Background:
- It is not clear the best configuration for deep learning structures in speech emotion recognition
- Limited databases
- No well defined network structure that works well across conditions

Our Work:
- We study various factors affecting performance in DNN for speech emotion recognition
  - Amount of training data
  - Depth of the network
  - Use of residual networks
  - Activation
  - Batch normalization

The MSP-Podcast Corpus
- Emotional corpus collected at UT-Dallas
- Multiple sentences from speakers appearing in various podcasts (2.75s – 11s)
- Annotated on Amazon Mechanical Turk for emotional dimensions
- V1.0: 20,045 labeled utterances (34 hrs, 15 min)
  - Test set: 6,069 segments from 50 speakers
  - Dev set: 2,226 segments from 15 speakers
  - Train set: 11,750 segments

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

Motivation
- Background:
  - Limited databases
  - No well defined network structure that works well across conditions

Our Work:
- We study various factors affecting performance in DNN for speech emotion recognition
  - Amount of training data
  - Depth of the network
  - Use of residual networks
  - Activation
  - Batch normalization

Data Augmentation
- As the training set size increases, the performance increases
- We expect to see further improvements with more data (ongoing effort)

Data Augmentation
- Speech rate data augmentation
- Data augmentation provides a small benefit for very deep layers when the training set size is small
- 20 layers trained with 1,000 turns
  - ccc=0.46 w/o data augmentation
  - ccc=0.48 w data augmentation

Future Work
- We are annotating more data
- Explore using GANs for data augmentation
- Study end-to-end networks