Building A Naturalistic Emotional Speech Corpus by Retrieving Expressive Behaviors From Existing Speech Corpora

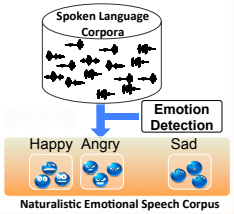
Soroosh Mariooryad, Reza Lotfian and Carlos Busso

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

interspeech 2014 singapore
15th Annual Conference of the International Speech Communication Association, 14 - 18 Sept 2014, MAX Atria @ Singapore EXPO

Motivation

- Limited naturalistic, spontaneous emotional corpora
  - Difficult to collect in controlled recordings
- Explore existing resources for speech processing
  - Natural interactions
  - Conversation elicits emotional responses
- Approach:
  - Emotion detection
  - Perceptual evaluation



Databases

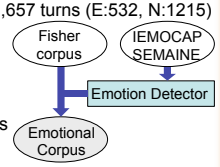
Emotional Databases

IEMOCAP
- 10 trained actors
- 5 dyadic sessions
- Spontaneous improvisations
- 5,496 turns (E:4375, N:1121)

SEMAINE
- User and operator interaction
- Emotion induction with SAL
- Sensitive artificial listener
- 1,657 turns (E:532, N:1215)

Non-emotional Databases

Fisher English Database
- 5,000 conversations, 2000 hours
- 800,000 speech segments
- Recorded for ASR systems for conversational speech



Retrieving Expressive Behaviors

Feature Extraction
- INTERSPEECH 2011 feature set
- OpenSMILE toolkit
- 4368 high level descriptors

Feature Selection
- Inter-class and intra-class distance measure (4368-->500)
- Maximizing the performance of the SVM (500-->100)
- Only on the SEMAINE database

Classifiers
- Linear kernel SVM
- Ranking based on confidence score

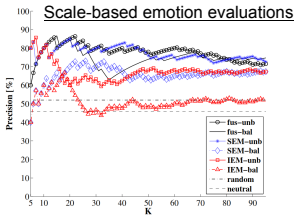
Set	# Turns	Corpus(Training)	Training Data
Emotional Sets	top 100 top 100 top 100 top 100 top 100 top 100	IEMOCAP IEMOCAP SEMAINE SEMAINE Fusion Fusion	Balanced Unbalanced Balanced Unbalanced Balanced Unbalanced
Neutral Sets	top 50 top 50	IEMOCAP SEMAINE	Unbalanced Unbalanced
Random Set	50	--	--

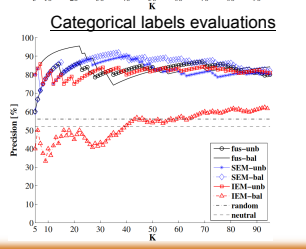
Crowdsourcing (Mechanical Turk)
- 5 evaluations per sentence, 536 utterances
- 30 utterances per HIT
  - Degree of emotion *(neutral vs. emotional)*
  - Categories (angry, happy, neutral, sad,…)
  - Dimensions (arousal, valence, dominance)

Analysis of Retrieved Emotional Content

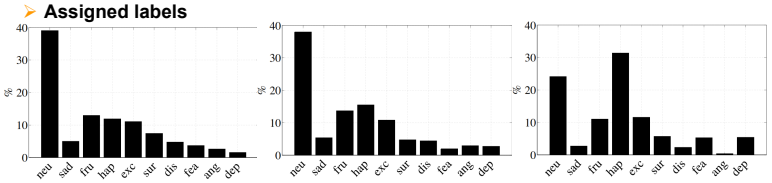
Precision Analysis
- Precision of K most emotional samples
- Neutral vs. Emotional (different evaluations)
- Higher precision than neutral and random
- Balanced training achieves lower precision

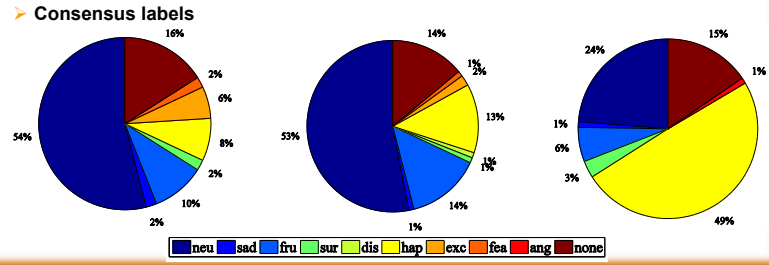
Assigned Categorical Emotional Labels
- Neutral and random sets exhibit similar distributions
- There are expressive behaviors in Fisher database (60% of assigned labels)



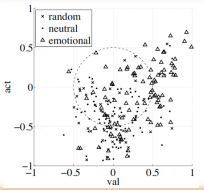


50 random sentence | Top 100 neutral | Top 100 emotional (fusion)





Distribution in valence-activation
- Neutral, random and emotional
- Mostly positive behaviors
  - Fisher collection protocol
- Emotional classifiers
  - more effective on high arousal-high valence



Discussion

Conclusions
- Building naturalistic emotional corpora by retrieving emotional behaviors from existing ASR or SID databases
- Emotion detectors trained with emotional corpora are effective to retrieve expressive sentences

Future Directions
- Create large natural emotional spontaneous databases
- Retrieve emotionally balanced behavior
- Building retrieval systems expert on specific emotions
- Unsupervised normalization techniques
  - Using *iterative feature normalization* (IFN) method

Acknowledgment

This work was funded by NSF (IIS-1217104, IIS-1329659) and Samsung Telecommunications America