and $\mathbf{x_2}$ different speaker

$$r = \ln \frac{\rho\left(x_1, x_2 | H_1\right)}{\rho\left(x_1 | H_0\right) \cdot \rho\left(x_2 | H_0\right)}$$

Emotion recognition

- **Multitask learning used to jointly predict arousal, valence and dominance [Parthasarathy and Busso, 2017]**
  - Two layers, training with CCC
  - 6,373 acoustic features (OpenSmile)
- **Target Regions**
  - Region 1: low v, high a
  - Region 2: high v, high a
  - Region 3: low v, low a
  - Region 4: high v, low a

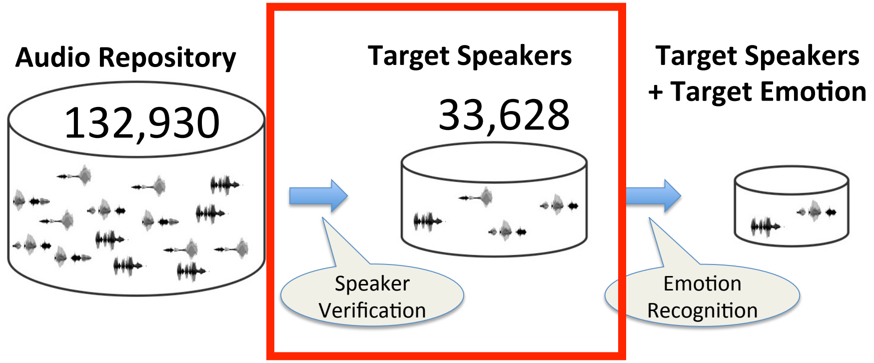


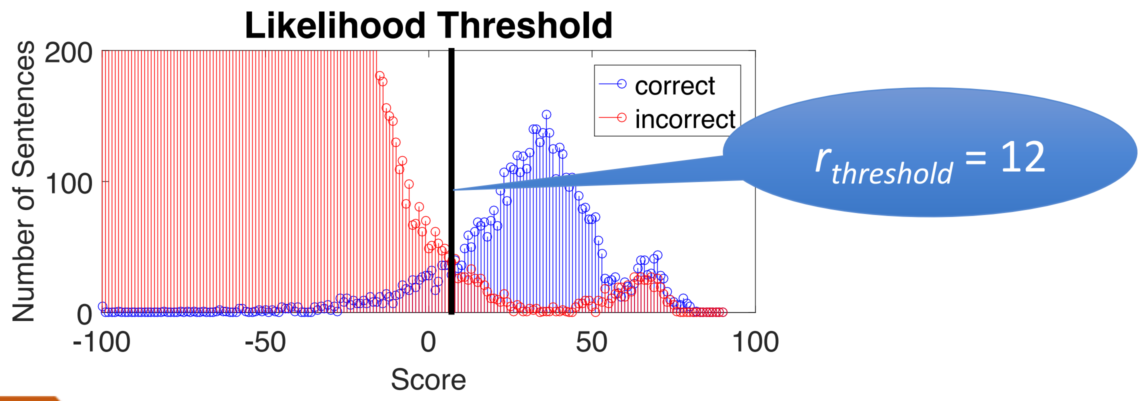


Srinivas Parthasarathy and Carlos Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in Interspeech 2017, Stockholm, Sweden, August 2017, pp. 1103-1107.

Experimental Results

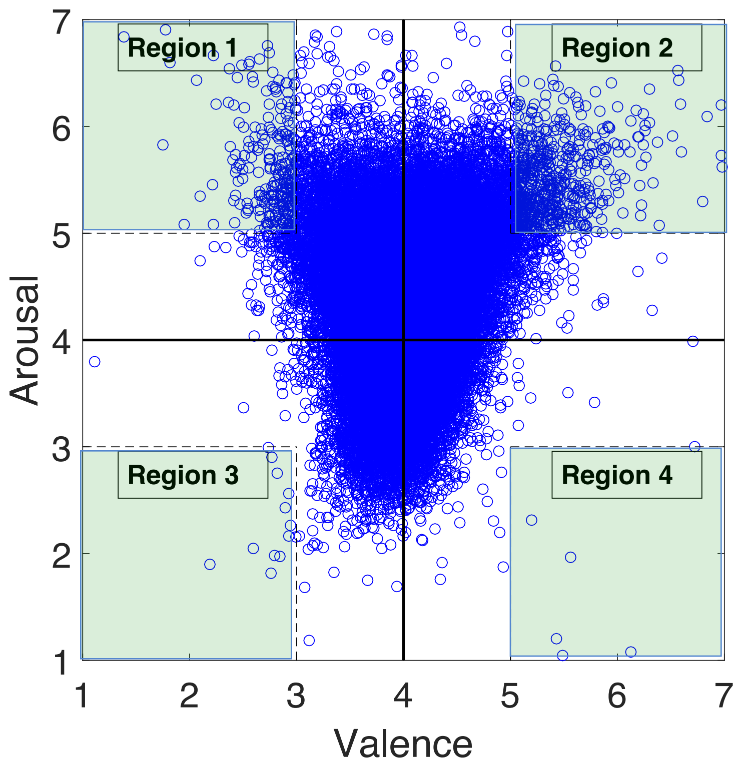
- **Speaker verification**
  - 146 target speakers
  - Train models with 150s per speaker
  - Test on the rest of the data from the 146 speakers
  - Threshold greater than $r_{threshold} = 12$
    - 33,628 unique segments identified

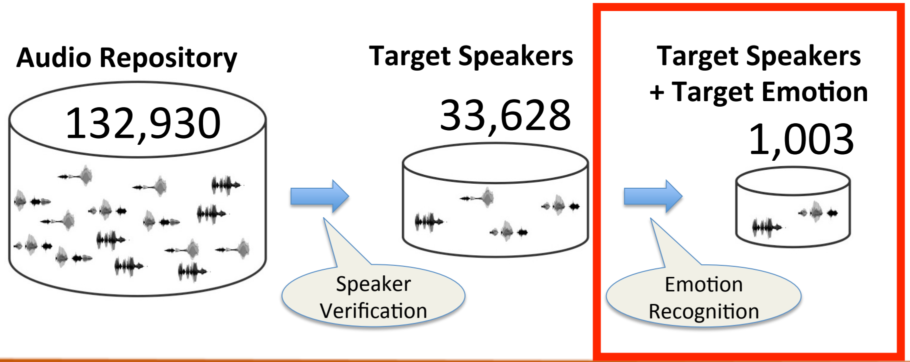




Experimental Results

- **Emotion recognition**
  - We analysis 33,628 segments using our multitask learning framework
  - 1,003 unique segments in the target regions
    - Region 1: 294
    - Region 2: 681
    - Region 3: 15
    - Region 4: 13





Analysis of Results

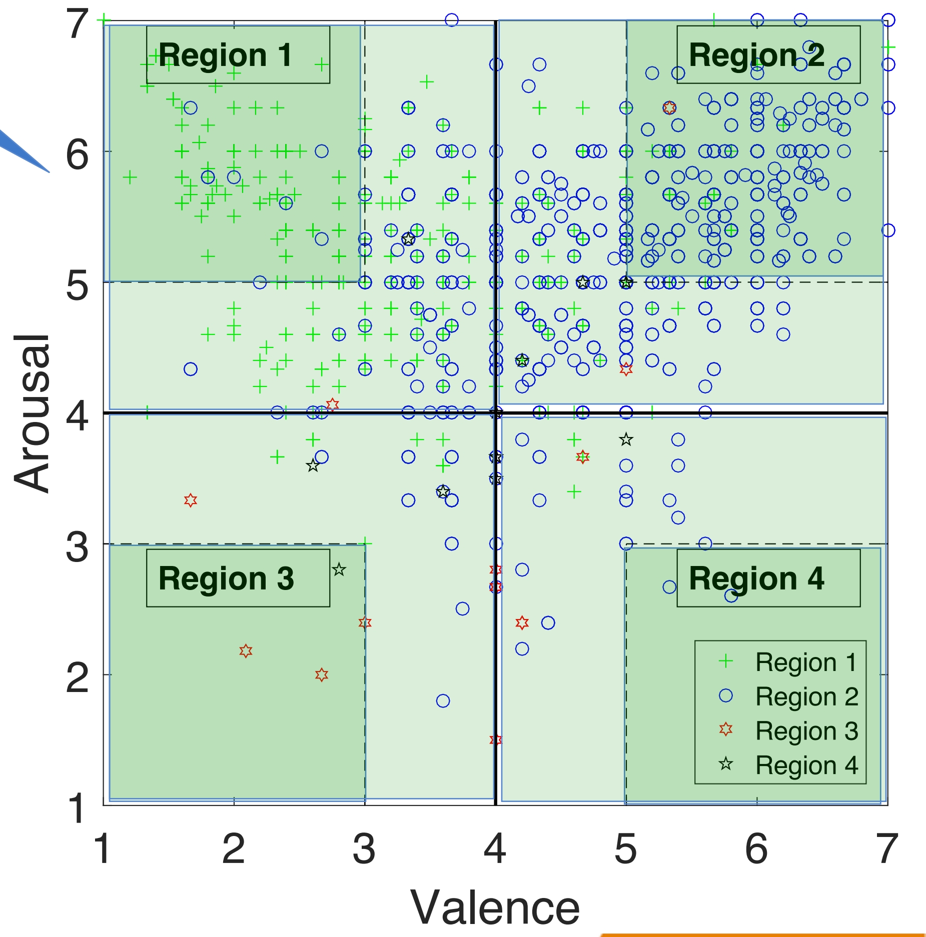
$ccc_{arousal}$ =0.532
$ccc_{valence}$=0.364

- **Emotion recognition**
  - Annotate segments with crowdsourcing
  - Precision rate

<u>Target region (45.8%)</u>
- Region 1: 37.5%
- Region 2: 50.6 %
- Region 3: 23.1%
- Region 4: 0%

<u>Target quadrant (77.4%)</u>
- Region 1: 73.3%
- Region 2: 80.2%
- Region 3: 61.5%
- Region 4: 36.4%

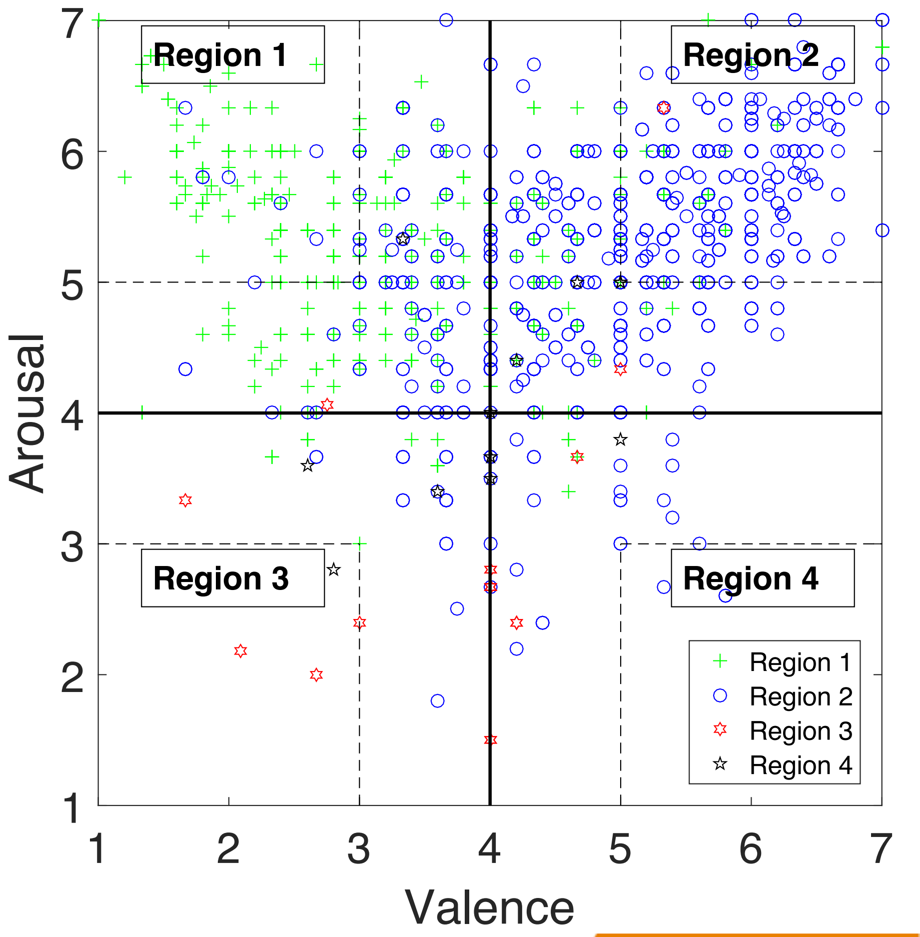
Few samples in these regions



Analysis of Results

- **Speaker verification**
  - 1,401 speaker verification evaluations satisfy ratio
    - A segments can have more than one speaker
  - We annotate the speaker identity of the 1,003 turns
    - 80.9% accuracy (1,135 evaluations correct)
    - Emotional speech challenges speaker verification systems
  - Speaker verification performance per target region
    - Region 1: 81.6%
    - Region 2: 80.9%
    - Region 3: 80.0%
    - Region 4: 33.3%

Few samples in these regions



Conclusions

- **Speaker verification and emotion recognition systems successfully combined**
- **This study**
  - built the infrastructure to pursue this research direction
  - revealed the limitations of speaker verification tasks in the presence of emotional speech
- **Future Work**
  - Further improve emotional recognition and speaker verification systems
  - Compensation schemes for speaker verification systems in the presence of emotion

Thank you

- **This work was funded by NSF CAREER award IIS-1453781**

Interested on our research?
msp.utdallas.edu