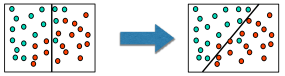
SUPERVISED DOMAIN ADAPTATION FOR EMOTION RECOGNITION FROM SPEECH

Mohammed Abdelwahab and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, University of Texas at Dallas

Motivation

- Performance of speech emotion recognition degrades with mismatched conditions
- Model adaptation can mitigate problems

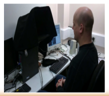


- We address the following questions:
  - How much labeled data is needed?
  - How important is speaker diversity?
  - Can acted data be used to train models?
  - What is best approach for supervised adaptation?

Databases

SEMAINE (training)

- 10 speakers, dyadic recordings
- Emotion induction with SAL
- 2315 turns, 6-8 evaluators
- Continuous time evaluations
  - Activation (calm vs. active)
  - Valence (negative vs. positive)
  - Average across time, raters



IEMOCAP (training)

- 10 trained actors in 5 dyadic sessions
- Spontaneous improvisations & scripted plays
- 6829 turns, 2 raters per turn
  - Activation and valence

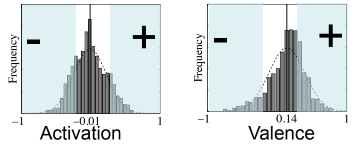
RECOLA (testing)

- 23 speakers in dyadic sessions
- Continuous time evaluations
  - Activation and valence
  - Average across time, raters
- We consider 899 turns, 6 raters per turn

Emotion Recognition Evaluation

Classification Problem

- Low vs high levels of arousal and valence
- We separately normalized the values, per corpora, using z-normalization.

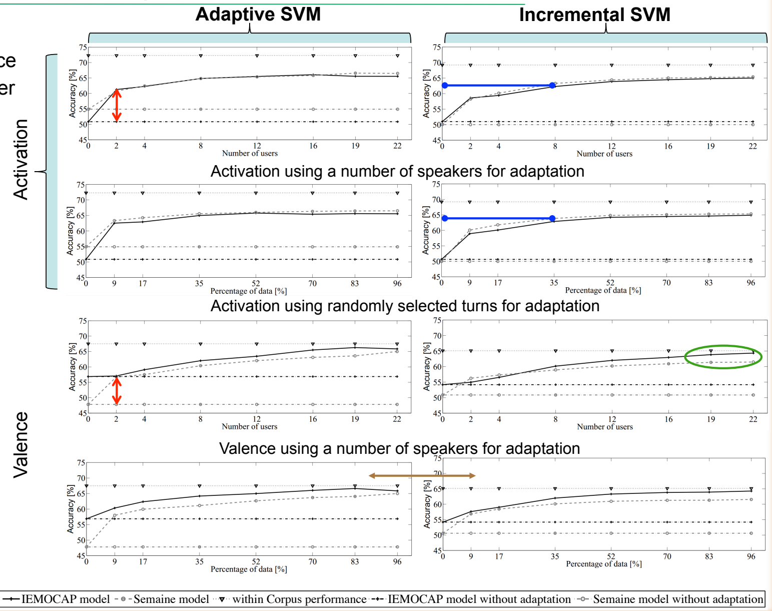


Feature Extraction of other acoustic features

- INTERSPEECH 2011 feature set
- OpenSMILE toolkit, 4368 high level descriptors

Feature Selection

- Correlation Attribute Evaluation
  - Ranked search method (4368 → 500)
- Correlation Feature Selection
  - Greedy stepwise method (500→50)



Activation using a number of speakers for adaptation

Activation using randomly selected turns for adaptation

Valence using a number of speakers for adaptation

Adaptation Schemes

SVM training

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \qquad \text{s.t.} \quad \xi_i \geq 0,\ y_i\mathbf{w}^T\phi(\mathbf{x}_i) \geq 1-\xi_i, \forall(\mathbf{x}_i, y_i)$$

Adaptive SVM [Yang et al., 2007]

- Minimizes:
  - Classification error over the training examples
  - Discrepancy between originals and adapted classifiers
- Decision boundary does not deviates much from original one
- It manages to separate new labeled data from target domain

$$f(x) = f^{old}(x) + \Delta f(x) = f^{old}(x) + \mathbf{w}^T\phi(\mathbf{x}_i)$$

Incremental SVM [Shalev et al., 2011]

- It allows to incrementally add more training data
- Only a subset of the data is considered at each step
- It discards old data while maintaining the support vectors
  - We use an effective stochastic sub-gradient descent algorithm for solving the optimization problem
  - Training examples are selected at random

$$\mathbf{w}_{t+1} \leftarrow (1-\frac{1}{t})\mathbf{w}_t + \eta_t \mathbb{1}[y_{i_t}\langle\mathbf{w}_t, \mathbf{x}_{i_t}\rangle < 1]y_{i_t}\mathbf{x}_{i_t}$$

Discussion

Conclusions

- We notice significant improvements even when we only use data from two subjects for adaptation (~9% of the data)
- Speaker variety is not a dominant factor in selecting the adaptation set
- A classifier built with acted data can perform as well as a classifier built with natural emotional databases
- Both SVM adaptation methods provide similar performance

Future Directions

- Unsupervised domain adaptation
- Feature Normalization

Acknowledgment

This work was funded by NSF (IIS-1217104 and IIS-1329659)