Understanding speech emotion features in LSTMs

Jonathan Chan

1. **Problem introduction**
2. **State of the art**
3. **Research roadmap & experiments**

Problem introduction

Speech emotion recognition

- Teaching computers to understand the emotion behind a spoken phrase/sentence

Domain Challenges

- Hard problem even for humans. Discerning between certain emotions in a short sentence, such as excited vs. surprised, is difficult.
- Emotion recognition accuracy relies heavily on the ability to generate representative features [1]

Deep Learning Challenges

- Audio data is sequential with high dimensionality
  - Deep learning models that take advantage of the structure of sequential data are just starting to see success (e.g., LSTMs)
- Dataset size is relatively small and can be either acted emotions or (under 50 hours of audio)
  - Deep learning models have a tendency to over-fit data
  - With thousands of training parameters, a model could easily memorize a small training set

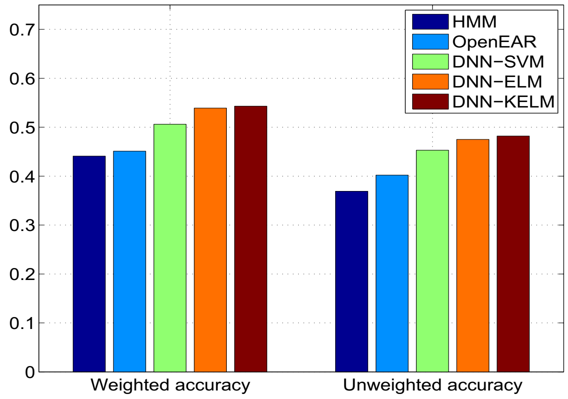
State of the art (IEMOCAP dataset)

"Traditional" machine learning approaches

- Eyben *et al*. developed an open source toolkit openEAR that utilizes an SVM with input features such as signal energy, MFCCs, LPCs, formants, voice quality, and chroma among others [2]

Deep learning approaches

- Han *et al.* explored using deep neural networks without recurrent networks in [3]
  - 3 fully connected hidden layers of 256 rectified linear units and an Extreme Learning Machine (ELM), which is a neural network with one hidden layer of 120 units
  - Input features are pitch, Mel-frequency cepstrum coefficients (MFCC), and deltas of MFCCs
- Lee et al. substituted a bidirectional LSTM for the fully connected neural network and continuing to employ the ELM in [4]
  - Accuracy increased from 52% to 64%
  - Added new input features: voice probability and zero crossing rate



Han et al. showing classification performance across models on IEMOCAP [3]

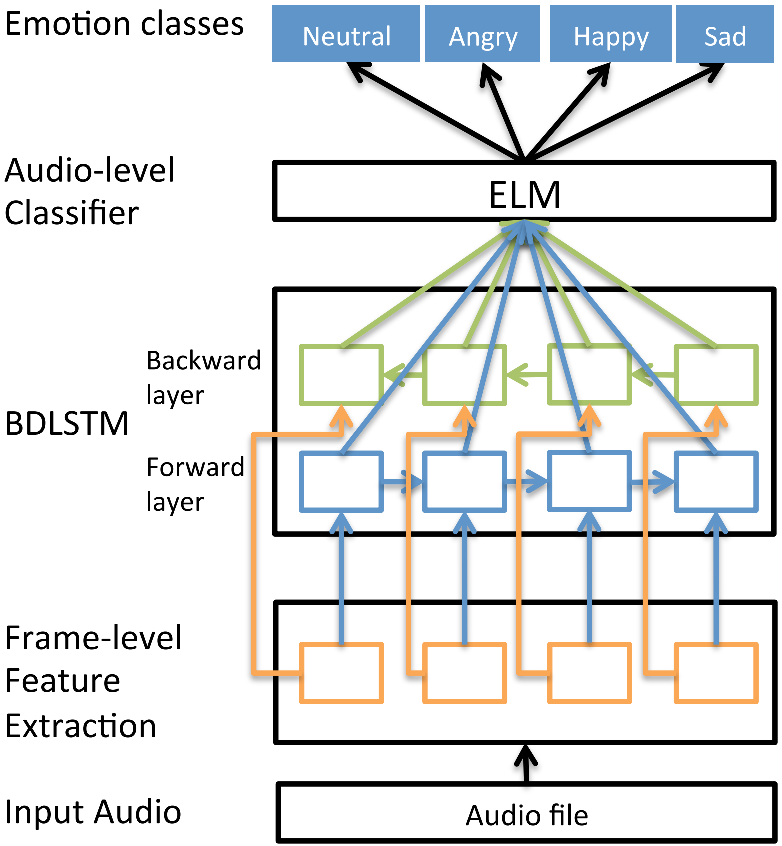
Research roadmap

Objective

- Understand the relative importance of the features used by *Lee et al.* and to explore how other commonly used features used in speech recognition affect the accuracy of the model

Roadmap

1. Merge IEMOCAP[5] and LDC [6] datasets
2. Divide dataset for speaker independent training/validation/test sets
3. Extract features: pitch, MFCCs, delta MFCCs, zero-crossing rate
4. Build voice probability feature with a simple model
5. Build bi-directional LSTM [7]
6. Build extreme learning machine
7. Train model, hyperparameter search, and evaluate accuracy on test
    - With spectrogram
    - With MFCCs
    - With MFCCs and delta MFCCs
    - With MFCCs, deltas MFCCs, and pitch
    - With MFCCs, deltas MFCCs, pitch, and zero-crossing rate
    - With MFCCs, deltas MFCCs, pitch, zero-crossing rate, and voice probability



Bibliography

[1] Yelin Kim, Honglak Lee, and Emily Mower Provost. "Deep learning for robust feature generation in audiovisual emotion recognition." In 2013 IEEE Int. Conference on Acoustics, Speech and Signal Processing, pp. 3687-3691.
[2] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR - introducing the Munich open-source emotion and affect recognition toolkit," in *Proceedings of ACII 2009*. IEEE, 2009, pp. 1–6.
[3] Kun Han, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." In *Interspeech*, pp. 223-227. 2014.
[4] Jinkyu Lee and Ivan Tashev. "High-level feature representation using recurrent neural network for speech emotion recognition." In *Sixteenth Annu. Conference of the Int. Speech Commun. Association*. 2015.
[5] Busso Carlos *et al.* "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42, no. 4 (2008): 335-359.
[6] Liberman, Mark, et al. "Emotional Prosody Speech and Transcripts" LDC2002S28. DVD. Philadelphia: Linguistic Data Consortium, 2002.
[7] Alex Graves, Navdeep Jaitly and Abdelrahman Mohamed, "Hybrid speech recognition with bidirectional LSTM," In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pages 273–278.