Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-021-02979-3

ORIGINAL RESEARCH

Impact of autoencoder based compact representation on emotion detection from audio

Nivedita Patel[1] · Shireen Patel[1] · Sapan H. Mankad[1]

Received: 17 June 2020 / Accepted: 15 February 2021

Abstract
Emotion recognition from speech has its fair share of applications and consequently extensive research has been done over the past few years in this interesting field. However, many of the existing solutions aren't yet ready for real time applications. In this work, we propose a compact representation of audio using conventional autoencoders for dimensionality reduction, and test the approach on two benchmark publicly available datasets. Such compact and simple classification systems where the computing cost is low and memory is managed efficiently may be more useful for real time application. System is evaluated on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Toronto Emotional Speech Set (TESS). Three classifiers, namely, support vector machines (SVM), decision tree classifier, and convolutional neural networks (CNN) have been implemented to judge the impact of the approach. The results obtained by attempting classification with Alexnet and Resnet50 are also reported. Observations proved that this introduction of autoencoders indeed can improve the classification accuracy of the emotion in the input audio files. It can be concluded that in emotion recognition from speech, the choice and application of dimensionality reduction of audio features impacts the results that are achieved and therefore, by working on this aspect of the general speech emotion recognition model, it may be possible to make great improvements in the future.

Keywords Audio · Emotion · RAVDESS · TESS · Autoencoder

1 Introduction

Speech is one of the major communication methods used by humans (Mustaqeem and Kwon 2019). Emotions are forms of expression for humans and therefore, emotion is naturally used in everyday speech by human beings for expressing their sentiments clearly. Speech contains both linguistic and non linguistic information (Mansour et al. 2019). A speech signal contains information like intended message, speaker identity and emotional state of the speaker (Bhaykar et al. 2013). Efficient communication through language and speech has enabled sharing of ideas, messages, and perceptions to one another. In voice based signals, there are two factors of primary importance: acoustic variation and words that are spoken. Acoustic features such as the pitch, timing, voice quality, and articulation of the speech signal highly correlate with the underlying emotion due to the effects of arousal in the nervous system, increased heart rate, etc. The variation of these features forms the basis of emotion recognition in speech.

Speech emotion recognition is the task of extracting the emotions of the speaker from his or her speech signal. Detecting these emotions provide insight into deeper complexities that help to navigate through real time situations. Emotion recognition from speech is one of the major challenges in the field of human computer interaction. The formulation of powerful emotion recognition systems are thus beneficial and the objective of a good emotion recognition system is to be able to mimic human perception in the way that humans are able to detect emotions such as anger, sadness, and happiness while talking to one another (Basu et al. 2017). Despite extensive research in emotion recognition from speech, there are still several challenges such

✉ Sapan H. Mankad
sapanmankad@nirmauni.ac.in

Nivedita Patel
17bce084@nirmauni.ac.in

Shireen Patel
17bce088@nirmauni.ac.in

[1] CSE Department, Institute of Technology, Nirma University, Ahmedabad, India

Published online: 03 March 2021