様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ３ 日現在

機関番号：１１３０１
研究種目：若手研究
研究期間：2018～2019
課題番号：１８Ｋ１８１３６

研究課題名（和文）音声アシスタントとの対話を介した非接触・暗黙的なヘルスモニタリング技術の研究

研究課題名（英文）A study on an implicit health-monitoring technique through the dialog with a speech assistant

研究代表者
千葉　祐弥（CHIBA, Yuya）

東北大学・工学研究科・助教

研究者番号：３０７８０９３６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本課題ではまず，音響的非言語シグナルからユーザ状態の推定を行う手法を検討した．感情音声合成を用いてデータを増強することで音声感情認性能が従来よりも大幅に改善することを示した．また，近年注目を集める系列モデルに対してセグメント特徴量の導入とネットワークのマルチストリーム化を行うことでさらに性能を向上させた．最終的な識別精度は73.4%であり，これは人間による判別に肉薄する結果である．また，対話型システムを継続的に使うための対話制御技術に関しても検討を行った．人間同士のマルチモーダル雑談コーパスを用いて言語・非言語情報の分析を行うことで，関係性の段階による対話戦略の違いを明らかにした．

研究成果の学術的意義や社会的意義
当初の目的である健康状態の推定には至らなかったものの，雑音重畳や統計的音声合成によりデータ増強を行うことでユーザ状態推定の頑健性を向上できることを示した．この手法は音声感情認識だけでなく，様々なユーザ状態推定に応用可能であるため，音声を用いたアプリケーションの多くで有用である．また，識別器の改善により音声感情認識自体も人間による判断に匹敵する性能が得られることを示した．加えて，対話型アプリケーションが継続的に使われるための言語/非言語的ふるまいの変化に関して，包括的な対話制御モデルを構築するための手がかりとなる結果を得た．

研究成果の概要（英文）：In this study, we examined estimation methods of user's emotion by using acoustic non-verbal signals at first. The experimental results showed that data augmentation based on emotional speech synthesis significantly improves recognition accuracy comparing with the conventional method. Then, we proposed an approach based on multi-stream attention-based BLSTM with segmental features, and further improved the recognition performance. The proposed method obtained 73.4% of recognition accuracy, which is comparable to human evaluation (75.5 %). In addition, we examined a dialog management method prompting to use continuously. We confirmed that the difference of dialog strategies between groups of different relationships by the analysis of verbal/non-verbal information using a multimodal chat-talk corpus of human-human conversation.

研究分野：音声対話システム

キーワード：ユーザ状態推定　音声感情認識　音声対話システム

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

医療における診察には問診が重大な役割を担っており，問診の精度向上のためには日々のきめ細かな健康情報の記録が鍵となる．健康情報の記録のための技術としてウェアラブルデバイスがあるが，常時健康情報の記録が可能である反面，装置の着用が必要となるため，継続的な利用につながらないという問題が指摘されている．このような背景から，接触型デバイスを用いず無意識のうちに健康情報を記録するヘルス・ライフログ技術の実現が強く望まれている．　一方，近年 Amazon Echo のようなマイクロホンやカメラを備えた対話型ホームデバイスが普及しはじめており，音声によるコンピュータの操作が一般的になりつつある．音声を伴うコミュニケーションでは，内的状態，感情，健康状態などを伝達する非言語シグナル (例えば声の調子や顔色など)の利用が可能であり，これらを有効活用することで，非接触なヘルス・ライフログシステムの実現が見えてくる．

２．研究の目的

本研究では，近年注目を集める対話型ホームデバイスとの対話において，マイクロホンやカメラを用いて利用者の声の調子や顔色といった非言語シグナルを取得することで，非接触かつ暗黙のうちに日々の健康情報が記録可能な対話型ヘルスモニタリングシステムの実現を目指した．目標とするシステムを実現するためには，非言語シグナルに基づくユーザの生体情報の獲得と包括的な健康状態の評価が必要となる．しかしながら，非言語シグナルとユーザの状態の関係性の分析は音声分野における嗄声などの限定的な検討にとどまる．本研究課題は，対話型デバイスとのインタラクションからこのような非言語シグナルを抽出し，非接触かつ暗黙のうちに利用者の日々の健康情報をモニタリングし記録する，対話型ヘルスモニタリングシステムを実現するための要素技術の構築を目指した．

３．研究の方法

当初の研究計画では，1)健康状態に依存したマルチモーダル対話データベースの構築，2)音声・画像情報からのユーザの生体情報の取得，3)深層学習に基づく包括的な健康状態推定の 3 つの項目を遂行する予定であった．しかしながら，ユーザの健康状態の推定に用いる非言語シグナルの取得に関する頑健性に問題があり，まず取得可能な非言語シグナルの状態推定への効果を確認する必要性が生じた．また，現状の対話型システムはユーザに日常的に利用させることが難しく，継続的な利用を促す対話制御手法の構築が必要となった．そこで，本研究課題では，健康状態そのものを対象とする前に，ユーザのメンタルヘルスに関係性の深いと考えられる感情状態を対象として非言語シグナル，特に音響的非言語シグナルの効果と識別モデルの頑健性の向上に関する検討を行うこととした．また，同時にマルチモーダル対話システムの対話制御手法に関しても検討を行った．下記にその概要を示す．

(1)．音声感情認識システムによる非言語シグナルの効果の検証

① 音声感情認識における非言語シグナルの比較と頑健性の向上
音声感情認識に一般的に用いられる特徴量を対象として性能の比較を行った．また，音声感情認識・雑音重畳によるデータ拡張によって状態推定の頑健性の向上を図った

② マルチストリーム音声感情認識による特徴量の効果的な利用
音声感情認識において注目を集める Attention-based BLSTM を改良し，異なる性質を持つ非言語シグナルを効果的に利用する手法を検討した．

(2)．非タスク指向対話システムのための継続的な利用を促す対話制御手法の検討
ユーザと関係を構築し，継続的な利用を促す対話制御モデルを実現するため，人間同士の対話データを収録しその対話戦略を分析した．

４．研究成果

(1)．① 音声感情認識における非言語シグナルの比較と頑健性の向上

音声感情認識においては感情音声コーパスである，Japanese Twitter-based Emotional Speech (JTES)を用いた．音声感情認識においては，従来一般的に用いられている Low-level descriptors (LLDs)の統計量を特徴量として用いた．この特徴量にはそれぞれ 384 次元（IS2009），1582 次元 (IS2010)，6373 次元（IS2016)の特徴量セットが存在する．これを用いて特徴量セットの性能を比較した．また，JTES を用いた感情音声合成を行い，データ拡張を行った．音声の合成においては新聞記事読み上げコーパス（JNAS)の文章をそれぞれの話者・感情で生成した．これによって，発話の多様性を示す尺度であるトライフォンの種類やアクセントラベルの種類を，本来のデータセットと比較してそれぞれ 2 倍，4 倍に増強した．結果を図 1 に示す．検証により，まずは IS2010 の特徴量セットの認識性能が最も高いことが明らかになった．これは当該の特徴量セットが IS2009 よりもより網羅的な情報を含み，かつ過学習を行さない情報量であることを示唆する結果である．また，データ拡張によって IS2009 の特徴量セットにおいて 4 ポイント，IS2010

の特徴量セットにおいて 3 ポイント，IS2016 の特徴量セットにおいて 4 ポイントの性能改善を達成した．これは，比較手法である線形補間に基づくデータ拡張 (SMOTE) よりも大きな性能改善である．また，近年注目を集める Attention-based BLSTM を用いた手法においても検証を行い，データ拡張の効果を確かめた．結果の分析より，拡張データに対して分散補償を行うことで認識性能が改善することを確かめた．分散補償は生成された合成音声の自然性を向上する手法であり，この結果はより自然な合成音声を用いることでさらなる精度の改善が得られる可能性があることを示唆している．したがって今後は高品質な音声の生成が可能な手法として近年注目を集める Tacotron2 などの手法を利用して更なる精度改善を目指す．

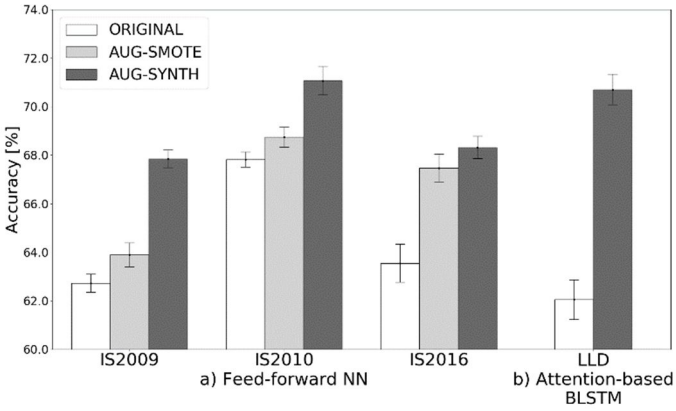

図 1 特徴量の効果の比較とデータ拡張による効果の検証

続いて，実運用化における感情音声認識システムの頑健性を向上させるため，マルチコンディション学習を利用したモデル学習を行った．雑音感情音声の生成には前述した感情音声コーパス JTES と，JEIDA-NOISE コーパスを用いた．JEIDA-NOISE コーパスは展示会場や自動車内といった，音声アプリケーションを用いる場面として自然な環境の背景音を収録した音声コーパスである．実験では，SN 比が 0dB，5dB，10dB，15dB，20dB になるように感情音声に対して雑音音声を重畳した．このデータセットを用いて，事前学習を行った場合 (PT)，直接ネットワークを学習した場合 (MC)，スペクトルサブトラクションによって雑音成分を除去した音声を用いて学習した場合 (NR) の 3 つのマルチコンディション学習手法を比較した．結果を図 2 に示す．結果より，まずはクリーンな音声で学習したモデル (Clean) は雑音の重畳された音声に対して著しく感情認識性能が低下することが見て取れる．一方で，マルチコンディション学習を行ったその他の条件では，雑音音声に対しても性能が低下しないことがわかる．したがって，雑音音声感情認識においてマルチコンディション学習が有効であることが示された．しかしながら，マルチコンディション学習を行ったモデルでは，クリーン音声に対して性能が低下するという問題が示された．したがって，今後は音声合成・音声認識の分野で検討されている話者コードベクトルを用いる適応手法を雑音環境に適用することで環境適応モデルを学習する手法を検討する．

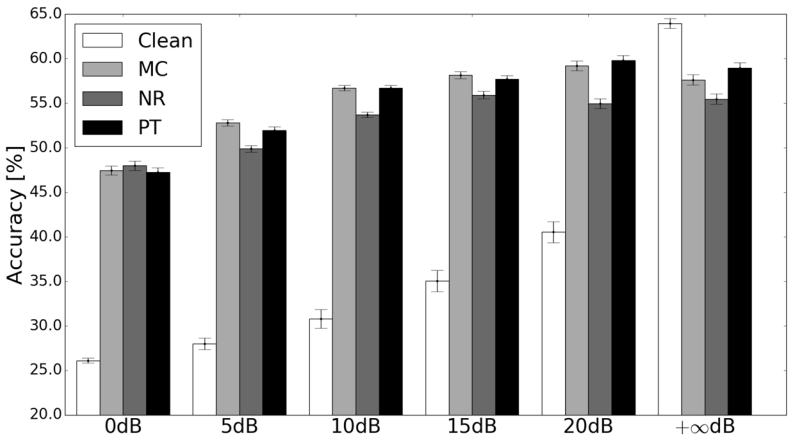

図 2 雑音音声感情認識実験の結果

(1).② マルチストリーム音声感情認識による特徴量の効果的な利用

音声感情認識の研究においては系列モデルの利用が注目を集めている．特に，注意機構付き BLSTM は，強い感情表現が含まれる局所的な区間に焦点を当てることができるとされており，研究が盛んである．従来手法の多くでは，スペクトル画像や Low-Level Descriptors (LLD) 系列が入力として用いられる．しかしながら，この方法では，従来の音声感情認識で有効とされてきた特徴量の変動を陽には考慮できない，単一のストリームで入力系列を扱うため異なる性質を持

つ非言語シグナルに対して個別の注意を学習することができないといった問題があった．そこで，本研究項目では局所領域において LLD 系列を集約したセグメント特徴量を入力とすることで，局所的な特徴量の変動とその時間変化の両方を考慮できる手法を検討した．さらに，入力系列のマルチストリーム化によって，パワー，基本周波数，スペクトルの各要素を効果的に利用可能なネットワークを構築した．図 3 に提案する音声感情認識ネットワークを示す．

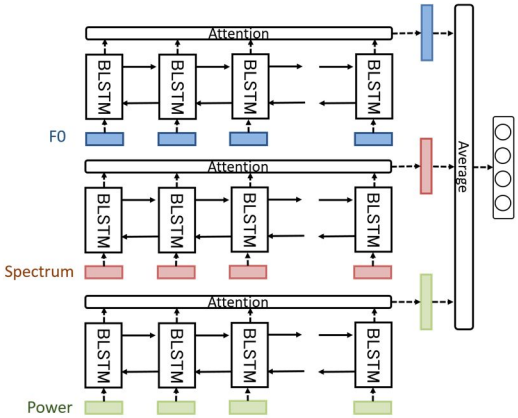

図 3 マルチストリーム音声感情認識の概要

表 1 に実験結果を示す．結果より，特徴量のセグメント化により 0.9 ポイント，ネットワークのマルチストリーム化によってさらに 1.6 ポイントの性能改善を達成した．提案する手法をすべて組み合わせた場合，ベースライン手法と比べて有意に性能が改善し，最終的に 73.4%の認識精度が得られた．我々が行った対象コーパスの評価では，音響情報のみを用いた場合に人間は 75.5%の精度で感情の判別が可能であることが明らかになっており，この結果は人間の判断結果に匹敵するものであると言える．これらの結果から，特徴量の局所的な統計量とその時間変化の考慮と，異なる特徴量群に対する個別のアテンションの適用は発話中の感情表現をとらえるのに有用であることが示唆された．

表 1 マルチストリーム音声感情認識による音声感情認識の結果 [%]

Condition	Accuracy (Avg.±SE)
Vanilla Attention-based BLSTM	70.7±0.63
+ feature segmentation	71.8±0.68
+ multi-stream	73.4±0.53

今後はさらなる精度の向上のため，言語情報を導入した認識手法を検討する予定である．

(2).継続的な利用を促す対話型システムの対話制御の検討

ユーザと関係を構築し，継続的な利用を促す対話システムを構築するため，人間同士の対話の分析を行った．分析に当たってまずマルチモーダル雑談対話コーパスの収集を行った．収集においては，可能な限り品質の高い音声・動画像データを収録するため，対話者同士をそれぞれ別の防音室に配置し，ダイナミックマイクとウェアラブルカメラ（GoPro HERO7）を用いた．対話者として大学院生 19 名（男性 15 名，女性 4 名）が収録に参加した．対話者は「相手と親しくなること」を目的に Jourald et al.によって設定された 60 個の自己開示項目を中心に選択された 5 つの話題に関して対話を行った．収録した対話データは 5 名のクラウドワーカが書き起こし，それぞれの発話に対して分析のための対話行為タグをアノテーションした．対話行為タグのアノテーションは機械学習の推定結果をもとに，1 名の実験者が人手で修正を行うことで付与した．

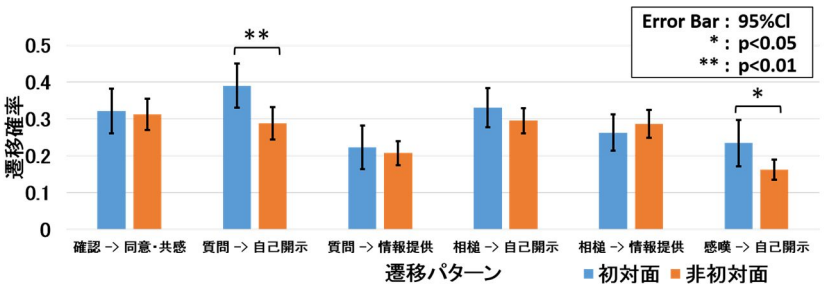

図 4 親しさの差を考慮した対話行為系列の分析結果

分析では，対話者の群を初対面のグループと面識のある話者のグループに分け，関係が構築された際の話者の挙動に関して調査を行った．まず，対話行為タグの出現頻度の分析では，「自己開示」・「質問」・「要求」に関して両グループ間に有意差があることが示された．この結果は，初対面のグループではより多くの質問を行い，多くの自己開示を行う傾向にあることを反映している．また，関係が構築されると対話相手に対して抵抗なく要求できるようになることも反映している．続いて，対話行為の遷移をバイグラムの出現頻度で比較した．結果を図 4 に示す．結果より，面識のある話者同士の対話では，質問に対して自己開示を行うといった典型的な対話行為の遷移が起こりにくくなることが示唆された．

非言語情報に関しても同様に分析を行った．発話ごとに同調に関する韻律特徴量[Kawahara et al. 2015]の抽出を行い，Convolutional Neural Network (CNN)を用いた表情認識により表情ラベルを付与した．初対面のグループと面識のあるグループにおいてそれぞれの特徴量を比較

したところ，関係性の違いによって笑顔の出現頻度や共起頻度，韻律特徴量の同調傾向が有意に異なることが明らかになった．これにより言語情報に加えて，非言語情報においても親密性の違いによって話者の対話行動が異なることが明らかになった．
今後は，獲得された知見をもとに対話戦略を統計的にモデル化する．これによって関係構築がなされたユーザとそうではないユーザに対して異なる応対が可能なシステムの構築を行い，継続的な利用の観点から対話システムの評価を行う．

(3).得られた成果の国内外における位置づけとインパクト
音声感情認識を含めたユーザの状態推定は国内では京都大学，北陸先端大学，NTT，国外では南カリフォルニア大学などの研究機関が取り組んでいるが，その研究の規模は音声認識・対話研究に比べると大きくない．しかしながら，人の気持ちや感情に寄り添うことのできる次世代の対話型インターフェースを実現するためには必須の要素技術であり，本研究課題の成果は対話システム研究の進展に寄与するものである．特に，本研究で検討したデータ拡張手法は音声感情認識だけでなく，様々なユーザ状態推定に関しても応用可能であるため，音声を用いたアプリケーションの多くで有用である．また，音声感情認識自体においても特徴量と識別ネットワークの改善により，上限性能であると考えられる人間に匹敵する識別性能が得られることを明らかにした．加えて，人間同士の対話における言語/非言語的ふるまいの分析によって，対話型アプリケーションが継続的に使われるための挙動を解明する手がかりを示した．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 山中麻衣，千葉祐弥，能勢隆，伊藤彰則
２．発表標題 話者・環境適応と波形正規化を用いた音声感情認識の精度改善
３．学会等名 音響学会春季研究発表会
４．発表年 2019年

１．発表者名 多田駿介，千葉祐弥，能勢隆，伊藤彰則
２．発表標題 相互自己開示に基づく対話システムにおける傾聴的応答生成の効果の検証
３．学会等名 音響学会春季研究発表会
４．発表年 2019年

１．発表者名 山中麻衣，千葉祐弥，能勢隆，伊藤彰則
２．発表標題 感情音声データベース JTES の主観ラベル付与に向けた予備的検討
３．学会等名 音響学会春季研究発表会
４．発表年 2019年

１．発表者名 M. Yamanaka, Y. Chiba, T. Nose, A. Ito
２．発表標題 A Study on a Spoken Dialogue System with Cooperative Emotional Speech Synthesis Using Acoustic and Linguistic Information
３．学会等名 IIH-MSP（国際学会）
４．発表年 2018年

１．発表者名
S. Tada, Y. Chiba, T. Nose, A. Ito

２．発表標題
Effect of Mutual Self-disclosure in Spoken Dialog System on User Impression

３．学会等名
APSIPA ASC（国際学会）

４．発表年
2018年

１．発表者名
山中麻衣，千葉祐弥，能勢隆，伊藤彰則

２．発表標題
音響情報と言語情報を用いた協調的発話感情付与に基づく音声対話システムの検討

３．学会等名
音響学会秋季研究発表会

４．発表年
2018年

１．発表者名
多田駿介，千葉祐弥，能勢隆，伊藤彰則

２．発表標題
相互自己開示によりユーザの印象を向上させる音声対話システムの構築と評価

３．学会等名
音声言語情報処理研究会

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考