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 ２０ 日現在

機関番号：１３９０３
研究種目：基盤研究(B)（一般）
研究期間：2019～2021
課題番号：１９Ｈ０４１３６

研究課題名（和文）階層化生成モデルとマルチタスク深層学習の融合に基づく次世代音声合成技術

研究課題名（英文）A next-generation speech synthesis technology based on integration of hierarchical generative models and multi-task deep learning

研究代表者

南角　吉彦（Nankaku, Yoshihiko）

名古屋工業大学・工学（系）研究科（研究院）・准教授

研究者番号：８０３９７４９７

交付決定額（研究期間全体）：（直接経費）　　13,400,000円

研究成果の概要（和文）：近年の深層学習に基づく音声合成は、入力（テキスト）から出力（音声波形）への変換を直接的にモデル化するEnd-to-Endアプローチによって、非常に高品質な音声を合成可能である。その一方で、次のような３つの問題点があった。１．従来手法に比べ学習に大量のデータが必要、２．直感的に理解しやすい音声特徴に基づいた合成音声のコントロールが困難、３．入力と出力を繋ぐ中間的な特徴表現や不完全なデータを利用する枠組みが未確立。本研究では、このようなEnd-to-Endアプローチにおける本質的な問題を解決する次世代音声合成技術を開発し、評価実験によってその有効性を示した。

研究成果の学術的意義や社会的意義
本研究は、統計的音声合成の先駆けとなった隠れマルコフモデルに基づく音声合成手法と近年の深層学習に基づく手法の融合を目指したものであり、これまでの研究分野の知見を活かしつつ、次世代へ発展させるという学術的意義を持った提案となっている。提案した音声合成の枠組みは、高い自然性を維持しつつ、話者性や感情、発話スタイルなどを自由自在にコントロール可能な、より柔軟な音声合成技術の基盤となるものであり、今後、音声対話や音声翻訳などの高度なアプリケーションへの活用によって、より豊かな音声コミュニケーションの実現が期待できる。

研究成果の概要（英文）：Recent deep learning-based speech synthesis techniques can generate very high-quality speech based on the end-to-end approach that directly models the transformation from input (text) to output (speech waveform). On the other hand, it still has three problems: 1) it requires a large amount of training data compared to conventional methods, 2) it is difficult to control the synthesized speech based on intuitively understandable speech features, and 3) a framework for using intermediate feature representations to connect input and output and incomplete data has not yet been established. This study developed a next-generation speech synthesis technology that solves these essential problems in the end-to-end approach and demonstrated its effectiveness through evaluation experiments.

研究分野：音声情報処理

キーワード：音声合成　深層学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年、多くの分野で深層学習が驚異的とも言える成果を上げ、音声合成の分野においても深層学習に基づく手法が盛んに研究されている。特に、入力から出力まですべてを単一のニューラルネットワークで構成し、全体最適化を行う End-to-End 音声合成の研究が注目を集めている。従来の音声合成システムは、言語解析部、音響モデル、波形生成部など複数のモジュールで構成されていたのに対し、End-to-End 音声合成は、テキストと音声波形のセットを大量に用意し、音声合成の問題である「テキストから音声波形を生成する」という機能を単一のニューラルネットワークによって直接的に実現するものである。すべてのモジュールが統一された基準によって全体最適化される点、およびモジュール間における中間表現を明示的に定義することなく、学習によって自動獲得している点が End-to-End アプローチにおける利点と考えられる。高品質な合成音声が得られる一方で、End-to-End 音声合成には大きく以下の３つの問題点があると考えられる。

問題１：従来手法と比べ、大量の学習データが必要

予備的な実験において End-to-End 音声合成の学習には数十時間以上、最低でも１０時間程度の音声データが必要であることが分かっている。一方で、従来の党右傾モデルに基づく音声合成では品質にもよるが１時間程度の音声データでも十分に合成システムの構築が可能であり、End-to-End の音声合成システムが、従来手法に比べて、より大量の音声データを必要であることが確認されている。一般的な音声合成においては、音声認識とは異なり単一話者の音声データを大量に用意する必要があり、さらには、音声データや書き起こしテキストのクオリティが合成音声の品質に強く影響するため、高品質な音声合成器の構築にはそれなりに音声収録にコストをかける必要がある。そのような状況において、数時間と数十時間のデータ量では、収録にかかる労力は大きく異なるため、手法としての利便性は大きく異なると言える。

問題２：直感的に理解しやすい音声特徴に基づいた合成音声のコントロールが困難

End-to-End 音声合成では、内部構造を単一のニューラルネットワークとしてブラックボックス的に学習することにより、構造的制約を与えないことが性能改善につながる一方で、人間にとって意味のある特徴が明示的にモデル化されていないために、それらの特徴をコントロールすることが困難となっている。例えば、声質やイントネーションをユーザの好みに合わせて合成時に若干変化させたい、言語解析のエラーに起因する読み間違いを簡易に修正したい、といった要求がある場合、通常の End-to-End 音声合成ではモデルのどの部分が、どのような役割を担っているかが不明であるため、そのような調整が困難である。End-to-End 音声合成においても複数話者の音声データを用いて単一のニューラルネットワークを学習し、“話者ベクトル”によって話者性をコントロールする方法などが提案されているが、よりプリミティブに音響特徴（スペクトル、基本周波数、継続長）や言語特徴（読み、アクセント、ポーズなど）を操作することは難しくなっている。

問題３：入力と出力を繋ぐ中間的な特徴表現や不完全なデータを利用する枠組みの欠如

従来手法においてモジュール間の中間表現として用いられる音響特徴や言語特徴は、テキストと音声波形の関係をモデル化する際の有用な手掛かりと考えられるが、End-to-End 音声合成では原理的にはそのような中間表現を利用しない。実際には各モジュールに対応するニューラルネットワークを従来法と同様に中間表現を用いて独立に構築し、その後、それらを結合して全体最適化を行うというアプローチもあるが、各モジュールの事前学習において入出力の表現形式を明示的に固定するため、End-to-End アプローチの利点が損なわれる可能性がある。また、通常の End-to-End 音声合成においては、学習データが必ず入力と出力のペアになっている必要があり、音声波形、テキスト、また中間表現も利用する場合は音響特徴、言語特徴も含めた組み合わせのうち、部分的にデータが欠落した不完全なデータを活用する枠組みが明確になっていない。例えば、日本語の音声合成において“かな漢字混じり”の表記を入力とする場合、モデル学習には大量のテキストや言語特徴のデータが必要となり、多くの場合、音声波形と対になっていないデータも活用する必要がある。

２．研究の目的

本研究では従来のデファクトスタンダードであった HMM（隠れマルコフモデル）音声合成と、近年の深層学習に基づく End-to-End 音声合成を融合し、前述した End-to-End 音声合成の問題を解決した次世代音声合成技術の確立を目指す。具体的な核となるアイデアは大きく２つある。

アイデア１：統計的生成モデルに基づくニューラルネットワークの構造化

近年の研究では、音声合成における各モジュールをより高性能なニューラルネットワークに置き換える試みが盛んになされている。このプロセスにおいて、多くの研究は、音声合成全体に対する End-to-End アプローチと同様に、各モジュールを入力から出力への変換器として捉え、データの物理的な意味や生成過程を考慮していない。過剰な構造化はモデル学習における制約となるが、依然としてデータに内在する本質的な構造を捉えることは学習の本質であり、限られたデータから高性能なモデルを学習するためには適切な構造化が必要と考えられる。この問題に対し、本研究では従来の HMM 音声合成において有用性が確認されている統計的生成モデルの

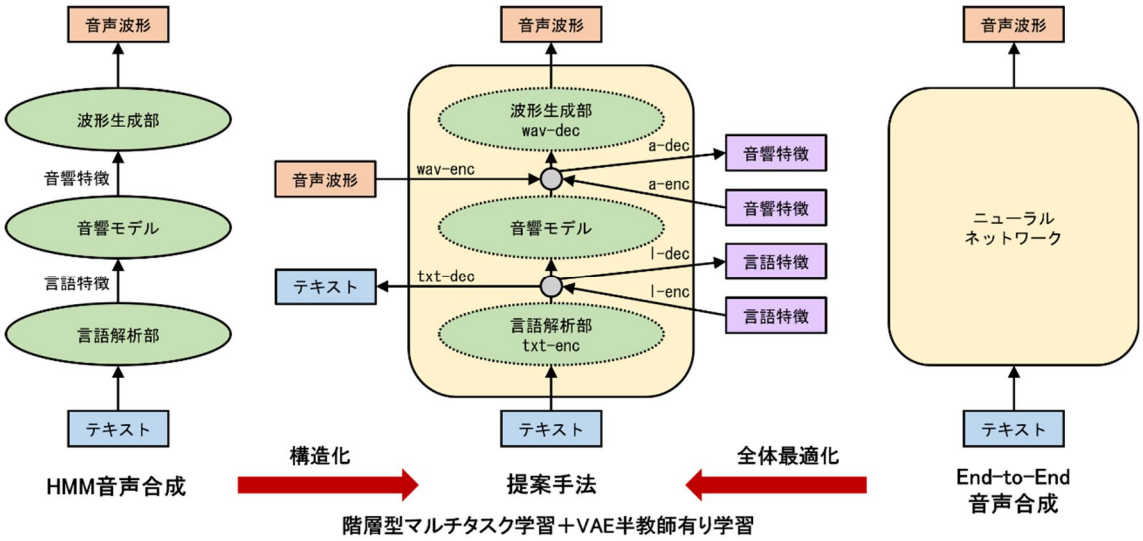

図１：提案する End-to-End 音声合成システムの概要

構造をニューラルネットワークに組み込むことを検討する。データの生成過程を考慮した構造化によって、少量のデータによる効率の良い学習が期待でき（問題１）、さらにはニューラルネットワークの入出力や潜在変数に対し、物理的な意味を与えることができる（問題２）。

音響モデルの構造化：従来の HMM 音声合成では、音響モデルとして統計的生成モデルである隠れセミマルコフモデル（HSMM）によってスペクトル、基本周波数、継続長などの音響特徴のモデル化が行われていた。近年の深層学習に基づく手法では、Attention メカニズム（注意機構）に基づく Sequence-to-Sequence モデルによる置き換えが行われている。HSMM では離散変数によって系列全体に渡る相関を考慮しつつ、明示的なアライメントを計算しているのに対し、Attention 機構では入力と出力の対応関係が局所的な重み（確率）として表現され、時間的な“ねじれ”を許す過剰な自由度を持つモデルとなっている。この問題に対し、本研究では Attention 機構を構造化し、HSMM の構造を内包した Sequence-to-Sequence モデルを提案する。これにより、HSMM と同様に系列データ全体を考慮したアライメント推定が可能であると同時に、構造化によって継続長に関する分布を保持するモデルとなるため、明示的な継続長のコントロールが可能となる。

波形生成部の構造化：波形生成部においても深層学習に基づくボコーダとして、WaveNet と呼ばれる波形を直接出力可能なニューラルネットワークに基づいたボコーダが提案されている。従来のボコーダは、その動作が完全に信号処理として表現されており、学習というフェーズは存在しないが、WaveNet ボコーダでは音響特徴は自己回帰モデルにおける補助特徴量として入力されており、モデルとして音響特徴の物理的な意味は全く考慮していないため、音声波形と音響特徴の関係は学習データから自動推定される。学習データの傾向を再現するという意味で非常に高品質な音声が得られる一方で、学習や合成の計算量が多く、学習データには含まれない範囲外の基本周波数の音声が合成できないなどの問題が観測されている。本研究では、スペクトルや基本周波数、有声・無声情報などを意味づけるモデル構造を与えることにより、これらの問題の解決を図る。

アイデア２: 音声合成のための中間表現を利用した半教師有り学習の枠組み

従来の標準的な HMM 音声合成では、各モジュールを繋ぐ音声特徴や言語特徴などの中間表現が、あらかじめ人間によって定義されていたのに対し、End-to-End 音声合成ではこれらの中間表現が自動獲得される点がアドバンテージになっている。提案手法では、中間的な特徴の有用性と End-to-End アプローチの利点の双方を生かすために、VAE（Variational Auto-Encoder）と階層型マルチタスク学習を統合した半教師有り学習の枠組みを導入する。VAE は、統計的生成モデルをニューラルネットワークにより表現したモデルとなっており、与えられた特徴量に対するコンパクトな表現を潜在変数として自動推定することができる。提案法では、VAE の構造を階層化し、さらにマルチタスク学習を適用することにより、与えられた中間的な特徴量を良く反映しつつ、最終的な波形レベルの再現性を良くするような潜在変数を自動推定することができる。学習時には、与えられた音響特徴や言語特徴を補助的な情報として利用することができ、合成時には所望の特徴を与えて潜在変数を推定することにより、合成音声をコントロールすることが可能となる（問題２）。さらには、テキスト、言語特徴、音響特徴、音声波形の任意の組み合わせのデータを利用してマルチタスクの損失関数を定義することにより、不完全なデータも含めて利用可能なデータを最大限に活用した学習が可能になる（問題３）。

３．研究の方法

本研究では、前述した２つのアイデアの有効性を示すため、音声合成システムを実装し、客観および主観評価を行った。また、評価と理論的検証を繰り返すことにより、性能改善を図った。

４．研究成果

(1) 音響モデルの構造化および階層化生成モデルへの組み込み

音響モデルの構造化として、隠れセミマルコフモデルの構造を内包した Sequence-to-Sequence モデルに基づく音声合成システム（HSMM-ATTN）を構築した。また、階層化生成モデルへの組み込みを考慮した自己回帰潜在変数に基づく音声合成手法（AR-VAE)を提案し、HSMM-ATTN と統合したシステム（ARHSMM-VAE）を構築した。音声合成の評価実験における結果を表１に示す。表中の MCD、F0MSE は、それぞれ、メルケプストラム歪み、基本周波数の平均 2 乗誤差を表し、値が小さいほど自然音声に近いことを表す。また、MOS は被験者による主観評価実験の結果であり、音声の自然性に関する５段階（１～５）評価の平均値を表す。手法 AS は分析合成音であり、本実験における上限を表す。比較手法として、近年提案された音声合成システム Tacotron 2、Fastspeech 2 を用いた。表より、提案法は Tacotron 2、Fastspeech 2 と比較して、より高品質な音声が合成可能であることが示された。また、学習データが少量（0.55hrs.）に場合においても、他手法と比べて音声品質の劣化が少ないことが分かる。提案手法では隠れセミマルコフモデルに基づく明示的な継続長のモデル化により、話速や韻律の制御が可能となる。図２に、実際に音声の継続長を制御した例を示す。全体的な発話速度を変化させるだけでなく、特定の単語の長さを変化させて音声を合成することが可能であることが分かる。

表１：構造化した音響モデルの評価

Models	9.5 hrs.			0.55 hrs.
	MCD	F0RMSE	MOS	MOS
AS（oracle）	—	—	4.25 ± 0.12	4.53 ± 0.11
Fastspeech 2	5.50	**0.222**	3.91 ± 0.15	3.47 ± 0.13
Tacotron 2	5.63	0.238	3.74 ± 0.13	failed
HSMM-ATTN	5.49	0.245	3.83 ± 0.14	2.95 ± 0.14
AR-VAE	**5.11**	0.231	4.07 ± 0.11	3.29 ± 0.13
ARHSMM-VAE	5.16	0.236	**4.15** ± 0.12	**3.51** ± 0.14

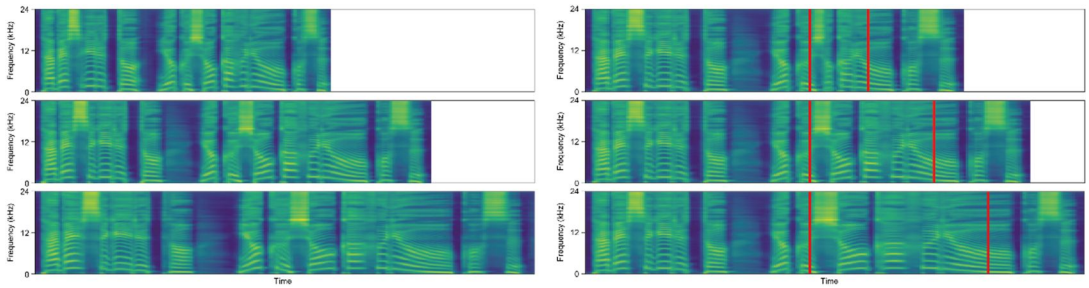

図２：音声継続長の制御例「食べ過ぎるとよく｜消化不良｜を起こす」
（左：発話全体の伸縮、右：特定単語の伸縮）

(2) 波形モデルの構造化

波形モデルの構造化においては、音声の周期・非周期成分に着目したニューラルネットワークの構造化を提案した。図３に提案したモデル構造を示す。提案モデルでは、音声における周期成分と非周期成分を生成するモジュールが個別のニューラルネットワークとしてモデル化される。また、周期成分を生成するニューラルネットワークの入力として、基本周波数に同期した正弦波を入力とすることにより、声の高さに関する制御性の改善が期待できる。図 4 に提案法における合成音声の自然性に関する評価結果を示す。評価は被験者による 5 段階の主観評価を行った。図中の NAT は自然音声を表す。図より、周期・非周期成分を考慮したモデル化によって合成音声の品質が改善していることが分かる。また、図５に音高（基本周波数）を制御した際の自然性に関する主観評価結果を示す。本実験では歌声合成において音高を１オクターブ高くした音声を合

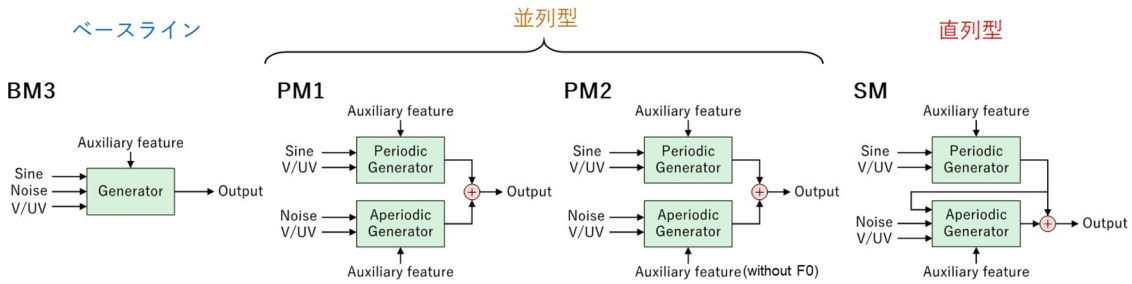

図３：周期・非周期成分を考慮した波形モデルの構造化

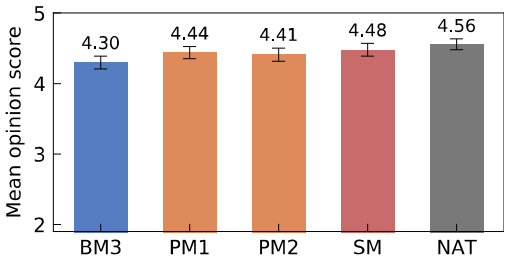

図４：波形モデルの合成音声品質評価

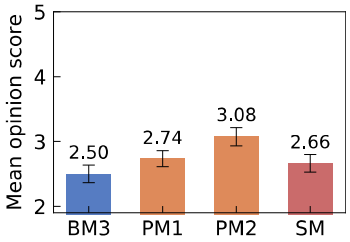

図５：音高制御時の品質評価

成した。ニューラルネットワークの構造化によって基本周波数に関する汎化性能が高まり、学習データに含まれない音高の合成音声が再現できることを確認した。

(3) 階層化生成モデルに基づく半教師あり学習

構築した階層化生成モデルの構造を図６に示す。本実験では、テキスト（文字列）、音素、音声（音響特徴量）の３層の階層構造を用いた。図のように、提案モデルは不完全なデータ（部分的な組み合わせ）に対しも対応したパスを用いて尤度計算が可能であり，任意のペアデータを学習に用いることができる。表２に実験に用いたデータの組み合わせを示す。通常の音声合成モデルの学習に用いられるテキストと音声のペアに加えて、学習に有用と思われる音素の情報および読みを推定するためのテキストと音素のペアを用いた。図７に提案手法と Tacotron 2 によるテキストからの読み（音素列）の推定精度を示す。読みの推定は被験者に合成音声を聞かせ、発声中の発話誤りをカウントした。結果より、提案手法によってテキストからの読み推定の精度が大幅に改善していることが分かる。これは、テキストと音声のペアに加えて、大量の辞書情報（テキストと音素のペア）を利用したことにより、音素列の推定誤りが減少したためと考えられる。図８に、音声の自然性に関する主観評価実験の結果を示す。比較手法として Tacotron 2 を用い、テキストを入力とした場合と音素を入力とした場合の２つのシステムを構築した。提案法は単一のモデルで音素入力もしくはテキスト入力が可能であるため、それぞれの入力において評価を行った。提案手法は音素入力、テキスト入力のいずれの場合も Tacotron 2 より高い性能を示した。また、読みの推定精度が影響しない音素入力の場合においても提案法が Tactron 2 より高い性能を示した。これは大量の辞書情報を利用することによって音声特徴量の生成においても高い汎化性能が得られたためと考えられる。

表 2：使用した不完全な学習データ

Combination of data			Number of training data
Text	Phoneme	Speech	
✔	✔	✔	11,580 utterances
✔		✔	47,498 utterances
✔	✔		378,614 words

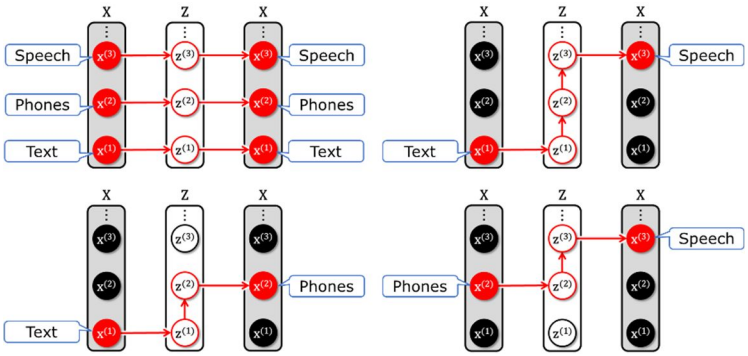

図６：階層化生成モデルに基づく半教師あり学習

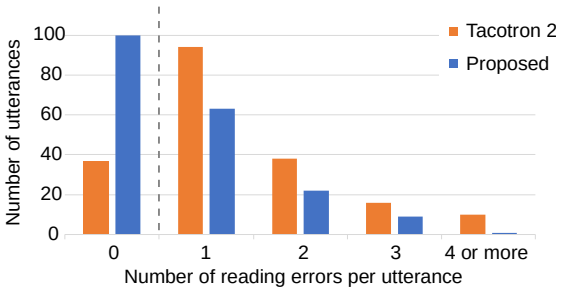

図７：テキストからの読み推定の評価

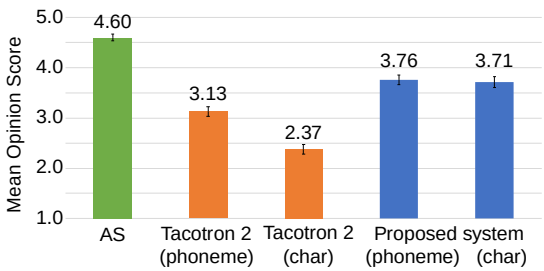

図８：半教師あり学習による合成音声の評価

(4) 今後の展望

本研究では統計的音声合成の先駆けとなった HMM 音声合成と、近年の深層学習に基づく End-to-End 音声合成を融合した次世代音声合成技術を開発した。提案した音声合成の枠組みは、高い自然性を維持しつつ、話者性や感情、発話スタイルなどを自由自在にコントロール可能な、より柔軟な音声合成技術の基盤となるものであり、今後、音声対話や音声翻訳などの高度なアプリケーションへ活用することによって、より豊かな音声コミュニケーションの実現を目指す。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Yukiya Hono, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda	４．巻 29
２．論文標題 Sinsy: A Deep Neural Network-Based Singing Voice Synthesis System	５．発行年 2021年
３．雑誌名 IEEE/ACM Transactions on Audio, Speech and Language Processing	６．最初と最後の頁 2803-2815
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2021.3104165	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 －

１．著者名 Yukiya Hono, Shinji Takaki, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda	４．巻 9
２．論文標題 PeriodNet: A Non-Autoregressive Raw Waveform Generative Model With a Structure Separating Periodic and Aperiodic Components	５．発行年 2021年
３．雑誌名 IEEE Access	６．最初と最後の頁 137599-137612
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2021.3118033	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32件（うち招待講演　2件／うち国際学会　8件）

１．発表者名 Takato Fujimoto, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda
２．発表標題 Autoregressive variational autoencoder with a hidden semi-Markov model-based structured attention for speech synthesis
３．学会等名 2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)（国際学会）
４．発表年 2022年

１．発表者名 法野行哉，高木信二，橋本佳，中村和寛，大浦圭一郎，南角吉彦，徳田恵一
２．発表標題 非周期性指標を考慮したニューラルボコーダの学習
３．学会等名 日本音響学会2022年春季研究発表会
４．発表年 2022年

１．発表者名
藤本崇人，橋本佳，南角吉彦，徳田恵一

２．発表標題
HSMM構造化アテンションに基づく音声合成のためのメモリ削減手法

３．学会等名
日本音響学会2022年春季研究発表会

４．発表年
2022年

１．発表者名
佐々木一匡，吉村建慶，高木信二，橋本佳，南角吉彦，徳田恵一

２．発表標題
声質・声の高さ・話速を変更可能なニューラルボコーダ構成法の検討

３．学会等名
日本音響学会2022年春季研究発表会

４．発表年
2022年

１．発表者名
平光啓祐，橋本佳，南角吉彦，徳田恵一

２．発表標題
深層学習に基づく音声合成における顔画像情報を用いたクロスモーダル話者適応

３．学会等名
日本音響学会2022年春季研究発表会

４．発表年
2022年

１．発表者名
藤本崇人，橋本佳，南角吉彦，徳田恵一

２．発表標題
隠れセミマルコフモデルによる構造化アテンションを用いた自己回帰型VAEに基づくsequence-to-sequence音声合成

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
法野行哉，加藤大誠，橋本佳，大浦圭一郎,南角吉彦，徳田恵一

２．発表標題
発声タイミングのずれを考慮したsequence-to-sequence歌声合成

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
法野行哉，橋本佳，大浦圭一郎,南角吉彦，徳田恵一

２．発表標題
DNN歌声合成のための調子はずれ補正

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
高木信二，牛田光一，橋本佳，南角吉彦，徳田恵一

２．発表標題
因子分析に基づくHSMMを利用した構造化アテンション音声合成

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
Yukiya Hono, Shinji Takaki, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

２．発表標題
PeriodNet: A non-autoregressive waveform generation model with a structure separating periodic and aperiodic components

３．学会等名
2021 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)（国際学会）

４．発表年
2021年

１．発表者名
藤本崇人，橋本佳，南角吉彦，徳田恵一

２．発表標題
学習時と合成時の一貫性を考慮したVAEに基づく自己回帰型sequence-to-sequence音声合成

３．学会等名
日本音響学会2021年春季研究発表会

４．発表年
2021年

１．発表者名
角谷健太，吉村建慶，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
隠れセミマルコフモデルに基づく構造化アテンションを用いたSequence-to-Sequence音声合成

３．学会等名
日本音響学会2021年春季研究発表会

４．発表年
2021年

１．発表者名
法野行哉，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
周期・非周期成分の分離に基づくニューラルボコーダによる音声波形のモデル化の検討

３．学会等名
日本音響学会2021年春季研究発表会

４．発表年
2021年

１．発表者名
岩田康平，高木信二，橋本佳，南角吉彦，徳田恵一

２．発表標題
勾配ブースティング決定木を用いた音声合成手法の検討

３．学会等名
日本音響学会2021年春季研究発表会

４．発表年
2021年

１．発表者名
久野宏彰，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
音声合成における特徴的な発話スタイルの転移学習

３．学会等名
第18回情報学ワークショップ

４．発表年
2020年

１．発表者名
大谷眞史，佐藤優介，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
音声合成における敵対的生成ネットワークを用いた複数言語・複数話者モデリングの検討

３．学会等名
第18回情報学ワークショップ

４．発表年
2020年

１．発表者名
岩田康平，高木信二，橋本佳，南角吉彦，徳田恵一

２．発表標題
勾配ブースティング決定木を用いた高速な音声合成手法の検討

３．学会等名
第18回情報学ワークショップ

４．発表年
2020年

１．発表者名
Yukiya Hono, Kazuna Tsuboi, Kei Sawada, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

２．発表標題
Hierarchical Multi-Grained Generative Model for Expressive Speech Synthesis

３．学会等名
Interspeech 2020（国際学会）

４．発表年
2020年

１．発表者名
藤本崇人，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
感情音声合成のためのDirichlet VAE

３．学会等名
日本音響学会2020年秋季研究発表会

４．発表年
2020年

１．発表者名
法野行哉，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
DNNに基づく音声ボコーダにおける周期・非周期成分のモデル化の検討

３．学会等名
日本音響学会2020年秋季研究発表会

４．発表年
2020年

１．発表者名
大谷眞史，佐藤優介，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
音声合成における敵対的生成ネットワークを用いた複数言語・複数話者モデリング

３．学会等名
日本音響学会2020年秋季研究発表会

４．発表年
2020年

１．発表者名
Takato Fujimoto, Shinji Takaki, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

２．発表標題
Semi-supervised learning based on hierarchical generative models for end-to-end speech synthesis

３．学会等名
2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)（国際学会）

４．発表年
2020年

１．発表者名
藤本崇人，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
End-to-End音声合成のための階層化生成モデルに基づく半教師あり学習

３．学会等名
日本音響学会2020年春季研究発表会

４．発表年
2020年

１．発表者名
大浦圭一郎，中村和寛，橋本佳，南角吉彦，徳田恵一

２．発表標題
周期・非周期信号を用いたDNNに基づくリアルタイム音声ボコーダ

３．学会等名
情報処理学会研究報告

４．発表年
2019年

１．発表者名
村田舜馬，藤本崇人，法野行哉，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一

２．発表標題
楽譜時間情報を用いたアテンション機構に基づく歌声合成の検討

３．学会等名
日本音響学会2019年秋季研究発表会

４．発表年
2019年

１．発表者名
大浦圭一郎，高木信二，中村和寛，橋本佳，南角吉彦，徳田恵一

２．発表標題
周期・非周期信号を用いた敵対的生成ネットワークに基づくリアルタイム音声ボコーダ

３．学会等名
日本音響学会2019年秋季研究発表会

４．発表年
2019年

１．発表者名
Keiichi Tokuda

２．発表標題
Statistical approach to speech synthesis: past, present and future

３．学会等名
Interspeech 2019（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Keiichiro Oura, Kazuhiro Nakamura, Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda

２．発表標題
Deep neural network based real-time speech vocoder with periodic and aperiodic inputs

３．学会等名
10th ISCA Speech Synthesis Workshop (SSW10)（国際学会）

４．発表年
2019年

１．発表者名
Takato Fujimoto, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

２．発表標題
Impacts of input linguistic feature representation on Japanese end-to-end speech synthesis

３．学会等名
10th ISCA Speech Synthesis Workshop (SSW10)（国際学会）

４．発表年
2019年

１．発表者名
Motoki Shimada, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

２．発表標題
Low computational cost speech synthesis based on deep neural networks using hidden semi-Markov model structures

３．学会等名
10th ISCA Speech Synthesis Workshop (SSW10)（国際学会）

４．発表年
2019年

１．発表者名 徳田恵一
２．発表標題 統計的音声合成の進展と展望
３．学会等名 音声研究会（招待講演）
４．発表年 2019年

１．発表者名 和田蒼汰，法野行哉，高木信二，橋本佳，大浦圭一郎，南角吉彦，徳田恵一
２．発表標題 歌声合成におけるニューラルボコーダの比較検討
３．学会等名 音声研究会
４．発表年 2019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
研究分担者	橋本　佳 (Hashimoto Kei) (10635907)	名古屋工業大学・工学（系）研究科（研究院）・准教授 (13903)	
研究分担者	徳田　恵一 (Tokuda Keiichi) (20217483)	名古屋工業大学・工学（系）研究科（研究院）・教授 (13903)	
研究分担者	大浦　圭一郎 (Oura Keiichiro) (20588579)	名古屋工業大学・工学（系）研究科（研究院）・研究員 (13903)	2019～2020

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