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

平成　３０　年　　５　月　２２　日現在

機関番号：　１１３０１
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２７２０

研究課題名（和文）状態推定に基づく多様な音声の認識・合成による「人にやさしい」対話システムの研究

研究課題名（英文）Research of Human-Kind Dialogue System with Recognition and Synthesis of Various Speech Based on State Estimation

研究代表者

能勢　隆（Nose, Takashi）

東北大学・工学研究科・准教授

研究者番号：９０５５０５９１

交付決定額（研究期間全体）：（直接経費）　　9,700,000円

研究成果の概要（和文）：本研究課題では、「人にやさしい」対話システムを実現するため、多様な音声の認識および合成手法の改善や高度化、および対話者の状態推定手法とその利用について検討を行なった。具体的には(1)音声対話における感情利用の妥当性、および感情推定法について検討した。(2)音韻と韻律コンテキストを考慮した拡張エントロピーに基づく文選択法の提案・評価を行なった。(3)対話意欲推定のために、対話の収録・分析を行なった。(4)感情音声合成・感情音声認識・感情推定に利用可能な大規模な感情音声コーパスを構築した。(5)多様で高品質な音声合成手法として分散補償およびテーラーメイド音声合成手法を提案・評価した。

研究成果の概要（英文）：In this research project, we improved and advanced techniques of recognition and synthesis of various speech, and studied a state estimation technique of system users and its applications to realize a dialogue system kind to users. Specifically, (1) We studied the validity of using emotions and a technique for emotion estimation. (2) We proposed and evaluated a sentence selection technique based on extended entropy where phonetic and prosodic contexts are taken into account. (3) We recorded and analyzed dialogue data for willingness estimation. (4) We constructed a large-scale emotional speech corpus that can be used for emotional speech synthesis/recognition and emotion estimation. (5) We proposed and evaluated variance compensation and taylor-made speech synthesis as a technique of synthesizing various and high-quality speech synthesis.

研究分野：　音声情報処理

キーワード：　音声対話　感情音声合成　感情認識　音声認識　感情音声コーパス

１．研究開始当初の背景

近年、日本や欧米において人間の生活を支援・活性化させるロボットやエージェントの開発競争が激化している。これらの実現のためには人間同士の場合と同様に音声や表情などによる自然なインタラクションが可能な音声対話システムが必須であり、音声合成・感情推定・音声認識などはこれらに不可欠な基盤要素技術となっている。

個別の要素技術については、音声合成分野では研究代表者らが取り組んでいる統計モデルに基づく音声合成法がオープンソースソフトウェアとして公開されており、その柔軟性とコストパフォーマンスの面から広く研究が行われている。事実、ここ数年は音声情報処理分野のトップカンファレンス（IEEE ICASSP, ISCA INTERSPEECH など）において関連セッションの過半数を占めるに至っている。またユーザの感情や意図などの獲得についてはINTERSPEECHにて定期的にスペシャルセッションが企画されるなど、活発な研究が行われている。

一方、音声認識分野では最近になってディープニューラルネットワーク（DNN）に基づく音声認識手法が確立され、話し言葉などの自発性の高い音声や雑音下における音声についても従来に比べ大幅に認識精度が向上している。

２．研究の目的

本研究課題では、今後の高齢化社会、高度化するデジタル社会において、子供からお年寄りまで、コンピュータと自然な形で調和・共生できる基盤環境を構築するため、これまで我々が音声合成・感情推定・音声認識において培ってきた個別の要素技術を発展・融合させ、実際の対話において要素技術の相互作用を利用できるよう拡張し、ユーザの気分や状態に応じて適切な応答を生成できる「人にやさしい」対話システムを実現することを目的とする。具体的には、以下の項目について研究・開発を行う.

(1)マルチモーダル対話システムのための要素技術の拡張

①対話音声・顔特徴量からのユーザの状態推定

実際の対話においてユーザの感情や発話意図、発話状態などを頑健に抽出するため、システムが受け取る音声および顔特徴量からDNNに基づきユーザ状態を定量的に推定する手法を提案する。

②状態フィードバックに基づく協調的対話音声合成

状態推定により得られたユーザ状態ベクトルと対話により得られる言語情報に基づき、システムとして適切な応答を生成するため、状態フィードバックを利用した対話音声合成法を提案する。また、実際の対話を分析し、対話に必要なスタイルを選定する。

③対話音声認識における状態適応による認識性能の向上

音声合成の場合と同様に、ユーザ状態ベクトルを音声認識の音響モデルにフィードバックすることでモデル適応を行い認識率の改善を行う。その際、ディープニューラルネットワークの利用も検討する。

(2)要素技術の相互作用を考慮した対話システムの構築・評価

(1)で拡張・高精度化した状態推定・音声合成・認識の各要素技術を利用して対話システムを構築し、実環境において得られる対話音声を用いて評価を行い、その挙動について問題点を分析し、改善について検討する。

３．研究の方法

(1)①対話音声・顔特徴量からのユーザの状態推定

これまでの研究により、感情表現や発話様式を含む音声に対し、重回帰隠れマルコフモデル(HSMM)を用いることで特定の話者に対しその度合を推定できることがわかっている。本研究ではこの枠組を利用するとともに、従来必要であったテキスト情報を不要としかつ不特定の入力話者に対応するため、新たにDNNに基づく手法を提案する。

この手法では、音声に含まれる感情表現や発話意図などを多次元ベクトルで表現し(状態ベクトルと呼ぶ)、入力である音声特徴量と出力である状態ベクトルの関係をDNNにより表現する。モデルの学習時はあらかじめ典型的な感情や発話意図などを含んだ多数の話者による音声を用意し、感情表現や発話意図とその度合を主観評価により決定した後、この値を用いてモデルパラメータの推定を行う。これにより与えられたユーザの入力音声に対し、状態ベクトルを推定することができる。なお、音声対話におけるユーザの状態は音声だけでなく、顔の表情などにも現れるため、顔画像から抽出した顔特徴量もユーザの状態推定に利用する。

＜ユーザの平静状態を利用した推定精度の向上＞

前述したユーザの状態推定手法は、あらかじめ用意した多数の話者の音声・顔画像データを事前知識として利用する方法であるが、ここで学習されるのは学習に用いた話者全体を表現するようなモデルであり、入力するユーザの音声や顔特徴量が学習データと大きく異なる場合には推定精度が低下する可能性がある。そこで、この問題を低減するために、実際の対話においてユーザの最初の数発話における音声・顔画像データを「平静状態」であると仮定し、これを用いてモデル適応を行うことを検討する。これにより、ユーザのばらつきに対してより頑健な推定システムが構築できると考えられる。

(1)②状態フィードバックに基づく協調的対話音声合成

これまでの対話システムにおける音声応答生成において、感情などを含んだ音声を合成しようとした場合、応答生成時に「楽しげ」「悲しげ」といった離散的なスタイルを対話文に応じてあらかじめ決定し、このスタイルによる音声を合成するのみであった。これに対し、提案する音声合成手法は、(1)①で推定されたユーザの状態ベクトルをそのまま利用する。

具体的には、研究代表者らが提案した重回帰 HSMM に基づくスタイル制御法を利用し、重回帰モデルの説明変数として状態ベクトルを用いる。例えば、学習時に状態推定で使用したデータと同じ基準で設定された目標話者の音声データを用いれば、発話した際のユーザの状態と同じ状態により発話された音声を応答として生成することができる。人間自身もこのような「協調的」な応答をしばしば用いることがあることから、提案法により協調的な音声合成が実現できると考えられる。

(1)③対話音声認識における状態適応による認識性能の向上

音声認識においては、音声合成の場合と同様に従来の隠れマルコフモデル(HMM)を重回帰 HMM に拡張し、入力特徴量から推定された状態ベクトルを音響モデルに反映する。これは一般的にはオンラインのモデル適応と考えられ、感情表現や発話意図などに応じて多様に変化するユーザの音声の音響的特徴に対し、頑健な音声認識を行うことができると考えられる。この手法ではあらかじめ代表的な感情表現や発話意図の音声に対しモデルの学習を行う必要があるが、これについては最尤線形回帰(MLLR)などのモデル適応を利用する。

一方で、さらに認識率を向上させるには、近年有効性が確認されている DNN に基づく音声認識の利用も重要である。そこで本研究では、DNN による尤度計算時に従来の特徴量に併せて状態ベクトルも入力として利用することで、状態適応を行う手法についても検討を行う。なお、これらの手法では状態ベクトルを発話から推定する必要があるが、状態ベクトルの推定には一定の発話長が必要であると考えられるため、推定精度と発話長の関係についても調査を行う。

(2)要素技術の相互作用を考慮した対話システムの構築・評価

平成 28 年度までに、これまでに述べた 3 つの要素記述を対話システム向けに拡張し、平成 29 年度にはこれらを統合することで実際の対話システムを構築する。まず、ユーザがシステムに対して発話すると、その音声を入力とし、音声特徴量の抽出を行う。提案システムではマルチモーダル情報を利用するため、音声に加え、WEB カメラや Microsoft Kinect などを使用して顔画像を取得し、これから顔特徴量を抽出する。

次に、抽出したこれらの特徴量を用いて(1)①で述べたユーザの状態推定を行い、感情表現や発話意図の種類や度合の推定を行う。これにより得られた状態ベクトルと、音声特徴量を用いて音声認識を行い、これを対話処理部に渡す。対話処理部では認識結果である言語情報と状態推定により得られたパラ言語情報に基いて出力するテキスト情報を生成し出力状態を決定する。音声合成部ではこれらの情報を用いて最終的な出力音声を生成し、ユーザへの応答とする。

このようにして構築した対話システムを用いて、複数の被験者により実際の対話実験を行う。この際の対話シナリオについては本研究では代表的なタスクを設定し、それに応じたシナリオを作成しておく。対話実験により各要素技術において適切な処理が行われているかを確認し、問題点があればそれぞれ改善を検討する。

４．研究成果

＜平成 27 年度＞

以下の項目について研究成果が得られた。

(1)感情音声対話の評価

音声対話において感情を用いることの妥当性、および感情推定法について検討を行なった。具体的にはまずは言語情報から感情を推定し、その感情に対して協調的な応答を返す場合と、ランダムな場合、入力感情とは逆の感情の応答を返した場合について対話の自然性や知的さなどの複数の主観的な評価を行い、協調的な応答が有効であることを示した。

(2)エントロピーによる文選択の拡張

音声合成部において自然性の高い音声を生成するため、従来の音韻バランスだけでなく、音韻と韻律(アクセントや文長) の両方のコンテキストを考慮した拡張エントロピーに基づく文 選 択 法を提案し、評価実験を行なった。その結果、ランダムに文 を 選択 した場合に比べ、有意に提案法のほうが自然性や再現性の高い合成音声が得られることがわかった。

(3)対話意欲推定

対話におけるユーザの状態推定として対話意欲を対象として、実際に対話の収録を行い、得られた音声について分析・推定を行なった。その結果、ある程度対話意欲の程度に顕著な差がある場 合 に ついては、対話意欲の度合いを音響的特徴などから精度よく推定できることがわかった。

(4)大規模感情音声コーパスの構築準備

感 情 を 伴 う音声対話における感情音声合成・感情音声認識・感情推定に利用可能な大規模な感情音声コーパスを構築するため、エントロピー を考慮した音韻・韻律バランス感情依存文の設計と予備的な評価を行なった。その後音声の収録を実際に行い、予備的な分析を行ない音声の特性について調査した。

＜平成 28 年度＞

以下の項目について研究成果が得られた。

(1)大規模感情音声データベースの構築

提案する「人にやさしい」対話システムにおいて任意の話者の感情音声やそれに含まれる感情を高い精度で認識するため、また感情豊かな音声を合成するために、前年度に引き続き感情音声データベースの構築を行い男女各 50 名、計 100 名について 4 感情（平常、喜び、怒り、悲しみ）の各 50 文、計 20,000 発話の収録を完了し、これを JTES(Japanese Twitter-based Emotional Speech)と名付けた。またこれらのデータの一部を用いて感情認識および感情音声合成においてデータベースの評価を行い、話者を増やすことで性能が向上することを確認した。

(2)音声収集のための Web ブラウザによる収録環境の構築

今後さらにデータベースを拡充するため、クラウドソーシングによる音声収集のための Web ブラウザによる収録環境を構築に取り組んだ。

(3)ベクトル量子化に基づく DNN 音声合成手法

音声合成部についてはスペクトル特徴量のベクトル量子化に基づく DNN 音声合成手法を提案し、主観品質が改善することを示した。

(4)音韻・韻律バランスコーパスの設計

聞き手にやさしい音声合成のために、話し言葉音声合成のための Web 上のテキストデータを用いた音韻・韻律バランスコーパスの設計についても検討した。

(5)アクセント結合規則の改良

自然なアクセントによる音声合成を行うため、日本語テキスト音声合成のためのアクセント辞典に基づくアクセント結合規則の改良を行った。

(6)雑談対話用例文の収集

音声対話システムによる雑談対話用例文の収集と人手 DB との比較を行い、有効性を示した。

(7)Kaldi インタフェースの構築

話し言葉音声を高い精度で認識するため、深層学習に基づく音声認識ツールキット Kaldi を、幅広く利用されている音声認識エンジン Julius 互換にするためのインタフェースを開発した。

＜平成 29 年度＞

以下の項目について研究成果が得られた。

(1)対話意欲の推定

感情情報に基づくボトルネック特 徴 量 を用いた対話意欲の推定に関する検討を行なった。具体的には、DNN を用いた識別器によりボトルネック特徴量を求め、これを用いることで対話意欲の推定精度が上がることを示した。

(2)DNN-HMM 音響モデル適応

感情音声データベース JTES を用いた感情音声認識における DNN-HMM 音響モデル適応の検討を行なった。具体的には感情混合モデルとしてコーパス適応を行うことにより、認識精度が改善することを示した。

(3)段階的口調付与による印象変化の検討

音声対話システムにおける段階的口調付与による印象変化の検討を行なった。具体的には、口調をデスマス体から非デスマス

体へと段階的に変化させること で、ユーザの対話システムの利用における対話の満足度が向上することを 3 日間連続した主観評価実験により示した。

(4)CRF によるアクセント結合推定の改善

CRF によるアクセント結合推定のための素性の改善に関する検討を行なった。具体的にはこれまでに行なった改良規則を用いて CRF における素性の改善を行い、これによりアクセント結合後のアクセント型の推定精度が向上することを実験により示した。

(5)音声入力による韻律制御

差分特徴量に基づく DNN 音声合成における音声入力による韻律制御について検討した。具体的には、音声入 力によって合成音声のピッチをユーザが実現し、自分の好みの韻律を容易に実現することを可能とした。

(6)複数話者データを用いた学習手法の比較

DNN 音声合成における複数話者データを用いた学習手 法の比較を行なった。

(7)ユーザの印象を向上させる対話システム

相互自己開示によりユーザの印 象 を 向 上させる対話システムの検討を行なった。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 11 件）

① HMM-Based Photo-Realistic Talking Face Synthesis Using Facial Expression Parameter Mapping with Deep Neural Networks, Kazuki Sato, Takashi Nose, Akinori Ito, Journal of Computer and Communications, 査読有, Vol. 5, No. 10, 2017, pp. 55-65

②Dimensional paralinguistic information control based on multiple-regression HSMM for spontaneous dialogue speech synthesis with robust parameter estimation, Tomohiro Nagata, Hiroki Mori, Takashi Nose, Speech Communication, 査読有, vol. 88, 2017, pp. 138-148

③統計モデルに基づく多様な音声の合成技術, 能勢隆, 電子情報通信学会論文誌 D, 査読有, vol. J100-D, no. 4, 2017, pp. 556-569

④Sentence Selection Based on Extended Entropy Using Phonetic and Prosodic Contexts for Statistical Parametric Speech Synthesis, Takashi Nose, Yusuke Arao, Takao Kobayashi, Komei Sugiura, Yoshinori Shiga, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 査読有, Vol. 25, Issue 5, 2017, pp.1107-1116

⑤クロスリンガル音声合成のための共有決定木コンテクストクラスタリングを用いた話者適応, 長濱大樹, 能勢隆, 郡山知樹, 小林隆夫, 電子情報通信学会論文誌 D, 査読有, vol. J100-D, no. 3, 2017, pp. 385-393

⑥Cluster-based approach to discriminate the user ’s state whether a user is embarrassed or thinking to an answer to a prompt, Yuya Chiba, Takashi Nose, Akinori Ito, Journal on Multimodal User Interfaces, 査読有, vol. 11, No. 2, 2017, pp. 185-196

⑦ Prosodically rich speech synthesis interface using limited data of celebrity voice, Takashi Nose, Taiki Kamei, Journal of Computer and Communications, 査読有, vol. 4, no. 16, 2016, pp. 79-94

⑧DNN を利用した Animation Unit の変換に基づく顔画像変換の検討, 齋藤優貴, 能勢隆, 伊藤彰則, 電子情報通信学会論文誌, 査読有, Vol. J199-D, no. 11, 2016, pp. 1112-1115

⑨ Efficient Implementation of Global Variance Compensation for Parametric Speech Synthesis, Takashi Nose, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 査読有, Vol. 24, Issue 10, 2016, pp. 1694-1704

⑩ Estimating the User's State before Exchanging Utterances Using Intermediate Acoustic Features for Spoken Dialog Systems, Yuya Chiba, Takashi Nose, Masashi Ito, Akinori Ito, IAENG International Journal of Computer Science, 査読有, Vol. 43, no. 1, 2016, pp. 1-9

⑪発話状態推定に基づく協調的感情音声合成による音声対話システムの評価, 加瀬嵩人, 能勢隆, 千葉祐弥, 伊藤彰則, 電子情報通信学会論文誌 A, 査読有, Vol. J199-A, no. 1,

2016, pp. 25-35

〔学会発表〕（計 73 件）

① Analysis of Efficient Multimodal Features for Estimating User's Willingness to Talk: Comparison of Human-Machine and Human-Human Dialog, Yuya Chiba, Takashi Nose, Akinori Ito, Proceeding of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, **査読有**, 2017, pp. 1-4

② Development and Evaluation of Julius-Compatible Interface for Kaldi ASR, Yusuke Yamada, Takashi Nose, Yuya Chiba, Akinori Ito and Takahiro Shinozaki, Proceeding of the Thirteenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, **査読有**, 2017, Vol.2, pp. 91-96

③Response Selection of Interview-Based Dialog System Using User Focus and Semantic Orientation, Shunsuke Tada, Yuya Chiba, Takashi Nose, Akinori Ito, Proceeding of the Thirteenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, **査読有**, 2017, Vol.2, pp. 84-90

④A Study on 2D Photo-Realistic Facial Animation Generation Using 3D Facial Feature Points and Deep Neural Networks, Kazuki Sato, Takashi Nose, Akira Ito, Yuya Chiba, Akinori Ito, Takahiro Shinozaki, Proceeding of the Thirteenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, **査読有**, 2017, Vol.2, pp. 112-118

⑤ A Study on Tailor-Made Speech Synthesis Based on Deep Neural Networks, Shuhei Yamada, Takashi Nose, and Akinori Ito, Proceeding of the Twelfth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, **査読有**, 2016, vol. 1 pp. 159-166

⑥ Construction and Analysis of Phonetically and Prosodically Balanced Emotional Speech Database, Emika Takeishi, Takashi Nose, Yuya Chiba, Akinori Ito, Proceedings of O-COCOSDA 2016, **査読有**, 2016, pp. 16-21

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

〔その他〕
ホームページ等

６．研究組織

(1)研究代表者
能勢　隆（NOSE, Takashi）
東北大学・大学院工学研究科・准教授
研究者番号：90550591

(2)研究分担者
伊藤　彰則（ITO, Akinori）
東北大学・大学院工学研究科・教授
研究者番号：70232428

千葉　祐弥（CHIBA, Yuya）
東北大学・大学院工学研究科・助教
研究者番号：30780936

(3)連携研究者
森　大毅（MORI, Hiroki）
宇都宮大学・大学院工学研究科・准教授
研究者番号：10302184