CSCE 633 Project Report

Jacob Fenger

1 Introduction

Within the past couple of years, there has been a rise human-machine technology such as Amazon Echo, Google Home, Apple HomePod, and other devices. Becoming more user-friendly is vital to the success of these devices. Interpreting emotions are a huge component of human interaction and devices that can successfully recognize certain emotions may result in increased usability. For example, a device could alter its response based on the emotional state of the user.

2 Literature Review

There have been a good number of studies that have tried to accurately capture emotion using speech and I will list several of them below:

One group of researchers utilized a combination of spectral and prosodic features to get an 88.35% recognition rate for five different emotions: anger, happy, surprise, netural, and sad [1]. They used these feature types to train a Gaussian mixture model and a support vector machine and then combined them to increase classification accuracy.

Another group of researchers had a similar approach in which they used both prosodic and acoustic features for classification. They used Principle Component Analysis (PCA) as well as Linear Discriminant Analysis (LDA) to find a better representation of the extracted features and then used several different models to test classification accuracy [2]. They achieved an average of 83.33% accuracy when using both PCA and LDA on the speech features.

An older paper from researchers proposed a method of speech emotion recognition by the use of hidden Markov models [3]. They managed to achieve an 86% recognition rate for seven discrete emotions.

3 Problem Formulation

For this project, I decided to see if I could accurately classify emotion based upon speech signals. While this project does not address any problems that have not been questioned by other researchers, it was a good introduction to this topic.

The data I used for this project comes from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database collected by the SAIL lab at the University of Southern California [4]. See section 5 below for a more detailed data description.

4 Proposed Solution

For my solution, I used the IEMOCAP database which contained speech signals from actors doing both improvisational and situational dialog which had certain emotion labels [4]. The dialog was segmented into turns which are defined as a continuous segment in which a single actor was speaking. The assumption I made was that the expressed emotion was consistent throughout the whole turn. I decided to just use improvisational turns which resulted in 4784 samples in total.

Once the samples were collected, I used the openSMILE feature extraction tool to extract features from each of the samples [5]. I generated features based on the INTERSPEECH 2009 Emotion Challenge feature set which had a configuration file within openSMILE. This feature set contained 384 features as statistical functionals applied to root-mean-square signal frame energy, mel-frequency cepstral coefficients 1-12, and several other low-level descriptor countours.

I also had to assign labels to each of the samples which was done by parsing certain files within the IEMOCAP database to find the labels that were assigned to each turn. Emotions were assigned an integer value and similar emotions were consolidated into a single category. For example: happiness, excitement, and surprise were all grouped into a single group called happiness. This was done because the distribution of the emotions was very skewed and there were not many samples with certain labels. This would have resulted in a very low recognition accuracy for many of the emotions. Figure 1 shows the distribution of the consolidated emotional groups.

Once features were extracted from the data and labels were assigned, I split the data up into an 80-20 train-test split. I then trained a support vector machine (SVM) with a radial basis function (RBF) kernel. All of the implementation was done in Python. The definiton of the RBF kernel is shown below:

$$K(x, x') = exp(-\gamma||x - x'||^2)$$

5 Data Description

As mentioned above, the data used for this project came from the IEMOCAP database [4]. The data for this database was collected from ten actors which performed three selected scripts as well as separate improvisational dialogs given scenarios designed to express certain emotions. Each actor was visually and vocally recored and markers were

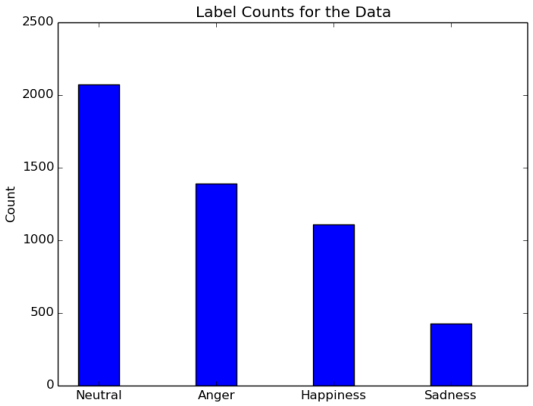

Fig. 1. Histogram of the Sample Labels After Consolidation

placed to capture hand and head motion during the recordings. There were five sessions in total where both female actresses and male actors were recorded. The recordings were manually segmented into turns (Described above) and each turn was assigned an emotional label.

6 Results

With the model I trained I was able to get a recognition accuracy of 44.5% with the test set. This was with an unweighted SVM which may have been the reason for such a low accuracy. Results could be improved if I were to weigh the samples with lower label representations with higher weight values. At the time of writing this, I am still working through this problem.

7 Future Work

The combination of using speech signal analysis as well as another medium would result in a much higher accuracy. One possible way to create a better classifier is to combine a visual emotion classifier with a speech classifier. IEMOCAP already contains visual and audio data so it could be an extension of what I have done so far.

Another approach to this problem that I have been considering is the use of natural language processing to capture emotion. Are the words spoken by people indicative of emotion? I am sure people choose the words they say based on the emotion they are feeling, but this is a complicated issue as there are many things that factor into this. It may depend on who the subject is talking to, the language they are speaking, or even the person themselves.

8 Conclusions

Capturing emotion via speech is a very difficult task to do as emotion is an inherently complex behavior. People express emotions in different ways and many emotional cues are non-verbal. In this document, I described an elementary method for classifying emotion via speech using a support vector machine on a feature space consisting of low-level features. There is still a lot of room for improvement.

References

[1] J. Z. Y. Zhuo, Y. Sun and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," Institute of Acoustics, Chinese Academy of Sciences, Tech. Rep., 2009.

[2] M. Y. M. Hoque and M. Louwerse, "Robust recognition of emotion from speech," The University of Memphis, Tech. Rep.

[3] G. R. B. Schuller and M. Lang, "Hidden markov model-based speech emotion recognition," Insitute for Human-Computer Communication, Technische Universitt Mnchen, Tech. Rep., 2003.

[4] C. L. C. Busso, M. Bulut, A. Kazemzadeh, E. Mower, S. Kim, J. Change, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," University of Southern California, Tech. Rep., 2008.

[5] F. G. F. Eyben, F. Weninger and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," *In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838*, 2013.