DepressionDetect

A Machine Learning Approach for Audio based Depression Classification

Ky Kiefer

Galvanize Data Science Immersive

kkiefer@umich.edu

kykiefer

kykiefer

Introduction

This effort addresses an automated device for detecting depression from acoustic features in speech. *DepressionDetect* is a tool aimed at lowering the barrier of entry in seeking help for potential mental illness and supporting medical professionals' diagnoses.

Early signs of depression are difficult to detect and quantify. These early signs have a promising potential to be quantified by machine learning algorithms that could be implemented in a wearable artificial intelligence or home device.

Data

- 50 hours of audio interviews from USC's DAIC-WOZ database.
- 189 virtual interview sessions averaging 16 minutes.

Figure 1: Virtual Interview with Ellie.

- Depression classification derived from a PHQ-8 psychiatric survey.
- Scores range from 0 to 24 (>9 labeled as "depressed").

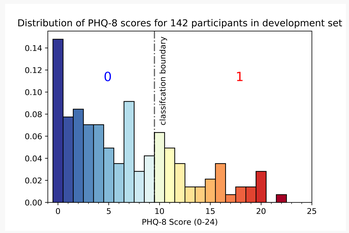

Figure 2: Distribution of PHQ-8 scores.

Methods

Underlying differences in prosody (rhythm, intonation, stress, etc.) exist between depressed individuals and non-depressed individuals.

The participants' speech was segmented from silence and the virtual interviewer using a Support Vector Machine (SVM).

The resulting segmented speech was represented as a spectrogram in the frequency-time domain via a short-time Fourier transform (STFT).

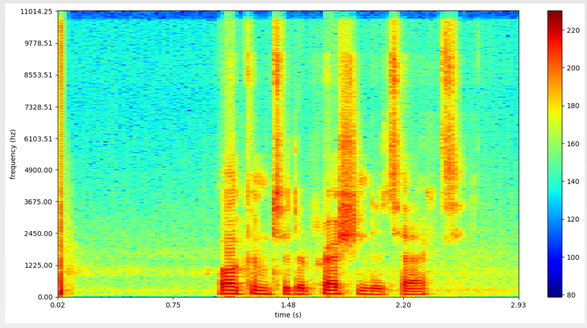

Figure 3: Spectrogram of a plosive (a consonant sound formed by completely stopping airflow), followed by a second of silence, and the spoken words, "Welcome to *DepressionDetect*".

Convolution Neural Networks

Convolutional Neural Networks (CNNs) take images as input.

A convolutional filter (kernel) is slid over each spectrogram input and prosodic features of speech are learned.

A spectrogram with 513 frequency bins x 125 time bins (spanning 4 seconds) is passed as input to the network.

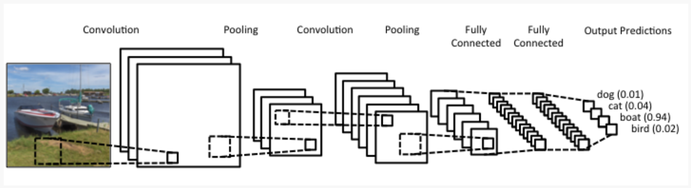

Figure 4: General CNN architecture.

Model Architecture

- 6-layer CNN
- 2 convolution layers with max-pooling
- 2 fully connected layers

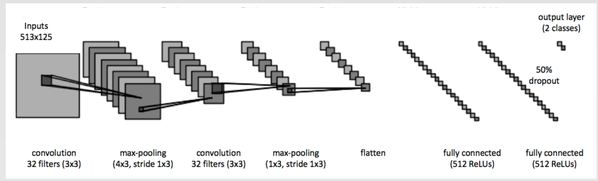

Figure 5: *DepressionDetect* CNN architecture.

Results

The network was trained on 3 hours of audio.

An equal number of spectrograms, from each class (depressed, not depressed) and each speaker, was fed to the network to prevent model bias.

The test (holdout) set was composed of 14 participants with 40 spectrograms per subject (representative of 160s of audio).

Predictions were first made on each of the 4 second spectrograms (Table 1). Then, a majority vote from the 40 spectrograms per participant was utilized to determine a participant's depression label based on160s of subject audio. (Table 2).

	Actual: Yes	Actual: No
Predicted: Yes	174 (TP)	106 (FP)
Predicted: No	144 (FN)	136 (TN)

Table 1: Test set predictions on 4 second spectrograms.

	Actual: Yes	Actual: No
Predicted: Yes	4 (TP)	2 (FP)
Predicted: No	3 (FN)	5 (TN)

Table 2: Test set predictions on 14 participants using majority vote.

Predictive power seems to increase with majority vote, but small sample size.

Conclusion

The model has an AUC score of 0.58, with state of the art emotion detection models ~0.7 using lower level audio representations (i.e. MFCCs).

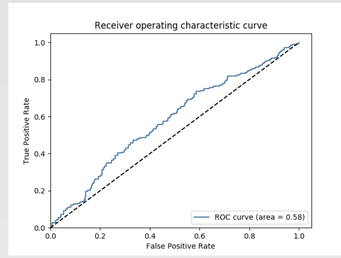

Figure 6: ROC curve on test set.

Results suggest a promising, new direction in using spectrograms for depression detection.

Spectrograms undoubtedly have powerful, learnable features but require larger sample sizes to diminish effects of noise (which lower level representations seek to eliminate).

To combat limited sample size, a web app was built accepting online audio donations for use in periodic model re-training.

Visit DataStopsDepression.com or scan the QR code below to become a data donor! It only takes a couple minutes to help fight depression!

Tech Stack

python, pyAudio Analysis, amazon web services, NumPy, SciPy, matplotlib, IPython Interactive Computing, pandas, scikit learn, Keras, theano, Flask