EMOTION RECOGNITION USING SYNTHETIC SPEECH AS NEUTRAL REFERENCE

Reza Lotfian and Carlos Busso
Multimodal Signal Processing (MSP) Laboratory
University of Texas at Dallas

Motivation
- 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
- Summary of 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)