Speech Emotion Recognition with a Reject Option

Kusha Sridhar, Carlos Busso

UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

Erik Jonsson School of Engineering & Computer Science at the University of Texas at Dallas, Richardson, Texas 75080, USA

GRAZ – AUSTRIA
SEPTEMBER 15th – 19th 2019
INTERSPEECH 2019

Motivation

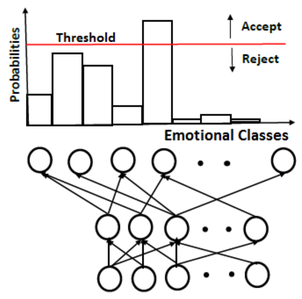
- Speech emotion recognition is a hard problem
  - Prediction are not always reliable
- Abstaining from prediction when in doubt can increase the reliability of a system
  - Selective classification on images have led to very low error rate (2%) for a test coverage of 60%
- The key challenge is to define mechanisms to quantify reliability to accept or reject an instance
  - e.g., Apply threshold on softmax output

First study on reject option for speech emotion recognition

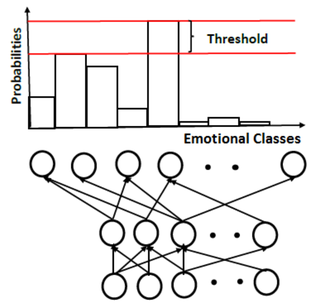
Reject Option for SER

Our Work

- SER system with a reject option
  - Accept or reject a sample based on the confidence of the classifier
  - Defined thresholds to interpret the confidence



Criterion 1



Criterion 2

- Goal is to improve the classifier performance while maintaining a high test coverage

Defining Thresholds

Criterion 1:

- Threshold on the neuronal activations
- Selective guaranteed risk (SGR) algorithm
  - Learn optimal risk bound on the classifier
  - Threshold on softmax outputs to achieve a desired error rate with high confidence

$$\hat{r}(f,g|S_m) = \frac{\frac{1}{m}\sum_{i=1}^{m} l(f(x_i),y_i)g(x_i)}{\hat{\phi}(f,g|S_m)}$$

$$Pr_{S_m}\{\hat{r}(f,g|S_m) < r^*\} > 99.99\%$$

$$\hat{\phi}(f,g|S_m) \triangleq \frac{1}{m}\sum_{i=1}^{m} g(x_i)$$

Criterion 2:

- Threshold on the difference between the two highest prediction values
- Large difference → clear prediction → accept

Optimization

- Empirical risk of classifier using SGR algorithm
- F1-Score

Architecture

- Two layers
- 1,024 nodes
- ReLU activation
- ADAM optimizer

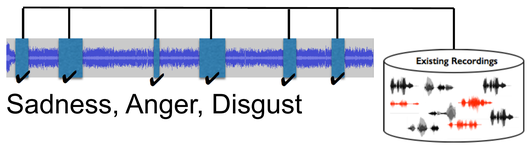
Task

- Categorical emotion recognition

Database and Features

The MSP-Podcast Corpus

- Emotionally rich speaking turns from speakers appearing in various podcasts (2.75s – 11s)
- Annotated for primary and secondary emotions (crowdsourcing)
- V1.4: 33,262 utterances with emotional labels (56h 29m)
  - Train set: 19,707 segments
  - Test set: 9,255 segments from 50 speakers
  - Validation set: 4,300 segments from 30 speakers
- Five-class problem
  - Happiness, Neutral, Sadness, Anger, Disgust
- Eight-class problem
  - Happiness, Neutral, Sadness, Anger, Disgust, **Surprised, Contempt, Fear**



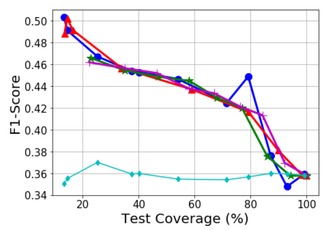
Acoustic Features

- Interspeech 2013 Computational Paralinguistic Challenge feature set (6,373 features extracted with OpenSmile)

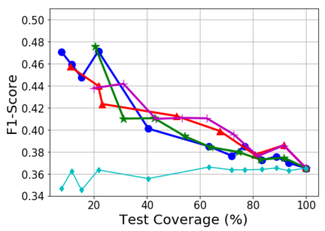
Results

5 classes

- **Hard labels**

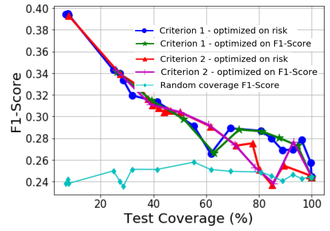


- **Soft labels**

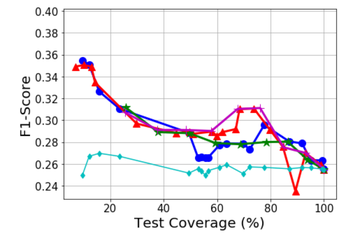


8 classes

- **Hard labels**



- **Soft labels**



- **Observations**
  - Confidence in predicting the accepted samples increases by rejecting ambiguous samples
  - With 75% coverage, we have relative gains up to 25.7% (5-class) and 20.6% (8-class)
  - Random selection of selected samples does not work

Analysis & Conclusion

Inter-Evaluator agreement of accepted/rejected samples

	Inter-evaluator agreement (Fleiss Kappa)		
	Coverage (%)	Accepted samples	Rejected samples
Hard labels (5-class)	100	0.2642	-
	75	0.2773	0.2590
	50	0.2897	0.2651
	25	0.3080	0.2633
Soft labels (8-class)	100	0.2680	-
	75	0.2723	0.2450
	50	0.2842	0.2496
	25	0.2983	0.2563

- **Observations**
  - Lower inter-evaluator agreement for rejected samples

Conclusions

- The reject option is a valuable feature, increasing the confidence in a SER system
  - Improvement in performance without compromising much on the coverage in the test set

References:
Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in neural information processing systems, 2017, pp.4878-4887

This work was supported by NSF under Grant CNS-1823166 and CAREER Grant IIS-1453781