Paralinguistic and Nonverbal Information Extraction from Speech Signal towards Empathetic Dialogue Systems

FUJIMURA, Hiroshi / 藤村, 浩司

(開始ページ / Start Page)
1
(終了ページ / End Page)
84
(発行年 / Year)
2022-03-24
(学位授与番号 / Degree Number)
32675甲第546号
(学位授与年月日 / Date of Granted)
2022-03-24
(学位名 / Degree Name)
博士(理学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)
(URL)
https://doi.org/10.15002/00025229

博士学位論文
論文内容の要旨および審査結果の要旨

論文題目	Paralinguistic and Nonverbal Information Extraction from Speech Signal towards Empathetic Dialogue Systems
氏名	藤村　浩司
学位の種類	博士（理学）
学位番号	第 546 号
学位授与年月日	2022 年 3 月 24 日
学位授与の要件	法政大学学位規則第 5 条第 1 項第 1 号該当者（甲）
論文審査委員	主　査　内田　薫　　教授
	副　査　伊藤　克亘　　教授
	副　査　黄　潤和　　教授

１．論文内容の要旨

本論文では、対話エージェントが人間のエージェントのように相手に共感し、話者の状況に合わせたコミュニケーションを可能するために、音声情報から共感に重要なパラ言語情報、非言語情報を抽出する。画像情報の使えない、コールセンターや AI スピーカーにおいて、対話エージェントが話者に共感し、適切に対応するためにはその対話システムが達成したいタスクに関する知識や一般知識の活用、パーソナライゼーション、感情の認識が必要である。このうちテキスト情報を用いたタスクに関する知識や一般知識と対話システムとの融合は多くの研究がなされているが、直接テキスト情報からは得られないパーソナライゼーションに関する情報や、感情の情報の抽出・活用に関してはまだ研究が足りていない。そこで、本論文では、パーソナライゼーションや感情のような共感に結び付く、音声から得られるパラ言語・非言語情報である、感情や人の話し方の特徴、話者の属性の検出・認識に焦点をあてる。それぞれを構成するアルゴリズムは検出・認識の精度を向上させると共に、リアルタイムコミュニケーションシステム構築に重要である計算量、リアルタイム・即時性、言語依存性に関しても考慮し提案する。２章では、話者属性の認識、３章では話し方に関わる音素認識、４章では同じく話し方に関わるフィラーや言いよどみなどの音響イベントの検出、５章では感情の認識について述べる。

２章では、話者属性の認識について述べる。既存の話者属性手法では音声区間検出と属性識別を別々に考えており、計算量が多く、即時性がなかった。提案手法では、短時間フレーム毎に話者の属性の識別と音声、非音声の識別を同時に行うニューラルネットワークを適用し、それぞれのフレームにおける識別結果を統合することにより、音声区間の検出と属性認識を逐次的に同時に解く。このアルゴリズムを適用することによって、音声区間検出、話者属性認識のリアルタイム性、即時性が向上した。

３章では、言い怠けに関連するスコアを求めるための音素識別について述べる。従来の音素スコアを算出するための音素識別では、前後の音素環境状態を用いて、決定木による

音素のクラスタリングを行う必要があった。しかし、音素のクラスタリングや識別器の構築には大量の学習データが必要となる。そこで、音素クラスタリングを内包する AdaBoost 識別器を提案し、決定木による音素クラスタリングを必要としない音素識別器を構築することが可能となった。

４章では話し方に関わるフィラーや言いよどみなどの音響イベント検出について述べる。フィラーや言いよどみ検出の既存手法では、終端まで音声認識を処理した後に、検出処理を行っていたため、リアルタイム性、即時性がなかった。また、すべての言いよどみパターンを辞書に登録することも困難であった。提案手法では、フィラーや言いよどみの音響現象を検出する LSTM-CTC モデルと音素ループによってあらゆる言いよどみ語を受理する WFST を用いて、新たに辞書に言いよどみ語を登録することなく、動的にフィラーや言いよどみを検出する。これにより、リアルタイム性、即時性が高く、言語依存性の低いフィラー・言いよどみ語の検出が可能となった。

５章では、感情の認識について述べる。感情認識に関してはラベル付きの感情音声データベースの量が少なく、大量のパラメータを持つニューラルネットワークの学習が困難であり、感情の特徴を捉えるために必要な、複数の時間解像度の特徴量を扱うことが難しかった。提案手法では複数の時間解像度それぞれに対して構築した LSTM の出力を、中央値特徴量を用いた XGBoost により統合し、感情識別性能を向上した。また、XGBoost の特徴量選択を用いることにより、必要な時間解像度を選択し、少ない計算量のモデルを構築した。

２．審査結果の要旨

本論文では、共感対話システムに重要なパーソナライゼーション、感情に関わるパラ言語情報、非言語情報を音声信号から抽出している。またそれら抽出アルゴリズムは共感対話システム構築に向けて、既存手法の性能、リアルタイム性、即時性、言語依存性についての課題を解決している。これらの成果は情報科学の発展に貢献するところが大きい。

本論文は学位請求者が、情報科学分野全般で高度な素養を持ち、新規性のある概念等を構成でき、また新しい研究領域あるいは新しい応用領域の開拓を行う能力を有することを示しており、博士論文としての評価基準を充分満たすものである。また、公聴会を通した口頭試問により、提出された論文を中心に関連する学問領域の試問を行った結果、学位請求者が合格に値する十分な学識を有していることを確認した。

よって、本審査小委員会は全会一致をもって提出論文が博士（理学）の学位に値するという結論に達した

（報告様式Ⅲ）