Speech Emotion Recognition with a Reject Option

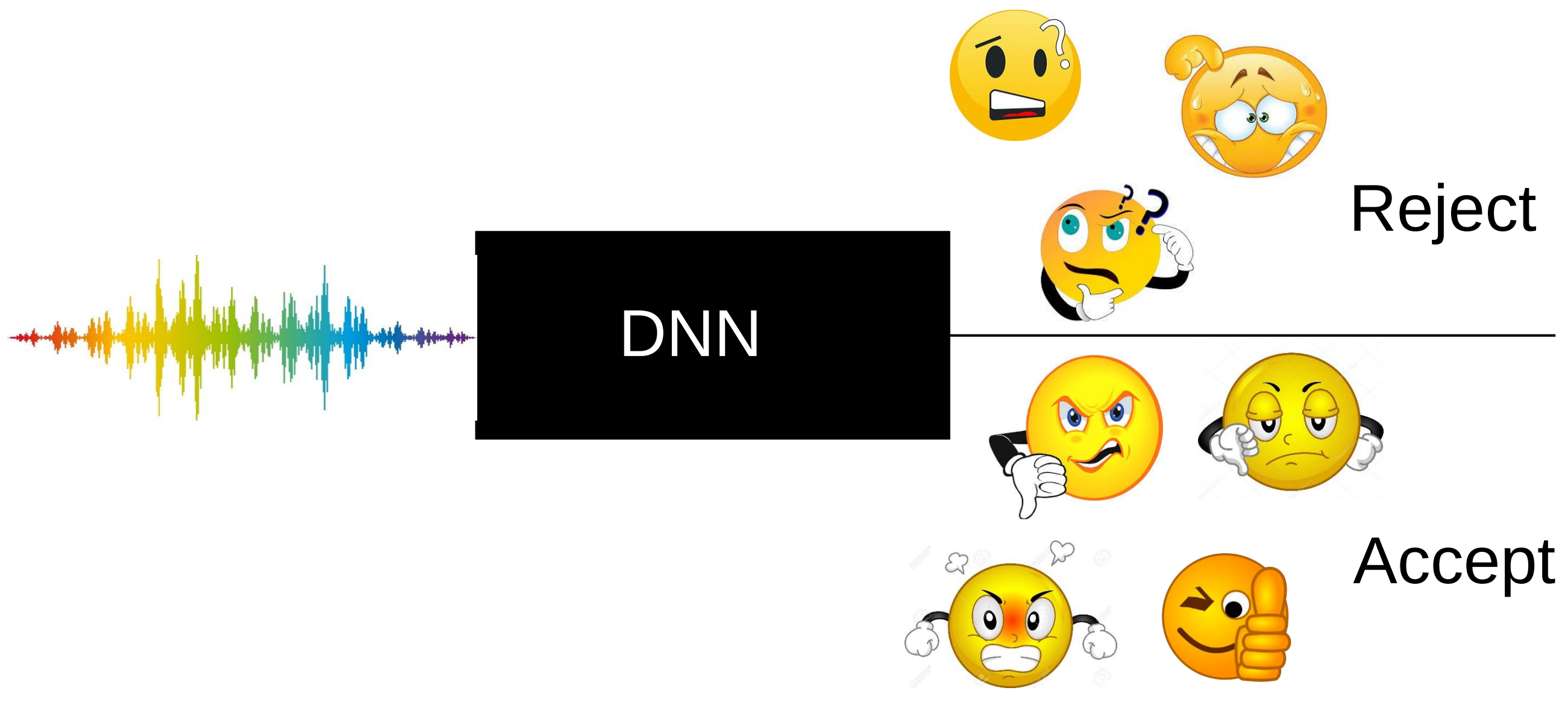
THE UNIVERSITY OF TEXAS AT DALLAS

Kusha Sridhar, Carlos Busso

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

Motivation

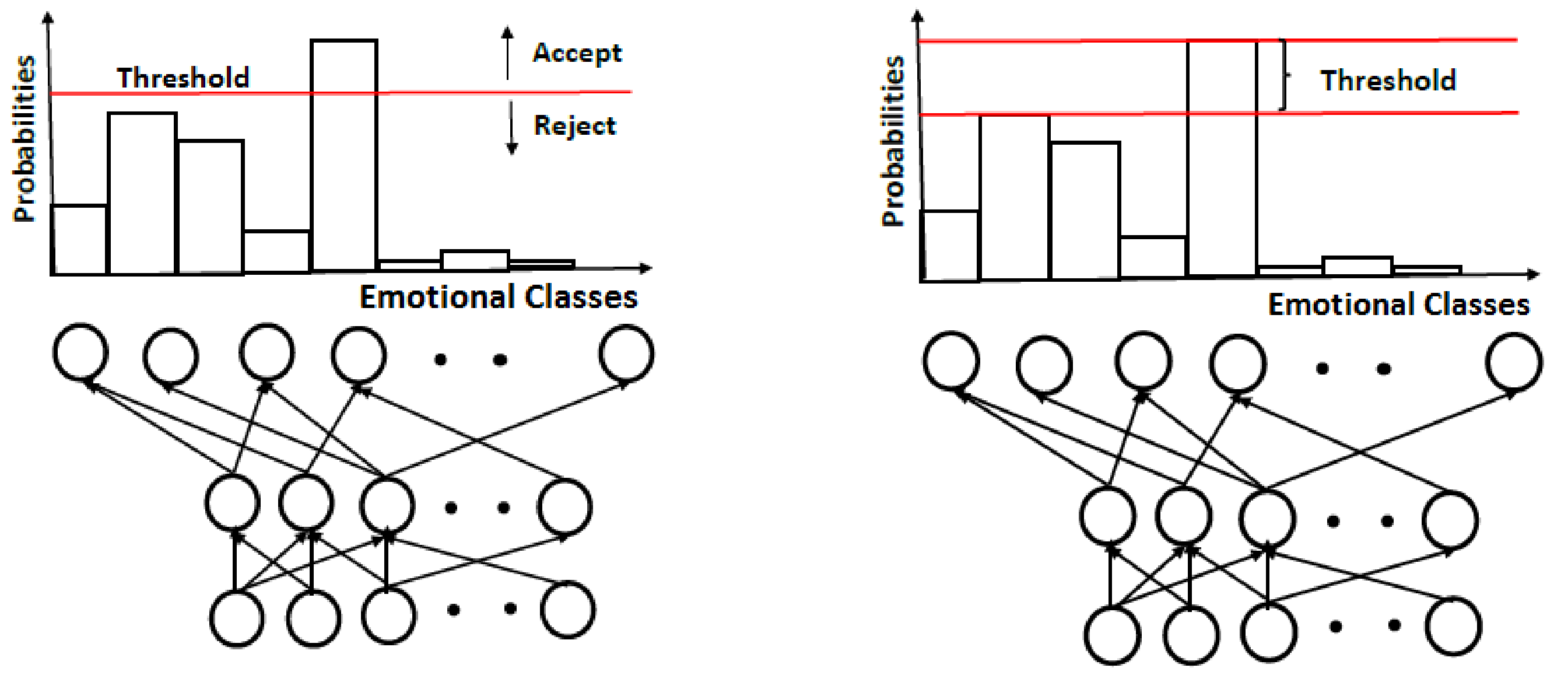
- Abstaining from prediction when in doubt helps application specific tasks
- Selective classification on images have led to very low error rate (2%) for a test coverage of 60%
- To accept or reject an instance – Apply threshold on softmax output / model the output uncertainty of the network



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



  - Classifier performance improved while maintaining a high test coverage

Defining Thresholds

Criterion 1:

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

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

$$Pr_{S_m}\{\hat{r}(f,g|S_m) < r^*\} > 99.99\% \ ; \ \hat{\phi}(f,g|S_m) \triangleq \frac{1}{m}\sum_{i=1}^{m} g(x_i)$$

Optimization

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

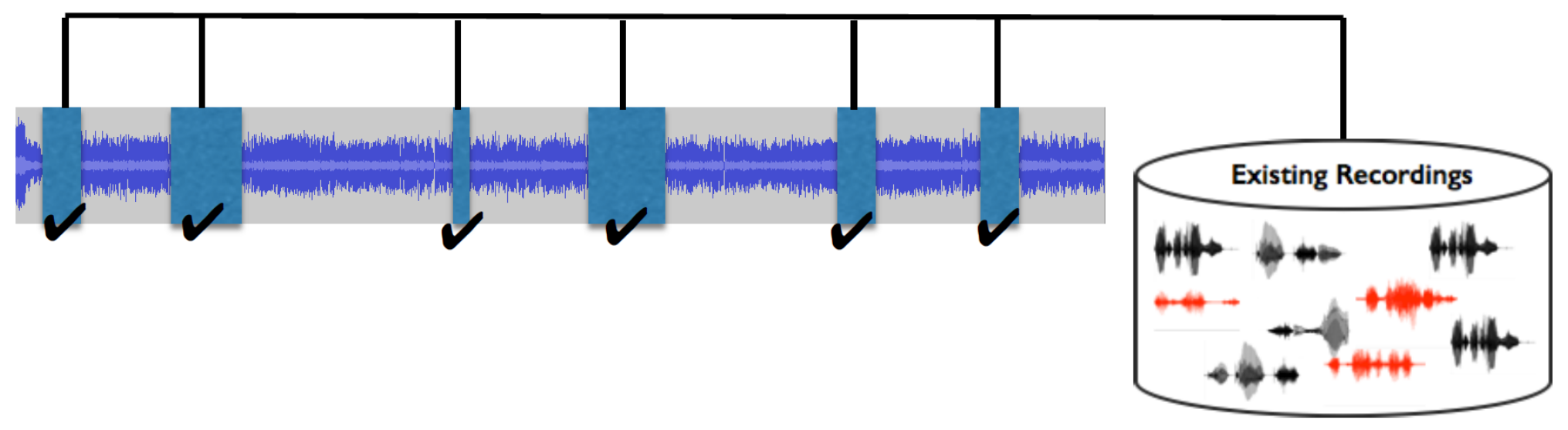
Criterion 2:

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

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 on Amazon mechanical Turk.
- 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



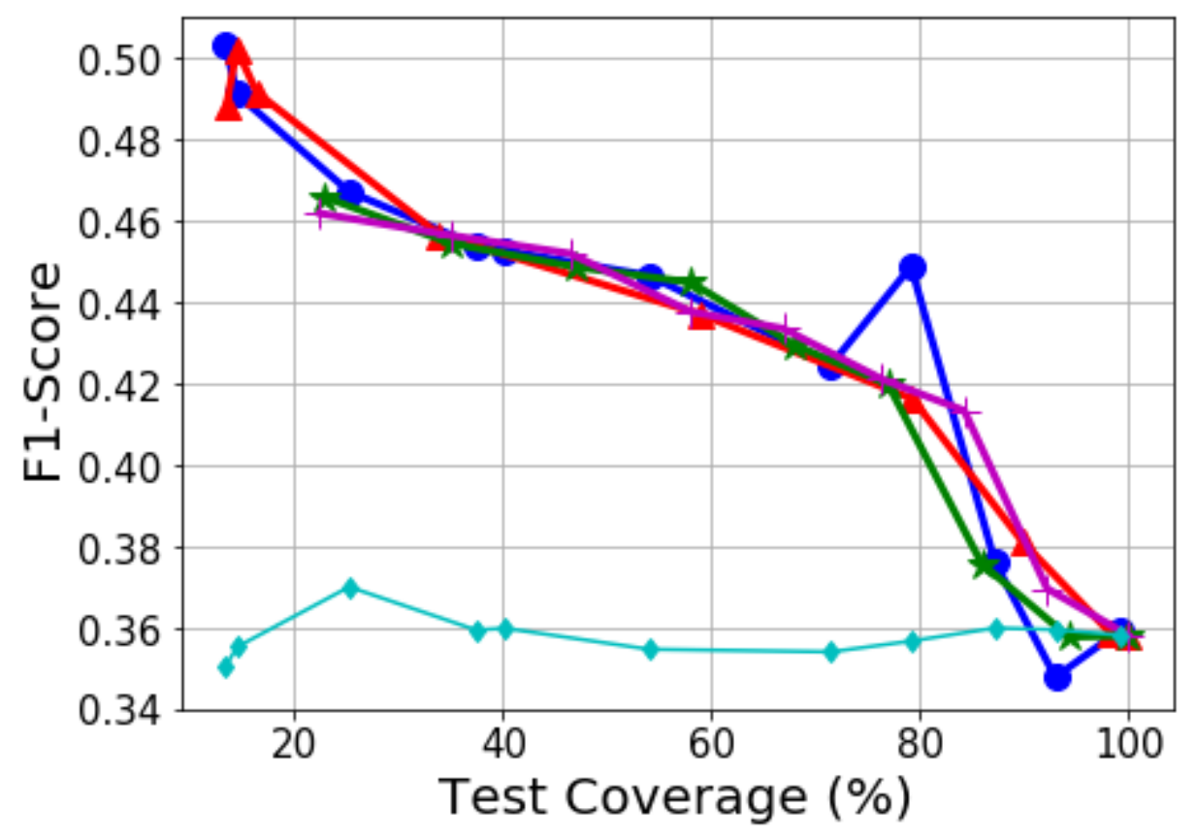
Acoustic Features

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

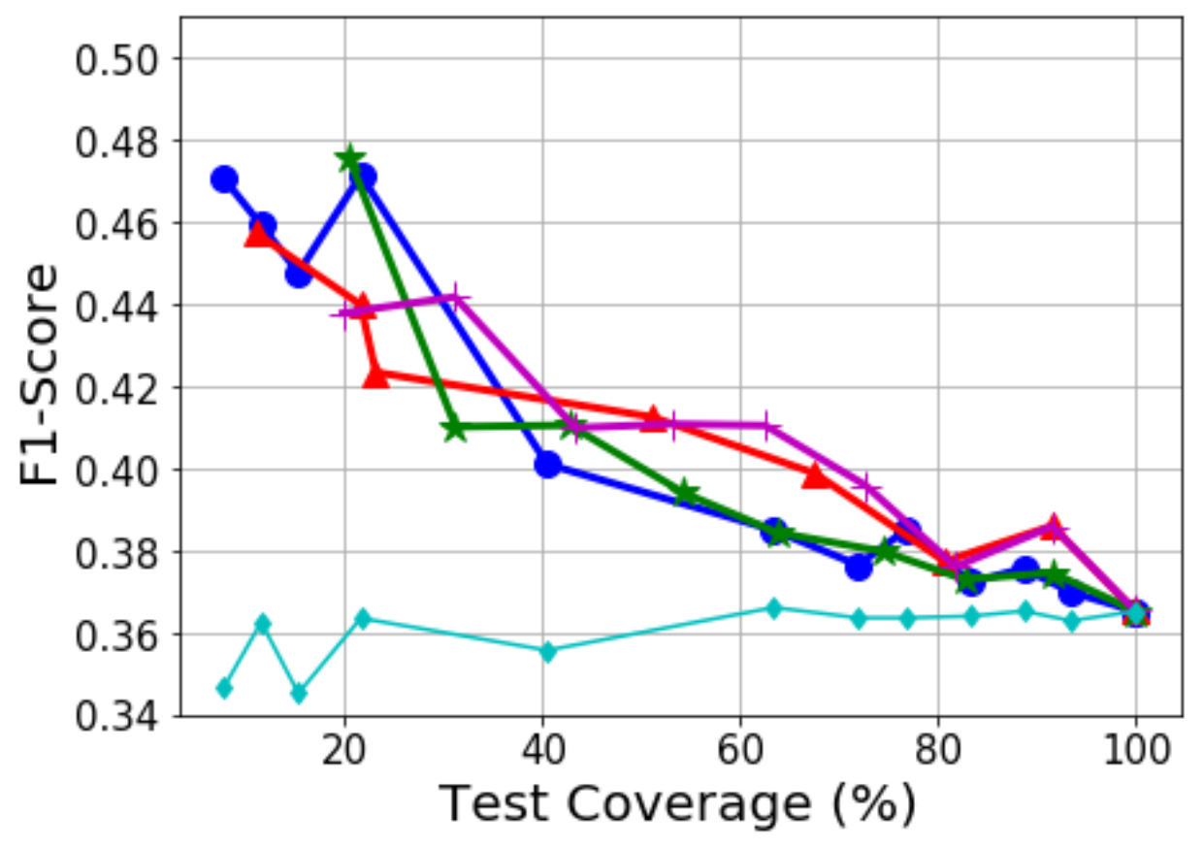
Results

5 classes
(Happy, Neutral, Sad, Angry, Disgust)

- **Hard labels**

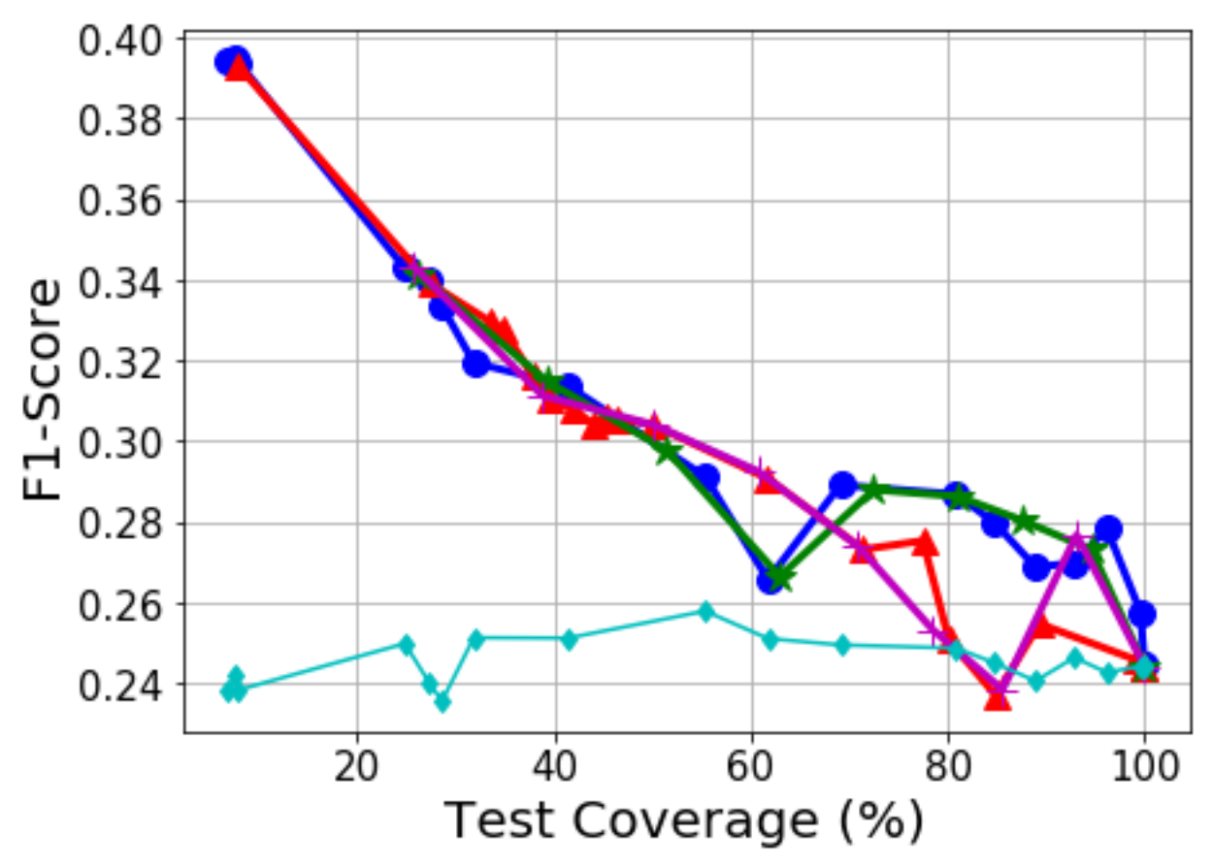


- **Soft labels**

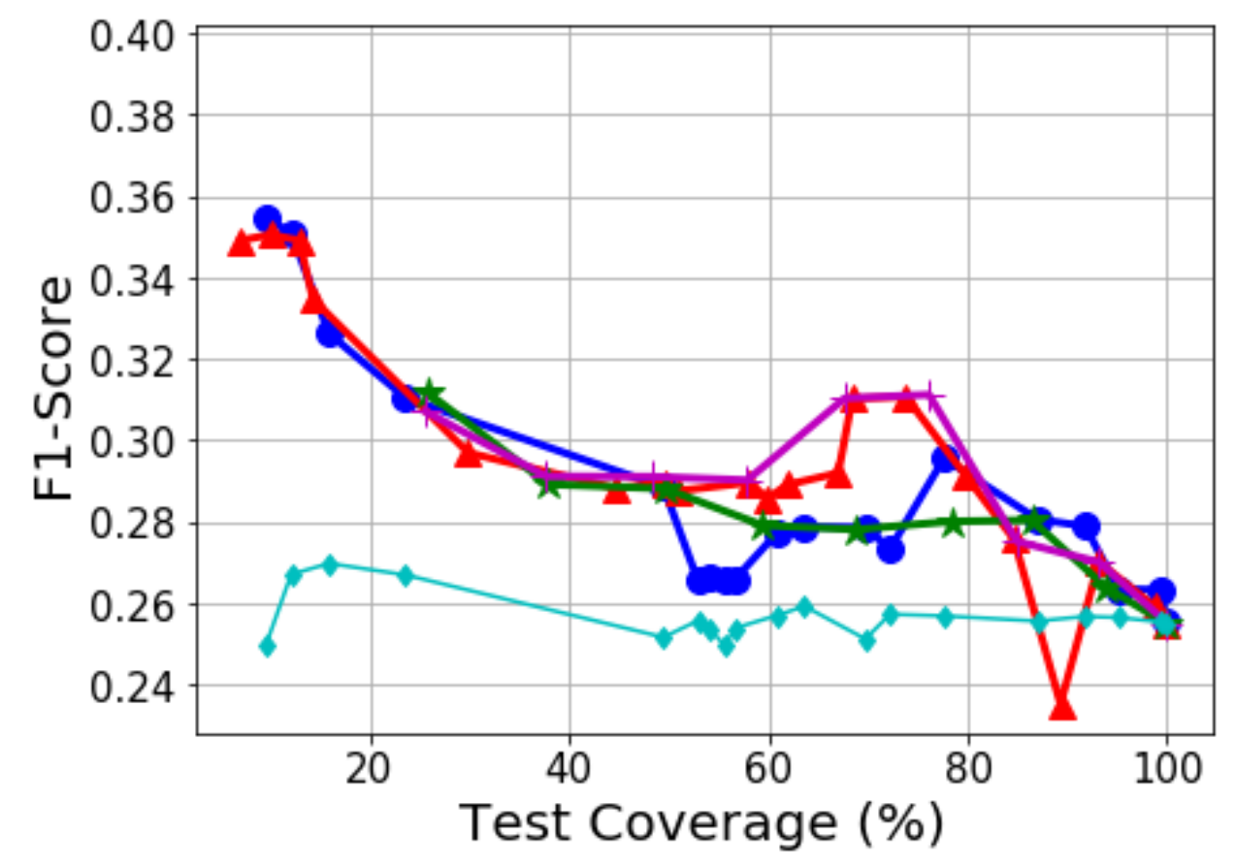


8 classes
(Happy, Neutral, Sad, Angry, Disgust, Surprised, Contempt, Fear)

- **Hard labels**



- **Soft labels**



Analysis & Conclusion

Inter-Evaluator agreement of accepted/rejected samples

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

- Relative gains in F1-Score at 75% test covarage
  - 25.71% with 5 classes (criterion1, risk opt)
  - 20.63% with 8 classes (criterion 2, F1-Score opt)
- Performance improvement:
  - 5 classes: Hard > Soft and 8 classes: Soft > Hard
- Lower inter-evaluator agreement for rejected samples

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 funded by NSF career award IIS-1453781