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　５ 年　 ５ 月 ２２ 日現在

機関番号：１２６０１
研究種目：研究活動スタート支援
研究期間：2021～2022
課題番号：２１Ｋ２１３０５

研究課題名（和文）Continual Learning に基づく持続的に学習可能な音声合成

研究課題名（英文）Sustainably Developable Speech Synthesis Based on Continual Learning

研究代表者
齋藤　佑樹（Saito, Yuki）

東京大学・大学院情報理工学系研究科・助教

研究者番号：２０９０７９０１

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：本研究では，継続的に学習可能な音声合成の基盤構築を目的とする．この遂行のために，(1)単一話者のテキスト読み上げドメインにおける音声合成のcontinual learningの基礎検討，(2)(1)を多話者音声合成に拡張するための学習アルゴリズムの設計・評価，(3)(1)を対話音声合成に拡張するための学習アルゴリズムおよびニューラルネットワーク構造の設計・評価，(4)(3)を多ドメイン対話音声合成に拡張するためのデータベース整備を実施した．

研究成果の学術的意義や社会的意義
人間は過去の経験に関連付けることで新たな知識を逐次的・効率的に学習できるが，現状の音声合成は与えられた音声データの高精度な再現を目的とした学習を1度行うのみであり，それにより得られた知識を保持しつつ，新たな環境に適応するための機構を有さない．そこで本研究では，AI が継続的・階層的・追加的に知識を学習するための枠組みである Continual Learning に基づく音声合成の学習法を提案し，既知のデータに対する再現精度を保持しつつ，追加で与えられるデータに対する予測性能も改善可能な音声合成理論を新たに構築した．

研究成果の概要（英文）：The purpose of this research is to build a foundation for continually trainable speech synthesis technologies. To accomplish this purpose, we 1) proposed continual learning (CL) for single-speaker speech synthesis, 2) developed an algorithm to extend 1) to multi-speaker speech synthesis, 3) developed a neural network method to extend 1) to empathetic dialogue speech synthesis, and 4) developed a speech corpus to extend 3) to multi-domain empathetic dialogue speech synthesis.

研究分野：知能情報学

キーワード：音声合成　深層学習　Continual Learning

１．研究開始当初の背景

音声合成は，コンピュータで人間の音声を人工的に生成・変換する技術であり，人間のように音声情報伝達可能な AI の実現を主目的として研究されている．音声合成の社会的な応用例としては，人間と対話可能なロボットや発話障害者支援などが挙げられる．故に，音声合成は，仮想空間と現実空間の融合で人と物が繋がり，より効率的で快適な社会を目指す Society 5.0 を担う次世代 AI を実現するための基盤技術と言える．

実社会は多様なドメインに属する問題の複合系であるため，人間のように複数ドメインの音声を合成可能な単一の AI エージェント が要求される．現状の AI エージェントは，深層学習に基づく音声合成（DNN 音声合成）の進展により，単一言語の読み上げ音声など，ドメインを限定して十分なデータを用意できれば高品質な音声を合成可能になりつつある．しかし，現状の DNN 音声合成は「一度の学習で用いられたドメインの音声データを高精度に再現する技術」に留まり，人間のように複数ドメインの音声を継続的・逐次的に学習する能力を有さないため，過去の学習で得られたドメイン固有の知識の忘却（DNN の破滅的忘却）が生じる．故に，実社会に適応可能な AI エージェントの実現には，ドメイン固有の知識を獲得・蓄積しつつ，破滅的忘却を回避可能な音声合成が必要である．

そこで本研究では，AI が継続的・階層的・追加的に知識を学習するための枠組みである Continual Learning (CL) に基づく音声合成の学習法を提案し，既知のデータに対する再現精度を保持しつつ，追加で与えられるデータに対する予測性能も改善可能な音声合成理論を新たに構築する．

２．研究の目的

本研究の学術的課題は「①破滅的忘却を回避可能な音声合成をいかに実現するか？」と「②音声におけるドメイン固有の知識をどのように獲得するか？」の 2 点である．これらの 2 点に対し，本研究では，CL の枠組みに基づく，継続的・階層的・追加的に学習可能な音声合成の実現を目的とする．具体的には，①類似ドメインデータを用いた音声合成の CL に関する調査と評価，②相違ドメインデータを用いた音声合成の CL への理論拡張の 2 つを実施する．

３．研究の方法

(1) 類似ドメインデータを用いた音声合成の CL に関する調査と評価：CL は強化学習の分野で登場した研究領域であり，AI エージェントが継続的・階層的・追加的に学習可能な枠組みを提唱する．主な技術的課題としては「有限の計算資源を用いた学習過程で得られた知識を可能な限り保持」しつつ「新たなドメインを学習することで，古いドメインと新しいドメインの両方に対する性能を改善」することが挙げられる．これまでの CL の研究は認識問題を対象としたものが多く，音声合成という「時系列データを対象とした回帰問題」に対して既存の枠組みがどれだけ適用可能かは明らかではない．本研究ではまず，学習法（過去のドメインで用いた学習データの再利用や DNN の知識蒸留など）や DNN の構造（ドメインの変化に応じたネットワーク接続の動的変化など）といった観点から既存の CL に関する技術の有効性を単一話者による単一言語の読み上げ音声合成において評価する．

(2) 相違ドメインデータを用いた音声合成の CL への理論拡張：(1)を拡張し，多様なドメインの音声を単一のモデルで効率的に学習可能な音声合成の実現を目指す．具体的には，データが属するドメインの変化点を明示的に与える教師ありの枠組みと，DNN を用いて変化点をデータドリブンで自動的に検知する教師なしの枠組みの 2 つを検討する．提案技術の有効性は主に合成音声の品質に関する主観評価で検証し，読み上げ音声データで学習した音声合成からの他ドメインへの適応性能と，過去のドメインに対する忘却防止性能の改善を目指す．

４．研究成果

本研究課題では，以下の 4 項目に関する研究成果を挙げた．

(1) 単一話者の多ドメイン音声合成のための CL の検討：近年広く用いられている音声合成モデルである FastSpeech 2 ベースの音声合成において，単一話者の多様なドメインのテキスト読み上げ音声が逐次的に与えられる CL を想定した実験を実施した．結果から，(1)破滅的忘却の影響は，合成音声の韻律・スペクトル包絡特徴量の予測において特に顕著であること，(2)リハーサル法 (Rehearsal Method: RM)が破滅的忘却に起因する合成音声の品質劣化を緩和さ

Method	MOS
MTL	**3.33±0.06**
STL	3.22±0.06
CL	3.16±0.06
CL+RM (M = 400 MB)	**3.28±0.06**
CL+KD ($\lambda = 0.05$)	3.19±0.06

図 1 合成音声の自然性に関する Mean Opinion Score (MOS) テストの結果

せることを示した（図 1）．本研究成果は，日本音響学会 2021 年秋季研究発表会で公表した．

(2) (1)の内容を多話者音声合成に拡張するための検討：学習データに含まれる既知話者の音声特徴量の分布と，学習データに含まれない（既知話者の特徴を補間して得られる）未知話者の音声特徴量の分布を近づけるような制約を考慮したマルチタスク(MT)敵対的学習により，未知話者の高品質な音声を合成できる技術を提案した．実験的評価の結果から，提案技術が合成音声の品質を改善する傾向にあることを示した（図 2）．本研究成果は，電子情報通信学会 2022 年 3 月音声研究会と，国際会議 APSIPA ASC 2022 で公表した．

Algorithm	Naturalness MOS	Speaker similarity DMOS
FS2	3.13 ± 0.12	2.38 ± 0.12
GAN	3.38 ± 0.12	2.40 ± 0.13
MT	**3.50 ± 0.12**	**2.48 ± 0.12**

図 2 合成音声の自然性に関する MOS テストと話者類似性に関する Degraded MOS (DMOS) テストの結果

(3) (1)の内容をより実社会に適した対話タスクに適用するためのアルゴリズムの検討：人間のように対話相手の感情に共感して発話スタイルを制御する「共感的対話音声合成」というタスクにおいて，これまでの対話履歴を考慮して音声合成モデルを学習するアルゴリズムと，対話履歴の音声言語情報から発話スタイルに関する文脈情報をデータ駆動で獲得するための注意機構（図 3）を設計した．評価結果から，従来の言語情報のみを考慮する学習アルゴリズムよりも表現力豊かな音声合成が実現できることを確認した．本研究成果は，電子情報通信学会 2022 年 3 月音声研究会と，国際会議 INTERSPEECH2022 で公表し，Google Travel Grants for Students in East Asia を受賞した．

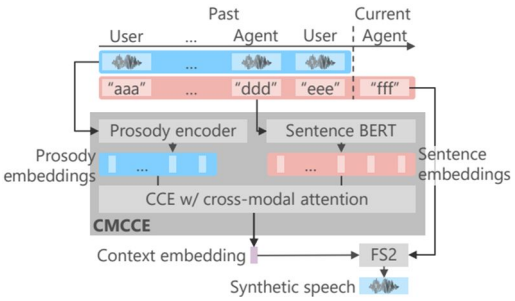

図 3 共感的対話音声合成のための対話履歴を考慮した深層学習モデル

(4) (3)を多ドメイン対話音声合成に拡張するためのデータベース整備：(3)は，共感的対話のドメインとして「個別指導塾での教師と生徒の対話」(図 4) に着目したものである．一方で，人間は図 5 に示すように，対話ドメインに応じて自らの発話スタイルを適切に制御し，対話相手と共感的に対話する能力を有する．本研究ではさらに，「コールセンターでのオペレータと顧客の対話」を考え，カジュアルな対話ドメインとフォーマルな対話ドメインを両方カバーするための共感的対話音声合成コーパスを構築した．本コーパスにより，多ドメイン共感的対話音声合成のための CL 研究の推進が期待される．本コーパスは，非商用利用での研究開発目的であればオンラインで誰でも入手できるようになっている．また，本コーパスを用いた多ドメイン共感的対話音声合成の実験結果から，機械学習による共感的発話の発話スタイル再現性は，対象としているドメインによって差が生じるが，ドメインの差をデータ駆動で学習することで，合成音声の品質劣化傾向に対処できる可能性が示唆された．本研究成果は，電子情報通信学会 2023 年 3 月音声研究会で公表され，国際会議 INTERSPEECH2023 (2023 年 8 月開催)で発表予定である．

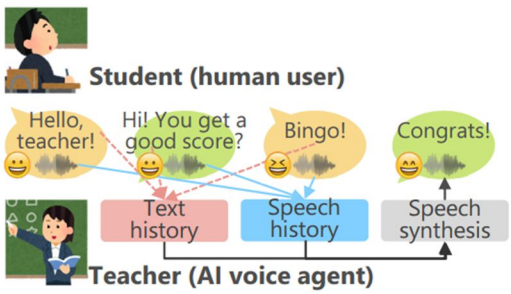

図 4 教師と生徒の共感的対話と，その音声合成の概念図

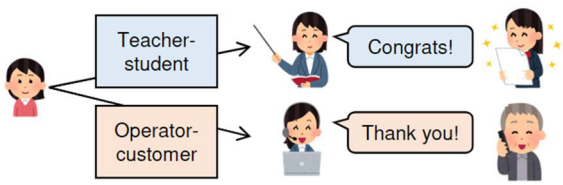

図 5 人間の多ドメイン共感的対話の概念図

以上より，本研究では音声合成に CL を導入するための基本方針と，多ドメイン音声合成のための CL へ拡張するための音声コーパスと機械学習アルゴリズムの基礎が構築された．今後は，自然言語処理における ChatGPT をはじめとした大規模言語モデルに基づく AI チャットボットのように，人間とのインタラクションを通じて対話的・かつ継続的に学習可能な音声合成技術の実現に向けて研究を進めていく予定である．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 齋藤 佑樹, 猿渡 洋
２．発表標題 End-to-End音声合成のContinual Learningにおける破滅的忘却の影響の調査
３．学会等名 日本音響学会 2021年秋季研究発表会
４．発表年 2021年

１．発表者名 仲井 佑友輔, 宇田川 健太, 齋藤 佑樹, 猿渡 洋
２．発表標題 多話者音声合成のためのAdversarial Regularizerを考慮した学習アルゴリズム
３．学会等名 電子情報通信学会 2022年3月音声研究会
４．発表年 2022年

１．発表者名 Yuto Nishimura, Yuki Saito, Shinnosuke Takamichi, Kentaro Tachibana, Hiroshi Saruwatari
２．発表標題 Acoustic Modeling for End-to-End Empathetic Dialogue Speech Synthesis Using Linguistic and Prosodic Contexts of Dialogue History
３．学会等名 INTERSPEECH 2022（国際学会）
４．発表年 2022年

１．発表者名 Yusuke Nakai, Yuki Saito, Kenta Udagawa, Hiroshi Saruwatari
２．発表標題 Multi-Task Adversarial Training Algorithm for Multi-Speaker Neural Text-to-Speech
３．学会等名 APSIPA ASC 2022（国際学会）
４．発表年 2022年

１．発表者名 齋藤 佑樹，飯森 英治，高道 慎之介，橘 健太郎，猿渡 洋
２．発表標題 多ドメイン共感的対話音声合成に向けた音声コーパスの構築
３．学会等名 第9回 音声・音響・信号処理ワークショップ（SPEASIP）
４．発表年 2023年

１．発表者名 Yuki Saito, Eiji Iimori, Shinnosuke Takamichi, Kentaro Tachibana, Hiroshi Saruwatari
２．発表標題 CALLS: Japanese Empathetic Dialogue Speech Corpus of Complaint Handling and Attentive Listening in Customer Center
３．学会等名 INTERSPEECH 2023（国際学会）
４．発表年 2023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