# あなたの声はどんな声？どんな声でしゃべりたい？

**人と人とのコミュニケーションには，物理的・能力的・心理的な状態に起因するさまざまなかたちの制約が存在します．私たちは，ある音声を異なる音声や望まれる音声へと変換する技術の研究を通じてこのような制約を取り除き，あらゆる人が不自由なく快適にコミュニケーションを行える環境を実現することをめざします．本稿では，「近年著しく発展し続けている深層学習を活用することによって音声をどのように変換できるのか？」を題材に，従来技術の課題と私たちの取り組みを紹介します．**

田中 宏（たなか こう）　金子 卓弘（かねこ たくひろ）
北条 伸克（ほうじょう のぶかつ）　亀岡 弘和（かめおか ひろかず）
NTTコミュニケーション科学基礎研究所

## 非言語情報を「変換」する

音声は，言語情報だけでなく話者性などの非言語情報も伝達できるという大きな特徴を有し，利便性に特に優れたコミュニケーション媒体であるため，人々がお互いにコミュニケーションを取るうえで基本的なツールの1つとなっています．発話することで，自分・相手の意図や感情を，伝える・理解することができるため，音声の特徴（例えば，抑揚や声質・リズム）をその時々で変化させることで，相手へ与える印象を変えることもできます．しかしながら，一個人の生成できる音声の表現力は身体的・能力的・心理的制約により制限されてしまいます．この制約を超え，発話者が所望の音声で思いのままに表現できるよう能力の拡張を行う技術が音声変換です．その適応先は，話者性の変換や発声障がい者補助，感情などの発話スタイル変換，語学学習のための発音・アクセント変換など，多岐にわたります．これらの利用シーンに応じて，変換したい音声特徴・学習データ・リアルタイム性に関する要件など，さまざまな前提条件が想定されます．私たちは，高品質であること，少量データ・非パラレルデータ*で学習可能であり効率的であること，リアルタイムに音声変換が動作すること，声質だけでなく抑揚やリズムといった超分節的特徴などの柔軟な変換が可能であること，上記の4点が音声変換において重要な要件であると考えています．以降，これら4点に着目した私たちの具体的な取り組みについて紹介していきます（**図1**）．

## 音声×深層生成モデルの取り組み

従来技術において代表的なものは，混合ガウス分布に基づく統計的声質変換[(1)]です．入力音声の特徴量から目標音声の特徴量への変換関数を得るために，事前に時間整合をとった入力音声と目標音声の同一発話文（パラレルデータ）を用意することで，両音声の特徴量の同時確率を記述したモデルです．また，近年では，前述のパラレルデータを必要とする枠組みにおいて，性能改善のため，ニュー

＊非パラレルデータ：入力音声と目標音声とで発話内容が異なるデータ（非同一発話文）を示します．

ラルネットワークを用いた手法や非負値行列因子分解などを用いた事例ベースの手法の検討も進められています．しかしながら，これら従来技術には，①学習データとして同一発話内容の音声ペアが必要であったり，②変換可能な音声特徴が声質に限られていたり，③音声の特徴量から波形を合成する際に古典的なボコーダを用いているため，合成される音声と実音声は容易に聞き分け可能であるなど，技術的制約があります．

一方，画像認識や自然言語処理の分野において，2014年以降，変分自己符号化器（VAE: Variational Auto-Encoder），敵対的生成モデル（GANs: Generative Adversarial Networks），系列変換モデル（Seq2Seq: Sequence-to-Sequence model）など，非常に興味深い深層学習モデルが台頭してきました．Seq2Seqの1モジュールである自己再帰型モデル（AR: Auto-Regressive model）とVAE，GANsを総して三大深層生成モデルと呼ばれることもあり，画像処理や自然言語処理，音声信号処理などさまざまな分野・タスクでその有効性が確認されています．また，2015年中期の機械翻訳タスクにて，注意機構（Attention mechanism）[(2)]がニューラルネットワークに導入され，その有効性とともに瞬く間に注目を浴びました．

NTTコミュニケーション科学基礎研究所では，前述の従来音声変換技術の課題を克服しさまざまな利用シーンに柔軟に対応可能な多用途音声変換システムを実現すべく，敵対的学習や系列変換学習などの方法論をベースに，①声質だけでなく長期依存特徴である韻律やアクセントの変換を行える「音声系列変換機能」，②発話内容に制約のない非パラレルデータを用いて声質変換を行える「非パラレル声質変換機能」，③合成音声波形から実音声波形へ波形空間上で変換を行い，出力音声の高音質化を実現する「波形ポストフィルター機能」などの新機能を創出しました．結

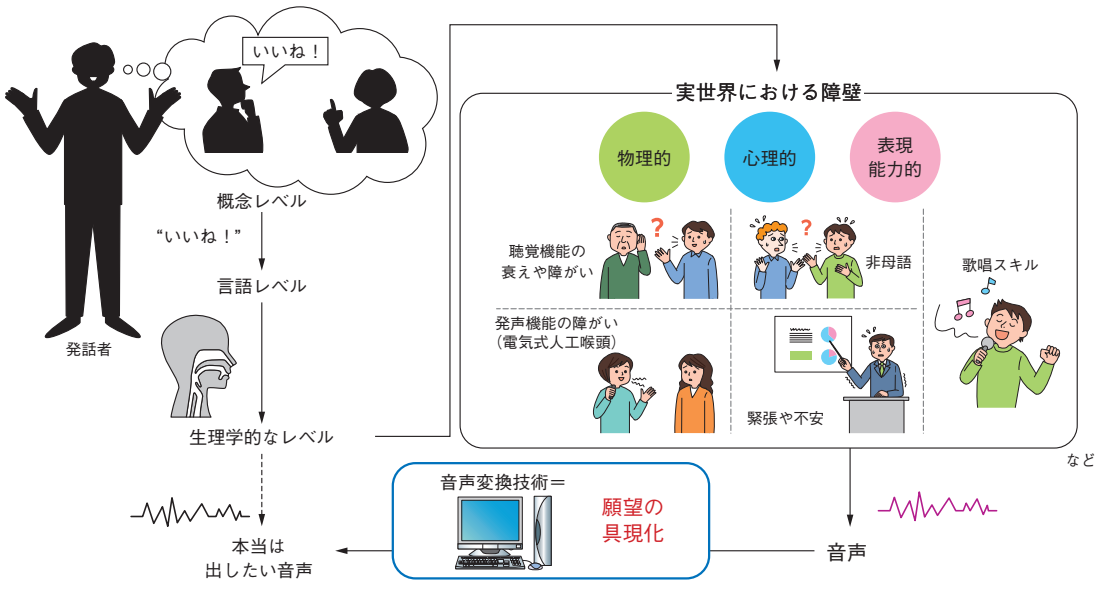


図1　音声 × 深層生成モデルの取り組み

果として，人の声と聴き分けられない品質の音声変換や任意の目標話者へ変換，非パラレルデータを用いることができるという高効率化，リアルタイム化を実現しています．また，近年盛んに研究されている分野横断型の研究として，目標話者の顔画像を用いて条件付けを行い，音声を変換する「クロスモーダル音声変換機能」も実現しています．

## 安定学習可能・高精度な音声系列変換

ここでは，前述の「音声系列変換機能」[(3)]を掘り下げようと思います．自然言語処理において離散値として扱われる単語（シンボル）と違い，音声は連続値の系列として観測されます．系列変換を用いた音声変換アプローチでは，大幅な高品質化が見込める一方で，連続値の系列を扱う際の学習の難しさが課題となります．この課題を克服するため，従来の音声系列変換では，音声認識とテキスト音声合成を組み合わせるアプローチが主流でした．音声認識を用いて，入力音声から単語のようなシンボルを認識し，離散値の系列であるシンボル系列上で変換を行い，変換されたシンボル系列から所望の音声を合成する，というアプローチです．離散値であるシンボルを用いて変換するので比較的安定した学習が可能な一方で，モデルを学習するために音声だけでなくテキストラベルも必要となります．また，テキスト化という処理を挟むことにより，テキスト化が困難な笑い声などの変換も容易ではありません．

NTTコミュニケーション科学基礎研究所では，テキストラベルを用いずに安定学習可能な音声系列変換を実現しました．**図2**に示されるとおり，後述の文脈保持機構を新たに導入し，音声データのみからすべてのモジュールを学習します．テキストラベルは用いないものの，文脈保持機構が存在するため，音声の文脈構造が崩れる変換を許さないという制約が働き，適切な学習を導くことができるようになります．文脈保持機構において，入力音声の復元をタスクとするモジュール（図2左）は，入力音声を復元できるように，入力音声の情報を変換の過程で保持するよう働きかけます．いわゆる自己符号化と呼ばれる技術となります．一方で，目標音声を予測するよう新たに導入されたモジュール（図2右）は，変換の過程で得られる中間特徴量を，入力音声も目標音声も予測可能な特徴量にするよう働きかけます．いわば，入力音声を“話者共通の音声空間”へ写像させるような制約となります．これにより，文脈情報を保持するため音声認識に近いことを行いますが，音声認識により得られるシンボルよりも，“リッチな”中間特徴量を得ることができ，高精度な変換を手助けします．

## 合成音声波形から実音声波形への波形変換

音声変換のもう1つの課題は，音声の特徴量から波形を合成した際に，波形合成器の精度の影響を受けるという点にあります．世の中にあふれる音声を聞いて，「これは機械により合成された音声だな」と判断できることもあるかと思います．そういった“合成音声っぽさ”をなくし，より高音質で肉声感のある自然な音声波形への波形の直接補正をめざします．

波形を深層学習で扱う際の難しさは2つあります．1つは，例えば，16 kHzサンプリング音声には1秒間に1万6000個の点があります．合成音声と自然音声のペアデータを用

いて対応点を求めようとすると，非常に困難なタスクとなることは想像が容易かと思います．もう１つは，「位相」という扱いが非常に難しい特徴量の存在です．そのため，音声変換では位相情報を捨てて振幅情報のみを扱うのが定石となっています．

NTTコミュニケーション科学基礎研究所では，前述の課題を克服し，深層学習を用いた音声波形の直接補正を実現しました[(4)]．**図3**に示されるとおり，補正法のコアモジュールである循環型敵対学習モデルにおける変換器は２つの指標に沿って学習されます．１つは，敵対学習の学習指標である「識別誤差の最小化」です．合成音声と自然音声との定量化しづらい違いを識別モデルに認識させ，その差をできるだけなくすように合成器を学習します．また，１つの識別器ではなく，複数の識別器を用いてさまざまな角度から違いを認識させることにより，高精度な変換を実現しています．もう１つは，循環モデルの学習指標である「再構成誤差の最小化」です．合成音声を自然音声化し，再度，合成音声化することで，入力である合成音声を復元します．この際に，完全に復元するように誤差を最小化します．ここがポイントで，位相まで含めて波形を再構成するよう制約付けることで，位相を適切に考慮した学習が可能となります．これらにより高音質化を実現していますが，さらに嬉しいことに，循環型敵対的モデルは，パラレルデータを必要としない枠組みであるため，合成音声と自然音声を適当にかき集めることで，所望のモデルを学習することができます．

## 今後の展開

NTTでは，より多様な音声への変換に向けて，技術改良と実証を進めていきます．現在までに，具体的な目標話者が存在する音声への高品質な変換には成功していますが，「もっと可愛い声で喋りたい」や「もっと渋

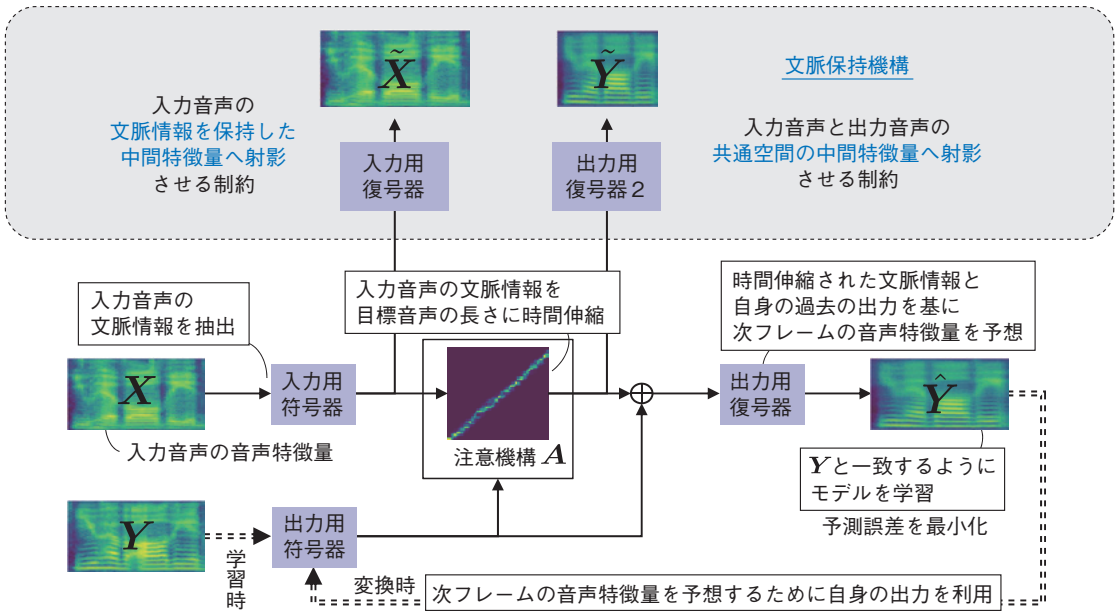


図２　安定学習可能・高精度な音声系列変換

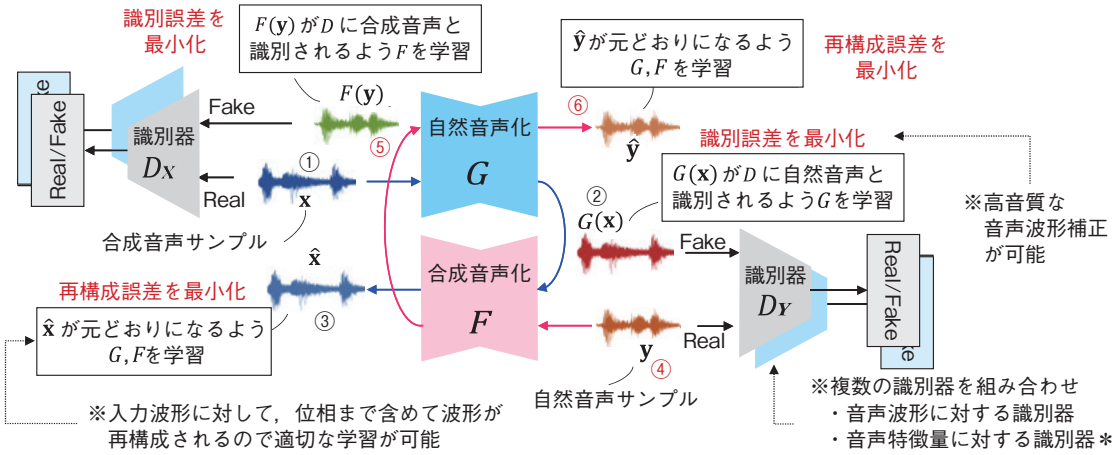


図３　合成音声波形から実音声波形への波形変換

い声で話したい」といった，目標話者ではなく音声のイメージを伴う変換を実現するには，音声の知覚的な空間を掌握し，空間上の潜在変数でうまく補間できるよう，技術自体の性能を高めていく必要があります．ユーザのあらゆる願望にこたえ，さまざまな利用シーンに柔軟に対応可能な多用途音声変換システムの実現に向けて，研究開発を推進していきます．

**■参考文献**


⑴ T. Toda, A. W. Black, and K. Tokuda: “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans TASLP, Vol. 15, No. 8, pp. 2222-2235, 2007.

⑵ D. Bahdanau, K. Cho, and Y. Bengio: “Neural Machine Translation by Jointly Learning to Align and Translate,” in Proc. ICLR 2015, San Diego, U.S.A., May 2015.

⑶ K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo: “AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms,” in Proc. IEEE ICASSP 2019, Brighton, U.K., May 2019.

⑷ K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka: “Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks,” in Proc. IEEE SLT 2018, Athens, Greece, Dec. 2018.



**（上段左から）田中　宏／ 金子　卓弘**

**（下段左から）北条　伸克／ 亀岡　弘和**


深層学習の目覚ましい発展により，メディア生成・認識においてさまざまなことが高精度に実現可能となってきました．私たちは，あらゆる人が所望の音声で思いのままに表現でき，不自由なく快適にコミュニケーションを行える環境を実現することをめざしています．


**◆問い合わせ先**

NTTコミュニケーション科学基礎研究所
メディア情報研究部
TEL　0774-93-5020
FAX　0774-93-5026
E-mail　cs-liaison-ml＠hco.ntt.co.jp