USC University of Southern California

USC Viterbi School of Engineering

Ming Hsieh Department of Electrical Engineering

A Hierarchical Framework for Modeling Multimodality and Emotional Evolution in Affective Dialogs

Angeliki Metallinou, Athanasios Katsamanis and Shrikanth Narayanan
Signal Analysis and Interpretation Lab (SAIL), USC

Motivation

- Model emotional evolution
  - Within emotions
  - Between emotions, speakers
- Context-Sensitive
- Multimodal Framework
- Flexible and Extensible

Database & Features

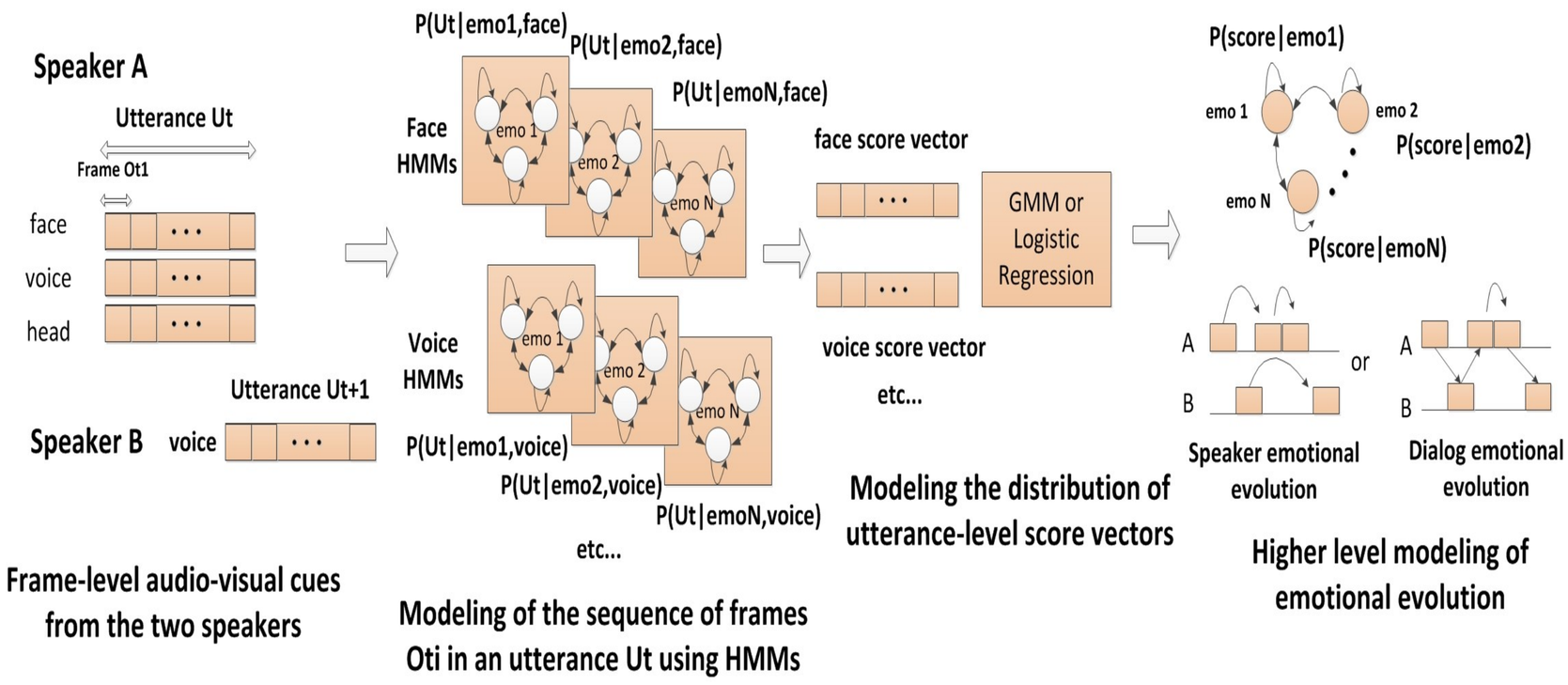
- IEMOCAP
  - sail.usc.edu/iemocap/
  - Affective, dyadic, naturalistic
  - Emotion annotations
    - Categorical, dimensional
- Speech
  - Pitch, energy, MFCC
- Facial Expressions
  - MoCap
  - Face normalization
  - PFA

- Head, hand movement

Results

- Valence: face+speech+context
- Activation: speech+head movement

Framework Overview



Utterance Models

- Emotion-specific HMMs
  - utterance-level
  - discriminatively trained

Score Vector Models

- Estimate log-likelihoods of utterance HMMs λ

$$s_{it} = \log P(U_t|\lambda_i), i = 1, .., N$$

- Model score distribution

$$P([s_{it}]_{i=1}^{N}|emo_i)$$

- Exploit relations between emotions
- Normalize score vector

Context Modeling

- Speaker emotional flow
- Dialog emotional flow
- Higher-level HMM
  - States are emotions emo_i
  - Emission: $P([s_{it}]_{i=1}^{N}|emo_i)$
  - Transition probabilities:
    - $P(emo_i|emo_j)$

Multimodal Fusion

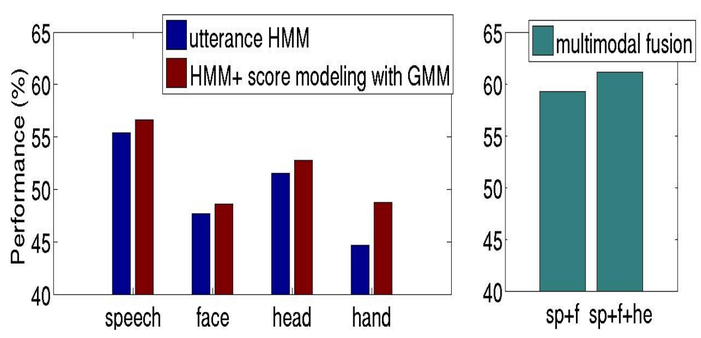
- Feature level
- Model level

$$s_t = [\log P(U_t|\lambda_i)^f, \log P(U_t|\lambda_i)^v]_{i=1}^{N}$$

- Score level

$$w_f \cdot \log P([s_{it}^f]_{i=1}^{N}|emo_i) + w_v \cdot \log P([s_{it}^v]_{i=1}^{N}|emo_i)$$

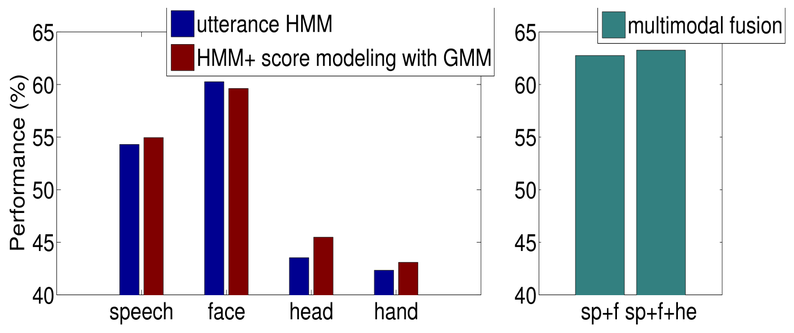
Activation (3 levels)



f+sp: HMM+GMM+fuse(fv):score-level fusion

f+sp+he: HMM+GMM+fuse(fvh):score-level fusion

Valence (3 levels)



f+sp: HMM(fv)+GMM: feature-level fusion

f+sp+he: HMM(fv)+GMM+fuse(h): feature and score-level fusion

Valence Context Modeling

	speaker	classifier	modeling	F1(m ± std)
No higher level modeling	audio-visual	HMM(fv)+GMM+fuse(h)	--	63.26 ± 4.05
	audio	HMM + GMM(v)	--	52.75 ± 2.11
		in total		58.92 ± 2.65
Speaker modeling HMM 1st order	audio-visual	HMM(fv)+HMM, fuse(h)	HMM^{1st}_{sp}	66.09 ± 3.39
	audio	HMM+HMM(v)	HMM^{1st}_{sp}	54.88 ± 3.52
		in total		61.30 ± 2.93
Dialog modeling HMM 2nd order	audio-visual	HMM(fv)+HMM,fuse(h)	HMM^{2nd}_{dial}	62.01 ± 2.14
	audio	HMM+HMM(v)	HMM^{2nd}_{dial}	57.14 ± 2.80
		in total		59.98 ± 2.61
Mixed modeling	audio-visual	HMM(fv)+HMM,fuse(h	HMM^{1st}_{sp}	66.08 ± 3.39
	audio	HMM+HMM(v	HMM^{2nd}_{dial}	57.14 ± 2.80
		in total		62.31 ± 2.18

For questions contact: metallin@usc.edu

Ming Hsieh Institute
Ming Hsieh Department of Electrical Engineering