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　３１　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：　2016～2017
課題番号：　１６Ｈ０６６８１

研究課題名（和文）音声制約の自動獲得に基づく高品質音声合成に関する研究

研究課題名（英文）High-quality speech synthesis based on automatically-retrieved speech constraints

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）　　2,300,000円

研究成果の概要（和文）：音声合成技術は，人工的に音声を合成する技術である．合成音声の品質を改善するために，本研究では，敵対的学習を用いた統計的音声合成法を提案した．音質劣化の主要因は，音声パラメータの過剰な平滑化により生じたものである．提案法の学習基準は，通常の基準と敵対基準の重み付き和で得られる．敵対的学習は，自然・合成音声の分布間距離を最小化するため，過剰平滑化を効率的に緩和できる．実験的評価から，(1) 提案法はハイパーパラメータに対して頑健に働くこと，(2) Wasserstein 距離最小化に基づく提案法が，もっとも音質改善効果に有効であること，(3) ボコーダフリー音声合成に展開できることを示す．

研究成果の概要（英文）：A method for speech synthesis incorporating generative adversarial networks (GANs) is proposed. One of the issues causing the quality degradation of speech synthesis is an oversmoothing effect often observed in the generated speech parameters. In the proposed framework incorporating the GANs, the discriminator is trained to distinguish natural and generated speech parameters. Since the objective of the GANs is to minimize the divergence (i.e., distribution difference) between the natural and generated speech parameters, the proposed method effectively alleviates the oversmoothing effect on the generated speech parameters. We evaluated the effectiveness and found that 1) the proposed method can generate more natural spectral parameters regardless of its hyperparameter settings. 2) a Wasserstein GAN minimizing the Earth-Mover's distance works the best in terms of improving the synthetic speech quality, and 3) the method can be extended to the vocoder-free speech synthesis.

研究分野：　音声合成

キーワード：　音声合成　アンチ・スプーフィング　深層学習　話者認証

１．研究開始当初の背景

音声合成技術は，人間・コンピュータ又は人間・人間コミュニケーションへの応用を目論んだ技術であり，本研究ではテキスト音声合成・声質変換を指す．任意のテキストから音声を合成するテキスト音声合成技術は，近い将来，実世界コミュニティの一員として振る舞うであろう人工知能の実装に必要とされ，また，ある音声を別の音声に変換する声質変換技術は，身障等の身体的制約を超えて人間の音声機能の拡張を可能にする．これらの技術では，あたかも人間の自然音声のような高品質音声の合成が要求される．現在の主流である統計的音声合成は，少量の音声データから音声合成器を構築できる一方で，著しく品質の低い音声を合成する．この品質劣化問題に対して，音声に関する制約（発声器官制約に由来する音響的制約）の導入の有効性が知られている．

申請者はこれまでに，変調スペクトル制約に基づく音声合成法を提案した．変調スペクトルは，音声学に基づいて定式化された特徴量であり，直感的に記述すれば，声色や音高の時間的なゆらぎを表す特徴量である．従来の音声合成基準に対してゆらぎの制約を与えることで音質改善が可能であり，2015 年に開催された音声合成の国際コンペティションにおいて，申請者の合成音声は世界最高品質であると評価された．しかしながら，合成音声と自然音声の品質差は未だに大きく，我々は合成音声と自然音声を明確に区別できてしまう．この問題に対し，更なる制約導入の有効性が期待されるとともに，「従来の音声学に基づく制約の不十分さに鑑みた新たなる『真の』確率モデル制約とは何か（それをどのように取得・記述するか）」という本質的な考察が必要とされている

２．研究の目的

本研究では，新たな制約の自動獲得に基づいた高品質音声合成の学習アルゴリズムを確立する．そのために，音声合成による「声のなりすまし」を検出するアンチ・スプーフィング技術（図 3）を利用する．アンチ・スプーフィングでは，音声学に基づく特徴量や機械学習に基づいて自動獲得される特徴量を用いて，自然音声と合成音声を識別する．このアンチ・スプーフィングを音声合成のための新たな制約とすることにより，従来の音声学的特徴量を利用しつつ自動獲得された新たな特徴量（機械学習的特徴量）を考慮可能であるため，合成音声の品質改善効果が期待できる．

３．研究の方法

上記の背景及びこれまでの研究成果をもとに，本研究は下記の 3 点を実施する．

1．学習アルゴリズムの確立

本アルゴリズムでは，①自然音声と合成音声を識別するアンチ・スプーフィングと，②アンチ・スプーフィングに敵対する音声合成を同時学習する．両者は敵対する関係にあるため，最終的な合成音声の品質は，学習パラメータに強く依存すると予想される．まず，この学習パラメータに対する知見を獲得するとともに，基本的な学習アルゴリズムを確立する．

2．音質改善に有効な音声学的特徴量の選択

本アルゴリズムは，アンチ・スプーフィングにおいて有効な音声学的特徴量を音声合成の学習に利用可能にする．しかしながら，その特徴量が合成音声の品質改善に有効である保証はない．そこで，音声学的特徴量を考慮した合成音声を評価することで，新たな音声学的特徴量を獲得する．

3．特徴量の自動獲得と学習アルゴリズムへの導入

1 と 2 では音声合成の性能を直接的に改善・評価するが，本節はアンチ・スプーフィングを改善することで間接的に音声合成の性能を改善する．アンチ・スプーフィングモデルによる自己符号化等を用いて特徴量を自動的に獲得した後，学習アルゴリズムに導入し品質改善効果を検証する．

４．研究成果

1．学習アルゴリズムの確立

研究実施期間の序盤にて，学習アルゴリズムの確立を行った．音声合成の学習基準は，通常の最小二乗誤差基準と，アンチ・スプーフィングに敵対する基準の重み付き和で構成される．本研究では，このアンチ・スプーフィングの項の導入により，通常の最小二乗誤差基準のみと比較して合成音声の音質改善を達成した．また，重み付き和における重みは，学習のハイパーパラメータであり，音声品質を左右する重要な要素である．これに対し，種々の重みにおける主観評価実験を行い，合成音声に対する音質改善効果は，この重みに対して頑健に得られることを明らかにした．

2．音質改善に有効な音声学的特徴量の選択

アンチ・スプーフィングにおける有効な特徴量を，1 で提案したアルゴリズムに導入した．具体的には，アンチ・スプーフィングにおいてより識別に有効な音声の時間動的特徴量を導入し，音質への影響を調査した．その結果，アンチ・スプーフィングに有効な特徴量が，必ずしも，音質改善に有効であるとは限らないことを明らかにした．また，従来の音声合成技術は，ボコーダと呼ばれる音声波形生成処理を行うため，アンチ・スプーフィングでは，ボコーダによるアーティファクトを検出することで，容易に識別可能であった．そこで本研究では，ボコーダを使用せず，かつ，1 で確立したアルゴリズムを導入した

方法として，ボコーダフリー敵対的音声合成アルゴリズムを確立した．これにより，ボコーダの使用により生じていた識別の容易さを緩和させ，音質改善に成功した．

3．特徴量の自動獲得と学習アルゴリズムへの導入

当初の予定では，特徴量の自動獲得法として，自己符号化法などの利用を検討していた．しかしながら，2 において「より識別に有効な特徴量が，より音質改善に有効であるとは限らない」との結論が得られたこと，また，従来の統計的音声信号処理において，敵対距離関数（自然音声と合成音声間の分布間距離）と処理性能の関係性が議論されているため，敵対的学習において最小化される分布間距離に関して議論した．具体的には，提案アルゴリズムを，近似イエンシェンシャノン距離，カルバックライブラー距離，ワッサーステイン距離最小化などに基づく手法に展開し，それぞれの音質を評価した．その結果，ワッサーステイン距離最小化に基づく提案アルゴリズムが，最も音質改善効果に有効であることを示した．

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 7 件）

[1] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 26, No. 1, pp. 84--96, Jan. 2018.

[2] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Voice Conversion Using Input-to-Output Highway Networks," IEICE Transactions on Information and Systems, Vol.E100-D, No.8, pp.1925--1928, Aug. 2017.

[3] Shinnosuke Takamichi, "Modulation spectrum-based speech parameter trajectory smoothing for DNN-based speech synthesis using FFT spectra," Proc. APSIPA ASC, Kuala Lumpur, Malaysia, Dec. 2017.

[4] Shinnosuke Takamichi, Daisuke Saito, Hiroshi Saruwatari, Nobuaki Minematsu, "The UTokyo speech synthesis system for Blizzard Challenge 2017," Proc. Blizzard Challenge Workshop, Stockholm, Sweden, Aug. 2017.

[5] Hiroyuki Miyoshi, Yuki Saito, Shinnosukea Takamichi, Hiroshi Saruwatari, "Voice Conversion Using Sequence-to-Sequence Learning of Context Posterior Probabilities," Proc. INTERSPEECH, pp. 1268--1272, Stockholm, Sweden, Aug. 2017.

[6] Shinnosuke Takamichi, Tomoki Koriyama, Hiroshi Saruwatari, "Sampling-based speech parameter generation using moment-matching network," Proc. INTERSPEECH, pp. 3961--3965, Stockholm, Sweden, Aug. 2017.

[7] Yuki Saito, Shinnosuke Takamichi, Hiroshi Saruwatari, "Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis," Proc. ICASSP, pp. 4900--4904, New Orleans, U.S.A., Mar. 2017.

〔学会発表〕（計 12 件）

[1] 宇根 昌和，齋藤 佑樹，高道 慎之介，北村 大地，宮崎 亮一，猿渡 洋，"雑音環境下音声を用いた音声合成のための雑音生成モデルの敵対的学習，"情報処理学会研究報告，2017-SLP-118，no. 1，pp. 1--6，Sep. 2017.

[2] 三好 裕之，齋藤 佑樹，高道 慎之介，猿渡 洋，"コンテキスト事後確率の Sequence-to-Sequence 学習を用いた音声変換と Dual Learning の評価，" 電子情報通信学会技術研究報告，SP2017-17，vol. 117，no. 160，pp. 9--14，Jun. 2017.

[3] 高道 慎之介，郡山 知樹，猿渡 洋，"Moment-matching network に基づく音声合成における音声パラメータのランダム生成，" 情報処理学会研究報告，2017-MUS-115，no.15，pp. 1--6，Jun.，2017.

[4] 齋藤 佑樹，高道 慎之介，猿渡 洋，”DNN テキスト音声合成のための Anti-spoofing に敵対する学習アルゴリズム，” 情報処理学会研究報告，2017-SLP-115，pp. 1--6，Feb.，2017.

[5] 齋藤 佑樹，高道 慎之介，猿渡 洋，"Anti--spoofing に敵対する DNN 音声変換の評価，" 電子情報通信学会技術研究報告，SP2016-69，vol. 116，no. 414，Jan. 2017.

[6] 齋藤 佑樹，高道 慎之介，猿渡 洋，"多重周波数解像度の STFT スペクトルを用いた敵対的 DNN音声合成，" 日本音響学会 2018 年春季研究発表会講演論文集，3-8-14，2018.

[7] 齋藤 佑樹，高道 慎之介，猿渡 洋，"敵対的DNN音声合成におけるダイバージェンスの影響の調査，" 日本音響学会 2017 年秋季研究発表会講演論文集，1-8-7，2017.

[8] 高道 慎之介，郡山 知樹，猿渡 洋，"Moment matching network を用いた音声

パラメータのランダム生成の検討," 日本音響学会 2017 年春季研究発表会講演論文集, 2-6-9, 2017.
[9] 三好 裕之, 齋藤 佑樹, 高道 慎之介, 猿渡 洋, "コンテキスト事後確率の Sequence-to-Sequence 学習を用いた音声変換," 日本音響学会 2017 年春季研究発表会講演論文集, 1-6-15, 2017.
[10] 齋藤 佑樹, 高道 慎之介, 猿渡 洋, "敵対的 DNN 音声合成における F0・継続長の生成," 日本音響学会 2017 年春季研究発表会講演論文集, 2-6-6, 2017.
[11] 齋藤 佑樹, 高道 慎之介, 猿渡 洋, "Highway network を用いた差分スペクトル法に基づく敵対的 DNN 音声変換," 日本音響学会 2017 年春季研究発表会講演論文集, 1-6-14, 2017.
[12] 齋藤 佑樹, 高道 慎之介, 猿渡 洋, "DNN 音声合成のための Anti-Spoofing を考慮した学習アルゴリズム," 本音響学会 2016 年秋季研究発表会講演論文集, 3-5-1, 2016

〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
1.
https://sites.google.com/site/shinnosuketakamichi/publication
2.
http://sython.org/demo/icassp2017advtts/demo.html
3.
http://sython.org/demo/sp201701advvc/demo.html

６．研究組織
(1)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
(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）