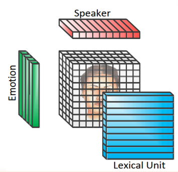
Factorizing Speaker, Lexical and Emotional Variabilities Observed in Facial Expressions

Soroosh Mariooryad and Carlos Busso
Multimodal Signal Processing (MSP) Laboratory
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083, U.S.A.

Motivation

- **Variabilities in Facial Expressions:**
  - **Speaker** (i.e., who is speaking)
    - Intrinsic cultural, physiological and idiosyncratic characteristics
  - **Lexical Content** (i.e., what is being spoken)
    - Underlying articulatory process
  - **Emotional Content** (i.e, how is being spoken)
    - Externalization of emotional cues
- **Goals:**
  - Decode the variability in the face
  - Propose solutions for robust emotion recognition systems



Methodology

Database (IEMOCAP) [Busso et al. 2008]

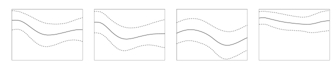
- ~12 hours of data, read, scripted and spontaneous
- Speech and motion capture markers (53)
- **Speaker**
  - 10 speakers (5 male, 5 female)
- **Lexical Content**
  - The 10 most frequent syllables and words

Syllables	AY	Y_UW	AX	N_OW	T_AX	AX_T	L_AY_K	DH_AX	G_OW	AX_N_D
Words	I	YOU	KNOW	A	TO	THE	LIKE	AND	DO	ME

- **Emotional Content**
  - The four most frequent emotions (Happiness, sadness, anger and neutral)

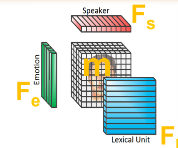
Trajectory Model (marker m)

- Interpolation-Resampling
- Mean trajectory (μ_m)
- Variations (Σ_m)
- Models for word "WELL"



(a) Neutral (b) Happy (c) Angry (d) Sad

Factor Analysis



- m = markers
- F = factors
  - speaker, lexical and emotional contents
- **Goal:** Measure the contribution of the factors in the variability of the features
- Mutual Information

$$IG(m, F) = H(m) - \sum_{f \in F} P(f) H(m \mid f)$$

- Proposed Relevance Measure (RM)

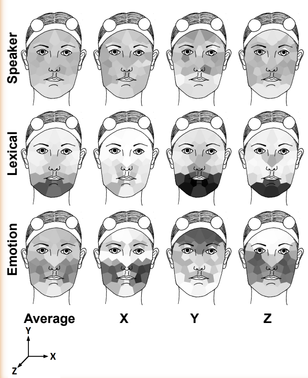
$$RM(m, F) = tr(\Sigma_m) - \sum_{f \in F} P(f) tr(\Sigma_m \mid f)$$

- Normalizing to compensate for different initial uncertainties

$$RM_n(m, F) = \frac{RM(m, F)}{tr(\Sigma_m)}$$

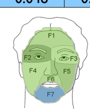
Factor Analysis Results

Distribution of the factors (lexical-independent): $RM_n(m, F)$



Div#	Syllable Level			Word Level		
	Speaker	Syllable	Emotion	Speaker	Word	Emotion
F1	0.068	0.014	**0.069**	0.070	0.016	**0.071**
F2	**0.053**	0.014	**0.053**	0.056	0.015	**0.057**
F3	0.033	0.013	**0.063**	0.035	0.015	**0.064**
F4	0.075	0.031	**0.107**	0.077	0.038	**0.102**
F5	0.080	0.032	**0.113**	0.081	0.040	**0.109**
F6	0.062	0.073	**0.117**	0.063	0.089	**0.114**
F7	0.038	**0.153**	0.048	0.040	**0.184**	0.043

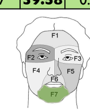
Emotion (F1–F6)
Lexical (F7)



The effect of lexical-dependent models

- Δ(%) = The difference of $RM_n(m, F)$ in lexical-independent and lexical-dependent

Div#	Syllable Level		Word Level	
	Emotion	Δ(%)	Emotion	Δ(%)
F1	0.070	1.44	0.069	-2.28
F2	0.053	0.00	0.053	-7.01
F3	0.068	7.93	0.063	-1.58
F4	0.115	7.47	0.103	0.98
F5	0.122	7.56	0.111	1.83
F6	0.123	5.12	0.115	0.87
F7	0.067	**39.58**	0.063	**46.51**



Conclusions

- **Conclusions:**
  - Emotion mostly affects the middle and upper face regions
    - Lexical independent model
  - Lexical influence is localized in the orofacial region
    - Constraining on the lexicon increases emotion variability
    - Lexical dependent model
- **Future Directions**
  - Fusing lexical dependent and lexical independent models
  - Find suitable lexical unit (e.g., visimes instead of syllables)
    - Finding lexical unit with similar trajectories (e.g., clustering)
- **References:**

C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, December 2008