国立研究開発法人情報通信研究機構（NICT）は高品質かつ高速に動作する 21 言語のニューラル音声合成技術を開発した。CPU コアひとつで１秒の音声を既存モデルの約８倍に相当する 0.1 秒で高速合成することが可能となった。今後、商用ライセンス等を通じて多言語音声翻訳やカーナビゲーションなどの音声アプリケーションへの導入が期待される。

研究機関名	国立研究開発法人情報通信研究機構（ＮＩＣＴ）		
研究内容	情報通信技術の研究開発を基礎から応用まで統合的な視点で推進し、大学、産業界、自治体、国内外の研究機関などと連携して、研究開発成果を広く社会に還元し、イノベーションを創出することを目指す。		
所 在 地	〒184-8795 東京都小金井市貫井北町４丁目２-１		
Ｔ Ｅ Ｌ	042-327-7429	Ｕ Ｒ Ｌ	https://www.nict.go.jp/

【本技術の概要】

国立研究開発法人情報通信研究機構（NICT エヌアイシーティー、理事長: 徳田 英幸）は、ユニバーサルコミュニケーション研究所において、高品質かつ高速に動作する 21 言語のニューラル音声合成（注１）技術の開発に成功した。本技術の開発により、CPU コアひとつで 1 秒の音声をわずか 0.1 秒で高速合成（既存モデルの約 8 倍の速さ）することが可能となった。また、ネットワークに接続されていないミドルレンジ（注２）スマートフォン端末上でテキスト入力からわずか 0.5 秒の高速生成が可能となった（図 1 参照）。

開発した 21 言語の音声合成モデルは、NICT が運用しているスマートフォン用の多言語音声翻訳アプリ VoiceTra（ボイストラ）（注３）のサーバに搭載され、一般公開されている。今後は、商用ライセンス等を通じて多言語音声翻訳やカーナビなどの音声アプリケーションへの導入が期待される。

なお、本成果は、2024 年９月に、International Speech Communication Association （ISCA）が主催する国際会議 INTERSPEECH 2024 の Show & Tell にて発表された。

（注１）人間の脳の働きを模した方法でデータを処理するようにコンピュータに教える人工知能の一手法であるニューラルネットワークを用いた音声合成のこと。テキスト系列を入力し音声波形を出力する。

（注２）パソコンや家電製品などの一連のシリーズの中で、性能や価格が中位の製品を指し、スマートフォンでは全体の 85%のシェアを占める。

（注３）Voice Tra（ボイストラ）について

（VoiceTra は、国立研究開発法人情報通信研究機構の登録商標）

＜利用に当たって＞

・サポートページ: https://voicetra.nict.go.jp/

・使い方の動画：https://voicetra.nict.go.jp/picture.html

【背景】

NICT のユニバーサルコミュニケーション研究所は、多言語音声翻訳技術の研究開発に取り組んでおり、研究成果を音声翻訳実証実験のために運用しているスマートフォン用音声翻訳アプリ「VoiceTra」で公開し、商用ライセンスを通じた社会実装を行っている。翻訳されたテキストを人間の声として読み上げるテキスト音声合成の音質は、ニューラルネット技術の導入により飛躍的に向上し、肉声に匹敵するまでになったが、膨大な計算が必要であり、ネットワークに接続されていないスマートフォンでの合成は、不可能であるという課題があった。また、多言語同時通訳においては、話者の発話終了を待たずに次々と翻訳音声を出力する必要から、音声認識や機械翻訳と同様、テキスト音声合成の更なる高速化が求められていた。

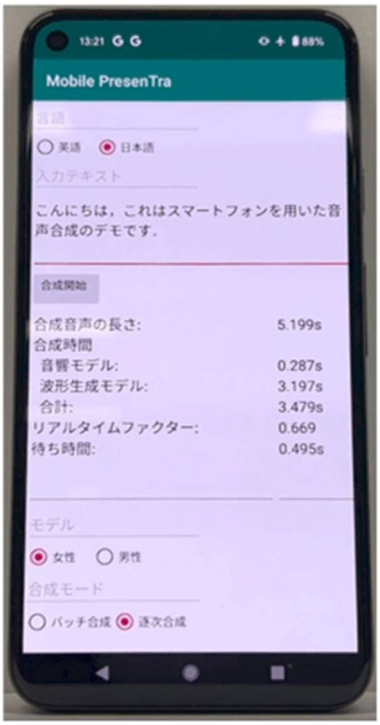

図１．ミドルレンジスマートフォンに実装した音声合成モデル

（動画: https://youtu.be/gD8HqE4lcbw）

引用元：https://www.nict.go.jp/press/2024/06/25-1.html

【今回の成果】

テキスト音声合成モデルは、入力テキストを中間特徴量(注4)へと変換する「音響モデル」(注5)と、中間特徴量を音声波形へと変換する「波形生成モデル」(注6)から構成される。ニューラル音声合成の「音響モデル」では、機械翻訳の分野や、音声認識や ChatGPT を始めとする大規模言語モデル等にも幅広く使われているニューラルネット（Transformer 型エンコーダ＋Transformer 型デコーダ）が主流だったが、近年画像識別の分野で使われ始めた高速・高性能なニューラルネット（ConvNeXt 型エンコーダ＋ConvNeXt 型デコーダ）を音響モデルに導入し、従来方式と比較して、品質を損なわず３倍の高速化を達成した[報告書 1]。また、肉声に匹敵する音声を合成可能な従来の「波形生成モデル」（HiFi-GAN）を発展させる形で、信号処理方式[報告書 2-4]を学習可能なニューラルネットとして表現するモデル（MS-HiFi-GAN）を 2021 年に導入し、合成品質を損なわず合成速度を２ 倍にすることに成功した[報告書５]。

そして、2023 年には同モデル（MS-HiFi-GAN）を更に高速化するモデル（MS-FC-HiFi-GAN）の開発に成功し、従来方式（HiFi-GAN）と比較して、品質を損なわず合成速度を 4 倍にすることを実現した[報告書 6、7]。

これらの成果の集大成として、上記で開発した「音響モデル（Transformer 型エンコーダ＋ConvNeXt 型デコーダ）」と「波形生成モデル（MS-FC-HiFi-GAN）」を用いた新しい高速・高品質なニューラル音声合成モデルを開発した（図 2 参照）。これにより、CPU コアひとつで 1 秒の音声をわずか 0.1 秒で高速合成することが可能となった。これは、既存モデルの約 8 倍の速さである。さらに、「波形生成モデル」のみを逐次合成する方式を実装することで（図 3 参照）、合成品質を損ねることなく、ネットワークに接続されていないミドルレンジスマートフォン端末上でも、テキスト入力からわずか 0.5 秒の高速生成が可能となった。これにより、これまでのサーバ経由での合成が不要となり、インターネット通信を必要とせず、通信コストを抑えたスマートフォンや PC 等での高品質ニューラル音声合成が可能となる。また、逐次合成処理により、多言語同時通訳においても翻訳テキストを即座に合成することが可能となった。2024 年 3 月から、VoiceTra の 21 言語[注7]の音声には、この音声合成技術が用いられ、一般公開されている。

本研究により開発した多言語合成音声は、2024 年 6 月 28 日（金）～29 日（土）の NICT オープンハウス 2024 における多言語同時通訳のデモ展示にて使用された。

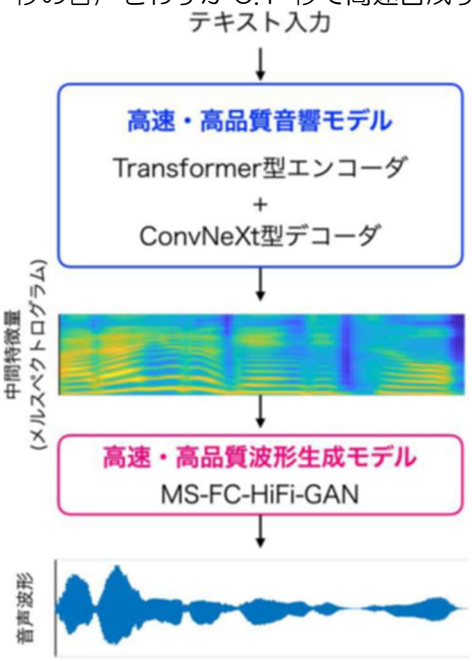

図 2．開発した高速・高品質な
ニューラル音声合成モデルの模式図

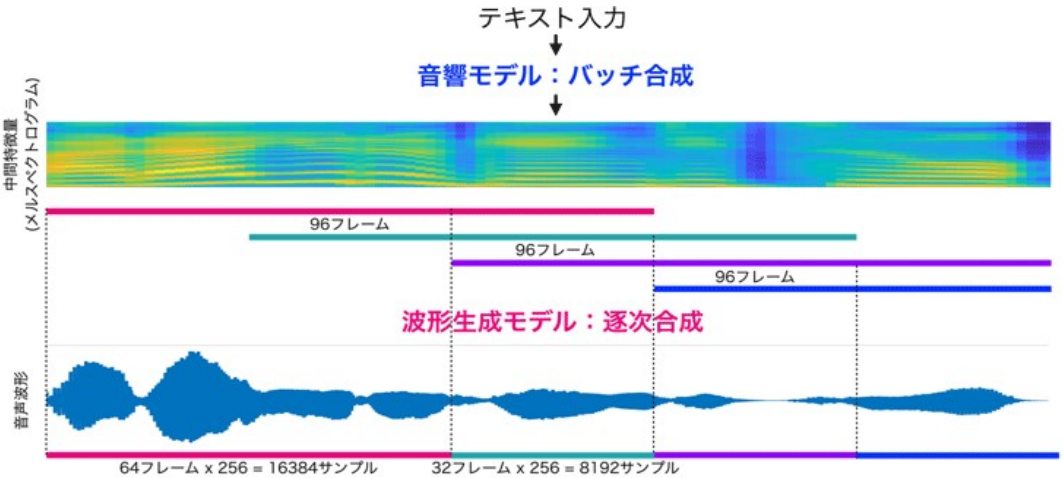

図３．波形生成モデルのみを逐次合成することにより待ち時間短縮を実現

引用元：https://www.nict.go.jp/press/2024/06/25-1.html

（注４）テキスト音声合成では、テキストを入力して音声波形を出力することが目的であるが、音声波形よりも時間解像度の低い中間特徴量を経由する方式が主流である。つまり、下記で説明する音響モデルを用いて入力テキストから中間特徴量を出力し、波形生成モデルを用いて中間特徴量から音声波形を出力する二段階の変換により入力テキストから音声波形を出力する（図２参照）。ここでの中間特徴量は、音声波形を短時間に区切り、区切られたフレームごとの音声信号に対して信号処理によって分析したものであり、音響特徴量とも呼ばれる。ニューラル音声合成においては、音声波形を短時間に区切り、区切られたフレームごとに周波数分析を施したメルスペクトログラムという音響特徴量が使われることが主流である（図 2 参照）。

（注５）テキスト音声合成における音響モデルとは、テキスト系列を入力して中間特徴量系列を出力する計算モデルである。10 年前までは隠れマルコフモデルに基づく方式が主流であったが、近年はニューラルネットに基づく方式が主流である。

（注６）中間特徴量系列を入力して音声波形を出力する計算モデルである。2016 年までは信号処理に基づく方式が主流であったが、2016 年からはニューラルネットに基づく方式が主流である。

（注７）21 言語とは、 日本語、英語、中国語、韓国語、タイ語、フランス語、インドネシア語、ベトナム語、スペイン語、ミャンマー語、フィリピン語、ブラジルポルトガル語、クメール語、ネパール語、モンゴル語、アラビア語、イタリア語、ウクライナ語、ドイツ語、ヒンディ語、ロシア語

【今後の展望】

今後は、商用ライセンスを通して、多言語音声翻訳やカーナビゲーションを始めとするスマートフォンアプリケーション等への社会実装を行う。

【論文紹介】

掲載誌: Proceedings of INTERSPEECH 2024

論文名: Mobile PresenTra: NICT fast neural text-to-speech system on smartphones with incremental inference of MS-FC-HiFi-GAN for low-latency synthesis

著者: Takuma Okamoto, Yamato Ohtani, Hisashi Kawai

【成果報告書】

[1] T. Okamoto, Y. Ohtani, T. Toda and H. Kawai, ″ConvNeXt-TTS and ConvNeXt-VC: ConvNeXt-based fast end-to-end sequence-to-sequence text-to-speech and voice conversion,″ in Proc. ICASSP, Apr. 2024, pp. 12456-12460.

[2] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga and H. Kawai, ″Subband WaveNet with overlapped single-sideband filterbanks,″ in Proc. ASRU, Dec. 2017, pp. 698-704.

[3] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga and H. Kawai, ″An investigation of subband WaveNet vocoder covering entire audible frequency range with limited acoustic features,″ in Proc. ICASSP, Apr. 2018, pp. 5654-5658.

[4] T. Okamoto, T. Toda, Y. Shiga and H. Kawai, ″Improving FFTNet vocoder with noise shaping and subband approaches,″ in Proc. SLT, Dec. 2018, pp. 304-311.

[5] T. Okamoto, T. Toda and H. Kawai, ″Multi-stream HiFi-GAN with data-driven waveform decomposition,″ in Proc. ASRU, Dec. 2021, pp. 610-617.
[6] T. Okamoto, H. Yamashita, Y. Ohtani, T. Toda and H. Kawai, ″WaveNeXt: ConvNeXt-based fast neural vocoder without iSTFT layer,″ in Proc. ASRU, Dec. 2023.
[7] H. Yamashita, T. Okamoto, R. Takashima, Y. Ohtani, T. Takiguchi, T. Toda and H. Kawai, ″Fast neural speech waveform generative models with fully-connected layer-based upsampling,″ IEEE Access, vol. 12, pp. 31409-31421, 2024. （神戸大学研修員山下陽生の研修成果）

専門家による目利きコメント

今回、NICT が開発した２１言語ニューラル音声合成技術は、オフラインのスマートフォン上でも高速動作できる。しかも、テキスト音声合成の音質は、肉声に匹敵するレベルに到達しつつあることから今後、多言語音声翻訳やカーナビゲーションといったスマートフォンアプリなどへの実装が大いに期待される。

お問い合わせ

＜ 本件に関する問合せ先 ＞
ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室　　岡本 拓磨
E-mail: ict@khn.nict.go.jp

＜広報（取材受付）＞
広報部　報道室
E-mail: publicity@nict.go.jp