EVALUATION OF SYLLABLE RATE ESTIMATION IN EXPRESSIVE SPEECH AND ITS CONTRIBUTION TO EMOTION RECOGNITION

Mohammed Abdelwahab 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.

2014 Spoken Language Technology Workshop
December 7-10, 2014 • South Lake Tahoe, Nevada, USA
IEEE

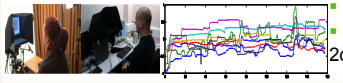
Motivation

- Evaluating the role of speaking rate in emotion recognition
- We ask two questions:
  - Can we reliably estimate syllable rate from emotional speech?
  - Does syllable rate provide complementary information over other acoustic features?
- Approach:
  - Evaluate performance of syllable estimation methods over expressive speech
  - Evaluate the contributions of syllable rate features on speech emotion recognition

Resources

SEMAINE Database

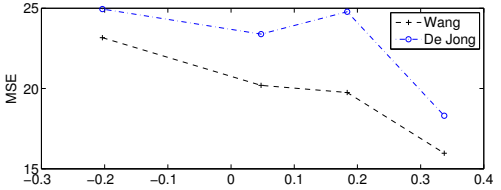
- Emotion induction with SAL
  - Sensitive artificial listener
- 24 speakers, 2830 turns
- Continuous time evaluations
  - Activation (calm vs. active)
  - Valence (negative vs. positive)
  - 6-8 raters
  - Average across time, evaluators



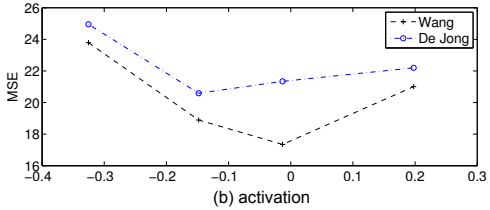
Syllable rate Estimation

- **Wang and Narayanan [2007]**
  - Extended *mrate* algorithm
    - consider only prominent sub-bands
    - added temporal
- **De jong and Wempe [2009]**
  - It detects syllable nuclei by peaks in intensity
  - Finds voiced speech using F0 contour
  - Counts number of peaks with drop of 2dB before and after peaks

Syllable Rate Estimation Analysis



(a) valence



(b) activation

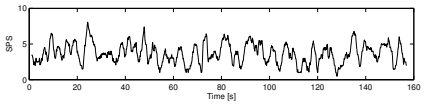
- We used forced alignment to define reference values for the syllable rate of the SEMAINE turns

$$\text{MSE}\% = \frac{\|\text{Reference rate} - \text{Estimated rate}\|^2}{\|\text{Reference rate}\|^2} \cdot 100\%$$

Emotion Recognition Evaluation

Syllable Rate Features

- Syllable rate estimation 2s windows with 20 ms steps
- 10 sentence level statistics (mean, variance, kurtosis, etc)



Feature Extraction of other acoustic features

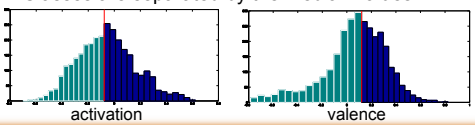
- INTERSPEECH 2011 feature set
- OpenSMILE toolkit, 4368 high level descriptors

Feature Groups

- Energy
- F0
- Voice Quality
- RASTA
- Spectral
- MFCC

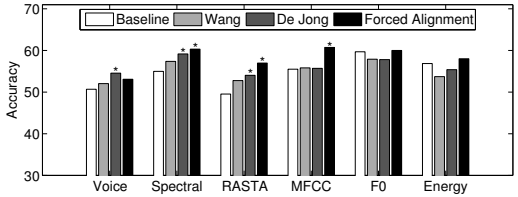
Classification problem

- Low versus high level of valence and activation
- Classes are separated by the median values

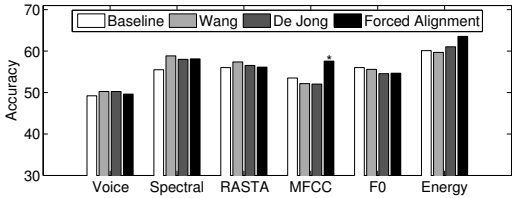


Feature Selection

- Correlation feature Selection (4368 → 400)
- Forward feature (400→50)
- We force syllable rate features



(a) valence



(b) activation

Discussion

Conclusions

- The analysis revealed a drop in accuracy for sentences perceived with low level of valence or activation
- Syllable rate features provide supplementary information
- Advances on robust speech rate estimations can benefit speech emotion recognition systems

Future Directions

- Develop a syllable rate estimation method that is robust against emotional content
- Replicate this analysis on other emotional databases with more extreme emotions

Acknowledgment

This work was funded by NSF (IIS-1217104 and IIS-1329659)