AI技術を使った実用的な声質変換システムの開発

中鹿 研究室

中鹿 亘
Toru NAKASHIKA

突然ですが、アニメ「名探偵コナン」にたびたび登場する必須の探偵アイテム『蝶ネクタイ型変声機』をご存じでしょうか。ネクタイの裏にあるダイヤルを調節し、番号を選ぶだけでさまざまな人の声が出せる道具です。ある人の発話を、話す内容は変えずに、あたかも別の人が話しているかのように音声だけをすり替える――。そんなSFの世界のような技術を研究しているのが中鹿亘助教です。

声質変換システムとは

中鹿助教は、今日のブームが到来する前から研究していた最先端の人工知能（AI）技術を使って、実用的な声質変換システムを開発しました。このシステムは、例えば入力として男性のAさんが『こんにちは』と言った場合に、声質だけ女性のBさんに似せる変換処理を行い、女性の声で『こんにちは』と出力させる機械です。100%Bさんに似せるだけでなく、Aさんを30%、Bさんを70%と声質を混合させることも可能です。

これまで音声信号処理の分野において、多くの声質変換の研究が行われてきました。ただ、従来のいずれの方法も、入力側の話者と出力側の話者とを関連づけた上で、音質を変換させる複雑なモデルを機械に学習させる必要がありました。また、入力側と出力側とが同一の内容、かつ同一のタイミングで発話した音声（対データ）をあらかじめ用意する必要もありました。

これに対して中鹿助教が開発したシステムは、対データが必要なく、どのようなデータでも自由に学習させられます。そのため相手を選ばずに、任意の話者から任意の話者へと変換することができるのです。

防犯対策などに応用

利用シーンを想定してみましょう。例えば、女性の一人暮らしの防犯対策として、インターホンや電話などの応答時に、自分の声を男性の声に変えることができます。また、アニメの声優が入れ替わっても、その主人公の声は時代を超えて永遠に受け継がれるでしょう。将来は、自分と同じ声で話すロボットが登場するかもしれません。

わずかな音声データさえあれば変換処理は可能です。声質はその場で自在に加工できるため、「実用的なシステムとして普及させたい」と中鹿助教は考えています。最近では、複数の人の音声データを入力し、自動的に声質のグループ分けを行う技術も開発しました。これにより、任意のグループの声に近づけてほしいといった、「ユーザー好みの声」にさらにチューニングしやすくなったそうです。

提案する声質変換の仕組み

キーワード

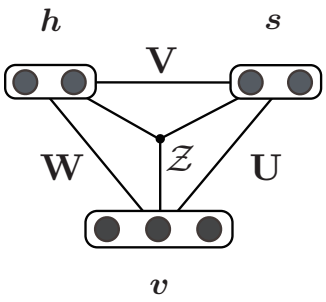
パターン認識、メディア情報処理、機械学習、ディープラーニング、統計的信号処理、音声認識、画像認識、声質変換、音楽ジャンル推定、自動採譜

所　属	大学院情報理工学研究科 情報・ネットワーク工学専攻
メンバー	中鹿 亘 助教
所属学会	日本音響学会、米電気電子学会（IEEE）、電子情報通信学会、人工知能学会
E-mail	nakashika@uec.ac.jp

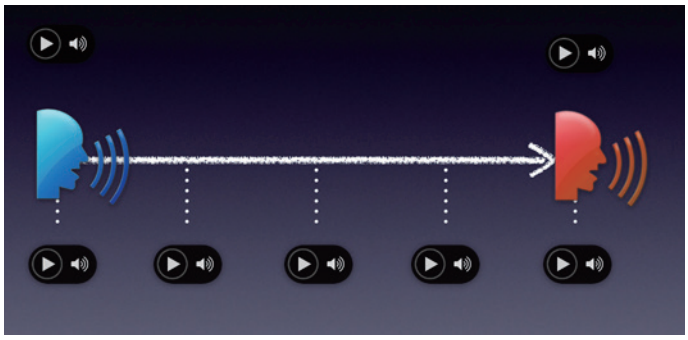
話者性だけ変換する

原理は複雑なため、ここでは簡単な概念の説明にとどめます。前述の男性Aさんが『こんにちは』と言った場合、これは、「こんにちは」というテキスト（潜在的音韻情報）と、〝男性Aらしさ〟（話者性）の二つで構成されている（男性A『こんにちは』＝「こんにちは」＋男性Aらしさ）と考えることができます。この考え方を応用すると、女性Bさんの声で『こんにちは』と出力するためには、潜在的音韻情報「こんにちは」は変えずに、話者性だけを女性Bに変えればよいことが分かります。

中鹿助教は、AI技術の一つであるディープラーニング（深層学習）で用いられている従来の確率モデル「制限ボルツマンマシン」を音声向けに改良した「適応型制限ボルツマンマシン」を開発し、音声を音韻情報と話者性とに分離することに成功しました。この独自モデルを使うことで、入力した音声に対して音韻情報は保存したまま、話者性だけを切り替えることができるのです。これが声質を変換する仕組みです。実験の結果、対データを使う従来の複雑な声質変換システムに匹敵する精度を持ちながら、利便性は大幅に向上することができました。



声質変換モデル
v：音声、s：話者性、h：潜在的音韻情報


試作した声質変換システム

AI技術に詳しい中鹿助教は、ほかにも従来の変換された実数データではなく、元の複素数データのまま、音声や画像情報を扱える、新しいディープラーニングの手法なども開発しています。

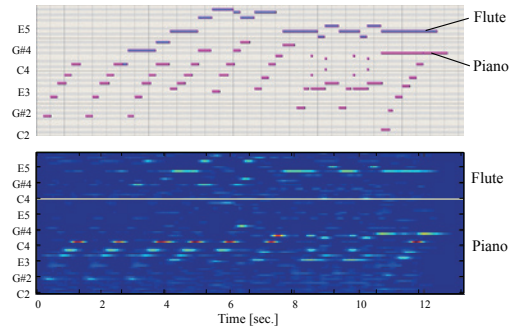
音楽の信号処理や画像処理も

音声のほかに、音楽の信号処理も手がけています。音楽を聴いてそれを楽譜に起こす、いわゆる「耳コピー」と呼ばれる手法がありますが、中鹿助教はこれをコンピュータに行わせる「自動採譜」技術を開発しました。例えば、ピアノやバイオリン、ビオラなど複数の楽器を聴き分けて推定し、それぞれ楽譜（ピアノロール）として表示できるそうです。

画像処理の研究もしています。手書き文字を認識する新しいモデルを開発したり、ぼやけた画像でも鮮明に拡大できる超解像技術を開発したりしています。手書き文字認識のモデルは、例えば、人が書いた数字の「3」の文字について、「3」を正確に言い当てるだけでなく、この逆の操作として、モデルが認識している「3」の画像を作り出すことができます。


(a) DRM　(b) Mean

提案モデルで生成した数字画像（左）と単純なアプローチによって得られた数字画像（右）



正解のピアノロール（上）と提案法による楽器カテゴリ・音高推定結果（下）

これは異なる2変数間の結びつきをディープラーニングの枠組みによってモデル化したもので、文字認識を超えた新しい用途が開拓できそうです。中鹿助教によれば、「例えば、口パクの動画からこれに合致する音声を自動で作り出したり、逆に、ある音声から口パクの動画を作ったりすることが可能かもしれない」そうです。

人間になじむシステムをつくる

我々はモノを見て音を聴き、それらを脳で処理しています。こうして周囲にあるものを認識したり、状況を理解したりして判断しています。中鹿助教は、このような複合的な「マルチメディア信号処理」の研究を通じて、人間の情報処理のメカニズムを解明し、人間の知能を模倣させることによって「人間になじむシステムを作りたい」と考えています。その先には、きっとさまざまな声が自在に出せる『蝶ネクタイ型変声機』のような、夢のあるシステムの実現が待っているかもしれません。

【取材・文＝藤木信穂】