Expert Systems with Applications 41 (2014) 5858–5869

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A new approach of audio emotion recognition

Chien Shing Ooi [a,*], Kah Phooi Seng [b], Li-Minn Ang [b], Li Wern Chew [c]

[a] Department of Computer Science & Networked System, Sunway University, 46150 Petaling Jaya, Malaysia
[b] School of Engineering, Edith Cowan University, WA 6027, Australia
[c] Intel Microelectronics (M) Sdn. Bhd., 11900 Pulau Pinang, Malaysia

ARTICLE INFO

Keywords:
Audio emotion recognition
RBF neural network
Prosodic features
MFCC feature

ABSTRACT

A new architecture of intelligent audio emotion recognition is proposed in this paper. It fully utilizes both prosodic and spectral features in its design. It has two main paths in parallel and can recognize 6 emotions. Path 1 is designed based on intensive analysis of different prosodic features. Significant prosodic features are identified to differentiate emotions. Path 2 is designed based on research analysis on spectral features. Extraction of Mel-Frequency Cepstral Coefficient (MFCC) feature is then followed by Bi-directional Principle Component Analysis (BDPCA), Linear Discriminant Analysis (LDA) and Radial Basis Function (RBF) neural classification. This path has 3 parallel BDPCA + LDA + RBF sub-paths structure and each handles two emotions. Fusion modules are also proposed for weights assignment and decision making. The performance of the proposed architecture is evaluated on eNTERFACE'05 and RML databases. Simulation results and comparison have revealed good performance of the proposed recognizer.

1. Introduction

Speech signals can rapidly deliver information or messages by human. Audio emotion recognition is a way to identify the emotional state of human from these speech signals. It is very useful for many applications such as safety in automotive (Nass et al., 2005), diagnosis tool (Edwards, Jackson, & Pattison, 2002), customer satisfaction assessments in call centers (Petrushin, 1999), etc.

In previous research, audio emotion recognition has been studied on different aspects. One of the aspects is the investigation on the emotion representation of audio features. Audio features such as pitch (Busso, Lee, & Narayanan, 2009; Devillers, Vidrascu, & Layachi, 2010), log energy, zero crossing rate (Chien Hung, Ping Tsung, & Chen, 2010; Chih-Chang, Chien-Hung, Ping-Tsung, & Chen, 2010), spectral features (Wong & Sridharan, 2001), voice quality (Lugger & Bin, 2007), jitter(Xi et al., 2007), etc. have been discovered useful in emotion recognition. However, it is insufficient to classify emotions correctly with only single type of audio features due to similarities may in certain emotions. Another aspect of the related research is based on the classification techniques. Few efforts have been reported that different types of classifiers such as Support Vector Machine (SVM) (Hu, Xu, & Wu, 2007; Morrison, Wang, De Silva, & Xu, 2005), neural network (Bhatti, Yongjin, & Ling, 2004; Bulut, Lee, & Narayanan, 2008; Khanchandani & Hussain, 2009; Nicholson, Takahashi, & Nakatsu, 2000; Petrushin, 1999) and Hidden Markov Model (HMM) (Bhaykar, Yadav, & Rao, 2013; Kammoun & Ellouze, 2006; Tin Lay, Say Wei, & De Silva, 2003; Yi-Lin & Gang, 2005; Zeng, Tu, Pianfetti, & Huang, 2008) are integrated in their systems. However, accuracy of classification is rather low especially when more than two number of emotion enclosed in their systems.

Recent trends in research of audio emotion recognition emphasized the use of combination of different features to achieve improvement in the recognition performance. System and prosodic features represent mostly mutually exclusive information of the speech signal. Therefore, these features are complementary in nature to each other. Combination of complementary features is expected to improve the intended performance of the system. For instance, researcher Yeh, Pao, Lin, Tsai, and Chen (2011) developed a system to recognize 5 emotions recently using up to 128 audio features. Their design used their exclusive segmentation method and feature selections method to recognize emotions every significant portion of the input signal. However, their system is not language-independent. Only Mandarin utterances were considered in their system. Another recent effort that utilizes combination of complementary features is reported by Lee, Mower, Busso, Lee, and Narayanan (2011). They used two databases, IEMOCAP (Busso et al., 2009) and AIBO to build a model consists of multiple layers of binary classifications. There are 384 audio features

* Corresponding author. Tel.: +60 125064977.
E-mail addresses: ocshing@gmail.com, 11057221@imail.sunway.edu.my (C.S. Ooi).

http://dx.doi.org/10.1016/j.eswa.2014.03.026
0957-4174/

(inclusive several statistical coefficients) extracted in their system such as zero crossing rate, root-mean-square energy, voice quality, pitch, MFCC. SVM classifier was used in each layer to distinguish two different emotions. This approach is language-independent and able to boost the accuracy due to binary classification. However, less than five emotions can be recognized. Wu and Liang (2011) also designed an architecture to extract various prosodic features such as pitch, duration, intensity, formants, and MFCC on affective speech based on semantic labels, and classified using Gaussian Mixture Models (GMM), SVM and neural network. Despite good recognition rate obtained using multi-layer classification scheme in their research, only 4 emotions (neutral, happy, angry and sad) were considered and own non-standard databases were used in performance evaluation.

In this paper, a new architecture of audio emotion recognition is presented. Different prosodic and spectral features are analyzed and useful features are identified to assist the design of this architecture. The universal six emotions such as Happy, Angry, Sad, Disgust, Surprise and Fear are considered in this paper. The proposed architecture has two main paths. The first path has an Audio Features Analyzer to extract different audio features and an audio feature-level fusion module. The Audio Features Analyzer is designed based on intensive research analyses on prosodic audio features. Prosodic audio features such as pitch, log-energy, zero-crossing rate (ZCR) and Teager Energy Operator (TEO) are found to be useful. On the other hand, the second path is designed after useful spectral audio features are identified. This path consists of MFCC features extraction followed by three parallel sub-paths for three sets of emotion groups. They are Emotion Group 1 (Angry and Happy), Emotion Group 2 (Sad and Disgust) and Emotion Group 3 (Surprise and Fear). An audio decision-level fusion is also proposed to fuse the information from both audio paths 1 and 2. A weight assignment mechanism is also designed. A decision making mechanism is also included in the fusion module to decide the final emotion.

This paper is organized as follow: a brief review of speech features, classification techniques, and databases are given in Section 2. Section 3 provides the details about the proposed architecture of audio emotion recognition. Experimental to evaluate the performance of the proposed system and results are presented in Section 4. Finally, the conclusion is presented in Section 5.

2. Brief review

This section provides a brief review on some important speech features and processing techniques, classification techniques and widely used databases for speech emotion recognition.

2.1. Speech features

Different speech features represent different speech information, e.g. emotion, speaker, in highly overlapped manner. These have motivated intensive research of audio emotion recognition in discovering the significant manner of the speech features on specific emotions. Speech features can be classified into 3 groups: vocal tract system, prosodic, and excitation source features.

Vocal Tract System features usually can be extracted from a short segment of speech signals. This kind of features represents the distribution of energy of a range of speech frequency. There were also some research works based on various vocal tract features and their combination. For instance, another vocal tract feature called Log Frequency Power Coefficients (LFPC) has been used by Tsang-Long, Yu-Te, Jun-Heng, and Pei-Jia (2006) along with Perceptual Linear Prediction (PLP), MFCC, and LPCC to recognize emotions such as Angry, Happy, Sad, Bored and Neutral. Highest recognition rate obtained from their experiments is 84.2% on their own Mandarin database. Linear prediction cepstral coefficients (LPCC) and MFCC which are two popular spectral features were also used by Nwe, Foo, and De Silva (2003a, 2003b) to classify the universal six emotions. Using their own database called Burmese-Mandarin corpus, LPCC and MFCC, respectively provided 56.1% and 59% of average classification rate in their experiment. In one of the recent literatures, Krishna Kishore and Krishna Satish (2013) used Sub-band based Cepstral Parameter (SBC) and MFCC to recognize six emotions (i.e. Angry, Fear, Happy, Sad, Disgust and Neutral) on SAVEE database. Their best achieved result is 79% of recognition rate. Another recent effort from Bhaykar et al. (2013) experimented the performance of MFCC feature alone on speaker dependent and speaker independent situation. With IITKGP-SESC (Koolagudi, Maity, Kumar, Chakrabarti, & Rao, 2009) and IITKGP-SEHSC (Koolagudi, Reddy, Yadav, & Rao, 2011) databases, Angry, Disgust, Fear, Happy, Neutral, Sarcastic, and Surprise are the seven emotions used in their effort. The reported results showed that although speaker dependent case could score 89.20%, speaker independent case only obtained 48.18% of recognition rate.

Among the prosodic features, pitch (or fundamental frequency) information is the most widely used for determining emotions (Busso et al., 2009; Devillers et al., 2010). It can well discriminate emotions compared to other features. Besides pitch, it was also reported (Kammoun & Ellouze, 2006) that log energy is also one of the most considered parameters of to evaluate speaking styles and emotions. Experiments using log-energy feature in Kammoun and Ellouze (2006) was reported that Angry emotion can be distinguished from Fast, Lombard, Question, Slow and Soft emotions using SUSAS database (Hansen & Bou-Ghazale, 1997). Another prosodic feature, zero crossing rate (ZCR) of speech signal was also a good parameter that related to emotions in the previous research (Chien Hung et al., 2010; Chih-Chang et al., 2010), stated that angry emotion has the higher mean value than happy emotion due to the higher frequency of vibration present in speech signals. Formants also could represent emotions based on the previous research, especially on the first and second position (Goudbeek, Goldman, & Scherer, 2009). For instance, it was reported by Pribil and Pribilova (2012) that Angry has highest value in formant frequency compared to Joy and Sad, while Sad has the lowest value. In literature, very few attempts (Cummings & Clements, 1995; Ling, Hu, & Wang, 2005) have been made to explore the excitation source information for developing any of the speech systems. Thus excitation source features are not reviewed here and they are not considered in our research.

2.2. Classification

Different classification methods have been developed for speech-related application such as speech recognition, emotion classification, speaker verification, etc. The classification methods used in audio emotion recognition typically can be divided into linear and nonlinear classifications. Linear classification performs the classification by making a decision based on weighted linear combination of the object characteristics, while non-linear classification is based on non-linear weighted combination of object characteristics. Non-linear classifiers are more widely used and effective in classifying the overlapped emotional characteristics of different emotions.

One of the most popularly classification methods for audio emotion recognition is HMM (Bhaykar et al., 2013; Kammoun & Ellouze, 2006; Tin Lay et al., 2003; Yi-Lin & Gang, 2005; Zeng et al., 2008). It is based on probability algorithm to model sequential data. Neural Network has also been widely applied in audio emotion recognition. It can be divided into 3 categories which