Abstract

The goal of this research is to design a computational model for recognizing emotional states using speech across multiple languages. Recognition of emotion in speech in general relied upon specific training data, and a different speaker and/or language may present significant challenges. Traditions on multilingual speech emotion recognition have been done to be extensively focused on reducing the differences between different speakers and/or languages by using different approaches, like normalization strategies, domain adaption methods, transfer learning algorithms, and so on. However, those previous studies usually required pre-knowledge on the target data that is challenging in practice; and most specific, they do not consider and clarify the effective features in common for different data sets. On the other hand, cross-lingual studies have revealed that human speech perception of emotion shares universal rules among people who speak different languages and provided a set of common emotions independent of languages, like anger, happy, sad, fear, disgust and surprise. By contrast, limited research focus has been gained to study the process of human emotion perception for multilingual speech emotion recognition. Thus, it motivated us to answer exactly what commonalities were in the speech perception of emotion across multiple languages.

In this regard, a three-layer model conducted with acoustic features, semantic primitives, and emotion dimensions was studied, taking inspiration from the fact that there exist multiple layers in a process of human emotion recognition. To achieve the research goal, 215 statistical acoustic features derived from prosodic and spectral domains, and seventeen semantic primitives were first explored. The shared appropriate parameter sets of acoustic features and semantic primitives were then determined by a feature selection algorithm. The implemented three-layer model suggested 22 acoustic features and four semantic primitives to be universal in the process of emotion perception in multilingual speech.

Based on studies implemented in this research it was demonstrated that the combination of acoustic features of prosodic and spectral domains is beneficial for multilingual speech emotion recognition; the proposed feature selection method can handle well in determining associations in a three-layer model, providing appropriate features across languages. The computational model for recognizing emotional states using speech that was implemented in this research was speaker and language independent where it does not require any pre-knowledge on the target speaker and/or language. This three-layer model yielded a comparable performance by multilingual and monolingual speech emotion recognition across three different languages; in addition, the cross-lingual tasks interestingly provided comparable results to those obtained by human evaluations in cross-lingual studies.

The outcome of this research is potential to contributing the recognition of emotion from speech irrespective of speakers and languages in many areas, such as affective speech-to-speech translation, call centre application, and healthcare.

Keywords: Multilingual perception of emotion, three-layer model, emotion space, modulation spectral feature, prosodic feature