A Personalized Emotion Recognition System Using an Unsupervised Feature Adaptation Scheme

ICASSP 2012
IEEE International Conference on Acoustics, Speech, and Signal Processing
ICASSP 2012: March 25-30, 2012
Kyoto, Japan

Tauhidur Rahman and Carlos Busso
Multimodal Signal Processing (MSP) Laboratory
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083, U.S.A.

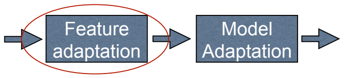
UT Dallas MSP Multimodal Signal Processing Laboratory

Motivation

- Emotional models and classifiers do not generalize with mismatched conditions (training and testing)
  - Speaker dependent models give better performance than Speaker independent models [Austermann et al. 2005]
- The challenge is to build a robust classifier that can recognize the expressive speech of unseen speakers
- The goal of this study is to adapt an emotion recognition system to a target user

Personalized emotion recognition system

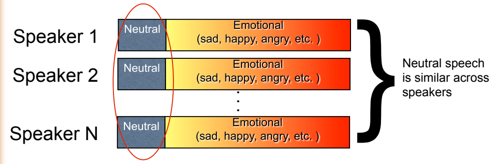
- An intriguing approach is the use of feature and/or model adaptation for emotion recognition [Kim et al., 2011]



Iterative Feature Normalization (IFN)

Optimal normalization:

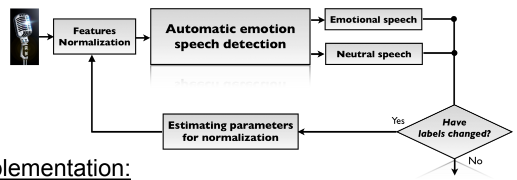
- Normalization parameters are estimated from neutral subset
- Parameters are applied to the entire emotional corpus



- Variability between emotional classes is preserved
  - Parameters are estimated only from the normal set
- Assumptions:
  - A portion of neutral speech from each speaker is available
  - Speaker Identity in the corpus is known

IFN Approach:

- Classify speech as emotional or neutral
- Use neutral samples to estimate normalization parameters
- Repeat “n” times (or until the labels do not change)

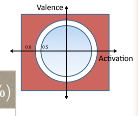


Implementation:

- Z normalization $\hat{f}_i = \dfrac{f_i - \mu_i^{neu}}{\sigma_i^{neu}}$
- 384 sentence level features (Interspeech 2009 emotion challenge)
- Linear kernel SVM with sequential minimal optimization (SMO)

Controlled Conditions

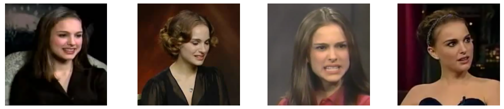
- IEMOCAP database [Busso et al. 2008]
  - ~12 hours of data, Read, scripted and spontaneous
  - Happiness, sadness, anger, neutral, etc.
  - Activation, valence and dominance
- Neutral versus emotional speech
- Classes are balanced during training testing



Normalization	Weighted Accuracy (%)
Without Normalization	69.81
IFN	71.81
Perfect Normalization	72.75

Uncontrolled Conditions

- Realistic recordings from a popular video sharing website
  - Data from a speaker during various uncontrolled conditions
    - Unbalanced data, environmental conditions, different ages
  - 90 minutes of speech from one speaker (837 5 sec files)
  - 3 subjects annotated the data [0 neutral - 1 emotional]
- The emotion detection system trained with IEMOCAP data



Normalization Type	WA(%)	UA(%)
Without Normalization	36.32	50.76
Unsupervised Feature Adaptation	80.28	70.02

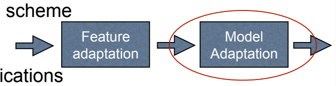
WA: Weighted Accuracy UA: Unweighted Accuracy

Conclusions

- The proposed front-end framework is able to reduce the mismatches in the training and testing conditions
- The approach is demonstrated in controlled and in uncontrolled recording conditions
  - 2% improvement (UA) with IEMOCAP database
  - 20% improvement (UA) with realistic recordings

Future Directions:

- Model adaptation for emotion recognition
  - Coupled with the proposed front-end unsupervised feature normalization scheme



- Explore different applications
  - Automatic call center, emotional profile of individuals

References:

C. Busso, A. Metallinou, and S. Narayanan, "Iterative feature normalization for emotional speech detection," in ICASSP 2011. Prague, Czech Republic, May 2011, pp. 5692-5695.