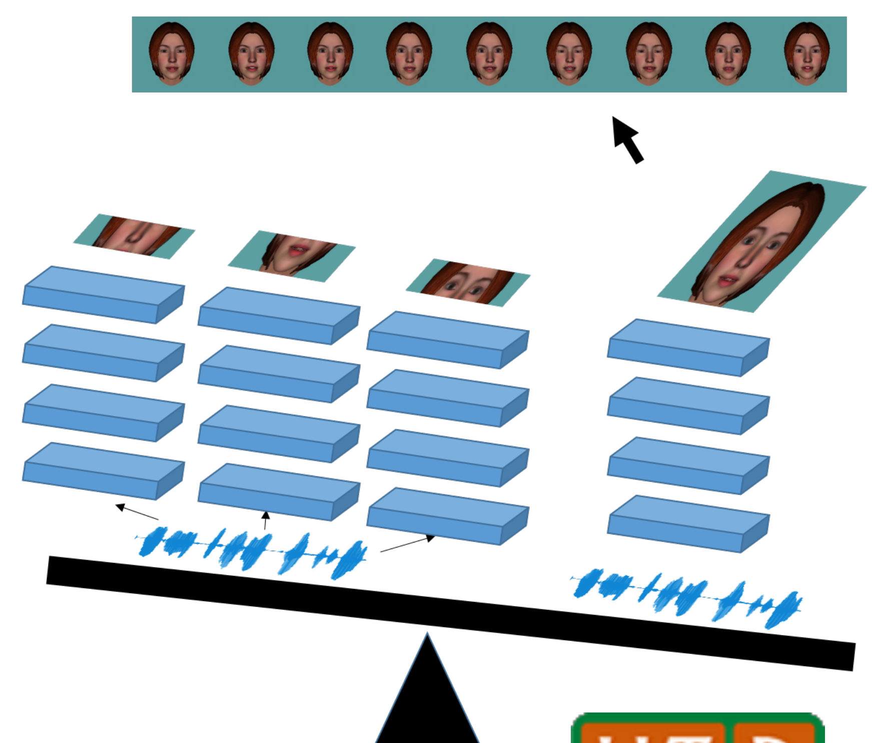
Joint Learning of Speech-Driven Facial Motion with Bidirectional Long-Short Term Memory

NAJMEH SADOUGHI AND CARLOS BUSSO

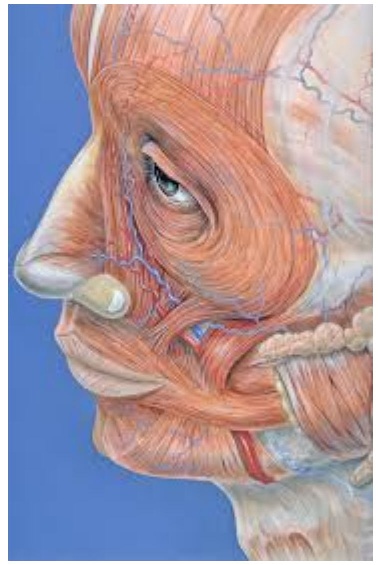
Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science



Motivation

- Generate expressive facial movements for virtual agent (VA)
  - Facilitate the communication
  - Naturalness

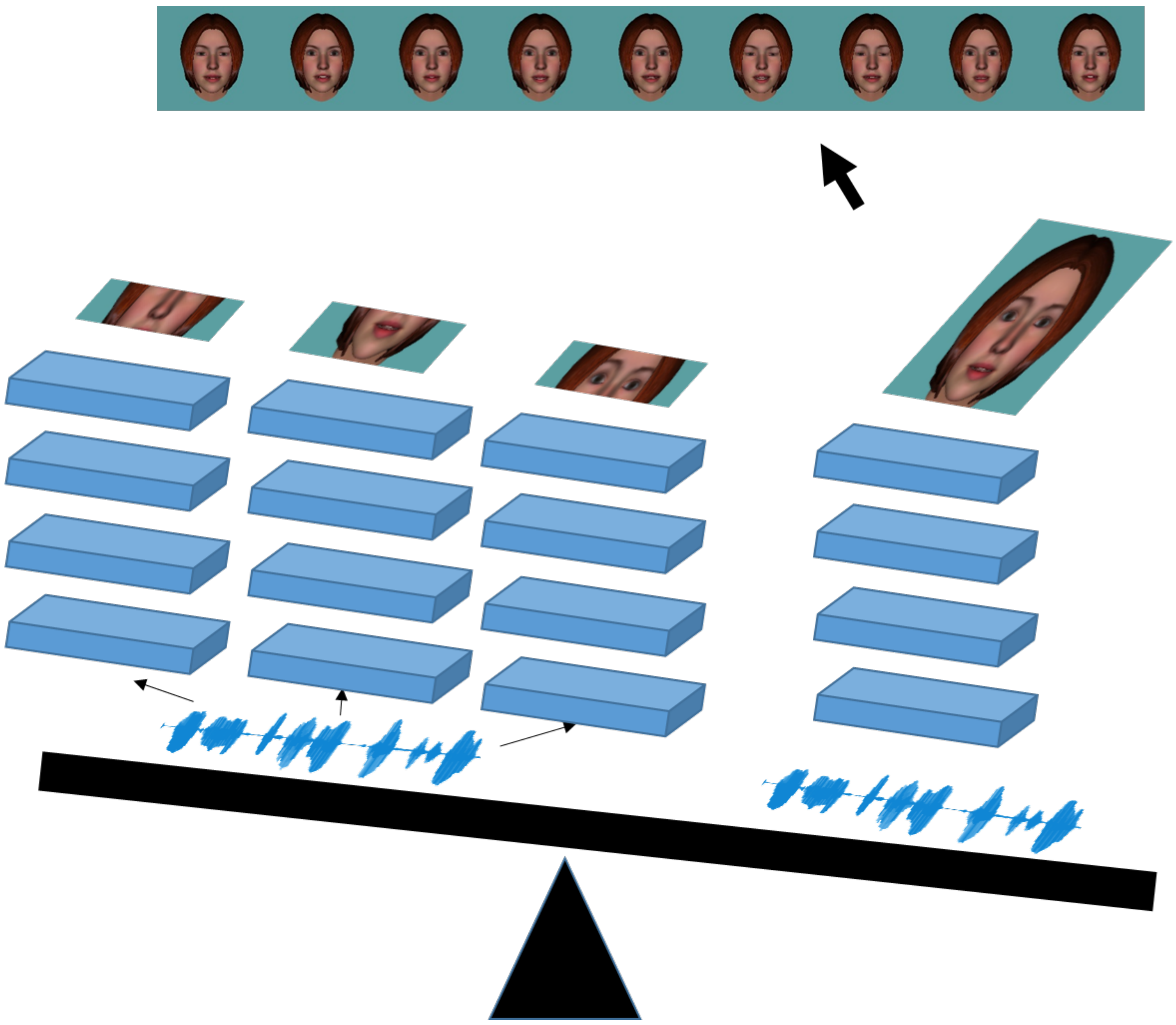
- Facial movements
  - **Articulation, emotion**, race, personality
- Articulation
  - Lower face region [Busso and Narayanan, 2007]



- Emotion
  - Upper face region
- Muscles throughout the face are connected
- Emotion manifestation through multiple regions

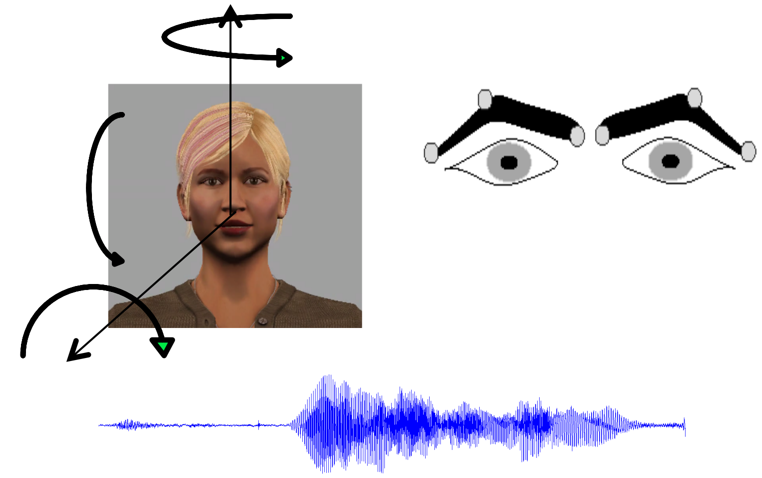
Overview

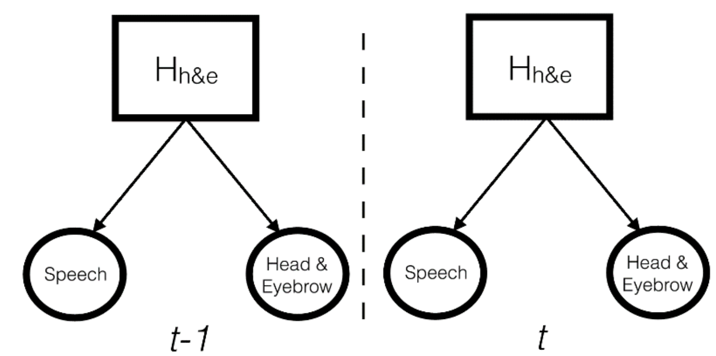
- Hypothesis: There are principled relationships between different facial regions

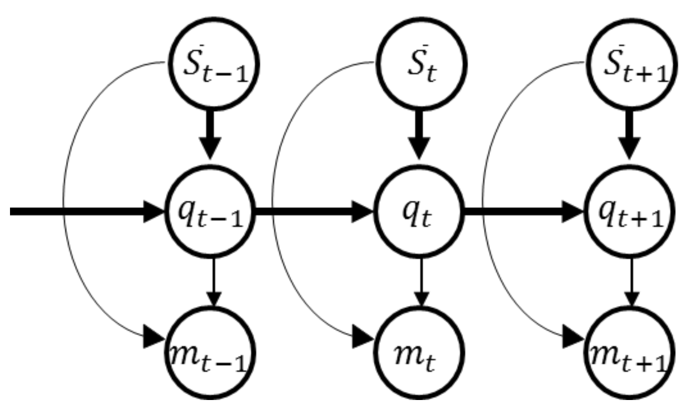


Related Work

- Joint models:
  - Eyebrow & head motion
- Generating more realistic sequences than separate models
- Mariooryad and Busso [2012]
- Ding et al. [2013]



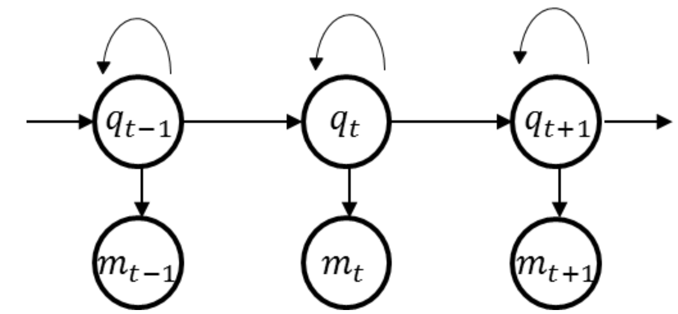


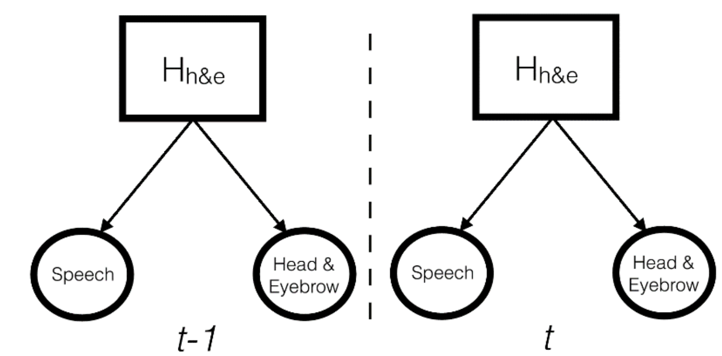


[Mariooryad and Busso 2012]

Model Selection

- HMMs, dynamic Bayesian networks:
  - Generative Models
  - Generate outputs with discontinuities
  - Require post processing smoothing





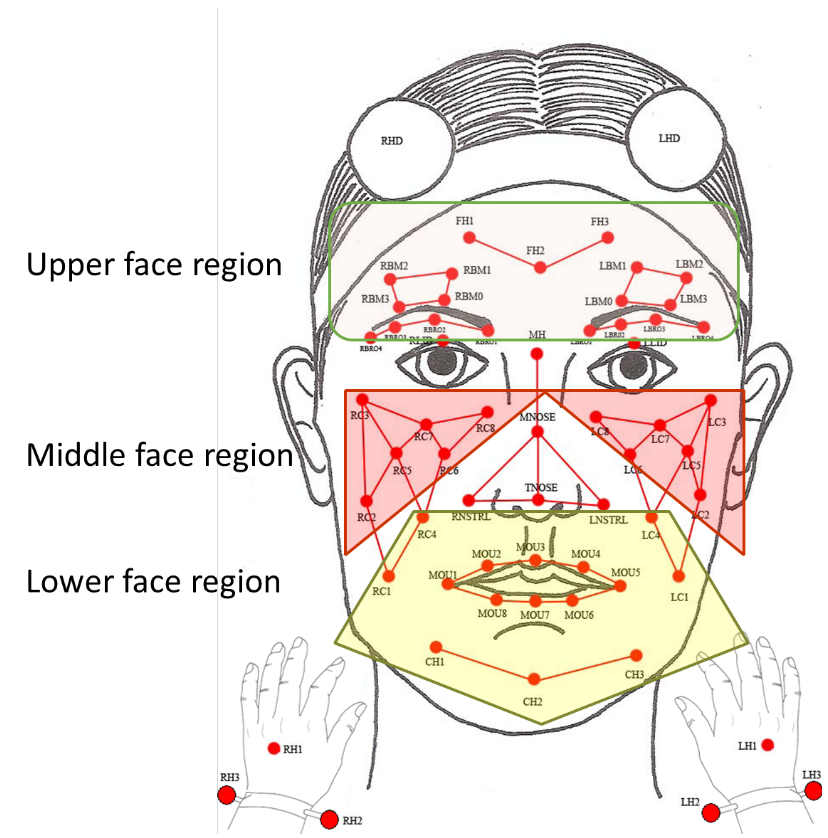
- Predictive deep model with nonlinear units:
  - Discriminative model
  - They have shown to outperform HMMs for lips movement prediction by Taylor et al.[2016], Fan et al. [2016]

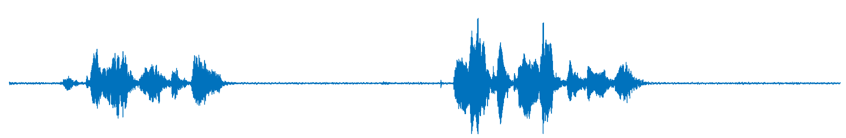
Corpus: IEMOCAP

- Video, audio and MoCap recording
- Dyadic interactions
- Script and improvisation scenarios
- 10 actors
- The position of the facial markers

Features

- 19 markers for the upper facial region
- 12 markers for the middle facial region
- 15 markers for the lower facial region

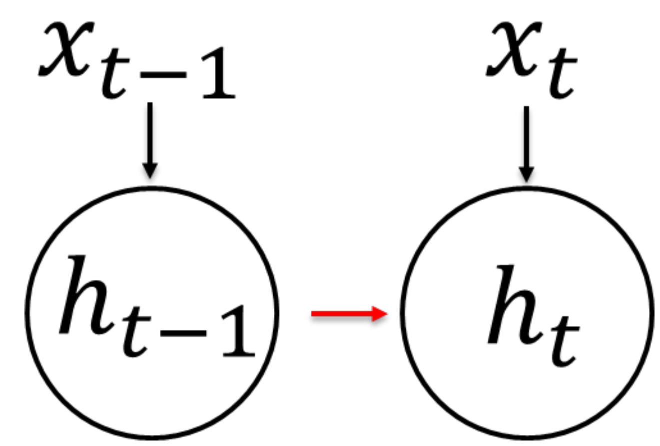




- 25 Mel-frequency cepstral coefficients (MFCCs)
- Fundamental frequency
- Intensity (25ms windows every 8.33ms)
- 17 LLDs eGeMAPS [Eyben et al., 2016]

Recurrent Neural Network

- RNNs learn temporal dependencies
  - Temporal connections between consecutive hidden units between time frames

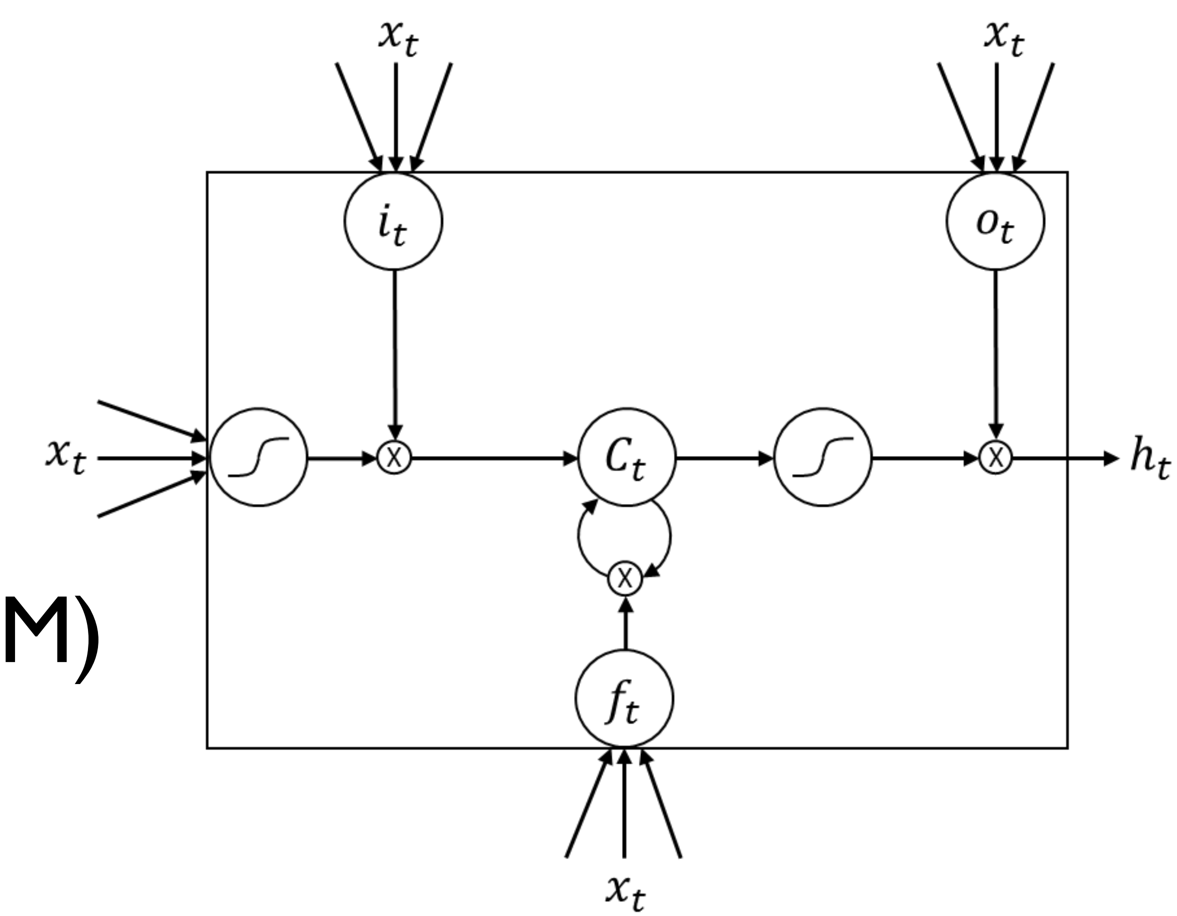


$\boxed{length(x) \Uparrow}$

$\downarrow$

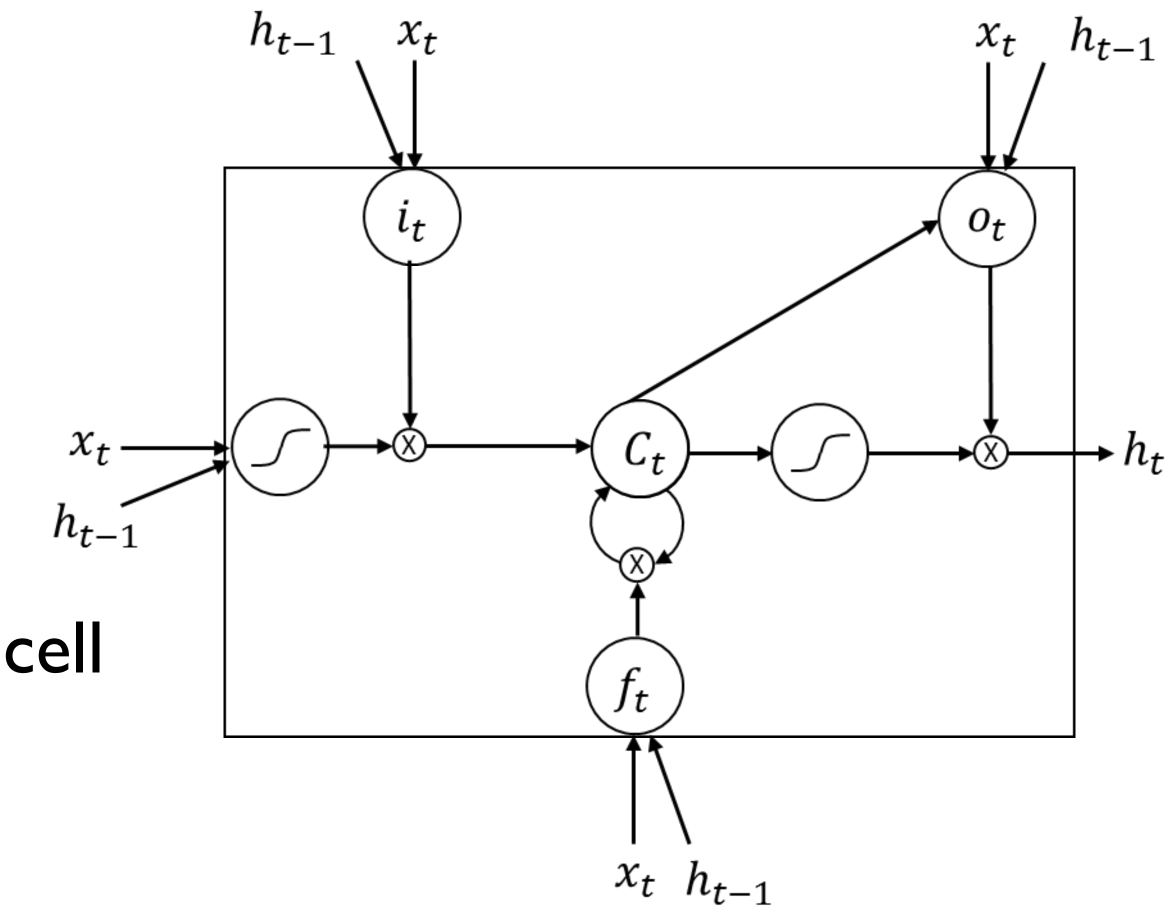
Vanishing or Exploding Grad.

- Long Short Term Memory (LSTM)
  - Extension of RNNs
  - They handle this problem



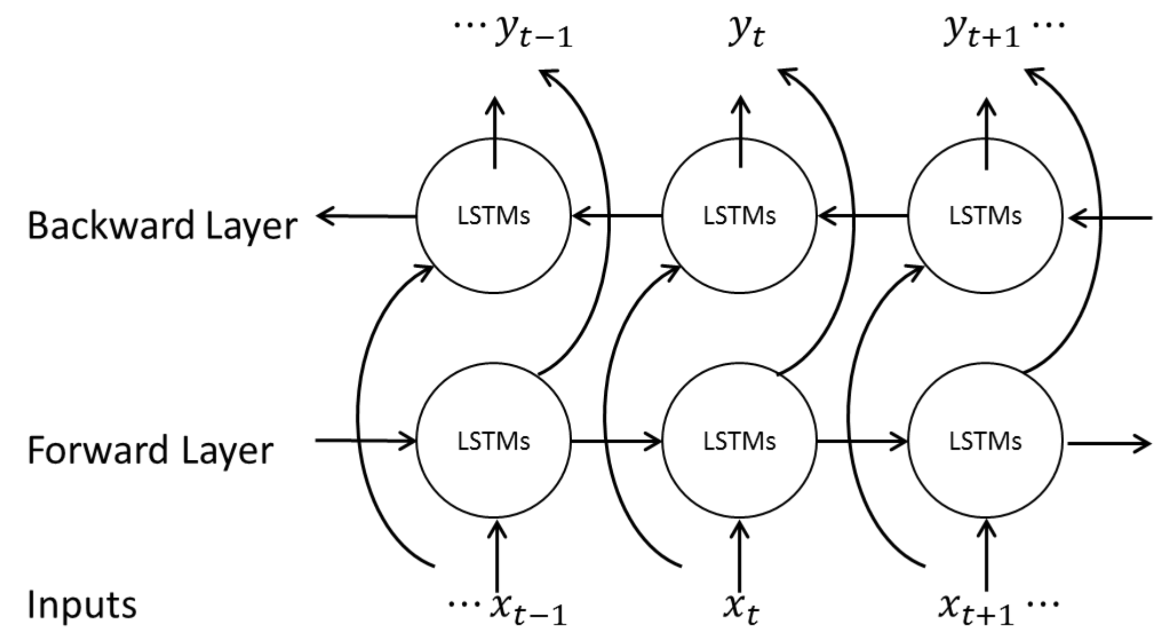
Long Short Term Memory

- LSTM utilizes a cell
- LSTM uses three gates
- Input gate:
  - How much of input to store in the cell
- Forget gate:
  - How of the previous cell being retained in the cell
- Output gate:
  - How much of cell to be used as output



$$o\ h_t = o_t \odot \tanh(C_t)\ + V_o C_t + b_o)$$

Bidirectional LSTM

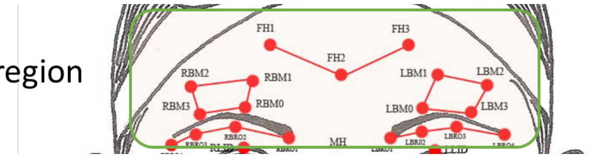


- An extension of LSTM
- Uses the previous and future frames to predict at t
- Consists of training forward and backward LSTMs
- Generates smoother movements
- Can be used in real time (post-buffer)
- We use it off-line, generating the whole turn sequence

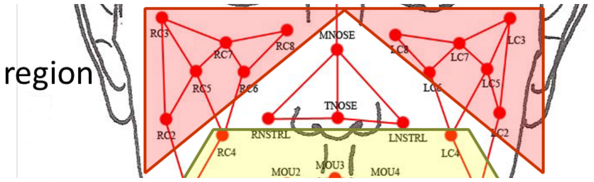
Separate Models (Baseline)

- Separately synthesize the lower, middle and upper face regions
- Independently create the facial markers trajectories for each region
- Local relationships within regions are preserved
- Possible intrinsic relationship across regions are neglected
- Assumption:
  - Relationships across the three regions are not important

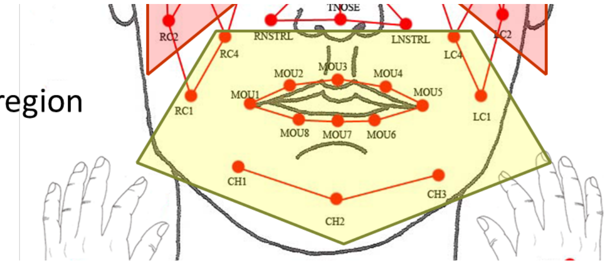
Upper face region



Middle face region

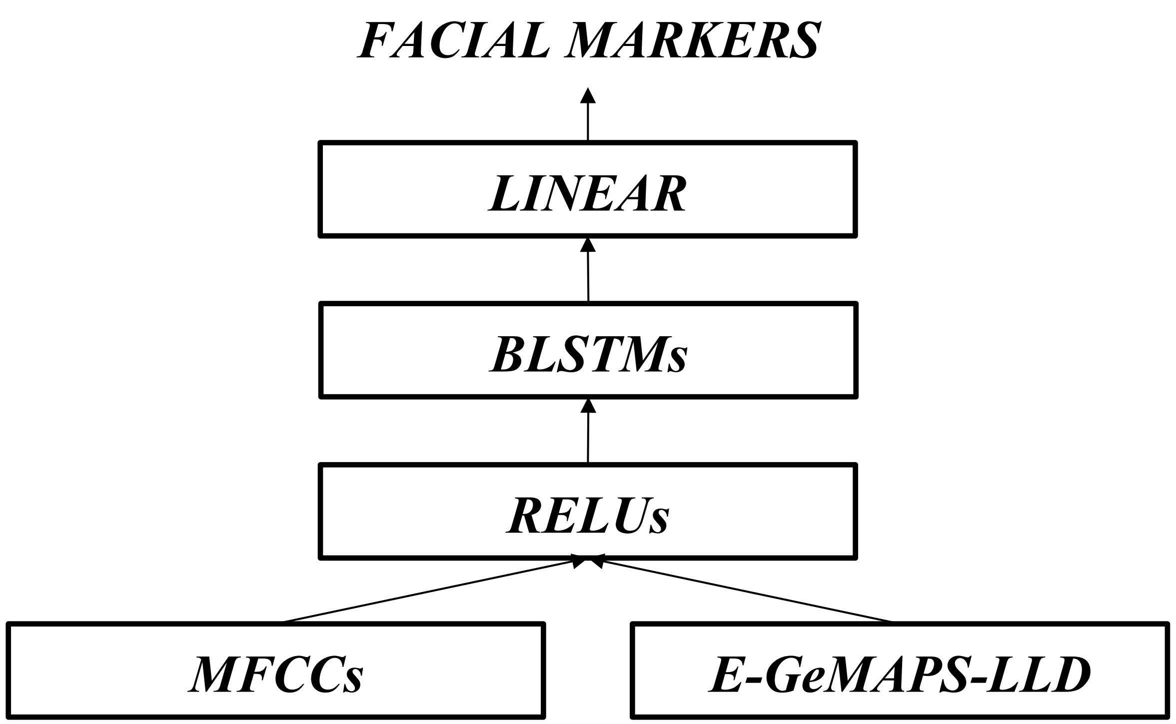


Lower face region

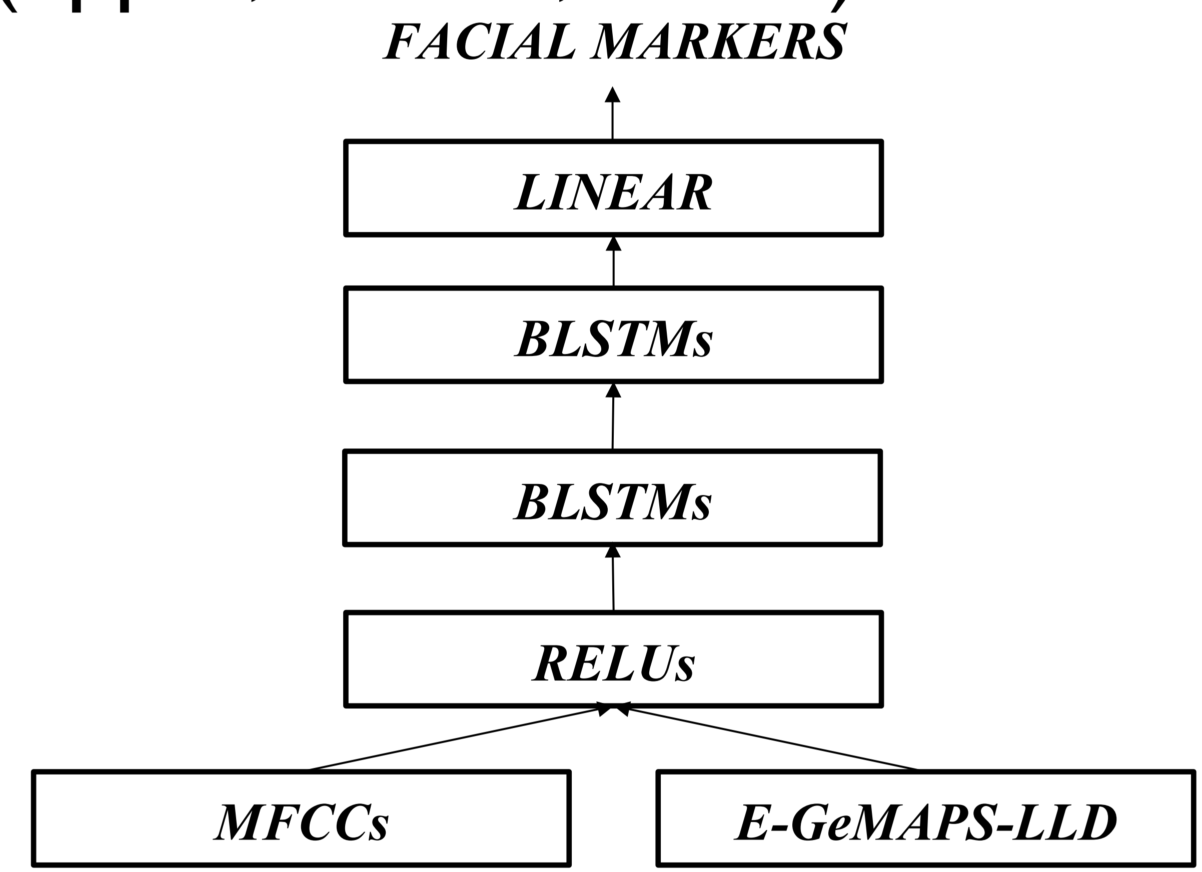


Separate Models (Baseline)

- One model per facial region (upper, middle, lower)

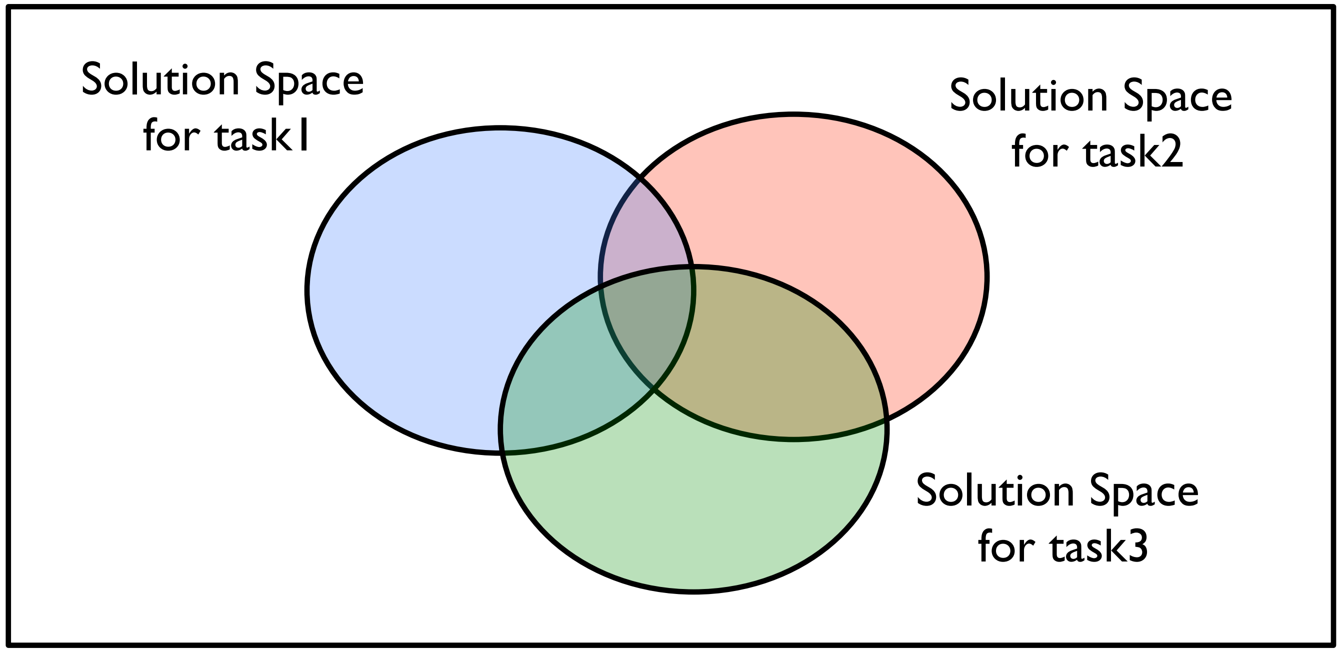


Structure 1



Structure 2

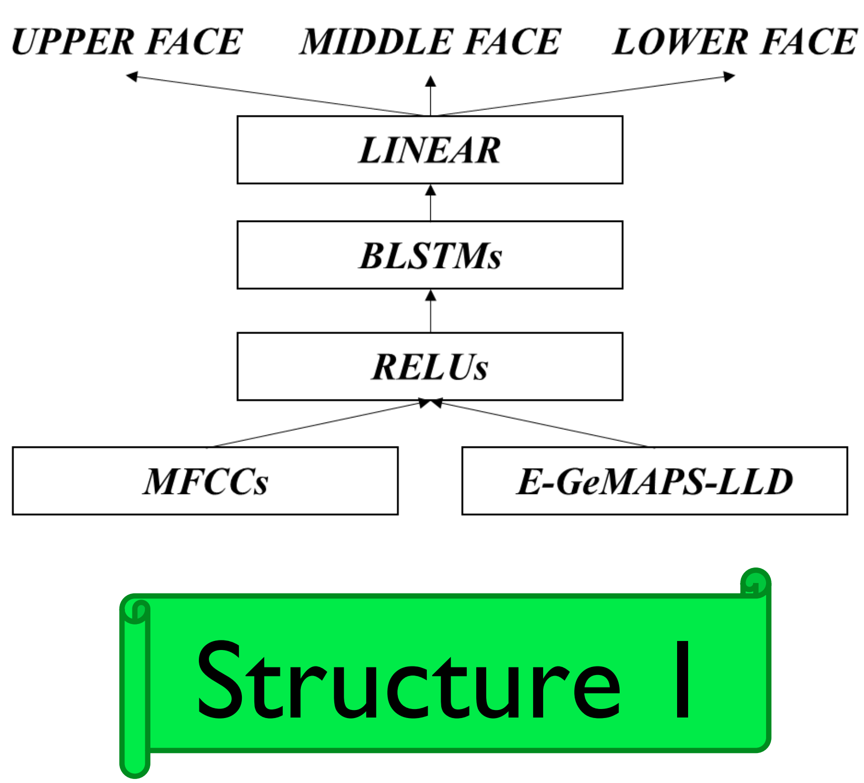
Joint Models – Multitask Learning

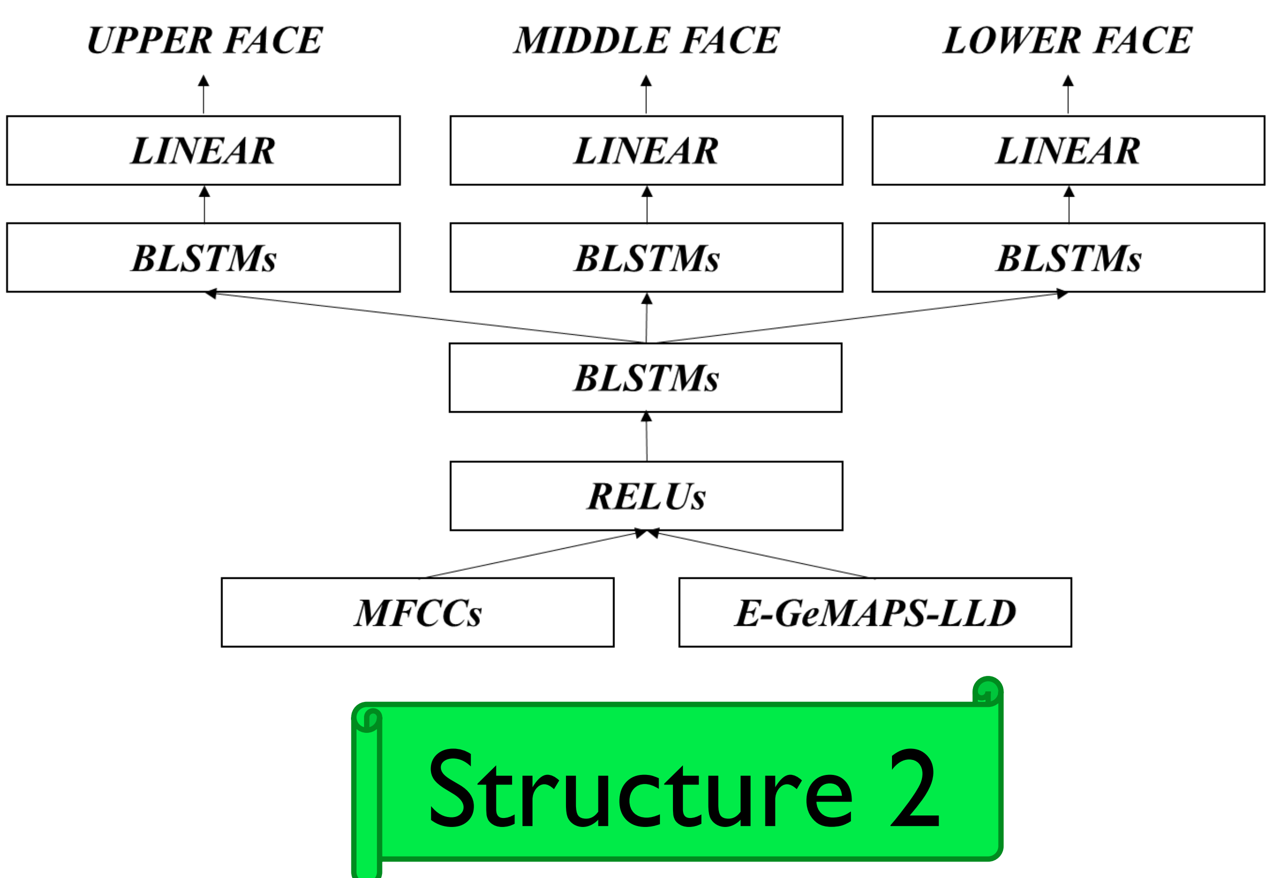


- Multitask learning
  - Jointly solve related problems using shared layer representation
- Three related tasks:
  - lower, middle and upper face movement predictions
- From a learning perspective
  - Two tasks regularize each task systematically
  - Learn more robust features with better generalization

Joint Models – Multitask Learning

- Part of the networks is shared between all the tasks
- Assumption:
  - Facial movements of different regions have principled relationships





Cost Function & Objective Metrics

- Concordance correlation coefficient
- Our objective:
  - $1\text{-}\rho_c$
- Advantage:
  - Increase correlation
  - Decrease mean square error (MSE)
  - Increase range of movements

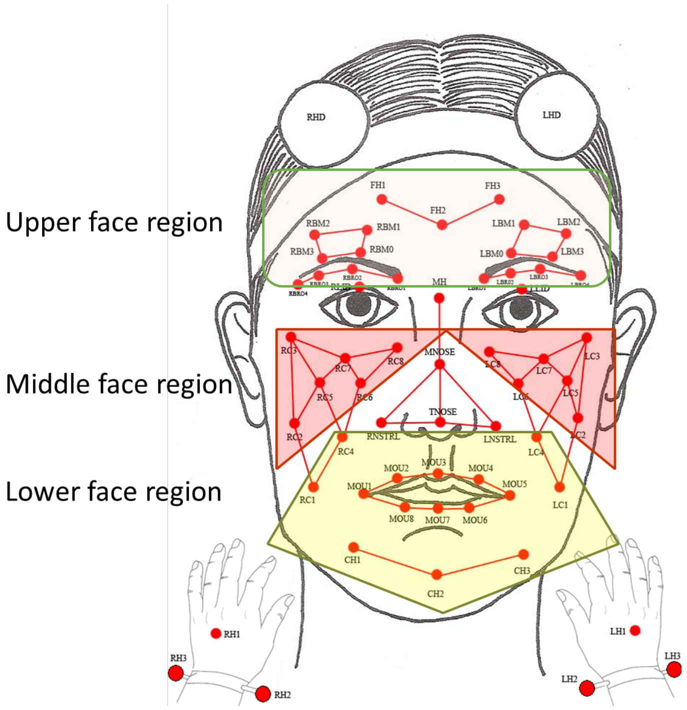
Predicted value: x

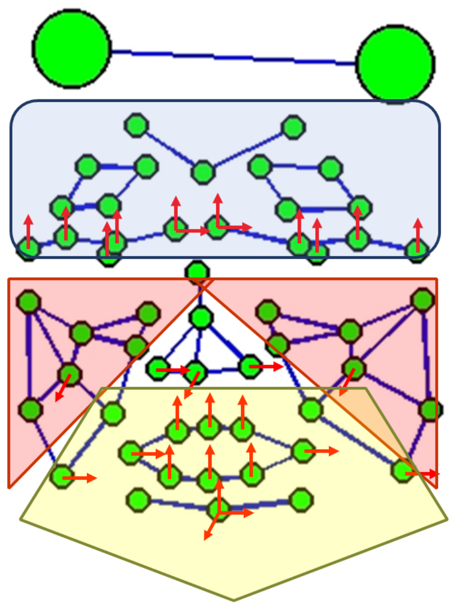
True value: y

$1\text{-}\rho_c$

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + \left(\mu_x - \mu_y\right)^2}$$

Rendering with Xface





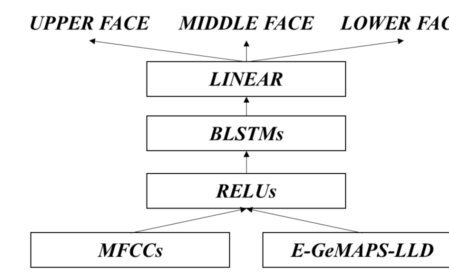
- Xface uses the MPEG4 standard to define facial points
- Most of the markers in the IEMOCAP database follow MPEG4 standard
- We follow the same mapping proposed by Mariooryad and Busso [2012]

Objective Evaluation

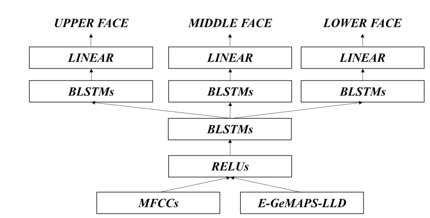
$\uparrow \rho_c$

$\downarrow$ MSE

- 60% training, 20% validation, 20% test
- Concatenate all the turns for evaluation
- ρ_c increases for most cases for the joint model
- MSE decreases for several of the cases for the joint models
- For separate model: 1024 units is better than 512 units
- Separate models require more memory



Joint-1

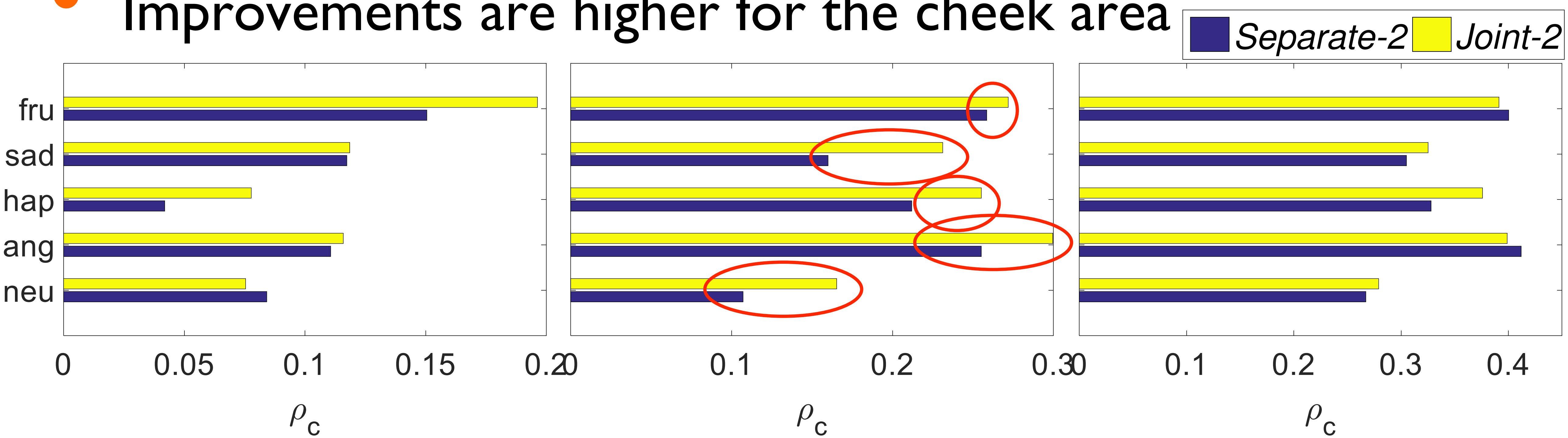


Joint-2

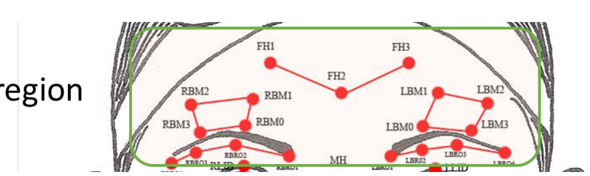
Model	# nodes per Layer	# params	Upper face		Middle face		Lower face	
			ρ_c	MSE	ρ_c	MSE	ρ_c	MSE
Separate-1	512	12.8 M	0.140	1.47	0.268	1.36	0.401	1.12
Joint-1	512	4.4 M	0.150	1.32	0.274	1.30	0.390	1.26
Separate-1	1024	50.8 M	0.149	1.41	0.277	1.16	0.411	1.05
Joint-1	1024	17.1 M	0.160	1.40	0.297	1.24	0.413	1.14
Separate-2	512	31.7 M	0.135	1.44	0.260	1.24	0.392	1.04
Joint-2	512	23.2 M	0.160	1.37	0.307	1.14	0.411	1.06

Emotional Analysis

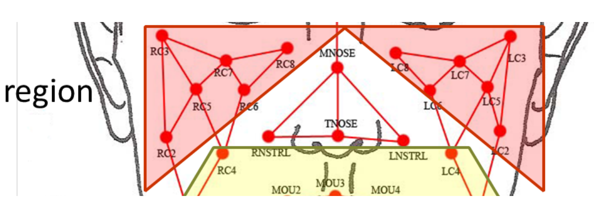
- 113 (neutral), 161 (anger), 86 (happiness), 131 (sadness), 247 (frustration)
- Separate-2 (512) vs Joint-2 (512)
- Improvements are higher for the cheek area



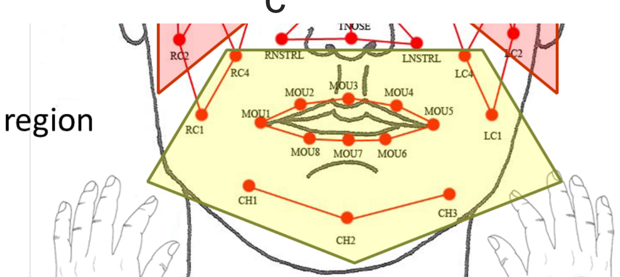
Upper face region



Middle face region

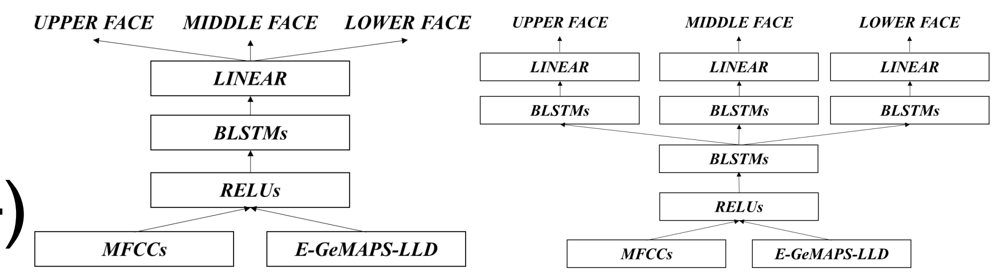


Lower face region



Subjective Evaluation

- Limit the cases for subjective evaluations (5 cases)
  - Original
  - Separate-1 (1024)
  - Joint-1 (1024)
  - Separate-2 (512)
  - Joint-2 (512)

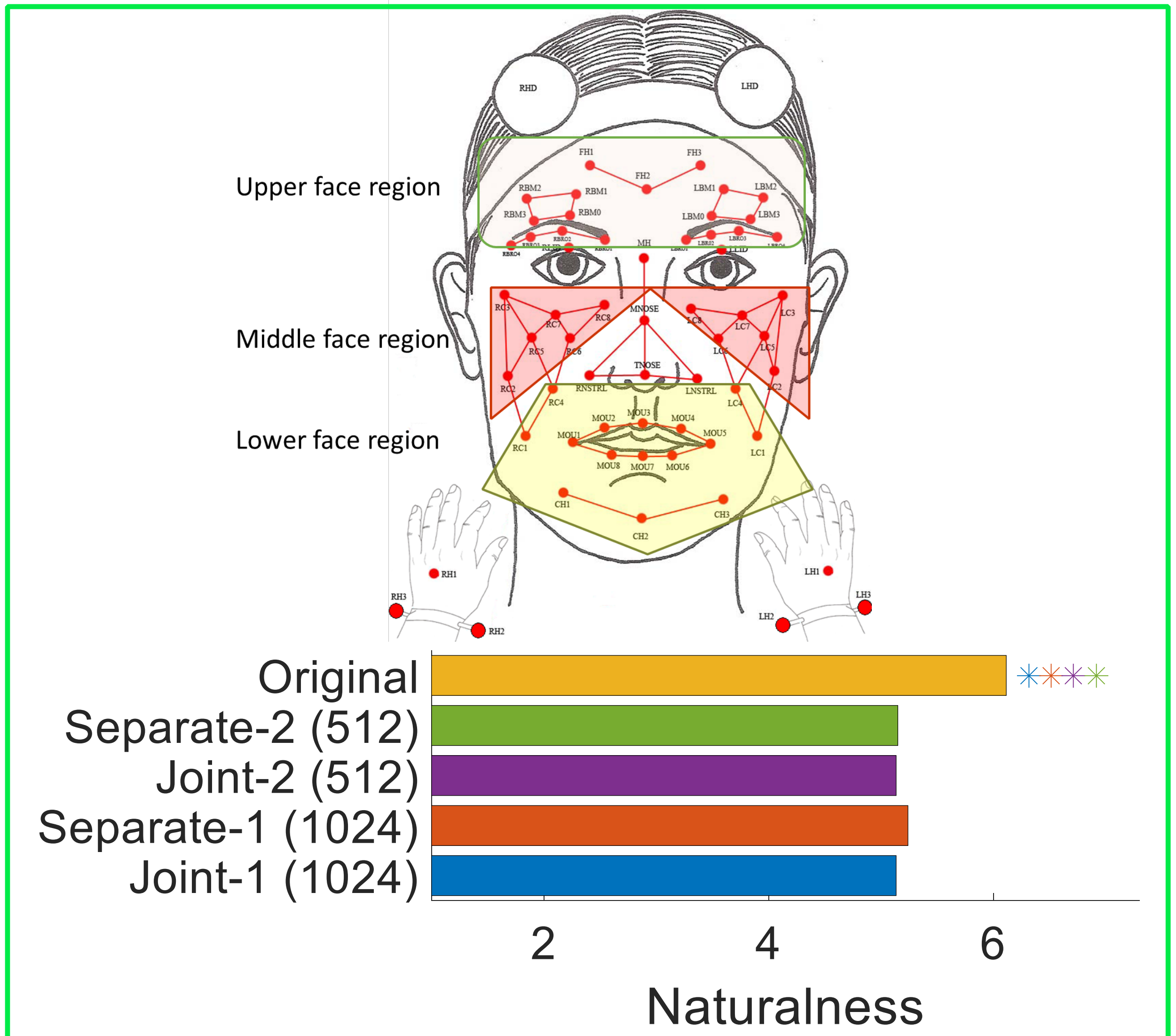


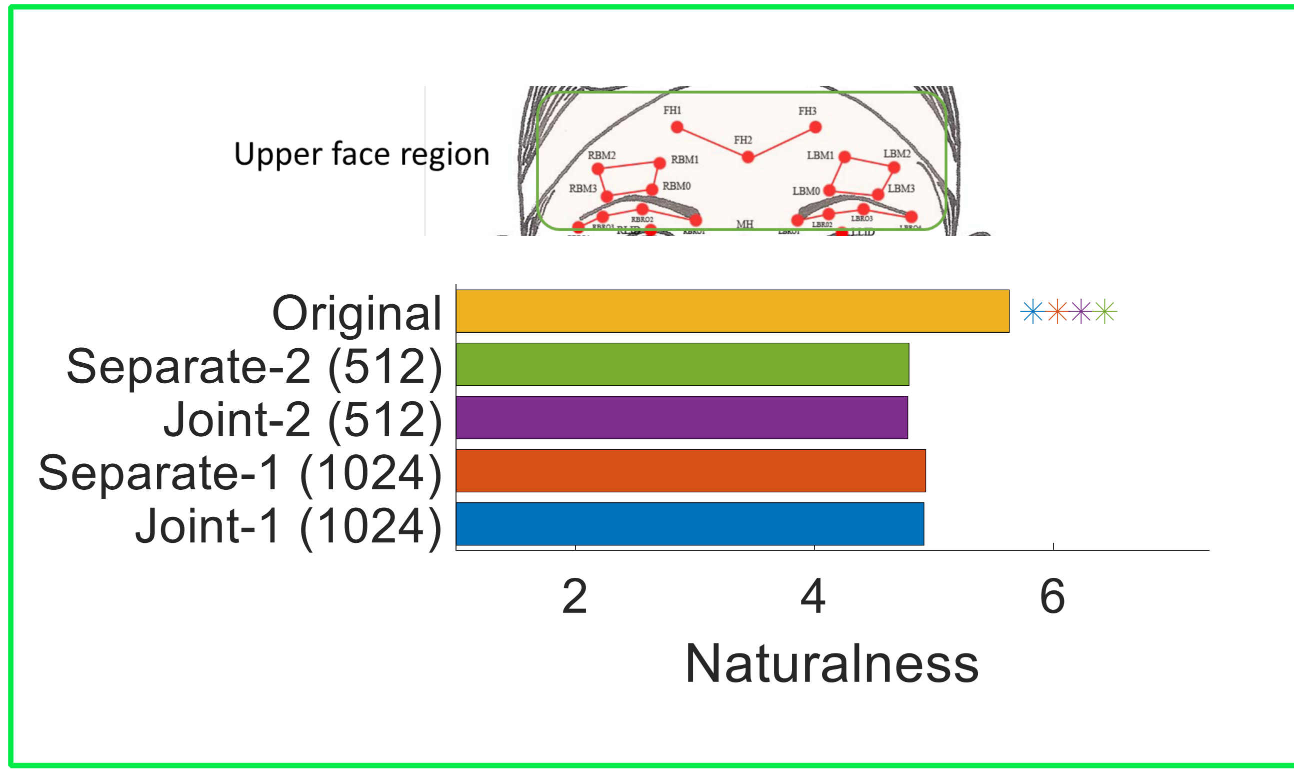
Joint-1 Joint-2

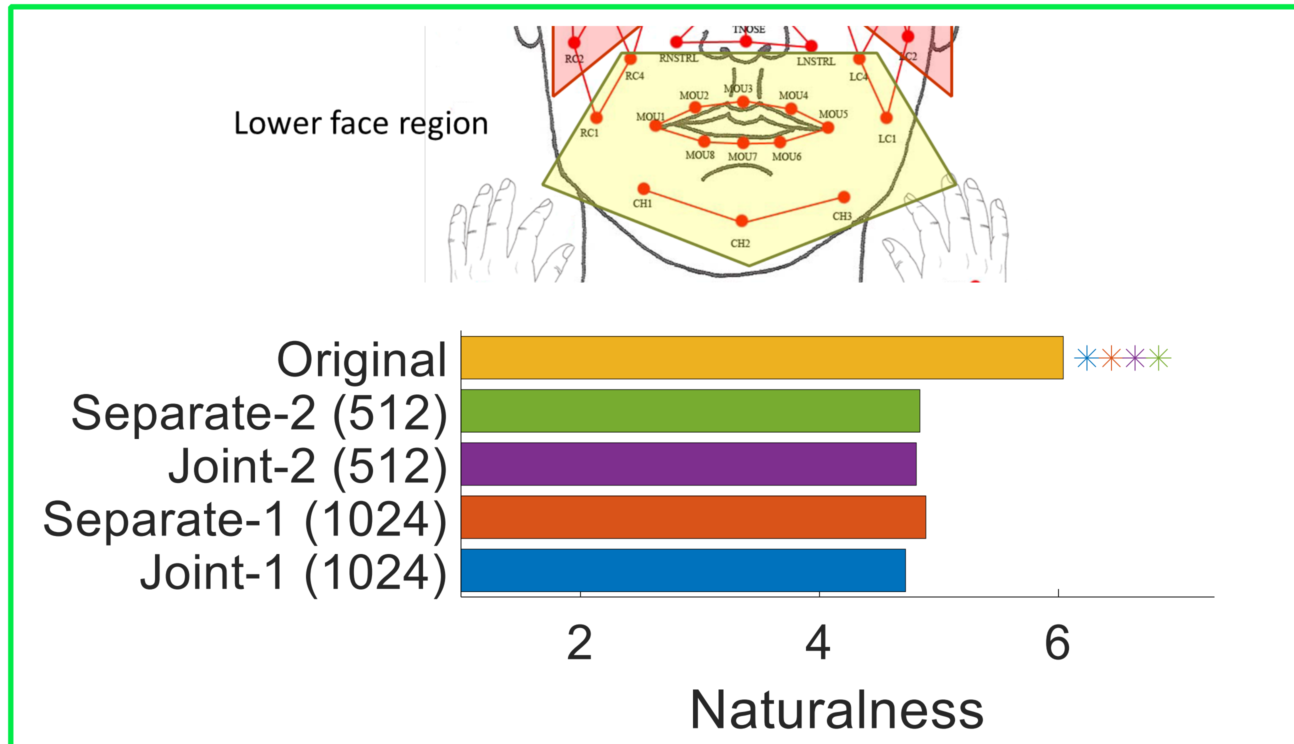
- Randomly select 10 videos (10 x 5)
- Head is still
- 20 subjects from AMT
- Naturalness scores 1-10

Subjective Evaluation

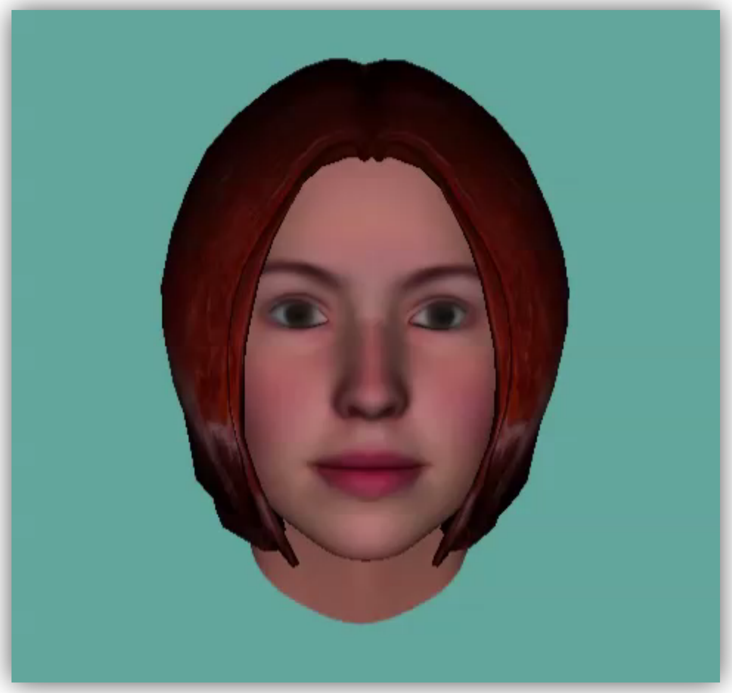
- Cronbach's alpha = 0.672







Sample videos



Original

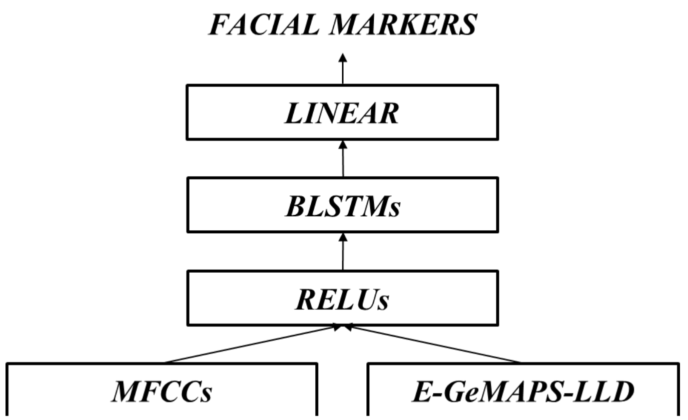
Separate-2 (512)

Joint-2 (512)

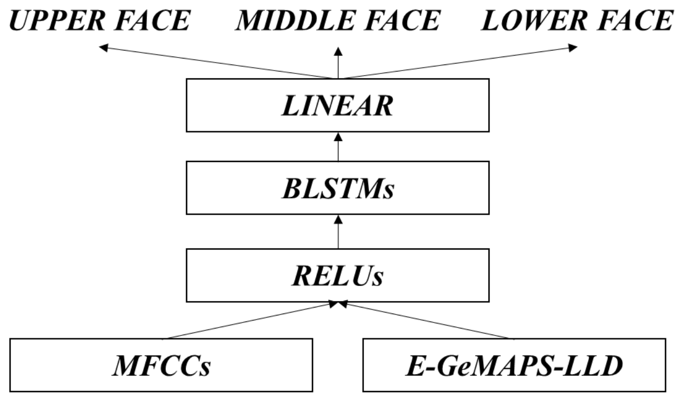
Videos

Summary

- This paper explored multitask learning with BLSTMs
- Joint models jointly learn:
  - The relationship between speech and facial expressions
  - The relationship across facial regions, capturing intrinsic dependencies
- Baseline: models that separately estimate movements for different facial regions



Separate model



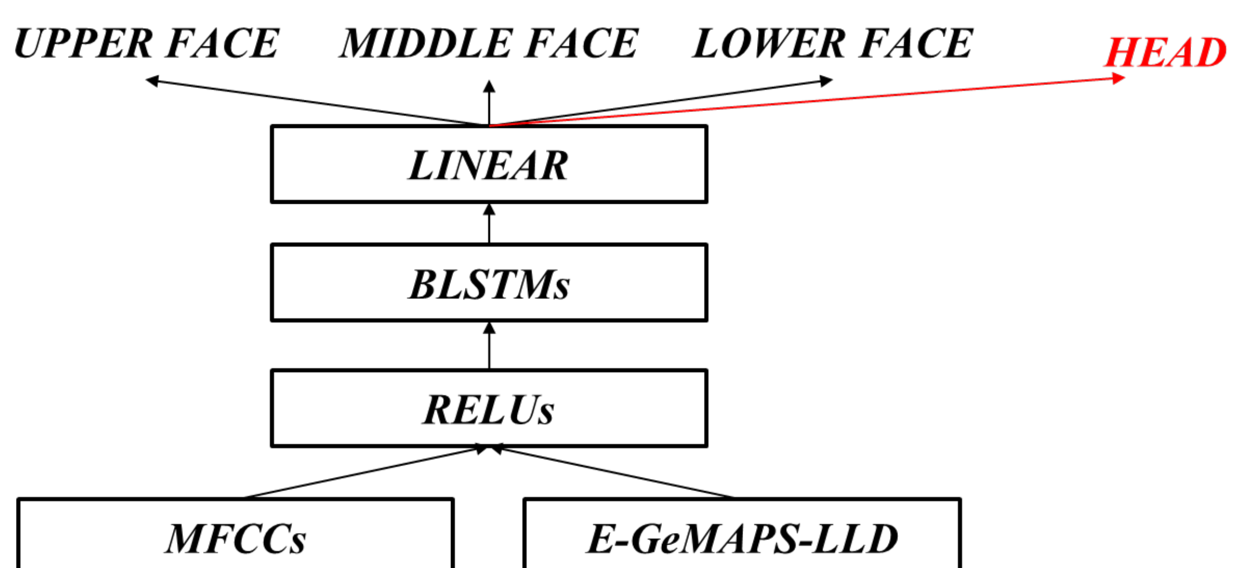
Joint model

Conclusions

- Objective evaluation showed improvements for the joint models in different facial regions
- The improvement are higher for the Joint-2 model, which has shared layers and task specific layers
- Sharing the layers reduces the number of parameters
- Subjective evaluations did not reveal any significant difference between the joint and separate models
- We believe that this result is due to the lack of expressiveness of Xface

Future works

- We will explore more sophisticated toolkits to present our results, including photo realistic videos [Taylor et al., 2016]
- We will also evaluate generating head motion driven by speech as an extra task in the multitask learning framework
- We will explore more advanced modeling strategies to better learn the relationships between speech and facial movements



Questions?

This work was funded by NSF grants (IIS: 1352950 and IIS: 1718944)