

Visual Contribution to Word Prominence Detection in a Playful Interaction Setting

Martin Heckmann

Honda Research Institute Europe, Carl-Legien-Strasse 30, 63073 Offenbach, Germany, martin.heckmann@honda-ri.de. www.honda-ri.de

Overview

Target

- Discrimination of words in wide and narrow focus condition

Scenario

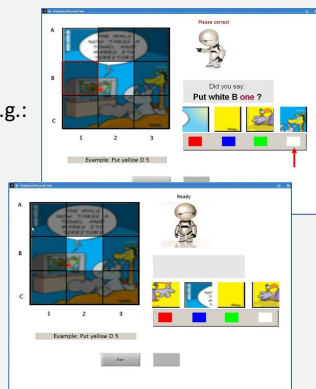
- Users interacting with a computer in a game
- Assembly of a cartoon out of tiles, e.g.:
Put green B one

- Occasional misunderstandings of the system

- Visual and acoustic feedback
- Users instructed to correct only using change in prosody

→ Turn:

- original utterance (wide focus)
- + correction (narrow focus)



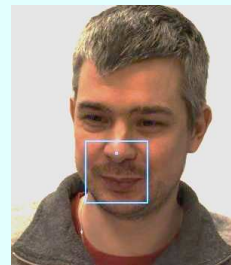
Results

- Visual features capture word prominence
- Two focus conditions can be discriminated with
 - ~80% correct acoustically
 - ~65% correct visually
 - ~85% correct audio-visually

Database & Features

Database

- Audio-visual recordings
 - Distant microphone, no visual markers
 - Video: 1280 × 1204 pixel @ 25 fps
 - Audio: 16 kHz mono
- 3 speaker, British English (BE) or British English/German (BE/G) bilingual
 - A: female, BE/G bilingual, 137 turns
 - B: male, BE/G bilingual, 146 turns
 - C: male, BE, 94 turns



Features

Audio

- dur duration of the word
- en energy relative to the mean of the utterance
- f0 mean fundamental frequency

Video

Image transformations calculated in the 80×80 pixel mouth region

→ keeping 50 coefficients with highest mean energy

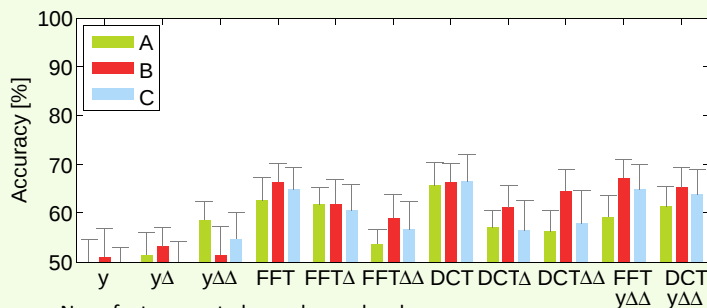
- FFT Fast Fourier Transform
- DCT Discrete Cosine Transform
- y nose y position
- $\Delta, \Delta\Delta$ first and second derivative

Classification Experiments

Classification system

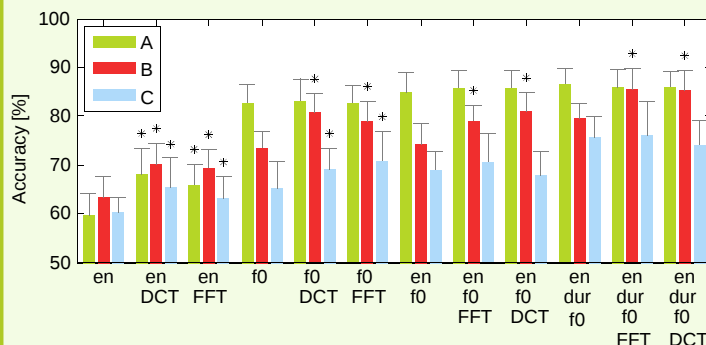
- Speaker dependent Support Vector Machine (SVM)
- (30 fold cross validation, 75% train, 25% test)

Results: Visual features



- Nose features not above chance level
- FFT or DCT ~65% correct

Results: Audio-visual features



- Significant AV gain for en+FFT or en+DCT
- With f0 significant gain for speaker B and C
- With all acoustic features significant gain only for speaker B

Conclusion

Realization of focus

- Speakers produced words differently in the two focus conditions
- Acoustic features indicate high word prominence

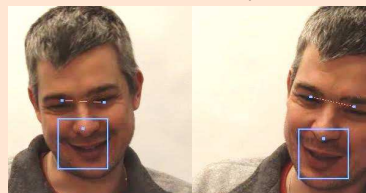
Classification Experiments

- Discrimination of two focus classes with ~65% correct for individual features (acoustic or visual)
- Exception f0: 65-83% (depending on speaker)
- Exception nose features: at chance level
- Significant AV gain when combining energy and FFT or DCT (~63% → ~69%)
- Significant gain when combining f0 or all acoustic cues with FFT or DCT only for speaker B where f0 was weaker (e.g. 79% → 86%)

→ Large speaker variation

Outlook

- Correction of head tilt to improve visual features



- More speaker ...

References

- Swerts, M. & Krahmer, E. Facial expression and prosodic prominence: Effects of modality and facial area *Journal of Phonetics*, Elsevier, 2008, 36, 219-238
- Dohen, M., Loevenbruck, H., Harold, H. & others Visual correlates of prosodic contrastive focus in French: Description and inter speaker variability *Speech Prosody*, 2006
- Cvejic, E., Kim, J., Davis, C. & Gibert, G. Prosody for the Eyes: Quantifying Visual Prosody Using Guided Principal Component Analysis *Proc. INTERSPEECH*, 2010
- Heckmann, M. Audio-visual Evaluation and Detection of Word Prominence in a Human-Machine Interaction Scenario *Proc. INTERSPEECH, ISCA*, 2012