

令和元年6月18日現在

機関番号：62615

研究種目：若手研究(B)

研究期間：2016～2018

課題番号：16K16096

研究課題名（和文）テキスト音声合成のためのニューラルネットワークに基づく波形ダイレクトモデリング

研究課題名（英文）Direct modeling of speech waveform using a DNN for text-to-speech synthesis

研究代表者

高木 信二（Takaki, Shinji）

国立情報学研究所・コンテンツ科学研究系・特任助教

研究者番号：50735090

交付決定額（研究期間全体）：（直接経費） 3,000,000円

研究成果の概要（和文）：本課題では、従来のテキスト音声合成に含まれるヒューリスティックに用いられてきた処理を取り除き、Deep Neural Networkを用いた音声波形のダイレクトモデリング手法に基づくテキスト音声合成の実現を目的とする。ヒューリスティックな処理を除いた単純な窓掛とフーリエ変換を用いて得られたスペクトルのモデル化、位相情報も含めたスペクトルのモデル化、スペクトル誤差を用いた音声波形のモデル学習を検討し、音声波形のダイレクトモデリング手法を実現した。

研究成果の学術的意義や社会的意義

音声インターフェースの核となる技術であるテキスト音声合成の性能改善のため、Deep Neural Networkを用いた音声波形モデリングが盛んに研究されている。本課題では、非常に注目されているこの研究トピックについて取り組み、テキスト音声合成の性能改善を行った。テキスト音声合成を用いる既存のシステムの性能改善、性能改善に伴う応用アプリの普及等多くの波及効果を期待できる。

研究成果の概要（英文）：The purpose of this work is to realize text-to-speech synthesis based on direct modeling of speech waveform using a deep neural network. In this work, we exclude heuristic processing included in conventional text-to-speech synthesis. Modeling of amplitude spectra obtained by utilizing simple windowing and Fourier transform, modeling of spectra including phase information and direct modeling of speech waveform were investigated. We realized a direct modeling method of speech waveform for text-to-speech synthesis.

研究分野：音声情報処理

キーワード：音声合成 DNN

## 様式 C - 19、F - 19 - 1、Z - 19、CK - 19 (共通)

### 1. 研究開始当初の背景

テキスト音声合成システムの性能は飛躍的に向上しており、iPhone の Siri のようなアプリや音声対話システム等に実用化され始めている。これらの音声合成器は統計的アプローチに基づき構築することが主流となっており、盛んに研究が行われている。その中でも統計モデルとして Deep Neural Network (DNN) を用いた手法は飛躍的な性能向上が多数報告されていることから国際的に注目を集めている。

多くのテキスト音声合成器はボコーダと呼ばれる音声の分析（音響特徴量の抽出）・波形の再合成器に基づいており、分析により得られた音響特徴と言語特徴の対応関係を表現する統計モデル（以降、音響モデルと呼ぶ）を学習する。これら音声合成器により合成された音声と人間の音声の間には自然性に知覚可能な違いが依然として存在し、その原因としてボコーダの利用が挙げられる。ボコーダに基づく音声合成器では、音響特徴量を高精度に予測できたとしても、ボコーダで仮定している音源、スペクトルの分離可能性を利用する限り、音声合成器の性能向上に限界が存在する。そこで抜本的な解決として、このようなヒューリスティックな処理や仮定が取り除かれた、ボコーダを用いない統計的手法に基づく音声合成器を提案する。

### 2. 研究の目的

本研究では従来の音声合成器に含まれるヒューリスティックに用いられていた処理や仮定を取り除き、DNN を用いることでより高精度な音声合成器の実現可能であることを明らかにし、最終的に音声波形を統計的アプローチに基づき直接モデル化することを目的とする。本研究では、大きく分け以下の3つの課題に取り組む。

#### (1) スペクトル包絡のモデル化

音声合成器では音声分析器によりスペクトル包絡を抽出し、得られたスペクトル包絡を音響モデル構築に用いることが多い。正確かつ安定したスペクトル包絡を推定することは音声分析器の目的の一つであるが、知見に基づくヒューリスティックな処理を数多く含む。DNN を用いることで、音声合成のためのスペクトル包絡の高精度モデル化を実現する。

#### (2) 位相情報のモデル化

音声情報処理分野では、スペクトルにおける位相情報は利用されないことが多く、音声合成においても通常利用されない。ここでは、従来の DNN を拡張し複素数を扱うことが可能な複素 DNN の音声合成への適用を行い、これまでヒューリスティックに破棄されてきた位相情報も統計的アプローチに基づきモデル化する。

#### (3) 音声波形のダイレクトモデリング

音声波形のダイレクトモデリングを行う。上記スペクトルモデリングを踏まえ DNN の学習について検討を行い、ボコーダを用いない音声合成器の実現を目指す。

### 3. 研究の方法

理論整備とモデル学習アルゴリズムの導出、および、計算機による評価実験を繰り返すことで、目的に記載した3つの課題に取り組む。評価実験により明らかになった点をフィードバックすることでテキスト音声合成のさらなる改善手法を検討する。

### 4. 研究成果

目的に記載した3つの課題について、以下に主な成果を記述する。

#### (1) スペクトル包絡のモデル化

音声合成器のためのより高精度なスペクトル包絡モデル化を検討した。ヒューリスティックな処理を除いた、単純な窓掛とフーリエ変換を用いて得られたスペクトルを用い、高精度なスペクトルのモデル化を実現した。実現された高精度スペクトルモデルを用いることで、ボコーダを用いない音声合成器構築の検討も行った。主観評価実験の結果から、ボコーダを用いたテキスト音声合成器と比較し、提案手法による合成音声の品質の向上を示した。

#### (2) 位相情報のモデル化

振幅スペクトルだけでなく位相情報も含めた複素スペクトルのモデル化を検討した。複素スペクトルを直接扱うことが可能な複素ニューラルネットワークの一種である複素 RBM を提案した。提案モデルを用いた複素スペクトルからの特徴量抽出の検討を行った。分析再合成実験を行いボコーダと比較し、提案手法の有効性を確認した。

#### (3) 音声波形のダイレクトモデリング

DNN を用いた音声波形のダイレクトモデリングを提案した。音声の振幅スペクトル・位相スペクトルの情報を有効活用するため、スペクトル領域での誤差を用いた DNN の学習法を提案した。テキスト音声合成実験を行い、提案手法により、さらなる音声合成の品質向上を示した。

### 5. 主な発表論文等

〔雑誌論文〕(計3件)

[1] Toru Nakashika, Shinji Takaki, Junichi Yamagishi, "Complex-Valued Restricted Boltzmann Machine for Speaker-Dependent Speech Parameterization from Complex Spectra," IEEE/ACM Transactions on Audio, Speech and Language Processing, 27(2), pp.244--254, Oct 2018.

[2] Xin Wang, Shinji Takaki, Junichi Yamagishi, ``Investigating Very Deep Highway Networks for Parametric Speech Synthesis,' ' Speech Communication, vol.96, pp.1--9 Feb 2018.

[3] Xin Wang, Shinji Takaki, Junichi Yamagishi, ``Investigation of Using Continuous Representation of Various Linguistic Units in Neural Network based Text-to-Speech Synthesis,' ' IEICE Transactions on Information and Systems, vol.E99-D, no.10, pp.2471--2480, Oct 2016.

〔学会発表〕(計 12 件)

[1] 高木 信二, 亀岡 弘和, 山岸 順一, ``CWT スペクトル誤差に基づく DNN 音声波形モデルの学習', ' 音声研究会, 信学技報, 118(497), pp.131--135, Mar 2019.

[2] 高木 信二, 中鹿 亘, 山岸 順一, ``スペクトル系列誤差に基づく DNN 音声波形モデルの学習', ' 日本音響学会秋季研究発表会, pp.1131--1132, Sep 2018.

[3] 高木 信二, ``ディープラーニングによるテキスト音声合成の進展', ' 日本音響学会春季研究発表会, pp.1497--1498, Mar 2018.

[4] Takuhiro Kaneko, Shinji Takaki, Hirokazu Kameoka, Junichi Yamagishi, ``Generative Adversarial Network-based Postfilter for STFT Spectrograms,' ' INTERSPEECH, pp.3389--3393, Aug 2017.

[5] Toru Nakashika, Shinji Takaki, Junichi Yamagishi, ``Complex-valued restricted Boltzmann machine for direct learning of frequency spectra,' ' INTERSPEECH, pp.4021--4025, Aug 2017.

[6] Shinji Takaki, Hirokazu Kameoka, Junichi Yamagishi, ``Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis,' ' INTERSPEECH, pp.1128--1132, Aug 2017.

[7] Xin Wang, Shinji Takaki, Junichi Yamagishi, ``An Autoregressive Recurrent Mixture Density Network for Parametric Speech Synthesis,' ' IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), SP-L4.1, pp.4895--4899, Mar 2017.

[8] 高木 信二, ``とても Deep なテキスト音声合成', ' 音声研究会, 信学技報, 116(414), pp.41--46, Jan 2017.

[9] 高木 信二, SangJin Kim, 亀岡 弘和, 山岸 順一, ``DNN に基づくテキスト音声合成のための FFT スペクトルを用いた位相復元に基づく音声波形生成', ' 第 114 回音声言語情報処理研究会 (SIG-SLP) SLP-21, Jul 2016.

[10] Hieu Thi Luong, 高木 信二, SangJin Kim, 山岸 順一, ``DNN に基づくテキスト音声合成における話者・ジェンダー・年齢コード利用の検討', ' 音声研究会, 信学技報, 116(279), pp.37--42, Oct 2016.

[11] Wang Xin, 高木 信二, 山岸 順一, ``巨大特定話者データを用いた HMM・DNN・RNN に基づく音声合成システムの性能評価', ' 第 112 回音声言語情報処理研究会 (SIG-SLP) SLP-2, Jul 2016.

[12] Shinji Takaki, SangJin Kim, Junichi Yamagishi, ``Speaker Adaptation of Various Components in Deep Neural Network based Speech Synthesis,' ' 9th Speech Synthesis Workshop (SSW9), pp.167--173, Sep 2016.

〔図書〕(計 0 件)

〔産業財産権〕

出願状況 (計 0 件)

名称：  
発明者：  
権利者：  
種類：  
番号：  
出願年：  
国内外の別：

取得状況 (計 0 件)

名称：  
発明者：  
権利者：  
種類：  
番号：  
取得年：

国内外の別：

〔その他〕  
ホームページ等

## 6．研究組織

### (1)研究分担者

研究分担者氏名：

ローマ字氏名：

所属研究機関名：

部局名：

職名：

研究者番号（8桁）：

### (2)研究協力者

研究協力者氏名：

ローマ字氏名：

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、研究者個人に帰属されます。