A STUDY OF SPEAKER VERIFICATION PERFORMANCE WITH EXPRESSIVE SPEECH

Srinivas Parthasarathy, Chunlei Zhang, John H.L Hansen, Carlos Busso

Center for Robust Speech Systems (CRSS)– Multimodal Signal Processing Lab (MSP)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas - 75080, USA

Motivation

Background:

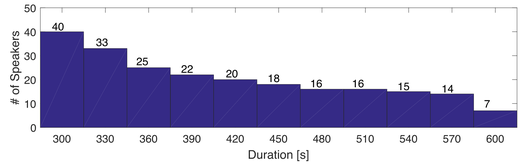
- Expressive speech introduces variations in acoustic features
  - Impacts performance of speaker verification systems
- Previous Work
  - Drop in performance when system trained with neutral speech and tested on expressive speech
- Limitations
  - Acted datasets
  - Limited number of speakers

Our Work:

- Analyze the effect of emotion on speaker verification performance
  - Naturalistic data from multiple speakers

MSP-PODCAST

- Emotional corpus being collected at UT-Dallas
  - We use subset – 40 speakers with >300s speech
- Multiple sentences from speakers appearing in various podcasts (2.75s – 11s)
- Annotated on Amazon Mechanical Turk
  - Primary Emotions : Angry, Neutral, Happy etc.
  - Arousal (1 –very calm vs 7 – very active)
  - Valence (1 – very negative vs 7 – very positive)
  - Dominance (1 – very weak vs 7 – very strong)



Number of speakers with at least x seconds of data

Methodology

Training Criterion:

- We use 150s of neutral speech, per speaker, for training the model

Criterion 1	Neu + 3 < x < 5	636
Criterion 2	Any + 3 < x < 5	188
Criterion 3	Neu + 2 < x < 6	194
Criterion 4	Any + 2 < x < 6	43

x ϵ {Arousal, Valence, Dominance}

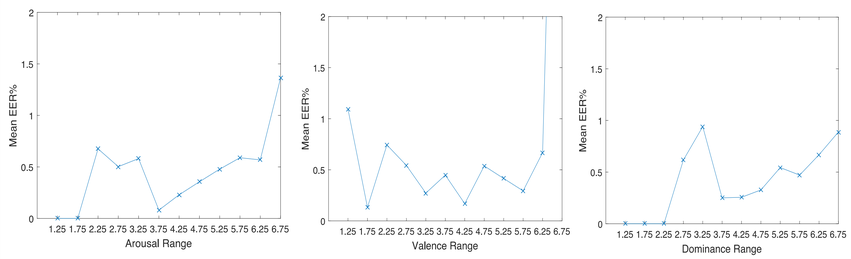
- i-Vector framework with probabilistic linear discriminant analysis (PLDA) back-end
- We extract a 13-dimensional MFCC with $\Delta + \Delta\,\Delta$ (39-D feature vector)
- We train a 256-mixture UBM using training data
- Dimension of i-Vector empirically set to 200

$$M = m + Tx$$

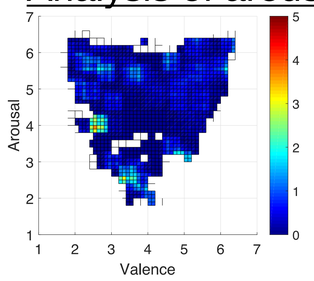
  - M – GMM super Vector
  - m – Mean vector constructed from UBM
  - T – Low- rank projection matrix
  - x – i-vector
- Dimension reduction with LDA
  - 200 → 39

Results

Individual emotions

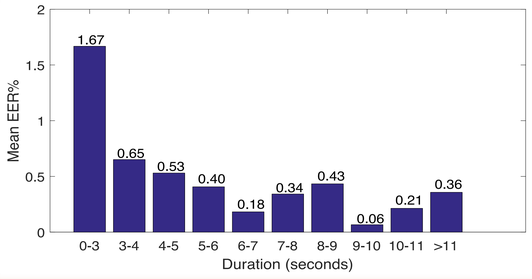


Analysis of arousal-valence space



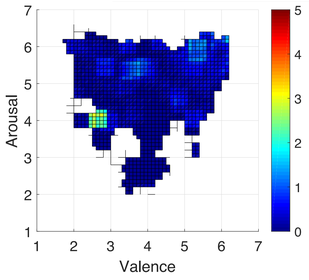
- Average of 0.5 x 0.5 window, shifted by 0.1
- EER low for neutral values of arousal, valence
- Higher EER as values deviate from neutral speech

Effect of duration of sentence



- Performance drop for short segments

After normalizing duration



- Create 5s training
  - Splitting long sentences
  - Concatenating turns with similar emotions
- Similar observations

Conclusions

- Speaker verification affected by expressive speech
- Higher errors on speaker verification when we deviate from neutral speech

Future Work

- We are annotating more data
- Study compensation techniques for emotional variability

80 speakers