Predicting Speaker Recognition Reliability by Considering Emotional Content

Srinivas Parthasarathy, Carlos Busso

Multimodal Signal Processing Lab (MSP)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas - 75080, USA

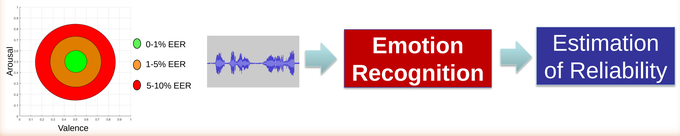
Motivation

Background:

- Emotional variability in speech impacts speaker recognition performance
  - Drop in performance when system trained with neutral speech and tested on expressive speech
- Can we define regions in the emotional space where speaker recognition is reliable?
- Can we automatically predict these regions?

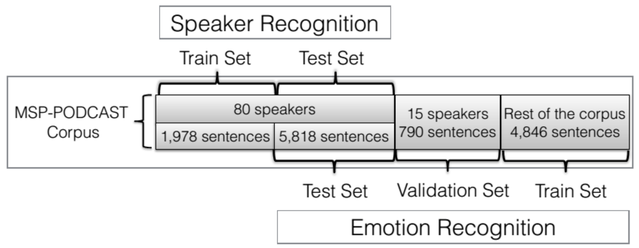
Our Work:

- Automatically predict reliable/unreliable regions for speaker recognition
- Build emotion-based classifiers to recognize reliability regions



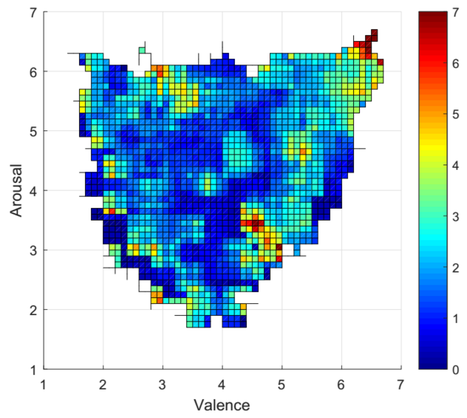
MSP-PODCAST

- Emotional corpus being collected at UT-Dallas
  - Multiple sentences from speakers appearing in various podcasts (2.75s – 11s)
  - We use subset – 95 speakers with >300s speech
  - Annotated on Amazon Mechanical Turk for emotional dimensions
- Data partition for speaker recognition and emotion recognition tasks
  - Unified test set, speaker independent partitions



Speaker Recognition

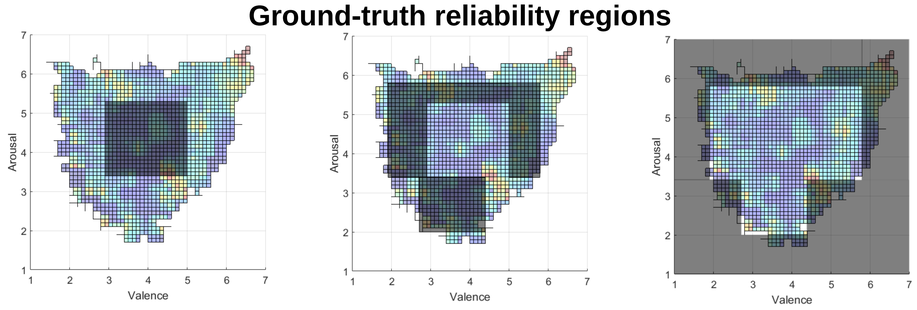
- i-Vector framework with probabilistic linear discriminant analysis (PLDA) back-end
- We extract a 13-dimensional MFCC with Δ + Δ Δ (39-D feature vector)
- We train a 256-mixture UBM using training data
- Dimension of i-Vector empirically set to 200
- Trained with neutral speech
- Tested with emotional,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



- Each bin shows EER of sentences within 0.4 x 0.4 window of the bin
- Consider bins with at least 10 speakers with 1 test utterance

Reliability for Speaker Recognition

Ground-truth reliability regions



Reliable (x < 2% EER) | Uncertain (2% < x < 3% EER) | Unreliable (x > 3% EER)

Acoustic Features

- Interspeech 2013 Computational Paralinguistic Challenge feature set
- 6,373 features

Classification Framework

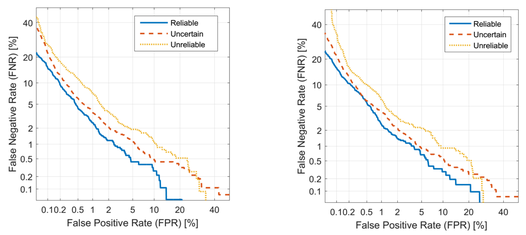
- Trained, validated, tested on speaker independent sets
- 1,024 nodes, 2 hidden layers, ReLU activation
- Dropout 0.5, early stopping on validation set

Classification results

		Predicted Class		
		Reliable	Uncertain	Unreliable
Correct Class	Reliable	905	822	136
	Uncertain	973	1398	426
	Unreliable	201	540	417

F1-score 0.45

DET curves



Ground truth | Predicted

EER

Type	Reliable	Uncertain	Unreliable
Original	1.47	2.03	2.85
Predicted	1.64	2.04	2.76

Conclusions

- Emotion classification system has average F1-score of 0.45
- Ground truth EER of unreliable regions 2 times the EER of reliable regions
- Predicted EER of different regions very similar to ground truth EER

Future Work

- We are annotating more data
- Analyze reliability across multiple speaker recognition systems
- Improve models for emotion recognition task
- Study compensation techniques for unreliable sentences