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　２９　年　　５　月　２３　日現在

機関番号：　１３９０３
研究種目：　挑戦的萌芽研究
研究期間：　2014～2016
課題番号：　２６５４００８３

研究課題名（和文）「しゃべって」つくる音声インタラクションシステム

研究課題名（英文）A speech interaction system created by "speech"

研究代表者
徳田　恵一（Tokuda, Keiichi）

名古屋工業大学・工学（系）研究科（研究院）・教授

研究者番号：２０２１７４８３

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では，音声インタフェースのコンテンツ制作において，コンテンツ製作者が「しゃべる」ことにより，その音声情報を利用してコンテンツを制作するインタフェースの構築方法について検討した．コンテンツ製作者は実際にしゃべることで意図した音高や話速を再現した音声インタフェースのコンテンツを制作することが可能となる．実験結果から，本研究の提案インタフェースを用いることでより魅力的な生き生きとした音声インタフェースを構築することが可能となった．

研究成果の概要（英文）：This work proposes a content creation interface using information extracted from speech spoken by content creators. In the proposed interface, various voice characteristics, e.g. pitch, speaking rate, are extracted from speech, and the characteristics are affected to contents of speech interface. Experimental results clearly show that the proposed method can produce attractive speech interface.

研究分野：　音声情報処理

キーワード：　音声合成　音声認識　音声対話　音声インタフェース

１．研究開始当初の背景

我々はこれまでに，多種多様な音声を容易に合成できる音声合成方式として，「隠れマルコフモデル（Hidden Markov Model；HMM）に基づく音声合成」方式を提案してきた[1]．また，音声インタラクションシステム構築ツールキット“MMDAgent”を開発・公開し，双方向音声案内デジタルサイネージを名古屋工業大学正門前に常設設置して運用するなど，音声インタフェース（音声対話システム）に関する研究に取り組んできた．しかし，音声インタフェース固有の「魅力」である生き生きとしたインタラクティブなやり取りの実現に関しては未だ十分とは言えない．その原因の一つとして，音声インタフェースにおけるコンテンツ（ここでは認識するキーワードや応答発話内容，ジェスチャー等音声インタフェースの対話内容を総じて「コンテンツ」と呼ぶ）の制作におけるインタフェースが挙げられる．人間は声の大きさや高さ，話速などによって，感情や強調などの様々な情報をやり取りすることが可能だが，これらの情報を全てコンテンツに組み込むことは通常のコンテンツ制作インタフェースなどでは容易ではない．そこで，コンテンツ制作者が所望の情報を含むような音声を「しゃべる」ことによって，その発話内容と音声に含まれる情報を直感的にコンテンツに組み込むことができるようなコンテンツ制作インタフェースを構築することによって，生き生きとした音声インタフェースを構築することが可能となるとの考えに至った．

２．研究の目的

本研究では，音声インタフェースのコンテンツを「しゃべる」ことによって制作することを可能にするコンテンツ制作インタフェースを構築するための方法論（方式やインタフェースを含む）を確立する．また，音声から声の大きさや話速，感情，強調などの様々な情報を適切に獲得する手法について検討し，それらの情報が音声インタフェースの「魅力」にどのような影響を与えるかについて，提案手法の有効性を検証する．

３．研究の方法

本研究の目的を達成するために，我々は以下の２つの課題に取り組んだ．

(1) 音声からの様々な情報の獲得

(2)獲得情報のコンテンツへの反映

音声から獲得可能である情報としては，声の大きさ，声の高さ，話速，また，これらを総合的に取り扱うことで獲得可能となる韻律，感情，強調，ニュアンス等が考えられる．音声からコンテンツ制作者の意図を汲み取るためにはどの特徴量を用いるかを選択する必要がある．本研究では特に声の高さと話速に注目し，これらの情報をコンテンツへと反映する方法について検討を行った．コンテンツへの反映については特に音声合成による合成音声への反映方法について検討を進めた．

４．研究成果

(1) 音声対話システムにおける音声入力を用いたコンテンツ制作インタフェース

本研究では，通常のテキスト音声合成機能を保持しつつ，音声入力を用いてHMM音声合成システムのパラメータを操作することにより，ユーザの細かな意図を反映させた応答音声の合成を行う．音声入力を行うだけであれば，音声対話システムやHMM音声合成システムについての知識を必要としないため，多くのコンテンツ制作者が利用可能であると考えられる．

① 入力音声を用いた音声合成

音声を分析することで音高を表す対数基本周波数及びその動的特徴量が得ることができる．また，音声認識技術を応用することで，入力音声の発話内容，音素境界，状態境界を得ることができる．これらの特徴量を応答音声へ反映させることにより，よりユーザの表現した応答文の生成に近づくことが期待できる．

対数基本周波数および継続長の情報を合成音声に反映させる手法として，HMM音声合成における音声パラメータ生成アルゴリズムにこれらの情報を利用することを考えた．入力音声の音高と音声対話システムの合成音声の音高は，性別が異なるなど，大きく異なる場合が考えられる．この時，入力音声の音高そのものを再現した場合には，音声対話システムが様々な音高で対話を行うこととなり不自然なものとなると考えられる．そこで，入力音声の対数基本周波数そのものではなく，入力音声の平均的な対数基本周波数からどの程度高いか・低いかといった音高の変動を表す抽象化された情報を利用することとした．抽象化された対数基本周波数情報と継続長情報を用いて音声合成システムのパラメータを修正することで，入力音声の話速や音高をユーザの意図した応答音声の制作を可能とした．

② 音声入力を用いたコンテンツ制作インタフェースの作成

先述した音声合成手法を用いて，音声対話システムにおける音声入力を用いたコンテンツ制作システムの構築を行った．本システムは音声対話システムであるMMDAgentにおける対話文の登録での運用を想定して構築を行った．また，専門的な知識のないユーザでも容易に扱えるシステムの構築が必要であると考え，グラフィカルユーザーインターフェース（Graphical User Interface；GUI）を採用した．本システムでは，音声対話システムにおけるキーワードと応答文の登録を行うことができる．キーワードと応答文を登録することで，MMDAgent上のエージ

ェントがキーワードを認識するとキーワードに応じた応答文を発話する．応答文の登録時には音声入力を用いた応答音声の生成が可能である．GUI 上で音声入力を行い，入力した音声を分析し応答音声へ反映する．この機能により，ユーザは音声に関する専門的な知識を持たずとも容易に表現したい音声の生成が可能となる．そして，入力した音声の特徴を反映した応答音声と元の応答音声を MMDAgent 上のエージェントに発話させることができる．この機能により，コンテンツ制作者は応答音声を確認しながらコンテンツを制作することが可能となる．

③ 評価実験

コンテンツの表現力向上，及びユーザが魅力を感じるようなコンテンツの制作に対して，音声入力を用いたコンテンツ制作インタフェースが有用であるか検証するため主観評価実験を行った．今回の実験では被験者に音声対話システムである MMDAgent に対して，作成したコンテンツ制作インタフェースを用いてコンテンツの制作を行ってもらった．被験者に対してコンテンツ制作を行った後にコンテンツ制作インタフェースと合成音声について 5 段階評価と自由記述欄によるアンケート調査を行った．合成音声の評価は従来の方法で生成された通常音声と音声入力を用いて補正した補正音声の比較を行った．

コンテンツ制作インタフェースに対する評価の平均を図 1 に，合成音声に対する評価の平均を図 2 に示す．コンテンツ制作インタフェースに対する評価は全体的に高い評価が得られた．特にコンテンツ制作の楽しさに対する評価は高く，音声入力を用いた音声応答生成はコンテンツ制作において有用であると考えられる．合成音声においては自然性，音声の明確さについては有意な差が見られなかったが，作りたい音声が作れたかという項目では補正音声が通常音声と比べ高い評価が得られた．このことから，音声入力を用いた応答音声生成は音声の自然性，明確さを劣化させることなく，コンテンツ制作者が表現したい音声をより適切に制作することができるといえる．

(2) 深層学習に基づく音声合成における入力音声からの情報獲得と合成音声への反映

統計モデルに基づく音声合成手法においては，HMM を利用した手法が広く利用されてきたが，近年，ディープニューラルネットワーク（Deep Neural Network; DNN）に代表される深層学習を利用した手法が提案された[2]．深層学習に基づく音声合成手法は高い性能を示しており，注目を集めている．そこで，我々は深層学習に基づく音声合成における入力音声からの情報獲得と合成音声への反映についても検討した．

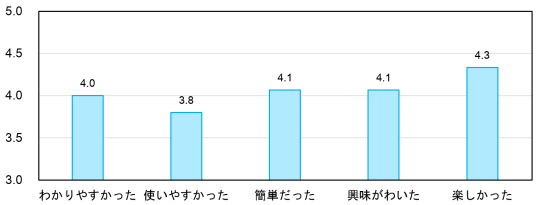

図 1 コンテンツ制作インタフェースに対する評価

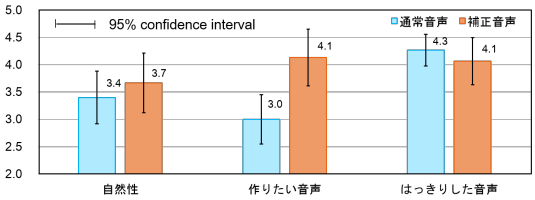

図 2 合成音声に対する評価

① オーディオブックを用いた表現豊かな音声合成

様々な話者や発話表現・発話スタイルを実現する音声合成システムを構築するためには，様々な話者性や感情などの発話スタイルを含む音声データを大量に用意する必要がある．そこで，本研究ではオーディオブックを音声データとして利用することを考える．オーディオブックは音声とそれに対応したテキストを大量に入手できるため，音声合成システムの学習データとして適している．さらに，オーディオブックでは，同じ登場人物や感情であっても様々な表現が含まれており，表現豊かな音声合成システムの構築に適していると考えられる．

オーディオブックには非常に多様な発話表現が含まれている．本研究では様々な発話スタイルの違いを DNN の入力特徴として表すことを検討する．あらかじめ人手で発話スタイルを割り当てるのは高いコストを必要とするため，本研究では，表現豊かな音声を合成するために有用な入力特徴を自動的に推定する手法について検討した．

② フレーズコードを用いた DNN に基づく音声合成

本研究では，音声データの各フレーズに対して乱数コードを割り当て，DNN の入力特徴量として利用する．それぞれのフレーズは異なる固有のフレーズコードを持ち，フレーズコードによってその音声データの発話表現や発話スタイルを表しているものとする．音声合成時にはフレーズコードを設定することで，音声データに現れた発話表現・発話スタイルを再現した合成音声を生成することが期待できる．フレーズコードを用いた DNN の学習は，音声データから抽象化された音声の情報をモデル化しているといえ，フレーズコードを推定することは入力音声の抽象化された音声情報を獲得しているといえる．さらに，抽象化された音声情報をフレーズコードとして入力することで，表現豊かな合成音

声を生成することが可能であり，音声インタフェースのコンテンツ制作においても利用できると考えられる．

③ 評価実験

提案法の有効性を評価するために，主観評価実験を行った．音声合成システムのための学習コーパスとして，英語の子供向けのオーディオブック 833 ページ(255 分)を用いた．またテストデータとして，学習データに含まれない子供向けのブック 348 ページを使用した．また，フレーズコードには 320 次元の乱数を用いた．

主観評価実験は，発話スタイルの表現性に関する XAB テストを行った．XAB テストでは，まず目標音声を被験者に聞かせ，続いて合成音声のペアを聞かせて，どちらが再現性に優れているかを選択させる．ここではフレーズコードを用いた DNN 音声合成システムとフレーズコードを用いていない従来の DNN 音声合成システムを比較した．フレーズコードは学習用音声データから選択したフレーズコードを利用した．実験結果から，フレーズコードを用いたシステムが 83.3%選択され，有効性を示した．フレーズコードを用いなかったシステムは合成音声が平坦な音声となったが，フレーズコードを用いたシステムは合成音声がフレーズコードの発話スタイルを表現した音声となった．この結果より，フレーズコードを入力に用いることで，指定した発話スタイルの表現が可能だといえる．

(3) 得られた成果の位置づけ，インパクト，今後の展望

本研究では音声インタフェースにおけるコンテンツの表現力向上，及びユーザが魅力を感じるようなコンテンツの制作を実現するために，音声入力を用いたコンテンツ制作インタフェースの研究開発に取り組んだ．実験結果から，提案インタフェースによってコンテンツ制作者が意図する，より表現豊かなコンテンツの制作を可能にした．製作者が実際に「しゃべる」ことでコンテンツ制作を行う本手法は，製作者の意図を反映させるためには有効な手法であり，今後のコンテンツ制作において重要な枠組みとなることが期待できる．また，本研究で試作したコンテンツ制作インタフェースは限定的にではあるが一般ユーザに公開し，高い評価を得ると同時に，多数の意見を収集することができた．今後は収集した意見を反映することで，コンテンツ制作インタフェースとしてさらなる改善に取り組む．

さらに，本研究では，音声データから抽象化された情報を獲得すること，獲得した情報を合成音声へと反映させる手法として，深層学習に基づく音声合成手法にも取り組んだ．深層学習に基づく音声合成手法は，本研究の期間中に急速な発展をとげた手法ではあるが，そのようの最先端の手法を取り込むことで，新たな枠組みへと展開することができた．本手法により，より多様な表現の実現と同時に合成音声の品質改善を実現することができた．今後は上記のコンテンツ制作インタフェースと組み合わせることで，コンテンツ製作者の意図をより反映したコンテンツ制作の実現に取り組む．

＜引用文献＞

① K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden Markov models,” Proceedings of the IEEE, vol. 101, no. 5, pp. 1234-1252, 2013.

② H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” ICASSP 2013, pp. 7962-7966, 2013.

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 5 件）

① K. Nakamura, K. Hashimoto, Y. Nankaku, and K. Tokuda, “Integration of spectral feature extraction and modeling for HMM-based speech synthesis,” IEICE Transactions on Information and Systems, vol. E97-D, no. 6, pp. 1438-1448, June, 2014.
(DOI: 10.1587/transinf.E97.D.1438)

② S. Takaki, Y. Nankaku, and K. Tokuda, “Contextual additive structure for HMM-based speech synthesis,” Selected Topics in Signal Processing, IEEE Journal, vol. 8, issue 2, pp. 229-238, April, 2014.
(DOI: 10.1109/JSTSP.2014.2305919)

〔学会発表〕（計 68 件）

① K. Sawada, C. Asai, K. Hashimoto, K. Oura, and K. Tokuda, “The NITech text-to-speech system for the Blizzard Challenge 2016,” Blizzard Challenge 2016 Workshop, Cupertino, USA, September, 2016.

② K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Trajectory training considering global variance for speech synthesis based on neural networks,” ICASSP 2016, pp. 5600-5604, Shanghai, China, March, 2016.

③ T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Simultaneous optimization of multiple tree structures for factor analyzed HMM-based speech

synthesis," Interspeech 2015, pp. 1196-1200, Dresden, Germany, September, 2015.

④ K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," ICASSP 2015, pp. 4455-4459, Brisbane, Australia, April, 2015.

〔図書〕（計 2 件）

① 山岸順一，徳田恵一，戸田智基，みわよしこ，"おしゃべりなコンピュータ ～音声合成技術の現在と未来～，" 情報研シリーズ 19，丸善ライブラリー，2015 年 4 月.

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

○ソフトウェア

音声対話システム構築ツールキット MMDAgent
http://www.mmdagent.jp/

HMM 音声合成ツールキット HTS
http://hts.sp.nitech.ac.jp/

音声信号処理ツールキット SPTK
http://sp-tk.sourceforge.net/

HMM 音声合成エンジン hts_engine API
http://hts-engine.sourceforge.net/

日本語テキスト音声合成システム Open JTalk
http://open-jtalk.sourceforge.net/

６．研究組織

(1)研究代表者

徳田 恵一（TOKUDA, Keiichi）
名古屋工業大学・工学(系)研究科・教授
研究者番号：20217483

(2)研究分担者

李 晃伸（LEE, Akinobu）
名古屋工業大学・工学(系)研究科・教授
研究者番号：80332766

南角 吉彦（NANKAKU, Yoshihiko）
名古屋工業大学・工学(系)研究科・准教授
研究者番号：80397497

山本 大介（YAMAMOTO, Daisuke）
名古屋工業大学・工学(系)研究科・准教授
研究者番号：00402470

(3)連携研究者

(4)研究協力者

大浦 圭一郎（OURA, Keiichiro）

橋本 佳（HASHIMOTO, Kei）