人の発声に近い音質とToSpeakの機能性を両立した深層学習に基づく次世代音声合成技術

Next-generation Speech Synthesis Technology based on Deep Learning with High-fidelity Human-like Speech and ToSpeak Functionality

田村　正統 TAMURA Masatsune　　蛭田　宜樹 HIRUTA Yoshiki　　松本　剣斗 MATSUMOTO Kento

音声合成技術は，深層学習の導入により基礎技術が急速に変化し，新規参入のベンダーも含めて競争が激化している。

東芝デジタルソリューションズ（株）は，人の発声と遜色のない高音質の実現と，音声合成ミドルウェアToSpeakの機能性とを両立した新しい音声合成技術を目指し，深層学習に基づく次世代方式の開発を進めている。DNN（Deep Neural Network）コンパクト化技術を適用し，組み込み用途に利用可能な小サイズの実現とともに，逐次的な波形生成機能，韻律の作り込み機能，問題箇所の調整機能などを実装し，ビジネス応用での顧客要望に応えられる次世代音声技術を開発した。

Thanks to the introduction of deep learning, the underlying technologies behind speech synthesis have changed rapidly, leading to fierce competition including new vendors in the market.

With the goal of achieving new speech synthesis technology that delivers both high fidelity comparable to human speech and ToSpeak functionality, Toshiba Digital Solutions Corporation is currently developing a next-generation system based on deep learning. This next-generation speech technology was developed to meet customer demands and business applications by applying deep neural network (DNN) compacting technology to achieve a smaller size that can be used in embedded applications, and includes sequential waveform generation, prosody creation, and acoustic problem correction functionality.

1. まえがき

近年，音声合成技術の研究開発は，深層学習の導入により急速に変化しており，生成される合成音声の品質も格段に向上してきている。自己回帰ニューラルネットにより波形を直接モデル化するWaveNet[(1)]の技術，及びエンドツーエンド型のフレームワークを用いたTacotron2[(2)]の報告以降，その方向性は従来の単位選択型の音声合成や統計モデルに基づく音声合成から大きく変わり，新規参入のベンダーも含めて開発競争が激化している。

東芝デジタルソリューションズ（株）の音声合成ミドルウェアToSpeakは，統計モデルに基づく音声合成[(3)]と統計的特徴パラメーター選択に基づく技術[(4)]を用いており，軽量さと音質の高さを特徴としてきた。しかし，高度なコンテンツ応用などのハイエンド領域では，録音音声と並べても遜色がない，更に高品質な合成音声が求められ，DNNを応用した新方式の開発が期待されている。そこで，DNNに基づく次世代音声合成技術を新たに開発した。

音声合成システムは大きく，入力テキストを解析して得られた言語情報から音響特徴量を生成するフレームワーク部と，音響特徴量から音声波形を生成するボコーダー部に分けられる（**図1**）。開発した技術は，フレームワーク部では品質改善とともに，逐次的な音響特徴量生成や，韻律の作り込みを可能にする構成，違和感のある箇所の調整機能など，ToSpeakの機能性を維持した構成を実現した。更に，東芝の研究開発センターがAIP（特定国立開発法人 理化学研究所 革新知能統合研究センター）と共同で開発したDNNコンパクト化技術[(5)]を適用し，音質を大きく落とさずにコンパクトなサイズを実現した。ボコーダー部は，高いサンプリングレートに対応した軽量かつ高品質なモデルを構築し，録音音声と遜色のない音声合成を可能とした。

ここでは，開発したDNNに基づく次世代音声合成技術

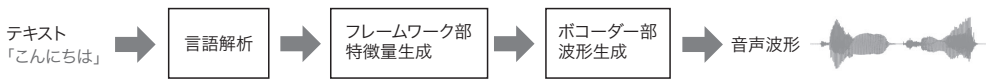


図1．音声合成システムの構成
フレームワーク部とボコーダー部より構成される。
Speech synthesis system configuration

について，技術の位置付けを示すとともに，フレームワーク部とボコーダー部のそれぞれの技術と評価結果を述べる。

2. 次世代DNN方式の音声合成技術の特長

表1に，開発した次世代DNN方式と，現行製品であるToSpeak G3, ToSpeak GxNEO（以下，GxNEOと略記），及び他社DNN方式との比較を示す。次世代DNN方式は，高い音質とともに，当社製品の特長であるコンパクトなメモリーサイズや軽量な演算量など，小規模なリソース性能を引き継ぎ，ツールやクラウドサービスによる提供だけでなく，柔軟にシステムに組み込めるミドルウェアとして提供できる。

3. フレームワーク部

フレームワーク部の構成を**図2**に示す。入力テキストを言語解析した結果から中間表現系列を得るエンコーダーと，韻律特徴量デコーダー及びスペクトル特徴量デコーダーで構成され，これら全体にDNNモデルを用いる。韻律特徴量とスペクトル特徴量のデコーダーを分離することで韻律の制御性を高め，逐次的に生成可能なスペクトル特徴量デコーダーにより応答時間を短縮する。更に，生成された特徴量に問題がある場合に対応する調整機能を実現している。

3.1 フレームワーク部の処理

エンコーダーは，音素コンテキスト情報を入力し，中間表現系列へ変換する。音素コンテキスト情報は，音素種別や，文内の位置，アクセント型など，音声合成に必用な各音素の言語属性を数値化した情報であり，ベクトルとしてエンコーダーに与えられる。

韻律特徴量デコーダーは，中間表現系列から韻律特徴量を生成する。音素継続長，各フレームの対数基本周波数，及びエネルギーを生成する。音素継続長は，畳み込み構造により，対数基本周波数とエネルギーは，単方向ゲート付き再帰ユニット（GRU：Gated Recurrent Unit）により生成する。韻律生成までの処理は，文全体に一括処理される。

表1. 音声合成システムの比較

Speech synthesis system comparison

項　目	ToSpeak G3	Tospeak GxNEO	次世代DNN方式	他社DNN方式
時期	2009～	2016～	2023～	2021～
方式	HMM	統計的 特徴量選択	DNN	DNN
音質	△	○	◎	◎
リソース	小	小～中	小～中	中～大
特徴	組み込み向け 30言語対応	コンテンツ応用 VoiceTrackMaker	ミドルウェア	ツール・ クラウドサービス

HMM：Hidden Markov Model（隠れマルコフモデル）

スペクトル特徴量デコーダーは，中間表現系列と韻律特徴量からスペクトル特徴量を逐次的に生成する。中間表現系列を音素継続長に基づいてフレーム単位に変換した後，エネルギーの情報を付加して入力し，先頭フレームから順に逐次的に生成するために，軽量畳み込み[6]とGRUを用いている。これにより，応答時間を短縮しつつ，滑らかに変化するスペクトル特徴量が得られる。

3.2 特徴量の調整

学習したモデルから例文を生成すると，まれに，音素の異音や不自然な韻律など音質的な問題が見つかる場合がある。こうした指摘事項に対応する調整機能が必用とされる。DNNにより学習されたモデルは，一般的に解釈の困難な重み情報のデータとなり，人手によって修正することは困難である。開発した技術では，調整対象を特定して特徴量を調整するルールを記述可能にすることで，調整機能を実装している。

3.3 DNNコンパクト化技術の適用

フレームワーク部の学習時に，東芝研究開発センターとAIPが開発したDNNコンパクト化技術を適用する。Adam（Adaptive Moment Estimation）による最適化，ReLU（Rectified Linear Unit）活性化関数，及び重み減衰項（Weight decay）の設定の組み合わせにより自動的に重み係数を削減する手法であり，画像認識など様々なモデルに適用されている。この技術をフレームワーク部に適用し，音質を維持したままモデルサイズ及び計算量を削減できることを確認した。

4. ボコーダー部

ボコーダー部は，フレームワーク部で各デコーダーにより生成された音響特徴量を入力し，音声波形を生成する。開

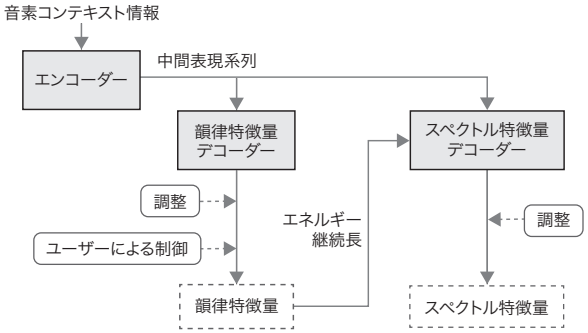


図2. フレームワーク部の構成

中間表現系列を得るエンコーダーと韻律特徴量デコーダー，スペクトル特徴量デコーダーで構成される。

Framework section configuration

発した技術では，比較的軽量に動作し，44.1 kHzの高いサンプリングレートでも録音音声に近い波形生成を可能とするニューラルボコーダーを実現した。GxNEOのスペクトル特徴量を用いる点，韻律特徴量のピッチマークに基づく補助特徴量を用いる点，及び因果的畳み込みにより逐次的な波形生成を可能とした点が特徴である。

4.1 波形生成モデルの構成

図3はボコーダー部の概要図である。ボコーダー部は，スペクトル特徴量と，韻律特徴量から音声波形を生成する。波形生成モデルは，HiFi-GAN（Generative Adversarial Network）[7]を参考に，フレーム単位の特徴量から音声サンプル単位に変換したものである。転置畳み込み（Transposed Convolution）によるアップサンプリングと，膨張畳み込み（Dilated convolution）による残差ブロックを繰り返す構造によって構成されており，韻律特徴量のピッチマークに基づく補助特徴量も用いている。

4.2 音響特徴量の比較

ボコーダー部の入力特徴量として，固定レート（窓長）による分析特徴量と，ピッチ周期分析に基づく特徴量の2種類を検討した。一般に，ニューラルボコーダーはメルスペクトログラムなど固定レート特徴量を用いることが多い。一方，ピッチ同期特徴量はGxNEOで用いられ，音声波形の高い再現性を備えている。そこで，ここではこれらの組み合わせを検討した。

4.3 因果的畳み込み

WaveNetなど自己回帰型の波形生成モデルは，過去の特徴量及び出力波形だけを用いて波形生成される。一方，非自己回帰型のモデルでは，前段の畳み込み層の出力は未来の情報も用いて波形生成されるため，各畳み込み層で入力文全体が計算されるまで合成音声が得られず，応答速度が低下する。そこで，開発した技術では，畳み込みの各層において過去の出力だけを用いる因果的畳み込みの導入を検討し，応答速度を改善した。

5. 評価実験

開発した技術の有効性を確認するため，従来法との比較と，フレームワーク部，ボコーダー部それぞれの評価を行った。実験では，女性話者と男性話者各1名について，約3,500文の音声コーパスを用いて行った。音声コーパスの文章は，現代日本語書き言葉均衡コーパス[8]から，音素バランス文を含む文セットを作成した。評価文章として，コーパスに含まれる学習内外の文章を用い，一般聴取者16名によるクラウドソーシングの評価を実施した。それぞれの合成音を5段階（5：良い，4：やや良い，3：普通，2：やや悪い，1：悪い）で評価し，MOS値（Mean Opinion Score）を求めた。

5.1 従来法との比較

図4は，従来法のGxNEOと，GxNEOのフレームワーク部を用いてボコーダー部をニューラルボコーダーに置き換えたDNNハイブリッド構成，全体にDNNモデルを用いたDNNフル構成での評価結果を示している。この図から，GxNEO，DNNハイブリッド構成，DNNフル構成の順に，段階的に音質が改善することが分かる。

5.2 フレームワーク部の有効性評価

フレームワーク部の評価結果を**図5**に示す。スペクトル特徴量として，39次のメルLSP（Line Spectrum Pair），19次の帯域雑音強度，及びエネルギーを用い，韻律特徴量として，音素継続長及び対数基本周波数を用いた。評価音声は，開発した技術のフレームワーク部を用いて生成した合

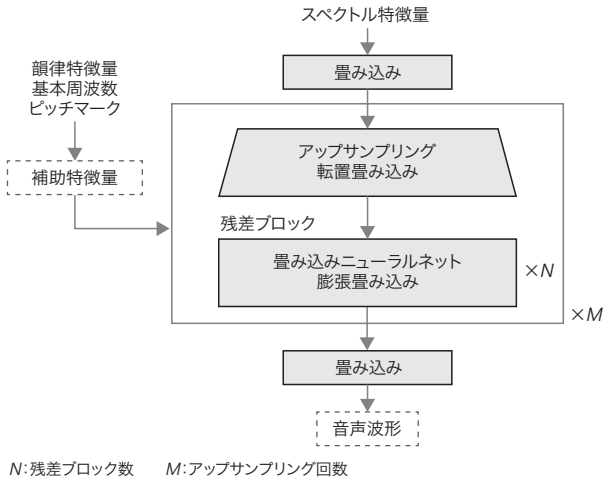


図3. ボコーダー部の構成

畳み込み層で構成され，韻律特徴量とスペクトル特徴量から音声波形を生成する。

Vocoder section configuration

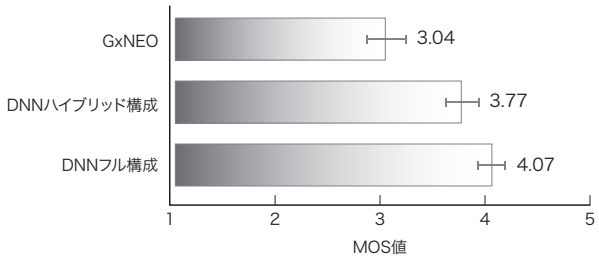
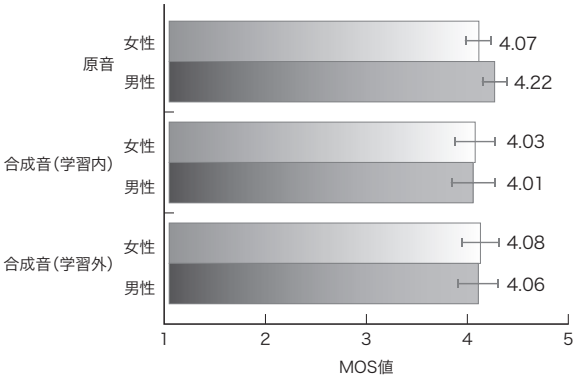


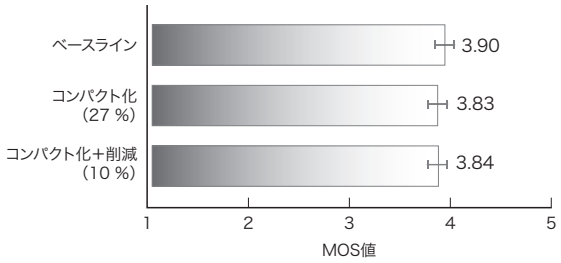
図4. 評価結果の従来法との比較

GxNEO，DNNハイブリッド構成，DNNフル構成の順に，音質が改善されている。

Comparison with conventional evaluation method



(a) フレームワーク部の評価結果



(b) DNNコンパクト化の評価結果

図5. フレームワーク部の評価結果

原音と遜色のない音質が得られており，10 %までパラメーターを削減しても，音質に大きな差は見られない。

Framework section evaluation

成音声を用い，原音と比較した。開発した技術の評価結果を図5(a)に示す。話者の違いや学習内外の文章によって大きく評価値は変わらず，原音と遜色のない，高い評価結果が得られていることが分かる。このことから，開発した技術は，韻律特徴量デコーダーとスペクトル特徴量デコーダーの分離や，逐次生成型のモデル構造により，機能性を保った中で，原音と遜色のない高い音質の合成音声が得られることが示された。

5.3 DNNコンパクト化技術適用の評価

DNNコンパクト化技術の適用による音質の違いを確認する評価実験結果を図5(b)に示す。ベースラインと，DNNコンパクト化技術により27 %に削減したモデル，更に10 %まで削減したモデルの合成音声を比較した。DNNコンパクト化技術を適用し，10 %までパラメーター削減を行っても評価値に大きな差が見られないことが分かる。ベースの約200 Mバイトのサイズに対し，削減後は約20 Mバイトまで小さくなるため，エッジデバイスに対しても次世代DNN方式による高音質な音声合成が提供可能になる。

5.4 ボコーダー部の評価

ニューラルボコーダーの評価として，特徴量の違いによる品質比較と，因果的畳み込み導入による品質比較を行った。固定レート特徴量は，39次のメルLSP，20次の帯域雑音強度，及び基本周波数を用い，ピッチ同期特徴量は，39次のメルLSP，19次の帯域雑音強度，及びピッチマークを用いた。また，ボコーダー部の入力は，開発した技術のフレームワーク部により生成した特徴量を用いた。女性話者による評価結果を**図6**に示す。ピッチ同期特徴量の比較，因果的畳み込みの比較，どちらも品質の差は小さく，原音に近い結果になっていることが分かる。

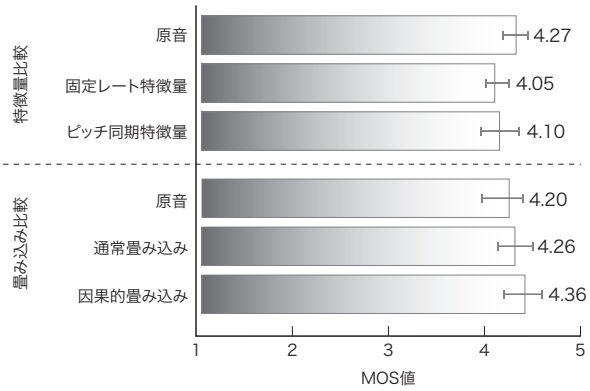


図6. ボコーダー部の評価結果

ピッチ同期特徴量，因果的畳み込みのどちらも，原音に近い結果が得られている。

Vocoder section evaluation

6. あとがき

DNNに基づく次世代音声合成技術について述べた。フレームワーク部は，韻律・スペクトル特徴量デコーダーの分離，逐次生成デコーダー，及び調整機能を特徴とし，コンパクト化技術の適用によりメモリーサイズを削減できることを示した。ボコーダー部は，原音と遜色のない高品質な波形生成が可能なことを示した。今後は，感情音声合成や，多言語，クロスリンガル技術などの多様性向上の技術開発を進めていく。

文　献

(1) Oord, A. V. D. et al. Wavenet: A generative model for raw audio. arXiv, 2016, arXiv:1609.03499v2, 2016, 15p. <https://arxiv.org/pdf/1609.03499.pdf>, (accessed 2023-04-28).

(2) Shen, J. et al. “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions.” Proceedings of 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2018). Albert, Canada, 2018-04, IEEE. 2018, p.4779–4783.

(3) 森田眞弘，ほか．多様な声や感情を豊かに表現できる音声合成技術．東芝レビュー．2013，**68**，9，p.10–13.

(4) 田村正統．統計的なパラメータ選択で声の再現性を高める次世代音声合

成技術．東芝レビュー．2016, **71**, 5, p.60–63. <https://www.global.toshiba/content/dam/toshiba/migration/corp/techReviewAssets/tech/review/2016/05/71_05pdf/f01.pdf>,（参照 2023-04-28）.
⑸ Yaguchi, A. et al. “Adam Induces Implicit Weight Sparsity in Rectifier Neural Networks.” Proceedings of 17th IEEE International Conference on Machine Learning and Applications (ICMLA 2018). Orlando, FL, 2018-12, IEEE. 2018, p.318–325.
⑹ Wu, F. et al. “Pay Less Attention with Lightweight and Dynamic Convolutions.” Proceedings of International Conference on Learning Representations (ICLR 2019). New Orleans, LA, 2019-05, ICLR. 2019.
⑺ Kong, J. et al. “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis.” Proceedings of 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Online, 2020-12, NeurIPS. 2020, p.17022–17033.
⑻ 国立国語研究所, “現代日本語書き言葉均衡コーパス（BCCWJ）”. 言語資源開発センター．<https://clrd.ninjal.ac.jp/bccwj/>,（参照 2023-04-28）.

田村　正統　TAMURA Masatsune, D.Eng.
東芝デジタルソリューションズ（株）
デジタルエンジニアリングセンター AI・自動化技術サービス部
博士（工学）　電子情報通信学会・日本音響学会・IEEE 会員
Toshiba Digital Solutions Corp.

蛭田　宜樹　HIRUTA Yoshiki
東芝デジタルソリューションズ（株）
デジタルエンジニアリングセンター AI・自動化技術サービス部
日本音響学会会員
Toshiba Digital Solutions Corp.

松本　剣斗　MATSUMOTO Kento
東芝デジタルソリューションズ（株）
デジタルエンジニアリングセンター AI・自動化技術サービス部
日本音響学会会員
Toshiba Digital Solutions Corp.