Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 635 – 643

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining

Jasmine Bhaskar [a,*], Sruthi K[a], Prema Nedungadi [b]

[a]Department of Compute science , Amrita University, Amrita School of Engineering, Amritapuri, 690525, India
[b]Amrita Create, Amrita University, Amrita School of Engineering, Amritapuri, 690525, India

Abstract

One of the greatest challenges in speech technology is estimating the speaker's emotion. Most of the existing approaches concentrate either on audio or text features. In this work, we propose a novel approach for emotion classification of audio conversation based on both speech and text. The novelty in this approach is in the choice of features and the generation of a single feature vector for classification. Our main intention is to increase the accuracy of emotion classification of speech by considering both audio and text features. In this work we use standard methods such as Natural Language Processing, Support Vector Machines, WordNet Affect and SentiWordNet. The dataset for this work have been taken from Semval -2007 and eNTERFACE'05 EMOTION Database.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Emotion classification;text mining;speech mining;hybrid approach

1. Introduction

Emotion recognition plays a major role in making the human-machine interactions more natural. In spite of the different techniques to boost machine intelligence, machines are still not able to make out human emotions and expressions correctly. Emotion recognition automatically identifies the emotional state of the human from his or her speech. One of the greatest challenges in speech technology is evaluating the speaker's emotion. Usually emotion

* Corresponding author. Tel.: +91 9447597307;
E-mail address: jasmine@am.amrita.edu

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.112

recognition tasks focus on extracting features from audio. There are different types of temporal and spectral features that can be extracted from human speech. The features like statistics relating to the amplitude and pitch, formants of speech, Mel Frequency Cepstral Coefficients (MFCCs) etc are fed as inputs to the classification algorithms [23]. Speaker's emotion can also be detected using text mining technique on audio material after translating it into text.

Existing human-machine interaction systems can identify "what is said" and "who said it" using speaker identification and speech recognition techniques. These machines can evaluate "how it is said" to respond more correctly and make the interaction more natural, if provided with emotion recognition techniques. Emotion recognition is useful for applications such as Entertainment, e- Learning, and diagnostic tool for therapists, call centre applications etc.

Usually in emotion classification, researchers consider the acoustic features alone. Though features like pitch, energy and speaking rate change with emotional state, strong emotions such as anger and surprise have high pitch and energy. In that case, it is very difficult to distinguish the emotions such as anger and surprise using acoustic features alone. But, if we classify speech solely on its textual component, we will not obtain a clear picture of the emotional content.

In this hybrid approach, we intend to analyze both speech and the corresponding text component in order to detect the speaker's emotion. This method aims to enhance the efficiency of emotion classification by consolidating the features of both audio and text into a single feature vector which is then given to the classifier. The emotional states considered in this experiment are: Happy, Sad, Fear, Disgust, Surprise and Anger. Accordingly the classifier assigns each speaker to one of the above emotional states. Before applying the hybrid approach, both text and speech are handled separately and classified .This allows the comparison of these methods with our proposed hybrid approach.

In the past, some work is done on music mood classification based on lyrics and audio features[1, 2, 3]. The uniqueness of the proposed method is in the choice of the features considered and in the generation of a single feature vector for classification. We propose to use lexical resources like SentiWordNet and WordNet-Affect in order to generate the feature vector for text classification and multi-class SVM for emotional classification.

2. Related Works

This section discusses three aspects of emotion classification: emotion classification from audio, emotion classification from text and emotion classification using both text mining and speech mining

2.1. Emotion classification from the audio

Current research has highlighted new approaches to emotion classification from speech based on audio features. The work by Shen. et .al [21] recognize five different emotion states like disgust, boredom, sadness, neutral and happy using the features: pitch, energy, LPCC, MFCC and LPCMCC. They have explained and compared different combination of features. Their experiment was based on Berlin emotional database. They got different accuracy in different combination of features. In their work [16]Casale.et .al described the working in DSR environment. Features considered for this experiment were extracted by using ETSI ES 202 211 V.1.1.1 S standard front end. In their experiment, they have used two different speech corpora: EMO-DB in German and SUSAS in American English. Their result showed that Support Vector Machine (SVM) trained with Sequential Minimal Optimization (SMO) algorithm leads to better performance.

2.2. Emotion classification from the Text

Mishne[5] worked on classifying blog text according to the mood reported by its author during the writing. Mishne considered different textual features like frequency counts of words, emotional polarity of posts, length of posts, PMI, emphasized words and special symbols like emoticons and punctuation marks. PMI - Point wise Mutual Information provides a numerical weight for keywords based on its relation to a particular mood. SVM (Support Vector Machine) classifier was used in his work for classification. Text mining over transcribed audio recordings was performed in [7], in order to find the speakers emotion. The dataset (audio conversation of the customers) for this experiment was collected from a call centre. The researchers used different feature selection methods in this work. The unsupervised and supervised method further clarified text classification. The evidence demonstrated the