Analysis and Compensation of the Reaction Lag of Evaluators in Continuous Emotional Annotations

Soroosh Mariooryad and Carlos Busso

Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas

Sep 4, 2013

busso@utdallas.edu

Emotional Descriptors

- Emotional labels at sentence level
  - One descriptor assigned to a segment
    - sentence, turn, chunk, word
  - Long segments: variations are not captured
- Continuous labels
  - Track emotional content continuously over time
  - They capture localized emotional behaviors
  - Facilitate emotion analysis at different resolutions

Continuous Emotional Labels

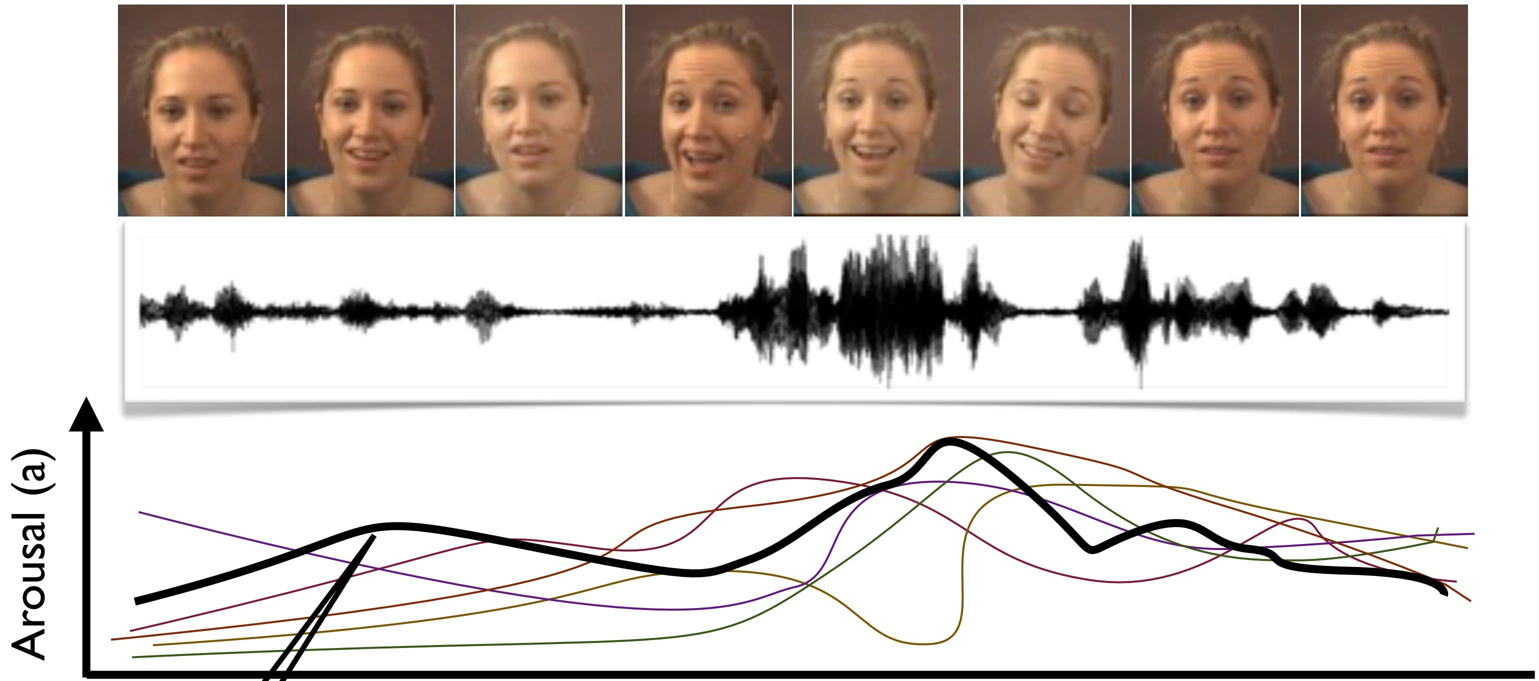
- Record position of a cursor controlled by user
- Examples of these GUIs:
  - FEELTRACE [Cowie et al. 2000] and Gtrace [Cowie et al., 2012]
  - MoodSwings [Kim et al., 2008] and EmuJoy [Nagel et al., 2007]

Feeltrace - picture from Cowie et al. (2000)

Thursday, September 5, 13

Discrete Classification Problem

- Approach: estimate mean across evaluators



Mean trajectory

$$\text{Segment} \in \begin{cases} \text{Class 1} & \text{if } \bar{a} < 0.5 \\ \text{Class 2} & \text{if } \bar{a} \geq 0.5 \end{cases}$$

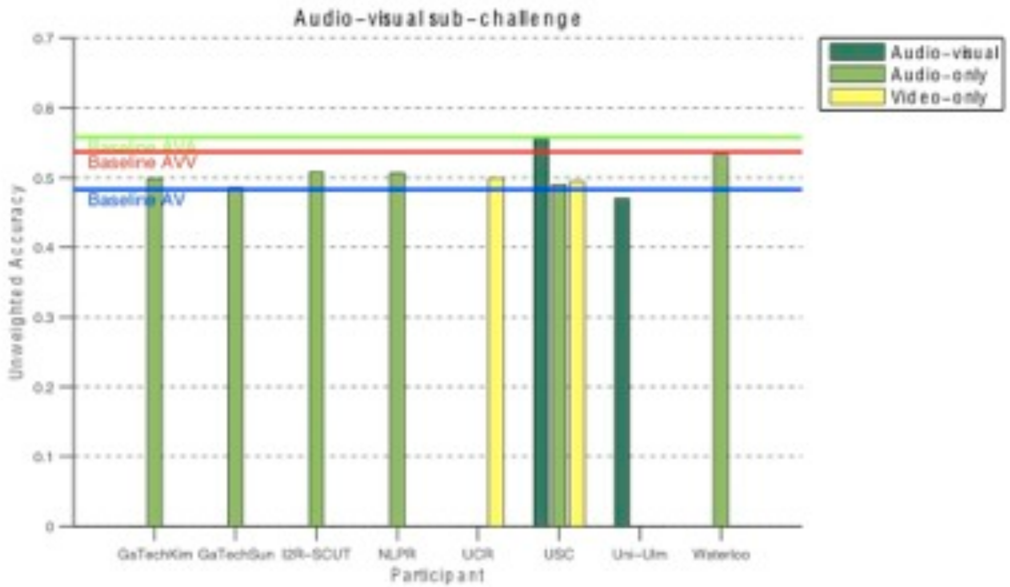
(Or regression model)

UTD

Motivation

- Emotional classification results in naturalistic database was very low - SEMAINE [McKeown et al., 2012]
    - Challenging task - spontaneous emotions

AVEC 2011

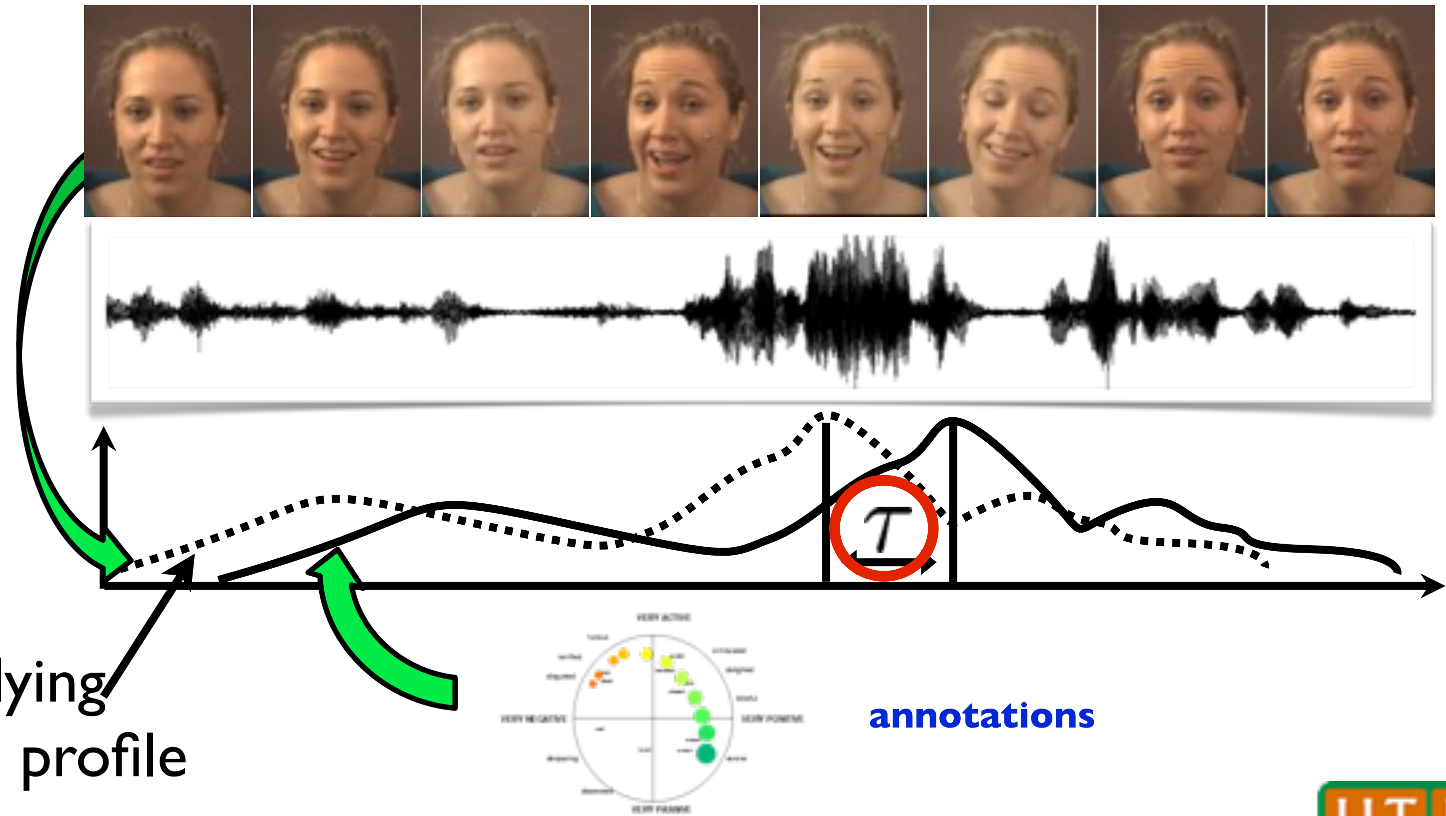


Is there any other reason for low performance?

Source: http://sspnet.eu/avec2011

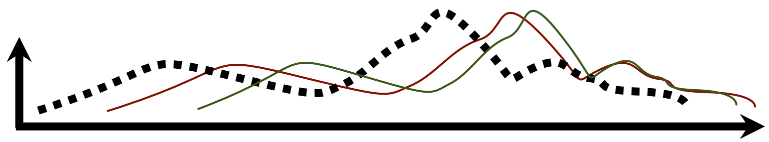
Evaluator Reaction Lag

- Emotion assessment
  - Sense the stimuli, appraise the emotional message, define their judgment, moving the cursor

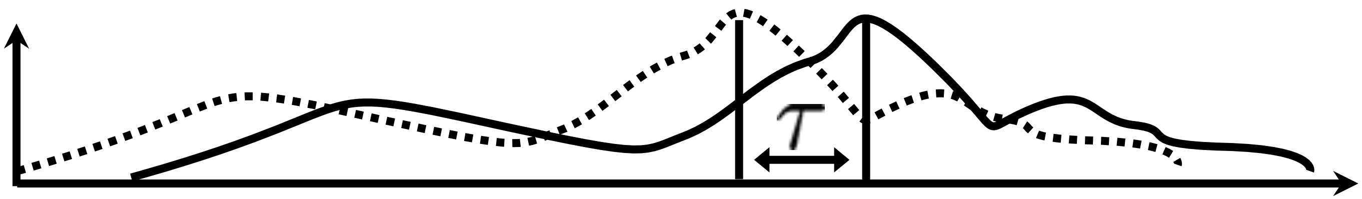


Problem Formulation

- How to formulate the estimation of the reaction lag?
  - Constant reaction lag or time-variant
  - Annotator-dependent or annotator-independent



- Assumptions in this work
  - Constant reaction lag across time
  - Annotator-independent (mean across evaluators)
    - Preliminary results on annotator-dependent

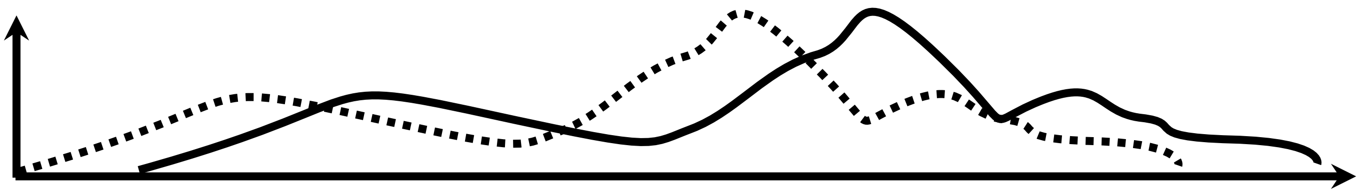


Estimating Reaction Lag

- Proposed approach based on mutual information (MI)
  - Capture the dependency between two random variables
- Find the optimal reaction lag

$$\hat{\tau} = \arg_{\tau} \max I[EMO; ANN^{\tau}]$$

  - EMO = emotional content of the stimulus
  - ANN^{τ} = shift version of emotional annotation



Estimation of Emotional Content

$$\hat{\tau} = \arg_{\tau} \max I[EMO; ANN^{\tau}]$$

- *EMO* represented by facial features capturing the deviations from neutral behaviors (EMO^F)
- Why acoustic features are not included?
  - During silence, speech features are not available
  - Single frame does not convey enough emotion cues
- Distributions are estimated with k-means
  - $P(EMO^F)$
  - $P(ANN^{\tau})$
  - $P(ANN^{\tau}, EMO^F)$

SEMAINE Database

Source: McKeown et al. (2012)

user

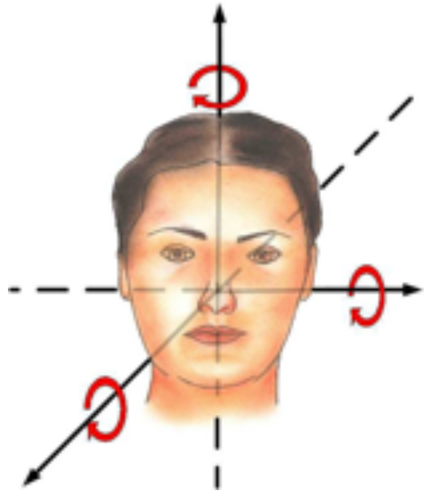
operator (stimulus)

- Emotionally colored interactions
- Annotations: FEELTRACE (activation, valence)
- 44 sessions, 9 unique speakers (users)
  - Sessions with annotations and correctly extracted facial features

Facial Features

- Facial features extracted with CERT [Bartlett et al. 2006]
  - Action Units from FACS (deviation from neutral faces)
  - Head rotation (Jaw, Yaw and Pitch)

AU	Description	AU	Description
AU 1	Inner Brow Raise	AU 15	Lip Corner Depressor
AU 2	Outer Brow Raise	AU 17	Chin Raise
AU 4	Brow Lower	AU 18	Lip Pucker
AU 5	Eye Widen	AU 20	Lip stretch
AU 6	Cheek Raise	AU 23	Lip Tightener
AU 7	Lids Tight	AU 24	Lip Presser
AU 9	Nose Wrinkle	AU 25	Lips Part
AU 10	Lip Raise	AU 26	Jaw Drop
AU 12	Lip Corner Pull	AU 28	Lips Suck
AU 14	Dimpler	AU 45	Blink/Eye Closure



- For EMO^F, we use $K \in \{2, 4, 6, 8, 10, 16, 20\}$ over the joint feature space

busso@utdallas.edu

$P(ANN^\tau)$

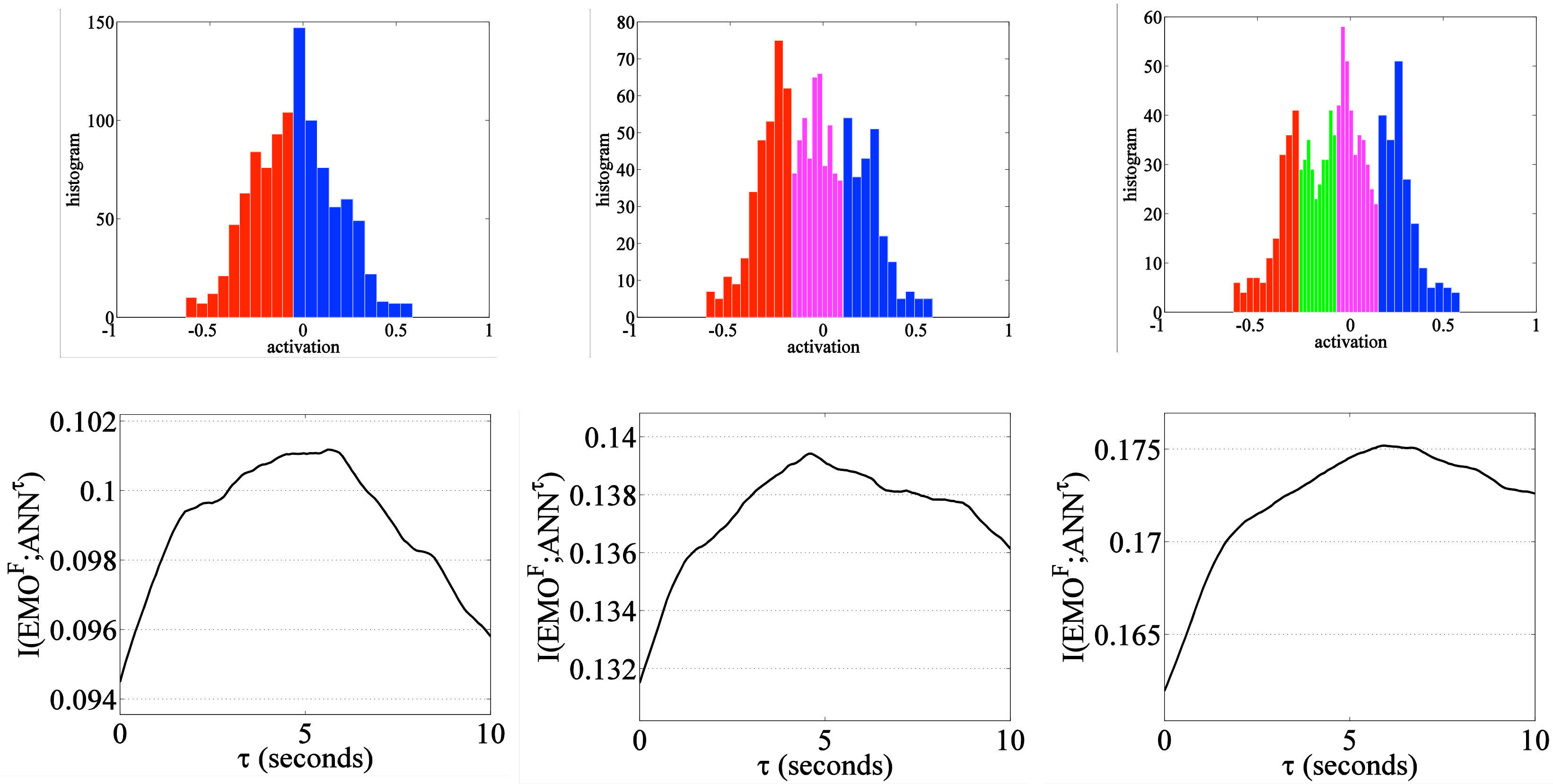
Analysis of the Reaction Lag

- Activation

K = 2 | K = 3 | K = 4

Clusters (ANN)

Lag analysis



UTD

Thursday, September 5, 13

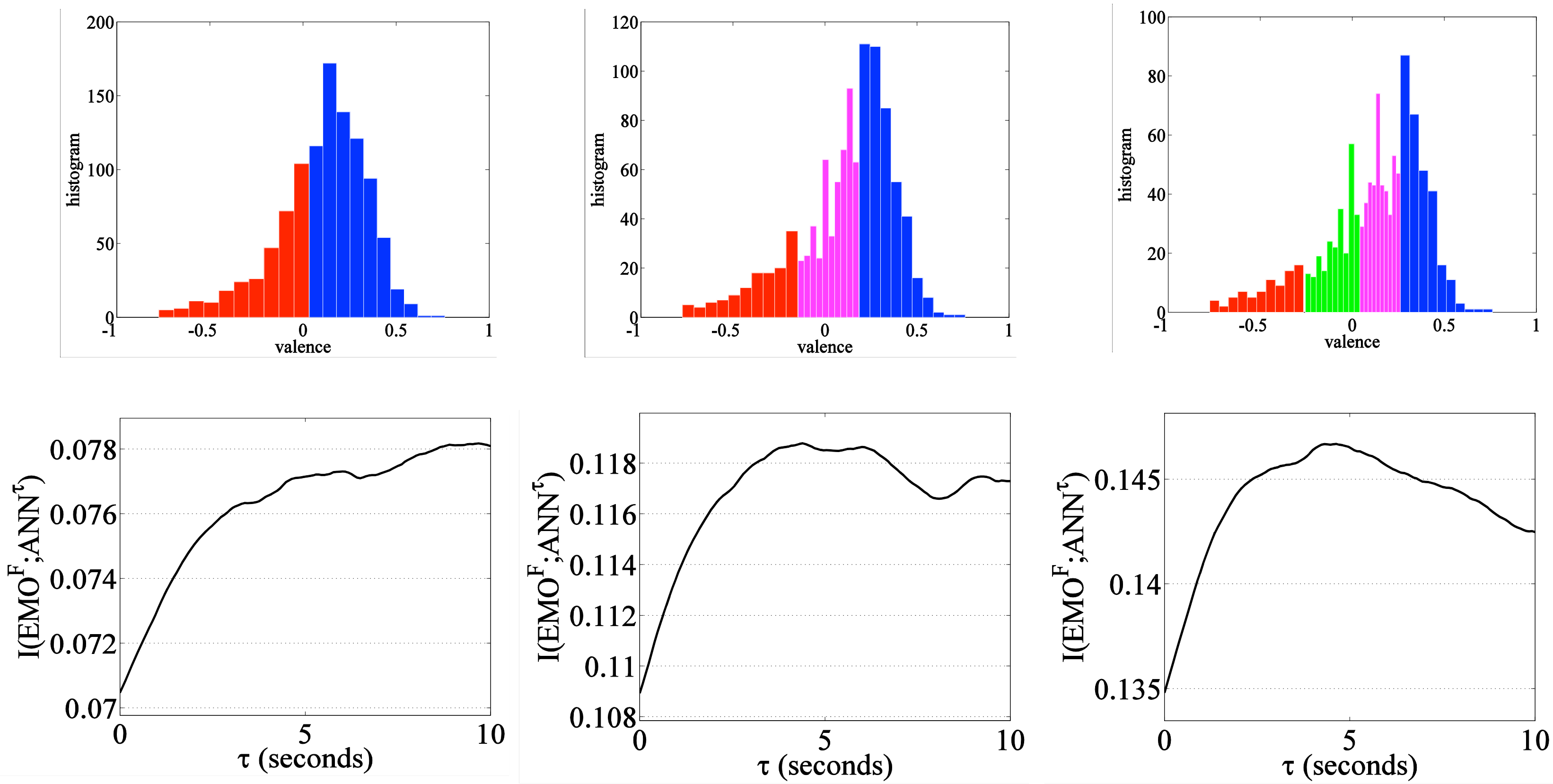
Analysis of the Reaction Lag

- Valence

K = 2 **K = 3** **K = 4**

Clusters (ANN)

Lag analysis



UTD

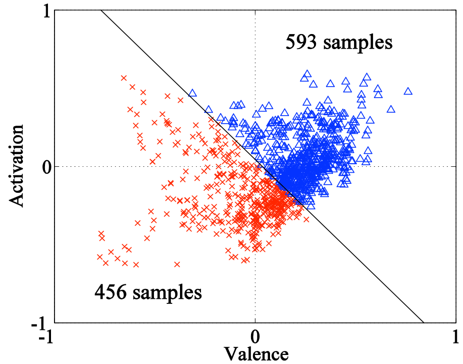
Thursday, September 5, 13

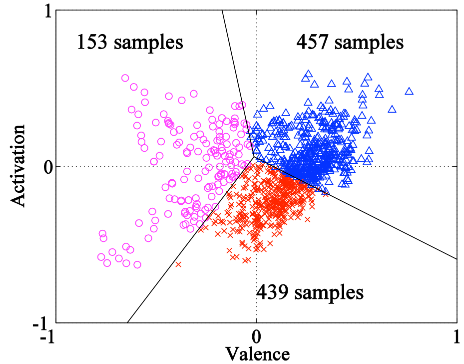
Analysis of the Reaction Lag

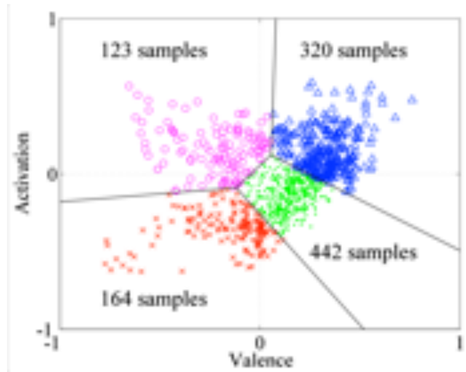
- [Activation, Valence]

K = 2 **K = 3** **K = 4**

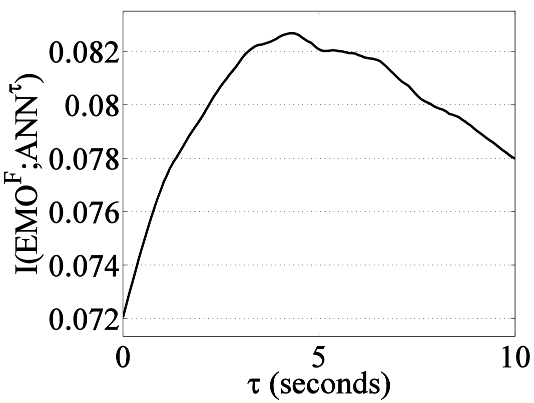
Clusters (ANN)

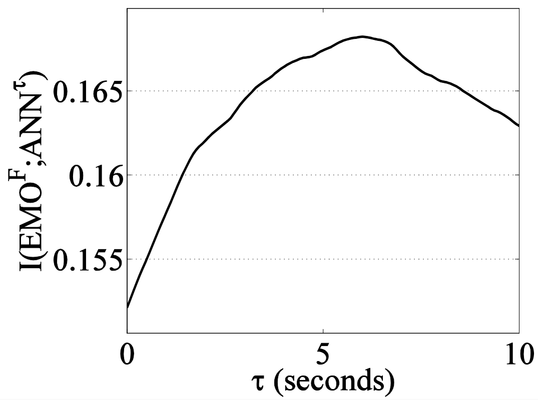


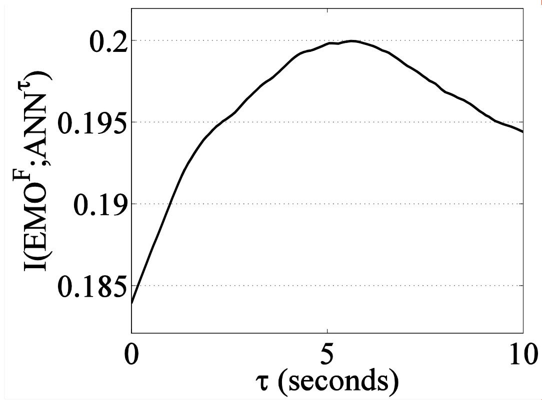




Lag analysis



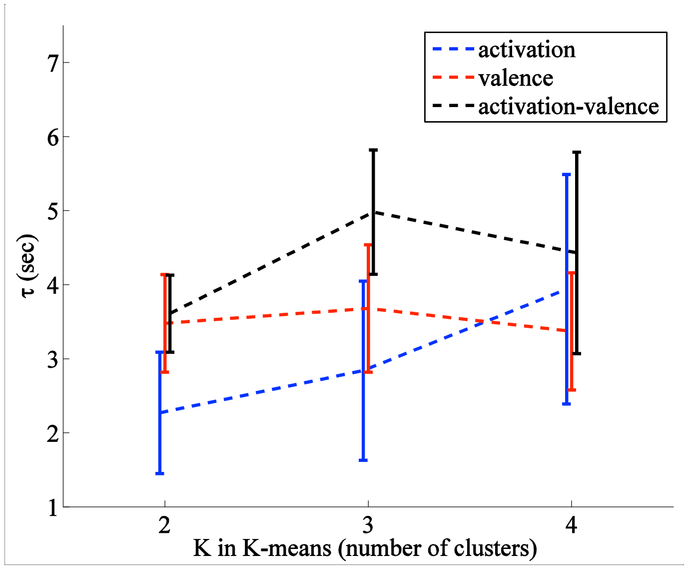




Experimental Setting

- The optimal delay is defined as the first time the mutual information does not increase
  - Priority to shorter reaction lag

Attribute	K=2		K=3		K=4	
	mean	std	mean	std	mean	std
Act	2.27	0.82	2.84	1.21	3.94	1.55
Val	3.48	0.66	3.68	0.86	3.37	0.79
Act-Val	3.61	0.52	4.98	0.84	4.43	1.36



Validation with Emotion Recognition

- 1049 turns (at least 300ms long) - 9 subjects
- SVM with 9-fold speaker independent cross-validation
- Evaluation settings
  - Activation, valence, and [activation, valence]
  - Discrete emotional labels with K=2, 3,4 classes
  - Reaction lag: 0, 1, 2 and 3 sec + optimal delay
- Facial features
  - [AUs+head] x 6 statistics (e.g., quantiles, mean and std)
- Acoustic features

UTD

Acoustic Features

- openSMILE 4368 features [Eyben et al. 2010, Schuller et al. 2011]
  - Spectral

Rasta-style filtered auditory spectrum bands
MFCCs
Spectral energy 25-60Hz, 1k-4KHz
Spectral roll-off point 0.25 0.50 0.75 0.90
Spectral Flux, entropy, variance, skewness, kurtosis, slope

  - Energy

Sum of auditory spectrum (loudness)
Sum of Rasta-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate

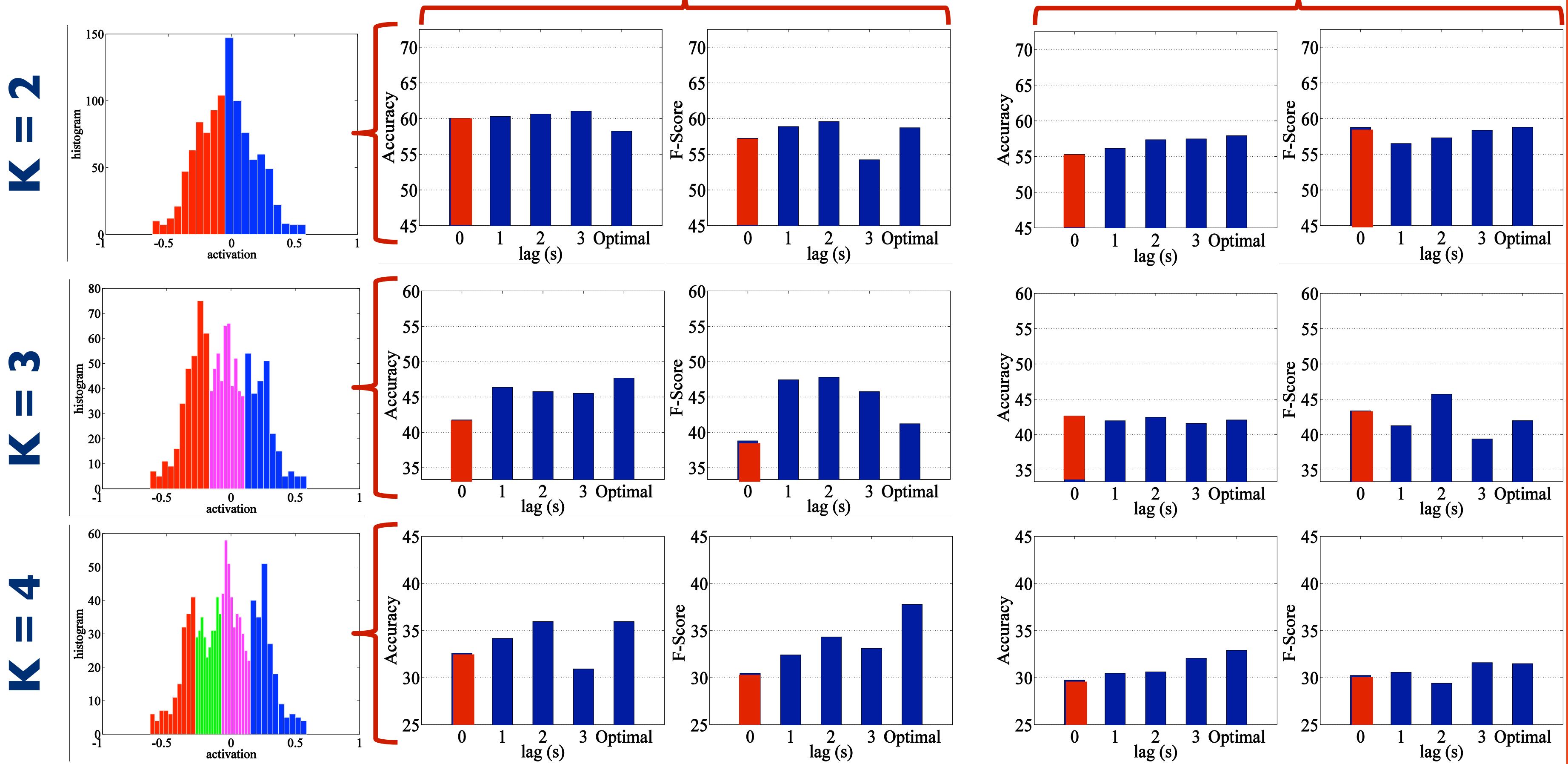
  - Voice

F0
Probability of voicing
Jitter (local, delta)
Shimmer

- Feature selection with CFS (~ 99 features)

UTD

Recognition Experiments - Activation



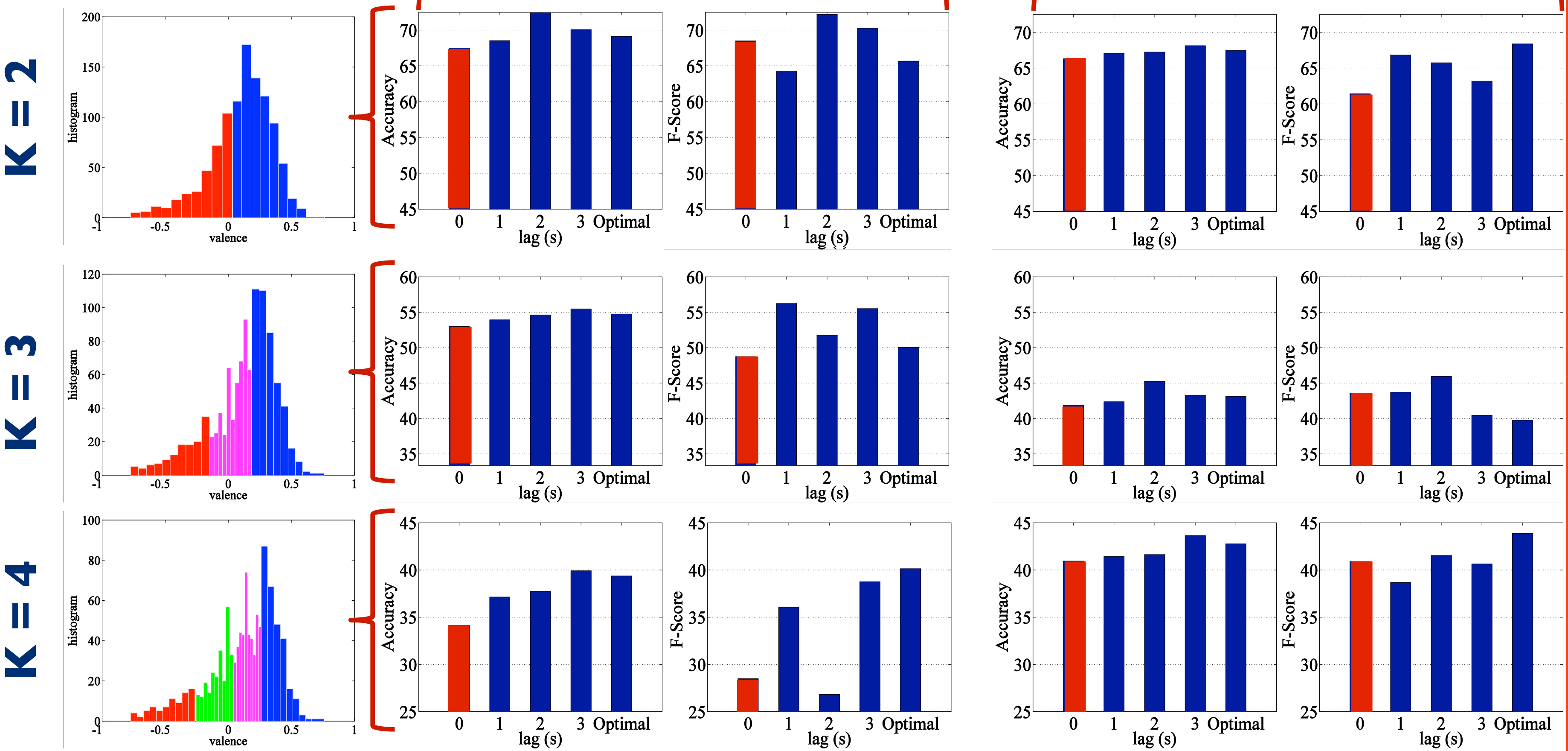
UTD

Thursday, September 5, 13

Recognition Experiments - Valence

Face

Speech

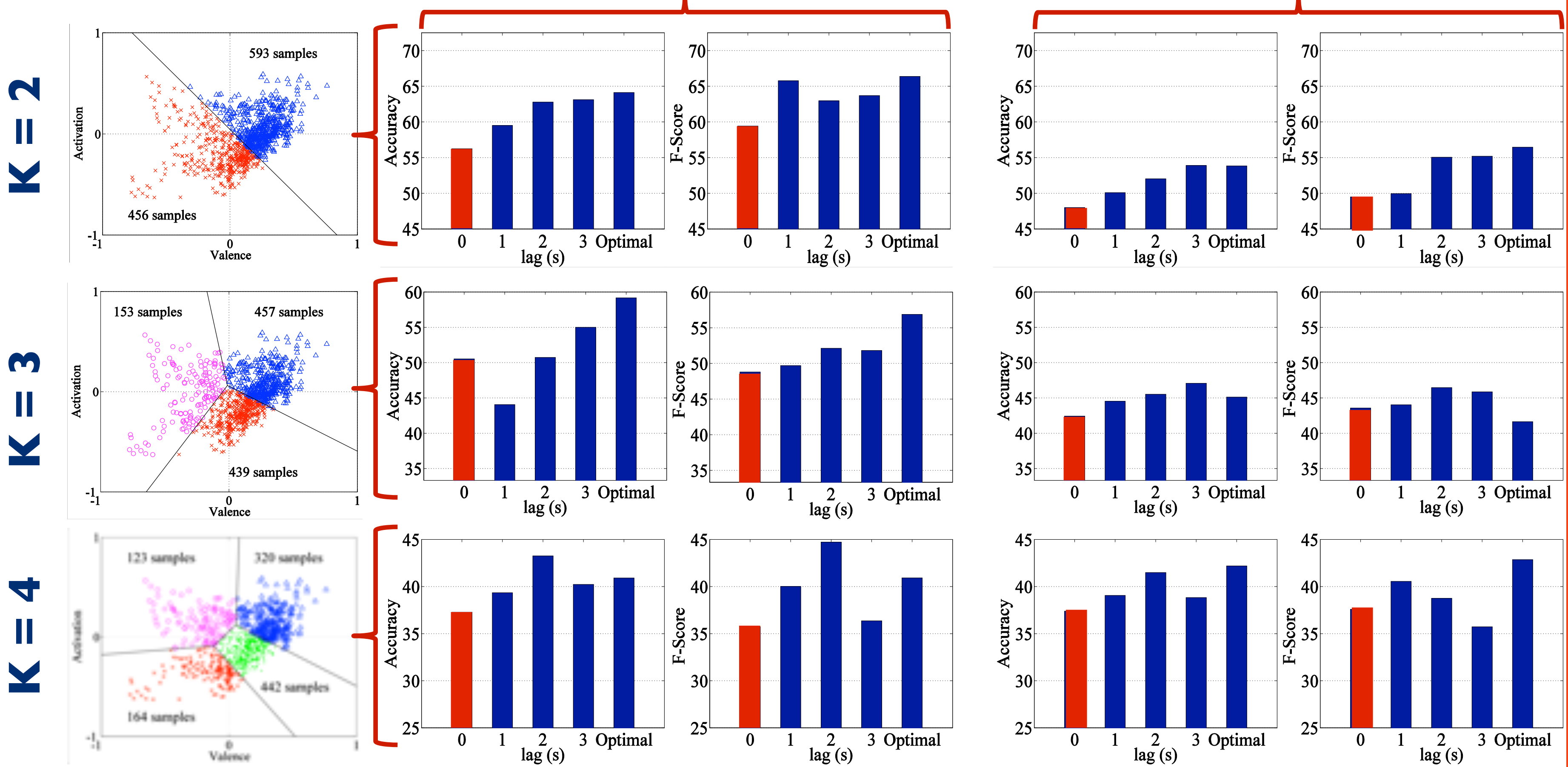


busso@utdallas.edu

Recognition Experiments - [Act,Val]

Face

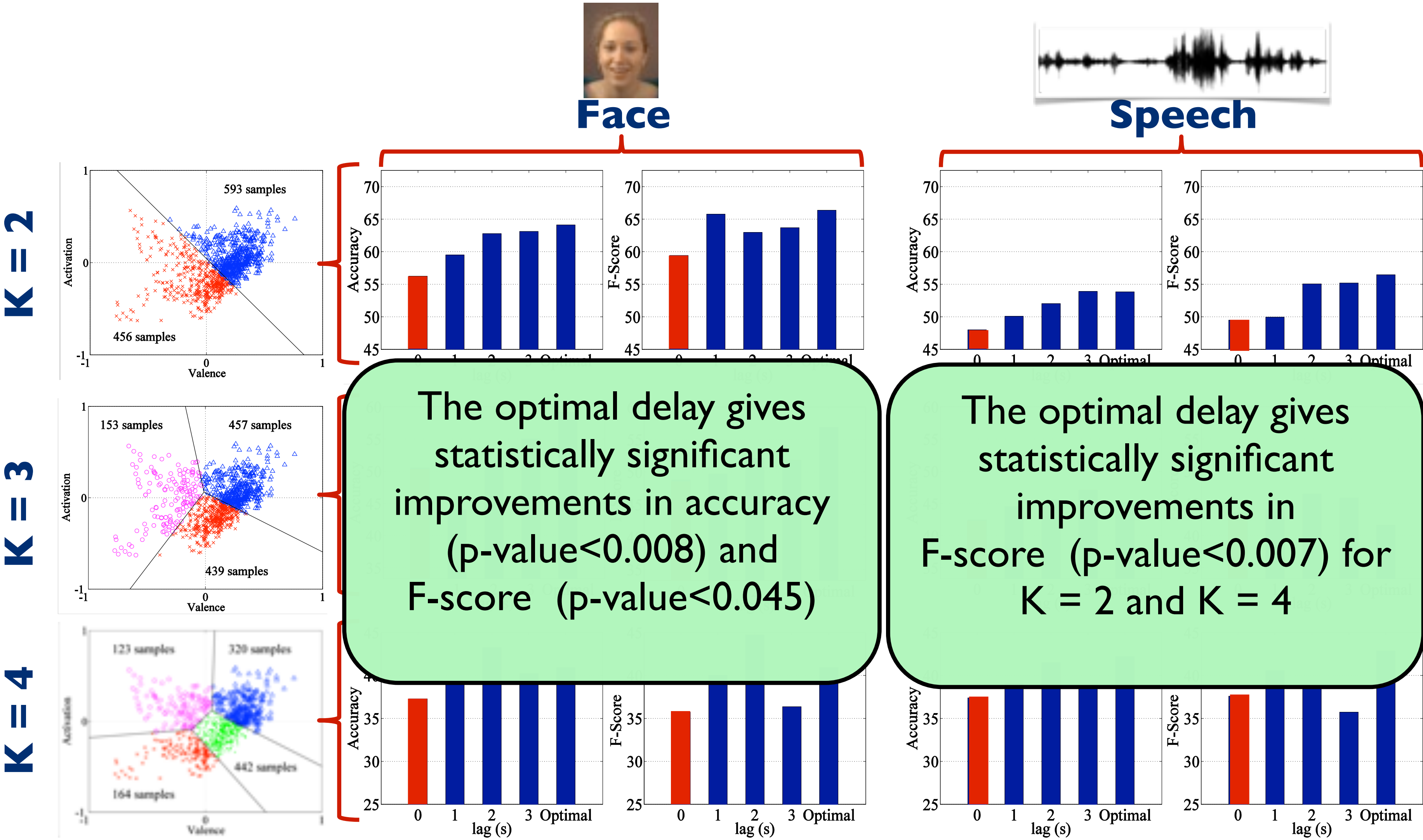
Speech



UTD

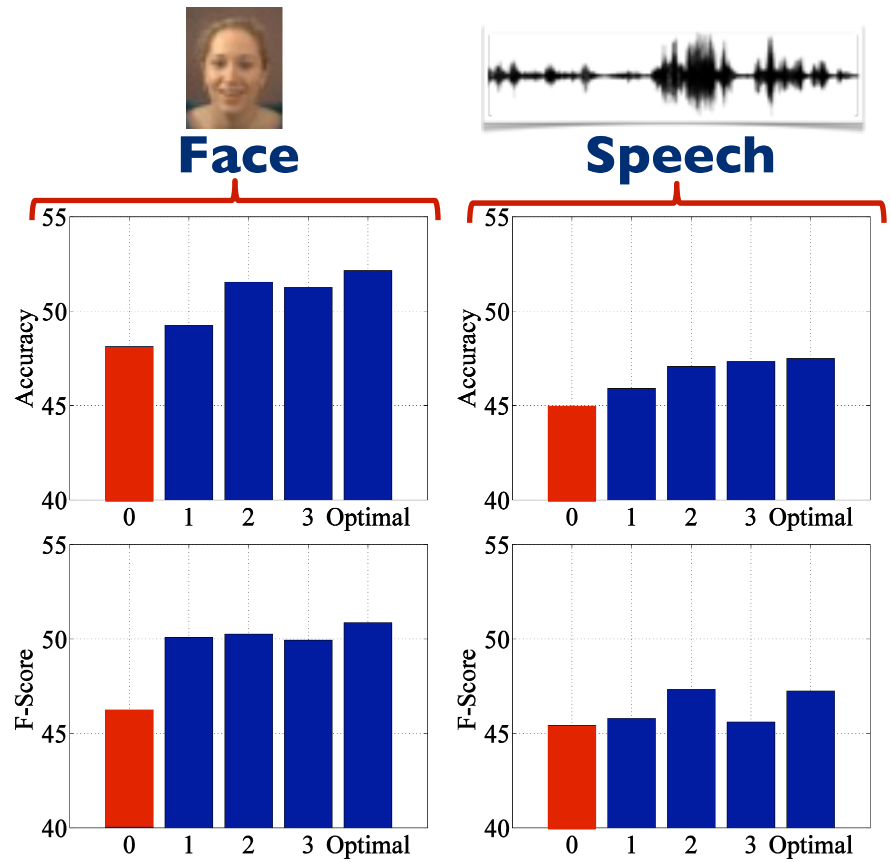
Thursday, September 5, 13

Recognition Experiments - [Act,Val]



Recognition Experiments - Average

- Across all settings
  - Act, Val, [Act-Val]
  - K = 2,3,4

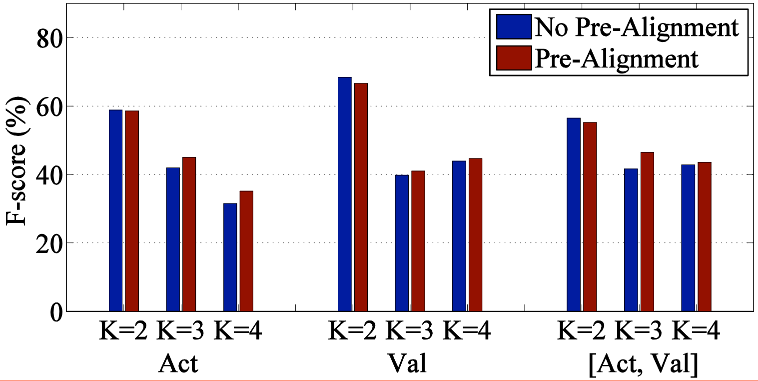


- Optimal delay estimated from training set yields the best performance across all settings on the test set

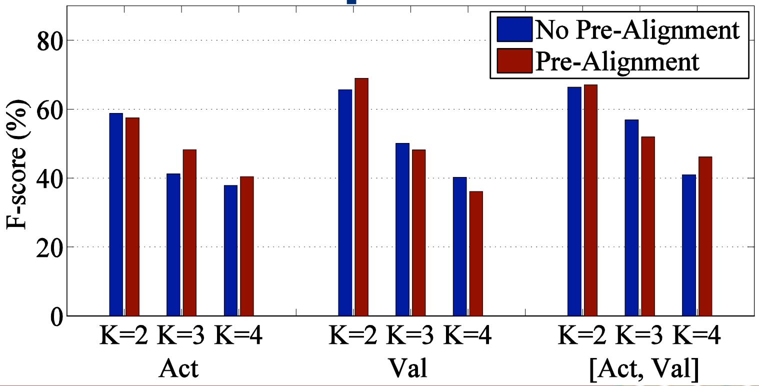
Experiments – Pre-Aligning the Annotations

- Evaluator dependent lag
  - Assumption: phase between two annotators is fixed and is less than 1 sec
  - Pre-Aligning the labels of multiple annotator to maximize the correlation between them within [-1, 1] seconds
  - F-score improves 1.06% (face) and 0.26% (speech)

Face



Speech



Conclusions

- The mutual information analysis unveils and quantifies the reaction lag with respect to facial features
- Compensating for the reaction lag improves the performance of both facial and vocal emotion recognition systems
  - Shift-delayed emotional annotations achieved statistically significant improvements

We are using the wrong labels!

busso@utdallas.edu

Future Work

- Reaction lag analysis with respect to speech features
- Reaction lag analysis in evaluator-dependent fashion
  - Find optimum delay per annotation
- Considering time-variant reaction lag
  - Time warping methods e.g., dynamic probabilistic canonical correlation with time warping (DPCTW) [Nicolaou et al., 2012]

Multimodal Signal Processing (MSP)

Thanks!

Work funded by Samsung Telecommunications America and NSF

http://msp.utdallas.edu/

busso@utdallas.edu