Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 49 (2015) 24 – 32

Procedia Computer Science

ICAC3'15

Recognition of Human Emotions from Speech Processing

V. V. Nanavare[a], S. K. Jagtap[b]

[a]PG Student,Department of E&TC, SKN College of Engineering,Pune-41,India.
[b]Associate Prof., Department of Electronics & Telecommunications Engg, SKN College of Engineering,Pune-41,India

Abstract

Human beings are communicating with each other by expressive gestures of emotions and feelings which are identified by experience and knowledge. These expressions may be conveyed in speech form or through body language. Emotions are part and parcel of human life and among other things, highly influence decision making. In this paper the kinds of features that might carry more information about the emotional meaning of each utterance are considered. The features that contribute to emotions may be different for different spoken languages. The approach is to calculate which features carry more information and to combine these features to get a better recognition rate. It also depends on which emotions we want a machine to recognize and its purpose. Active learning tries to select the most informative examples to build a training set for a predictive model. In this paper, we used the hidden Markov model to model the phonetic units corresponding to sentences taken from the training base. The results obtained are very encouraging given the size of the training set and the number of people taken to the registration. This algorithm is based on the flexibility of the hidden Markov model for sentences by means of dynamic programming.

Peer-review under responsibility of organizing committee of the 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)

Keywords: Hidden Marko Model(HMM),MFCC,DTW.

1. Introduction

Recognizing human emotion by computer has been an active research area in the past a few years. An efficient human emotion recognition system will help to make the interaction between human and computer more natural

* Corresponding author. Tel. +919421027499; fax: +91-20-24354938.
E-mail address: skjagtap.skncoe@sinhgad.edu (S. K. Jagtap)

Peer-review under responsibility of organizing committee of the 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15)
doi:10.1016/j.procs.2015.04.223

and friendly [1]. Recently, gesture of emotions in speech to communicate with the machines is an upcoming challenge. The emotions can be observed through the variations in several prosodic parameters of a natural language. Acoustic correlates of emotional speech are often listed in terms of features such as utterance intensity, f0 contour, and voice quality as well as timing and speech rate. It is well established that voice quality correlates with emotion [2]. Emotional conversion plays an important role in many applications. The direct uses include human machine interaction for any lay man literate or illiterate and especially visually impaired persons, entertainment like computer games, and business like call centers etc. For indirect uses, the conversion of emotional features in speech can improve the performance of speech recognition systems, artificial intelligence systems, etc. Emotion conversion is a technique where the parameters of the input speech emotion are analyzed and manipulated to the target emotion and then the final output is resynthesized using the new parameters. To avoid building a separate voice for each required emotion, a transformation can be applied to modify the acoustic parameters of neutral speech such that the modified utterance conveys the desired target emotion [3]. In recent years, data-driven voice conversion methods have been explored for modeling and transforming both short-term spectra and prosody. In [4], GMM based spectral conversion techniques were applied to emotion conversion but it was found that spectral transformation alone is not sufficient for conveying the required target emotion. In [5], an emotion conversion system for English was described which is independent of the underlying synthesis system.
In the previous studies three main rules were commonly used, FO conversion, duration conversion, and spectral conversion [2].

In this paper, we study the audiovisual recognition of human emotion regardless of the subject's cultural background, language, and race. The audio features including prosodic, MFCC, and formant frequency features are extracted from the speech to map the emotional speech to the corresponding feature space [1]. This classification is used as a basis for the comparison throughout this work also expecting further comparisons. Most approaches in nowadays speech emotion recognition use global statistics of a phrase as basis. However; also first efforts in recognition of instantaneous features exist. We present two working engines using both alluded alternatives by use of continuous hidden Markov models, which have evolved as a far spread standard technique in speech processing [6].

2. LITERATURE SURVEY

There are a lot of approaches to speech recognition. Algorithms and feature extraction are based on the acoustic-phonetic approach. Algorithms such as template matching come under the pattern recognition approach, while algorithms that depend on knowledge sources, stochastic of speech signals and neural networks are based on the artificial intelligence approach. However, an important approach to speech recognition is stochastic modelling, in particular stochastic modelling using hidden markov models. Among these the most popular and accurate algorithm is the template based dynamic time warping [7]. Speech recognition process includes front end processing part that converts a speech signal into features useful for further faithful processing. Feature extraction stage plays important role in obtaining the important features from the speech signal which are comparatively insensitive to talker and channel variability. The features obtained after this crucial step decreases the data rate in the later part of the speech recognition and reduces the redundancy residing in the speech signal. Wide range of feature extraction techniques is dependent on the standard signal processing techniques, such as linear predictive coding, filter banks, or cepstral techniques. Some novel methods are also involved which are based on the human perception of the speech signal [2]. Feature extraction based upon auditory model system has shown the better performances than conventional signal processing schemes.

The process of the speech recognition is highly affected by the surrounding in which the system resides. Noise is the major obstacle which contaminates the maximum possible response of the system. Feature extraction plays significant role in the optimality of the system response. Fig. 1 gives the schematic representation of the speech recognition system. In speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. The feature extraction is usually performed in three stages. The first stage is called the speech analysis or the acoustic front end. It performs some kind of spectro temporal analysis of the signal and generates raw features describing the envelope of the power