博 士（情報科学） 藤 原 敬 記

学 位 論 文 題 名

音声対話システムにおける頑健な言語理解と自然な音声対話に関する研究

学位論文内容の要旨

現在，情報機器の高速化・高度化を背景に身の回りの様々な場所でコンピュータが利用されている．それらの入出力装置としてはタッチパネルやディスプレイ，あるいはキーボードやマウスといったものが一般的である．しかし一方で，これらの入出力装置は車載情報機器などでの使用においては，安全性や操作性などを考えた場合，必ずしも最適なインターフェースとはいえない．そこで近年マンマシンインターフェースとして，音声対話システムが注目されてきている．音声対話システムは，音声認識，言語理解，問題解決，対話制御，音声合成など様々な技術を統合したシステムである．音声は，人間が日常用いているコミュニケーション手段であり，ユーザにとって親しみやすく，また特別な訓練が必要ないという利点がある．さらに手や目を占有しない (ハンズフリー・アイズフリー) ので他の作業を行いながらでも使用可能であると言った利点もある．

実環境での音声対話システムの使用においては様々な問題が存在する．そのひとつに音声認識誤りがある．音声認識は長年に渡って研究されてきており，孤立発話や読み上げ音声などに対しては，実用レベルに達しているといえる．実際にニュースの字幕作成において音声認識が実用化されている．また現在では市販のソフトウェアとして販売されているものも少なくない．しかし，アナウンサーが原稿を読み上げている音声やきちんと発声している音声に対しての認識率は高いものの，日常会話の話し言葉などの自発発話の音声に対する認識率はそれほど高くない．自発発話の音声認識が難しいのは，自発発話音声には，言い直し，言いよどみ，繰り返し，間投詞，不正確な発音などが含まれ，音響的にも言語的にも読み上げ音声とは大きく異なるためである．そのため，実環境での音声対話システムの使用を考えた場合には，誤認識を完全に回避することは難しい．誤認識が起きると，システムはユーザの期待する応答とかけ離れた応答を行い，対話がスムーズに進まなくなることも多くある．そこで本研究では，音声認識器が誤認識した場合でも，正しくユーザの意図を推定することができる頑健な音声言語理解手法を提案する．提案手法では，音声認識器が誤認識した場合でも多くの場合，複数候補 (N-best) 中に正解が含まれていること，システムが誤認識した場合にはユーザは大体訂正反応を示すこと，タスク指向対話には強い一貫性がありユーザは基本的に意味的・文脈的に関係した内容以外を発話しないことを利用する．また，提案手法では予め全ての認識可能単語を理解候補として保持し，言語理解部の対話戦略において音声認識結果中の単語との意味的関連性などを考慮している．これにより音声認識結果の N-best 中に正解の一部が含まれていない場合でも，複数のユーザ発話の認識結果に基づくことで正しい意図を推定することが可能となっている．評価

データにおいて, 提案手法における対話単位での理解率は 72.2%(21,430/29,670 対話), 単語単位での理解率は 87.1%(77,544/89,010 単語) であり, 従来手法の最新認識結果の上位候補を優先するシステムの 57.9%(17,178/29,670 対話),75.4%(67,084/89,010 単語) と比較しても有効であることが明らかになった.

提案した言語理解手法により言語理解部における性能の向上は確認された. しかし, 実際のシステムの使用においては, ユーザ意図として可能性のある理解候補を全て考慮し計算を行っているため, ユーザが発話してからシステムが応答するまで若干のタイムラグがあり, 人間同士の自然な対話とはいえなかった. これは, 人対機械におけるコミュニケーションにおいては, 音声認識や言語理解の精度, 合成音声の質などだけではなく, テンポやリズム, システム発話の韻律情報 (以降, 対話リズムと呼ぶ) も重要であると言う知見と一致する. この対話リズムは, 発話者の状態 (対話の盛り上がり, 対話意図, 感情など) により大きく変化すると考えられる. 対話リズムに与える影響が最も強い要素として対話意図が考えられるが, 対話意図と対話リズムの関係を分析した研究はほとんど存在しない. そこで, 人間同士のコミュニケーションの自然さで音声対話が可能な音声対話システムを実現するために, 実際の人間同士のタスク指向型の音声対話を収集し, 対話意図と対話リズム (発話タイミング・F0・発話速度) に関する分析を行った. その結果, 発話タイミングは, 発話者の思考時間, 発話内容の重要度, 対話相手の予想 (期待) と同じ対話意図かどうかが影響することが明らかになった. また F0 と発話速度は発話内容の重要度, 対話相手の予想 (期待) と同じ対話意図かどうかが影響することが明らかになった.

人間同士のタスク指向対話における対話意図と対話リズムの関係について分析を行ったが, 今後さらにコンピュータが普及することを考えると, より多様な状況での音声対話システムの使用が予想される. 近年では, 喜怒哀楽に代表される感情音声に対する研究も行われてきており, 様々な発話様式への取り組みがみられる. 一方人間の発話を考えると, 人間は様々な音韻要素・韻律要素を柔軟に制御し様々な発話様式を実現している. しかし, 日常会話などにおいて人間は常にはっきりと発話しているわけではない. 発話内容や発話様式, 対話意図などによりそれほどはっきりと発話しないことがある. つまり自然音声には明瞭である部分とそうではない部分があり, この状態が連続的に変化することで, 自然性や人間らしさと感じている可能性がある. そこで, この明瞭性の変化を合成システムへ導入することで, 人間の多様な発話様式にも対応可能な音声対話システムの実現を目指す. 本研究では, 明瞭性を変化させるために従来あまり扱われていなかった比較的明瞭でない部分, つまり音声の「あいまい」な部分に注目する. なお本研究における「あいまい」とは重要な情報の含まれない部分などの「個々の音韻ははっきりしないが部分全体としてなめらかである」状態を示しており, より長い発話単位 (文, 呼気段落など) 全体で表現される, 伝えたい内容の了解度を失ってしまうような, 文全体に渡った「不明瞭さ」ではない. 特に文意に関わる重要な部分の明瞭性は高く保ったままである. 本論文では, 予備調査によりあいまいな音声に特に顕著な変化として観測された,F0, パワー, フォルマント周波数を後処理加工することで合成音声の明瞭性を制御することを試みた. また制御した合成音声を聴取実験により評価した. その結果, 発話内容により明瞭性を変化させた合成音声は,「丸みのある」,「やわらかい」といった人間性に関係する形容詞について無加工の音声より強い印象を持つことが判明した. また,「落ち着いた」,「冷静な」といった印象も強くなることが明らかとなった.

学位論文審査の要旨

主　査　教　授　荒　木　健　治
副　査　教　授　山　本　　強
副　査　教　授　長谷山　美　紀
副　査　准教授　伊　藤　敏　彦

学位論文題名

音声対話システムにおける頑健な言語理解と自然な音声対話に関する研究

著者は，音声対話システムにおいて音声認識誤りが発生した場合でもユーザの発話内容を正しく理解できる頑健な言語理解手法についての提案を行った．音声対話システムは，音声認識，言語理解，問題解決，対話制御，音声合成など様々な技術を統合したシステムである．音声は人間が日常用いているコミュニケーション手段であり，ユーザにとって親しみやすく，また特別な訓練を必要としないという利点がある．さらに手や目を占有しない (ハンズフリー・アイズフリー) ので他の作業を行いながらでも使用可能であるといった利点もある．一方，音声認識誤りが起きると，システムはユーザの期待する応答とかけ離れた応答を行い，対話がスムーズに進まなくなるという問題がある．そこで著者は，音声認識誤りによる理解誤りを回避するために，音声認識誤りが発生した場合でも多くの場合，複数候補 (N-best) 中に正解が含まれていること，システムが誤認識した場合にはユーザは大体訂正反応を示すこと，タスク指向対話には強い一貫性がありユーザは基本的に意味的・文脈的に関係した内容以外を発話しないことを利用した．また，提案手法では予め全ての認識可能単語を理解候補として保持し，言語理解部の対話戦略において音声認識結果中の単語との意味的関連性などを考慮している．このことにより音声認識結果の N-best 中に正解の一部が含まれていない場合でも，複数のユーザ発話の認識結果に基づくことで正しい意図を推定することが可能となっている．評価データにおいて，提案手法における対話単位での理解率は 72.2%(21,430/29,670 対話)，単語単位での理解率は 87.1%(77,544/89,010 単語) であり，従来手法の最新認識結果の上位候補を優先するシステムの 57.9%(17,178/29,670 対話),75.4%(67,084/89,010 単語) と比較しても有効であることが明らかになった．

次に著者は，人間と対話システムが自然にやりとりできるためのテンポやリズム，システム発話の韻律情報などを「対話リズム」と定義し，この対話リズムと対話意図との関係について調査を行った．これは，上述の対話システムは理解精度の点では従来手法よりも向上しているにもかかわらず，人間同士の対話ほどの自然性が見られなかったことと，近年の知見において明らかになった，人対機械の対話においては理解精度だけではなく対話リズムも重要であることに基づいている．著者は，

対話リズムを構成する要素として，発話タイミング，F0，発話速度を考え，これらと対話意図の関係について分析を行っている．その結果，発話タイミングは，発話者の思考時間，発話内容の重要度，対話相手の予想 (期待) と同じ対話意図かどうかが影響することが明らかとなった．また F0 と発話速度は発話内容の重要度，対話相手の予想 (期待) と同じ対話意図かどうかが影響することが明らかとなった．

次に著者は，多様な状況にも対応可能な多様な発話様式を持つ対話システムについて研究を行った．人間は様々な音韻要素・韻律要素を柔軟に制御し様々な発話様式を実現している．しかし，日常会話などにおいて人間は常にはっきりと発話しているわけではなく，発話内容などによりそれほどはっきりと発話しないことがある．著者は，自然音声には明瞭である部分とそうではない部分があり，この状態が連続的に変化することで，自然性や人間らしさと感じている可能性を考え，この明瞭性の変化を音声合成へ導入することで，人間の多様な発話様式の実現を試みた．F0，パワー，フォルマント周波数を後処理加工することで合成音声の明瞭性の制御を行い，作成した合成音声を聴取実験により評価した．その結果，発話内容により明瞭性を変化させた合成音声は，「丸みのある」，「やわらかい」といった人間性に関係する形容詞について無加工の音声より強い印象を持つことが判明した．また，「落ち着いた」，「冷静な」といった印象も強くなることが明らかとなった．

著者は論文全体を通じて，研究領域の現状の分析，新規提案内容の記述，有効性の主張，研究領域における位置づけを正確に行ったと判定する．

以上を要約すると，著者は音声対話システムにおいて，認識信頼度と対話履歴を用いる言語理解手法を提案し，音声認識誤りが発生した場合でも正しくユーザの発話内容を理解することが可能であることを示した．また，対話リズムに注目し，対話意図と対話リズムの関係について分析を行った．さらに多様な発話様式を実現するための試みとして明瞭性の制御を行った．本研究を通じて情報メディア工学，音声言語処理工学の発展に貢献するところ大なるものがある．よって，著者は北海道大学博士 (情報科学) の学位を授与される資格あるものと認める．