多様なユースケースに適用可能な音声合成エンジン「Saxe」

近年では，深層学習等の技術進展，AI（人工知能）による人の活動の支援・代替といった社会的背景の変化に伴い，音声合成技術が必要とされるユースケース，要求される機能・性能が変化しつつあります．新たなユースケース適用への課題として，多種多様な「文脈に応じた読み分け」「話者性の再現」「動作環境」への対応があげられます．NTTメディアインテリジェンス研究所ではこれらの課題に対し，DNN（Deep Neural Networks）に基づく音声合成エンジン（開発コード「Saxe（サックス）」）を開発しています．本稿では技術概要，および適用事例について紹介するとともに，今後の展開について述べます．

井島 勇祐（いじま ゆうすけ）　小林 のぞみ（こばやし）
薮下 浩子（やぶした ひろこ）　中村 孝（なかむら たかし）
NTTメディアインテリジェンス研究所

はじめに

音声合成技術とは，入力されたテキストに対応する音声を生成する技術で，テキスト音声合成技術（TTS：Text-to-Speech Synthesis）とも呼ばれます．NTTでの音声合成に関する研究開発の歴史は長く，これまでに開発してきた音声合成技術は，web171（災害用伝言板），177（天気予報電話サービス），IVR（自動電話応答システム）といった電話サービスをはじめとした，「情報を正しく伝えること」を目的としたサービスで幅広く使われています．

一方近年では，深層学習をはじめとしたさまざまな技術進展，AI（人工知能）による人の活動の支援・代替の進展といった社会的背景の変化に伴い，音声合成技術が必要とされるユースケース，要求される機能・性能も変化しつつあります．これまでの「情報を正しく伝えること」を目的としたユースケースでは，「定型的な文章を」「特定の話者の声で」音声を生成することが求められていたのに対し，人の活動を支援・代替するユースケースでは，「多種多様な文章を」「所望の話者の声で」「多様な動作環境で」音声を生成することが求められています．NTTメディアインテリジェンス研究所ではこれらの課題に対し，DNN（Deep Neural Networks）に基づく音声合成エンジン（開発コード「Saxe（サックス）」）を開発し，多様なユースケースへの実応用を推進してきました．本稿ではその技術概要と適用事例について紹介し，最後に今後の展開について述べます．

技術概要

(1) 文脈に応じた同形異音語の高精度な読み分け

音声合成は，大きく分けて，入力されたテキストから読みやアクセントを推定する「テキスト解析部」と，推定された読みやアクセ

ントから合成音声を生成する「音声合成部」から構成されます（**図1**）．このうちテキスト解析部では，誤った読みやアクセントを推定してしまうと合成音声の聴取者に正しい情報を伝達することができないため，入力されたテキストに対して高精度に読みやアクセントを推定することが求められます．しかし日本語では同じ表記でも文脈によって異なった読みやアクセントとなる「同形異音語（例えば，「辛い（カライ/ツライ）」，「寒気（サムケ/カンキ）」など）」が存在しており，高精度な読みやアクセントの推定に向けた大きな課題となります．

そこで私たちは，明らかな読み誤りに対して正しい読みを推定する「読み曖昧性解消技術」を実現しました．この技術は，言語的な知見を活かした辞書と規則によって曖昧性のある語の読みを推定します．例えば，「カレー」という語が周辺に出現していれば「カライ」に加点する，という規則をあらかじめ用意しておくことで，「この店のカレーは辛いだけではない」という文における「辛い」という語は「カライ」が正しい読みであると推定します（**図2**）．ここで，「カライ」として考えられる語の表記を網羅的に書きつくすことは困難であるため，語のカテゴリ（例えば「食べ物」）なども規則として利用できる枠組みとすることで，規則数の削減と網羅性の向上を実現しています．この技術により，省メモリかつ高精度で正しい読みを推定することが可能となりました．

⑵　低コストで多様な話者性を再現するDNN音声合成技術

高精度な読みやアクセントの推定が要求されるテキスト解析部に対し，音声合成部では，顧客の要望などに応じた所望の話者の音声を高精度に再現することが要求されます．しかし，所望の話者で高品質な音声合成を実現するためには，その話者が発声した大量の音声データ（例えば，波形接続型音声合成方式

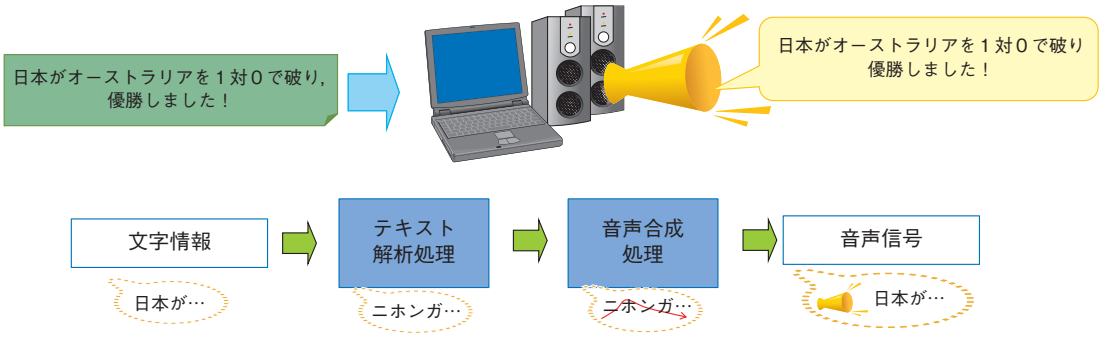

図1　音声合成技術の概略

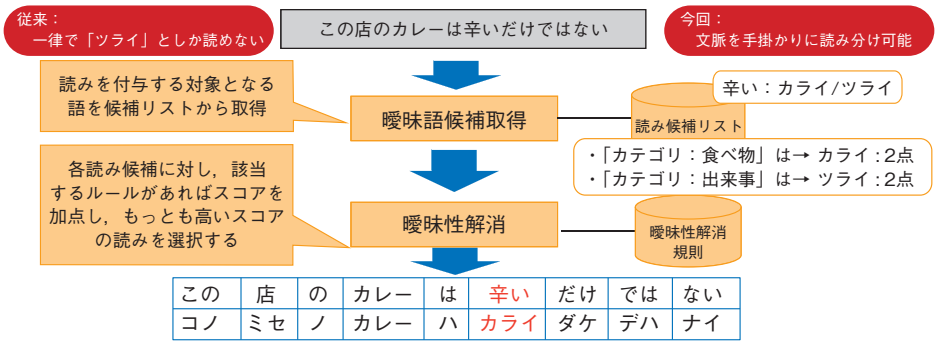

図 2　読み曖昧性解消技術の概要

Cralinet[(1)]では，高品質な合成音声を生成するためには数時間～20時間程度）が必要となります．そのため音声対話システム等で，さまざまなキャラクターの音声合成を実現するためには，音声収録やデータベース構築等のコストが大きな課題となっていました．

この課題に対して，私たちはこれまで整備してきた多数話者の音声データベースとDNNとを活用することで，20～30分の音声データ（2時間程度の音声収録）から所望の話者での高品質な音声合成を実現しています．この方式の特長は，複数の話者の音声データを1つのDNNでモデル化することです（**図 3**）．読みやアクセントといった音声を生成するために必要な情報は，あらかじめ用意してある多数話者の音声データから学習し，所望の話者の声質や話し方の特徴は，所望の話者の音声データから学習します．これにより，所望の話者の音声データは少量でも高品質な音声合成を実現しています[(2)]．さらに，画像生成等で有効性が示されているGAN（Generative Adversarial Networks）を組み合わせることで，合成音声の品質，話者の再現性のさらなる向上を実現しています[(3)]．

⑶　多様な環境で動作するDNN音声合成技術

音声合成が実際に利用される環境によっては，さまざまな制約（ネットワークに接続できない，高速なレスポンスが求められる等）により，計算リソース（CPU，ROM，RAM等）が潤沢である計算機サーバ上ではなく，計算リソースが非常に限られたスマートフォンやロボット等のデバイス上での動作が求められます．この課題に対して私たちは，合成音声の品質を可能な限り保ちながら，計算リソースが限られたデバイスにおいて実用的な速度で動作する，組み込み用DNN音声合成ライブラリを開発しました．具体的には，サーバ向けのライブラリに加えて，スマート

フォンやタブレットといったデバイス上で動作する「省リソース端末向けライブラリ」，さらには，マイコンや家電，高級玩具等といった計算リソースが大きく制限されたデバイス上でも動作する「超省リソース端末向けライブラリ」の3種類のラインアップをそろえています（**図4**）．

特にマイコン等においては，FPU（浮動

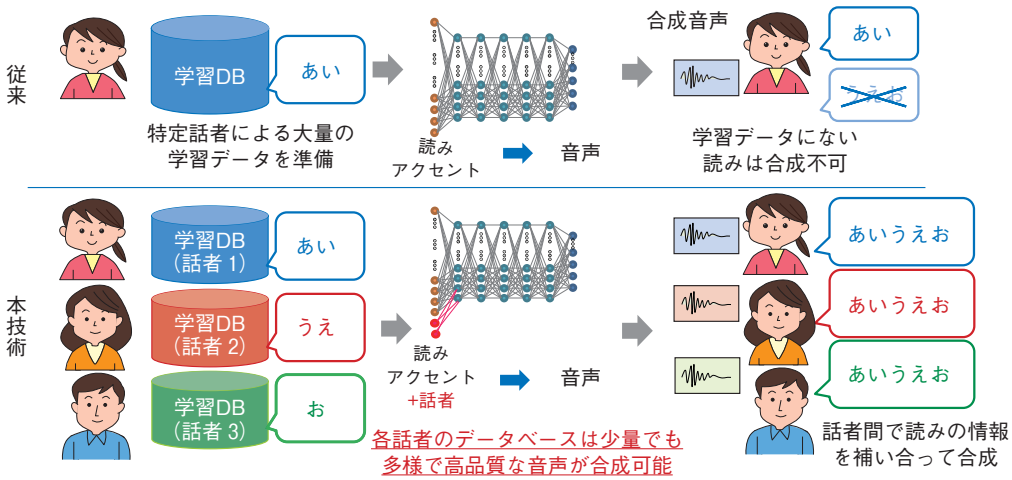

図3　多様な話者性を再現可能なDNN音声合成技術

特長	高品質な音声合成，多言語対応（日英中韓）	スマートフォン等で高品質な音声合成が可能な軽量音声合成ライブラリ	整数演算のみの低スペックCPUでの動作可能な超軽量音声合成ライブラリ
CPU	×86_64	ARM Cortex-Aシリーズ	ARM 9，MIPS 1など
メモリ	2 GB（ROM）/4 GB（RAM）	十数MB（ROM/RAM）	数MB〜十数MB（ROM/RAM）
用途	コンテンツ作成，ロボット，音声対話，エンタメ，情報提供等幅広く	高速応答が必要なロボット，ネットワークが使用できない環境での利用	安価，低スペックな機器での利用

図4　多様なデバイスで動作する音声合成エンジン

小数点演算ユニット）が搭載されていないことが多く，行列演算がほとんどを占めるDNNの推論処理をどのように高速化するかがポイントとなります．「超省リソース端末向けライブラリ」では，固定小数点演算を用いることで，浮動小数点演算を用いずに高速なDNNの推論処理を実現しています．加えて，テキスト解析部にも高速化の工夫を行うことで，FPUが搭載されていないデバイス上でも高速かつ省メモリ（ROM：7MB～）で動作する音声合成ライブラリを実現しています．

適用事例

⑴ CGアナウンサーのニュース読み上げ音声への適用

私たちの研究開発した音声合成技術をサービス提供するNTTテクノクロス「Future-Voice Crayon」[4]は，2020年2月よりテレビ朝日の「AI×CGアナウンサー　花里ゆいな」の音声合成として採用されました[5]．ニュース番組のため，アナウンサーに近い豊かな表現力に加えて，さまざまなカテゴリのニュース原稿の正しい読み上げ能力が求められます．前述の「読み曖昧性解消技術」により自動的にカテゴリに合った読みを付与し，これまでかかっていた人手による読み・アクセント修正の稼働削減に寄与しています．

また，本事例におけるCGアナウンサーの声は，テレビ朝日の複数名のアナウンサーの声を混合して作成しました．特定の人物の権利に依存しない独自の声をつくり出した取り組みとして，音声合成技術の新たな可能性を示しました．

⑵ ドコモAIエージェントAPI

NTTドコモ「ドコモAIエージェントAPI®」[6]は，音声・テキストユーザインタフェース（UI）をパッケージ化した対話型AIのASPサービスで，本サービスの音声合成エンジンとして私たちの音声合成技術が搭載されています．本APIでは，50種類以上の音声がプリセット話者として準備されており，利用者は都度の音声収録や権利処理等の手間なく，さまざまなキャラクターや環境に合わせた音声UIの実装が可能となっています．ここでは前述の「DNN音声合成技術」により，小学生からお年寄りの声まで，さまざまなバリエーションの話者性の再現をかなえています．

⑶ 減災コミュニケーションシステム

NTTデータの「減災コミュニケーションシステム®」[7]は，地方自治体から住民に向けて行政・防災情報等を伝達するための告知放送システムで，自治体庁舎内の送信システムや遠隔操作端末などから，地域内に配備した屋外スピーカ装置やタブレット端末，スマートフォン・携帯電話などへ情報を配信します．配信された情報を基に屋外スピーカ装置，タブレット端末，戸別受信端末等の各デバイスで音声合成を行い，合成音声で情報の伝達を行います．

今後の展開

本稿では，NTTメディアインテリジェンス研究所の音声合成技術の近年の技術開発とその実用事例について述べました．これらの取り組みにより，音声合成技術は入力されたテキストから所望の話者の合成音声を生成するという観点では，一定のレベルまで到達しています．

一方で，現在の音声合成技術と人の発声とを比較すると，まだまだ大きな差が存在します．例えば，アナウンサーや声優であればニュースやセリフを読むときは，テキストに含まれる意図等を理解したうえで，感情を込めたり声色で表現したりしますが，現在の音声合成技術では意図の解釈はできておらず，常に同じ調子の合成音声しか生成できません．人の活動を支援・代替することに対する期待が高まっている今，音声合成技術がより広く世の中に普及するためには，人と同等か，それ以上の表現が可能な音声合成を実現する必要があると考えています．今後は，そうした文脈・意図・感情に即した表現や，聞き手の属性・受容性を考慮した表現が可能な技術に取り組むことで，さらなる適用先拡大を図っていきたいと考えています．

■参考文献

⑴ 間野・水野・中嶋・宮崎・吉田：“顧客へのリアルな音声応答を実現するテキスト音声合成技術「Cralinet」，” NTT技術ジャーナル，Vol. 18，No. 11，pp. 19-22，2006.
⑵ N. Hojo, Y. Ijima, and H. Mizuno: “DNN-based speech synthesis using speaker codes,” IEICE Trans. on Information and Systems, Vol. E101-D, No. 2, pp. 462-472, 2018.
⑶ H. Kanagawa and Y. Ijima: “Multi-Speaker Modeling for DNN-based Speech Synthesis Incorporating Generative Adversarial Networks,” Proc. of 10th ISCA Speech Synthesis Workshop, pp. 40-44, 2019.
⑷ https://www.futurevoice.jp/
⑸ https://news.tv-asahi.co.jp/news_international/articles/000175834.html
⑹ https://docs.sebastien.ai/
⑺ https://www.nttdata.com/jp/ja/lineup/disaster_mitigation_c/

（左から）小林　のぞみ／井島　勇祐／
薮下　浩子／中村　　孝（右上）

音声合成は多くのユースケースで活用されており今後も拡大が期待されます．本稿で紹介した適用事例をはじめNTTグループ各社を通して，音声合成をお試しいただけます．ぜひご活用ください．

◆問い合わせ先
NTTメディアインテリジェンス研究所
心理情報処理プロジェクト
TEL　046-859-4301
FAX　046-855-1054
E-mail　gosei-produce-p＠hco.ntt.co.jp