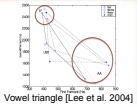
Emotion Recognition Using Synthetic Speech as Neutral Reference

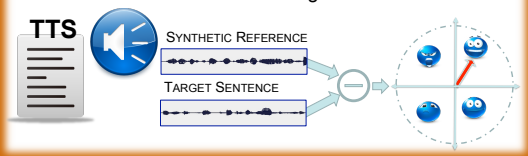
Reza Lotfian and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory
University of Texas at Dallas

Motivation


Vowel triangle [Lee et al. 2004]

- Lexical dependent models improve emotion recognition accuracy
- Practical approaches can only model small lexical units
  - Phonemes, syllables or few key words
- Can we leverage Text-to-Speech (TTS) system?
- IDEA: Synthetic speech as neutral reference model
  - Contrast different acoustic features
  - Unveil local emotional changes

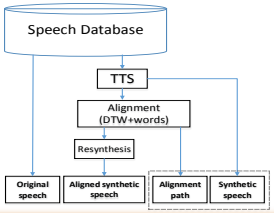


Proposed Approach

- Underlying assumptions:
  - Synthetic speech is a good representative of neutral speech
  - Lexical content (transcript) is available (ASR may required)
- TTS generates synthetic speech with same lexical content
  - Festival - cluster unit selection
- Alignment of original and synthetic speech
- Step 1: Match word boundary
  - Forced alignment
- Step 2: DTW to estimate alignment within words
  - MFCCs from original and synthetic speech

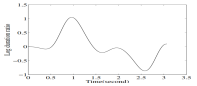
Feature set 1: Relative features

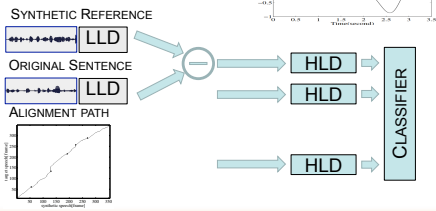
- We compare low level descriptors (LLD)
  - Original speech features
  - Aligned synthetic features
- Subtraction LLDs to generate a trace
- Estimate high level descriptors (HLD)



Feature set 2: Duration features

- Estimated from the alignment path
- Localized ratio between number of frames
- Convert speech rate signal into log scale
- Smooth the curve
  - Hamming filter

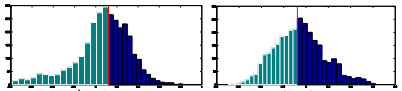




Experimental Evaluation

- Database: SEMAINE [McKeown et al., 2012]
  - Natural dyadic interaction (user, operator)
  - Sensitive artificial listener (SAL) framework
  - 10,799 1-sec segments

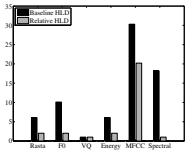
- Binary Classification Problems
  - Valence (negative versus positive)
  - Arousal (calm versus active)



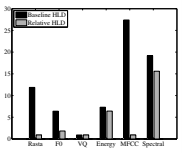
- Feature Extraction
  - OpenSMILE toolkit
  - IS 2011 set [Schuller et al., 2011]
  - 60 LLDs plus derivates (e.g., MFCCs)
  - Only 17 high level functionals
- Features set:
  - Baseline HLD: (2040=60x2x17)
  - Duration HLD: (34=1x2x17)
  - Relative HLD: (2040=60x2x17)
- Feature Selection (2-stage approach)
  - Information gain ratio (2040--> 500)
  - Correlation feature selection (500-->100)
- Classifiers: Linear kernel SVM (SMO)
  - Leave one subject out (LOSO)
  - Weighted average across speakers

Feature set	Arousal [%]	Valence [%]
Baseline HLD	63.8	59.9
Duration HLD	57.7	56.8
Relative HLD	63.3	58.6
Relative HLD & Duration HLD	63.5	60.3
Baseline HLD & Relative HLD	64.1	61.6
Baseline HLD & Relative HLD & Duration HLD	66.0	62.7

Valence



Arousal



Discussion

Results:

- Performance of Relative HLD similar to Baseline HLDs
- Best performance when all feature sets are considered
  - Absolute gain of 2.1% (arousal) and 2.8% (valence)
  - Proposed features provide complementary information
- Duration HLDs achieves performance above chances
- Over 26% of selected features come from Relative HLD

Future Directions

- Re-synthesizing of speech before estimating features
- Evaluating different speech synthesis approaches
- Building a family of synthetic speech

Acknowledgements: Study funded by NSF (IIS 1329659)