氏　　名	LI, Xingfeng		
学位の種類	博士(情報科学)		
学位記番号	博情第 416 号		
学位授与年月日	令和元年 6 月 24 日		
論文題目	A Three-Layer Model Based Estimation of Emotions in Multilingual Speech		
論文審査委員	主査	赤木　正人	北陸先端科学技術大学院大学　教授
		党　　建武	同　教授
		鵜木　祐史	同　教授
		北村　達也	甲南大学　教授
		LI, Junfeng	中国科学院音響研究所　教授

論文の内容の要旨

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

論文審査の結果の要旨

本論文は，ヒトの感情知覚を説明する三層構造感情知覚モデルの構築と，複数言語で話された音声に含まれる感情自動推定への本モデルの適用に関する研究報告である。

音声に含まれる感情の認識は，多くの研究機関で研究されているホットな話題である。しかし提案されている認識システムの多くは，単一言語に最適化されたシステムである。言語が異なる環境での音声コミュニケーションの重要性が議論されているにもかかわらず，感情認識システムにおける多言語への展開・適用は，言語ごとの再学習に頼らざるを得ない状況となっている。一方，ヒトは，異なる言語で話された内容が理解できない音声であっても，その音声に含まれる感情は知覚することが可能である。すなわち，ヒトの基本的な感情知覚は言語によらない。

本論文では，ヒトの優れた感情知覚能力を模擬することを目指して，感情知覚を説明するモデルである Brunswik レンズモデルにもとづいた計算機上の実行モデルを議論し，第一層である音声特徴量，第二層であるセマンティックス，第三層である感情因子の各層を持つ三層構造感情知覚モデルを構築した。構築においては，ヒトを用いた聴取実験による基礎データの収集，各層の表現法および最適な特徴量セットの決定，Fuzzy Inference System を用いた層間の結合を行い，最終的に 22 個の音響特徴，4 種類のセマンティックス，2 種類の感情因子からなる三層構造感情知覚モデルが実現した。本モデルは，感情知覚のメカニズムを解明する点から構築が試みられているため, End-to-End モデルのようなデータ依存の

深層学習では対応できない分野へも応用可能であると考えられる。

モデルの評価を行うために，日，中，独三か国語の感情音声データセットを用い，音声中の感情推定実験，および，推定値を用いた感情認識実験を実施した。音声中の感情推定実験では，モデルの第三層を構成する 2 種類の感情因子の値について，ヒトによる聴取実験で得られた値と本モデルによる推定値の比較を行った。実験の結果，どの言語においてもその差は小さいものとなり，ヒトの知覚結果と同等の性能が得られた。感情認識実験では，本モデルによる 2 種類の感情因子の推定値から得られる感情カテゴリを認識結果として，同じ感情音声データセットを用いた競合他者の認識実験結果と比較した。比較の結果，すべての既報データを上回る性能を実現した。

以上のように，本研究は新しい概念のもとで，話された言語によらず音声中の感情を高精度で推定する手法を実現したものであり，学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。