挑戦する研究者たち
CHALLENGERS

亀岡弘和

NTTコミュニケーション科学基礎研究所
上席特別研究員

科学技術は先達が少しずつ積み上げてきた成果．それをさらに良くするのが，今を生きる私たち研究者の使命である

「アニメのキャラクターの声が想像と違った」「言い淀みがひどいのでスピーチに自信が持てない」「病気や怪我などで失ってしまった自分の声を取り戻したい」等，発話にまつわるさまざまな思いや不自由さがあります．コミュニケーションにおけるさまざまな制約をAI（人工知能）の機械学習や信号処理の力により取り除き，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の実現をめざして本分野の最前線で活躍するNTTコミュニケーション科学基礎研究所 亀岡弘和上席特別研究員に研究の進捗と研究活動の醍醐味を伺いました．

コミュニケーション機能を拡張するメディア情景分析・生成技術を追究

2度目のご登場ですね．手掛けている研究について教えていただけますでしょうか．

私たちの日ごろのコミュニケーションにおいては，障がいや加齢などによる物理的な制約，外国語の会話などにおける能力的な制約，緊張状態などの心理的な制約などが伴い，思いどおりに話せないことがあると思います．私が取り組んでいるのは，そうしたコミュニケーションにおけるさまざまなかたちの障壁や制約を克服するための信号処理・機械学習技術の開発です．

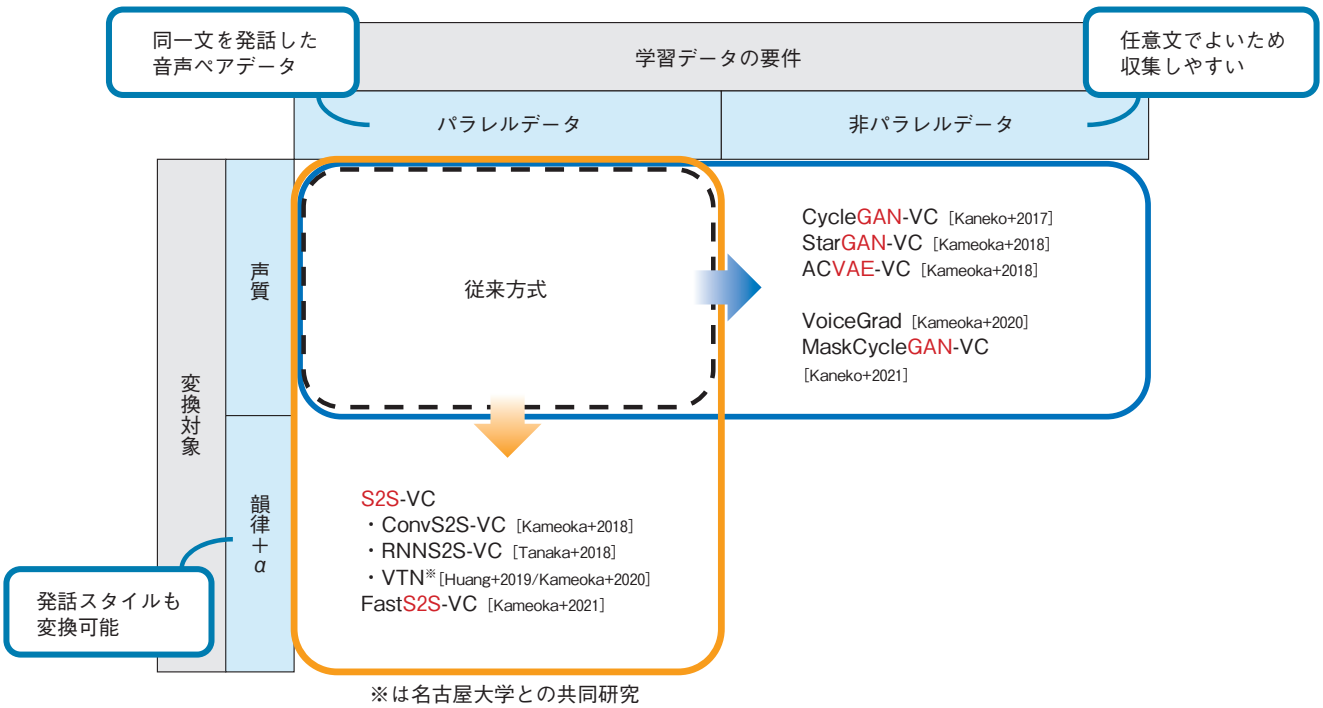
コミュニケーションには発信者と受信者が存在しますが，それぞれが望む表現でメッセージを受け渡しできるようにするため，発信者から送信される信号を状況に適した表現にリアルタイム変換するシステムの構築をめざしています．このようなシステムを実現するうえで今のところ核となると考えられるのが音源分離技術と音声変換技術で，前者が受信者側の聴覚機能を補完する技術に相当し，後者が発信者側の発声機能を補完する技術に相当します．音源分離は前回お話しした「外界音を対象とした要素分解」に該当し，混合音に含まれる複数の音を分離抽出し，残響や雑音を取

り除いたりすることで対象音を強調することが目的になります．音声変換についても前回少しだけ触れましたが，発話内容を保持したまま音声の特徴を所望のものに変えることが目的になります．

さらに，音だけでなく動画像やテキストなどの多種のメディアを有効活用した新たなコミュニケーション方式の可能性を模索しています．例えば，顔に合った音声を生成したり，声に合った顔画像を生成したりすることで，コミュニケーションに広がりを与えられないかということを考えています．

前回，さらに追究したいとお話しされていた高い品質と自然性を意識した音声生成についてはいかがですか．

前回も少しだけ触れましたが，これまで音声変換にかかわる基礎技術およびその周辺技術を多く開発してきました．私たちが音声変換の研究に着手したのは2016年ごろだったのですが，当時主流となっていた従来方式では，同一の文章を発話した音声ペアを用い，同じ音素を発している時刻が合うように一方の音声を時間伸縮したうえで，音声変換器，つまり元音声の特徴を目標音声の特徴に変換する変換則を学習するアプローチがとられていました．このような同一の文章を発話した音声ペアのデータをパラレルデータと言います．もちろん，パラレルデータを多く集められる条件ではこのアプローチは有効なのですが，例えば目標音声が特定の有名人の声の場合など，パラレルデータを容易に取得できない場面も多々あります．そこで，当時機械学習やコンピュータビジョンなどの分野ですでに脚光を浴びていた変分自己符号化器（VAE：Variational Autoencoder）や敵対的生成ネットワーク（GAN：Generative Adversarial Network）といった深層生成モデルに目を付け，任意の文章を発話した元音声および目標音声のサンプルからでも音声変換器を学習することが可能な非パラレル音声変換手法を考案しました（**図1**）．これらの手法は学習にパラレルデータを必要としないので，音声変換の利用場面を大きく



※は名古屋大学との共同研究

図1　深層生成モデルによる柔軟な音声変換の実現

広げることが期待されます．

また，当時のほとんどの従来方式では，変換対象の音声特徴が声質に限られており，抑揚やリズムなどの発話スタイルを変換するまでには至りませんでした．そこで私たちは声質だけでなく発話スタイルも変換できるような方式を創り出したいと考え，当時すでに機械翻訳や音声認識，テキスト音声合成などで多大な効果が示されていた系列変換（S2S：Sequence-to-Sequence）学習と呼ぶ枠組に着目しました．S2S学習は，長期依存関係をとらえながらあるベクトル系列から（異なる長さの）別のベクトル系列に変換するニューラルネットワークモデルを学習する枠組です．ポイントは注意（Attention）機構と呼ぶモデル構造にあり，これにより変換元と変換先の音声特徴量系列の要素間の対応付け規則とともに要素間の変換則を学習することが可能になります．当時S2S学習アプローチを音声変換に適用する試みは私たちが知る限りほとんどなされていなかったのですが，これをいち早く試したところ，期待していたとおり声質以外にも抑揚や発話リズムも柔軟に変換できるようになることが実験を通じて分かり，同僚たちと興奮したのを覚えています．

現在の音声変換の最先端手法は，ほぼ例外なく，元音声からメルスペクトルと呼ぶ音声特徴量ベクトルの系列を抽出して変換するステップと，変換したメルスペクトルの系列から音声波形を生成するステップからなります．前述のVAE，GAN，S2S学習に基づく手法はいずれも特徴量変換を行う前段のステップに相当する技術でしたが，後段のステップを波形生成と言い，ニューラルネットワークで表現した場合の波形生成器をニューラルボコーダと言います．音声研究に詳しい方であればご存じかもしれませんが，WaveNetと呼ぶ高品質波形生成法が2016年にDeepMindから発表され，以降多くの研究者により盛んに高速化，高品質化，学習効率向上の研究が行われています．これまで私たちは特徴量変換技術の研究をメインに進めてきましたが，最近は波形生成の高品質化と低遅延化のための研究にも力を入れ始めているところです．

以上のそれぞれの成果はICASSPやInterspeechなどの国際会議やIEEE Transactions on Audio, Speech, and Language Processingなどの学術論文誌に採録され，これまでのところ合計で1000回以上引用されています．私たちの最近のアクティビティも徐々に認知されてきていると感じます．

音声変換・音源分離技術の高精度化・効率化・柔軟化のための機械学習基盤の構築

コミュニケーションに悩む人には朗報ですね．具体的な応用先を聞かせていただけますか．

まず音声変換技術に関してですが，これまで実験的に良い感触が得られた応用先としては，話者変換以外ですと，英語の訛りの変換，ささやき声の変換，電気喉頭音声の変換，感情表現変換，言い淀みの変換などがあります．英語の訛りの変換は，話し手の英語を聴き手にとって聴き取りやすい訛りに変換することで，会話を円滑化するのに役立つと考えています．例えば，日本人にとっては（もちろん人によってですが）いわゆる日本語訛りの発音の方がネイティブな発音よりも聴き取りやすい場合もあるかと思いますので，あえて訛りを付与するといったような使い方も考えられます．ささやき声の変換は，ささやき声を自然発声風の音声に変換することを目的としたタスクです．例えば１人で電車内や喫茶店にいて声を発するのがはばかられるような場面で電話やオンラインミーティングをしたい場合があると思いますが，これが実現できると，周囲に声を聞かれないようにささやき声で話しても，相手側には普通の声として届けられるようになります．電気喉頭音声の変換は，電気喉頭音声を健常な音声に変換することを目的としたタスクです．電気喉頭音声は，喉頭摘出手術などで声帯を失ってしまった発声障がい者が電気式人工咽頭を用いて発した音声で，抑揚に乏しく機械的な音声になってしまいがちですが，音声変換技術によりそういった音声を健常者のような音声に変換することが可能になります．ほかにも発話スタイルを変化させることで感情表現も変換することも，「あー」や「えーと」などのような言い淀みやフィラー

を自動的に省略して発話全体を流暢にすることもある程度可能であることが分かってきました．これらの音声サンプルはデモサイトで聴くことができますので，ご興味がある方はぜひアクセスしてみてください[(1)〜(3)]．また，音声分離については，前述のVAEを音源信号のモデル化に用いた多チャネル音源分離法を以前提案し，その高速化と高精度化に向けた検討を行ってきました．これまで多チャネル音源分離の研究分野ではせいぜい5音源程度の混合信号しか扱われていませんでしたが，私たちが提案した手法により18音源もの混合信号を高精度に分離できることを実証し，世界でも例をみないレベルの性能を達成することができたと考えています．こちらの音声サンプルもデモサイトで聴くことができますので，ぜひアクセスしてみてください[(4)]．

従来の手法と提案された手法を比較しながら聴いてみると，この技術の素晴らしさが良く分かりますね．

ありがとうございます．さらにこれらの検討に加えて，音以外のメディアを活用しながら音声を生成・制御したり音声を活用して音以外の信号を生成・制御したりするクロスモーダル信号生成の研究にも取り組んでいます．例えば顔に合った音声の生成や声に合った顔画像生成などです（**図2**）．この取り組みでは，音声コミュニケーションをより豊かにすることだけなく，コミュニケーション機能拡張において直感的な制御を可能にすることをめざしています．

具体的な取り組みの一例として，音声のみから話者の顔を予測し，予測した顔を画像として出力するクロスモーダル顔画像生成や，声質変換において目標声質を（話者IDの代わりに）顔画像により指定することができるクロスモーダル声質変換などの検討を行いました．これらのデモンストレーションを，NTTコミュニケーション科学基礎研究所（CS研）オープンハウスで実演したところ好評を博し，各種メディアで取り上げていただきました．

また，音声のみから話者のアクションユニット（顔面筋パラメータ）系列を推定する新たな試みも行いました．私たちの知る限りこのような試みはほかになかったため，どの程度の精度を達成できるかは全くの未知数でしたが，実験を通じてこれがある程度可能であることを明らかにしました．また，音声から推定したアクションユニット系列を用いて顔画像変換を行うことで，声に合わせて静止顔画像の表情を動かすことができるようになります（**図3**）．今後これをさらに高精度化しうまく活用すれば，自分の話し方や声質が会話相手にどのような印象を与えているかを視覚的にフィードバックできるようになり，プレゼンテーション能力やコールセンターなどの接客能力の向上を支援することに役立つかもしれません．

これらの取り組みについてもそれぞれデモサイトを用意していますので，ご興味のある方はアクセスしてみてください[(5)，(6)]．

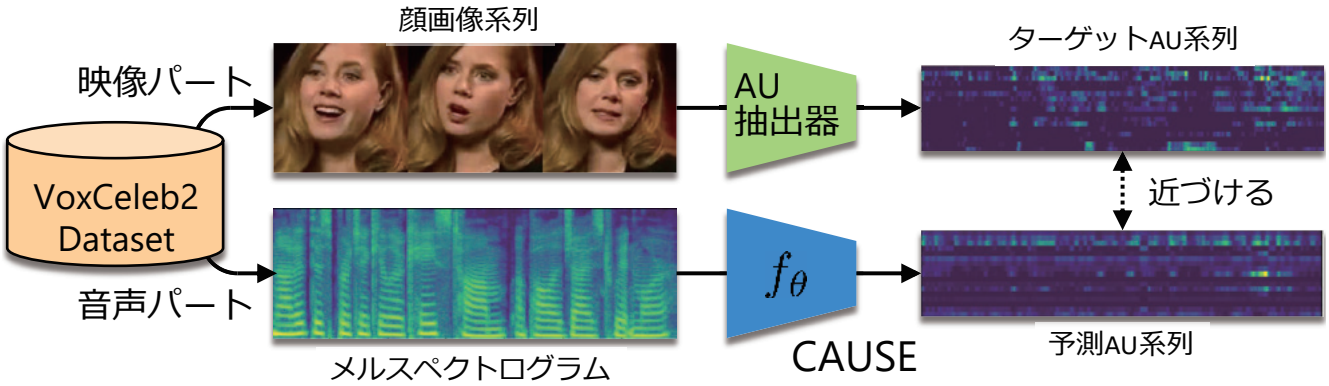
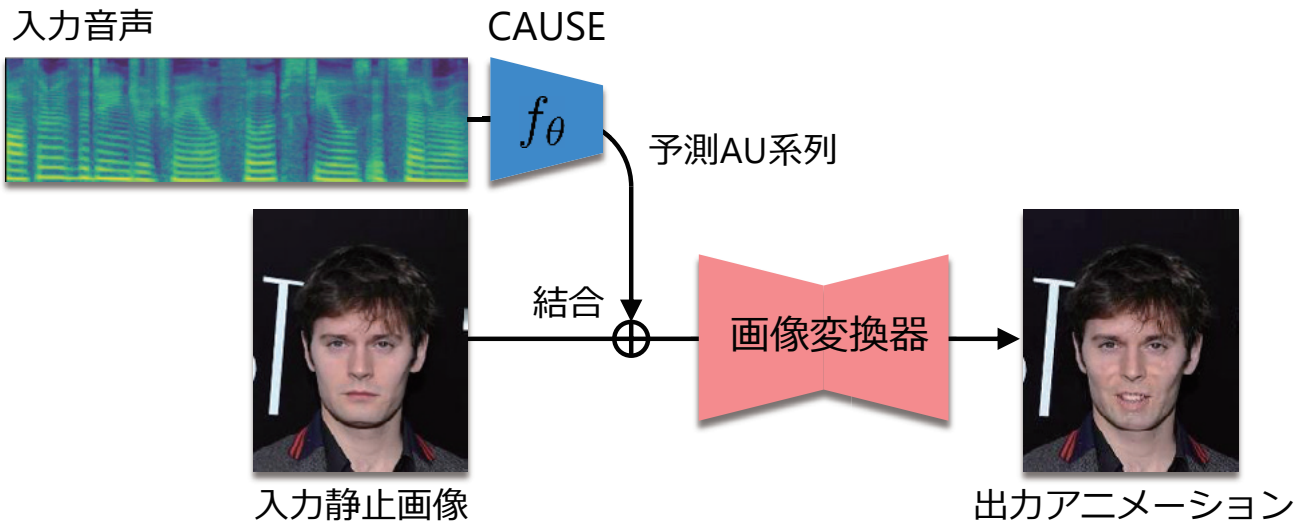


図2　クロスモーダルアクションユニット系列推定器（CAUSE：Crossmodal Action Unit Sequence Estimator）の学習



※　顔画像はVoxCeleb2 Dataset [Chung+2018*]，CelebA Dataset [Liu+2015*]のものを使用

Z. Liu, P. Luo, X. Wang, and X. Tang: "Deep Learning Face Attributes in the Wild," in Proc. ICCV, pp. 3730-3738, 2015.
J. S. Chung, A. Nagrani, and A. Zisserman: "VoxCeleb2: Deep Speaker Recognition," in Proc. Interspeech, pp. 1086-1090, 2018.

図３　CAUSEと顔画像変換器を用いた音声からの顔表情制御

素人発想・玄人実行に努める

研究者として大切にされてきたことを教えてください．

金出武雄先生の著書のタイトルにある「素人発想，玄人実行」は，普段から私が研究者として心掛けていることの１つです．専門的な知識が増えてくるとつい「研究のための研究」に陥り，重箱の隅をつついたような研究テーマを設定してしまいがちになります．それはそれで重要な研究テーマに発展し得る場合もあると思うのですが，１つひとつの研究テーマに対して，素直に面白いと思えるかどうか，本当に世の中の役に立つのかどうか，をできるだけ冷静に自問自答するようにしています．例えばコミュニケーション機能拡張の研究においては，日常生活の中で，普段あまり意識しないような違和感や不自由さはないか，それらを解決するための打開策はないか，ということを常日頃考えています．今，人工知能（AI）や機械学習の分野は発展がものすごく速い激動の時代に入っていますので，常に最新のトレンドや研究動向を追うことはもちろん大切ですが，一度冷静になり，自分の中にある「内なる声」に耳を傾けることも大切だと思っています．

そして最近，AIの研究をやっていて改めて実感するのは，当たり前のことですが，できるだけ多く手を動かすこと，つまりコーディングと実験をできるだけ多く行うことの大切さです．もともと私は１つひとつの問題に対し，じっくりと仮説と理論を立て，定式化したうえでその解法を編み出す研究スタイルが得意なほうでしたが，深層学習やニューラルネットワークを用いた研究では，それとは対照的に，実験による仮説検証を速いスピードでとにかく何回も何回も繰り返すことが重要であると感じています．ニューラルネットワークの挙動は必ずしも直感どおりではなく，生物を相手にしているような感覚がすることがあり，触れれば触れるほど感覚が身に付いてくると感じるので，今は必ず１日１回はコーディングと実験を欠かさず行うようにしています．深層学習ではニューラルネットワークに学習サン

プルをたくさん入力して，学習データに合った振る舞いを学ばせていくわけですが，その訓練プログラムをコーディングしている自分自身が，たくさんのコーディングと実験をとおしてニューラルネットワークの振る舞いを学んでいっている感覚がして，とても新鮮で面白いです．

今後はどのようなことに取り組まれますか．また，後進の研究者にも一言お願いいたします．

まず，感性語による要望にこたえるような音声変換です．例えば「可愛い声」「優しい声」「堂々とした声」にしたい，といったような要望があったときに，それにこたえるような音声変換です．これまで扱ってきた音声変換では変換目標の音声特徴は一意に定義しやすいものでしたが，これらの例からも分かるように感性語の定義は曖昧で人によって異なります．このように定義が曖昧で主観的であるような感性語をいかにして定量化できるかが鍵になりますが，現在その課題に同僚と一緒に取り組み始めています．

また，この音声変換システムの実用化を想定したとき，他人の声になりすます等により悪用される可能性は否めません．今後は，音声変換システムの悪用防止のための研究も視野に入れています．

それから，実用化に向けてはホワイトボックス化したモデルをつくる必要性も感じています．音声変換の例でいえば，リアルタイムに音声を変換するシステムを実際に使用する場合，想定外な変換が行われないように保証する必要があります．変換のされ方によっては話し手の意図に反した印象を相手に与えかねない可能性があるからです．ニューラルネットワークは，学習データに合った振る舞いを学習する能力は非常に長けている一方で，内部がブラックボックス的であるがゆえに，学習データにないデータが入力されたときの振る舞いをなかなか予見することができず，制御するのが必ずしも簡単ではありません．したがって，音声変換モデルを安心して利用できるようにするためのモデル構造や制御メカニズムの研究も今後必要になると考えています．

最後に，後進の研究者の皆さんに向けてですが，研究者の使命は「世の中を良くする」ことだと思っています．皆で協力し，常に便利さや快適さを追究する人間の隠れた欲求にこたえるために知恵を絞り，人が安心・安全に，幸せに生きていける世の中にしていってもらいたいと思います．

研究をしていると辛いことも多いと思います．月並みな言葉かもしれませんが，ネガティブな面ばかりに目を向けず，研究を楽しむことが大事だと思います．今NTT研究所ではリモート勤務の方が多くなっていると思いますが，そういう方々は特に，雑談目的でもいいのでオンラインミーティングを頻繁に開いて同僚や先輩とコミュニケーションを図る場をたくさん設けるよう意識してみてください．話をしているうちに楽しくなってきて，刺激も得られると思います．そして，研究者どうし，相互のリスペクトも忘れずにいたいですね．私は学生と一緒に研究することも多く，論文原稿をチェックする機会がありますが，時折，提案技術の優位性を主張したいがために従来技術を必要以上におとしめるような記述を目にします．しかし，科学技術は先達が英知を結集して少しずつ積み上げてきたものであって，それをさらに良くしようとするのが研究者の仕事です．だからこそ，良いところを見つけ，さらに良くしようという視点で先行研究を眺め，研究に臨んでいただきたいです．

■参考文献

⑴ https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/s2s-vc/index.html
⑵ https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/acvae-vc3/index.html
⑶ https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/stargan-vc2/index.html
⑷ https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/mvae-ss/index.html
⑸ https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/crossmodal-vc/index.html
⑹ https://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/cause/index.html