一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

# 局所的な句構造の情報を用いたニューラル音声合成

海木 延佳† サクティ サクリアニ†‡ 中村 哲†‡

†奈良先端科学技術大学院大学 〒630-0192 奈良県生駒市高山町 8916-5
‡理化学研究所 革新知能統合研究センター 〒565-0456 大阪府吹田市河田 4-5-6
E-mail: †{kaiki.nobuyoshi.ka1,ssakti,s-nakamura}@is.naist.jp

**あらまし** 自然な韻律をもつ日本語音声を合成するため、局所的な句構造に基づくフレーズ成分を表す韻律記号を end-to-end 音声合成に新たに導入すること提案する．本稿では、フレーズ成分を表現するために、１）句境界に係り受けの深さを表す韻律記号を追加するモデルと、２）韻律生成制御機構に基づき、フレーズ成分とアクセント成分の重畳型モデルを反映させた韻律記号を採用するの２つのモデルを提案する．この２つのモデルを用いた音声合成により，右枝分かれ境界において、１）フレーズ境界を示すポーズが生成されること．２）F0 のフレーズ成分の立て直しが生じることが観察された．アクセント成分のみの韻律記号を用いた従来モデルに対し、これら２つの提案モデルの効果を検証するため対比較の聴取実験を行った．この結果、日本語 end-to-end 音声合成に文の局所的な句境界の情報や、韻律の生成モデルを取り入れることにより、発話者の意図をより正しく反映した自然な韻律を持つ合成音声が生成できることが確認された．

**キーワード** ニューラル end-to-end テキスト音声合成，局所的な句構造，韻律記号

# Neural speech synthesis using local phrase dependency structure information

Nobuyoshi KAIKI† Sakriani SAKTI†‡ and Satoshi NAKAMURA†‡

†Nara Institute of Science and Technology, Takayama-cho, Ikoma, Nara 630–0192, Japan
‡RIKEN AIP, Chuo-ku, Tokyo 103-0027, Japan
E-mail: †{kaiki.nobuyoshi.ka1,ssakti,s-nakamura}@is.naist.jp

**Abstract** In order to synthesize Japanese speech with natural prosody, we introduce an end-to-end TTS with new prosodic symbol representing phrase components based on local phrase dependency structures to end-to-end text-to-speech synthesis (TTS). In this paper, we propose two TTS models: 1) a model with prosodic symbols that represent the depth at phrase boundaries, and 2) a model with prosodic symbols that reflects a folded model of phrase and accent components based on a prosodic generation control mechanism. In synthesized speech at left-branching boundary using these two models, 1) pause indicating the phrase boundary is generated. 2) the re-rebuilding phrase component of F0 may occur. To verify the effect of these two proposed models on a conventional model using prosodic symbols using only accent components, we conducted a subjective evaluation on the AB test. As a result, it was confirmed that by using local phrase boundary information of sentences and prosodic generation model in Japanese end-to-end text-to-speech synthesis, synthetic speech with more natural prosody that reflects the intention of the utterance could be generated.

**Keywords** Neural end-to-end text-to-speech speech synthesis, Local phrase dependency structure, Prosodic symbol

## 1. はじめに

近年、ニューラルネットの研究・開発の進展と共に、それを活用した Text-to-Speech の合成音声の自然性は著しく向上している．特に、テキスト文字列や音素を入力として、直接音声を合成する end-to-end 音声合成が広く研究されている．

End-to-end 音声合成の品質は言語によって異なり、日本語においては、アクセント情報の追加により、自然性が大きく向上することが報告されている[1],[2]．しかし、特に朗読や対話といった合成音声の韻律の自

然性はまだ不充分であると思われる.

これまでに、規則による合成音声の品質向上のため、文の句構造に基づいたいくつかの規則化がなされてきた[3]～[8]. F0 パターンモデルには、局所的なアクセント句による成分（アクセント句成分）と大局的な下降特性が保たれる同一の韻律的なまとまりにわたる成分（話調成分，フレーズ成分）の重畳によって表すものが用いられている．大局的な下降特性は隣接句間の依存関係，いわゆる「係り受け」関係によって表現され,句境界直前の句が直後の句を直接修飾する場合(左枝分かれ境界）は，大局的な下降特性は保たれ，より後方の句を修飾する場合（右枝分かれ境界）は，いわゆる「立て直し」現象が生じる．またポーズ挿入の分析・規則化では、読点が含まれる右枝分かれ境界で長いポーズが挿入されやすいことが示されている[9].

藤本ら[1]は、Tacotron2 にピッチの高低と音素の Onehot ベクトルを同時に入力することにより、合成音声の自然性が向上することを示した．しかしながら、単語の品詞などの韻律情報を含むフルコンテキスト情報の入力は、合成音声の自然性にあまり寄与しないことが報告されている.

栗原ら[10]は、読み仮名のみの Tacotron2[11][12]の入力に比べ、読み仮名とフレーズ区切りを示す記号を含む韻律記号を加えた場合、合成音声の自然性が向上することを示した．しかしフレーズ区切り記号は人手により挿入され、挿入の方法は明示されていない．またフレーズ区切りの韻律記号が韻律に与える効果も明確にされておらず、更なる検証が待たれる.

本稿では、より自然な合成音声を生成するため、文の構文構造に基づく、句境界の係り受けの深さの情報を、日本語 end-to-end 音声合成に適用することを試みた．また、韻律生成制御機構のモデル化を念頭に、性質の異なるアクセント成分とフレーズ成分の２つの成分の韻律制御記号を取り入れるモデルの検討も行った.これら構文の局所的な句構造を表す韻律記号を使った２つのモデルに対し、新たに作成した朗読音声の音声データセットで学習し、得られた合成音声を評価した結果を報告する.

## 2. 提案モデル（句構造を表す韻律記号の追加）

本稿では、栗原ら[2]の提案した表１の韻律記号に対し、表２に示す文の局所的な句構造を表現する韻律記号を追加・修正し、音素記号と共に Tacotron2 の入力とすることを試みた．文の構文木の局所的な句構造を表現するため、アクセント句境界を表す＃の代わりに、アクセント句の句境界の係り受けの深さを示す＃１～＃６を全てのアクセント境界に配置した．＃１は句境界直前の句が直後の句を直接修飾する場合（左枝分かれ境界）を示し、＃２～＃６は、より後方の句を修飾する場合（右枝分かれ境界）を示す．図１に統語的にあいまいな文「警官は走って逃げる泥棒を追いかけた」を入力した場合の想定される２つの構文木と、各モデルに対応する音素記号と韻律記号を示す．構文木１の場合は、「警官は」は後方の「追いかけた」に係るため、「警官は」と「走って」の間の句境界には＃４が配置される．また「走って」は「逃げる」に直接かかるため、「走って」と「逃げる」の間の句境界は＃１が配置される．ただし処理上、係り受けの深さが６以上の場合は＃６として句構造をまとめて表現している.

テキストに読点がある場合、ポーズが挿入されることが多いが、読点があってもポーズが挿入されないことや、読点があっても、係り受けが深い句境界ではポーズが挿入されることもあること報告されている[9].この現象を End-to-end モデルでも表現するため、韻律記号のポーズ“_”の代わりに読点“、”を利用した.本実験では、アクセント情報のみで句構造を表す韻律記号を用いないモデルを従来モデル、句構造の情報を表す韻律記号を追加入力するモデルを提案モデル１と略称する.

表１ 韻律記号（baseline）

| | 韻律記号 |
|---|---|
| アクセントの立ち上り | ^ |
| アクセント核 | ! |
| アクセント句境界 | # |
| 文末（句点） | （ |
| 文末（疑問符） | ? |
| ポーズ | _ |

表２ 句構造の情報を加えた韻律記号（proposed1）

| | 韻律記号 |
|---|---|
| アクセントの立上り | ^ |
| アクセント核（立下り） | ! |
| 係り受けの深さ（アクセント句境界） | #1, #2, #3, #4, #5, #6 |
| 読点 | , |

表３ 韻律生成モデルに基づく韻律記号（proposed2）

| | 韻律記号 |
|---|---|
| アクセント指令の立上り | / |
| アクセント指令の立下り | ¥ |
| フレーズ指令（係り受けの深さ） | #2, #3, #4, #5, #6 |
| 読点 | , |

更に、韻律の生成機構に基づくモデルを提案する（提案モデル２）.表３にこのモデルに利用する韻律記号を示す．アクセント成分は、アクセント指令の立上りを示す制御記号“／”と、立下りを示す制御記号“￥”を用いる．フレーズ成分は、係り受けの深さがフレーズ指令の強さを示すと仮定し、係り受けの深さを示す韻律記号＃２～＃６を用いた．ここで、句境界直前の句が直後の句を直接修飾する場合（左枝分かれ境界）、フレーズの立て直し現象は起こらないと仮定し、＃１は除いた．

input: 警官は走って逃げる泥棒を追いかけた

phoneme: k e e k a N w a h a sh i cl t e n i g e r u d o r o b o o o o i k a k e t a
baseline (accent): k e ^ e k a N w a # h a ^ sh i ! cl t e # n i ^ g e ! r u # d o ^ r o b o o o # o ^ i k a k e ! t a (

proposed1 (accent+phrase dependency structure):
構文木1: k e ^ e k a N w a #4 h a ^ sh i ! cl t e #1 n i ^ g e ! r u #1 d o ^ r o b o o o #1 o ^ i k a k e ! t a
構文木2: k e ^ e k a N w a #1 h a ^ sh i ! cl t e #3 n i ^ g e ! r u #1 d o ^ r o b o o o #1 o ^ i k a k e ! t a

proposed2: (based on the processes of generation control mechanism):
構文木1: k e / e k a N w a ¥ #4 h a / sh i ¥ cl t e n i / g e ¥ r u d o / r o b o o o ¥ o / i k a k e ¥ t a
構文木2: k e / e k a N w a ¥ h a / sh i ¥ cl t e #3 n i / g e ¥ r u d o / r o b o o o ¥ o / i k a k e ¥ t a


図１　２つの構文木候補に対する、従来モデルと提案モデルで用いる音素と韻律記号の比較

Fig. 1　The comparison of phonemes and prosodic symbols utilized in the baseline and the proposed method based on two syntax tree candidates.

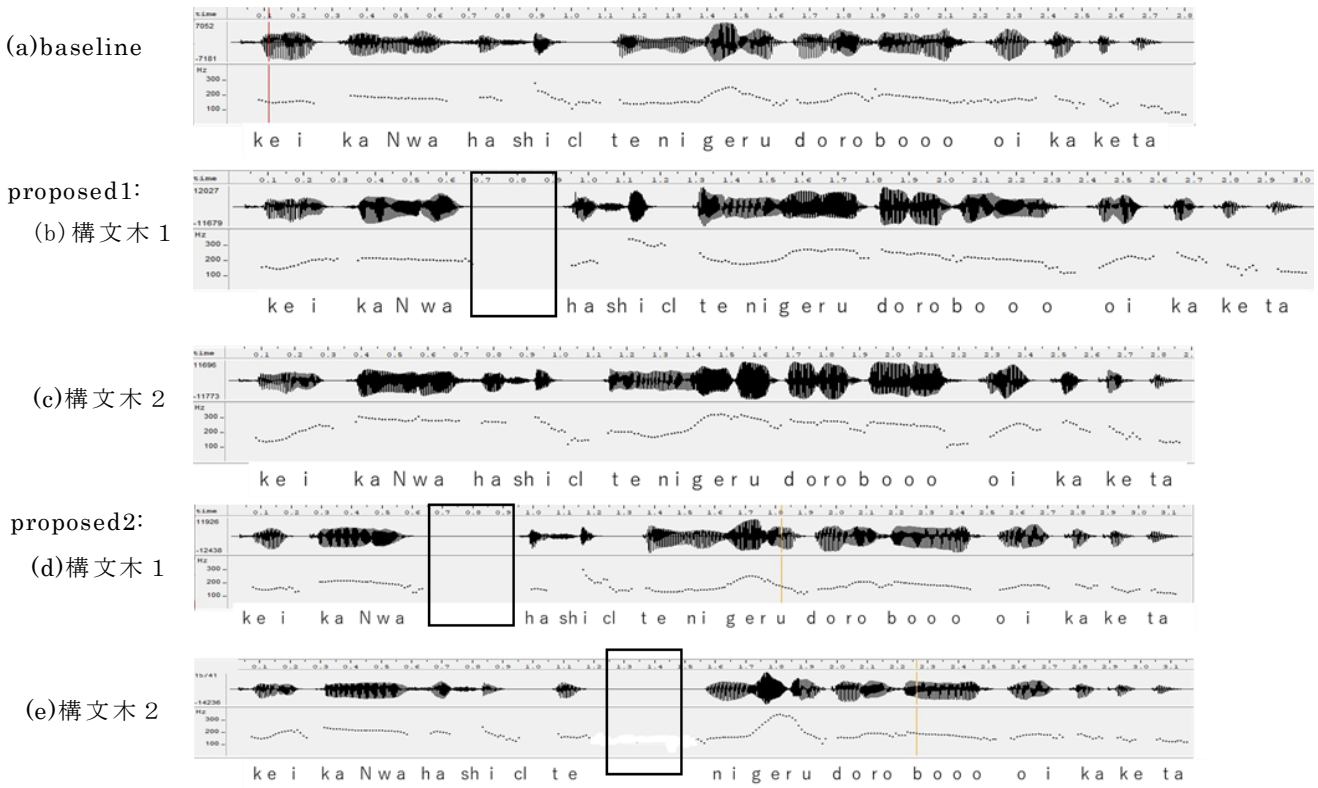


図２　構文木の句構造を反映したポーズ生成

Fig. 2 Pause generation that reflects the phrase dependency structure of syntax trees.

## 3. 実験用音声データセット

本稿では、朗読音声の自然性向上を目的として、単文の読み上げ音声ではなく、１人の話者が物語を朗読した音声データを用いた. 実験用データセットは、Web に公開されているアラビアンナイト口語訳のテキストとその朗読音声（190 文章、平均 8 分 21 秒、全 26 時間 26 分）[13]を利用し、新たに作成した。

テキストは、手作業で不要な記号や改行などを取り除いたのち、句点、改行等に基づき、文単位に分割した. 文の読み、アクセント型やその境界などの情報は、Open Jtalk [14]を用い自動的に付与した. また文の句構造を示す句境界の係り受け情報は、Open Jtalk を用いた形態素解析結果を ChaboCha[15]に入力し、得られた構文木表現に基づき自動的に付与した.

音声データは、ESPnet[16] の CTC Segmentation[17] を用い文単位に自動的に分割した. さらに、Open Jtalk で得た読みに対応する、音素単位のアライメント[18] を文単位にとり、Open Jtalk で得られた他の情報と併せてフルコンテキストラベルを作成した. これらの処理中、文分割の誤り、音素アライメントの誤りとなった文、及び１文が 20 秒を超える文を除いた計 11,615 文をデータセットとして用いた.

## 4. 実験

本実験では、入力シーケンスからメルケプストラムを生成するモデルは、ESPNet2 で提供されている日本語 Tacotron2[11][12]を用いた. またメルケプストラムから音声波形を生成するボコーダは、Parallel-Wavegan[19]を用いた.

３章で述べた新規に作成した朗読データベース 11,615 文を利用し、表１、２、３の韻律記号と音素記号を入力として用いた３つのモデル作成した. 11,615 文のうち、学習用に 11,115 文、検証・テスト用に各 250 文を用いた. モデル作成に当たっては、充分に収束するまで学習を行った.

### 4.1. 客観評価

本節では、学習した提案モデルの合成音声が、構文木の句構造を反映し、読点のない右枝分かれ境界で、１）フレーズ区切りを示すポーズが自動的に挿入されることや、２）F0 の「立て直し」現象が起こることを示す.

従来モデルと２つの提案モデルの合計３モデルを用い、統語的にあいまいな文「警官は走って逃げる泥棒を追いかけた」を入力し、合成音声を生成しその音声波形と F0 概形を調べた. この統語的にあいまいな文は、図１のように、２つの構文木を想定することができる.

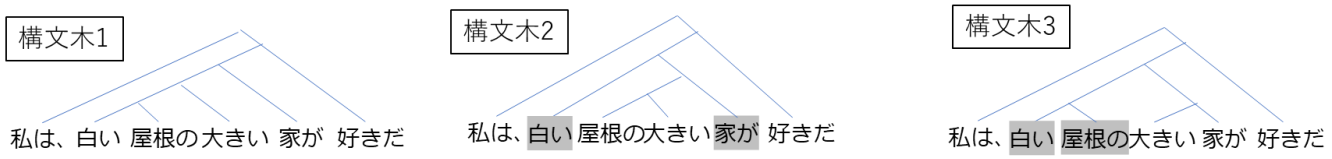


図 3　３つの異なる木構造

Fig. 3 Three different phrase dependency tree.

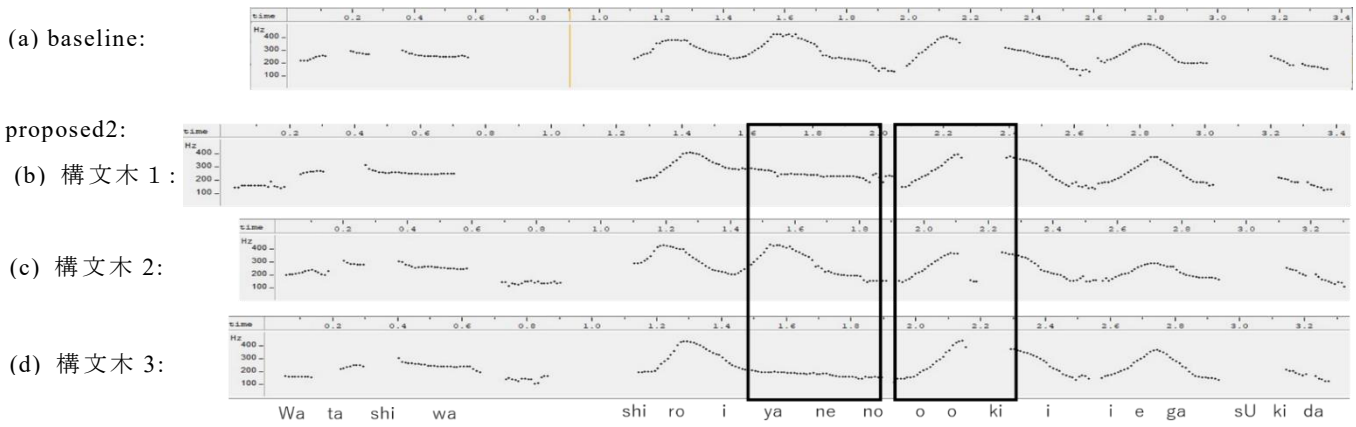


図 4　３つの異なる木構造を入力して合成した音声の F0 概形

Fig. 4　F0 counter of synthesis speech of three different phrase dependency tree.

図２(a)にアクセント制御のみの制御記号を用いた従来モデルにより生成した合成音声波形と F0 概形を示す．また図２(b)～(e)に提案モデル１を、２を用い生成した２つの構文木に対応する合成音声波形と F0 概形を示す．構文１の場合、右枝分かれ境界となる「警官は」と「走って」の句境界において、従来モデルの合成音声ではポーズは生成されない．これに対し提案モデル１，２とも構文を明確にするフレーズの境界を示すポーズが生成されている．構文２の場合、右枝分かれ境界となる「走って」と「逃げる」の句境界が右枝分かれ境界において、モデル２の合成音声では同様に構文を明確にするフレーズの境界を示すポーズが生成されるが．従来モデル、提案モデル１ではポーズは生成されていない．

次に、右枝分かれ境界で、F0 の「立て直し」が起こる例を示す． 図３に、統語的にあいまいな文「私は、屋根の大きい白い家が好きだ」で想定できる３つの構文木を示す． 図４(a)にアクセント制御記号のみに基づく従来モデルにより生成した合成音声の F0 概形を示す．また図４(b)～(d)に、提案モデル２により生成した図３の３つの構文木の合成音声の F0 概形を示す．「白い」が「家が」係る構文木２の場合、右枝分かれ境界となる「白い」と「屋根の」の句境界で、フレーズの「立て直し」現象が起こり、木構造に沿った F0 概形を示している．しかしながら、「白い屋根の」が「家が」に係る構文木３の場合、右枝分かれ境界となる「屋根の」と「大きい」の句境界は、左枝分かれ境界となる他の構文木１，２と同様の F0 概形となり、F0 の「立て直し」現象は見られなかった．但し、構文木３のアクセント句「大きい」の F0 最大値は他の構文木のアクセント句の F0 最大値よりも若干大きかった．

これらの結果は、右枝分かれ境界で、従来モデルでは、ポーズの生成や、F0 の「立て直し」現象が起きないが．新たに提案した２つのモデルでは、ポーズの生成や、F0 の「立て直し」現象が起こることが示された．しかし、右枝分かれ境界でこれらの現象が必ず起こるものではなく、学習に用いた音声データに従い学習・生成されている．

### 4.2. 主観評価

句構造の情報を利用する韻律記号を導入・学習した提案モデルにより合成した音声の総合的な評価を行うため、合成音声の自然性の聴取実験を行った．

#### 4.2.1. 評価データ

評価データには、モデル作成時に利用した学習用、検証用に用いたデータ以外の評価用 250 文からランダムに抽出したデータを用いた．但し、構文木の効果を聴取実験で検証するため、以下の 3 つの条件を１つでも満たす文を除いた 20 文（以下、朗読文）を用いた．

１）句境界が、句境界直前の句が直後の句を直接修飾する左枝分かれ境界のみで構成される文

２）アクセント句が３つ以下の文

３）８秒を超える音声が合成される文

従来モデル、２つの提案モデルの３つのモデル毎に各 20 文、計 60 文の合成音声を作成し．聴取実験用の評価データとした．

#### 4.2.2 評価方法

３つのモデル間の韻律の自然性の比較を行うため、対比較実験を行った．評価者は１３名で、全て日本語母国語話者である．評価は１対の合成音声を聴取し、どちらの合成音声の韻律の自然性が高いかの強制判定を行った．各２０文を従来モデルと提案モデル１、従来モデルと提案モデル２、提案モデル１と提案モデル２の計６０対をランダムに提示し、聴取実験を行った．また１対の合成音声のモデルの提示順序はランダムとした．

#### 4.2.3 評価結果・考察

主観評価実験の結果を表５に示す．この結果より以下が判明した．

１）文の句構造の情報を用いた韻律記号を用いる提案モデル 1、2 の合成音声は、アクセント成分のみの韻律記号を用いた従来モデルの合成音声に比べ、有意(1%水準)に自然性が高い（提案モデル１：68%、提案モデル２：62%）．

２）従来モデルと提案モデル１，２の比較において、提案モデル１は提案モデル２の合成音声より自然性が高いと判定された、しかし提案モデル１と２の差に有意性は認められなかった．また提案モデル１と２の合成音声の直接比較で提案モデル２の合成音声は、提案モデル１に対し、53%自然性が高いと判断された．しかし同様に、モデル間の有意差は認められなかった．今回の主観評価では、提案モデル１と提案モデル２の合成音声の自然性の差を判定することはできなかった．

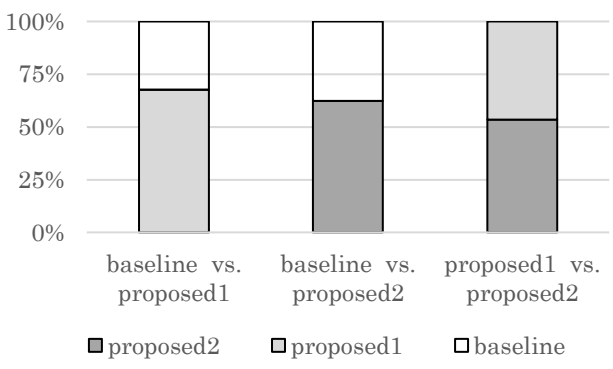


図５　韻律記号の有効性に関する自然性評価結果
Fig. 5 The subjective evaluation (AB test) of naturalness for prosodic symbols.

## 5. おわりに

本稿では、自然な韻律をもつ日本語音声を合成するため、局所的な句構造に基づくフレーズ成分を表す韻律制御記号を end-to-end 音声合成に新たに導入した．文の局所的な句構造を表現するために、１）句境界に係り受けの深さを表す韻律記号を追加するモデルと、２）韻律生成の制御機構に基づき、フレーズ成分とアクセント成分の重畳型モデルを反映させた韻律制御記号を採用するの２つのモデルを提案した．

この２つのモデルを用いた合成音声を調べた結果、文の構文木の構造を反映した右枝分かれ境界で、以下の韻律の生成が確認された．

1) 読点がないにも関わらず、ポーズが生成されること．
2) F0 のフレーズ成分の立て直し現象が生じること．

この２つの提案モデルと従来モデルの３つのモデルを用いた end-to-end 音声合成の合成音声の主観評価実験を行った．局所的な句構造を表す句境界の係り受けの深さを表すの韻律制御記号を新たに導入した提案モデル１では、従来のアクセント情報のみの韻律制御記号を用いた従来モデルに比べ、合成音声の韻律の自然性が 68%向上した．さらに、韻律生成制御機構に基づく韻律制御記号を用いた提案モデル２の場合、従来モデルに比べ、合成音声の韻律の自然性が 62%向上した．

これらの実験結果から、日本語 end-to-end 音声合成に文の局所的な句構造を表す情報や、韻律の生成モデルを取り入れることにより、発話者の意図をより正しく反映した自然な韻律を持つ合成音声が生成できることが確認された．

今後、更により自然な音声合成音の生成を目指し、当該句境界の句構造のみでなく、その前の句の構造や、並列関係などの詳細な句構造検討や、品詞情報などの区間のより詳細な情報の利用を図っていく．

## 謝　辞

本研究に利用させていただいた朗読音声とテキストを広く公開し、営利目的以外の利用を許可していただいている武葉槌様に感謝いたします．

本研究の一部は JSPS 科研費 JP17H06101 および JP21H03467 の助成を受けたものです．

## 文　献

[1] T. Fujimoto, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda “Impacts of Input Linguistic Feature Representation on Japanese End-to-End Speech Synthesis,” Proc. of 10th ISCA Speech Synthesis Workshop(SSW), pp.166-171, Vienna, Austria, Sep. 2019

[2] K. Kurihara, N. Seiyama, T. Kumano, “Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS,” IEICE Trans. Inf. & Syst., Vol.E104–D, no2, Feb. 2021

[3] 箱田和雄，佐藤大和，“文音声合成における音調規則，” 信学論(D), vol.J63-D, no.9, pp.715-722, Sept. 1980.

[4] 広瀬啓吉，藤崎博也，河井恒，山口幹雄，“基本周波数パターン生成過程モデルに基づく文章音声の合成，” 信学論(A), vol.J72-A, no.1, pp.32-40, Jan. 1989.

[5] 箱田和雄，中嶌信弥，広川智久，“文章音声の音調結合型導出規則の検討，” 信学技報，SP89-5, May 1989.

[6] 匂坂芳典，“F0 パタン概形制御の定量的検討，” 信学技報，SP89-111, Jan. 1990.

[7] 阿部匡伸，佐藤大和，“音節区分化モデルに基づく基本周波数パタンの 2 階層制御方式，” 音響誌，vol.49, no.10,pp.682–690, Oct. 1993.

[8] 海木延佳，匂坂芳典，“局所的句構造に基づく F0 制御，” 信学論(D—II), vol.J83-D-II, no.9, pp.1853-1860, Sept. 2000.

[9] 海木延佳，匂坂芳典，“局所的な句構造によるポーズ挿入規則化の検討，” 信学論(D—II), vol.J79-D-II, no.9, pp.1455-1463, Sept. 1996.

[10] 栗原清、清山信正、熊野正、今井篤、“読み仮名と韻律記号を入力とする日本語 end-to-end 音声合成の音質評価、” 信学技報、SP2018-49, Dec. 2018.

[11] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” in Proc. Interspeech, Aug. 2017, pp. 4006–4010.

[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, RJ Skerry-Ryan, Rif A. Saurous, Y. Agiomyrgiannakis, Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," ICASSP 2018

[13] 武葉槌，“ホッとするお話の朗読 アラビアンナイト口語訳について，” https://o-keil.com/okinu-baba/wordpress/?p=818

[14] “Open JTalk,” http://open-jtalk.sourceforge.net/.

[15] “CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer,” http://taku910.github.io/cabocha/

[16] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.E.Y. Soplin, J. Heymann, M. Wiesner, and N. Chen, “Espnet: End-to-end speech processing toolkit,” Proc. Interspeech, pp.2207–2211, 2018, https://github.com/espnet/espnet.

[17] 西鳥羽二郎、“CTC Segmentation の紹介，” https://tech.retrieva.jp/entry/2020/10/02/143338

[18] “Montreal Forced Aligner、” https://montrealcorpustools.github.io/Montreal-Forced-Aligner/

[19] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” in Proc. Interspeech, Aug. 2017, pp. 4006–4010.