Dynamic versus Static Facial Expressions in the Presence of Speech

Ali N. Salman and Carlos Busso

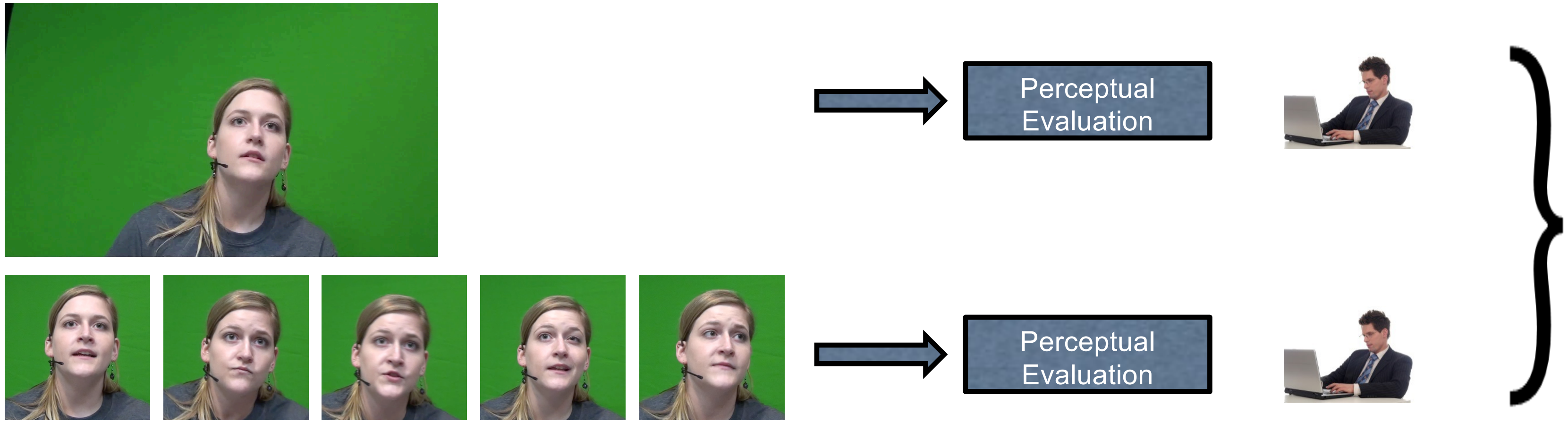
NEC \Orchestrating a brighter world
UTD CRSS
THE UNIVERSITY OF TEXAS AT DALLAS
UT Dallas MSP Multimodal Signal Processing Laboratory

Why study emotion ?

- **Emotions play a crucial role in human interaction**
  - Emotional (vs. cognitive) reasoning
  - Emotion is reflected in our body
  - Our emotions change the minds of others
  - People rely on emotion for making decisions
- **Knowing the user's emotional state should help to adjust system performance**
- **User can be more engaged and have a more effective interaction with the system**

Facial expression analysis

- **Study contextual information, including lexical content in the expression of emotion**



Similar emotions?

Is the emotion in isolated frames in a video a good representation of the emotional perception of the entire video?

MSP-IMPROV corpus

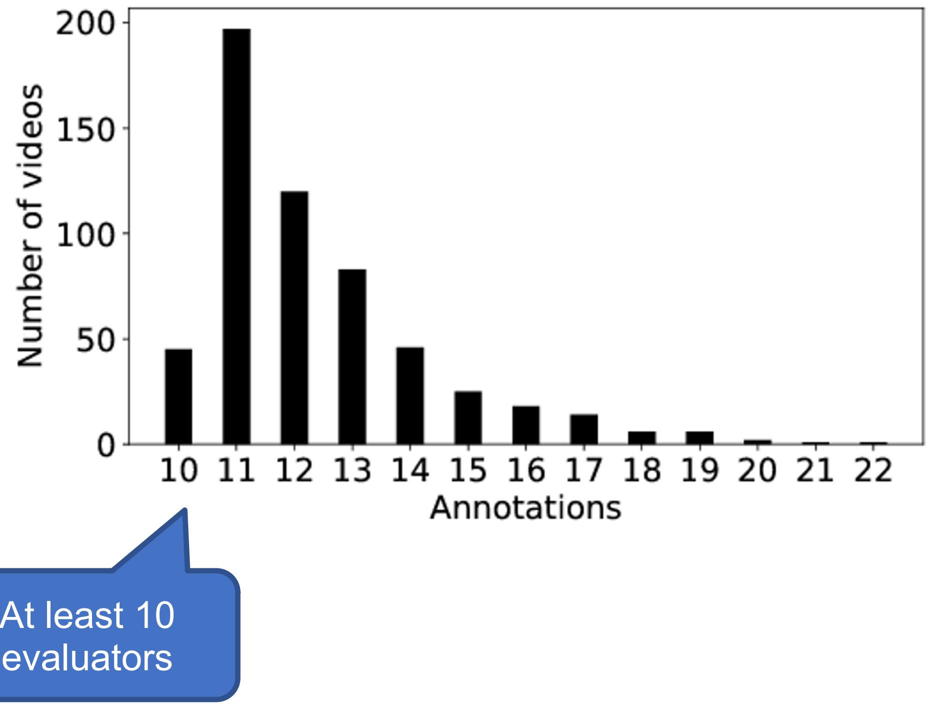
- **Collection of the Corpus**
  - 6 dyadic session pairing one male and one female actor (6 female , 6 male total)
  - Collected in 13ft x 13ft ASHA CRSS sound booth
  - High resolution digital cameras recording both actors (1440x1080 pixels)
  - Audio recorded with 48khz and 32 bit PCM and TASCAM US-1641 interface
  - Green Screen and LED lighting behind actors

Setup

C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," IEEE Transactions on Affective Computing, vol. 8, no. 1, pp. 119-130 January-March 2017.

MSP-IMPROV Corpus (Cont.)

- **The key feature of this corpus is annotations with different conditions:**
  - audiovisual presentations
  - audio only presentations
  - video only presentations

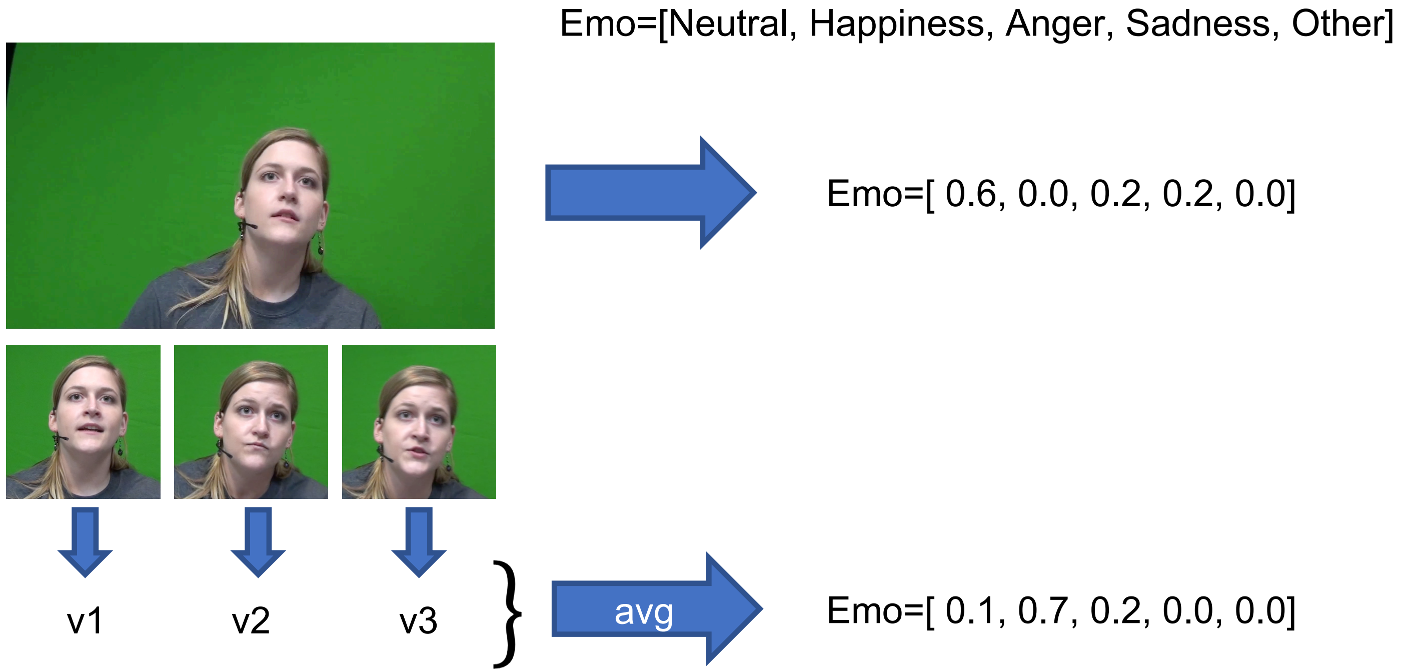


- **Annotations**
  - Categorical based annotations
    - Happiness, anger, sadness, neutral, other
  - Attribute based annotations
    - Valence, Arousal, and Dominance (VAD)

E. Mower Provost, Y. Shangguan, and C. Busso, "UMEME: University of Michigan emotional McGurk effect data set," IEEE Transactions on Affective Computing, vol. 6, no. 4, pp. 395-409, October-December 2015.

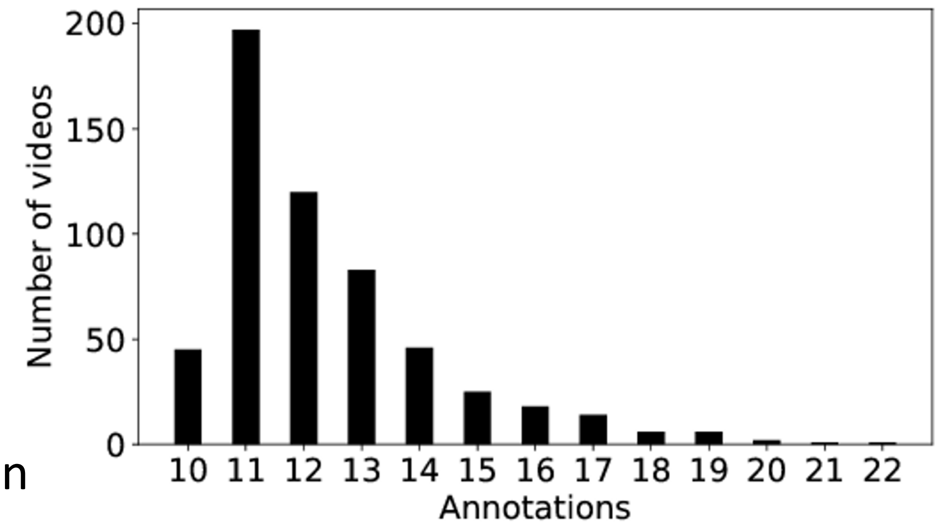
Experimental Setting

- **We consider 5 settings**
  - GROUND
    - Video sequence
  - REFERENCE
    - Video sequence (other annotators)
  - FRAME
    - Randomized static frames
  - FER
    - Deep learning model
  - RANDOM



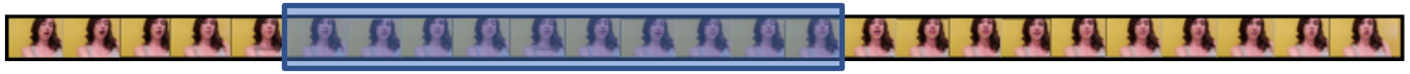
Experimental Setting (cont.)

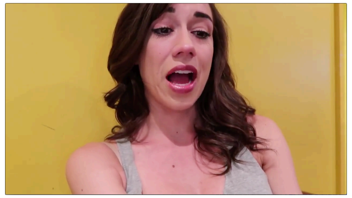
- **GROUND & REFERENCE**
  - MSP-IMPROV (video only) contains at least 10 annotations each, up to 22.
  - Two sets:
    - Reference: 5 randomly selected annotations
    - Ground: Rest of the annotations (5 to 17)
  - Goal: inter-evaluator consistency
  - Categorical class
    - Majority vote to obtain consensus label
    - We normalized annotations to obtain a distribution
  - Emotional attributes
    - Average of arousal, valence and dominance scores
  - 564 Videos total

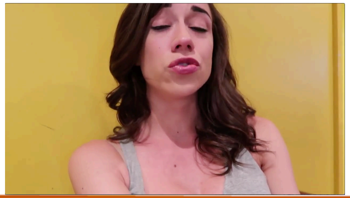


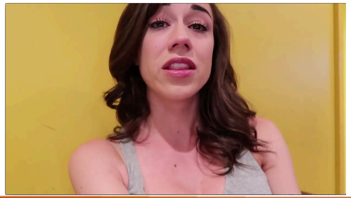
Experimental Setting (cont.)

- **FRAME**
  - Sampled at 3 FPS from target sentences in the MSP-IMPROV corpus (GROUND)
  - Emotional annotations
    - Randomize the order
    - Annotated using crowdsourcing
    - Five annotations per frame
    - Majority vote to decide emotional class
    - Normalization to obtain distribution
  - Average to decide VAD





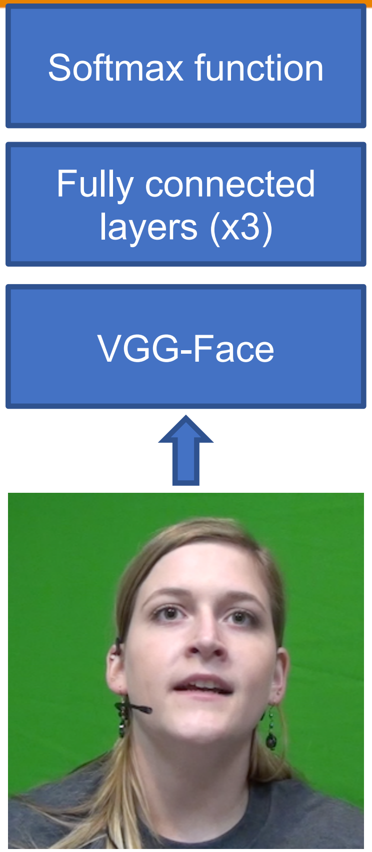




Experimental Setting (cont.)

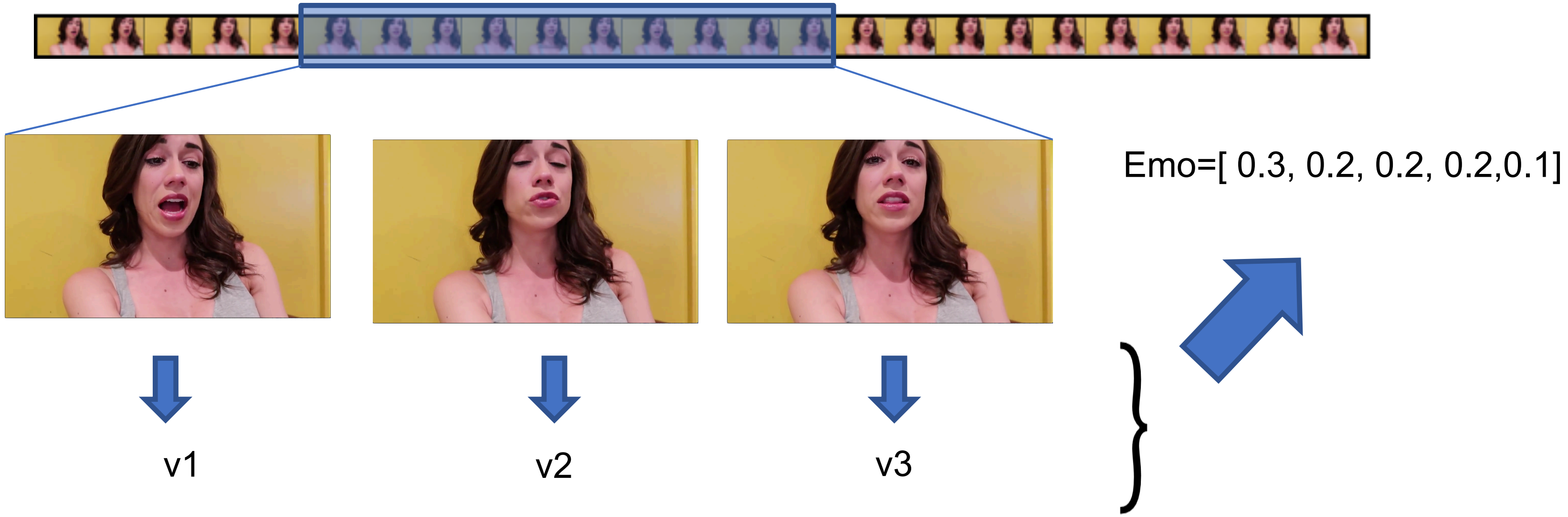
- **FER**
  - Deep learning model trained with the AffectNet corpus
  - VGG16 architecture – VGG-Face initial weights

Emotion	Precision	Recall	F1-Score
Neutral	0.64	0.70	0.67
Sadness	0.87	0.90	0.89
Happiness	0.75	0.70	0.72
Anger	0.76	0.70	0.73
Average	**0.75**	**0.75**	**0.75**



Experimental Setting (cont.)

- **Random**
  - Randomly picks an emotion for each frame (3 FPS)
  - Randomly pick a score for valence, activation, and dominance



Analysis

- **Euclidean distance (ED) for categorical emotion distributions**
- **Observations**
  - GROUND and REFERENCE sets have the lowest ED
  - ED increases for GROUND and FRAME
  - ED increases for GROUND and FER

Euclidian distance	GROUND	REFERENCE	FRAME	FER	RANDOM
GROUND	0	0.30	0.38	0.54	0.68
REFERENCE	0.30	0	0.45	0.57	0.72
FRAME	0.38	0.45	0	0.48	0.57
FER	0.54	0.57	0.48	0	0.77
RANDOM	0.68	0.72	0.57	0.77	0

Emotional perception of isolated frames is not representative of the emotional perception of the entire video

Emotional F-score for categorical representation

Label	Set	Precision	Recall	F1-score
Anger	REFERENCE	0.73	0.67	0.7
	FRAME	0.55	0.14	0.22
	FER	0.5	0.05	0.08
	RANDOM	0.16	0.16	0.16
Happiness	REFERENCE	0.91	0.84	0.87
	FRAME	0.67	0.97	0.79
	FER	0.78	0.77	0.78
	RANDOM	0.29	0.16	0.16
Neutral	REFERENCE	0.72	0.72	0.72
	FRAME	0.54	0.77	0.63
	FER	0.55	0.59	0.57
	RANDOM	0.29	0.16	0.2

Label	Set	Precision	Recall	F1-score
Sadness	REFERENCE	0.77	0.79	0.78
	FRAME	0.66	0.57	0.61
	FER	0.4	0.79	0.53
	RANDOM	0.21	0.11	0.14
Other	REFERENCE	0.18	0.21	0.19
	FRAME	0.5	0.07	0.12
	FER	0	0	0
	RANDOM	0	0	0
Average	REFERENCE	0.71	0.69	0.7
	FRAME	0.55	0.59	0.54
	FER	0.51	0.52	0.47
	RANDOM	0.23	0.15	0.16

F1-score of GROUND compared to the other sets

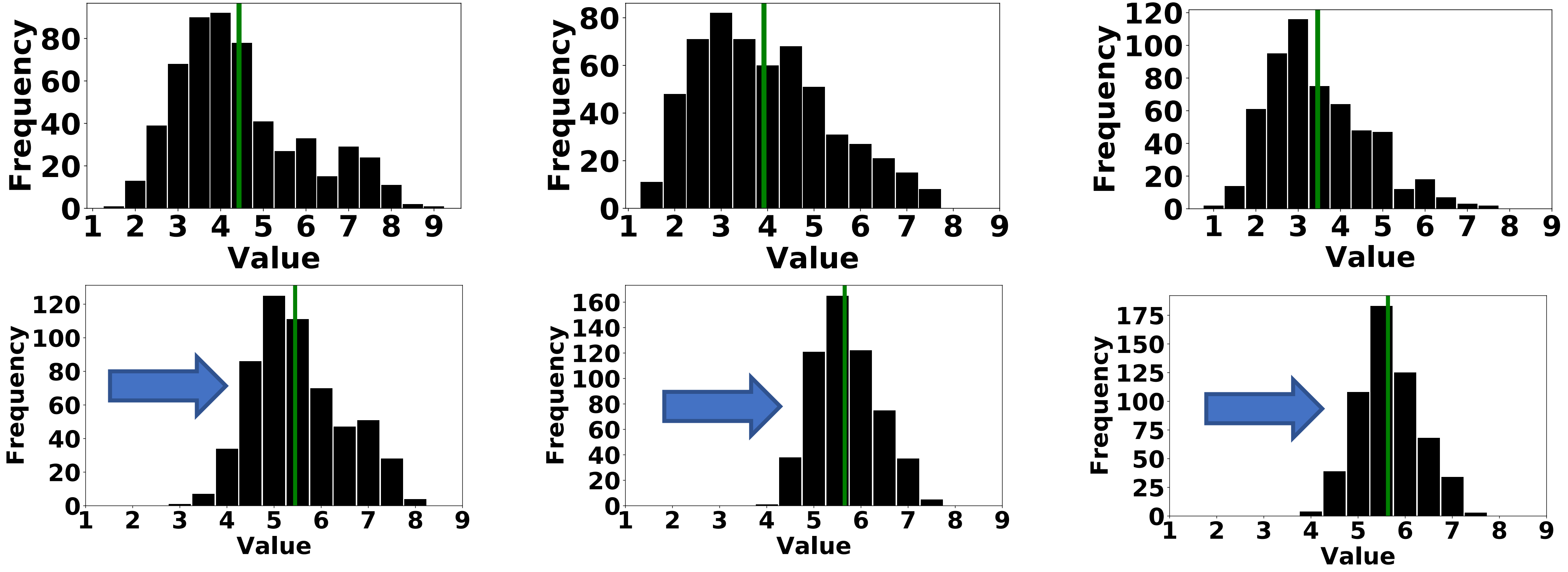
- **Observations**
  - Low F1-score for anger in static images
  - Other emotions are closer between video and frames
  - Similar trend with FER

Analysis of attribute representations

Average valence, arousal, dominance scores

GROUND

FRAME



Valence

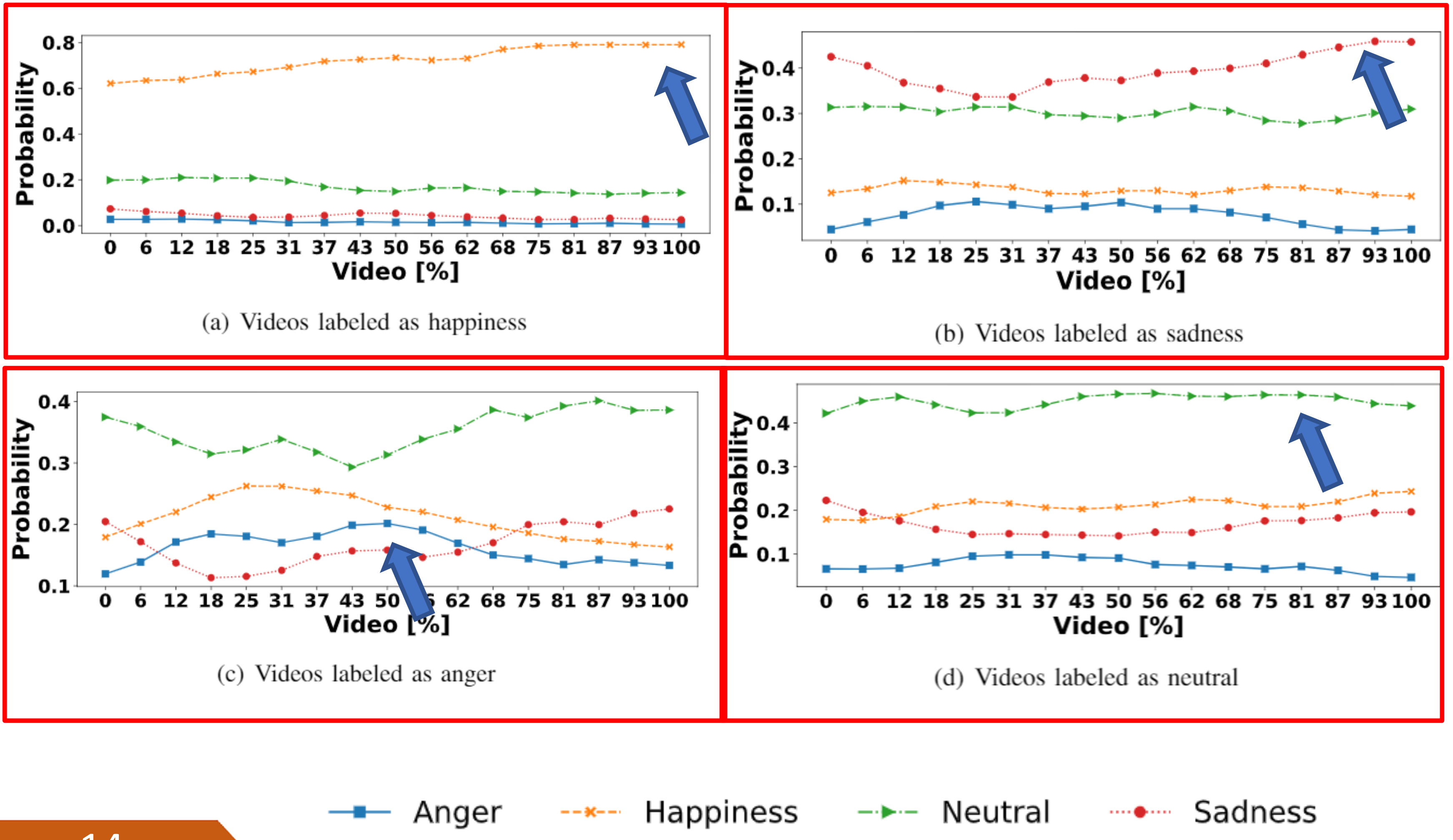
Arousal

Dominance

- **Observations**
  - Shift in the perception of emotion in static images
    - Arousal (more active); valence (more positive); dominance (more dominant)

Temporal Analysis

Temporal average distribution in the FRAME set



(a) Videos labeled as happiness

(b) Videos labeled as sadness

(c) Videos labeled as anger

(d) Videos labeled as neutral

Observations

- Happiness (a) has the highest confident reaching 80%
- Sadness (b) and Neutral (d) have less confidence, hovering around 40%
- Anger peaks at only 20% showing opposite behavior

Viseme Analysis

Table 1: Phoneme to viseme mapping

Phoneme	Viseme	Phoneme	Viseme
P		K	
B	/p/	G	
M		N	
EM		L	/k/
F	/f/	NX	
V		HH	
T		Y	
D		EL	
S	/t/	EN	
Z		IY	/iy/
TH		IH	
DH		AA	/aa/
DX		AH	
W		AX	/ah/
WH	/w/	AY	
R		ER	/er/
CH		AO	
JH	/ch/	OY	
SH		IX	
ZH		OW	
EH		UH	/uh/
EY	/ey/	UW	
AE		SIL	/sp/
AW		SP	

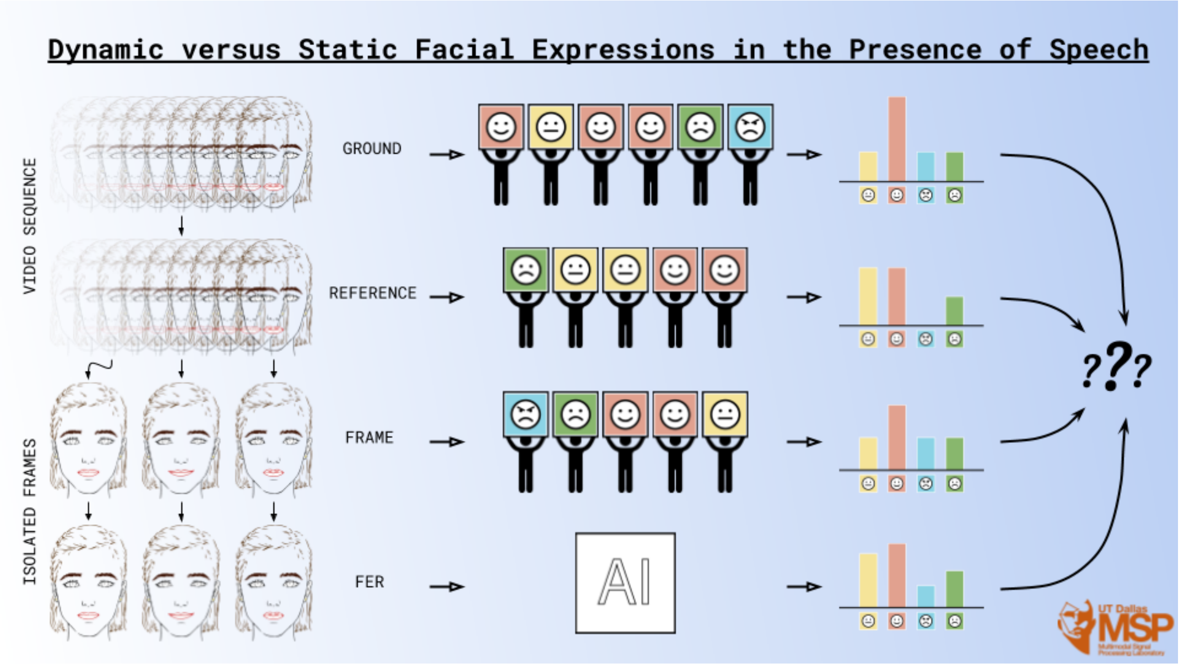
Viseme	Coverage	Primary Emotion	L2 Distance
/ah/	8.0%	39.3%	0.5849
/sp/	23.7%	44.1%	0.6371
/er/	1.9%	41.3%	0.6438
/iy/	9.1%	44.5%	0.6528
/t/	16.3%	46.0%	0.6719
/ch/	3.3%	43.3%	0.6740
/ey/	5.7%	41.4%	0.6787
/x/	4.9%	41.0%	0.6797
/w/	4.1%	36.0%	0.6929
/k/	14.0%	46.6%	0.7022
/aa/	1.6%	36.1%	0.7295
/f/	1.9%	37.3%	0.7593
/uh/	1.0%	38.2%	0.7598
/p/	4.3%	36.9%	0.7612

- **Viseme Level Analysis**
- **Observations**
  - Silence has second lowest ED
  - /p/ has highest L2 (bilabial sound)

P. Lucey, T. Martin, and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in Australian International Conference on Speech Science & Technology(SST 2004), Sydney, NSW, Australia, December 2004, pp. 265–270.

Conclusions

- **Observations**
  - Frame-based analysis without considering context or temporal information does not represent well the emotion of a video
    - Even if frame-based model is as good as human performance
    - Anger emotion is the most affected class
  - Speech articulations affect the perception of emotion



Our Research: msp.utdallas.edu

NEC \Orchestrating a brighter world

This work was funded by NEC Foundation Of America