ご所望の声でコミュニケーション

～深層生成モデルが切り拓く音声変換の可能性～

Communication with desired voice

— Deep generative model opens the way to innovative speech transformation —

メディア情報研究部

田中　宏 Kou Tanaka

プロフィール

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 研究員。2017年奈良先端科学技術大学院大学 情報科学研究科 博士課程修了。博士（工学）。同年NTTに入社以来、音声合成・音声変換の研究に従事。深層生成モデルを用いた音声信号処理に特に興味を持つ。奈良先端科学技術大学院大学優秀学生賞や日本音響学会第47回粟屋潔学術奨励賞を受賞。日本音響学会の会員。

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ており、人々がお互いにコミュニケーションを取るうえで特に重要なツールの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、相手に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。その適用先は、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、様々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。

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。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）を用意し、この両者の特徴量の同時確率を最大化することで、前者から後者への変換関数を得ます。また、近年では、上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ること、2）変換可能な音声特徴が声質に限定されること、と技術的制約があります。さらに、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、不自然でいかにも合成音といった音が生成されてしまいます。

一方、深層学習界隈において、2014年を皮切りに、画像処理や自然言語処理の分野から、非常に興味深いモデルが台頭してきました。変分自己符号化器 （Variational Auto-Encoder: VAE）や敵対的生成モデル（Generative Adversarial Networks: GANs）、系列変換モデル（Sequence-to-Sequence model: Seq2Seq）といったモデルです。Seq2Seqのモジュールの一つである自己再帰型モデル（Auto-Regressive model: AR）とVAE、GANsを総称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など様々な分野・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）がニューラルネットに導入され、その高い有効性が瞬く間に注目を浴びました。

音声変換の入力・出力はどちらも音声信号（連続値で構成された時系列データ）であることを意識しつつこれらの深層学習技術をうまく拡張することで、従来音声変換技術の課題を克服し様々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します。私たちの音声変換システムでは、1）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［2］」、2）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［3］」などの新機能を創出しました。また、合成音声波形から実音声波形へ波形空間上で深層学習を用いて変換を行い、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」を世界で初めて実現しました。結果として、人の声と聴き分けられないほど高品質の音声変換や任意の目標話者へ変換を行うためのモデルの多対多化、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その顔らしい音声に変換する「クロスモーダル音声変換機能」も実現しています。

※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）を示す。

▼関連文献

[1] T. Toda, A. W. Black, K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” *IEEE Trans TASLP*, Vol. 15, No. 8, pp. 2222-2235, 2007.
[2] T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo, “CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion,” in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2019.
[3] K. Tanaka, H. Kameoka, T. Kaneko, N. Hojo, “ATTS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms,” in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2019.