MSP-Face Corpus: A Natural Audiovisual Emotional Database

Andrea Vidal, Ali Salman,
Wei-Cheng Lin, Carlos Busso

UTD CRSS

THE UNIVERSITY OF TEXAS AT DALLAS

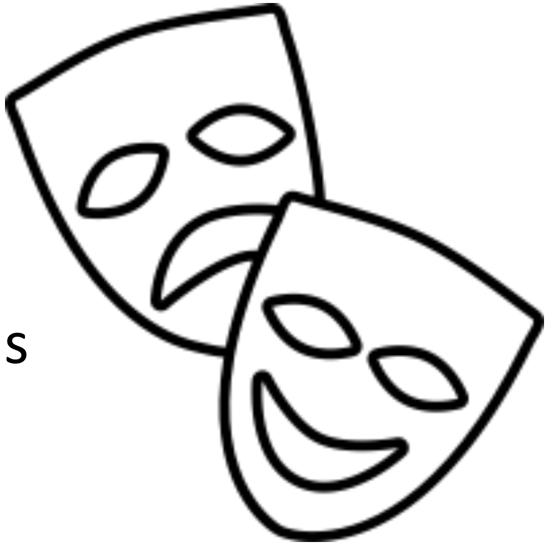
UT Dallas MSP Multimodal Signal Processing Laboratory

Outline

1. **Motivation**
2. **MSP-Face corpus**
   1. Description
   2. Annotation process
   3. Emotional content
   4. Baselines
3. **Conclusions**

How do people express their emotions?

- **Multimodal emotional databases**
  - Acted
  - Emotion response is elicited
  - Problem: It is not how people show and express their emotions



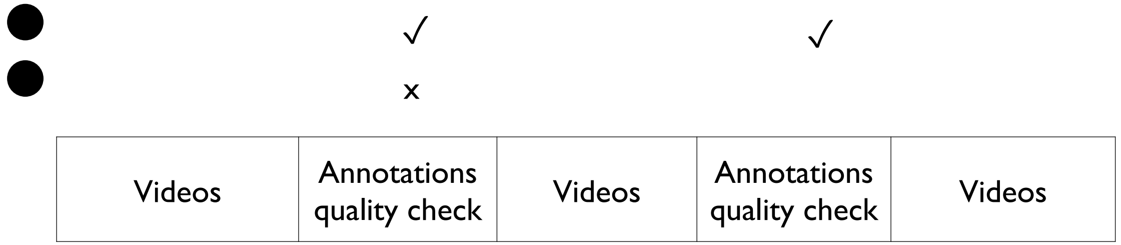
- **MSP-Face corpus**
  - Natural and spontaneous recordings
  - People talk in front of the camera
  - Multiple participants, broad range of emotions
  - Emotions labels obtained via crowdsourcing

MSP-Face corpus

- **Collection of online videos**
  - Frontal face
  - No background music
  - Single speaker
  - Video segments 3-10 seconds
- **Speakers**
  - Number of speakers: 491
  - Diversity of speakers
- **Duration of the database**
  - ≈70 hrs (27,325 video segments)
    - Labeled: ≈24.7 hrs (9,370 video segments)
    - Unlabeled: ≈46 hrs (17,955 video segments)

MSP-Face corpus annotation

- **Annotation Process**
  - Amazon Mechanical Turk (AMT) crowdsourcing
  - Qualified annotators
    - Live in The United States
    - More than 100 tasks accepted
    - More than 95% acceptance rate of tasks
  - At least 5 annotations per video
  - A quality check of the annotations is performed during the annotation process



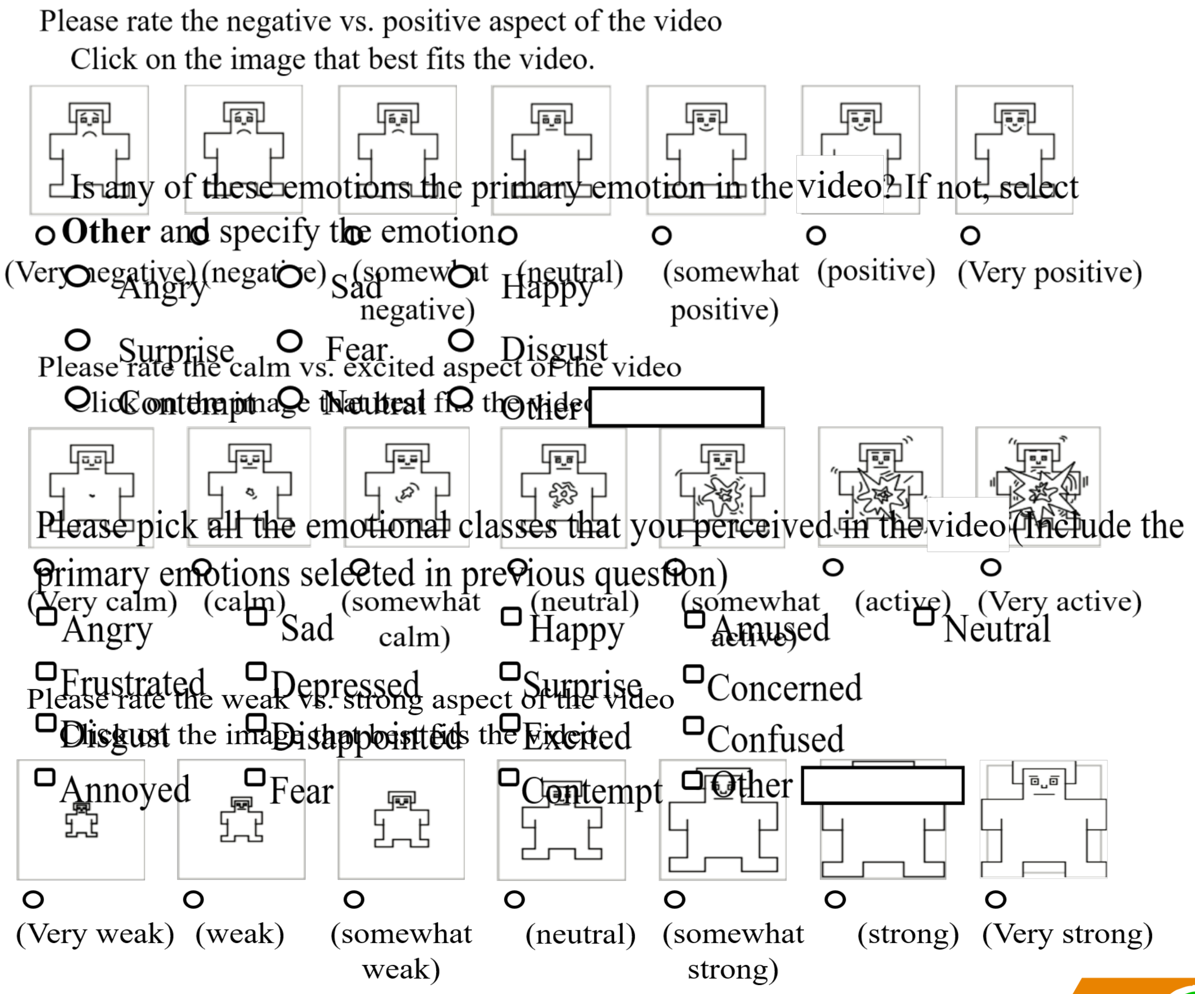
● ✓ ✓

● x

Videos	Annotations quality check	Videos	Annotations quality check	Videos

MSP-Face corpus annotation

- **Emotions**
  - Categorical emotions
    - Primary emotions
    - Secondary emotions
  - Attributes-based descriptors
    - Valence
    - Arousal
    - Dominance



Please rate the negative vs. positive aspect of the video
Click on the image that best fits the video.

○ (Very negative) ○ (negative) ○ (somewhat negative) ○ (neutral) ○ (somewhat positive) ○ (positive) ○ (Very positive)

Is any of these emotions the primary emotion in the video? If not, select **Other** and specify the emotion

○ Angry ○ Sad ○ Happy
○ Surprise ○ Fear ○ Disgust
○ Contempt ○ Neutral ○ Other

Please rate the calm vs. excited aspect of the video
Click on the image that best fits the video

○ (Very calm) ○ (calm) ○ (somewhat calm) ○ (neutral) ○ (somewhat active) ○ (active) ○ (Very active)

Please pick all the emotional classes that you perceived in the video (Include the primary emotions selected in previous question)

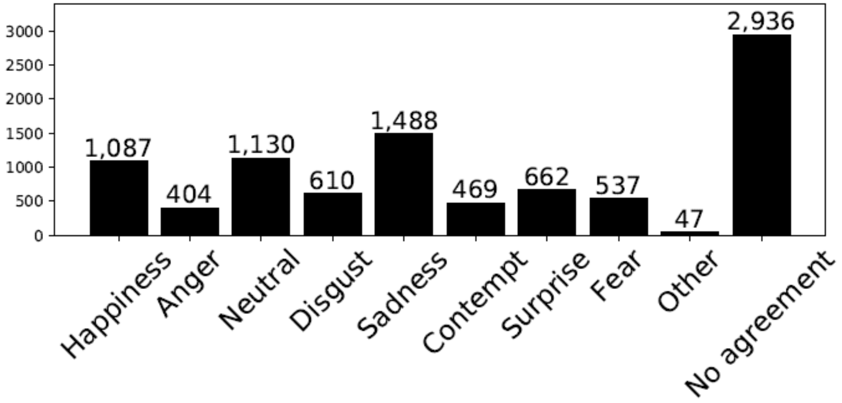
□ Angry □ Sad □ Happy □ Amused □ Neutral
□ Frustrated □ Depressed □ Surprise □ Concerned
□ Disgust □ Disappointed □ Excited □ Confused
□ Annoyed □ Fear □ Contempt □ Other

Please rate the weak vs. strong aspect of the video
Click on the image that best fits the video

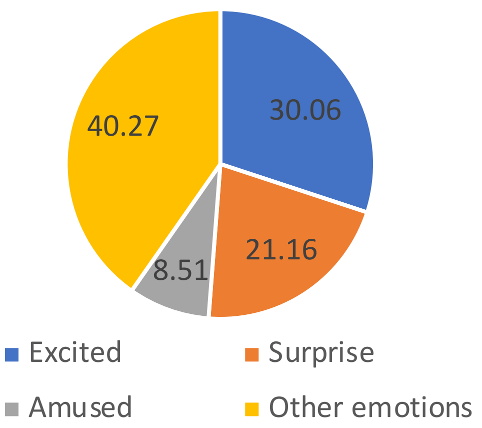
○ (Very weak) ○ (weak) ○ (somewhat weak) ○ (neutral) ○ (somewhat strong) ○ (strong) ○ (Very strong)

MSP-Face corpus emotional content

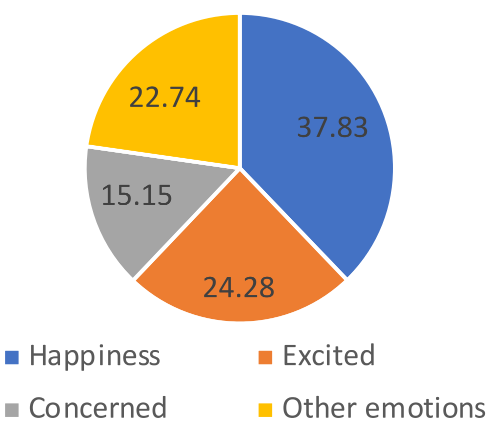
- **Primary categorical emotions**
  - Eight emotions
  - Consensus label is set by using plurality
  - All emotions have more than 400 samples
- **Secondary categorical emotions**
  - Give us a more understanding on the emotional content
  - Each primary emotion has assigned 1.12 secondary emotions



Happiness

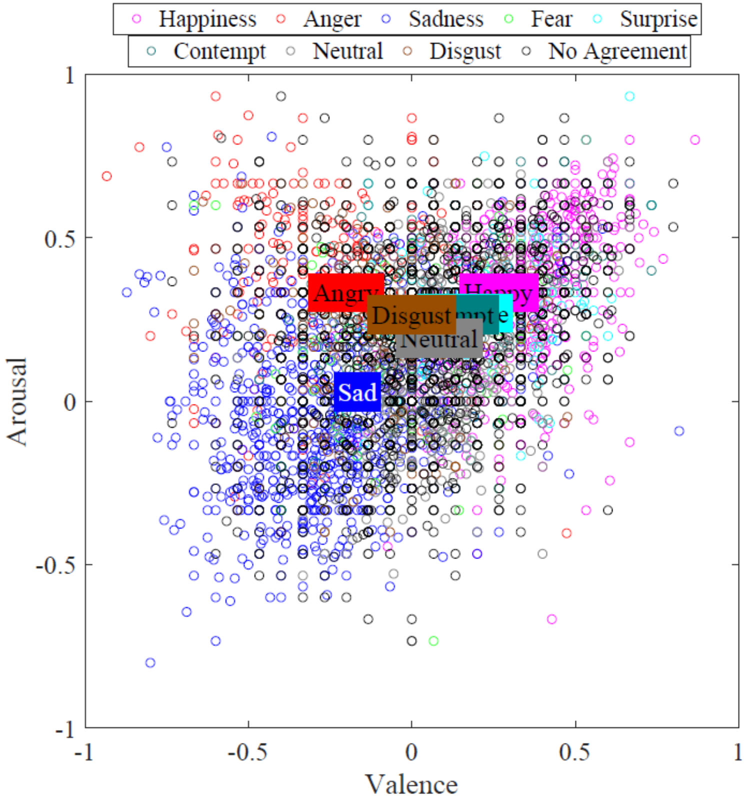


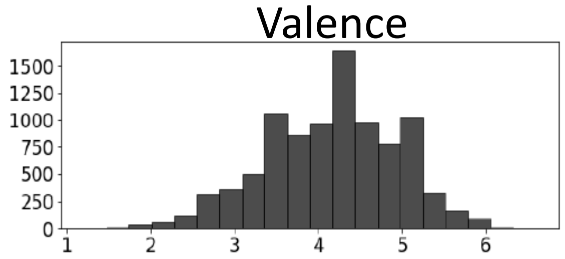
Surprise

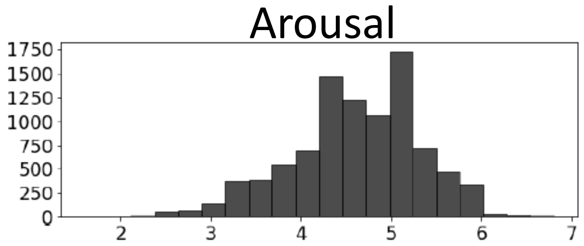


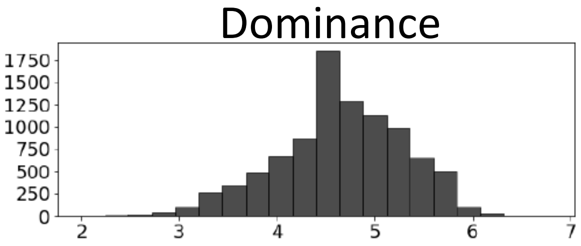
MSP-Face corpus emotional content

- **Attribute-based descriptors**
  - Balanced distributions
  - Broad range of emotional content
  - Emotional content covers most of the arousal-valence space
  - Variability of an emotion



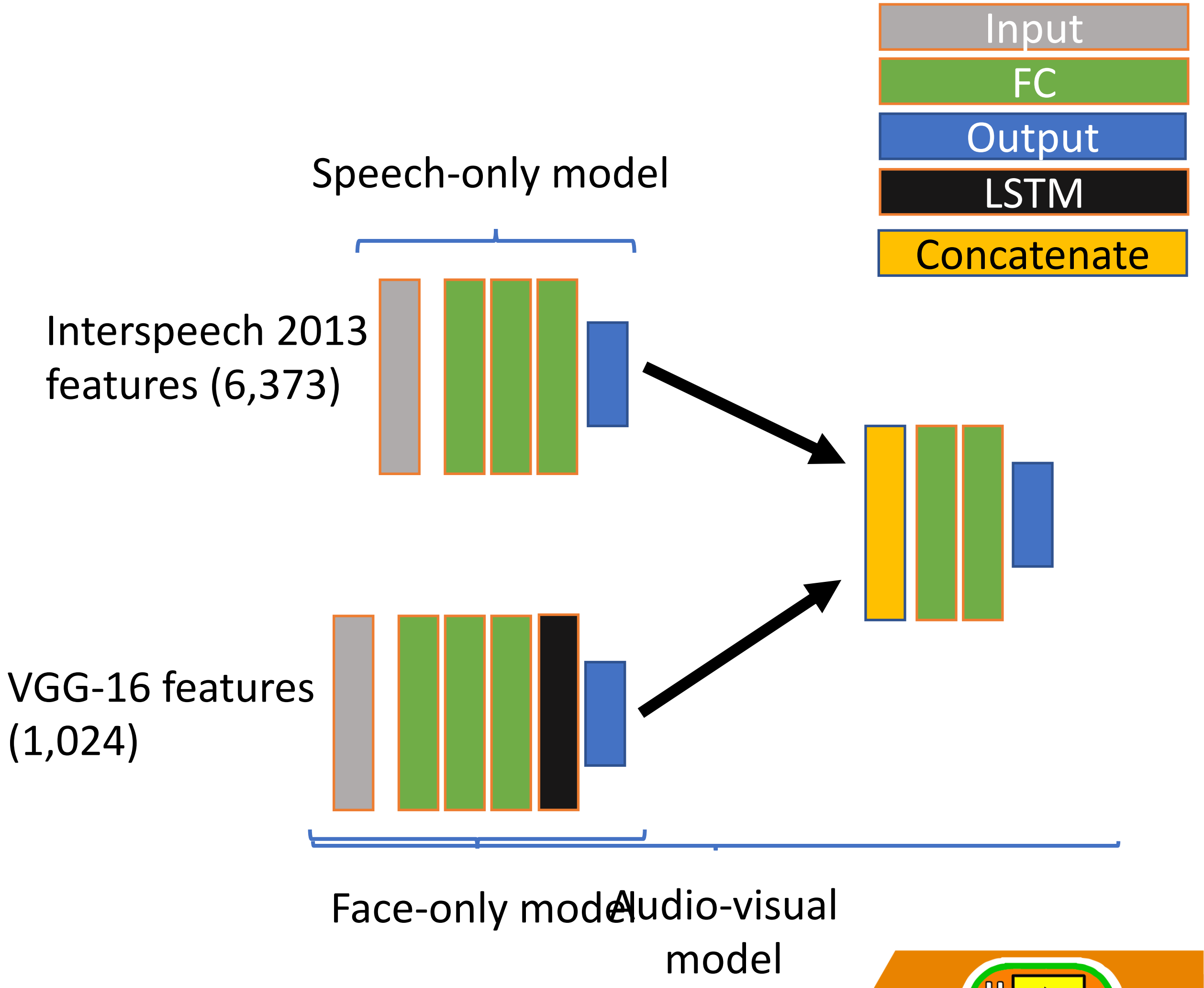






Emotion recognition experiments

- **Speech-only model**
  - Input: Interspeech 2013 features (6,373)
- **Face-only model**
  - Input: VGG-16 features (1,024)
- **Audio-visual model**
  - Input: Embeddings from the previous models.
- **Output of the models**
  - Categorical emotion for classification
  - Attribute-based descriptors for regression using Concordance Correlation Coefficient (CCC) as loss function.



Emotion recognition experiments

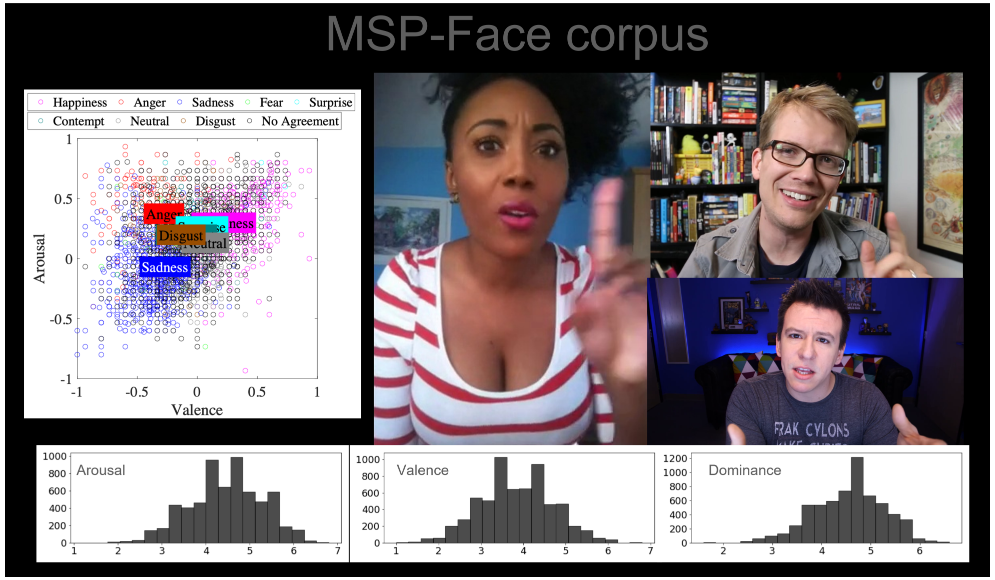
	Speech-only	Face-only	Audiovisual
Arousal-CCC	0.3794	0.2065	0.3961
Valence-CCC	0.2924	0.2677	0.3453
Dominance-CCC	0.3390	0.2085	0.3430
5 class F1-score (macro)	0.2835	0.3027	0.3010
5 class F1-score (micro)	0.3599	0.3494	0.3641
8 class F1-score (macro)	0.1629	0.1308	0.1690
8 class F1-score (micro)	0.2637	0.3161	0.2710

- **Speech modality regression results outperform face modality**
- **Classification results are comparable between the modalities**
- **In overall, the fusion of the modalities improves the performance of each modality separately**

Conclusions

MSP-Face corpus

- Database of natural and spontaneous recordings
- Speaker diversity
- ≈70 hrs of audiovisual database
  - ≈24.7 hrs (labeled)
  - ≈46 hrs (unlabeled)
- Unlabeled part is set to explore unsupervised methods



MSP-Face corpus applications

- Emotion recognition
- Generating visual agents with expressive behaviors

MSP-Face corpus available

- Annotations
- Source code of baselines
- Video links
- https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Face.html