韻律情報で条件付けされた 非自己回帰型End-to-End日本語音声合成の検討

藤井 一貴[1]　齋藤 佑樹[1]　猿渡 洋[1]

概要：
日本語音声合成において，合成音声のアクセントは合成音声の品質に大きく寄与するだけでなく，正確な情報伝達を行う上で重要な情報である．しかし，音素や文字などの生テキストに近い情報から音声を予測する End-to-End 日本語音声合成では，合成音声のアクセント誤りが頻出する．提案手法では非自己回帰型 End-to-End 音声合成モデルの代表である FastSpeech2 に入力する音素記号に，テキスト解析により得られた韻律情報を取り込むことで，アクセント誤りの改善を目指す．また，韻律情報を抽出する際のテキスト解析で用いる辞書の影響も調査する．実験的評価の結果より，提案手法が合成音声の韻律予測精度と自然性を有意に改善させることを示す．

Kazuki Fujii[1]　Yuki Saito[1]　Hiroshi Saruwatari[1]

1. はじめに

任意のテキストを入力として，それに対応する明瞭で自然な音声を合成することを目的とした技術のことを音声合成 (Text-to-Speech: TTS) [1] と呼ぶ．この技術は，各種機器の操作案内音声や警告音声，公共における案内音声や，今日においてはスマートスピーカーのようなユーザからの質問に音声で返答するタスクにも応用されている．テキスト音声合成は，テキストから音声を予測するという問題として考えることができる．従来の統計的パラメトリック音声合成 [2] では，この問題を直接解くことは困難であったため，1 つの問題を複数の問題に分割するパイプライン型の方式をとっていた．しかし，この方式は，各モジュールが独立して最適化されるために必ずしも全体として最適にならず合成音声の品質が劣化する傾向にある．そこで，部分問題への分割を行わず，全てを深層ニューラルネットワーク (Deep Neural Network: DNN) による予測に置き換えてテキストから音声波形を一気通貫方式で生成する End-to-End 音声合成 [3] が提案された．これにより，テキストから音声を予測する問題を全体的に最適化することが可能となり，高い品質の音声を合成可能でありながら，単一モジュールでの音声合成となったために非専門家の音声分野への新規参入が容易になった．特に，Shen らによって提案された自己回帰型 End-to-End 音声合成の Tacotron2 [4] は，人間の肉声に匹敵する品質の音声を合成可能である．しかし，End-to-End 音声合成は音声合成モデル全体が DNN で記述されているため人間によるモデルの解釈性が低下し，合成音声の制御や，誤りが生じた場合の訂正が困難である．

特に，End-to-End 日本語音声合成では音素列や文字列のみから正しいアクセントを推定することは非常に困難であり，日本語において合成音声のアクセント誤りが頻出してしまう問題がある．正しいアクセントの予測は合成音声の品質に大きく寄与するだけでなく，正確な情報伝達を行う上で重要である．故に，合成音声のアクセントに誤りが発生した場合，誤ったアクセントによって全く違う単語として捉えられてしまい正しい情報伝達を行うことができなくなる可能性がある．

End-to-End 日本語音声合成における合成音声のアクセント誤りを改善するために，テキスト解析由来のアクセント情報を音声合成モデルの入力特徴量として使用する手法がこれまでに提案されている．Okamoto らは自己回帰型 [5] と非自己回帰型 [6] の両方の End-to-End 音声合成方式において入力にフルコンテキストラベル（音素情報と文脈を考慮した文章のアクセント情報が内包されているファ

[1] 東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, Hongo 7–3–1, Bunkyo-ku, Tokyo, 113–8656 Japan.

イル）を用いる枠組みを提案した．しかし，フルコンテキストラベルは非専門家にとって非常に解釈性の低い言語特徴量表現であり，音声合成のユーザがフルコンテキストラベルの修正を基に合成音声の制御やアクセント誤りの訂正を行うことは困難である．Okamoto らの研究に関連して，Kurihara [7] らはフルコンテキストラベルから音素と韻律を表すひとまとまりの情報に変換を行った後，Tacotron2 の入力に使用する枠組みの提案を行った．Kurihara らによる取り組みによって，合成音声の韻律の予測精度改善と，より解釈しやすい記号形式での韻律操作が実現された．しかし，Tacotron2 は前述の通り自己回帰型の End-to-End 音声合成であるため，非自己回帰型の合成方式と比較すると学習と推論が低速であるという点や，推論の不安定さ（読み飛ばしや繰り返しを発生される可能性がある）という点で課題がある．

そのため本稿では，自然な韻律を持つ音声を高速に合成可能な解釈性の高い End-to-End 日本語音声合成システムの実現を目指し，韻律情報で条件付けされた非自己回帰 End-to-End 音声合成の手法を提案する．提案手法では，音素列と Kurihara らの手法 [7] における韻律記号からそれぞれ音素埋め込みと韻律埋め込みを抽出し，それらを加算した特徴量から合成音声のメルスペクトログラムを予測するように音響モデルを学習する．また，新語や複合語などに対しても頑健に韻律記号を週出するために，テキスト解析のための辞書に tdmelodic [8] を導入し，その有効性を検証する．提案手法に対して主観的評価，客観的評価の両方を行った結果，提案手法は自然な韻律を持つ合成音声を生成できることを報告する．

2. 従来手法

2.1 End-to-End 日本語音声合成

音声合成は，テキストからその音声を予測するという問題として捉えることができる．以下，この問題を音声合成問題と呼ぶ．統計的パラメトリック方式 [2] と呼ばれる，従来のテキスト音声合成では，音声合成問題を直接解くことが困難であったため，テキストから言語特徴量を抽出するモジュール，音声から音響特徴量を抽出するモジュール，音響モデルにより言語特徴量から音響特徴量を予測するモジュール，音響特徴量から音声波形を生成するモジュールに分割し，これらのモジュールを連結することで音声合成システムを実現していた．しかし，この手法ではそれぞれのモジュールが独立に最適化されるため，全体として必ずしも最適な音声合成システムが構築される保証がない．そのため，合成音声の品質が自然音声と比較して劣化する傾向にある．そこで，End-to-End 音声合成 [3] と呼ばれる，音声合成問題を DNN によって直接解く手法が提案された．End-to-End 音声合成ではテキストから音声を生成する単一の DNN を学習するため，音声合成システム全体としての最適化が可能であり，従来方式よりも高い品質の音声を合成できる．しかし，システム全体が DNN で記述されているため，人間によるモデルの解釈性が低下する．そのため，合成音声の制御や，誤りが生じた場合の訂正が困難となる．

End-to-End 音声合成の音響モデルは自己回帰型のものと非自己回帰型のものに分類可能であり，自己回帰型の代表例が Tacotron2 [4]，非自己回帰型の代表例が FastSpeech2 [9] である．これらの手法は，文献中では英語の音声合成で評価されており，Yasuda らは日本語音声合成へ適用するための調査を行った [10]．本研究では，学習と推論の速さ及び推論の安定性を重視し，FastSpeech2 を従来手法として検討する．

FastSpeech2 は，テキスト表記から音素表記に変換した後に音素を入力としてメルスペクトログラムを出力する音声合成手法である．出力されたメルスペクトログラムはボコーダにより音声波形へと変換される．入力された音素は音素埋め込みを経て，時間的位置を表すために Positional Encoding と呼ばれる値を足し合わした後にエンコーダへと入力される．エンコーダの出力は Variance Adaptor (VA) と呼ばれるモジュールへと入力される．VA は，音素継続長，ピッチやエナジーを予測する複数の Predictor を内包しており，学習時には学習データから抽出した音素継続長などの特徴量を用いて Predictor の学習を行い，推論時には学習された Predictor より各特徴量の予測が行われる．音素継続長やピッチ，エナジーなどを中間表現として予測することで，モデルの解釈性や合成音声の制御性を高めることができる．

2.2 日本語アクセントとその表現形式

日本語はピッチアクセント言語の一種であり，1 つのモーラに対してピッチの高低 (H/L) を変化させることでアクセントを表現する．これらの音の違いから，同じ読みであるが意味が異なる単語（例えば「箸」や「橋」，「端」など）の弁別を行っている．故に，日本語音声合成において，正しいアクセント情報は自然で正確な意味を持つ音声を合成する上で非常に重要な役割を持つ．

しかし，音素列のみを入力する End-to-End 日本語音声合成は，正しいアクセントを持つ音声の合成は非常に困難である．そこで，Okamoto ら [5] [6] は End-to-End 音声合成の入力にフルコンテキストラベルを用いる枠組みを提案し，音素以外のテキストに関する情報を入力することで合成音声の品質が改善することを示した．しかし，フルコンテキストラベルは非専門家にとって解釈性の低い言語特徴量であり，音声合成のユーザがフルコンテキストラベルの修正を基に合成音声の制御やアクセント誤りの訂正を行うことは困難である．Kurihara ら [7] はフルコンテキストラベルから音素と韻律を表すひとまとまりの情報に変換

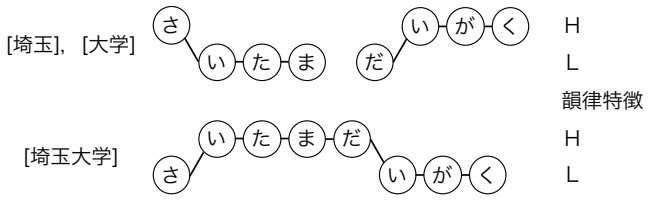

図 1 アクセント結合の例.

を行った後，Tacotron2 の入力に使用する枠組みの提案を行った．Kurihara らによる取り組みによって，フルコンテキストラベルよりも解釈が容易な情報を用いた韻律の制御が可能となったが，Tacotron2 は前述の通り自己回帰型の End-to-End 音声合成であるため，非自己回帰型の合成方式と比較すると学習と推論が低速であるという点や，推論の不安定さ（読み飛ばしや繰り返しを発生される可能性がある）という点で課題がある．

これらの先行研究は，テキスト解析により正しいアクセント情報が得られれば合成音声の品質が改善することを示したものである．しかし，日本語では **Fig. 1** に示すように，複合語におけるアクセント変化が生じる．故に，単語だけでなく，複合語や新語などを含む様々なテキストに対して正しいアクセント情報が推定できることが望ましい．

表 1 使用する韻律記号の定義

韻律記号	韻律記号の意味
[	ピッチ上がり
]	ピッチ下がり（アクセント核）
#	アクセント境界
^	<SOS>
$	<EOS>
?	<EOS>（疑問形）
_	アクセント情報に変化なし

3. 提案手法

本稿では，自然な韻律を持つ音声を高速に合成可能な解釈性の高い End-to-End 日本語音声合成システムの実現を目指し，韻律情報で条件付けされた非自己回帰 End-to-End 音声合成の手法を提案する．提案手法の概要を **Fig. 2** に示す．

3.1 韻律記号の推定

本研究では，Kurihara ら [7] の研究で用いられていた韻律記号を FastSpeech2 ベースの非自己回帰型 End-to-End 音声合成に適用可能な形で拡張する．提案手法では，まず，日本語テキストに対して OpenJTalk [11] を用いたテキスト解析を行い，フルコンテキストラベルを得る．その後，フルコンテキストラベルから Kurihara らのアルゴリズム [7] を用いて **Table 1** に示す韻律記号を抽出する．ここで，先行研究 [7] では入力音素列に **Table 1** の 2 行目から 7 行目と同様の韻律記号を直接付与して Tacotron2 に入力していたが，FastSpeech2 ではテキストに付与された韻律記号に対する継続長が定義できない．そこで，本研究では **Table 1** の 8 行目に示す記号 “_” を新たに導入し，音素列と同じ長さを持つ韻律記号列で FastSpeech2 を条件付けする．また，“<SOS>” と “<EOS>” はそれぞれ文頭と文末を意味する．

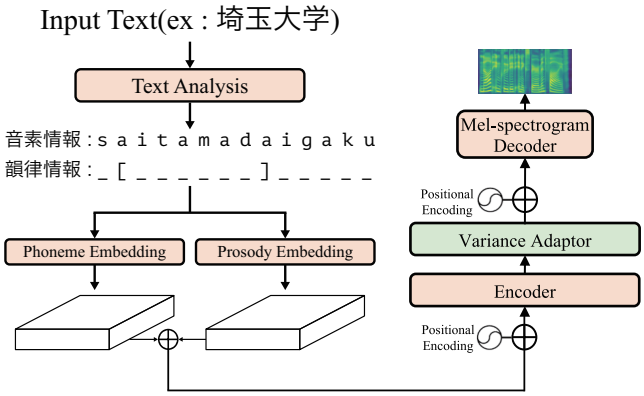

図 2 提案手法の概要．提案手法では，日本語テキストを入力として，音素情報と韻律情報の抽出を行う．抽出した情報は，音素埋め込みと韻律埋め込みへ変換後，加算されて FastSpeech2 の Encoder に入力される．

ここで，OpenJTalk が日本語解析に使用する辞書は NAIST-jdic [12] がデフォルトとなっている．本研究では，多様なテキストに対して正しいアクセントを推定可能にするために，OpenJTalk の辞書に tdmelodic [8] を導入する．tdmelodic は日本語テキストのアクセントを DNN により自動推定できるモジュールを使用して構築された辞書であり，新語や流行語，商標名などの標準的な辞書には搭載されていない単語や，複合語などのアクセントも多く収録している．故に，アクセントの誤推定による合成音声の品質劣化を緩和できる．

3.2 韻律記号で条件付けされた FastSpeech2

Section 3.1 で推定された音素情報と韻律情報は，音素埋め込みと韻律埋め込みに変換される．次に，2 つの埋め込みの和を取り，Positional Encoding を行なった後に FastSpeech2 の Encoder に入力する．以降の処理はオリジナルの FastSpeech2 と同じであるため省略する．つまり，FastSpeech2 に入力する韻律記号を修正することで合成音声のアクセントを直感的に制御・修正することが可能となる．アクセント制御に関しては **Section 3.3** にて議論を行う．また，オリジナルの FastSpeech2 が推論を行う際のデータフローは **Fig. 2** に示すうちの Prosody に関連する項目を無視したものと一致する．

3.3 考察

合成音声のアクセント制御可能性を確かめるために，予備実験として FastSpeech2 に入力する韻律記号を変更する

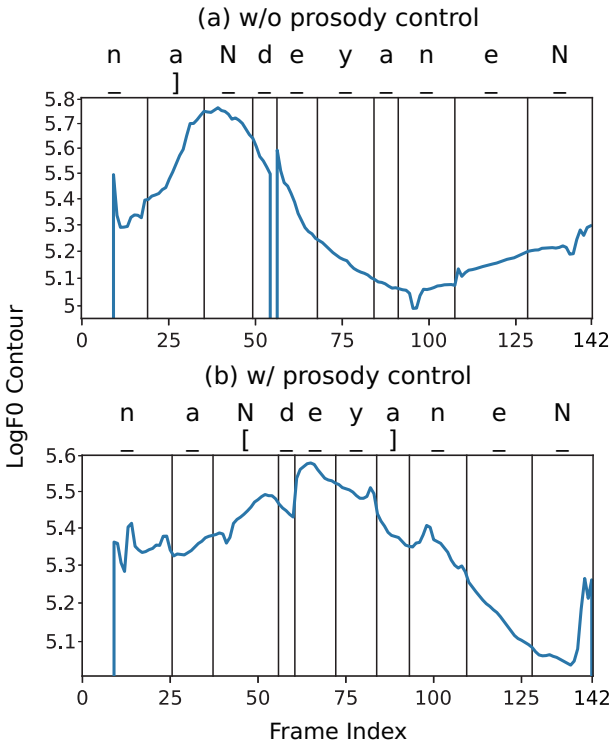

図 3 韻律記号の操作により F0 が変化する例．図の縦線は音素アライメントの区切りを表し，各音素に対応する韻律記号を併記している．

ことで合成音声のアクセントがその韻律記号に従うかどうかの確認を行う． **Fig. 3** に，韻律記号の操作により合成音声の F0 が変化する例を示す．この入力テキストは，日本語の関西方言である「なんでやねん」である．ここで，方言に対応していない辞書を用いてテキスト解析を行った結果として得られる韻律記号を FastSpeech2 に入力すると，, **Fig. 3**(a) に示すように「な」にアクセントがある音声が合成される．一方で，**Fig. 3**(b) に示すように韻律記号を修正すると，「で」で上昇し「ね」で下降する京阪式アクセントの音声を合成するように音声合成システムを制御できる．

4. 実験的評価

実験的評価では，提案手法により合成された音声のアクセントが自然なものであり，従来手法においてアクセント誤りを起こしていた合成音声の品質が改善されているかを確認する．

4.1 実験条件

本研究では，単一の女性話者による日本語音声約 10 時間により構成される JSUT コーパス [13] のうち，日本語常用漢字を全て網羅する BASIC5000 というサブセットを用いて実験を行った．音声のサンプリング周波数は 22050 Hz で，使用するメルスペクトログラムの次元は 80 とした．また，F0 の分析には音声分析システム WORLD [14] の

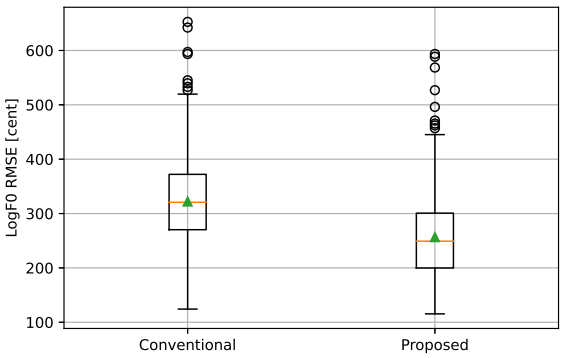

図 4 従来手法と提案手法の客観評価結果 (logF0 RMSE)

Python ラッパーである PyWORLD [*1]を用いた．学習データと評価データの数はそれぞれ 4488 文, 512 文とした．

本稿で使用する FastSpeech2 は，GitHub 上の実装 [*2]をフォークして日本語に対応させたリポジトリ [*3]を基に実装を行った．学習に使用する音素アライメントの情報は，JSUT コーパスの各音声に対して汎用大語彙連続音声認識エンジンである Julius [15] を用いて得た．この際使用する音響モデル等はデフォルトのものとした．FastSpeech2 の音素埋め込みと韻律埋め込みの次元は 256 に設定し，DNN 学習の Optimizer は学習率が 0.001 の Adam [16] とした．メルスペクトログラムから音声波形を生成するニューラルボコーダには，Hi-Fi-GAN [17] を使用した．HiFi-GAN のモデルは，GitHub 上で公開されている事前学習済みの UNIVERSAL_V1[*4]を用いた．

比較手法は，提案手法を実装する基となった GitHub 上の実装 [*2]をそのまま利用して学習を行なったものとして，隠れ層数や隠れ素子数などの DNN アーキテクチャは提案手法のものと同じとした．音声を合成する際に使用する各手法のモデルの学習ステップは 100000 とした．以降，実験結果における “Conventional” は従来手法を，“Proposed” は提案手法を意味する．

4.2 韻律の予測精度に関する客観評価

客観評価の指標として，自然音声と合成音声の間の対数基本周波数 (logF0) の Root Mean Squared Error (RMSE) [cent] を用いた．また，本実験では logF0 の予測精度のみに着目するため，客観評価における合成音声は自然音声の音素継続長を用いて生成した．

4.2.1 従来手法と提案手法の比較

初めに，従来手法と提案手法の合成音声の logF0 の予測精度に着目した客観評価を行う． 客観評価の結果を **Fig.**

[*1] https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder
[*2] https://github.com/ming024/FastSpeech2
[*3] https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[*4] https://github.com/jik876/hifi-gan

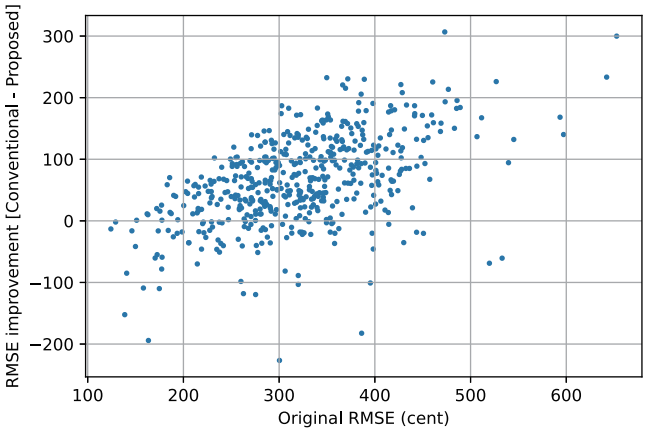

図 5 “Conventional” に対する “Proposed” の logF0 RMSE 改善量を示す 2 次元散布図

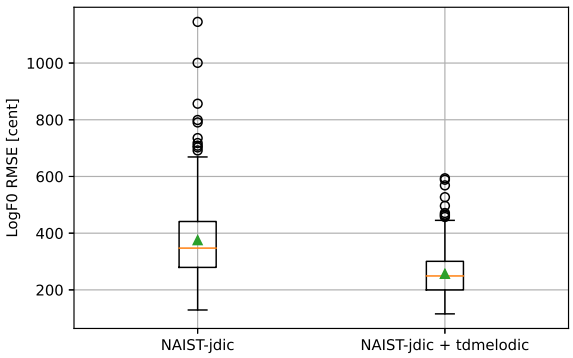

図 6 テキスト解析に用いる辞書を変更した場合における “Proposed” の logF0 RMSE の比較

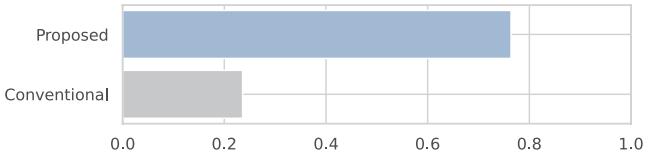

図 7 韻律の類似性に関するプリファレンス XAB スコア

4 に示す．この図は，評価結果を手法ごとに箱ひげ図によりプロットしたものである．縦軸の単位は，1 オクターブを 1200 分割した音高の単位である cent で，RMSE の結果を表す軸であることからこの値は小さいほど韻律の予測精度が高いと言える．評価結果より，提案手法により合成された音声に対する logF0 RMSE が従来手法よりも低い値を取る傾向にあり，非自己回帰型の End-to-End 音声合成方式においても，韻律記号を考慮することで合成音声の韻律予測精度が改善することが示された．

ここで，この評価結果を詳細に分析するために，従来手法に対する提案手法による logF0 RMSE の改善を示す 2 次元散布図を作成した． この結果を，**Fig. 5** に示す．ここで，横軸は “Conventional” により予測された合成音声に対する logF0 RMSE を，縦軸は “Conventional” と “Proposed” の logF0 RMSE の差分を表しており，散布図上で縦軸の値が 0 より大きな点は提案手法により韻律の予測精度が改善した評価データであることを意味する．この結果から，提案手法により logF0 RMSE が改善した評価データは 426 文（全体の約 83.2 %）であり，劣化したサンプルは 86 文（全体の約 16.8 %）であることがわかった．以上より，(1) 従来手法で logF0 RMSE が高い（即ち，合成音声のアクセント誤りが生じている可能性が高い）サンプルは，提案手法によりその値が改善され，(2) 従来手法で logF0 RMSE が低い（即ち，合成音声のアクセントを正しく予測できている可能性が高い）サンプルは，提案手法と同程度，もしくは劣化した値を取ることが示された．つまり，従来手法で韻律を正しく予測できているサンプルは，提案手法においても同様に韻律を予測できたものと，もしくは不適切な韻律記号の追加により韻律の予測精度を劣化させたものの 2 種類に分類されることを示唆した．

4.2.2 テキスト解析における辞書の影響の調査

Section 4.2.1 の結果より，提案手法が韻律の予測精度を改善させることがわかった．ここで，提案手法において OpenJTalk の辞書を変更することの影響を調査する実験を行う．この実験における比較手法は，提案手法においてテキスト解析に NAIST-jdic のみを用いたものとした．その他の実験条件は **Section 4.2.1** と同じとした． 客観評価の結果を **Fig. 6** に示す．評価結果より，OpenJTalk の辞書に tdmelodic を追加導入することで合成音声の韻律予測精度が改善することが示された．そのため，以降の主観評価ではテキスト解析に NAIST-jdic と tdmelodic の両方の辞書を用いた．

4.3 合成音声の品質に関する主観評価

合成音声の品質を評価するための主観評価を実施した．主観評価は，クラウドソーシングサービスのプラットフォームであるランサーズ [18] を用いて実施した．

4.3.1 韻律の類似性に関する評価

韻律の類似性を評価するためのプリファレンス XAB テストでは，受聴者に対してリファレンス音声 X(自然音声のメルスペクトログラムから HiFi-GAN を用いて合成した音声) を提示した後に，従来手法もしくは提案手法により合成した音声である A と B を提示し，どちらが X に近いかを選択させた．従来手法と提案手法で合成した音声は A と B のどちらかにランダムに割り当てた．受聴者に対しては，A と B の韻律が X に近い方を選択して評価を行うように指示した．この評価での受聴者の数は 25 人で，1 人の評価回数は 10 回であるため，合計の評価回数は 250 である．

Fig. 7 に評価結果を示す．評価結果より，提案手法は従来手法よりも高い XAB スコアを獲得しており，二項検定の結果から，手法間のスコアには $p < 10^{-10}$ で有意差が

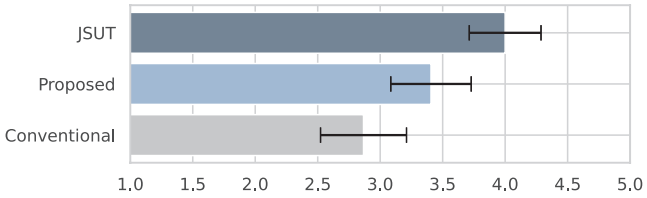

図 8 合成音声の自然性に関する MOS 値

あることを確認した．即ち，提案手法は従来手法よりも高い精度で韻律を予測できることが主観的にも示された．

4.3.2 自然性に関する評価

合成音声の自然性を評価するための MOS テストでは，“Conventional” と “Proposed” に加え，自然音声のメルスペクトログラムから HiFi-GAN ボコーダにより合成した音声 (“JSUT”) の計 3 手法を比較した．受聴者は 3 手法のいずれかで合成された音声をランダムな順番で聴き，その自然性を 5 段階（1：非常に悪い–5：非常に良い）で評価した．受聴者に対しては，合成音声の自然性（人間らしさ）に着目して評価を行うように指示した．この評価での受聴者の数は 50 人で，1 人の評価回数は 33（うち 3 問は評価基準を定めるためのダミー問題）であるため，評価回数は各手法に対して 150 回である．

Fig. 8 に MOS テストの結果を示す．図のエラーバーは 95%信頼区間を表している．評価結果より，合成音声の自然性は “JSUT”,“Proposed”,“Conventional” の順で高く評価され，t 検定の結果から，全ての手法間に MOS 値の有意差（“Conventional”-“Proposed” 間が $p < 10^{-5}$，“Proposed”-“JSUT” 間が $p < 10^{-6}$，“Conventional”-“JSUT” 間が $p < 10^{-20}$）があることが示された．即ち，提案手法は従来手法と比べて韻律の予測精度だけでなく，合成音声の自然性も有意に改善させることが示唆された．

5. おわりに

本稿では，正しいアクセントの音声を高速に合成可能な音声合成システムの実現を目指し，韻律情報で条件付けした非自己回帰型 End-to-End 日本語音声合成の手法を提案した．実験的評価の結果より，韻律記号の導入とテキスト解析時の辞書の追加が合成音声の品質を有意に改善させることを示した．今後は，アクセント誤りが生じた場合にユーザからのフィードバックを用いてその訂正が可能な音声合成の枠組みを検討する．また，与える韻律情報の形式を変更することも検討する．具体的には，アクセントの潜在表現 [19] を導入した学習法の検討を行う．

謝辞 本研究は，JST ムーンショット型研究開発事業，JPMJMS2011（実証実験），キオクシア株式会社（アルゴリズム開発）の支援を受けたものです．

参考文献

[1] Y. Sagisaka, “Speech synthesis by rule using an optimal selection of non-uniform synthesis units,” in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.
[2] H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.
[3] 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*, Stockholm, Sweden, Aug. 2017, pp. 4006–4010.
[4] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. Saurous, Y. Agiomyrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4779–4783.
[5] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, “Real-time neural text-to-speech with sequence-to-sequence acoustic model and WaveGlow or single gaussian WaveRNN vocoders,” in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1308–1312.
[6] ——, “Transformer-based text-to-speech with weighted forced attention,” in *Proc. ICASSP*, Barcelona, Spain, May. 2020, pp. 6729–6733.
[7] K. Kurihara, N. Seiyama, and T. Kumano, “Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS,” *IEICE Transactions on Information and Systems*, vol. E104.D, no. 2, pp. 302–311, Feb. 2021.
[8] H. Tachibana and Y. Katayama, “Accent estimation of Japanese words from their surfaces and romanizations for building large vocabulary accent dictionaries,” in *Proc. ICASSP*, May. 2020, pp. 8059–8063.
[9] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “FastSpeech 2: Fast and high-quality end-to-end text to speech,” in *Proc. ICLR*, Vienna, Austria, May. 2021.
[10] Y. Yasuda, X. Wang, and J. Yamagishi, “Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech synthesis,” *Computer Speech & Language*, vol. 67, p. 101183, May. 2021.
[11] “OpenJTalk,” http://open-jtalk.sourceforge.net/.
[12] “NAIST Japanese Dictionary,” https://ja.osdn.net/projects/naist-jdic/.
[13] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, Sep. 2020.
[14] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016.
[15] A. Lee, T. Kawahara, and K. Shikano, “Julius — an open source real-time large vocabulary recognition engine,” in *Proc. EUROSPEECH*, Aalborg, Denmark, Sep. 2001, pp. 1691–1694.
[16] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. ICLR*, San Diego, California, U.S.A., May. 2015.
[17] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative

adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.
[18] "Lancers," https://www.lancers.jp/.
[19] K. Yufune, T. Koriyama, S. Takamichi, and H. Saruwatari, "Accent modeling of low-resourced dialect in pitch accent language using variational autoencoder," in *Proc. SSW*, Budapest, Hungary, 2021, pp. 189–194.