Title – Speech Emotion Analysis

Abstract

Speech emotion recognition systems have high prediction latency because of the high computational requirements for deep learning models and low generalizability mainly because of the poor reliability of emotional measurements across multiple corpora. To solve these problems, we present a speech emotion recognition system based on a reductionist approach of decomposing and analyzing syllable-level features. Mel-spectrogram of an audio stream is decomposed into syllable-level components, which are then analyzed to extract statistical features. The proposed method uses formant attention, noise-gate filtering, and rolling normalization contexts to increase feature processing speed and tolerance to adversity. A set of syllable-level formant features is extracted and fed into a single hidden layer neural network that makes predictions for each syllable as opposed to the conventional approach of using a sophisticated deep learner to make sentence-wide predictions. The syllable level predictions help to achieve the real-time latency and lower the aggregated error in utterance level cross-corpus predictions. The experiments on TESS and RAVDESS (RA) databases show that the method archives real-time latency while predicting the sentiments.

Introduction

The project ‘Sentiment Analysis using Voice’ will help in identifying the fraud calls like scammers asking for bank details ,personal details of less educated people ,who believe these frauds easily and believe they are actually the officials from the agency ,and got scammed by them .The project can also be used in detecting negative emotions and harmful thoughts on more research and on including some additional features.We will start from already given datasets and test the models on it after the process ,testing on data after that will start and deployment of model using a appropriate interface . The project will come in handy when used on calls for real time analysis of call , for detecting the call is fraud or valid on time.It will start from basic after learning process , understanding sound and it’s features in machine from different sources , and how to handle sound data , how to preprocess it and how to build model for it.The final step involves deployment and testing of model on real voice data provided by one of our team members.

In a large proportion of these audio calls, people depict their opinions about products, movies, social issues, political issues, etc. The capability of detecting the sentiment of the speaker on calls can serve two basic functions: (i) it can enhance the retrieval of the particular data in question, thereby, increasing its utility, and (ii) the combined sentiment of a large number of audio on a similar topic can help in establishing the general sentiment. It is important to note that automatic sentiment detection using text is a mature area of research, and significant attention has been given to product reviews, we focus our attention on dual sentiment detection in calls based on audio and text analysis. We focus on audio calls because the nature of speech in these calls is more natural and spontaneous which makes automatic sentiment processing challenging.

Literature review

* From paper” A Study of Support Vector Machines for Emotional Speech Recognition” In this paper, efficiency comparison of Support Vector Machines (SVM) and Binary Support Vector Machines (BSVM) techniques in utterance-based emotion recognition is studied. Acoustic features including energy, Mel-Frequency Cepstral coefficients (MFCC), Perceptual Linear Predictive (PLP), Filter Bank (FBANK), pitch, their first and second derivatives are used as frame-based features.
* In paper “Audio and Text based multimodal sentiment analysis using features extracted from selective regions and deep neural networks” An improved multimodal approach to detect the sentiment of products based on their multimodality natures (audio and text) is proposed. The basic goal is to classify the input data as either positive or negative sentiment. Learning utterance-level representations for speech emotion and age/gender recognition. Accurately recognizing speaker emotion and age/gender from speech can provide better user experience for many spoken dialogue systems. In this study, we propose to use Deep Neural Networks (DNNs) to encode each utterance into a fixed-length vector by pooling the activations of the last hidden layer over time.
* The paper “Towards Real-time Speech Emotion Recognition using Deep Neural Networks” proposes a real-time SER system based on end-to-end deep learning. Namely, a Deep Neural Network (DNN) that recognizes emotions from a  
  one second frame of raw speech spectrograms is presented and investigated. This is achievable due to a deep hierarchical architecture, data augmentation, and sensible regularization. Promising results are reported on two databases which are the ENTERFACE database and the Surrey Audio-Visual Expressed Emotion (SAVEE) database.
* In paper“ Sentiment extraction from natural audio streams” a system for automatic sentiment detection in natural audio streams such as those found in YouTube is proposed. The proposed technique uses POS (part of speech) tagging and Maximum Entropy modelling (ME) to develop a text-based sentiment detection model. Additionally, we propose attuning technique which dramatically reduces the number of model parameters in ME while retaining classification capability. Finally, using decoded ASR (automatic speech recognition) transcripts and the ME sentiment model, the proposed system is able to estimate the sentiment in the YouTube video. In our experimental evaluation, we obtain encouraging classification accuracy given the challenging nature of the data. Our results show that it is possible to perform sentiment analysis on natural spontaneous speech data despite poor WER (word error rates).