Speech Emotion Analysis

Researchers:
Ryan Culhane | rpc21@duke.edu
Reza Soleimani | rs508@duke.edu
Andre Wang | jw542@duke.edu
Michael Xue | myx2@duke.edu

Project Leads:
Dr. Vahid Tarokh, Duke University
Dr. Jie Ding, University of Minnesota

Project Manager:
Enmao Diao, Duke University

RHODES INFORMATION INITIATIVE AT DUKE UNIVERSITY

Data+

Background

From the Google Assistant to Amazon Alexa, the ways humans engage with machines have changed drastically in the past few years. An intriguing next step in making such human-machine interactions more natural is integrating emotion.

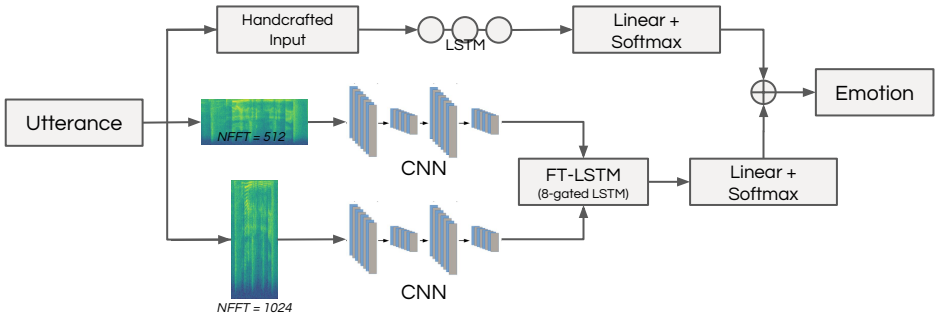
Objectives

1. **Speech Emotion Recognition (SER):** recognize emotion from an utterance

2. **Text-to-Speech Synthesis (TTS):** integrate emotion into speech generated from text

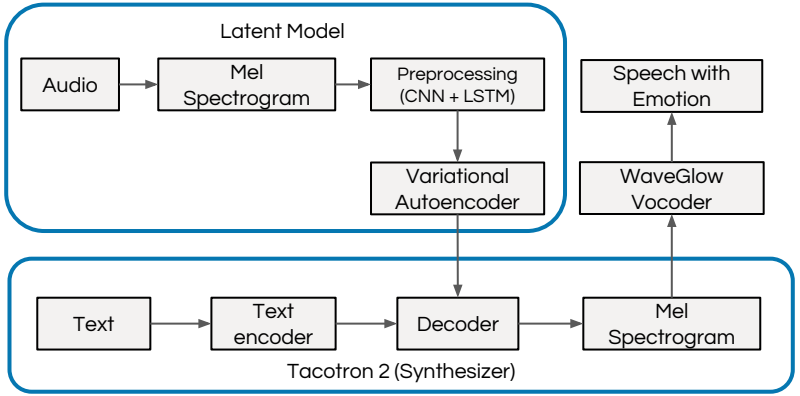
Proposed Models

Speech Emotion Recognition



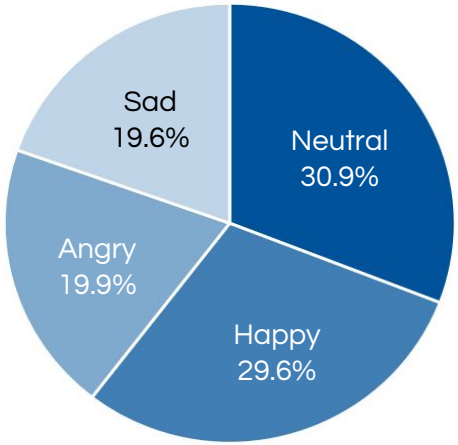
From a given utterance, we create a handcrafted input by splitting it into equal-length segments and extracting handcrafted features from each segment. We also construct spectrograms with two different frequency resolutions and pass them through a CNN in order to learn features. Each of these inputs are passed through an LSTM, followed by a linear layer. Finally, the outputs are added to classify the emotion. **We find that combining handcrafted and learned features raises classification accuracy considerably.**

Text-to-Speech Synthesis



In the latent model, a variational encoder allows us to learn the latent spaces of emotions. When samples from the latent space are input into the decoder of the original Tacotron 2 model, which is able to convert text to speech, we are able to incorporate emotion into generated speech.

Experimental Results



The IEMOCAP database contains over 12 hours of improvised and scripted speech from professional actors. We trained our recognition model on utterances from four, roughly balanced emotions: neutral, happy, angry, and sad.

Speech Emotion Recognition

Model	WA	UA
D. Dai et al. (2019)	65.4%	66.9%
S. Mao et al. (2019)	65.9%	66.9%
R. Li et al. (2019)	-	67.4%
Proposed model	**69.9%**	**70.5%**

Confusion matrix for emotion prediction from highest validation accuracy

Text-to-Speech Synthesis

Scan to listen to examples of synthesized speech, or visit: *https://rpc21.github.io/data-plus-results/*