Speech Emotion Classification

•••

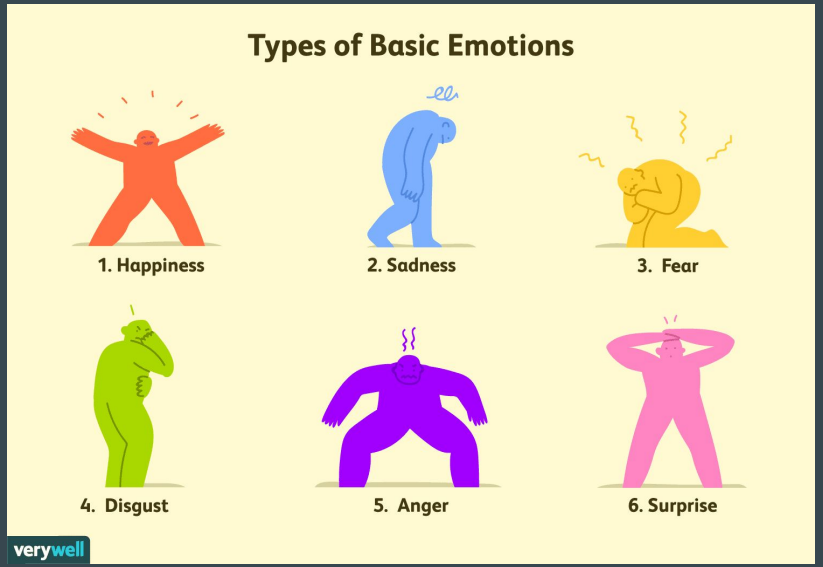
Support Vector Machine vs. Deep Learning

Qiyang Ma, Nathaniel Lounsbury, Christopher Woloshyn

Introduction

- Classify emotions from an audio file input.
- Train an SVM Classifier.
- Train Deep Learning architectures.
- Assess robustness by introducing noise to the data.
- Compare results.

https://www.verywellmind.com/an-overview-of-the-types-of-emotions-4163976

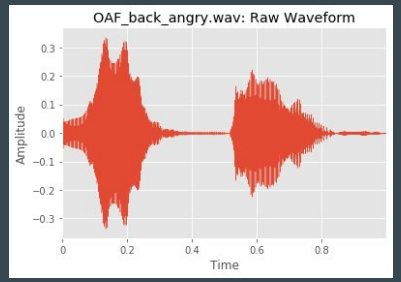


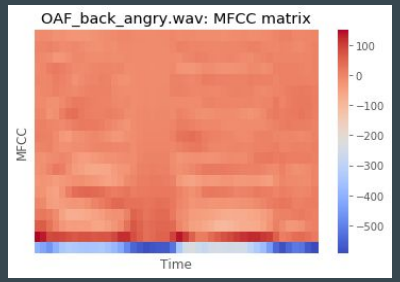
*Also Neutral

Preprocessing

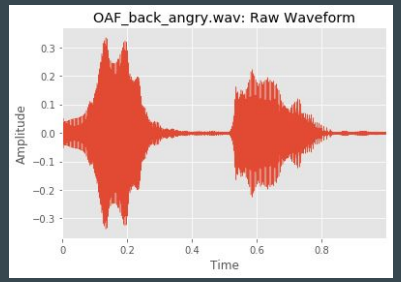
- Toronto Emotion Speech Set (TESS).
  - 2800 files.
  - 2 actors.
  - 7 emotions.
- Mel Frequency Cepstral Coefficients (MFCC).
  - Based on human audio perception (phonemes).
  - Compared with wavelet function.
  - Consistently produces the best results.*
- Multiple Preprocessed datasets.
  - No noise.
  - Light random noise.
  - Heavy random noise.

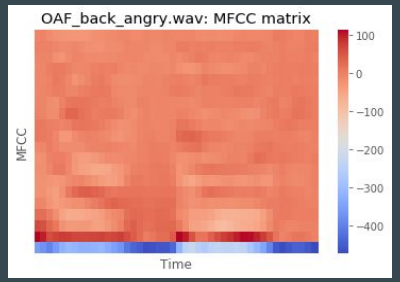
No noise



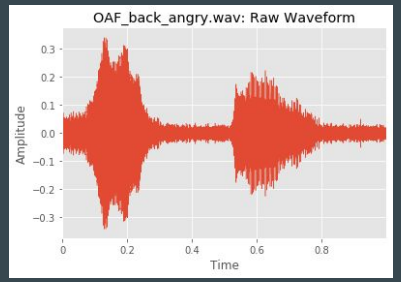


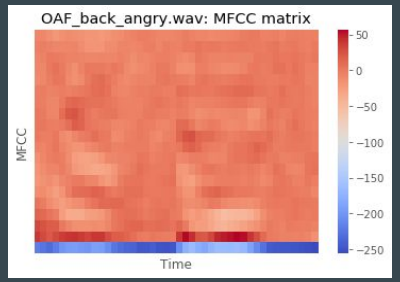
Light noise



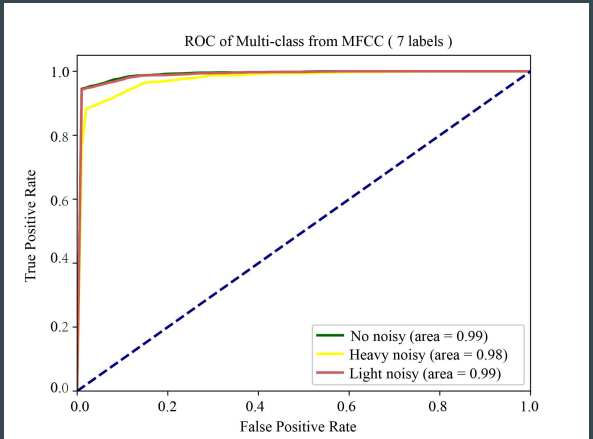


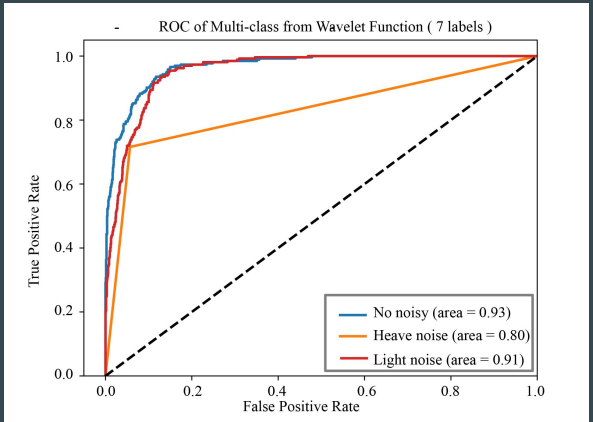
Heavy noise





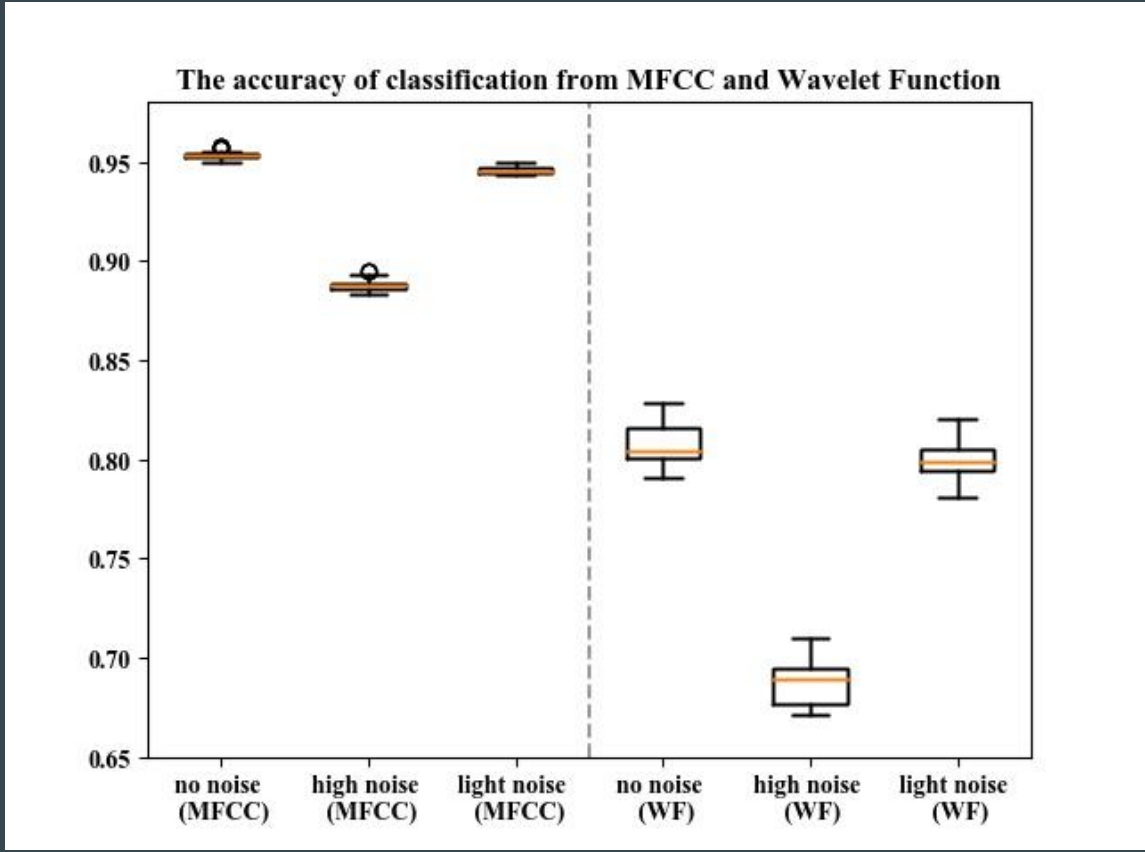
* https://research.ijcaonline.org/volume101/number12/pxc3898271.pdf





Support Vector Machine

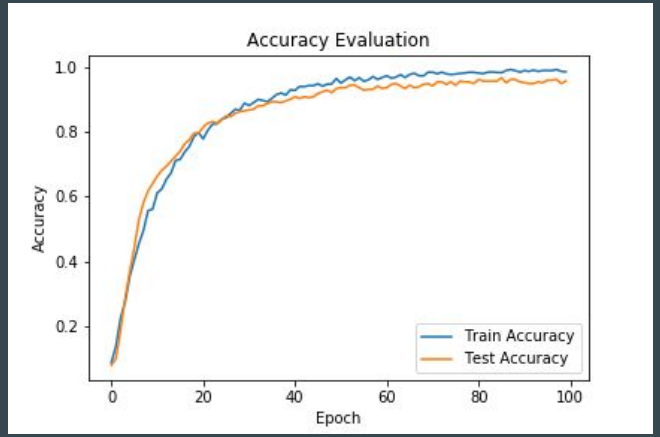
- Parameters
  - C = 50
  - Kernel = Radial Basis Function (‘rbf’)
  - Gamma = ‘scale’
- Results (MFCC Pre; 7 Labels, 5-fold)
  - .99 accuracy with no noise
  - .99 accuracy with light noise
  - .98 accuracy with heavy noise
- Results (Wavelet Pre; 7 Labels, 5-fold)
  - .93 accuracy with no noise
  - .91 accuracy with light noise
  - .80 accuracy with heavy noise

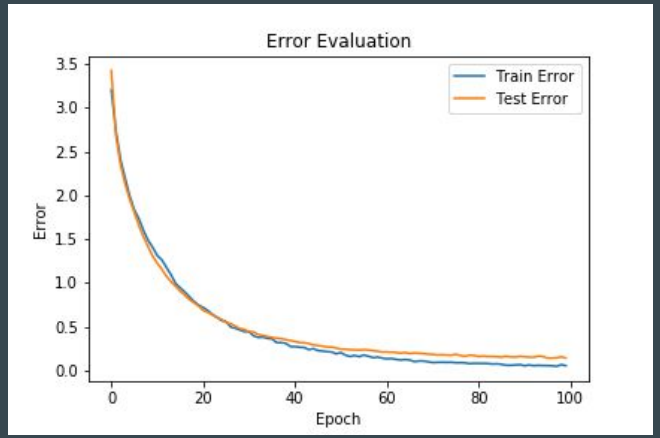


Preprocessing comparison of 20 SVMs each.

Deep Learning

- Multi-layer Perceptron (MLP)
- Convolutional Neural Network (CNN)
- Results (MLP)
  - .90 accuracy with no noise
  - .70 accuracy with light noise
  - .43 accuracy with heavy noise
- Results (CNN)
  - .96 accuracy with no noise
  - .58 accuracy with light noise
  - .09 accuracy with heavy noise

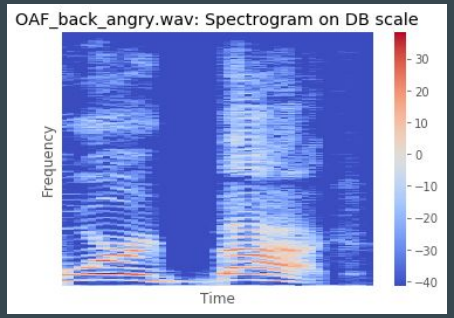




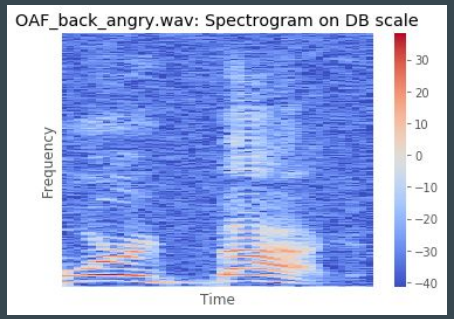
Conclusions

- MFCC preprocessing outperforms the wavelet function.
- Support Vector Machine is quite robust against adding noise.
- Deep learning completely breaks down when adding noise.
  - MLP outperforms CNN when noise is added to the system.
  - CNN extracts features from three dimensional inputs (e.g. images).

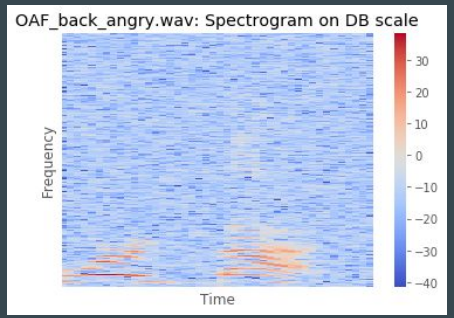
No noise



Light noise



Heavy noise



Questions!