

NTTコミュニケーション科学基礎研究所  
 首席特別研究員

**亀岡弘和** Hirokazu Kameoka

## 声の印象を自由にカスタマイズできる最先端の音声変換技術で、コミュニケーション機能のさらなる拡張をめざす

コミュニケーションには、障がいや加齢による衰え、不慣れな言語での会話など、物理的・能力的・心理的な状態に起因するさまざまな制約が存在します。そこで、状況に適した音声ヘリアルタイムに変換し、制約を解消したコミュニケーションを実現する「コミュニケーション機能拡張技術」が注目されています。音質の変換にとどまらず、訛り、ささやき声、電気音声など韻律の変換も可能とした研究、また最近では「かわいい声にしたい」「りりしい声にしたい」など人へ与える印象を大切にする社会的な要望にもこたえるべく、音声の主観的印象を自由に変換させる研究の第一人者である、NTTコミュニケーション科学基礎研究所 亀岡弘和首席特別研究員に、これまでの研究成果を振り返っていただくとともに、最新のオリジナルな研究の取り組み状況や具体的な成果、さらには研究に対する姿勢について伺いました。



### 画像生成など他分野で活用される技術を音声変換方式へいち早く取り入れ、オリジナリティの高い研究に臨む

現在、取り組まれている音声変換の方式研究について教えてくださいませんか。

私たちはコミュニケーション機能をいかにして拡張するかを目標に置き、機械学習や信号処理の研究を進めています。その中でも音声変換のテーマをメインに取り組んでいます。広い意味で音

声変換とは「声を入れて声を出す」というプロセスで、混成された音源からの分離というタスク、声色に相当する声質を別人のものへ置き換えるといったタスクなどがあります。また、声質以外にも訛りの変換、声を自力で出せなくなった人向けの電気音声から通常声への変換などがあります。通常の音声へ変換するほかにも、感情のない声から感情豊かな声に変換するものまで、幅広く研究を行っています（図1）。この声質変換のテーマについては、さまざまな最先端技術によるアイデアをたくさん提案しており、後ほど詳しく紹介します。

#### コミュニケーションにおける物理的・能力的・心理的な状態に起因する制約



#### 信号のリアルタイム変換により制約を解消したコミュニケーションを実現



#### 開発中の技術による変換例

- 音源分離**
  - 混合音声から聴取したい声
- 声質変換**
  - 話者Aから話者Bの声
  - 現状の最先端技術の要素となるアイデアをこれまで多く提案
- 声質以外も変換**
  - 訛りありからなしの声
  - 電気音声から通常声
  - 通常声から感情音声
  - ささやき声から通常声

系列変換 (S2S) モデルに基づく音声変換法を考案したことで声質だけでなく他の非言語特徴も変換可能に (IEEE-TASLPIC 3件採録)

図1 NTTグループにおける研究目標

音声変換 (Voice Conversion, VC) は、一般に、学習データを用いて音声特徴量の変換則 (変換モデル) を事前に獲得し、その変換モデルを入力音声に適用することで実現されます。図2に示すように、この研究分野は大きく2つの観点から分類できます。第一は、学習に必要とされるデータの性質に基づく「**パラレル方式**」と「**非パラレル方式**」です。

パラレル方式とは、音声変換モデルの学習に際して、同一文を異なる話者が発話した音声ペア (パラレルデータ) を用いる方式を指します。例えば、ある話者による「こんにちは」という発話と、別の話者による「こんにちは」という発話の組み合わせがこれに該当します。十分量のパラレルデータを収集できる場合、この方式によって高精度な変換モデルを得ることが可能です。しかし、パラレルデータの収集には多大なコストが伴うため、実用上の制約が大きいという課題がありました。

この制約を克服するために提案されたのが非パラレル方式です。この方式では、文単位で対応付けられていない任意の発話データからでも変換モデルの学習が可能となります。私たちが研究を開始した当初は、パラレル方式が主流でした。しかし、当時画像分野で注目され始めていたアンペアード・スタイル変換の枠組みを音声変換にも応用できるのではないかと着想し、非パラレル方式による手法を次々に提案しました。代表例として、CycleGAN-VCやStarGAN-VCが挙げられます。さらに近年では、私たちのチームメンバーが提案したPRVAE-VCが注目を集めています。

第二の分類は、変換対象を声質に限定するか、あるいはより広範な音声特徴へと拡張するかという観点です。従来の多くの音声変換手法は、声質、すなわち「特定の話し方らしい声色」の変換に限定されていました。これに対し、声質のみならず韻律やアクセントといった非言語的特徴の変換を可能にしたのが「**系列変換方式**」です。韻律には抑揚やリズムのパターンが含まれ、とりわけ感情表現や発話スタイル、話し方の癖などが該当します。さらに、英語におけるアクセントや発音様式も声質以外の音声特徴として含まれます。これらの特徴を変換するためには従来方式では限界があり、新たなアプローチが必要であると考えました。そこで、当時音声認識や機械翻訳で注目を集めていた系列変換モデルに着目し、その応用を検討しました。当時、音声変換に系列変換モデルを導入した事例はほとんど存在しませんでした。私たちがこの枠組みに基づく初めてのアプローチを提案し、声質に加えて韻律の変換が可能であることを示しました。

加えて、近年私たちが注目しているのが、変換音声の特徴を柔軟にカスタマイズ可能とする方式です。従来の多くの手法では、あらかじめ変換目標を定め、それに応じた学習データを収集・準備したうえで変換モデルを構築してきました。しかし実際の利用場面では、「このような声にしたい」といった多様な要望が状況に応じて生じます。その要望に対応する変換モデルが存在しない場合、既存の学習済みモデルを組み合わせ、利用場面に応じて柔軟に変換を実現するアプローチが求められます。この観点から私たちが開発したのが、世界で初めて拡散モデルを応用した音声変

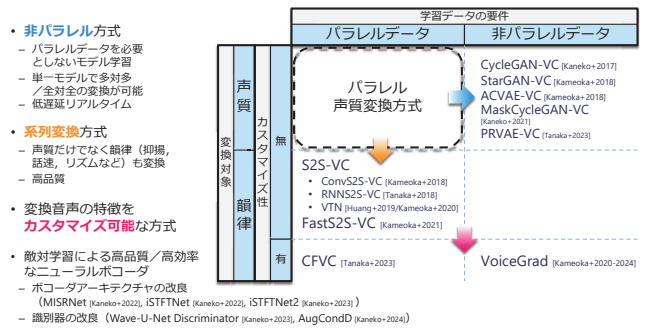


図2 高音質かつ低遅延なリアルタイム音声変換

換方式であるVoiceGradです。

さらに、音声変換においては、特徴量変換後に音声信号を生成するプロセスも重要です。この生成過程は、リアルタイムシステムにおける遅延時間や最終的な音質に直結します。そのため、変換モデルの研究と並行して、高速かつ高品質な音声生成を可能にするアーキテクチャの設計についても研究を進めています。

これら一連の研究は、名古屋大学や東京都立大学との共同プロジェクトとして、2024年度末までの5年間にわたり実施してきました。本プロジェクトは科学技術振興機構 (JST) のCRESTに採択され、明確な研究目標を共有する中で、多様な専門性を持つ研究者が集い、深くかつ多角的な議論を重ねることができました。また、大学から多くの学生を実習生として受け入れることで、研究を効率的に推進することができました。こうした協働体制のもと、各メンバーが共通の課題意識を持って研究を進められたことは、本プロジェクトの大きな成果であったと考えています。

### 非パラレル方式のPRVAE-VCにニューラルポコーダを改良したiSTFTNetを組み合わせて、低遅延かつ高品質なリアルタイム音声変換を実現できたそうですね。

音声特徴量から音声波形を生成するプロセスは、ニューラルポコーダによって実現されます。2017年ごろに登場したWaveNetは、その高い音質により大きな注目を集めましたが、波形生成に膨大な計算時間を要するため、リアルタイムアプリケーションへの適用が音声研究者の間で大きな議論となりました。その後およそ8年の間に、音質を維持しつつ高速化を実現する軽量化研究が急速に進展し、私たちにとても重要な研究課題の1つとなっています。特にリアルタイム音声変換においては、遅延はほとんど許容されず、わずかな軽量化でも大きな効果をもたらします。

音声は「音素」と呼ばれる最小の単位に分割できますが、その平均持続時間は約100 msとされています。自分の声が100 ms以上遅れて耳に戻ると、知覚される音素が実際に発話した音素と異なり、聴覚フィードバックに齟齬が生じます。その結果、脳が混乱し、発話が著しく困難になることが知られています。このため私たちは、100 msを大きく下回る50 ms以内の処理を目標としています。

近年では、HiFi-GANと呼ばれるニューラルポコーダが高品質

PRVAE-VC + iSTFTNet による 低遅延かつ高品質なリアルタイム音声変換

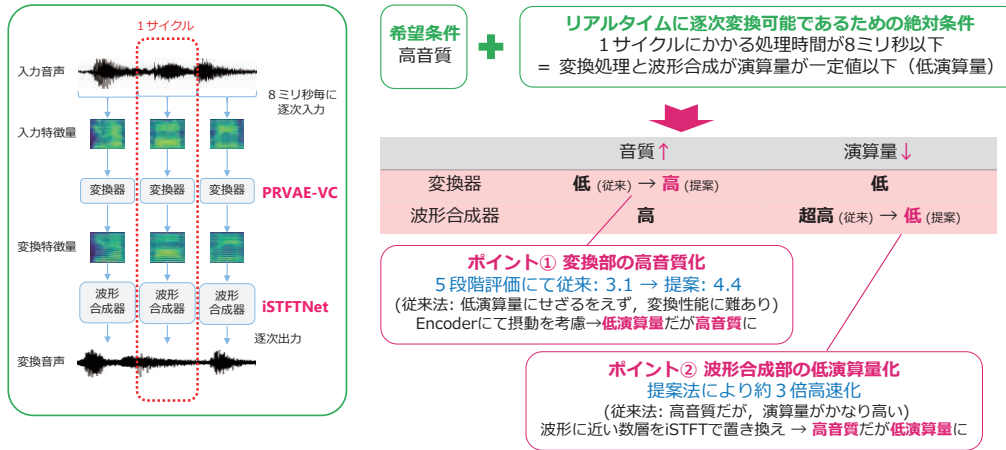


図3 高音質かつ低遅延なリアルタイム音声変換

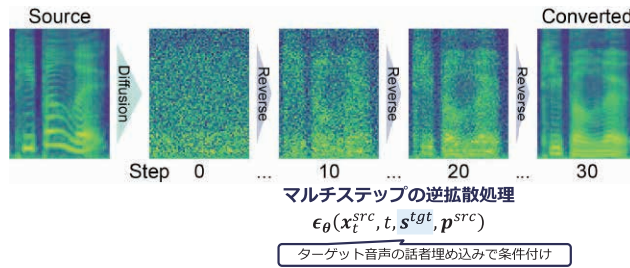


図4 VoiceGrad

かつ低遅延な波形生成技術として広く用いられています。しかし、標準的な計算資源を用いた場合、HiFi-GANで音声波形生成を50ms以内に収めることは依然として難しいことが分かっています。

これに対して、私たちのチームが2年前に提案したiSTFTNetは、50ms以内での波形生成を可能とする方式であることを実証しました。さらに、このiSTFTNetと前述の非パラレル方式PRVAE-VCを組み合わせることで、低遅延かつ高品質なリアルタイム音声変換を実現しました(図3)。参考までに、参考文献(1)~(3)の変換音声サンプルをお聴きいただければと思います。

この成果は、NTTコミュニケーション科学基礎研究所オープンハウス(4)やNTT R&Dフォーラムにおいて、新しいリアルタイム音声変換技術として展示され、多くの注目を集めました。

声の印象を自由にカスタマイズできるVoiceGradは将来楽しみな方式ですね。

拡散モデルに基づく方式であるVoiceGradを私たちが初めて提案(プレプリントとして発表)したのは2020年ごろのことです。その後、ようやく論文誌に掲載され、現在はこれを基盤として多様な研究を展開しています。本方式の中核となる拡散モデルは、現在では画像生成の分野で広く知られるようになっていますが、VoiceGradはその拡散モデルを初めて音声変換に応用した研究

です(図4)。提案当時、画像生成分野では敵対的生成ネットワーク(GAN)が主流であり、拡散モデルはまだ現在のように注目を集めていない時期でした。しかし、拡散モデルのアイデアを初めて目にした際、その強力さを直感的に理解し、「音声変換にも必ず応用できる」と確信して、いち早く取り入れました。

当時特に魅力的だと感じたのは、拡散モデルが画像や音声の生成を一種の最適化問題としてとらえるアプローチであった点です。最適化問題では目的関数を最小化(あるいは最大化)することが基本となりますが、主目的に加えて副次的な目的関数を設計し組み合わせることで、複数の目標を同時に満たす解を導くことが可能となります。これは正則化の一手法ですが、拡散モデルに基づくアプローチでは、この正則化の枠組みを利用して、利用者の要望に応じた柔軟な音声変換、すなわちカスタマイズ可能な音声変換を実現できるのではないかと考えました。

例えば、ある利用者が「より可愛い声」を望む場合や「力強い声」を求める場合でも、自身の声の特徴を残しつつ、目的に沿った声へと変換することが可能になります。言い換えれば、異なる要求を表す複数の目的関数を設計し、それらを同時に最適化することで、両立する音声変換が実現できるのです(図5)。

VoiceGradの検討を始めた当初は、従来の非パラレル方式と同様に、変換対象は声質に限られると考えていました。しかし研究を進めるうちに、VoiceGradが系列変換方式に近い能力を有し、

挑戦する研究者たち

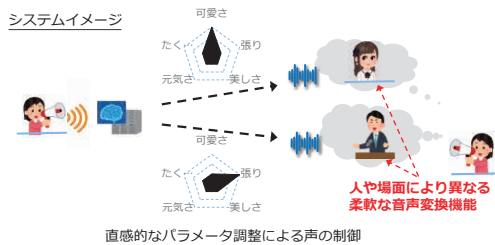


図5 音声の主観的印象制御の取り組み

声質のみならず韻律的特徴の変換も柔軟に実現できることが明らかになってきました。この点で、これまで私たちが提案してきたGANや変分自己符号化器(VAE)ベースの非パラレル方式とは一線を画す方式であると考えています。さらに最近の研究では、VoiceGradを改良することで、英語のアクセント変換も可能であることが確認されました。これは新たな知見であり、本方式が持つ潜在能力の高さを示すものです。

現時点では声質変換のサンプルのみを用意していますが、VoiceGradの変換性能に興味をお持ちであれば、ぜひ参考文献(5)よりお聴きいただければと思います。

## 👤 VoiceGradは将来性の高い音声変換方式。この技術的課題に挑むのは面白い

今後の展望を教えてください。

前述のように、VoiceGradは非パラレル方式であることに加え、声質のみならず韻律的特徴の変換も可能であり、さらに高いカスタマイズ性と優れた変換性能を併せ持つなど、多くの利点を備えています。そのため、現在私たちはVoiceGradの発展を今後の研究における重要な方向性として位置付けています。

一方で、技術的課題も明らかになりつつあります。その最たるものが「低遅延リアルタイム化」の実現です。現状のVoiceGradのアーキテクチャは、信号処理の用語で言うところの非因果的構造をとっており、未来の入力情報を参照するために必然的に遅延が生じます。これに対し、リアルタイム性を備えた音声変換を実現するためには、因果的構成——すなわち、ある時点の音声を変換する際に、その時点以前の入力情報のみを利用する方式——が不可欠です。図2右上に示した従来の非パラレル方式の手法はいずれもこの因果的制約を満たしており、低遅延リアルタイム動作が可能ですが、VoiceGradは现阶段では非因果的方式に依拠しており、未来の入力を必要とする点が課題となっています。したがって、因果的構成に基づくVoiceGradの実現が次なる研究課題です。

さらに、初期のVoiceGradは反復計算を必要としたため、推論——すなわち実際の音声変換——に多大な計算時間を要するという問題もありました。これらは現在取り組むべき主要な技術的課題であり、同時に非常に興味深い研究テーマでもあります。これらが解決されれば、VoiceGradは一層大きな可能性を拓くも

のと期待しています。

その試みの1つがFastVoiceGradです。これは依然として非因果的構造を残してはいるものの、従来は複数回の反復を要していた処理を、音質を損なうことなく単一ステップに短縮できることが明らかになってきました。現在はさらに因果的制約を組み込む研究を進めており、近年になって有望な実験結果も得られ始めており、実現への見通しが立ちつつあります。

## 👤 点と点を線でつなぎ空間を埋めていくことが大切

研究者として日頃心掛けていることを教えてください。

私は、知識や経験を単なる「点」として個別にとらえるのではなく、それらを「線」で結びつけていくことが極めて重要であると考え、常に意識しています。空間上に点を散在させるだけでは、その空間を埋め尽くすことは容易ではありません。しかし、それらの点どうしを線で結びつけていくことで、効率的に理解を広げ、空間全体を埋めていくことが可能になります。つまり、ある知識を得た際にそこで止めるのではなく、既存の知識と関連付けることで、理解をより汎用的かつ普遍的なものへと発展させることが、研究者にとって重要であると考えています。

この姿勢は、ニューラルネットワークの学習原理にも通じるものがあります。ニューラルネットワークが強力である理由は、個々の学習サンプルを単に暗記するのではなく、複数のサンプルが集まることで、その間を補間し、存在しないサンプルまでもあたかも存在するかのように学習できる点にあります。研究者が多様な知識や経験を取り込む際も同様に、それらをいかに結び付け、普遍的な理解として体系化するかが本質的に重要であると、私は常々感じています。

また、人工知能(AI)をはじめ、日々新しい研究成果が次々と発表されていますが、それらを単に「面白い」で終わらせるのではなく、自らがこれまで蓄積してきた知見との関係性を意識的に結び付けていくことで、理解はより普遍的で深みのあるものへと昇華していきます。私はそのような姿勢を持ち続けることが、研究活動において極めて大切であると考えています。

### ■参考文献

- (1) <http://www.kecl.ntt.co.jp/people/tanaka.ko/projects/prvaevc/>
- (2) <https://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/istftnet/>
- (3) <https://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/istftnet2/>
- (4) <https://www.tv-osaka.co.jp/news/53399/>
- (5) <https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/latentvoicegrad/index.html>