TEXAS A&M UNIVERSITY

Human Behavior Analytics

Final Project Report

Exploring Significance of Speech Features for Emotion Recognition

Submitted By:

Saurabh Kumar

Akhil Babu Manam

Venkata Aditya Chintala

1. Abstract

Emotion is a relatively brief conscious experience characterized by intense mental activity and a high degree of pleasure or displeasure. Emotion is experienced when the brain reacts to the information received via the body's nervous system. In 1970s, American psychologist Paul Eckman classified the basic human emotions into 6 categories: anger, disgust, fear, happiness, sadness, and surprise. In this project, we are developing a technology that improves the accuracy of emotion classification by finding hidden patterns in the speech signals.

2. Introduction

Emotion recognition is a highly researched field right now as it has many applications in the field of human computer interaction (HCI) and in designing systems that can act according to a person's state of mind. Human speech contains information related to emotional state of a person either in the language or in the way they utter speech. For example, there is some difference in the way people say "What?" when they are surprised from when they are angry. In this work, we plan to explore various features in speech literature for the task of emotion recognition, analyzing the uses and disadvantages that various features pose and compare their performances.

3. Related Works

Emotion recognition can be used potentially in various tasks related to HCI to make them more interactive and natural. Emotion recognition from various kinds of signals including speech, face expressions, physiological signals like heart rate, electrodermal activity etc. have been explored in the past. As it can be noted that people emote through their language and the way they speak, it can be assumed that understood that emotion recognition can be possible by analysing speech signals. Hence, many works have worked on finding the best features to represent speech to apply machine learning techniques to classify various emotions. But, recently deep neural architectures have improved the performance and usability of emotion recognition systems.

In the traditional way researchers have come up with various techniques for emotion recognition from speech which included

4. Feature Extraction

4.1 Low - Level descriptors (Frame - level features)

Speech is a non-stationary signal as the frequencies and the way speech is produced changes rapidly. Hence, speech is analyzed at a frame-level where the whole speech is divided into frames of certain length and features for each frame are computed. In this work, we have computed various frame level features using the PRAAT software in

which the frame length was set as 25 ms and frame shift was set as 10 ms. Various features we computed are as follows:

Mel - Frequency Cepstral Coefficients (MFCC)
MFCC are traditional features used for representing speech and are used in various tasks such as speech recognition, speaker recognition, speech emotion recognition etc. as they capture the vocal tract information from a speech signal. We extracted 13 dimensional MFCC features from speech signal with 1 of them being the intensity (energy) of the signal.

Linear Prediction Cepstral Coefficients (LPCC)
LPCC are features which are generally used in voice coding tasks where voice compression is the main goal. Linear prediction technique allows speech signal to be separated into vocal tract filter and excitation source approximations where LPCC coefficients are used to represent the vocal tract filter characteristics. There have been only a few works or no works that have explored speech emotion recognition using LPCC. Hence, we have explored LPCC features for this task.

Residual - Mel Frequency Cepstral Coefficients (RMFCC)
As the features mentioned above, MFCC and LPCC capture information related to the vocal tract, the other key part of speech production is being underrepresented or not used. But, the excitation source signal contains information related to emotion recognition as mentioned in [][][]. Hence, various works have explored features that can be extracted from the residual signal which is an approximation of the excitation source signal computed using linear prediction analysis. One of such works is RMFCC, which was first introduced in [] and has been used for various speech related tasks such as [][][].

4.2 Utterance - level Descriptors

As categorical emotion labels for a speech signal are given at an utterance level, there are two ways of constructing speech emotion recognition systems. One is to develop features at an utterance level and other is to consider the label given to an utterance as the label for all the frames within it and to develop systems using all frames as training data. In the latter systems, testing can be done using some sort of majority voting or median voting strategies. But, the disadvantage of such systems is that temporal information is not taken into account and statistics are only captured at a classifier level instead of feature level. Hence, we use the former approach where we develop features for each utterance. Even in such kind of approach, one approach that has been followed in the past was gaussian mixture modeling (GMM) where the differences in gaussian mixture components between each utterances are analyzed and used for classifying between emotions. But the disadvantages of this approach included requirement of a huge dataset for initial modeling of speech for starting point of GMM called the universal background model (UBM) and the inability to capture the time-series information. Hence,

we have used long short term memory (LSTM) based deep-features for capturing information from the low-level descriptors described in section 4.1 and have also extracted a few other features that are mentioned in the literature for speech emotion recognition. These included LSTM autoencoder representation, LSTM categorical embeddings, jitter, shimmer, harmonics-to-noise ratio (HNR) and probability of voicing. Apart from these, we have used the opensmile toolkit to extract features in the paralinguistic challenge 2010 configuration as these features are used in various emotion recognition and paralinguistic works in the past. The details of the features and extraction procedure followed for utterance-level features are described below.

LSTM - Autoencoder Representation

To capture the information in an utterance as a single vector, we train a LSTM network to predict itself and collect the hidden representation at the end of the encoder to represent the speech signal. During this process, all the utterances are required to have a unique length as the LSTM network is fixed. Hence, after extracting the LLDs, the number of frames in each utterance was analyzed. A histogram plot of number of frames vs number of utterances can be found in fig. … Since, there was a peak around 250 frames, we have computed the accuracy of LSTM Autoencoder using 150, 200, 250 and 300 frames where training was performed on first 80% of the training set and validation on the last 20% and it has been observed that 200 frames gave slightly higher performance. Hence, all the utterances were either truncated or padded with zeros to contain 200 frames only. An LSTM encoder is developed in this way for each Low-level descriptor which we call MFCC_LSTM_Autoencoder, LPCC_LSTM_Autoencoder and RMFCC_LSTM_Autoencoder. The dimension of these representations is designed to be 256.

LSTM - categorical Embedding

The processing of the Low-level descriptors was performed in a similar fashion as mentioned in LSTM - autoencoder representation framework where 200 frames were chosen to represent each utterance. The structure of the network contained two-stacked LSTMs with 512 and 256 hidden units respectively followed by a fully connected layer to a hidden layer of 256 nodes with a ReLU activation function. This layer is considered as the representation layer and it is in turn connected to the output layer with softmax activation containing 4 nodes. A network is trained for each LLD and they are called MFCC_cat_embedding, LPCC_cat_embedding and RMFCC_cat_embedding.

5. Data Description

In this project, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database for training and testing our model. IEMOCAP stands for Interactive Emotional Dyadic Motion Capture database and has the following features:

1. It is an acted, multimodal and multi speaker database
2. It was recently collected at SAIL lab at USC

3. It contains roughly 12 hours of audiovisual data, including video, speech, motion capture of face and text transcriptions.
4. The data is captured through dyadic sessions where actors improvise on certain scenes or scripted scenarios which are especially designed to elicit emotional expressions in the dataset.
5. The database is annotated by multiple annotators into categorical labels, such as:
    a. Anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state.
6. The dataset also contains dimensional labels such as:
    a. Valence, Activation and Dominance values.
7. The dataset sessions are manually segmented into utterances.
8. Each utterance is annotated by at least 3 human annotators.

In our project, we only use the speech data from the database to train and test our system. We use the speech data annotated into categorical labels as well as dimensional labels at utterance level. For classification task, we only use the following 4 categorical labels namely Anger, Happiness, Sadness and Neutrality. We also used the dimensionals labels of Valence, Activation and Dominance to train our model in order to achieve better results.

6. Results

The performance of each feature representation is evaluated using Leave-One-Subject-Out cross validation approach where weighted recall and unweighted recall are computed. The results for the same can be observed in tab. …. In case of jitter, shimmer, HNR, probability of voicing and opensmile features, data from one participant is left out for testing and remaining data from 9 participants is used for training the classifier. In case of LSTM based deep features, both the network and the classifier are only trained on 9 participants leaving one subject out for testing. Hence, the network has been trained only for 50 epochs where the model with best results is captured.

Classifier	**SVM**					
Metrics	**Weighted (Recall / Precision / F1-score)**			**Unweighted (Recall / Precision / F1-score)**		
Jitter, Shimmer, HNR, UV ratio	31.50 (5.52)	31.50 (5.52)	31.50 (5.52)	33.47 (2.61)	34.43 (3.42)	27.60 (rbf) (3.56)

Opensmile (PC 2010)	52.70 (3.19)	52.70 (3.19)	52.70 (3.19)	54.35 (2.73)	52.79 (3.56)	51.86 (NB) (3.80)
MFCC_Autoencoder	43.40 (2.43)	43.40 (2.43)	43.40 (2.43)	40.32 (1.97)	47.46 (5.21)	37.72 (rbf) (1.56)
LPCC_Autoencoder	34.20 (5.06)	34.20 (5.06)	34.20 (5.06)	36.15 (3.48)	36.83 (7.65)	29.70 (NB) (3.22)
RMFCC_Autoencoder	42.95 (5.14)	42.95 (5.14)	42.95 (5.14)	38.44 (4.54)	40.18 (12.64)	33.68 (poly) (5.86)
MFCC_cat_embed	54.13 (2.63)	54.13 (2.63)	54.13 (2.63)	52.17 (2.31)	54.72 (2.71)	52.32 (rbf) (2.44)
LPCC_cat_embed	43.42 (5.38)	43.42 (5.38)	43.42 (5.38)	38.18 (4.18)	35.39 (5.84)	32.97 (poly) (4.44)
RMFCC_cat_embed	51.53 (2.06)	51.53 (2.06)	51.53 (2.06)	49.09 (2.32)	51.36 (3.73)	48.25 (rbf) (3.09)

7. Conclusions and Future Work

By looking at the results, we can conclude that we have found features(which also include deep features) for emotion recognition in speech. We can also say that the results of the deep features are comparable with that of the OpenSmile features, and the performance of the model can only increase in the future because there will not be sparse data, also, as deep Neural Networks need vast amount of data to interpret subtle features, we can say that the results will eventually improve over time.