ICASSP 2011: May 22-27, 2011
Prague, Czech Republic

Iterative Feature Normalization for Emotional Speech Detection

Carlos Busso[1]
[1]Multimodal Signal Processing (MSP) Laboratory
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083, U.S.A.

Angeliki Metallinou[2] and Shrikanth S. Narayanan[2]
[2]Speech Analysis and Interpretation Laboratory (SAIL)
Viterbi School of Engineering,
University of Southern California,
Los Angeles, California 90089, U.S.A.

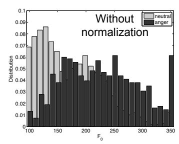
Introduction

- Recognition of emotion is an important problem
  - Development of new human machine interfaces
- A main challenges is to compensate the inter-speaker variability observed in expressive speech
  - Properties of speech are intrinsically speaker dependent
  - Expression of emotions presents idiosyncratic difference
- Goals:
  - Reduce speaker variability
  - Preserve the discrimination between emotions
- Concept:
  - Normalize emotional corpus such that neutral speech from each speaker presents similar trends

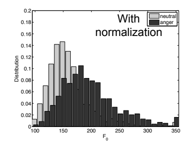
Motivation

- Optimal normalization:
  - Normalization parameters are estimated from neutral subset
  - Parameters are applied to the entire emotional corpus
- Variability between emotional classes is preserved

Case study: F0 mean



$$S_{F0}^{s} = \frac{F0_{ref}}{F0_{neu}^{s}}$$

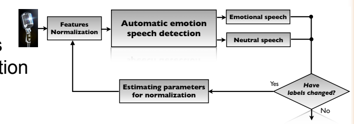


Assumptions:

1. A portion of neutral speech from each speaker is available
2. Speaker Identity in the corpus is known

IFN Approach

- Classify speech as emotional or neutral (speaker dependent)
- Use neutral samples to estimate normalization parameters



- Repeat *n* times (or until the labels do not change)

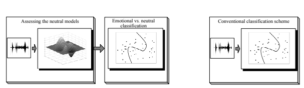
Databases & classifiers

- USC-EMA, EMO-DB, EPSAT
- Binary classifiers: neutral versus emotional speech
  - Samples for emotional classes are re-labeled as emotional
  - Average values over 400 realizations (chances 50%)
- Classifiers: Neutral models [Busso et al., 2009], Conventional classifiers
- Features: SQ75, SQ25, Smedian, Sdmedian, SVmeanRange, SDiqr, SVmaxCurv

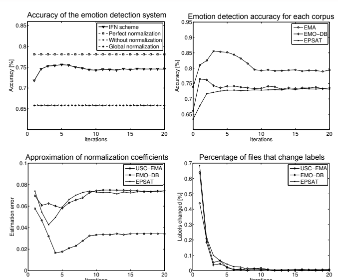
Experimental Results

Performance of the neutral model and conventional approach with different normalization schemes

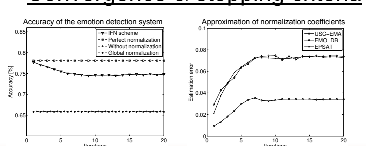
	Neutral Model				Conventional scheme			
Optimal Normalization [%]								
	Acc	Rec	Pre	F	Acc	Rec	Pre	F
All	**78.1**	**80.2**	**74.6**	**77.3**	**74.6**	**89.5**	**55.7**	**68.7**
EMA	86.6	92.1	71.9	80.8	81.7	95.2	56.0	70.5
EMO-DB	80.5	85.9	77.4	81.4	77.6	94.6	63.0	75.6
EPSAT	74.9	76.7	74.6	75.7	71.7	87.1	54.0	66.7
Without normalization [%]								
	Acc	Rec	Pre	F	Acc	Rec	Pre	F
All	**65.9**	**65.1**	**69.0**	**67.0**	**67.7**	**72.8**	**56.6**	**63.7**
EMA	73.9	72.7	54.2	62.1	68.7	64.5	45.8	53.6
EMO-DB	66.1	69.7	68.3	69.0	72.4	84.8	61.0	71.0
EPSAT	63.4	63.1	72.8	67.6	66.4	72.2	58.4	64.6
Global normalization (speaker dependent) [%]								
	Acc	Rec	Pre	F	Acc	Rec	Pre	F
All	**65.8**	**66.2**	**64.4**	**65.3**	**72.1**	**82.2**	**56.4**	**66.9**
EMA	73.6	69.1	57.5	62.8	77.2	72.4	66.7	69.4
EMO-DB	70.7	81.8	59.8	69.1	75.4	89.6	62.5	73.6
EPSAT	62.3	63.4	67.0	65.2	69.8	83.9	52.7	64.7
IFN approach (speaker dependent) [%]								
	Acc	Rec	Pre	F	Acc	Rec	Pre	F
All	**75.6**	**76.6**	**73.7**	**75.1**	**73.3**	**86.7**	**55.0**	**67.3**
EMA	85.2	88.5	71.1	78.8	80.8	97.8	51.9	67.8
EMO-DB	74.1	77.9	73.9	75.8	76.8	92.7	62.8	74.8
EPSAT	72.8	74.1	74.2	74.2	70.2	83.3	54.1	65.6



- Accuracy decreases without normalization
- Speaker dependent global normalization
  - Accuracy of the system is not improved
  - It affects emotional discrimination
- Iterative Feature Normalization Approach
  - 2.5% (1.3%) lower than optimal normalization
  - 9.8% (5.3%) higher than global normalization
  - Less than <5% of labels changed after the 5th ite.



Convergence & stopping criteria



- Normalization parameters are initialized with optimal values
- IFN approach converges to a suboptimal state due to misclassification
- Performance is still 8.7% higher than global normalization

Discussion

- The IFN scheme approximates optimal normalization
  - Minimize differences across speakers' neutral speech
  - Preserve emotional discrimination

Limitations & Future Directions

- It assumes that speakers' identities are known
  - Supervised or unsupervised speaker identification
- Other directions:
  - Study performance in multi-class emotion classification
  - Study performance in non-acted databases
  - Normalization of other acoustic features

References:

C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," IEEE Transactions on Audio, Speech and Language Processing, vol. 17, no. 4, pp. 582-596, May 2009.