Chapter 9

Analysis of Speaker’s Age Using Clustering Approaches With Emotionally Dependent Speech Features

Hemanta Kumar Palo
Siksha ‘O’ Anusandhan (Deemed), India

Debasis Behera
https://orcid.org/0000-0002-3983-0891
C. V. Raman College of Engineering, India

ABSTRACT

Emotions are age, gender, culture, speaker, and situationally dependent. Due to an underdeveloped vocal tract or the vocal folds of children and a weak or aged speech production mechanism of older adults, the acoustic properties differ with the age of a person. In this sense, the features describing the age and emotionally relevant information of human voice also differ. This motivates the authors to investigate a number of issues related to database collection, feature extraction, and clustering algorithms for effective characterization and identification of human age of his or her paralanguage information. The prosodic features such as the speech rate, pitch, log energy, and spectral parameters have been explored to characterize the chosen emotional utterances whereas the efficient K-means and Fuzzy C-means clustering algorithms have been used to partition age-related emotional features for a better understanding of the related issues.

INTRODUCTION

The understanding and identification of human affective states have been a study of research during the last few decades. These emotional attributes are manifested in the form of facial expressions, gestures, postures or during a spoken dialogue. Among these modalities, speech remains an important and lone

DOI: 10.4018/978-1-7998-1021-6.ch009
medium of communication via phone. Nevertheless, the acoustic features describing human speech emotions vary with age, gender, culture, physical and mental conditions of a speaker. Thus, the design and development of an efficient Emotional Speech Recognition (ESR) system for human-computer interaction remains a complex domain of research until today. In this regard, the age of a speaker plays an important role with varying acoustic properties that can influence the performance of an ESR system drastically (Lyakso, & Frolova, 2015A; Lyakso, et al., 2015B). Absence of adequate research in this direction motivates the authors to consider the age and emotional relevant acoustic cues that influence the performance of an ESR system. It is a well-known fact that the speech production mechanism goes many structural changes during different stages of a human life-cycle. The vocal tract and vocal folds of a child take time to develop until he becomes a fully grown-up adult. Thus, the prosodic characteristics that describe the voice of a child and his affective states arguably differ from that of adults. As a result, an ESR system destined for a particular age group may influence the accuracy of a machine learner if applied to other age groups (Palo, Mohanty, & Chandra, 2018D; Chaudhari, & Kagalkar, 2015A; Hämäläinen, et al., 2011). The system requires suitable features which are discriminating, reliable, and can provide emotional relevant information across different age. Alternatively, the features can be used to partition people of different age groups and their affective states with some efficient clustering algorithms. The outcomes of these ESR systems can provide psychological assistance to children affected with psychological trauma or negative emotions. Service industries in the area of telecommunications, meeting browser, speech translation, human-robotic interfaces, assessment systems, smart call center, smart workspaces, intelligent tutoring, dialogue and language learning systems, forensic labs, etc. can be managed more efficiently using these systems. The criminal investigators, lawmakers, psychologists, etc. can use the system to counsel or interrogate both victims and culprits irrespective of their age. An accurate evaluation of the intended emotional states will assist the law enforcement agencies in decision making during court hearings. The paralanguage information identification system based on the age of anti-social elements, radicals, fanatic individuals, and influential people can help the law administrators monitor the defaulters. It can serve as a deterrent to persons involved with intimidating calls, false alarms, kidnappings, etc. These factors motivate this piece of work and lay the foundation for the current investigation. It focuses on the issues, algorithms and the current trends that influence this field in the present scenario. Most of the issues related to database generation and collection, feature extraction and clustering algorithms, emphasizing a speaker’s age have been analyzed elaborately. A few simulation results to characterize and cluster the chosen state of emotions across different age groups have been graphically portrayed for a better understanding of the related issues.

The rest of the paper has been organized as follows. The related literature is provided in section 2. The issues that motivate this piece of research is briefed in section 3. An overview of the different database corresponding to different age groups has been made in section 4. The application domain of different feature extraction techniques related to ESR has been investigated in section 5 whereas section 6 provides information on the clustering algorithms discussed in this work. A brief overview of the fuzzy and k-means clustering algorithms has also been depicted in this section. Section 7 concludes the work with possible future directions.

**Related Literature**

The effect of age on speech information can be studied using discriminative parameters such as the pitch, jitter, formants, shimmer, Mel Frequency Cepstral Coefficients (MFCC), harmonic to noise ratio (HNR),