EMOTION CLASSIFICATION USING SPEECH ACOUSTIC FEATURES AT VARIOUS TEMPORAL LENGTHS

Jonathan Kim, Hrishikesh Rao, and Mark A. Clements
Georgia Institute of Technology

Expeditions in Computing

Introduction

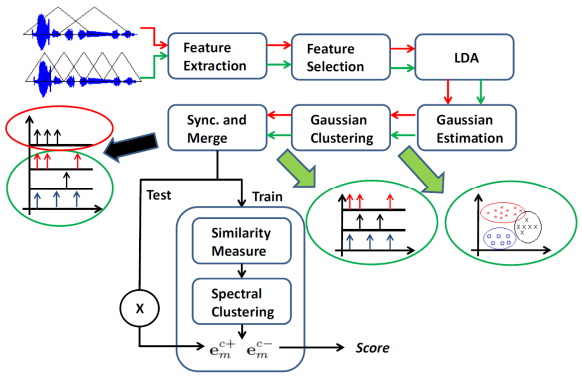
Human express their feelings through different forms of communication:

- Speech: explicit linguistic messages and implicit paralinguistic features.
- Gestures: hand motion, head orientation, etc.
- Writing: explicit messages, emoticons, etc.

Different emotional characteristics can be observed at different time scales [Busso et al. 2004].

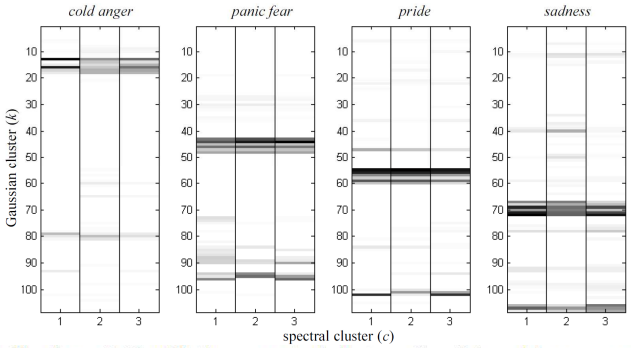
- At the phrasal level, average pitch and intensity are higher with the emotional state of *hot anger* than with other states [Sauter et al. 2010].
- At the phonemic level, spectral tilt and formant frequency amplitudes are significantly different for different emotional states [Lascarcyk et al. 2008].
- At a 30 ms analysis frame level, jitter and shimmer measurements are useful for the detection of *arousal* [Li et al. 2007].
- Moreover, other modalities use different analysis window lengths!

Proposed Method



Advantages:

- By using Gaussian clustering, it reduces the potential loss of information.
- By using the binary representation, the fusion of feature representation becomes an easy problem.



The visualization of $\mathbf{e}_m^{c+}$ with three spectral clusters $C = 3$ in cold anger, panic fear, pride, and sadness.

Database

The Geneva multi-modal emotion portrayals (GEMEP).

- 1,260 emotional utterances.
- 12,512 analysis frames with a 400 ms analysis window length.
- **12 emotional categories:** *amusement, anxiety, cold anger, despair, elation, hot anger, interest, panic, fear, pleasure, pride, relief, sadness.*
- **two affective dimensions:** *arousal, valence*.

Results

Experiment I: Classification without Fusion

Classification and detection results in unweighted average recall (UAR) using a Bayesian classifier (GMM) and the proposed method (BinF) at two temporal analysis lengths before fusion.

	Category		Arousal		Valence	
	UAR (%)	# mix.	UAR (%)	# mix.	UAR (%)	# mix.
GMM (400 ms)	36.7	96	76.7	16	76.0	16
BinF (400 ms)	**37.6**	72	**78.2**	12	**76.9**	12
GMM (phrase)	34.2	96	73.2	16	72.1	16
BinF (phrase)	**37.2**	36	**77.7**	12	**75.0**	12

Experiment II: Classification with Fusion

Classification and detection results in unweighted average recall (UAR) using the proposed method with fusion, and percentage points of improvement by fusion.

	Category	Arousal	Valence
UAR (%) (400 ms + phrase)	44.9	83.7	80.7
# mix.	108 (72+36)	24 (12+12)	24 (12+12)
Improvement (%)	7.3	5.5	3.8

Conclusions

An emotion classifier can improve its performance when speech is analyzed at the different timescales with fusion before a final classifier.

For a 12-way classifier, the unweighted accuracy is improved by 7.3 percentage points when compared to a system with a fixed analysis frame size.

For arousal and valence detectors, 5.5 and 3.8 percentage point improvements respectively are observed.