URL https://journal.ntt.co.jp/article/33343
DOI https://doi.org/10.60249/25041301

NTT人間情報研究所
特別研究員

井島勇祐 Yusuke Ijima

表現力の豊かな音声合成技術により，人々の生活をサポートする

障がい者支援や高齢者サポートなどアクセシビリティの向上や，コールセンタやカーナビなど私たちの生活をサポートしてくれる音声合成技術．現在ではナレーションやゲームのキャラクタ音声を自動生成，声色を保ったまま多言語の音声の生成ができるなどの高度でグローバルな展開を遂げています．声優，タレントなどだけでなく，なんらかの事情で声を失ってしまった方の音声や動画があればその方の音声を復元することも可能になり，社会的な貢献度も高い音声合成技術ですが，より人間らしく話すためにはどのような課題があるのでしょうか．今回は最新の音声合成技術「Zero/Few-shotクロスリンガル音声合成」を開発された井島勇祐特別研究員にお話を伺いました．

◆PROFILE：2009年東京工業大学大学院総合理工学研究科博士前期課程修了．同年，日本電信電話株式会社入社．以来，テキスト音声合成，音声変換をはじめとした音声情報処理に関する研究開発に従事．現在，NTT人間情報研究所勤務．2021年前島密賞奨励賞受賞．博士（工学）．

数秒の録音データでその人に似た音声合成ができる「Zero/Few-shotクロスリンガル音声合成」

■現在どのような研究をされているのか，お伺いできますでしょうか．

音声合成技術とは文字を音声に変換するメディア変換技術のことで，音声合成の処理は読みやアクセントを付ける「テキスト解析処理」と，抑揚・声色・音声波形をつくる「音声生成処理」から成り立ちます．私は学生時代からずっと音声処理について研究しており，当時は音声をテキスト化する音声認識，音声に含まれる感情を認識する音声感情認識を研究していましたが，NTTに入社してからは音声合成技術に関する研究をしています．音声合成技術自体は昔から研究されていますが，最近の技術進展により以前と比較すると自然な合成音声が生成できるようになってきています．

音声合成技術は，災害などで活用される安否確認（Web171など）・コールセンタのガイダンスなど電話での情報提供サービス，カーナビなど，ユーザに「情報を正確に伝える」ことが目的のサービスで多く使用されています．一方で，最近のLLM（Large Language Models）や音声合成の技術的な進歩により，エンタテインメント向けの応用，キャラクタなどとの対話サービスでの需要が増加しています．従来の「情報を正確に伝える」ことが目的のサービスでは，あらかじめつくり込まれた特定の話者の音声を生成することができれば十分でしたが，このような用途の場合，コンテンツなどの制作者の要望に沿った声を，低コストかつ高精度で再現できることも重要になってきました．

例えば音声合成技術で「この人の声で喋らせたいみたい」という要望があったとき，一般的な方法としてはその人の音声を収録し，収録音声に対して人手で発音（読みやアクセントなど）を付与して音声合成モデルを学習します．しかし，音声収録やアノテーションのためのコストやリードタイムがかかるため，コンテンツ制作などに気軽に活用するのは困難な部分があります．

そうした背景から大量の音声収録や，収録音声に対し読みやアクセントを付与する作業を行わずに，数秒の音声だけからその人に似せた声で音声合成が可能な「Zero/Few-shotクロスリンガル音声合成」を開発しました．Zero-shot音声合成は数秒の音声データがあればその人に類似した声で，さまざまなテキストの合成音声を作成することができますし，数分の音声データがあればその人の話し方の特徴などをより高精度に再現することが可能です．また，クロスリンガル音声合成技術では，その人の声色を保ったまま英語，中国語などの多言語の合成音声を生成することが可能です（**図**）．

この技術を活用することで，タレントや声優などの著名人，ユーザ個人の家族などの声で応対が可能なロボットやAIエージェントの実現が可能です．また，クロスリンガル音声合成技術を活用することで，外国人観光客などに対してキャラクタの声色を日本語から変えずに多言語での応対ができるようになります．

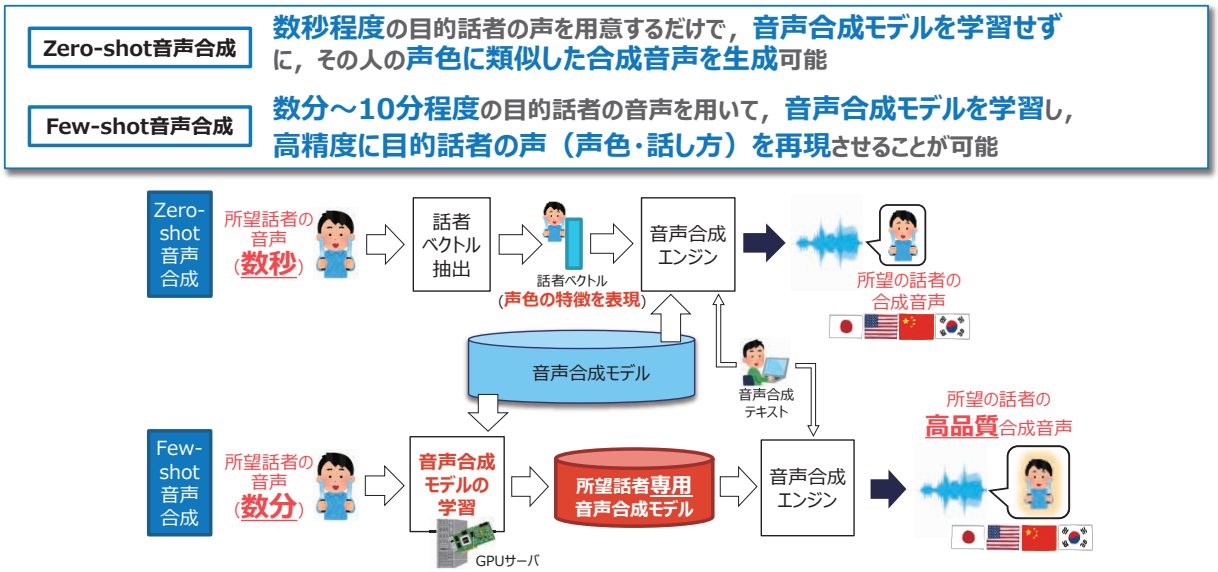

図　Zero/Few-shot クロスリンガル音声合成

技術開発だけで終わるのではなく，実際に使ってもらうことで研究に価値が生まれる

■この研究によりどのようなことが実現できるのか，社会にもたらす影響があればお伺いできますでしょうか．

企業の研究者として，研究成果を出すということは重要だと思いますが，それ以上に研究成果を実用化し，NTTグループ内外の皆様に実際に使っていただくということを重視しています．特に私たちの取り組んでいる音声，言語，画像などを取り扱うメディア処理の分野では，研究と実用化の境界が非常にあいまいで，技術進展も非常に速くなっています．こういった状況では，世に出すことができなかった研究は失敗だったとみなすぐらいの気持ちで取り組まなければならないと考えています．また，実際に使用していただくためには，NTTグループ内外の皆様との連携が不可欠で，さまざまな方々との連携を進めています．

その１つが病気などで声を失ってしまった方の声の再現です．病気や事故などで話すことが難しくなった方は多くいます．そういった方々のホームビデオなどの動画に含まれているご自身のわずかな音声から，ご自身の声を再現することができないか取り組んでいます．

また，コンテンツ制作での活用に向けて，NTT西日本との連携を進めています．この取り組みでは，キャラクタなどの権利を保有するIP（Intellectual Property rights）ホルダー，芸能事務所の方々と連携することで，キャラクタや著名人の声を活用したサービスの提供をめざしています．近年では音声広告やオーディオブックなどの音声コンテンツの需要が増えてきており，その際にキャラクタや著名人の声でという要望は多くあります．一方でそういった方々の稼働時間が限られているのが現状です．私たちの技術を活用し，稼働時間を仮想的に増加させることで，キャラクタ，著名人の活躍する機会を増やすことが期待されています．またクロスリンガル音声合成技術も活用することで，キャラクタや著名人の声色を保ったまま多言語でのコンテンツを作成することができるため，キャラクタなどの海外展開をサポートできないかと考えています．

それ以外にも，これまでは定型文での情報提供が中心に行われてきたコールセンタにおいて，お客さまとの応対の高度化に向けて，よりオペレータらしい自然な表現の合成音声を生成することができないかというお話もいただいています．

■研究における課題やポイントを教えてください．

現在の音声合成技術は，数秒の音声からその人に似た声を生成するなど，一昔前では難しかったことができるようになってきています．一方で，発声のプロであるアナウンサー，声優の方々と比較すると，合成音声の表現力には大きな差があると考えています．私は業務でそういった方々とお仕事をさせていただいていますが，そのたびに現在の音声合成技術とプロによる表現力の差を痛感させられます．

例えば，実際の音声収録の現場では，単に用意された原稿を読むだけではなく，音響監督やディレクターが理想とする音声を収録するために，音響監督などからの表現に対する指示を受け，そ

れに従った表現の修正を繰り返し行います．このようなことを実現するためには，細やかな指示の意図を理解し，それを音声の表現に反映させることが必要になります．しかし現在の音声合成技術ではそういった細やかな表現・ニュアンスの演じ分けなどは非常に難しいのが現実です．それ以外にも，台本や小説から登場人物の心境や人間関係を汲み取って，それを声によって表現することができる表現力の多様性にも多くの課題があります．前述のコンテンツ制作，コールセンタなどで要望されていることを実現するためには，こういった課題の解決が不可欠だと考えています．

音声合成技術でさまざまな音声の生成が容易にできるようになった一方，インターネット上から取得した著名人の音声を使用して，その方の声を使用したフェイク動画などの問題が顕在化しており，発声者の権利をどのように担保するかも非常に重要な問題です．数年前であれば，著名人の音声合成を活用したサービスを提供することに対して，ご本人や事務所からネガティブな印象を持たれることはあまりありませんでしたが，現在はネガティブな印象がかなり増加してきています．

音声合成技術をより広く活用していただくためには，こういった懸念に対して，法律的，技術的な側面での対応が必要不可欠なため，NTTグループ内でもさまざまな取り組みがなされています．NTT社会情報研究所では，法律的な観点から発声者の権利に関する研究を進めていますし，NTT西日本では技術的な観点から音声合成技術により生成された音声をトレースする技術検討が行われています．こういった方々と連携することで，単に音声をつくるだけではなく，権利の保護も含めた取り組みを進めたいと考えています．

■ビジネス観点での難しさについて教えてください．

ビジネスとして成り立たせようとしたときには，ランニングコストなどの観点も重要です．現在の技術では，例えばモデルサイズを大きくすることで品質向上や新しい機能の実現もできます．しかし数十万人，数百万人が利用するような場合は，サーバなどの維持管理費が膨大となるため，サービスとして成り立たないという可能性もあります．このバランスを含めて考えながら，研究開発をどう進めていくのかが難しいところです．

また，私たちのチームとして音声合成プロダクトを出すためには，実用化開発のタイミングで最終的にどの技術を次のプロダクトに採用するかを判断しなければいけません．プロダクトとしての競争力を維持し続けるためには，「必ず採用しなければならない技術」だけではなく，「実用化の段階で差別化の要素となり得る技術」も必要となります．限られたチームのメンバでそれらのバランスを取りながら研究開発を行っています．

■若き研究者の方や学生，ビジネスパートナーの方へのメッセージをお願いします．

企業の研究者として仕事をしている以上，自分が取り組んでいる仕事が革新的な研究成果を出せるのか，実用的なビジネスができるのか意識することが重要だと思います．もし，そのどちらもやれていないのであれば，取り組んでいる研究テーマ，業務内容を変えるなどの判断が必要なのかもしれません．現在のメディア処理やAIに関する分野は，研究と実用化の境界が非常にあいまいで，少し気を抜くと学術的にも実用にも供さない中途半端なことをやりがちになってしまいます．私も自身の業務がどういう役割で何を求められているのか，という意識を常に持つようにしており，どのような業務においてもその意識を忘れずに取り組むことが大切だと思います．

NTTは専門的な人材のバリエーションの広さが大きな魅力だと思っています．私は音声処理を専門としていますが，社内には音声処理をはじめさまざまな分野の専門家が非常に多くいます．私の専門とは違う分野のことをちょっと聞きたいときに，例えば「この技術をプロダクトに使いたいけれど，どうすればいいの？」と，チャットで聞けば，すぐ教えてくれるといったことができるのは，ほかの会社にはないNTTならではの魅力だと思います．

（今回はリモートにてインタビューを実施しました）