Development of a Real-time Embedded System for Speech Emotion Recognition

K Sekhar
Student,OUCE, OU, TS, India

ABSTRACT

Speech emotion recognition is one of the latest challenges in speech processing and Human Computer Interaction (HCI) in order to address the operational needs in real world applications. Besides human facial expressions, speech has proven to be one of the most promising modalities for automatic human emotion recognition. Speech is a spontaneous medium of perceiving emotions which provides in-depth information related to different cognitive states of a human being. In this context, we introduce a novel approach using a combination of prosody features (i.e. pitch, energy, Zero crossing rate), quality features (i.e. Formant Frequencies, Spectral features etc.), derived features ((i.e.) Mel-Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding Coefficients (LPCC)) and dynamic feature (Mel-Energy spectrum dynamic Coefficients (MEDC)) for robust automatic recognition of speaker's emotional states. Multilevel SVM classifier is used for identification of seven discrete emotional states namely angry, disgust, fear, happy, neutral, sad and surprise in 'Five native Assamese Languages'. The overall experimental results using MATLAB simulation revealed that the approach using combination of features achieved an average accuracy rate of 82.26% for speaker independent cases. Real time implementation of this algorithm is prepared on ARM CORTEX M3 board

Introduction

In the past decade, we have seen intensive progress of speech technology in the field of robotics, automation and human computer interface applications. It has helped to gain easy access to information retrieval (e.g. voice-automated call centers and voice search) and to access huge volumes of speech information (e.g. spoken document retrieval, speech understanding, and speech translation). In such frameworks, Automatic Speech Emotion Recognition (ASER) plays a major role, as speech is the fundamental mode of communication which tells about mental and psychological states of humans, associated with feelings, thoughts and behavior. ASER basically aims at automatic identification of different human emotions or physical states through a human's voice. Emotion recognition system has various applications in the fields of security, learning, medicine, entertainment, etc. It can act as a feedback system for real life applications in the field of robotics, where robot will follow human commands by understanding the emotional state of human. The successful recognition of emotions will open up new possibilities for development of an e-learning system with enhanced facilities in terms of student's interaction with machines. The idea can be incorporated in entertainment with the development of natural and interesting games with virtual reality experiences. It can also be used in the field of medicine for analysis and diagnosis of

cognitive state of a human being. With the advancement of the human-machine interaction technology, a user-friendly interface is becoming even more important for speech-oriented applications. The emotion in speech may be considered as similar kind of stress on all sound events across the speech. Emotional speech recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. With the advance of the human-machine interaction technology, a user- 1 friendly interface is becoming more and more important for speech-oriented applications.

Literature Review

In recent years, a great deal of research has been done to recognize human emotions using speech information. Researchers have combined new speech

processing technologies with different machine learning algorithms [1], [2] in order to achieve better results. In machine learning platform, speech emotion recognition belongs to supervised learning, following the generalized system model of data collection, feature exaction and classification. The extremely complex nature of human emotional states makes this problem more complicated in terms of feature selection and classification. Many Researchers have proposed important speech features which contain emotion information, such as prosody features [3] (pitch [4], energy, and intensity) and quality features [5], [6] like formant frequencies and spectra temporal features [7]. Along with these features, many state of-the-art derived features like MelFrequency Cepstral Coefficients (MFCC) [8], [9], Linear Predictive Coding have been suggested as very relevant features for emotion recognition. We have also considered some dynamic features like Mel-energy spectrum dynamic coefficients (MEDC) and have combined all the features to get a better result in emotion recognition. Many researches provide an indepth insight into the wide range of classification algorithms available, such as: Neural Networks (NN), Gaussian Mixture Model (GMM) [10], Hidden Markov Model (HMM) [11], Maximum Likelihood Bayesian Classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN) and Support Vector Machine (SVM) [12], [13], [14], [15]. We have chosen Support vector machine for our research work as it gives better results in emotion recognition domain of various databases like BDES (Berlin Database of Emotional Speech) and MESC (Mandarin Emotional Speech Corpora). Humans have been endowed by nature with the voice capability that allows them to interact and communicate with each other. Hence, the spoken language becomes one of the main attributes of humanity. Intensive progress of speech technology in the field of robotics, automation and Human Computer Interface (HCI) applications which is future of the world. Emotion recognition system has various applications in the fields of security, learning, medicine, entertainment, etc. A feedback system for real life applications can be developed in the field of robotics, where

robot will follow human commands by understanding the emotional state of humans. This research will open up new possibilities for development of an e-learning system with enhanced facilities in terms of student's interaction with machines. If incorporated

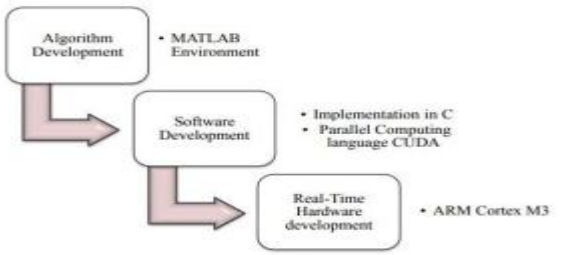

Fig. 1: Flow diagram of work done

entertainment world with the development of natural and interesting games the virtual world will become the real world for us. It can be used in the field of medicine for analysis and diagnosis of cognitive state of a human being. Microsoft and Google has been trying to implement speech interactive system but till today they are not completely successful with flaws in real time integration with an online database.

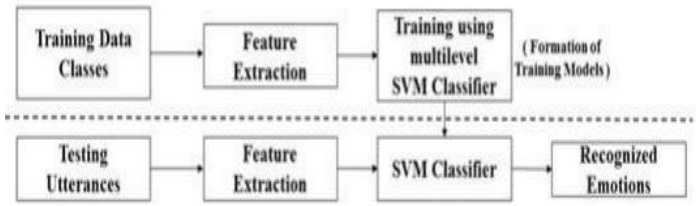

Fig. 2 Generalized system model for emotion recognition

Machine learning which concerns the development of algorithms, which allows machine to learn via inductive inference based on observation data that represent incomplete information about statistical phenomenon. Classification, also referred to as pattern recognition, is an important task in Machine Learning, by which machines "learn" to automatically recognize complex patterns, to distinguish between exemplars based on their different patterns, and to make intelligent decisions. A pattern classification task generally consists of three modules, i.e. data representation (feature extraction) module, feature selection or reduction module, and classification module. The first module aims to find invariant features that are able to best describe the differences in classes. The second module of feature selection and feature reduction is to reduce the dimensionality of the feature vectors for classification. The classification module finds the actual mapping between patterns and labels based on features. The objective of our work is to investigate the machine learning methods in the application of automatic recognition of emotional states from human speech.

Different Machine Learning Algorithms based on the input available at the time of training: Supervised learning algorithms are trained on labelled• examples, i.e., input where the desired output is known. The supervised algorithm attempts to generalize a function or 6 Fig. 2 Generalized system model for emotion recognition mapping from input to outputs which can then be used to speculatively generate an output for previous unseen inputs. Unsupervised learning algorithms operate on unlabeled• examples, i.e. input where the desired output is unknown. Here the objective is to discover structure in the data (e.g. through a cluster analysis), not to generalize a mapping from input to output. Semi-Supervised learning

combines both labeled and• unlabeled examples to generate an appropriate function or classifier. This problem belongs to the class of supervised learning of pattern recognition as we have to train the machine for particular classes with labelled data The speech samples which are going to be processed for emotion recognition should go through a preprocessing step that removes the noise and other irrelevant components of speech corpus for better perception of speech data. The preprocessing step involves three major steps such as pre -emphasis, framing and windowing. The pre-emphasis step is carried out on the speech signal using a Finite Impulse response (FIR) filter called pre-emphasis. The filter impulse response is given by The filtered speech signal is then divided into frames of 25ms with an overlap of 10ms. A hamming window is applied to each signal frame to reduce signal discontinuity and thus avoid 8 Fig. 3 Steps for feature extraction spectral leakage. Then speech emotion related features are extracted from the preprocessed speech data

Feature Extraction Different emotional states can be recognized using certain speech features which can be either prosody features or quality features. Some Prosody features which can be extracted directly; includes pitch, intensity and energy are the most widely used features in the emotion recognition domain. Though it is possible to distinguish some emotional states using only these features, but it becomes very inconvenient when it comes to emotional states with same level of stimulation [5]. The difficulty in distinguishing between joy and anger can be lowered by reflecting some quality features. Formants and Spectral energy distributions are the most important quality features to solve the classical problem of emotion recognition using speech. While prosody features are preferred on the arousal axis, quality features are favored on valence axis. Some other features which are derived from the basic acoustic features like MFCC and 7 coefficients from Linear

Predictive Coding are considered good for emotion recognition. Some dynamic features can be obtained from the variation of speech utterances in time domain by taking first order derivative of MFCC; named as MFDC. A method which combines all the above mentioned features is more promising than a method that uses only one type of features for the classification. Fig 3. Shows the overall model for feature extraction that has been used for both training of classifier and testing the unknown speech samples

Conclusions

The speech samples which are going to be processed for emotion recognition should go through a preprocessing step that removes the noise and other irrelevant components of speech corpus for better perception of speech data. The preprocessing step involves three major steps such as pre -emphasis, framing and windowing.

REFERENCES

[1] C. M. Lee, S. Member, S. S. Narayanan, and S. Member, “Toward Detecting Emotions in Spoken Dialogs,” vol. 13, no. 2, pp. 293–303, 2005.

[2] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech Commun., vol. 48, no. 9, pp. 1162– 1181, Sep. 2006.

ISSN: 2250-3676
Volume-5, Issue-3, 417-421

[3] I. Luengo and E. Navas, "Automatic Emotion Recognition using Prosodic Parameters" pp. 493–496, 2005.

[4] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," Int. J. Speech Technol., vol. 16, no. 2, pp. 143–160, Aug. 2012.