DepressionDetect: A Machine Learning Approach for Audio based Depression Classification

Kiefer, Ky
kkiefer@umich.edu

ABSTRACT

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

Automatic Depression Detection (ADD) is a relatively nascent topic that first appeared in 2009. DepressionDetect presents a novel approach focusing on two aspects that receive scant research attention: class imbalance and data representation (feature extraction).

1. INTRODUCTION

Early detection and treatment of depression is essential in promoting remission, preventing relapse, and reducing the emotional burden of the disease. Current diagnoses are primarily subjective, inconsistent across professionals, and expensive for the individual who may be in dire need of help. Additionally, 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 (AI) or home device.

2. DATASET

All audio recordings and associated depression metrics were provided by the DAIC-WOZ Database, which was compiled by USC Institute of Creative Technologies and released as part of the 2016 Audio/Visual Emotional Challenge and Workshop (AVEC 2016). The dataset consists of 189 sessions, averaging 16 minutes, between a participant and virtual interviewer called Ellie (Figure 1), controlled by a human interviewer in another room via a "Wizard of Oz" approach. Prior to the interview, each participant completed a psychiatric questionnaire (PHQ-8), from which a binary "truth" classification (depressed, not depressed) was derived.

.

Figure 1: Virtual Interview with Ellie.

3. ACOUSTIC FEATURES OF SPEECH

While some emotion detection research focuses on the semantic content of audio signals in predicting depression, I decided to focus on the prosodic features, which have also been found to be promising predictors of depression. Prosodic features can be characterized by a listener as pitch, tone, rhythm, stress, voice quality, articulation, intonation, etc. Encouraging features in research include sentence length and rhythm, intonation, fundamental frequency, and Mel-frequency cepstral coefficients (MFCCs).

3.1 Segmentation

The first step in analyzing a person's prosodic features of speech is segmenting the person's speech from silence, other speakers, and noise. Fortunately, the participants in the DAIC-WOZ study were wearing close proximity microphones in low noise environments, which allowed for fairly complete segmentation in 84% of interviews using pyAudioAnanlysis' segmentation module. When implementing the algorithm in a wearable device, speaker diarization (speaker identification) and background noise removal would require further development for a more robust product. However, in the interest of quickly establishing a minimum viable product, this desired further development was not addressed in the current effort.

3.2 Features Extraction

There are several ways to approach acoustic feature extraction, which is the most critical component to building a successful approach. One approach includes extracting short-term and mid-term audio features such as MFCCs, chroma vectors, zero crossing rate, etc. and feeding them

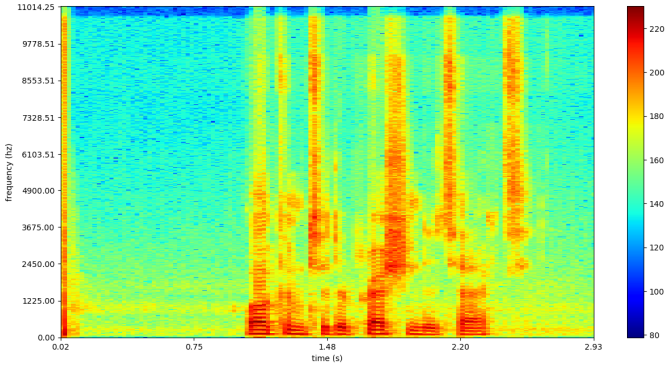

Figure 2: Spectrogram of a plosive, followed by a second of silence, and the spoken words, "Welcome to DepressionDetect".

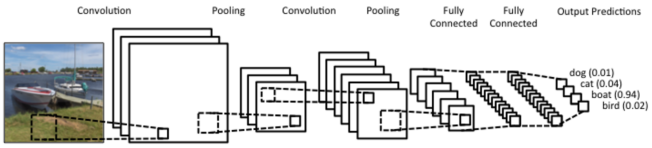

Figure 3: General CNN architecture.

as inputs to a Support Vector Machine (SVM) or Random Forest. Since pyAudioAnalysis makes short-term feature extraction fairly streamlined, my first approach to this classification problem involved building short-term feature matrices from 50ms audio segments of the 34 short-term features available from pyAudioAnalysis. Since these features are lower level representations of audio, the concern arises that subtle speech characteristics displayed by depressed individuals would go undetected.

Running a Random Forest on the 34 short-term features alluded to yielded an encouraging F1 score of 0.59, with minimal tuning. This approach has been previously employed by others, so I treated this as "baseline" comparative data for which to develop and evaluate a completely new approach involving convolutional neural networks (CNNs) with spectrograms, which I felt could be quite promising and powerful.

CNNs require a visual image. In this effort, speech stimuli is visually represented via a spectrogram. A spectrogram (Figure 2) is a visual representation of sound, displaying the amplitude of the frequency components of a signal over time. Unlike MFCCs and other transformations that represent lower level features of sound, spectrograms maintain a high level of detail (including the noise, which can present challenges to neural network learning).

4. CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNNs) are a variation of the better known Multilayer Perceptron (MLP) in which node connections are inspired by the visual cortex. CNNs (Figure 3) have proven to be a powerful tool in image recognition, video analysis, and natural language processing. More germane to the current effort, successful applications have also been applied to speech analysis.

CNNs take images as input. In the case of the spectrogram, I pass (or input) a grayscale representation, with the "grayness" representative of the audio power level at that specific frequency and time. A filter (kernel) is subsequently slid over the spectrogram image and patterns for depressed and non-depressed individuals are learned (based on the aforementioned "truth" dataset).

The CNN begins by learning features like vertical lines, but in subsequent layers, begins to pick up on features like the shape of frequency-time curve (perhaps representative of speaker intonation). Such learned features may provide an elegant and powerful representation of different prosodic features of speech, which in turn are representative of underlying differences between depressed and non-depressed speech.

However, with the highly detailed representations of speech provided in spectrograms, false noise signals (ambient noise, plosives, unsegmented audio from other speakers, etc.) can be inconveniently picked up by the network. One can mitigate this noise with different regularization parameters in the network (pooling layers, L1 loss functions, dropout, etc.), but unless your training data is abundant, it is challenging for the network to distinguish real predictors of depression from the false signal.

4.1 Class Imbalance

In the current dataset, the number of non-depressed subjects is about four times larger than that of depressed ones, which can introduce a classification "non-depressed" bias. Additional bias can occur due to the considerable range of interview durations from 7-33 minutes because a larger volume of signal from an individual may emphasize some characteristics that are person specific.

In an attempt to address these issues, each of the participant's segmented spectrograms were cropped into 4 second slices. Next, participants were randomly sampled in 50/50 proportion from each class (depressed, not depressed). Then, a fixed number of slices were sampled from each of the selected participants to ensure the CNN has an equal interview duration for each participant. This drastically reduced the training dataset size to 3 hours from original 35 hours of segmented audio, which was felt adequate for this exploratory analysis.

It should be noted a few different sampling methods were explored to try to increase the size of the training data, and all resulted in highly biased models in which only the "non-depressed" class was predicted. A revised sampling method should be considered as high-priority in future directions, which I hope to implement and explore in future directions to increase the training sample size.

4.2 Model Architecture

A 6-layer Convolutional Neural Network (CNN) model was employed consisting of 2 convolutional layers with max-pooling and 2 fully connected layers. Each spectrogram input is an image with dimension 513x125 representing 4 seconds of audio and frequencies ranging from 0 to 8kHz. The frequency range was tuned as a hyperparameter, since most human speech energy is concentrated between 0.3-

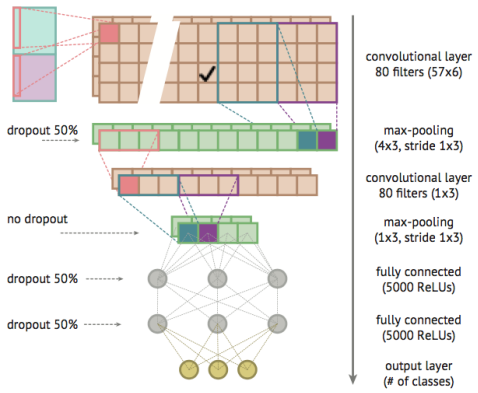

Figure 4: Environmental Sound Classification CNN architecture.

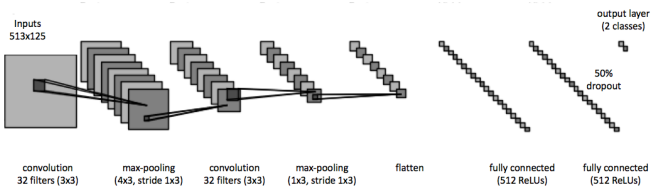

Figure 5: DepressionDetect CNN architecture.

3kHz. Each input is normalized according to decibels relative to full scale (dBFS).

Though there are some differences, the actual architecture employed in this effort was largely inspired by a paper on Environmental Sound Classification with CNNs. The network employed in this paper is shown in Figure 4, with DepressionDetect's displayed in Figure 5.

The CNN used here begins with an input layer being convolved with 32-3x3 filters to create 32 feature maps followed by a ReLU activation function. Next, the feature maps undergo dimensionality reduction with a max-pooling layer, which uses a 4x3 filter with a stride of 1x3.

A second similar convolutional layer is employed with 32-3x3 filters followed by a max-pooling layer with a 1x3 filter and stride of 1x3.

This layer is then followed by two dense layers. After the second dense layer, a dropout layer of 0.5 is used (meaning each neuron in the second dense layer has a 50% chance of turning off after each batch update).

Lastly, a softmax function is applied, which returns the probability that a spectrogram is in the depressed class or not depressed class. The sum probabilities of each class is equal to 1. A batch size of 32 (out of 2480 spectrograms) was used along with an Adadelta optimizer, which dynamically adapts the learning rate based on the gradient.

4.3 Training the Model

I created the model using Keras with a Theano backend and trained it on an AWS GPU-optimized EC2 instance.

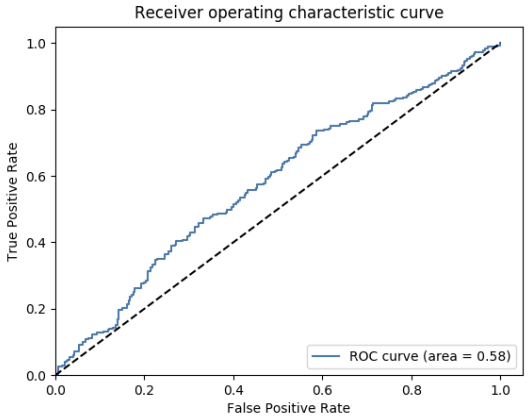

Figure 6: ROC curve of the CNN model.

The model was trained on 40 randomly selected 513x125 (frequency x time bins) audio segments from 31 participants in each category of depression (resulting in 2,480 spectrograms in total). The network was trained on just under 3 hours of audio in order to adhere by strict class (depressed, not depressed) and speaker balancing (160 seconds per subject) parameters. The network was trained for 7 epochs, after which it was observed to overfit based on train and validation loss curves.

4.4 Results

I assessed my model and tuned my hyperparameters based on AUC score (Figure 6) and F1 score on a training and validation set. AUC scores are commonly used to evaluate emotion detection models, because precision and recall can be misleading if test sets have unbalanced classes (although they were balanced with this approach).

The test set (which is distinct from the train and validation sets used to develop the model) was composed of 560 spectrograms from 14 participants (40 spectrograms per participant, totaling 160 seconds of audio). Initially, predictions were made on each of the 4 second spectrograms, to explore extent to which depression can be detected from 4 second audio segments. Ultimately, a majority vote of the 40 predictions per participant was utilized to label the participant as depressed or not depressed.

Table 1 provides a summary of the predictive power using 4 second spectrograms with Table 3 using the "majority vote" approach.

As stated above, a majority vote across the 40 spectrograms per participant was also explored as a means to predict each participant's depression category. Only 14 users were contained in the test set (in order to maximize the training data), so I wanted to be sure to include statistics on individual spectrogram predictions as well as the majority vote approach. As might be expected, model evaluation statistics improved somewhat when taking a majority vote. However, the sample size is quite small.

State of the emotion detection models exhibit AUC scores around 0.7 (my model had an AUC score of 0.58), utilizing the lower level features alluded to. Although, this rapidly developed model is not yet at a predictive state for practical usage "as is", these results strongly suggest a promising, new direction for using spectrograms in depression detection.

Table 1: Test set predictions on 4 second spectrograms		
Confusion Matrix	Actual: Yes	Actual: No
Predicted: Yes	174 (TP)	106 (FP)
Predicted: No	144 (FN)	136 (TN)

Table 2: Test set statistics on 4 second spectrograms			
F1 score	precision	recall	accuracy
0.582	0.621	0.547	0.555

Table 3: Test set predictions using majority vote		
Confusion Matrix	Actual: Yes	Actual: No
Predicted: Yes	4 (TP)	2 (FP)
Predicted: No	3 (FN)	5 (TN)

Table 4: Test set statistics using majority vote			
F1 score	precision	recall	accuracy
0.615	0.667	0.571	0.643

5. DONATE YOUR DATA

The model needs your help! Detecting depression is hard. Robust speech recognition models rely on hundreds of hours of audio data. The good news is that you can contribute! Visit www.DataStopsDepression.com to become a data donor! Your audio data will be incorporated in periodic model retraining with a batch algorithm.

The donation process:

1. Record a 40 second anonymized clip of you reading a provided paragraph (and see a cool spectrogram of your audio recording!).
2. You will be prompted to complete an 8 question psychiatric survey.

6. FUTURE DIRECTIONS

I ultimately envision the model being implemented in a wearable device (Apple Watch, Garmin) or home device (Amazon Echo). The device could prompt you to answer a simple question in the morning and a simple question before bed on a daily basis. The model stores your predicted depression score and tracks it over time, such that the model can learn from your baseline (perhaps using a Bayesian approach). If a threshold is crossed, it notifies you to seek help, or in extreme cases, notifies an emergency contact to help you help yourself.

This initial model provides a solid foundation and promising directions for detecting depression with spectrograms. Further work should train the model in more speakers. Low level audio transformations do a good job of reducing the noise in the data, which allows for robust models to be trained on smaller sample sizes. However, I still hypothesize they overlook subtleties in depressed speech.

I am prioritizing future efforts in the following areas (ordered by priority):

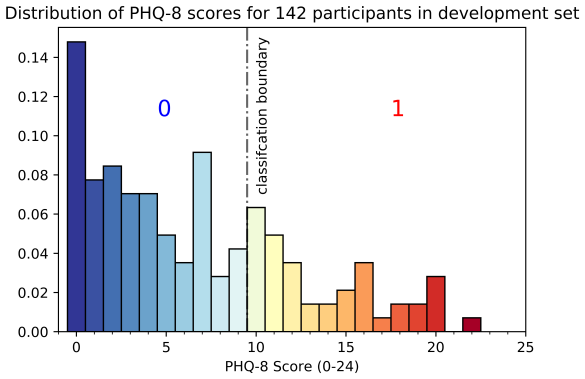

Figure 7: Distribution of PHQ-8 scores.

1. Sampling methods to increase training size without introducing class or speaker bias.
2. Treating depression detection as a regression problem (see below).
3. Introducing network recurrence (LSTM).
4. Incorporate Vocal Tract Length Perturbation (VTLP).

Depression moves across a spectrum (Figure 7), so deriving a binary classification (depressed, not depressed) from a single test (PHQ-8) is somewhat naive and perhaps unrealistic. The threshold for a depression classification was a score of 10, but how much difference in depression-related speech prosody exists between a score of 9 (classified as not depressed) and a 10 (classified as depressed)? For this reason, the problem may be better approached by using regression techniques to predict participants' PHQ-8 scores and scoring the model based on RMSE.

7. REFERENCES

1. Gratch, Artstein, Lucas, Stratou, Scherer, Nazarian, Wood, Boberg, DeVault, Marsella, Traum. The Distress Analysis Interview Corpus of human and computer interviews. InLREC 2014 May (pp. 3123-3128).
2. Girard, Cohn. Automated Depression Analysis. Curr Opin Psychol. 2015 August; 4: 75-79.
3. Ma, Yang, Chen, Huang, and Wang. DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. ACM International Conference on Multimedia (ACM-MM) Workshop: Audio/Visual Emotion Challenge (AVEC), 2016.
4. Giannakopoulos, Aggelos. Introduction to audio analysis : a MATLAB approach. Oxford: Academic Press, 2014.
5. Piczak. Environmental Sound Classification with Convolutional Neural Networks. Institute of Electronic System, Warsaw University of Technology, 2015.