Applied Acoustics 115 (2017) 6–14

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A bio-inspired emotion recognition system under real-life conditions

Farah Chenchah *, Zied Lachiri

LR-SITI Laboratory, National Engineering School of Tunis, Tunis, Tunisia

ARTICLE INFO

Article history:
Received 22 October 2015
Received in revised form 22 January 2016
Accepted 21 June 2016

Keywords:
Emotion recognition
HMM
PNCC
Noisy environment

ABSTRACT

This paper pulls together the advances of recognizing emotion theory with advances in speech feature in order to improve understand of emotion under real life condition. It presents the application of a recently proposed feature extraction method based on spectral features, used for speech emotion recognition purposes. Specifically, the performance of the proposed approach is evaluated on real condition speech signal (IEMOCAP database) with real world noise using various SNR levels. We examined an assessment of emotion error rate using classical descriptors (MFCC, PLP) and also new type of speech features considered as more robust to noise and reverberation distortions (PNCC with different variants). The results reveal that the used methods give better performance using MFCC under clean environment, and that PNCC shows an advantage compared to other features methods in noisy environment.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Understanding emotional states is an essential vector of communication that has recently interested researchers. Speech is a time-varying signal conveying multiple layers of information (words, speaker, language, ...). Speech signal provides an extensive array of information for conveying emotion. However it is very interesting to study the impact of emotion detection in a noisy environment.

Machine understanding of emotion in a real world condition is a challenge that may significantly fulfill the dream of creating a humanoid robot. This field is undoubtedly a multi-disciplinary subject involving psychology, cognitive science and pattern recognition [1]. But it is speech recognition as a field of signal processing that it must be accepted as a central concern when it goes to recognize emotional states [2]. Speech used for this purpose is over a wider variety of conditions, which may introduce distortions in the signal. It may be corrupted by a wide variety of noises, such as sounds from various devices in the vicinity.

An emotion is a mental and physiological state associated with a wide variety of feelings, thoughts, and behavior. Affection plays an important role in our everyday lives. It occurs in every relationship we care about—in the workplace, in our friendships, in dealings with family members, and in our most intimate relationships [3]. Information about emotions resides at multiple time scales, through multiple cues. Human-robot interaction technology is an extensive field and recognizing human emotion is one of the basic techniques. Emotion recognition technology is very popular because it adds to the appeal of electronic systems by contributing to the user's perception of the system's intelligence and adaptability [4].

Speech is a natural mode of communication among human beings. It carries multi-dimensional information such as intended message, speaking characteristics of a person, spoken language identity, speaker mood and emotions. The same textual message would be conveyed with different semantics by incorporating appropriate emotions. Humans understand the intended message by perceiving the underlying emotions in addition to phonetic information. Therefore, there is a need to develop a speech emotion recognition system that can recognize affective state independently from linguistic message.

The importance of recognizing emotion from speech is proved by an increasing number of related works that can be found in literature [5–11] but most of the existing speech emotion recognition systems are processed under ideal acoustic conditions. Therefore, the impact of noise on the classification performance is still not well studied [12]. The innovation of this study consists in recognizing affective states from speech signals under acoustic distortions caused by real noises. Detecting emotions under disordered influences is still a very challenging task. Even if noise is an important factor affecting the performance of most recognition systems, it was rarely studied in speech emotion recognition systems. Schuller et al. [13] was pioneer in studying noise problem in automatic emotion recognition by adding white noise to two well-known

* Corresponding author.
E-mail addresses: farahchenchah@yahoo.fr (F. Chenchah), zied.lachiri@enit.rnu.tn (Z. Lachiri).

http://dx.doi.org/10.1016/j.apacoust.2016.06.020
0003-682X/© 2016 Elsevier Ltd. All rights reserved.

public databases and was showing effects of noise conditions in recognition results. Improvement consisting in noise cancellation based on the adaptive threshold in wavelet domain has been proposed by Tawari and Trivedi [14]. However, several improvements have to be explored in this field, Huang et al. [15] applied two existing speech enhancement algorithm to deal with additive white Gaussian noise to better recognize speakers' emotional states. Georgogiannis and Digalakis [16] implement speech emotion recognition system capable of working in noisy environments, using non-linear Teager energy based features extracted from voice.

Gharavian et al. [17] extract formant, pitch, energy and MFCC features, then they used fast correlation based feature selection to reduce feature vectors dimension, finally they implement fuzzy ARTMAP neural network to classify emotions states. Oflazoglu and Yildirim [18] presents classic acoustic features based on low-level descriptors and statistics features using Support Vector machines (SVM) to recognize emotion from Turkish speech. Yuncu et al. [19] implement auditory model as descriptors and use statistical features such mean and standard deviations, classification was carried out by binary SVM decision tree, they use three databases: Berlin, Polish and Savee.

Hence, the main purpose of this article is to describe the implementation of the recognition of emotion from audio signal in more realistic settings. Thus we developed a system assessing the use of prosodic features such as MFCC and PLP and perform a novel kind of features PNCC with different configurations, under clean and noisy environment.

This paper is structured as follows: After introducing our emotion recognition approach in Section 2, we present the emotional database used in this work. The proposed system is described in Section 3. This is followed by a presentation of the used feature extraction methods in Section 4. Section 5 deals with the classification method. In Section 6 we present experimental set up and output results. Finally we conclude this paper with a perspective analysis of possible work in Section 7.

2. Emotional database

An important issue to be considered to implement emotion recognition system is the quality of the databases used to assess the performance of the systems. Speech corpora used for developing the present emotional speech systems is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database collected at SAIL lab at USC [20]. This database contains audio, video and motion-capture recordings of dyadic mixed-gender pairs of actors (approximately a total of twelve hours). It is split into five sessions that last approximately five minutes, each session consist in a dialogue, actors play improvisation of scripts or hypothetical scenarios. Each improvisation was designed to convey a general emotional theme; for instance sadness is conveyed through talking about a death of close relative, happiness is obtained when discussing a recent marriage. Scripts are extracted from theatrical plays and represent a more complex emotional flow. The aim is to obtain a presentation that mostly resembles to natural emotion expression.

After being divided into utterance, utterance was manually annotated in categorical labels: {angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, other} and in terms of dimensional axes: {valence, activation, dominance}.

Our speech emotion recognition system is performed using only four emotions which are: {anger, happiness, sadness and neutral}. Table 1 shows the distribution of the corpus over the four classes of emotions used in this analysis.

Table 1
IEMOCAP corpus utterances for the four classes of emotions.

	Angry	Happy	Sad	Neutral	Total
Male	320	150	338	335	1143
Female	311	173	320	271	1075
Total	631	323	658	606	2218

3. System description

The advances in this field are mainly based on one or more three steps: The first step consists in implementing an adaptive process of enhancement in training set based on testing set. The second step is based on ameliorating the feature vector in order to obtain better understanding of the characteristics of emotional speech. The third step used is based on trying to optimize the system results by using better adapted classifiers. This paper focuses essentially on the second method by trying to find feature adapted to emotion.

The framework of the proposed emotion recognition system from voice is illustrated in Fig. 1. The architecture of the system is based on two main levels: feature extraction and emotion classification. The objective of the step of feature extraction is to characterize the information by extracting the most relevant properties that are characteristics for emotion, and represent them in feature vector. The machine learning model has inputs features with known emotional labels of training set. The classifier obtained through this model aims to distinguish changes between emotional classes.

4. Feature extraction

Feature extraction models the speech signal in a sequence of feature vectors with a compact representation. To develop an automatic emotion recognition system, feature extraction is one of the key dimensions of design [21]. Various feature extraction techniques are used to extract the relevant characteristics of the speech signal and retain useful information; four different front-end methods are performed in this paper. This section deals with two most popular sets of front-end methods, MFCC coefficients [22] and PLP coefficients [23] and performs a third technique called Power Normalized Cepstral Coefficients with their different variants [24,25].

4.1. MFCC coefficients

Mel Frequency Cepstral Coefficients (MFCC) is a popular technique widely used for speech recognition as well as emotion recognition. MFCC is based on the known variation of the human ear's critical frequency bandwidth which uses a non-linear Mel scale frequency to approximate the behavior of the human auditory system. These coefficients are the results of a discrete cosine transform of output log energies of triangular filters bank, linearly spaced on the Mel frequency scale. The detailed procedure of computing MFCC is shown in Fig. 2: The input speech signal is segmented into time frames then hamming window is applied to each frame to eliminate discontinuities at the edges. Spectral coefficients are then calculated using FFT for each frame to extract frequency components, then passed through set of triangular filters spaced according to Mel scale. This scale is approximately linear up to 1 kHz, and logarithmic at greater frequencies. Logarithm of spectral amplitude is then taken and the output so obtained is finally converted back to time domain. Since the Mel spectrum coefficients are real numbers, we converted them using the Discrete Cosine Transform (DCT). Typically, only the first 13 cepstral

Download English Version:

https://daneshyari.com/en/article/754215

Download Persian Version:

https://daneshyari.com/article/754215

Daneshyari.com