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　４ 年　 ６ 月 ２１ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2018～2021
課題番号：１８Ｋ１８１００

研究課題名（和文）多方言音声合成のための地理情報を利用した音韻・アクセントモデリングに関する研究

研究課題名（英文）Pronunciation and accent modeling for multi-dialect speech synthesis

研究代表者
高道　慎之介（Takamichi, Shinnosuke）

東京大学・大学院情報理工学系研究科・助教

研究者番号：９０７８４３３０

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究は，あらゆる日本語方言の音声を人工的に合成することを目的とする．この遂行のために，(1) 一般家庭における収録音声でも頑健に音声合成を構築可能にする技術，(2) 方言の地理情報を使ってアクセントを制御する音声合成方式，(3) アクセントを構築する言語単位を言語知識なしに獲得する方法，(4) 方言アクセントを言語知識なしに獲得する方法，(5) 言語知識なしに方言音声合成を実現する方法，(6) 方言音声合成を実現するためのフリーな音声データベース公開を実施した．

研究成果の学術的意義や社会的意義
本研究は，あらゆる日本語方言の音声を人工的に合成することを目的とする．消滅の危機にある日本語方言について，その特性を計算機的に保存することは，音声言語文化の保存からコンテンツ制作まで幅広い範囲に有用である．本研究はこれに向け，方言の知識なしに方言音声を合成可能な方法について多角的に取り組み，さらに，一般に利用可能な方言データベースを整備した．

研究成果の概要（英文）：The purpose of this research is to artificially synthesize speech in any Japanese dialect. To achieve this goal, we have developed (1) a method that enables robust speech synthesis from noisy recorded speech, (2) a speech synthesis method that controls accents using geographic information of dialects, (3) a method for acquiring linguistic units for constructing accents without linguistic knowledge, (4) a method for acquiring dialectal accents without linguistic knowledge, (5) a method for realizing dialectal speech synthesis without linguistic knowledge, and (6) the release of a free speech database to realize dialectal speech synthesis.

研究分野：知能情報学

キーワード：音声合成　方言　韻律

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

１．研究開始当初の背景

あらゆる音声言語情報(音韻・アクセント情報など)を処理可能な音声言語処理技術は,音声言語文化の保存のみならず,音声コミュニケーションの拡張に重要な技術である.特に,任意のテキストから人工的に音声を合成する音声合成技術は,人間と人工知能の違いや言語・文化の違いを超えた音声コミュニケーションを可能にする技術である.日本語や英語を始めとする主要言語の読み上げにおいて,合成音声の自然性は人間の肉声のそれに迫りつつあった.これには,深層学習に基づく統計的音声合成技術の発展,及び,収録スタジオ等の理想環境で収録された大規模音声コーパスの利用が大きい.

一方,日本語方言の音声合成技術は,消滅方言文化の保存や,地理と時代に伴う方言の変化を計算機的に扱える点で非常に重要な技術だが,その実現には多くの課題がある.特に,当該方言のネイティブ話者が日本共通語の話者に比べ地理的に限定されており,理想収録環境における大規模音声コーパスの収集が,地理的・経済的に非現実的であることは,解決しなければならない大きな課題である.方言音声コーパスは既にいくつか存在するが,そのほとんどは,必要な音声言語情報とその質・量が不足しており,統計的音声合成の利用に耐えうるものではない.これに対し申請者は 2017 年に,一般家庭環境において収録された大規模方言音声コーパスを構築した.これは,クラウドソーシングとウェブ型音声収録プラットフォームを利用したもので,地理情報と方言音声データを紐づけたコーパスである.既存コーパスと比較して非常に大規模であるが,収録環境は理想的なものでないため,その直接的な利用は方言合成音声の品質を著しく劣化させる.

これに対し本研究課題では,一般家庭環境において収録された方言音声から音韻・アクセント情報をモデル化する高品質音声合成の研究開発を行う.その際に,方言の地理情報を用いることで,単一方言のみではなく,多方言の音韻・アクセント情報を一括でモデル化する.これは,地理情報と音声合成を紐づけることで,将来的に,任意の地理情報に対応した方言音声合成を可能にし,方言音声合成を利用した音声コミュニケーションシステムの研究を加速させるためである.

２．研究の目的

日本共通語を対象に統計的音声合成（機械学習と大規模音声コーパスの利用に基づく音声読み上げ）の実用化が進み，実社会の一員として音声コミュニケーションを行う人工知能の期待が大きくなっている．今後は，単なる日本共通語の読み上げに留まらない音声合成システムが期待される．日本語方言の音声合成は，地域産業の活性化，音声言語文化の保存，方言のキャラクタ性の自動生成など，多様な応用を期待できる技術である．しかしながら，日本語方言の音声合成は解決困難な課題を多く持つ．特に，地理的・経済的な理由から，所望の方言の大規模音声コーパスを理想収録環境において収集することは困難である．これに対し本研究では，(1)劣悪環境の収録音声を用いた高品質な統計的音声合成と，(2)地理情報に基づいた音韻・アクセント情報の多方言モデリングの研究を行う.

３．研究の方法

以降の研究は,構築済みの多方言音声データベースをもとに行われる.ただし,必要に応じてデータベースの拡大も行う.

1. **音声・雑音生成過程の同時推定に基づく高品質音声合成アルゴリズム**:一般家庭の収録音声を対象に,雑音生成過程と音声生成過程(音声合成)の同時推定を試みる.まず,空調音などの定常性雑音を対象に,その生成統計量を効率よく推定する敵対的学習アルゴリズムに基づく雑音生成モデルを導入する(2017 年にプロトタイプ版を発表済み).その後,対象を非定常雑音に拡張し,スペクトルの時変高次，統計量保存,時間構造アクティベーションを導入する.次に,音声・雑音生成過程の同時推定に取り組む.単純な雑音加算過程の考慮のみでは,片方の生成過程にもう一方の音成分が混入すると考えられる.そこで本研究では,音源分離の研究において用いられている,深層モデルに基づく音情報事前分布と独立低ランク性を積極的に導入することで,音声合成・雑音生成の同時学習を可能にする.
2. **地理情報を利用した多方言音韻・アクセントモデリング**:音声合成を行うためには,入力テキストの言語情報と,出力音声の音韻・アクセント情報を対応付ける必要がある.日本共通語の場合,各単語とその音韻・アクセント情報を対応付ける辞書データが豊富に存在するため,音声合成が可能となる.しかしながら,日本語方言の場合,そのような辞書データが存在するケースはまれである.また,日本共通語に登場しない方言独自の未知語が存在する.これらの問題に対して,phoneme-/prosody-aware subword embedding に基づく教師無し方言音韻・アクセント推定アルゴリズムを提案する.subword とは,近年,深層学習に基づく機械翻訳の分野で提案された技術であり,単語

と文字の中間的な分割区分により未知語を減少させる枠組みである.このアイディアを方言音声合成に導入し,subword から音韻・アクセント情報を推定する統計モデルについて研究する.また,複数方言の情報を一括でモデリングするため,方言地理情報からその音韻・アクセントを回帰的に推定する統計モデルを構築する.最終的な多方言音声合成は,コーパスに含まれる地理情報のみならず,それ以外の地理情報においてもある程度自然な音韻・アクセントをもつことが望ましい.故に,地理情報に基づいて音声言語情報を制御・補間するアルゴリズムを研究開発する.最も単純な方法として,方言地理学に基づく規則ベースの方言属性情報の付与が考えられる.一方,従来の統計的音声合成における制御アルゴリズムの問題点として,制御パラメータを変更した際に,自然音声と明確に異なる音声を合成してしまう欠点が知られている.これに対し我々は,回帰モデルを使用しない通常の統計的音声合成において敵対的学習の枠組みを導入している.この敵対的学習は,自然音声と合成音声間の分布間距離を最小化する役割を持ち,直感的には,人間の発話しうる範囲での音声パラメータ生成を可能にする.本研究課題では,この枠組みを地理情報による制御アルゴリズムに拡張し,未知の地理情報が与えられた場合においても,自然性を出来るだけ損なわない音韻・アクセント推定を目指す

４．研究成果

本研究課題では，以下の 6 項目に関する研究成果を挙げた.

1. **一般家庭における収録音声でも頑健に音声合成を構築可能にする技術**：一般家庭の収録音声を対象に，雑音生成過程と音声生成過程（音声合成）の同時推定を試みた．まず，空調音などの定常性雑音を対象に，その生成統計量を効率よく推定する敵対的学習アルゴリズムに基づく雑音生成モデルを導入した．本研究成果は，国際会議 APSIPA で発表した.
2. **地理情報を利用した多方言音韻・アクセントモデリング**： 音声合成を行うためには，入力テキストの言語情報と，出力音声の音韻・アクセント情報を対応付ける必要がある．日本共通語の場合，各単語とその音韻・アクセント情報を対応付ける辞書データが豊富に存在するため，音声合成が可能となる．しかしながら，日本語方言の場合，そのような辞書データが存在するケースはまれである．また，日本共通語に登場しない方言独自の未知語が存在する．これらの問題に対して，phoneme-/prosody-aware subword embedding に基づく教師無し方言音韻・アクセント推定アルゴリズムを提案した．subword とは，近年，深層学習に基づく機械翻訳の分野で提案された技術であり，単語と文字の中間的な分割区分により未知語を減少させる枠組みである．このアイディアを方言音声合成に導入し，subword から音韻・アクセント情報を推定する統計モデルについて研究した．また，複数方言の情報を一括でモデリングするため，方言地理情報からその音韻・アクセントを回帰的に推定する統計モデルを提案した．本研究成果は，国際会議 APSIPA で発表した.
3. **アクセントを構築する言語単位を言語知識なしに獲得する方法**：音声のアクセントの教師なし推定のために，音響モデル尤度に基づく subword 分割法と，隠れマルコフモデルに基づく教師なし音声合成法を提案した．2018 年度に提案した教師なし音声合成法は，方言の地理情報を利用することで多様な方言音声の合成を可能にした一方で，その言語単位が言語モデル的に分割されていたため，その品質に限界があった．音響モデル尤度に基づく subword 分割法は，音声合成の言語単位を音響モデル尤度（すなわち，音声特徴量の予測しやすさ）に基づいて決定する方法である．本研究では，深層生成モデルを出力分布として持つ隠れマルコフモデルを立て，その学習法・生成法を定式化し，音声合成の音質改善への貢献を確認した．本研究成果は，学術論文 speech communication にて発表した.
4. **方言アクセントを言語知識なしに獲得する方法**：変分オートエンコーダを利用した潜在変数獲得を試みた．日本語アクセントに関する離散的な潜在変数を教師なしに獲得する方法を提案し，日本語と同様に 4 段階の離散表現が，最も効率的であることを明らかにした.
5. **言語知識なしに方言音声合成を実現する方法**：4 の方法を利用して，当該方言の知識なしに方言音声合成を実現する方法を提案した．encoder-decoder 型のアクセント条件付けと，東京方言アクセント辞書を利用する方法を提案し，それぞれ有効性を確認した．本研究成果は，4 の成果と合わせて国際会議 ISCA SSW にて発表した.
6. **方言音声合成を実現するためのフリーな音声データベース公開**： 方言音声合成のさらなる研究開発には，誰でも利用可能かつ高品質な方言音声データベースの整備が不可欠である．本研究課題では，この整備を実施し，研究代表者ウェブサイトにて大阪弁と熊本弁のデータベースを公開した．各データベースは，https://sites.google.com/site/shinnosuketakamichi/research-topics/jmd_corpus にて公開しており，無償でダウンロード可能である．各データベースは 3 時間の音声データを有している.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Masashi Aso, Shinnosuke Takamichi, Norihiro Takamune, and Hiroshi Saruwatari	４．巻 125
２．論文標題 Acoustic model-based subword tokenization and prosodic-context extraction without language knowledge for text-to-speech synthesis	５．発行年 2021年
３．雑誌名 Speech Communication	６．最初と最後の頁 53--60
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計6件（うち招待講演　1件／うち国際学会　3件）

１．発表者名 湯舟 航耶, 郡山 知樹, 高道 慎之介, 猿渡 洋
２．発表標題 変分オートエンコーダを用いたアクセントの潜在変数表現の検討
３．学会等名 日本音響学会2020年秋季研究発表会講演論文集
４．発表年 2020年

１．発表者名 Masashi Aso, Shinnosuke Takamichi, Norihiro Takamune, Hiroshi Saruwatari
２．発表標題 Subword tokenization based on DNN-based acoustic model for end-to-end prosody generation
３．学会等名 ISCA SSW（国際学会）
４．発表年 2019年

１．発表者名 阿曽 真至, 高道 慎之介, 高宗 典玄, 猿渡 洋
２．発表標題 音響モデル尤度に基づく subword 分割の韻律推定精度における評価
３．学会等名 日本音響学会2020年春季研究発表会講演論文集
４．発表年 2020年

１．発表者名
阿曽 真至，高道 慎之介，高宗 典玄，猿渡 洋

２．発表標題
End-to-end 韻律推定に向けたDNN音響モデルに基づくsubword分割

３．学会等名
日本音響学会2019年秋季研究発表会講演論文集

４．発表年
2019年

１．発表者名
Takanori Akiyama, Shinnosuke Takamichi, Hiroshi Saruwatari

２．発表標題
Prosody-aware subword embedding considering Japanese intonation systems and its application to DNN-based multi-dialect speech synthesis

３．学会等名
APSIPA ASC（国際学会）

４．発表年
2018年

１．発表者名
Masakazu Une, Yuki Saito, Shinnosuke Takamichi, Daichi Kitamura, Ryoichi Miyazaki, Hiroshi Saruwatari

２．発表標題
Generative approach using the noise generation models for DNN-based speech synthesis trained from noisy speech

３．学会等名
APSIPA ASC（招待講演）（国際学会）

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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関