ECE214B - Miniproject 1: Depression Detection using Speech Signals

David Castro, Roy Jara, Neil Kane

UCLA

davcstro@g.ucla.edu, rmjara@g.ucla.edu, neilkane1317@g.ucla.edu

Abstract

Emotion recognition is an emerging field in speech processing and has seen various advancements in conjunction with fields in natural language processing. In addition to these tasks one subset is depression detection which can have significant medical impact. Using a CNN-LSTM model we are tasked with using different feature extraction techniques to improve a given depressed vs non-depressed classifier. Specifically we explored features from the spafe and parselmouth libraries to extract features like BFCC, GFCC, LFCC, along with pitch and formant information. We achieve an average F1 score of 61.87% computed from scores of 85.04% and 38.71% for depressed and non-depressed classes respectively. We discuss findings and potential avenues for continuing this work.

Index Terms: automatic depression recognition, speech processing, computational paralinguistics

1. Introduction

Depression is a mental disorder that can impede people from flourishing, and in the most dire of cases, lead people to take their own life. Since the advent of psychology, experts have devised ways to diagnose this disorder. The primary diagnostic tool today is the DSM-5, a questionnaire to be answered by the person receiving the diagnostic [1]. However, this diagnostic method has several barriers including inaccessibility, stigma, inability to scale, and lack of objectivity, impeding the proper diagnostic of all the people who might be experiencing depression. As a result, automatic methods have been proposed as a scalable solution, capable of providing objective measurements and bypassing any misleading answers from patients [2]. In this project, we propose an automated depression recognition system based on user speech.

The goal of this project is to identify clinical depression and non-depression in mandarin speakers from the Emotional Audio-Textual Depression Dataset (EATD) Corpus [3] through the use of a neural network. Different acoustic features can be used in order to train the system, resulting in different accuracies based on the features.

2. Background

There have been numerous studies seeking to identify speech characteristics seen in people with depression. Noticeable differences are seen in the pitch, formant frequencies, speaking rate, and energy. Due to the monotonic nature of depressive speech, pitch will have a much smaller variation or range in depressed speech [4]. Formant Frequencies (the resonant frequencies of the vocal tract) are lower than in non-depressed speech, possibly due to the stress causing tightness in the muscles along the vocal tract [4]. A slower speaking rate is not quite understood why it occurs in motor speakers, but it is considered a very strong feature for identifying depression [4]. Energy or loudness in depressive speakers is noted to be either overly loud or not varying their loudness [4]. Based on these speech characteristics, certain features can be targeted in order to recognize the patterns seen in depressive speech.

In addition to these knowledge based approaches another technique that can be utilized are various representations of cepstral coefficients. These are very common in speech recognition systems and follow a traditional digital signal processing approach by computing a frequency spectrum to capture both temporal and spectral information based on varying window lengths and hop lengths used. Also the number of FFT bins used in this computation has an impact on the types of information that can be recovered when considering the tradeoffs between frequency and temporal resolutions based on the window length used. A log of this spectrum is taken before computing an inverse which then produces our desired cepstral coefficients as a representation of the spectral envelope of the speech signal.

The baseline model in this project uses a mel spectrogram representation as the input feature. This representation better approximates human hearing by mimicking the frequency resolution of the human auditory system. In addition to this we will explore some other filter banks in an effort to gain some empirical results from different hearing models, specifically the Bark, Gammatone, and Linear filter banks. The motivation for these lies partially due to some existing results which will be discussed further in their respective section.

3. Project Description

The baseline model for the neural network is a CNN-LTSM. Different matrices of features can be combined in order to train the model. Many features were tested and each provided their different performances.

3.1. Prosodic Features

Prosodic features such as pitch and speaking rate were examined through a Parselmouth, a python library, based on how PRAAT calculates pitch and formant frequencies. Pitch samples were taken every 10ms and then grouped into frames by a sliding time window. From here the maximum pitch in the window, lowest pitch in the window, and the range between these two values was taken to form the feature matrix. The thought process for modeling speaking rate was taking the derivative of the formant frequencies. If someone is speaking slower, the formants will stay a bit more constant for longer duration. If there is a desire to combine the derivative of formant frequencies with other features (frame level features, e.g. MFCC), there needs to be consistency in the dimensions. This is where framing the formant frequencies also comes into place; same idea as mentioned for pitch. Here in each time window or frame, the formant frequencies are averaged and then the derivative is taken on the frame level and not the sample level. By doing this, the derivative of formant frequencies can now be combined with other frame level features.

3.2. Cepstral Features

3.2.1. GFCC

Following the direction of the mel spectrogram baseline feature representation, we also considered the Gammatone Frequency Cepstral Coefficients based on gammatone filter banks. These values serve a similar goal as the mel scale with various bandpass filters meant to mimic the human auditory system and critical bands to capture spectral details of the incoming speech signal. It is often used for speech signals that have higher frequency information and from the intuition we know about depressed vs nondepressed individuals we might expect there might be a difference in this metric. In addition to this aspect of the filter banks some research has shown that it is a more accurate model of how the cochlea processings varying frequency signals than the Mel Scale and was another motivator to consider this feature.

Additionally, by increasing the feature size (the number of frames to be processed together) from the baseline 120 frames to 240 frames, we were able to see significant improvements in the model's accuracy. This goes to show that the best performing hyperparameters for training models are dependent on the qualities of the input features.

3.2.2. BFCC

In addition to the GFCCs another cepstral metric utilized was Bark Frequency Cepstral Coefficients based on the bark scale filter banks. The shape of the filter banks serve to unify distances between frequency bands and match them with the perceptual differences between frequencies. In essence the idea is similar to what was seen with the Mel and Gammatone filter banks meant to model the human ear. Another motivator was that an existing paper on emotion recognition found promising results when using BFCC values [5]. Their research was tasked with classifying emotions for acted and natural speech and although different models were used from this project it leads to strong results showing the features extracted have identifiable differences. As a result these values were also considered for the task of depression detection which intuitively seems correlated with negative emotions that might be differentiable with the BFCC values.

3.2.3. LFCC

Lastly another cepstral value explored was Linear Frequency Cepstral Coefficients that make use of linear filter banks. This is a contrasting idea with the baseline Mel Scale and other two cepstral scales of bark and gammatone as here the idea is not to accurately mimic the human ear. We know that the ear has a greater resolution for lower frequencies and it is easier to tell different towns apart in these ranges however for higher frequencies the same true delta in frequency value can appear to sound the same. This places an inherent bias against these higher frequencies and as we have discussed already we might expect there to be a big difference exactly in these larger frequency values between depressed and non depressed individuals. Thus, unlike the human ear, for classification purposes we do in fact care about the finer details at the higher range of speech signals. By using a linear scale we don't have any bias (for or against) towards a frequency band and thus might be able to extract more useful information uniformly across all values.

3.2.4. Cepstral Overview

All together each of these cepstral values attempt to consolidate energy based windows prior to performing a log transformation and DCT. The main differences lie in how the energy is modeled within each window and is largely dependent on the shape of the filter banks used to convert the original spectrogram into one more "desirable" for feature extraction. Many of these approaches have tangible explanations largely surrounding the model of the human ear but also can be based on intuition. The main one in the context of this project is the idea that the frames may have significant energy or loudness differences between the two classes of depressed and nondepressed. Varying how these frames are ultimately consolidated may lead to learnable differences that hopefully improve the accuracy of the CNN-LSTM model. As a final step building on the concept of energy differences a natural progression is to consider the derivatives of these values as we might expect there to be different rates involved as well.

3.3. Low Level Descriptors

Finally, we used additional features from both Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [6] and Computational Paralinguistics Challenge Feature Set (ComParE-2016) [7]. These features were extracted using the implementations from the software OpenSMILE [8]. Out of the numerous available features, we focused on the most relevant for the task. These included: the pitch (F0), Root Mean Squared (RMS) energy, spectral rolloff, spectral entropy. Their derivatives were also used as input features to better grasp the speaker's speaking patterns and how they change from frame to frame.

3.4. CNN-LSTM Model + Variations

The baseline model, which we used for the majority of the project was a neural network consisting of a 1D Convolutional Layer (with Maxpooling, Batch Normalization, and Dropout), two LSTM layers, and a dense layer with sigmoid activation. As mentioned in [9] convolutional layers are excellent for learning from data since the kernels used to perform the convolution can adapt their weights to find representations not detected otherwise. In this case, these learned representations convolutional layer were then fed to two LSTM layers. The LSTM layers allow the overall network to capture sequential dependencies, which is ideal for sequential data such as speech. LSTM layers' ability to selectively choose which sequential dependencies are relevant given a certain input, allow for more efficient and precise learning when compared to Recurrent Neural Networks.

Furthermore, seeing how some of the previously mentioned features come from diverse procedures or diverse algorithms, we sought to adapt the baseline model so that different features can contribute to the final classification with different importance. As seen in Figure 1, we decided to use the CNN to process some spectral or cepstral feature sets and then concatenate its output with prosodic features (such as F0). The intention was to have the prosodic features have a more direct influence on the final prediction.

4. Summary and Discussion

The best accuracy came from the derivative of the formant frequencies (non-framed), and the LFCC with its derivatives where they both have around a 61.8% F1 accuracy. There was an expectation of the derivative of formant frequencies producing

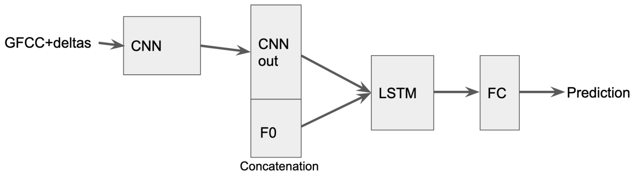

Figure 1: *Modified CNN-LSTM network.*

some of the better results because speaking rate is one of the strongest ways to identify speech from people with depression. Pitch, another feature for depression identification, did not see the same success as the other results, which can be due to how the pitch features were chosen. Since samples were taken from both male and female, only the range difference may be best to take and not maximum/minimum. One thing to note is the degradation in accuracy for the windowing of the formant frequencies. This is likely due to the use of too large of a window. A larger time window means a larger time frame for when the derivative is taken; not enough of the finer change is captured. Cepstral coefficients of energy frequency windows (e.g. MFCC, LFCC, BFCC, etc) each provided similar accuracy around the 50% to 60%. The similarity in accuracy occurs because they follow the same steps, but their emphasis on what frequency bands differ. Combinations of different features did not necessarily result in an increase in accuracy. This could be due to the extra feature not providing any new information to the neural network, but instead masks the more accurate feature.

5. Conclusion and Future Work

Depression and non-depression identification is an open ended research project. Different audio features from cepstral coefficients, and prosody were examined to see if characteristics seen in speech for those with clinical depression could be captured to improve the accuracy of the neural network. An average F1 accuracy high of 61.87% was achieved, but there is much more room for improvement. Some ideas worth testing include using different window lengths for frame based features, including alternative speaking rate features, like energy or loudness based features. Additionally, experimenting with alternative network architectures and different hyperparameters might be conducive to more accurate depression detection models. A much larger idea to explore would be to take into consideration the words used by the speaker to possibly capture semantic features that cannot be seen from audio alone. These ideas are just a bit of what can be explored in order to possibly achieve greater accuracy.

6. References

[1] N. Schimelpfening, "When were the earliest accounts of depression?" [Online]. Available: https://www.verywellmind.com/who-discovered-depression-1066770

[2] L. Lin, X. Chen, Y. Shen, and L. Zhang, "Towards automatic depression detection: A bilstm/1d cnn-based model," *Applied Sciences*, vol. 10, no. 23, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/23/8701

[3] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6247–6251.

[4] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639315000369

[5] D. Kamińska, T. Sapiński, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, no. 1, p. 3, Feb 2017. [Online]. Available: https://doi.org/10.1186/s13636-017-0100-x

[6] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[7] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, vol. 8. ISCA, 2016, pp. 2001–2005.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

7. Appendix A: Table of Experiments

Table 1: *Performance Metrics*

Name	AVG F1	F1 ND	F1 D
emobase-trimmed	0.506	0.8125	0.2
compare2016-trim+deltas	0.542	0.8841	0.2
MFCC_13	0.4552	0.9103	0
Pitch = Fmax, Fmin, MaxRange	0.4514	0.7778	0.125
bfcc	0.5595	0.816	0.303
gfcc	0.5775	0.8824	0.2727
ngcc	0.4397	0.8794	0
lfcc	0.5949	0.9231	0.2667
Deriv. 1-3 Formant Freq	0.6187	0.8504	0.3871
1st & 2nd Deriv. 1-3 Formant Freq	0.5993	0.8986	0.3
Deriv. 1st Freq (Framed)	0.4873	0.708	0.2667
Deriv 1st Format Freq & pitch (Framed)	0.5143	0.8217	0.2069
Deriv 1-3 Freq (Framed)	0.4719	0.8529	0.0909
Deriv 1-3 Format Freq & pitch (Framed)	0.4719	0.8529	0.0909
gfcc 7 + deltas	0.6081	0.8413	0.375
gfcc+F0+RMSEnergy+SpecEntropy	0.5775	0.8824	0.2727
gfcc + deltas	0.5611	0.9	0.2222
bfcc + deltas	0.4317	0.8633	0
lfcc + deltas	0.6184	0.8889	0.3478
ngcc + deltas	0.5298	0.9167	0.1429
gfcc7+deltas+F0	0.5786	0.8346	0.3226