FEATURE AND MODEL LEVEL COMPENSATION OF LEXICAL CONTENT FOR FACIAL EMOTION RECOGNITION

Soroosh Mariooryad and Carlos Busso

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

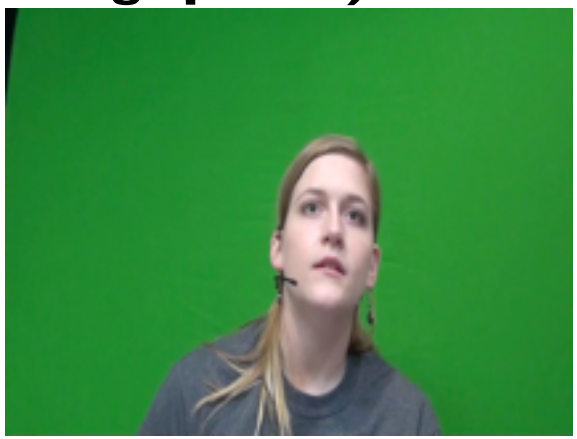
April 24, 2013

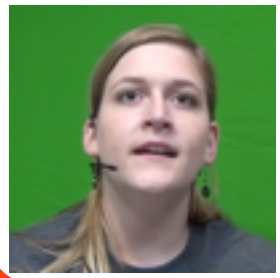
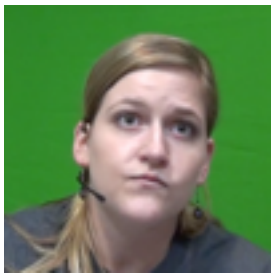
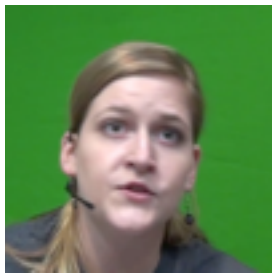
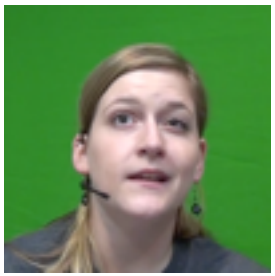
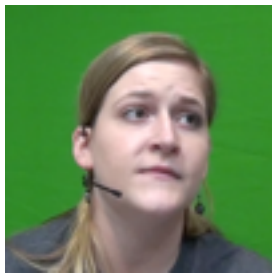
FG 2013 10th IEEE International Conference on
Automatic Face and Gesture Recognition
Shanghai, China—April 22-26, 2013

busso@utdallas.edu

Motivation

- Human communication is a sophisticated process
  - **Lexical Content (i.e., what is being spoken)**
  - **Speaker Characteristics (i.e., who is speaking)**
  - **Emotional Content (i.e., how is being spoken)**
  - Cultural background
  - Age
  - Physical state

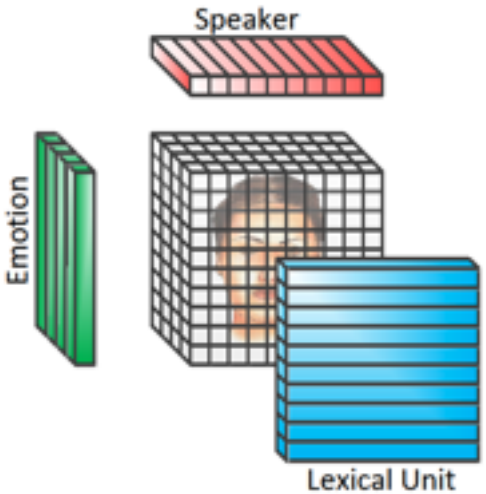


Motivation

- Each of these factors introduce variabilities
  - Lexicon: underlying articulatory process
  - Speaker: intrinsic cultural, physiological and idiosyncratic characteristics
  - Emotion: expressive behaviors
- Controlling each of the factors can enhance the analysis of the others
  - Emotion recognition

"Conditioning reduces entropy"



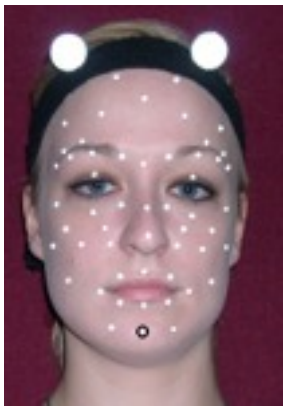
Goals

- Quantify and localize the speaker, lexical and emotional variabilities observed in facial movements
  - Identify facial areas that are more affected by speaker and lexical variabilities
- Guide the design of lexicon dependent models for robust emotion recognition system
- Simplifications:
  - We are using the transcriptions of the corpus
  - We use controlled recordings using motion capture system

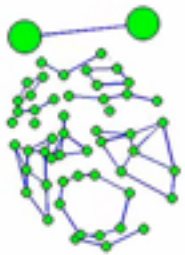
Outline

- Motivation
- **Decoding Facial Expressions**
- Emotional Recognition Experiments
- Conclusions

IEMOCAP Database



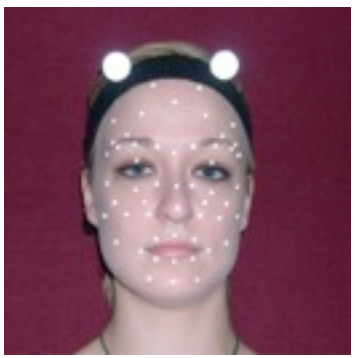
- 12 hours of spontaneous dyadic interactions
- Ten speakers (5 male, 5 female)
- Manually segmented and transcribed
  - Automatic forced alignment (HMM)
- Emotional content subjectively evaluated
  - The 4 most frequent: happiness, anger, sadness and neutral



Facial Features

- 53 Markers
  - Facial structure and marker placement variability





  - Z-normalized (reference speaker)

$$m'^{s}_{i,d} = (m^{s}_{i,d} - \mu^{s}_{i,d}) \times \frac{\sigma^{ref}_{i,d}}{\sigma^{s}_{i,d}} + \mu^{ref}_{i,d}$$

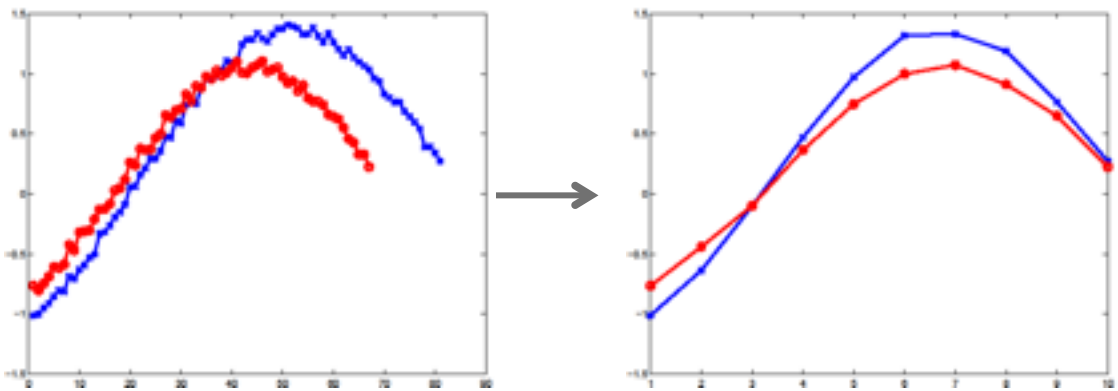
Methodology

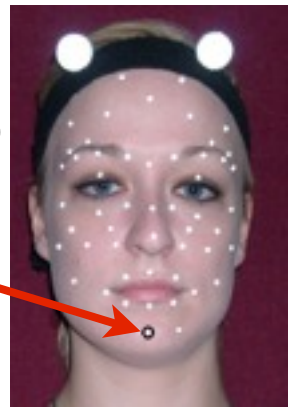
- Speaker
  - Ten speakers
- Emotional Content
  - The four most frequent emotions (happy, angry, sad, neutral)
- Lexical Content
  - The ten most frequent syllables and worlds

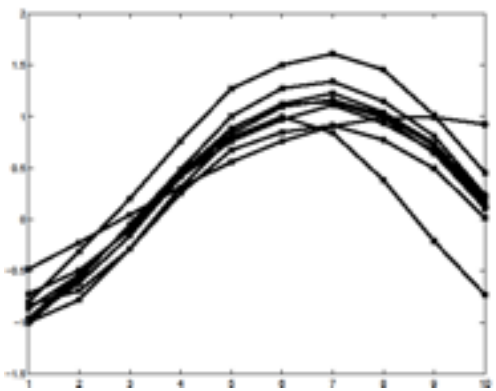
Syllables	AY 3232	Y_UW 2437	AX 1748	N_OW 1556	T_AX 899	AX_T 823	L_AY_K 733	DH_AX 693	G_OW 670	AX_N_D 599
Words	I 3152	YOU 2402	KNOW 955	A 952	TO 739	THE 690	LIKE 566	AND 541	DO 523	ME 503

Factor Analysis – Trajectory Models

- Trajectory Modeling of Facial Features at some lexical unit (phoneme, syllable or word)
  - Normalizing the duration (resampling and interpolation)



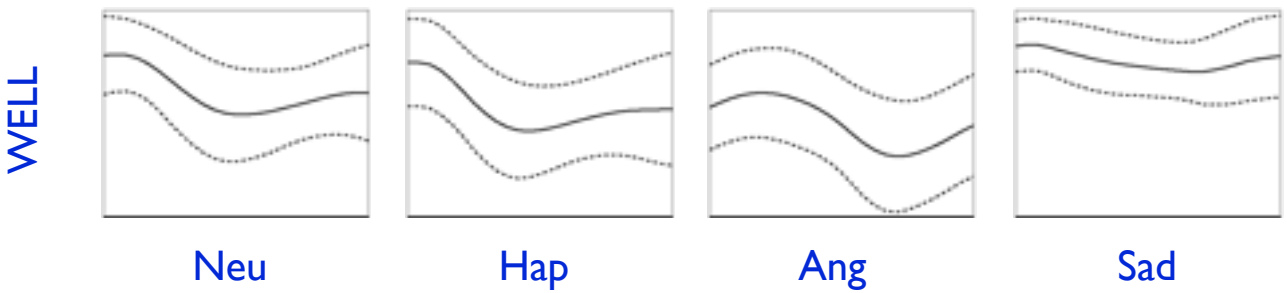




  - N = 10 samples
  - Mean (μ_A)
  - Covariance (Σ_A)

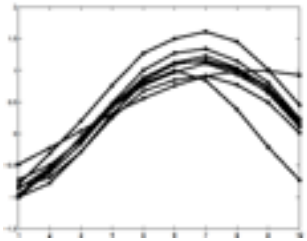
Factor Analysis – Trajectory Models

- Trajectory models
  - Word-level
- Models for word "WELL"

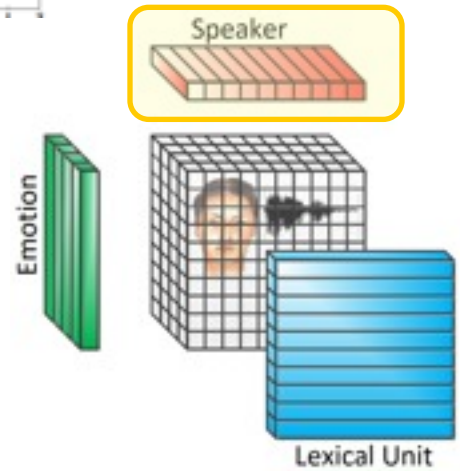


Factor Analysis – Measuring Variability Reduction

- Factor Analysis
  - Conditioning on the factors reduces the uncertainty
  - Measure variability reduction



- Uncertainty (Variability)
  - Uncertainty (trace of Σ_A)
- E.g., speaker factor



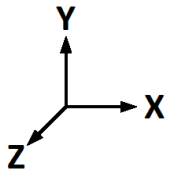
One matrix per lexical unit

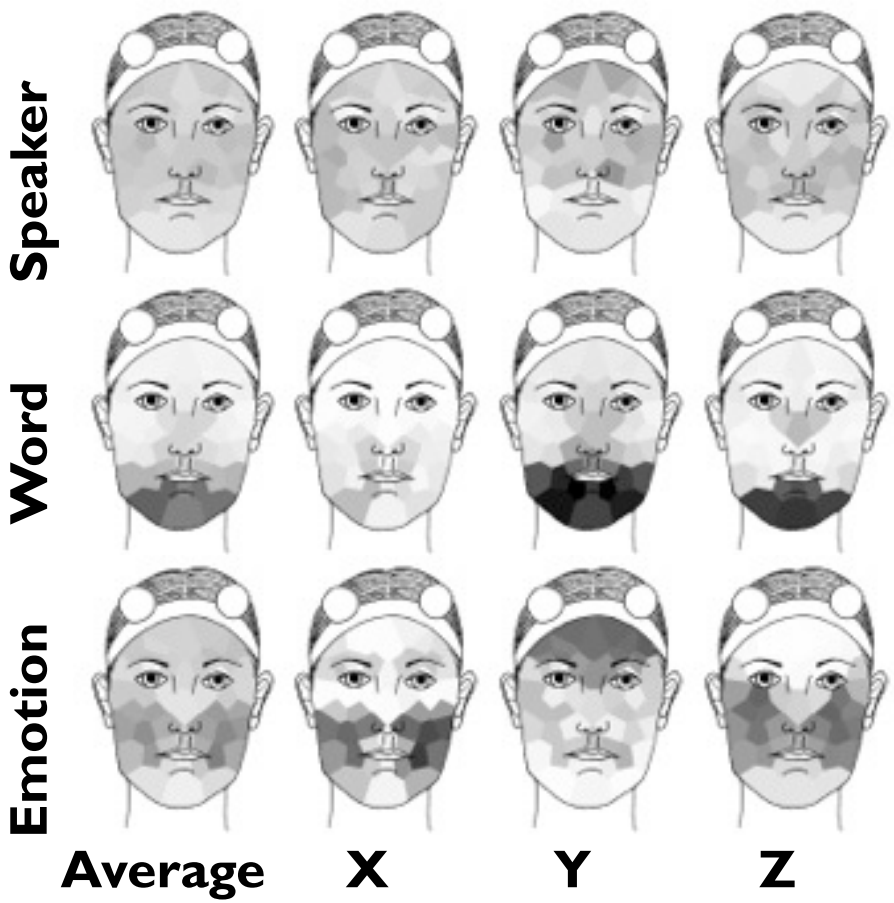
$$RM(A;F) = tr(\Sigma_A) - \sum_{f \in F} P(f) tr(\Sigma_A \mid f)$$

For all speakers, emotions, and lexical units

Factor Analysis - Results

- Word level models:
- Speaker
  - Uniformly distributed
  - Not dominant
- Lexical Content
  - Orofacial area (Y, Z)
- Emotion
  - Middle face (X, Z)
  - Upper region (Y)





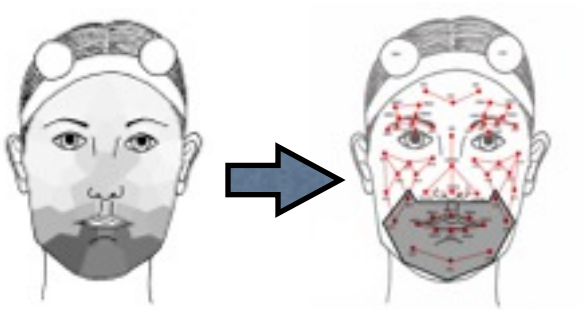
Outline

- Motivation
- Decoding Facial Expressions
- **Emotional Recognition Experiments**
- Conclusions

Emotion recognition experiments

- Hypothesis
  - Lexicon-dependent models for the orofacial area should gives better performance than generic, lexicon-independent models
- We only consider 15 markers from the orofacial area

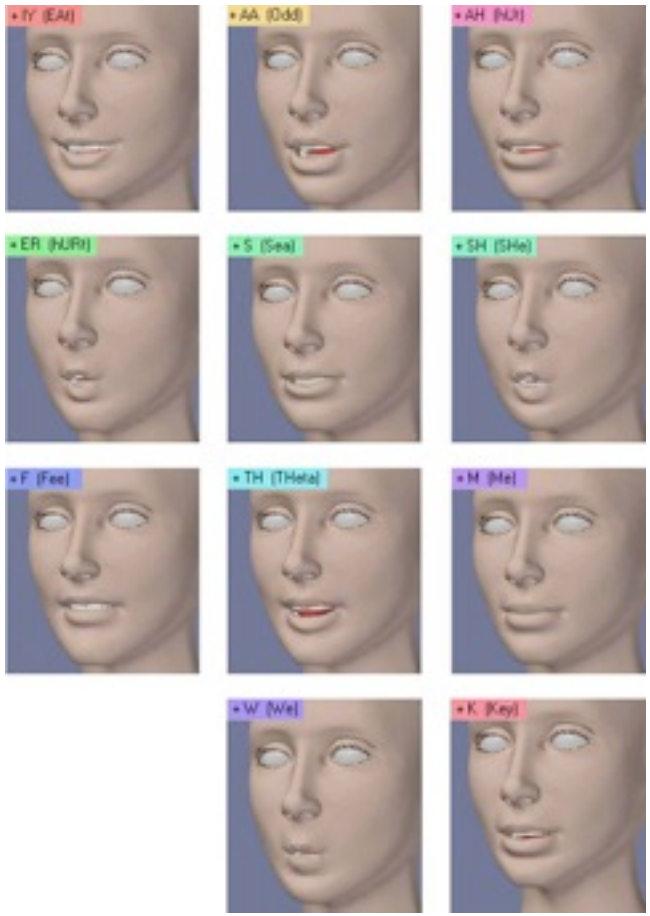
Lexical variability



- Visemes are selected as the lexical unit
  - Longer units (words/syllables) are not practical (too many)

Viseme-level lexical compensation

- Phoneme-to-viseme mapping



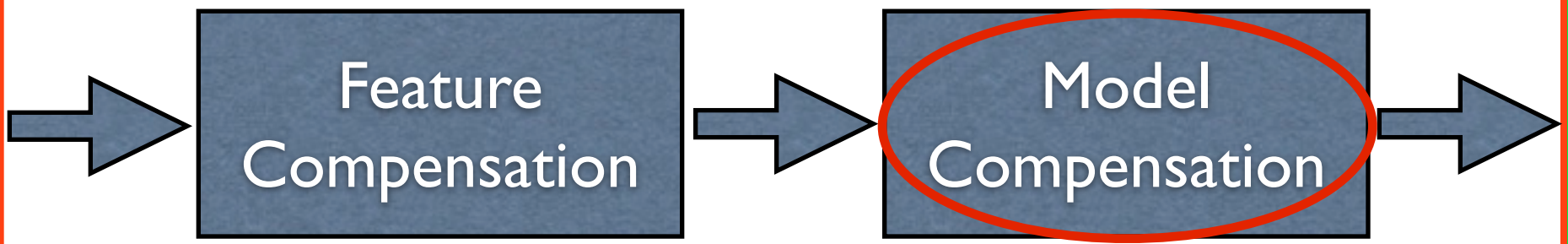
[http://freesdk.crydev.net/display/SDKDOC3/Phonemes+and+Visemes]

#	Phoneme	Viseme
1725	B	p
1258	P	
3077	M	
1510	F	f
1470	V	
5454	T	t
1676	TD	
713	TH	
2921	D	
662	DD	
216	DX	
2414	DH	
411	TS	
3952	S	
1930	Z	
2560	W	w
3071	R	
282	CH	ch
540	SH	
13	ZH	
524	JH	
2213	AH	ah
3670	AY	
7273	AX	

#	Phoneme	Viseme
1354	EY	ey
2312	EH	
2610	AE	
589	AW	
2703	K	k
1344	G	
6476	N	
3714	L	
1597	H	
2210	Y	
1322	NG	
274	KD	
3679	IY	iy
3487	IH	
1475	AA	aa
462	ER	er
1059	AO	x
56	OY	
874	IX	
2841	OW	
596	UH	uh
2254	UW	
1501	AXR	
-	SIL	sp

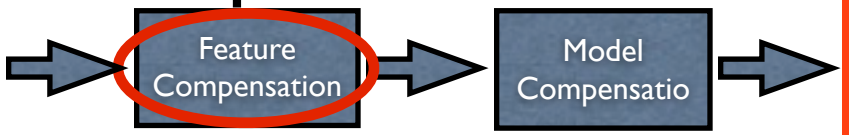
Lexical Compensation

- We explore compensation schemes at the feature and model level



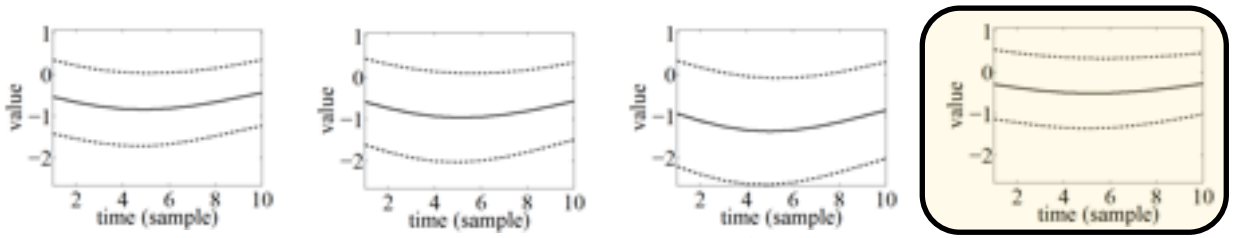
- Model Level Lexical Compensation
  - Viseme-dependent models (i.e., one classifier per emotion)
  - Drawback: we need to train 13 classifiers, and, therefore, the data is split in 13

Feature Level Lexical Compensation

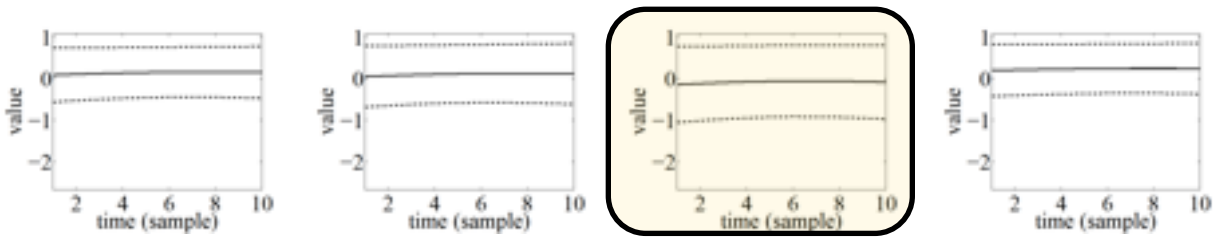


- Approach

/ey/



/t/



- Whitening transformation at viseme level
  - Whitening viseme *i*
  - Coloring to reference viseme

$$X^w = D_i^{-\frac{1}{2}} V_i' (X - \mu_i)$$

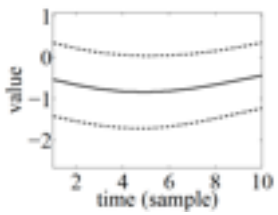
$$X^n = V_{ref} D_{ref}^{\frac{1}{2}} X^w + \mu_{ref}$$

$$\Sigma_A^i = V_i D_i V_i' \qquad \Sigma_A^{ref} = V_{ref} D_{ref} V_{ref}'$$

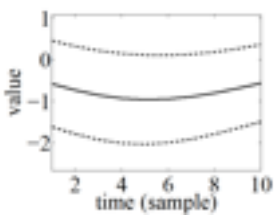
Lexical Normalization (Whitening Transformation)

- Lexical normalization

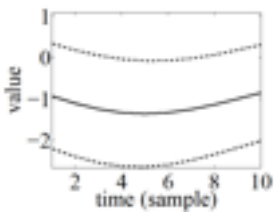
/ey/



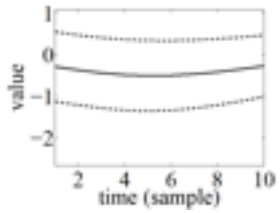
Neu



Hap

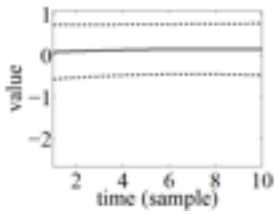


Ang

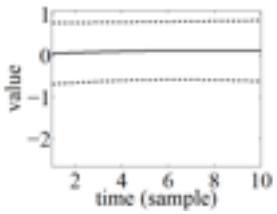


Sad

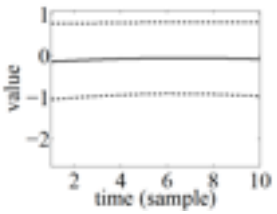
/t/



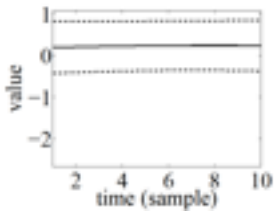
Neu



Hap

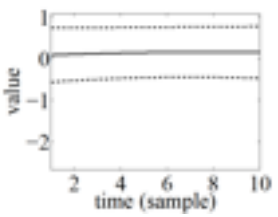


Ang

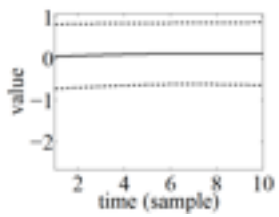


Sad

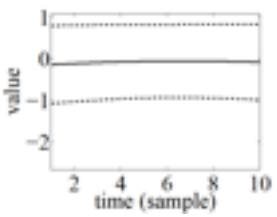
/ey/=>/t/



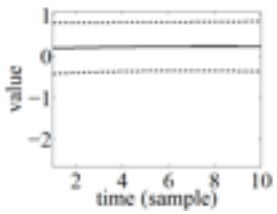
Neu



Hap



Ang



Sad

Emotion Recognition Experiments

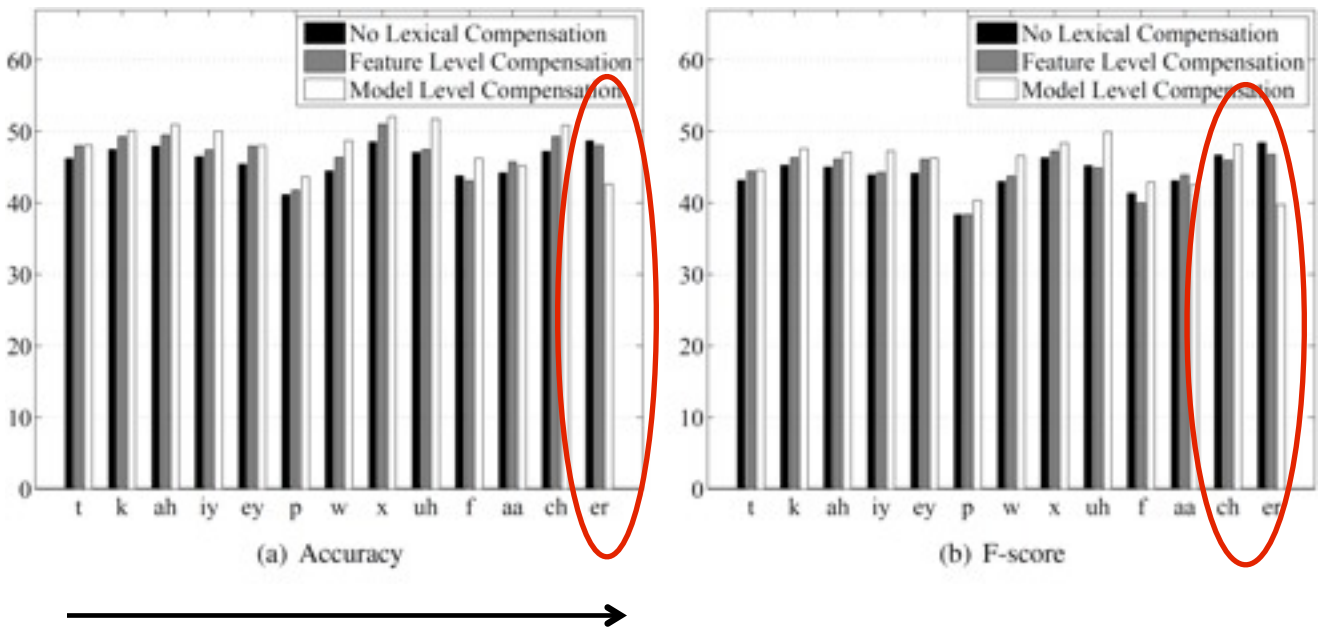
- SVM (WEKA)
  - Statistics: min, max, mean, std, median, lower quartile and upper quartile
- Phoneme (Viseme)-Level

Lexical compensation	Accuracy [%]	Precision [%]	Recall [%]	F-score
No compensation	46.34	43.74	43.99	0.439
Feature Level	47.89	44.74	44.86	0.448
Model Level	**49.07**	**46.13**	**46.32**	**0.462**

- Feature level: 1.55% (relatively 3.34%) improvement in A
- Model level: 2.73% (relatively 5.9%) improvement in A
  - Statistically significant for A, P, R and F (p-value < 0.0001)

Emotion Recognition Experiments

- Phoneme (Viseme)-Level
  - Lexical compensation improves accuracy for most of the visemes



Decreasing order of number samples per viseme

#	Phoneme	Viseme
462	ER	er
282	CH	ch
540	SH	
13	ZH	
524	JH	

Emotion Recognition Experiments - Utterance-Level

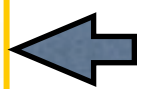
- Unbalanced emotional classes
  - Metrics: accuracy (A), precision (P), recall (R), F-score (F)
- Previous works
  - Average Recall (R) = 54.19% (HMM) [Metallinou et al. 2010]

Emotion	Happiness	Anger	Sadness	Neutral	All
Number	838	574	630	585	2627

Emotion Recognition Experiments - Utterance-Level

- Fusion of viseme-level recognizers
  - Feature level: 3.71% (relatively 7.1%) improvement in accuracy
  - Model level: 5.82% (relatively 11.01%) improvement in accuracy
- Statistically significant in A, P, R and F (p-value < 0.02).

Lexical compensation	Fusion	A [%]	P [%]	R [%]	F
No compensation	Majority	51.82	49.58	50.42	50.00
	Sum	52.55	50.39	51.30	50.84
	Product	51.55	49.35	50.27	49.81
Feature Level	Majority	55.42	52.63	53.62	53.12
	Sum	56.26	53.58	54.76	54.16
	Product	55.57	52.83	53.94	53.38
Model Level	Majority	57.41	54.77	56.22	55.49
	Sum	**58.37**	**55.96**	**57.38**	**56.66**
	Product	57.60	55.03	56.30	55.66

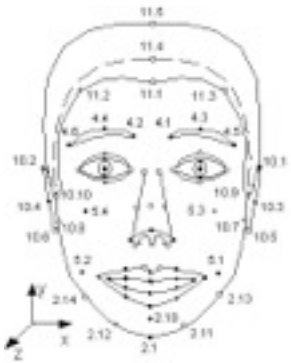


Outline

- Motivation
- Decoding Facial Expressions
- Emotional Recognition Experiments
- **Conclusions**
  - Lexicon-dependent models for classification are only needed on the orofacial region
  - Emotion classification performance improves when the lexical information is compensated at either feature or model level

Future Work

- Fusing lexical dependent (F7) and lexical independent (F1-F6) classifiers
- Blind lexical compensation
  - What to do when we do not have transcripts?
- Validate results on naturalistic databases
  - Features extracted from videos (Actions Units)



Multimodal Signal Processing (MSP)

Thanks!

Work funded by Samsung Telecommunications America and NSF

http://msp.utdallas.edu/