NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Speaker-Invariant Emotion Recognition with Adversarial Learning

Student: Bryan Leow Xuan Zhen

Supervisor: Professor Jagath C Rajapakse

Motivation

Recent advances in technology have given birth to intelligent speech assistants such as Alexa and Siri which can perform a myriad of tasks just from the end users' voice command. However, they still lack the capability to recognize human emotions when formulating a response. For such speech assistants to be useful to the general population, not only is it important for the underlying Speech Emotion Recognition system be able to recognise human emotions, but the representation captured should also be speaker-invariant.

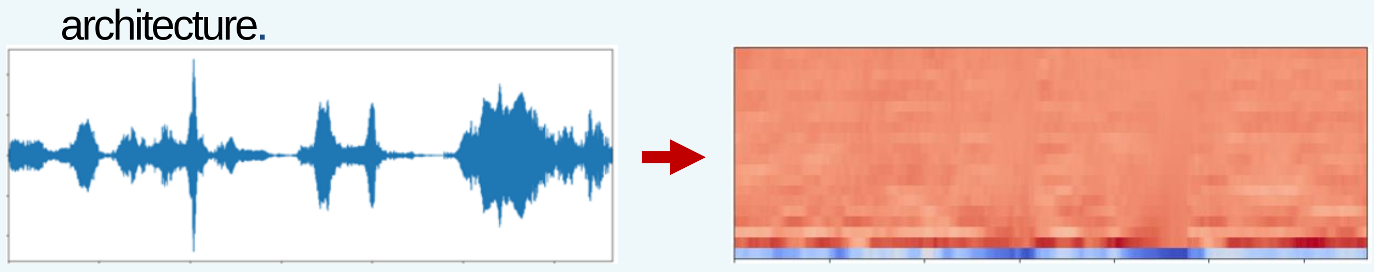
Project Objectives

This project aims to create a novel encoder that can recognize the unseen speakers' emotion from their audio recordings. The encoder is a part of the model architecture which utilized adversarial learning as a framework. The novel encoder should yield better performance than the current state-of-the-art encoders leveraging on the same framework. For demonstration purposes, we also create a web application.

Methodology

Audio Signal Processing

1. Silence and background noise from the raw audio recording is removed using *librosa* and *scipy* package
2. To perform batch training, every input must be of the same size. Hence, the audio recordings are padded to the longest audio length in the dataset.
3. The pre-processed audio recordings are then converted to Mel Frequency Cepstral Coefficients (MFCCs) before feeding into the encoder part of our model architecture.

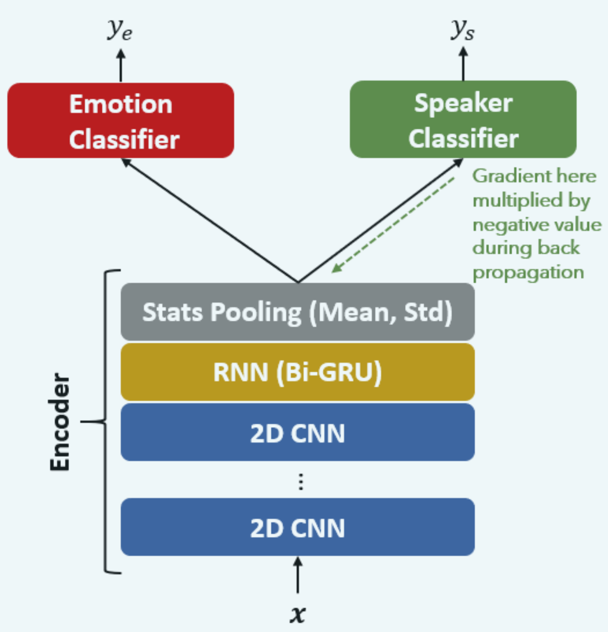


Conversion of a pre-processed audio recording to its MFCCs

Model Architecture

Our model architecture is composed of 3 components:

i. An encoder made from 2D Convolutional Neural Network (CNN) and Bi-directional Gated Recurrent Unit (biGRU) layers and a statistical pooling layer. Together, they generate speaker invariant representations
ii. An emotional classifier that predicts emotional labels
iii. A speaker classifier which removes speaker variability, hence making the entire training an adversarial learning



Experiments and Results

Our experiments are conducted on the Emo-DB and RAVDESS dataset using K-fold leave-two-speakers-out cross-validation and testing. In each fold, 2 speakers are used for validation and 2 speakers and used for testing, with the rest of the speakers used for training. The ratio of male to female speakers in training, validation and testing were set to 1:1.

We experimented with different number of CNN layers for our 2D CNN biGRU encoder. For every depth, we compare the performance of our 2D CNN biGRU encoder with different value of γ where a higher γ value correspond to a higher rate and degree of speaker adversarial training.

No. of CNN layers	γ	Emo-DB	RAVDESS
1	without AL	67.3±6.5	56.3±8.1
	1.25	72.1±8.4	60.6±10.3
	2.50	70.6±5.9	**61.2±8.4**
	3.33	**73.6±7.1**	60.6±10.6
2	without AL	68.3±9.6	56.3±9.8
	1.25	68.0±8.0	59.4±7.6
	2.50	71.0±7.6	60.9±8.5
	3.33	67.6±7.4	60.5±9.9
3	without AL	65.7±10.6	55.8±8.8
	1.25	65.1±7.2	53.7±6.7
	2.50	66.0±5.5	59.4±7.6
	3.33	66.6±5.6	58.9±8.5
4	without AL	59.6±5.3	54.0±11.3
	1.25	66.0±6.9	54.2±8.7
	2.50	62.9±9.6	54.1±9.3
	3.33	65.2±7.2	51.3±11.1

To understand how our 2D CNN biGRU encoder perform relative to other encoders utilising adversarial learning in speaker independent emotion recognition tasks, we perform the same K-fold-leave-two-speakers-out cross-validation and testing on those encoders. Our experiments show that our proposed architecture achieves the best performance compared to existing encoders on both Emo-DB and RAVDESS dataset.

Input features	Encoder	Emo-DB	RAVDESS
MFCC	TDNN biLSTM	61.4±8.7	52.4±10.9
LogMFB, energy, pitch	1D CNN GRU	44.2±4.7	30.5±4.6
MFCC	2D CNN biGRU	**73.6±7.1**	**61.2±8.4**

Web Application for Demonstration