Pattern Recognition 44 (2011) 572–587

Contents lists available at ScienceDirect

Pattern Recognition

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

Survey on speech emotion recognition: Features, classification schemes, and databases

Moataz El Ayadi [a,*], Mohamed S. Kamel [b], Fakhri Karray [b]

[a] Engineering Mathematics and Physics, Cairo University, Giza 12613, Egypt
[b] Electrical and Computer Engineering, University of Waterloo, 200 University Avenue W., Waterloo, Ontario, Canada N2L 1V9

ARTICLE INFO

Article history:
Received 4 February 2009
Received in revised form
25 July 2010
Accepted 1 September 2010

Keywords:
Archetypal emotions
Speech emotion recognition
Statistical classifiers
Dimensionality reduction techniques
Emotional speech databases

ABSTRACT

Recently, increasing attention has been directed to the study of the emotional content of speech signals, and hence, many systems have been proposed to identify the emotional content of a spoken utterance. This paper is a survey of speech emotion classification addressing three important aspects of the design of a speech emotion recognition system. The first one is the choice of suitable features for speech representation. The second issue is the design of an appropriate classification scheme and the third issue is the proper preparation of an emotional speech database for evaluating system performance. Conclusions about the performance and limitations of current speech emotion recognition systems are discussed in the last section of this survey. This section also suggests possible ways of improving speech emotion recognition systems.

1. Introduction

The speech signal is the fastest and the most natural method of communication between humans. This fact has motivated researchers to think of speech as a fast and efficient method of interaction between human and machine. However, this requires that the machine should have the sufficient intelligence to *recognize* human voices. Since the late fifties, there has been tremendous research on speech recognition, which refers to the process of converting the human speech into a sequence of words. However, despite the great progress made in speech recognition, we are still far from having a *natural* interaction between man and machine because the machine does not understand the *emotional* state of the speaker. This has introduced a relatively recent research field, namely speech emotion recognition, which is defined as extracting the emotional state of a speaker from his or her speech. It is believed that speech emotion recognition can be used to extract useful semantics from speech, and hence, improves the performance of speech recognition systems [93].

Speech emotion recognition is particularly useful for applications which require natural man–machine interaction such as web movies and computer tutorial applications where the response of those systems to the user depends on the detected emotion [116]. It is also useful for in-car board system where information of the mental state of the driver may be provided to the system to initiate his/her safety [116]. It can be also employed as a diagnostic tool for therapists [41]. It may be also useful in automatic translation systems in which the emotional state of the speaker plays an important role in communication between parties. In aircraft cockpits, it has been found that speech recognition systems trained to stressed-speech achieve better performance than those trained by normal speech [49]. Speech emotion recognition has also been used in call center applications and mobile communication [86]. The main objective of employing speech emotion recognition is to adapt the system response upon detecting frustration or annoyance in the speaker's voice.

The task of speech emotion recognition is very challenging for the following reasons. First, it is not clear which speech features are most powerful in distinguishing between emotions. The acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates adds another obstacle because these properties directly affect most of the common extracted speech features such as pitch, and energy contours [7]. Moreover, there may be more than one perceived emotion in the same utterance; each emotion corresponds to a different portion of the spoken utterance. In addition, it is very difficult to determine the boundaries between these portions. Another challenging issue is that how a certain emotion is expressed generally depends on the speaker, his or her culture and environment. Most work has focused on monolingual emotion classification, making an assumption there is no cultural difference among speakers. However, the task of multi-lingual classification has been investigated [53]. Another problem is that one may undergo a certain emotional state such as sadness for days, weeks, or even months. In such a case, other emotions will be transient and will not last for more than a few minutes. As a consequence, it is not clear which emotion

[*] Corresponding author.
E-mail address: moataz@pami.uwaterloo.ca (M. El Ayadi).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.09.020

the automatic emotion recognizer will detect: the long-term emotion or the transient one. Emotion does not have a commonly agreed theoretical definition [62]. However, people know emotions when they feel them. For this reason, researchers were able to study and define different aspects of emotions. It is widely thought that emotion can be characterized in two dimensions: activation and valence [40]. Activation refers to the amount of energy required to express a certain emotion. According to some physiological studies made by Williams and Stevens [136] of the emotion production mechanism, it has been found that the sympathetic nervous system is aroused with the emotions of Joy, Anger, and Fear. This induces an increased heart rate, higher blood pressure, changes in depth of respiratory movements, greater sub-glottal pressure, dryness of the mouth, and occasional muscle tremor. The resulting speech is correspondingly loud, fast and enunciated with strong high-frequency energy, a higher average pitch, and wider pitch range. On the other hand, with the arousal of the parasympathetic nervous system, as with sadness, heart rate and blood pressure decrease and salivation increases, producing speech that is slow, low-pitched, and with little high-frequency energy. Thus, acoustic features such as the pitch, timing, voice quality, and articulation of the speech signal highly correlate with the underlying emotion [20]. However, emotions cannot be distinguished using only activation. For example, both the anger and the happiness emotions correspond to high activation but they convey different affect. This difference is characterized by the valence dimension. Unfortunately, there is no agreement within researchers on how, or even if, acoustic features correlate with this dimension [79]. Therefore, while classification between high-activation (also called high-arousal) emotions and low-activation emotions can be achieved at high accuracies, classification between different emotions is still challenging.

An important issue in speech emotion recognition is the need to determine a set of the important emotions to be classified by an automatic emotion recognizer. Linguists have defined inventories of the emotional states, most encountered in our lives. A typical set is given by Schubiger [111] and O'Connor and Arnold [95], which contains 300 emotional states. However, classifying such a large number of emotions is very difficult. Many researchers agree with the 'palette theory', which states that any emotion can be decomposed into *primary* emotions similar to the way that any color is a combination of some basic colors. Primary emotions are Anger, Disgust, Fear, Joy, Sadness, and Surprise [29]. These emotions are the most obvious and distinct emotions in our life. They are called the *archetypal* emotions [29].

In this paper, we present a comprehensive review of speech emotion recognition systems targeting pattern recognition researchers who do not necessarily have a deep background in speech analysis. We survey three important aspects in speech emotion recognition: (1) important design criteria of emotional speech corpora, (2) the impact of speech features on the classification performance of speech emotion recognition, and (3) classification systems employed in speech emotion recognition. Though there are many reviews on speech emotion recognition such as [129,5,12], our survey is more comprehensive in surveying the speech features and the classification techniques used in speech emotion recognition. We surveyed different types of features and considered the benefits of combining the available acoustic information with other sources of information such as linguistic, discourse, and video information. We theoretically covered, in some detail different classification techniques commonly used in speech emotion recognition. We also included numerous speech recognition systems implemented in other research papers in order to have an insight on the performance of existing speech emotion recognizers. However, the reader should interpret the recognition rates of those systems carefully since different emotional speech corpora and experimental setups were used with each of them.

The paper is divided into five sections. In Section 2, important issues in the design of an emotional speech database are discussed. Section 3 reviews in detail speech feature extraction methods. Classification techniques applied in speech emotion recognition are addressed in Section 4. Finally, important conclusions are drawn in Section 5.

2. Emotional speech databases

An important issue to be considered in the evaluation of an emotional speech recognizer is the degree of naturalness of the database used to assess its performance. Incorrect conclusions may be established if a low-quality database is used. Moreover, the design of the database is critically important to the classification task being considered. For example, the emotions being classified may be infant-directed; e.g. soothing and prohibition [15,120], or adult-directed; e.g. joy and anger [22,38]. In other databases, the classification task is to detect stress in speech [140]. The classification task is also defined by the number and type of emotions included in the database. This section is divided into three subsections. In Section 2.1, different criteria used to evaluate the goodness of an emotional speech database are discussed. In Section 2.2, a brief overview of some of the available databases is given. Finally, limitations of the emotional speech databases are addressed in Section 2.3.

2.1. Design criteria

There should be some criteria that can be used to judge how well a certain emotional database simulates a real-world environment. According to some studies [69,22], the following are the most relevant factors to be considered:

Real-world emotions or acted ones?: It is more realistic to use speech data that are collected from real life situations. A famous example is the recordings of the radio news broadcast of major events such as the crash of Hindenburg [22]. Such recordings contain utterances with very natural conveyed emotions. Unfortunately, there may be some legal and moral issues that prohibit the use of them for research purposes. Alternatively, emotional sentences can be elicited in sound laboratories as in the majority of the existing databases. It has always been criticized that acted emotions are not the same as real ones. Williams and Stevens [135] found that acted emotions tend to be more exaggerated than real ones. Nonetheless, the relationship between the acoustic correlate and the acted emotions does not contradict that between acoustic correlates and real ones.

Who utters the emotions?: In most emotional speech databases, professional actors are invited to express (or feign) pre-determined sentences with the required emotions. However, in some of them such as the Danish Emotional Speech (DES) database [38], semi-professional actors are employed instead in order to avoid exaggeration in expressing emotions and to be closer to real-world situations.

How to simulate the utterances?: The recorded utterances in most emotional speech databases are not produced in a conversational context [69]. Therefore, utterances may lack some naturalness since it is believed that most emotions are outcomes of our response to different situations. Generally, there are two approaches for eliciting emotional utterances. In the first approach, experienced speakers act as if they were in a specific emotional state, e.g. being glad, angry, or sad. In many developed corpora [15,38], such experienced actors were not available and semi-professional or amateur actors were invited to utter the emotional utterances. Alternatively, a Wizard-of-Oz scenario is