Emotion Recognition from Speech

Chinmay Wadgaonkar, Harshvardhan Singh, Nikhil Anand

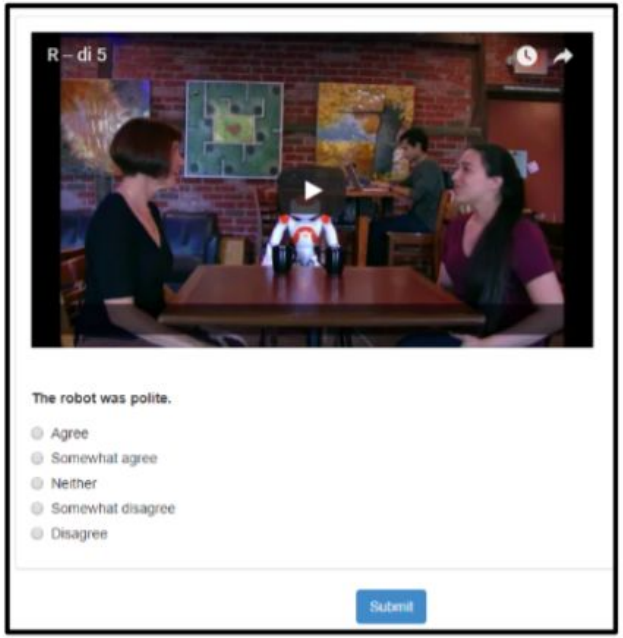
Problem

Classifying the emotions experienced by a person based on their speech and the intonations in their speech.

Influenced by the need for emotion recognition for human-robot interaction.

Collaborative Humans and Robots: Interaction, Sociability, Machine Learning and Art Lab (CHARISMA Lab) led by Dr. Heather Knight.

Examples of Human-Robot Interaction



DATA TYPE CONDITION	ADDRESSEE CONDITION
Database Search a. This person has 3 traffic violations. I would advise caution. b. This person has a clean criminal record, I would go for it.	**Robot Speaks To One (1:1)** g. You seem ready for love h. You don't see ready for love **OR** i. You seem ready for the job j. You don't seem ready for the job
Body Language Analysis c. You both do not look happy together. d. You both look happy together.	**Robot Speaks To One About the Other (1:1-aboutother)** k. This person is my least favorite l. This person is my favorite
Ecological (Control) e. This is your 5th visit this week. f. Did you bring a stamp card?	**Robot Speaks To Both (1:2)** m. You both are cute together n. You both are not cute together

Fig. 4. Participants were assigned to Data Type or Addressee conditions (as in Table II). Those in Data Type experienced statements a-f, while Addressee saw statements g-n. The within-subjects experiment designs allowed participants to condition variations explicitly.

S. Hedaoo, A. Williams, C. Wadgaonkar and H. Knight, "A Robot Barista Comments on its Clients: Social Attitudes Toward Robot Data Use," 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, Korea (South), 2019, pp. 66-74.

Novelty to the Field

A paper that used facial expressions for emotion recognition

Z. Liu *et al*., "A facial expression emotion recognition based human-robot interaction system," in *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 668-676, 2017. doi: 10.1109/JAS.2017.7510622

A paper that used speech and audio for emotion recognition using general machine learning techniques like SVMs and BayesNets

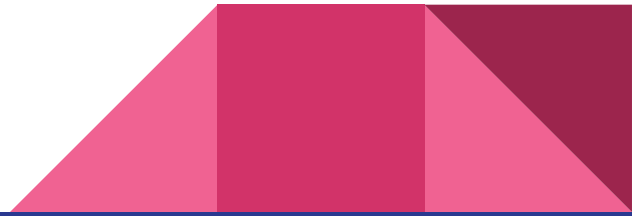
Rázuri, Javier Francisco & Sundgren, David & Rahmani, Rahim & Larsson, Aron & Moran, Antonio & Bonet, Isis. (2015). Speech emotion recognition in emotional feedback for Human-Robot Interaction. International Journal of Advanced Research in Artificial Intelligence. 4. 10.14569/IJARAI.2015.040204.

Datasets

Some popular datasets that we used were

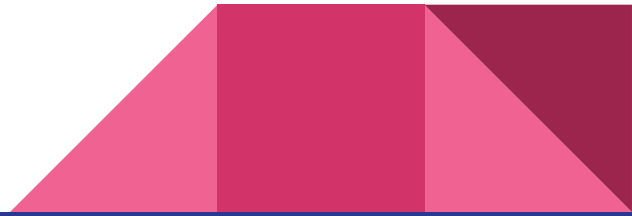
- SAVEE Dataset - All male speakers
- RAVDESS Dataset - 1440 Files - 12 Male/12 Female speakers
- TESS Dataset - All female speakers, varying age

We didn’t use the open source CREMA-D Dataset despite it having a lot of audio files since there was no clear structure and some of the audio files often had lots of background noise.



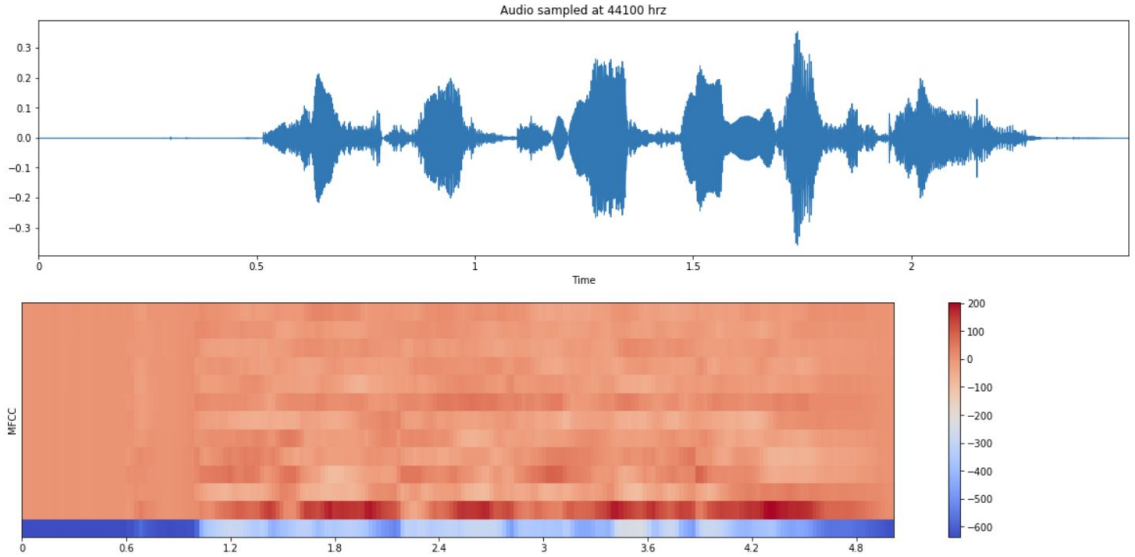
Emotions

- Calm
- Happy
- Sad
- Angry
- Fearful
- Surprise
- Disgust



Technical Approach - MFCC

- Mel-frequency cepstral coefficient (MFCC) -



Technical Approach - 2D Convolutional Model

- 2D Convolutional Model -

conv2d_5 (Conv2D)	(None, 30, 216, 32)	1312
batch_normalization_6 (Batch	(None, 30, 216, 32)	128
activation_6 (Activation)	(None, 30, 216, 32)	0
max_pooling2d_5 (MaxPooling2	(None, 15, 108, 32)	0
dropout_7 (Dropout)	(None, 15, 108, 32)	0
conv2d_6 (Conv2D)	(None, 15, 108, 32)	40992
batch_normalization_7 (Batch	(None, 15, 108, 32)	128
activation_7 (Activation)	(None, 15, 108, 32)	0
max_pooling2d_6 (MaxPooling2	(None, 7, 54, 32)	0
dropout_8 (Dropout)	(None, 7, 54, 32)	0
conv2d_7 (Conv2D)	(None, 7, 54, 32)	40992
batch_normalization_8 (Batch	(None, 7, 54, 32)	128
activation_8 (Activation)	(None, 7, 54, 32)	0
max_pooling2d_7 (MaxPooling2	(None, 3, 27, 32)	0
dropout_9 (Dropout)	(None, 3, 27, 32)	0
conv2d_8 (Conv2D)	(None, 3, 27, 32)	40992
batch_normalization_9 (Batch	(None, 3, 27, 32)	128
activation_9 (Activation)	(None, 3, 27, 32)	0
max_pooling2d_8 (MaxPooling2	(None, 1, 13, 32)	0
dropout_10 (Dropout)	(None, 1, 13, 32)	0
flatten_2 (Flatten)	(None, 416)	0
dense_2 (Dense)	(None, 64)	26688
dropout_11 (Dropout)	(None, 64)	0
batch_normalization_10 (Batc	(None, 64)	256
activation_10 (Activation)	(None, 64)	0
dropout_12 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 14)	910

Total params: 152,654
Trainable params: 152,270
Non-trainable params: 384

Technical Approach - ResNet

- ResNet - Since we are solving a classic image classification problem, it made sense to use ResNet to classify the emotions in audio.

Results

- 2D Convolutional Model

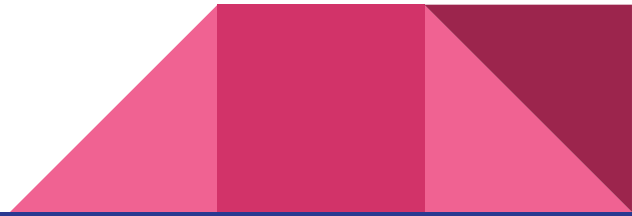
  Training Accuracy - **0.9130**

  Validation Accuracy - ***0.8212***

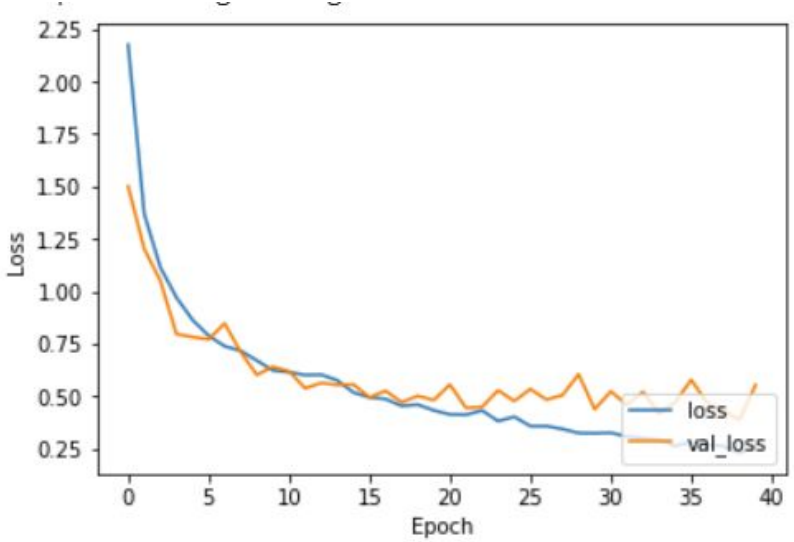
- ResNet

  Training Accuracy - **0.9961**

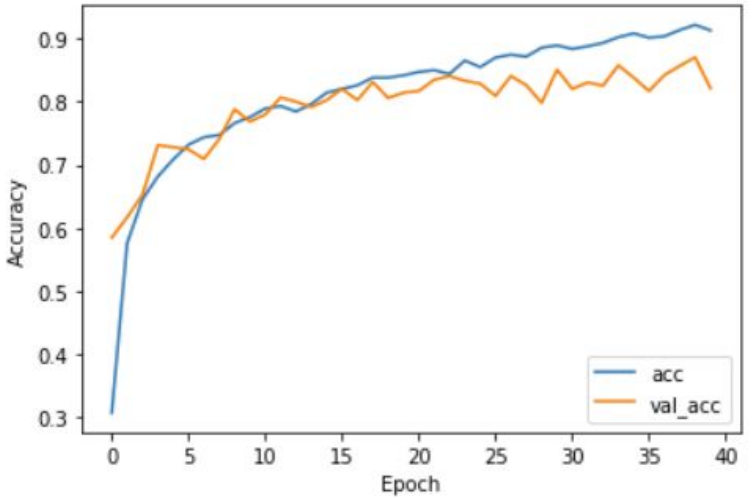
  Validation Accuracy - ***0.8362***



2D Conv Model Results



Loss



Accuracy

Conclusion and Possible Future Work

We compare using a 2D Convolutional Model and a ResNet model to classify the emotions. Though we came up with pretty good validation accuracies, it is evident that ResNet resulted in some overfitting.

However, the low latency of MFCC conversion and model prediction support the idea that this speech recognition can be used in real life settings and human-robot interactions.

Future Work

- Create separate models for Male and Female speakers
- Classify “sentiment” from these separate models