JOINT ANALYSIS OF THE EMOTIONAL FINGERPRINT IN THE FACE AND SPEECH: A SINGLE SUBJECT STUDY

Carlos Busso and Shrikanth S. Narayanan

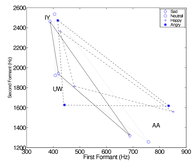
Speech Analysis and Interpretation Laboratory (SAIL)
Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089
busso@usc.edu, shri@sipi.usc.edu

Motivation

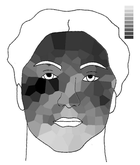
- Different communicative modalities are used to encode the affective states.
  - Speech, facial expression, head motion, and body posture.
- The same channels are simultaneously used to convey other communicative goals.
  - Linguistic, emotional, social and physiological goals.

Previous Work

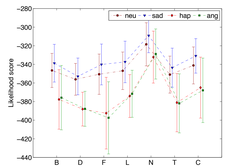
- Articulatory: low vowels (i.e., /a/) vs. high vowels (i.e.,/i/) [1].
- Facial: upper face region (i.e., forehead) vs. lower face region (i.e., lips) [2].
- Spectral: front vowels vs. nasal sounds [3].



(a) Articulatory (F1 vs. F2) [1]



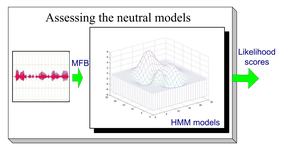
(b) Facial (Neutral vs. Happiness) [2]



(c) Spectral (likelihoods)[3]

Spectral-based neutral models *(Busso et al., 2007 [3])*

- Original goal: emotion recognition.
- HMM models trained with MFB.
- Models for broad phonetic classes.
- Output: likelihood score (decoding).
- Measurement of similarity with neutral speech.



	Description	Phonemes
F	Front vowels	iy ih eh ae ix
B	Mid/back vowels	ax ah axh ax-h uw uh ao aa ux
D	Diphthong	ey ay oy aw ow
L	Liquid and glide	l el r y w er axr
N	Nasal	m n en ng em nx eng
T	Stop	b d dx g p t k pcl tcl kcl qcl bcl dcl gcl q epi
C	Fricatives	ch j jh dh z zh v f th s sh hh hv
S	Silence	sil h# #h pau

- Fingerprint in these spectral features is stronger in some broad phonetic class.
- What happens in other channels (e.g., pitch, energy, face)?

Hypotheses

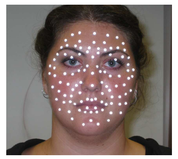
- There is interplay between linguistic and affective goals expressed in various communicate channels.
- When some channels are used to fulfill linguistic goals, other modalities with less restrictive articulatory constraints are used to convey affective goals.

Proposed Method

- Project the phonetic segmental boundaries to other communicative channels.
  - Spectral and prosodic speech features.
  - Facial expressions.
- Compare features from neutral and emotional speech.
  - Average: Ratio emotional/neutral (reported in the paper).
  - Distribution: Kullback-Leibler Divergence (KLD).
- The focus is on instantaneous behavior displayed during the phonetic boundaries.

Audio-visual database

- An actress read a corpus four times (sadness, happiness, anger, and neutral state).
- A VICON motion capture system tracked 102 facial markers (3 cameras).
- Phoneme transcription was estimated with forced alignment (HTK toolkit).

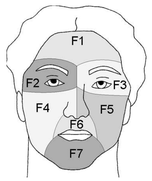


(a) Facial marker layout

(b) Motion capture system

Features

- Facial expression
  - Each marker was used as a facial feature.
  - The markers were aggregated in facial areas.
  - (F1) Forehead, (F2) Left eye, (F3) Right eye, (F4) Left cheek, (F5) Right cheek, (F6) Nasolabial, and (F7) Chin.
- Prosodic speech features
  - Pitch and RMS energy
- Spectral speech features
  - Likelihood score values (neutral models)

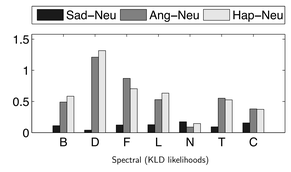


Experimental Results

Longer bars mean larger KLD (i.e., stronger differences between the distributions).

Spectral speech features

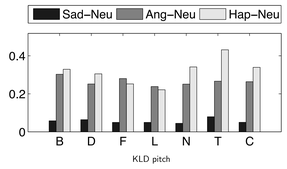
- Emotions with a high level of arousal present strong differences for vowels (F, B).
- The differences disappear for phonetic classes such as nasal sounds (N).
- Physical constraints in the articulatory domain restrict the degrees of freedom.



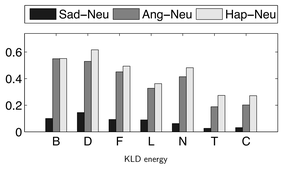
Spectral (KLD likelihoods)

Prosodic speech features

- These features predominantly describe the source of speech.
- Pitch: emotional modulation for stop and fricatives (T, C) is strong.
- Energy: the KLD for vowels (B, D, F) are higher than other classes.



KLD pitch



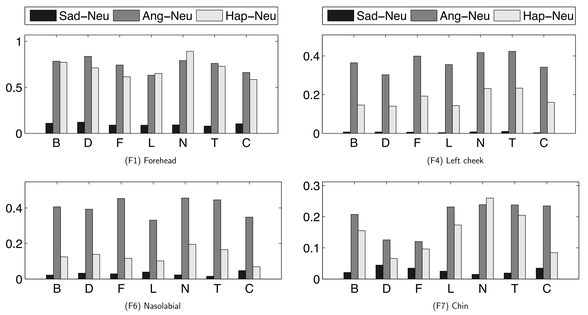
KLD energy

Facial expression

The displacement coefficient is defined to quantize the facial activeness

$$\Psi_u = \frac{1}{T_u} \sum_{i=1}^{T_u} D_{eq}(\vec{X}_i^u, \vec{\mu}^u) \quad (1)$$

- Nasal sounds present stronger emotional modulation in happiness and anger.
- Emotional modulation in upper facial region is higher than in orofacial area.
- Acoustic domain: happiness is the emotion with stronger modulation.
- Facial domain: anger is the emotion with stronger modulation.
- Different modalities are used to emphasize happiness and anger.



Discussion and conclusions

- The paper presents evidences about the emotional encoding process.
- Facial expression and pitch present stronger emotional modulation when the articulatory configuration does not have enough freedom to express emotions.
- Emotional bits are assigned to the modalities that are less constrained by other communicative goals.
- Emotional assignment compensates temporal limitation seen in other modalities.

Future work

- Analyze other phonetic descriptions to link acoustic and visual modalities.
  - Manner of articulation (i.e., fricative, stop).
  - Place of articulation (i.e., bilabial, dental, palatal).
- Validate the results in a database collected from more speakers (IEMOCAP).
- Design emotional models that capture the underlying relationships and interplays between modalities.

References

[1] C. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, 2004, pp. 889–892.

[2] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, December 2006, pp. 549–556.

[3] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.

Acknowledgements

This research was supported in part by funds from the NSF and Army.