様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月　 ６ 日現在

機関番号：１３９０３
研究種目：基盤研究(B)（一般）
研究期間：2013～2015
課題番号：２５２８００６１

研究課題名（和文）超巨大データに基づくユニバーサル音声モデル構築のための技術的・社会的基盤の確立

研究課題名（英文）aaa

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）　　13,100,000円

研究成果の概要（和文）：本研究では、音声データを永続的に収集・蓄積・共有・維持し続ける社会的基盤の確立を目指し、音声収録ツールの開発や音声データに対する適切なライセンスの設計により、より多くの音声データを収集できるような枠組みを提案した。また、多様な声質を表現可能な音声モデルの開発や整備されていない音声データから音声合成システムを構築する枠組みを提案し、様々な発話表現を実現可能な音声合成システムの構築に取り組んだ。

研究成果の概要（英文）：In this work, we proposed a framework that can collecting speech data from a lot of people by developing a tool for speech recording and an appropriate license for recorded speech to build social infrastructure that can continue to collect recorded speech data. Additionally, we proposed acoustic models that can generate speech with various characteristics and a framework to construct speech synthesis systems from speech database including various errors.

研究分野：音声情報処理

キーワード：音声合成　超巨大データ　音声モデル

１．研究開始当初の背景

音声は人間にとって最も基本的なコミュニケーションメディアであるということから、近年、音声インターフェースの普及やデジタルメディア作品の増加に従って、様々な場面で音声合成が利用され始めている。しかし、音声が重要な位置を占めるコンテンツに関しては、未だプロフェッショナルによる実音声が使われているのが現状である。また、音声案内システムや音声翻訳システム等において音声合成を用いたシステムが登場しているものの、未だ広く普及しているとは言い難い。この原因の一つに、多様性の不足があげられる。実際、大半の音声合成システムは固定の発話スタイルしか出力することができず、感情表現を含む人間の音声の多種多様性とは比較にならないほど表現能力が低い。この問題に対し、これまでに合成音声の品質を向上させるための手法だけでなく、多様な話者性を実現する手法、言語を超えて話者性を再現する手法、感情表現を可能とする手法など、様々な理論や手法が提案されてきた。個々の手法により合成音声の表現の幅は格段に広がってはいるものの、あらゆる話者性や発話スタイル、感情表現を自在に実現できるユニバーサルな音声モデルの構築には至っていない。これは、人間の音声の多様性を十分に表現可能な大量の音声データを集積することができていないことが本質的な原因であると考えられる。

２．研究の目的

音声合成技術が次のステージに向かうブレークスルーのためには、多種多様な音声を表現できるユニバーサル音声モデル構築のための技術的な基盤、そしてそれと呼応しあう音声データ集積のための社会的な基盤が必須であり、その二つが相互作用し発展し続ける必要がある。そこで本研究では「(1) 超巨大音声データベースを用いたユニバーサル音声モデル構築のための技術的基盤の確立」と、「(2) 音声データを永続的に収集・蓄積・共有・維持し続ける社会的基盤の確立」の２つを目的とする。

研究の第１段階は音声合成における肉声感の向上、多言語化、多様な話者性や発話スタイル、感情表現を実現する表現能力の向上だけでなく、超巨大データを取り扱うことが可能な技術的基盤を確立することである。これらは本研究目標の基礎にあたる部分であり、様々な課題を解決しながら研究を進めなければならない。並行して、より多くの音声データを収集・蓄積し続けるための社会的基盤について検討・設計する。これまでに医療分野、エンターテイメント分野、商業分野等で音声合成システムの構築が行われてきているが、音声データは個別に収集され、そのデータベースの管理・維持も個別に行われてきた。各分野において適切にインセンティブを設定し、わかりやすい共通のライセンス形態を定義することで、音声データを収集する社会的な仕組みを構築することができると考えられる。また、収集した音声データを用いて合成音声の品質や表現能力を改善することで、社会への還元を行い、さらなる音声データの提供を促すインセンティブとすることができる。このように、音声データの提供、音声合成技術の向上、社会への還元が循環的に発展するような社会的基盤の構築を目指す（図 1）。

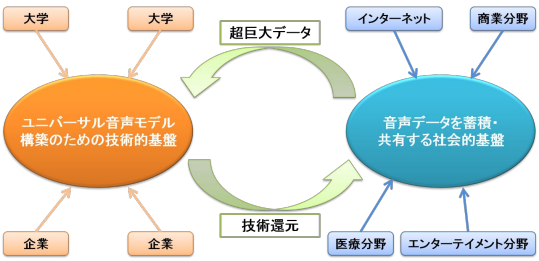

図 1. 技術的・社会的基盤が相互作用する循環系

３．研究の方法

本研究では、「(1) 超巨大音声データベースを用いたユニバーサル音声モデル構築のための技術的基盤の確立」と、「(2) 音声データを永続的に収集・蓄積・共有・維持し続ける社会的基盤の確立」の２つの目的を達成するために、以下の４つの課題に取り組む。

(1) 多種多様な言語、話者性、発話スタイル、感情表現を実現する枠組みの確立

(2) 整理されていない膨大な音声データを自動処理する枠組みの検討

(3) 新たな社会基盤の確立に向けた適切なインセンティブの設計

(4) 様々な分野で共通に使えるわかりやすいライセンス形態の設計と枠組み全体の検証・評価

これらの成果をまとめることにより、最終的に言語・話者・発話スタイル・感情表現を超えて、多様な音声を生成できるフレームワークの構築を行い、研究の成果を検証する。

４．研究成果

(1) 多種多様な言語、話者性、発話スタイル、感情表現を実現する枠組みの確立

あらゆる話者性や発話スタイル、感情表現を自在に実現できるユニバーサルな音声モデルを構築するために、以下の手法に取り組んだ。

① 因子分析に基づく音声モデル

話者や発話スタイル、感情などを低次元の特徴量として抽出可能な因子分析に基づく音声モデルを提案した。本手法では、音声合成システムの構築時に様々な話者、発話スタイル、感情を含む音声データを利用し、それらを低次元の特徴量として抽出しながら音声モデルを構築する。合成音声を生成するときには、低次元特徴量を設定することで様々

な音声表現を実現することができる。より効果的に因子分析に基づく音声モデルを構築するために、テキストと音声の特徴を関連付ける決定木構造の構築方法についても改善し、合成音声の品質が大きく改善することを示した。

② 発音情報未知言語のための音声合成システム構築法

音声合成システムを構築するためには音声データとテキストを用いるが、通常はテキストから「読み」を判別するようなテキスト解析が必要となる。しかし、このようなテキスト解析が定まっていないような言語については、書き文字であるテキストから直接音声を合成することが可能な音声合成システムが必要である。そこで、書き文字から音声合成システムを構築するための枠組みについて検討した。本手法では、発音情報未知言語の音声に対して英語などの異なる言語の音声認識器を用いて読みを付与する。また、テキストから英語の読みを予測する Grapheme-to-Phoneme を構築することで、あらゆるテキストに対して読みを推定可能とする。推定された読みと音声を利用することで、発音情報未知言語の音声合成システムを構築することができる。本手法により、テキスト解析処理が定まっていないような言語についても音声合成システムを構築可能となった。

③ ディープニューラルネットワークに基づく音声合成と声質変換

より高品質な合成音声の生成を目指し、音声合成のための新規理論としてディープニューラルネットワークに基づく音声合成について検討した。テキスト情報から音響特徴量への変換にはこれまで隠れマルコフモデル (HMM) と決定木構造が利用されてきたが、新たにディープニューラルネットワークを用いることでテキスト情報から精度良く音響特徴量へと変換することが可能となった。この時、最終的な出力を考慮した目的関数を設定し、ディープニューラルネットワークを学習することによって、合成音声の品質を大きく改善することを示した。また、声質変換においても同様の枠組みを適用したところ、変換音声の品質の向上、話者性の再現度の向上が確認され、ディープニューラルネットワークを用いた新規手法の有効性を示した。

(2) 整理されていない膨大な音声データを自動処理する枠組みの検討

ここではオーディオブックの音声データを用いて音声合成システムの自動構築について検討を行った。整備されていない大量の音声データを取り扱うにあっては、テキストと発話内容の不一致、ポーズ位置の間違い、言い間違いや言い淀みなどの問題を考慮する必要がある。本課題では、特に言い淀みや言い間違いによるテキストと発話内容の不一致を検出するための枠組みについて検討し、オーディオブックを用いた音声合成システムの自動構築に取り組んだ。

本課題では、テキストと発話内容の不一致を検出するために音声認識技術を利用する。音声認識結果のテキストとオーディオブックのテキストを比較し、テキストの一致率を計算する。音声認識結果が音声データの発話内容を正確に表していると仮定すると、この一致率が低い場合は発話内容とテキストの不一致を含んでいると考えられる。しかし、実際には音声認識結果にも誤りが含まれるため、一致率が高いからといって必ずしも不一致を含んでいないとはいえない。このため、適切な音声データとテキストを取捨選択する必要がある。

まず、一致率がある閾値を超える音声データのみを音声合成システムの学習コーパスとして利用することを考える。閾値を低く設定した場合、学習コーパスの量は大きくなるが発話内容とテキストの不一致が多く含まれる可能性が高い。閾値を高く設定した場合、発話内容とテキストの不一致は少なくなると考えられるが学習コーパスの量は小さくなる。このように、閾値によって学習コーパスの質と量がトレードオフの関係になり、これが音声合成システムの性能にどのように影響を与えるか実験によって評価した。実験結果から、一致率の閾値を 80%とした時に合成音声の品質が最も高くなった。不一致を多く含む場合は音声合成システムの性能に悪影響を与えたと考えられる。また、不一致が少ない場合は学習コーパスの量が小さくなるため、高性能な音声合成システムを構築するには不十分であったと考えられる。

次に、音声認識結果とオーディオブックのテキストを組み合わせることでより不一致の少ないテキストを生成する手法について検討した。オーディオブックのテキストを利用しながら音声認識を行うことによって、より発話にように近いテキストを生成可能となった。このテキストを利用することによって、オーディオブックからより高性能な音声合成システムを自動構築することが可能となった。

(3) 新たな社会基盤の確立に向けた適切なインセンティブの設計

音声データを収集・蓄積・共有・維持するための社会的基盤を構築するためには、様々な分野における音声データ提供に対して適切なインセンティブを設計する必要がある。本課題では、特にエンターテイメント分野に関してインセンティブを設定し、音声データを効率的に蓄積する枠組みについて検討した。

主にエンターテイメント分野向けのインセンティブとして、利用者が収録した音声に基づいて Web 上に音声合成気が作成され、自

由に使うことができる枠組みを設計した（図2）。まず、自分の音声合成気を作ってみたいと考えた人が音声収録ツールをダウンロード・インストールし、音声収録ツールの指示に従って音声収録を行う。収録された音声はサーバにアップロードされ、アップロードされた音声を利用して収録者の声を再現可能な音声合成機が自動構築される。構築された音声号席は、我々が開発したWeb上で動作する日本語テキスト音声合成システム Open JTalk 上で自由に使用できることを想定している。より自然な音声を合成するために、利用者は雑音のない静かな環境で音声収録を行うことや、より多くの分を収録することが期待される。このときサーバに集積された大量の音声データは、多様な音声合成システムの構築にはもちろん、音声認識や話者認識など様々な音声情報処理研究に応用することが可能である。さらに、医療分野やエンターテイメント分野に広く利用され、音声技術の社会普及の一助となることが期待される。

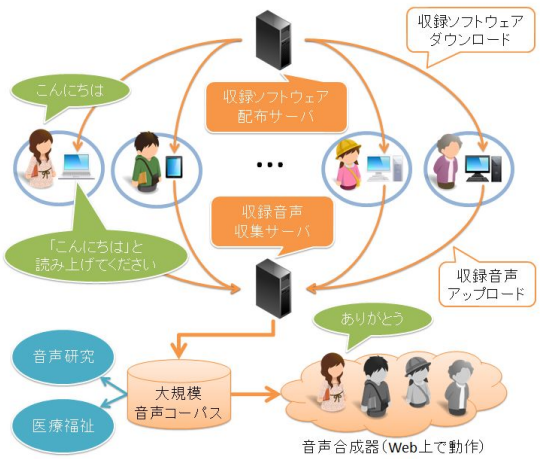

図2．音声合成システムの自動構築に伴う音声収集の枠組み

本枠組みを実現するために、より多くの人がどのような環境であってもストレスなく音声収録を行うことができる音声収録ツールを開発した（図3）。音声収録ツールは、プラットフォームの依存性が低いプログラミング言語（Java）と音声ライブラリ(PortAudio)を利用しており、多くの端末上で動作することを目的として設計した。老若男女問わず利用できるようにするために、起動時には音声収録のガイダンスが自動で開始されるようにした。また、収録テキストのルビ表示や、声量を表すレベルメータなど、初めて収録を行う人でもストレスなく利用できるような工夫を取り込んだ。作成した音声収録ツールは、名古屋工業大学の30人以上の学生および国立情報学研究所が推進する日本語ボイスバンクプロジェクトに参加した100人以上のボランティアに使用してもらい、音声収録に関する多くのフィードバックが得られた。実際の音声収録に基づく知見の獲得と音声収録ツールの改善を繰り返すことで、音声収録に関する注意点や問題点を整理・解決していくことができ、音声収録ツールの使いやすさは開発初期段階のものと比較して格段に向上した。音声収録ツールは今後オープンソース・ソフトウェアとして広く公開していく予定である。

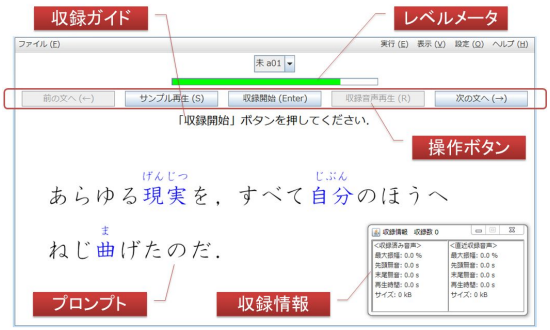

図3．開発した音声収録ツール

(4) 様々な分野で共通に使えるわかりやすいライセンス形態の設計と枠組み全体の検証・評価

まず、クリエイティブコモンズやオープンデータコモンズ等の代表的なライセンス形態と、提供された音声データを共有するために必要な要件について調査した。ライセンス設計においては、商用利用の可能性、本人の声の再現性、匿名性、再配布の可能性など、考慮すべき事柄が多く、このため、ライセンス形態が多岐に渡り、管理が複雑になることが予想される。そこで、既存のライセンス形態の調査結果をもとに、主として非商用向けのライセンス形態の設計に取り組んだ。現在までのところ、ライセンスは日本国内での利用を想定しているが、超巨大データベースを収集可能な基盤を構築するためには、日本語ライセンスの英語化に取り組み、提供された音声データをより広い範囲で共有可能とすることを目指す。

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 7 件）

① Shinji Takaki, Yoshihiko Nankaku, and Keiichi Tokuda, “Spectral modeling with contextual additive structure for HMM-based speech synthesis,” IEEE Journal of Selected Topics in Signal Processing, vol.8 pp.229-238, 2014. (DOI: 10.1109/JSTSP.2014.2305919)

② Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura, “Speech Synthesis Based on Hidden Markov Models,” Proceedings of the IEEE, vol.101, no.5, pp.1234-1252, 2013. (DOI:

10.1109/JPROC.2013.2251852)

〔学会発表〕（計 67 件）

① Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, “Trajectory training considering global variance for speech synthesis based on neural networks,” ICASSP 2016, China, March 2016.

② 吉村建慶，橋本佳，大浦圭一郎，南角吉彦，徳田恵一，“クラウドソーシングによる音声収集のための収録ソフトウェアの設計，” 日本音響学会 2016 年春季研究発表会，神奈川，2016 年 3 月.

③ 沢田慶，伊神和輝，浅井千明，佐藤雄介，橋本佳，大浦圭一郎，南角吉彦，徳田恵一，“統計的パラメトリック音声合成のためのオーディオブックを用いた学習コーパス自動構築，” 日本音響学会 2016 年春季研究発表会，神奈川，2016 年 3 月.

④ 沢田慶，橋本佳，大浦圭一郎，南角吉彦，徳田恵一，“発音情報が未知の言語におけるテキスト音声合成システム構築法の評価，” 第 17 回音声言語シンポジウム，愛知，2015 年 12 月.

〔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ー音声合成技術の現在と未来－，” 丸善ライブラリ，210 ページ，2015.

〔その他〕

ホームページ等

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

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

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

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

(5) 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・大学院工学研究科・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・大学院工学研究科・准教授
研究者番号：80332766

南角 吉彦（NANKAKU, Yoshihiko）
名古屋工業大学・大学院工学研究科・准教授
研究者番号：80332766

戸田 智基（TODA, Tomoki）
名古屋大学・情報基盤センター・教授
研究者番号：90403328

山岸 順一（YAMAGISHI, Junichi）
国立情報学研究所・コンテンツ科学研究系・准教授
研究者番号：70709352

(3)連携研究者

(4)研究者協力者

大浦 圭一郎（OURA, Keiichiro）

橋本 佳（HASHIMOTO, Kei）