日本音響学会誌 74 巻 3 号（2018），pp. 144–147

小特集―障害者を支える音声技術―

構音障害者のための話者性を維持した音声変換*

滝口哲也（神戸大学/JST さきがけ）**

43.70.Dn

1. はじめに

構音障害とは，「発話器官の形態の異常，運動機能の異常，あるいは麻痺等が原因となって，正しい構音ができずに音の誤りが生じる状態」のことをいう [1]。構音障害の症状は，その原因により様々であり，聞き取りがやや難しい発話から，聞き取りが困難な発話が存在する。聞き取り困難な発話においても，その原因（咽頭摘出者なのか，脳性麻痺者なのか等）により症状が異なる。そのような方々の発話コミュニケーションを支援するための音声研究への期待は大きい。

例えば，発話コミュニケーションを支援する音声研究として，声質変換手法を咽頭摘出者に応用し，電気式人工咽頭を用いた発話を自然性の高い音声へ変換する研究が行われている [2, 3]。また，ALS 患者のための音声合成システム [4] など構音障害者のための研究が行われている。

本稿では構音障害として脳性麻痺者と重度難聴者の発話特性を紹介し，続いて彼ら/彼女らの聞き取りが難しい発話を，聞き取りが容易な声へ変換する声質変換，音声合成の研究について紹介する。

2. 構音障害者の発話特徴

2.1 発話スタイルの不安定性

脳性麻痺者は，手足に麻痺がある場合はタブレットパソコンなどの端末操作を行うのも難しい。特に筋肉の緊張から意図しない動作が生じるため（不随意運動），同じ言葉を繰り返し発声しても不安定な発話となる。図–1 に脳性麻痺者が各単語を繰り返し発話した時の音声認識率を示す [5]。使用した音響モデルは特定話者モデルで，216 単語認識を行っている。緊張のため特に第一発話が不安定になり音声認識率が低くなっている。このような不随意運動による発話スタイルの不安定性は脳性麻痺発話の特徴の一つである。[5] では，この不安定要素を周波数領域における乗算性ひずみとして定式化した音声強調手法が述べられている。

2.2 スペクトルの違い

図–2 に脳性麻痺者と健常者の/i k i o i/という発話のスペクトログラムを示す。全体的に脳性麻痺者の音声は高域周波数成分のエネルギーが弱くなっている。特に脳性麻痺者の発話例では，子音/k/の成分が健常者と比較しても弱くなっているのが分かる。

次に図–3 に重度難聴者と健常者の/r i cl sh u N/という発話のスペクトログラムを示す。健常者と比較して重度難聴者は高域のエネルギーが弱くなっている。また，重度難聴者と健常者の発話時間は概ね等しいが，音素継続長が間延びする音素（例えば，促音/cl/，/sh/）や欠落する音素（例えば，/r/）がある。これらが音声の明瞭度を劣化させる一因である [6]。

3. 話者性を維持した音声変換

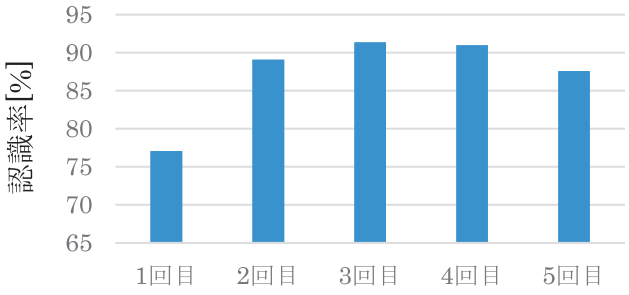
現在，声質変換，音声合成に関する研究が活発に行われているが，例えば声質変換手法を構音障害者に用いた場合，音声は “健常者の声” に変換されるが，構音障害者の話者性は完全に別の健常者の話者性へ置き換えられてしまう。もちろん，そのような応用も選択肢の一つとして重要ではある。一方，構音障害者の日常生活，自立生活の支援に注目した場合，構音障害者のなかには，「自分らしい声で話したい」というニーズもあり，障害者の話者性を維持した声質変換，音声変換が求められている。

3.1 話者性を維持した声質変換

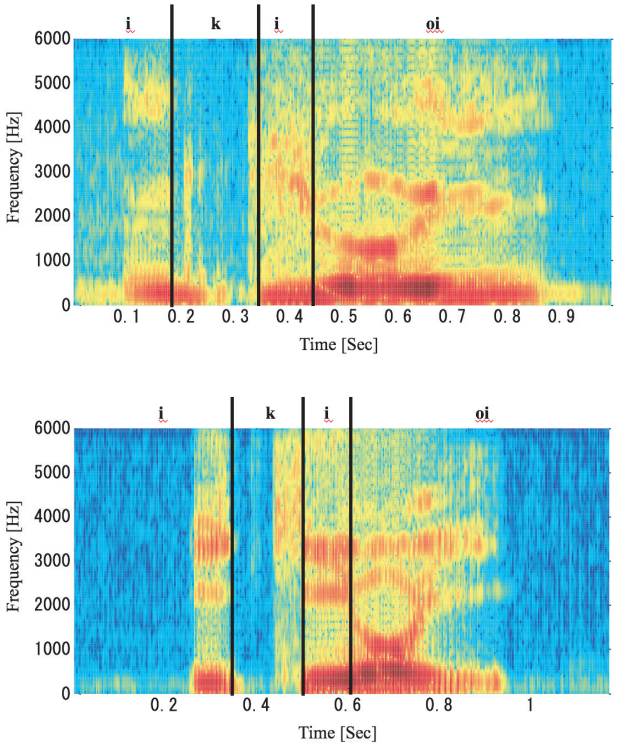
声質変換は，音声に含まれる音韻情報を維持し

* Individuality-preserving speech conversion for articulation disorders.

** Tetsuya Takiguchi (Research Center for Urban Safety and Security, Kobe University, Kobe, 657–8501) e-mail: takigu@kobe-u.ac.jp



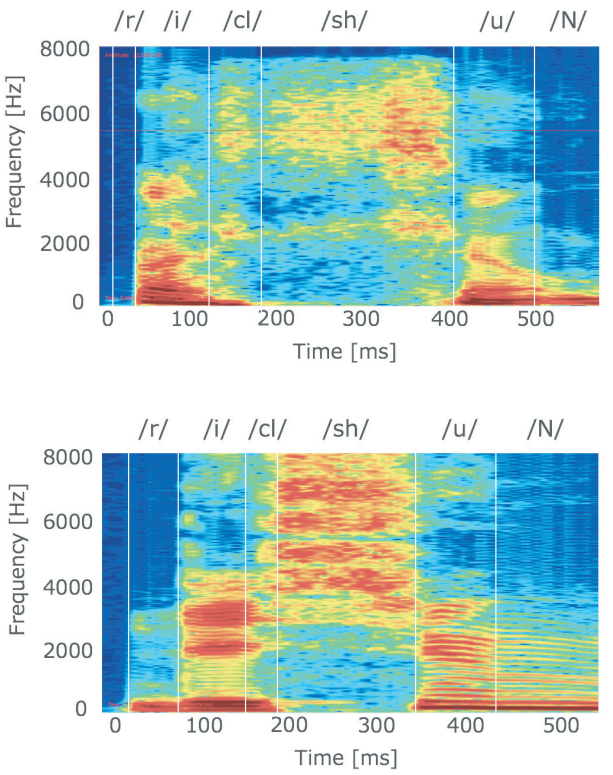
図–1　脳性麻痺者の繰り返し発話の音声認識率



図–2　脳性麻痺者（上）と健常者（下）の発話スペクトル

つつ話者情報を変換する手法であり，主として話者変換を目的としているため，入力話者の声質は別の話者の声質に変換されてしまう。[7] では非負値行列因子分解（Non-negative Matrix Factorization: NMF）による Exemplar-based な声質変換を用いて，入力辞書に構音障害者発話，出力辞書に構音障害者の母音と健常者の子音を組み合わせた Combined 辞書を用いることで，入力障害者音声の話者性を維持しつつ，より聞き取り易い声へ変換する方法を提案している。

しかしながら，[7] では同一フレーム内において，障害者の母音と健常者の子音の線形結合が生じてしまい，変換精度を劣化させてしまう。そのような問題を解決するため，[8] では，話者辞書から音素カテゴリに分けた副辞書を作成し，NMF による音素カテゴリ認識を行い，音素カテゴリに分割



図–3　重度難聴者（上）と健常者（下）の発話スペクトル

した副辞書を用いて声質変換を行っている。

構音障害者の発話は，2 章で述べたように，ある音素の欠落や，また母音，子音などの特性に合わせた変換（補正）の必要がある。しかし声質変換では，発話内容（音素系列）が正確には分からないため，構音障害者発話に対して細かい補正を行うことが難しい。

また，f_0（基本周波数）については通常の声質変換では，入力話者の f_0 パターンの平均値を線形変換により出力話者の平均 f_0 にシフトさせている。つまり抑揚パターンは，入力話者のものを使うことになる。しかし，例えば重度難聴者では健常者と比較して抑揚が少ないという特徴もあり，入力話者（重度難聴者）の f_0 パターンの概形自体を修正する必要がある。

また，音素継続長については，健常者と比較して長くなる傾向があるが，声質変換では正確な音素継続長が分からないため，細かい補正を行うのは難しい。

一方，テキスト音声合成では入力テキストが与えられるため，f_0，音素継続長の補正にも対応しやすい。次の節では，話者性を維持しながら聞き取り易い声を生成するテキスト音声合成について紹介する。

3.2 話者性を維持したテキスト音声合成

Hidden Markov Model（HMM）を用いた音声合成の枠組みにて，ALS 患者のための話者性を維持した音声構築が試みられている [4]。また，脳性麻痺者に関しても，健常者と脳性麻痺者の両方の音声を学習データとして用いて，二つの HMM 音声合成システムを作成し，各モデルパラメータを統合し，構音障害者の話者性を維持しながら聞き取り易い声を構築している [9]。この手法では，スペクトルに関しては，ある子音グループの高域周波数成分にて健常者の合成音を利用している。

脳性麻痺構音障害者の f_0 系列はしばしば不安定なものであるので，[9] では健常者の f_0 系列を基本として f_0 モデルを学習する。f_0 系列に構音障害者の話者性を付与するため，健常者の f_0 系列を以下の線形変換により変換する。

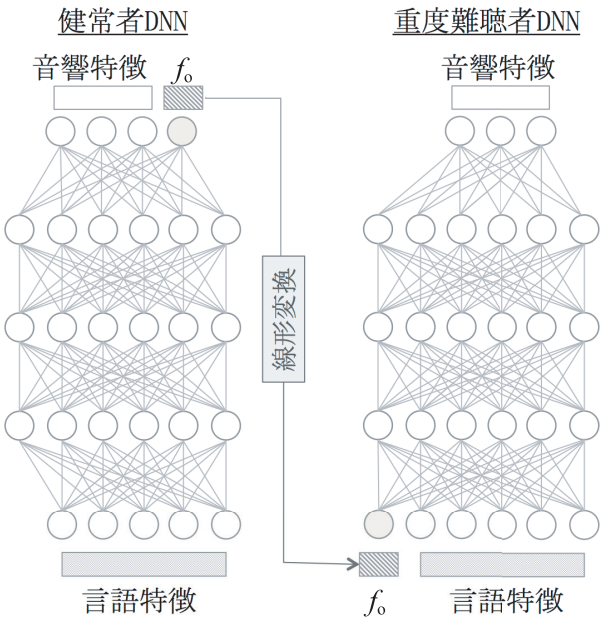
$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}(x_t - \mu^{(x)}) + \mu^{(y)}$$

上式において，x_t は健常者の t フレーム目の f_0，$\mu^{(x)}$，$\sigma^{(x)}$ は各々健常者の f_0 系列の平均と分散，$\mu^{(y)}$，$\sigma^{(y)}$ は各々障害者の f_0 系列の平均と分散である。上式は，声質変換の f_0 の線形変換と同様であるが，音声合成では健常者の（安定している）f_0 系列を用いて線形変換を行うことで，障害者の話者性を維持しつつ，聞き取り易い声に修正している。音素継続長も同様の手法により修正している。

3.3 Deep Neural Networks を用いた話者性を維持したテキスト音声合成

テキスト音声合成の枠組みとして，Deep neural networks（DNNs）を用いた音声合成が広く研究されてきている。[6] では，重度難聴者のための話者性を維持しつつ聞き取り易い合成音を生成する DNN に基づいた方法が述べられている。聴覚障害者の発話は，f_0 の変化が少なく比較的淡々としている場合や，一部の母音や子音を曖昧に発話する場合があるため，聞き取りが難しい場合がある。[6] では，HMM 音声合成 [9] と同様に健常者の音声パラメータを用いて構音障害者のモデルを修正することで，話者性を維持しつつ聞き取り易い合成音を作成している。

具体的にはまず，学習時には健常者と重度難聴者の二つの DNN を独立に学習する。健常者の DNN は入力に言語特徴量，教師に健常者の音響特徴量（スペクトル特徴）を使用して学習を行う。重度難聴者の DNN は入力に言語特徴量と重度難聴者の f_0 特徴量，教師として音響特徴量を用いて学習を行う。これにより，合成時に入力の f_0 に対応したフォルマント構造を持つスペクトルを推定する重度難聴者 DNN を構築する。



図–4 Deep neural networks を用いた話者性を維持しながら聞き取り易い合成音の生成

次に，二つの DNN を用いた f_0 修正及びスペクトル推定（音声合成プロセス）について**図–4** に示す。合成時は，健常者 DNN に言語特徴量を入力して得られる f_0 特徴量に，3.2 節の線形変換式を用いて（健常者の）f_0 系列の平均値を重度難聴者の値に修正する。この f_0 特徴量と言語特徴量を重度難聴者 DNN に入力として用いることで，入力 f_0 特徴量に対応したスペクトルを得ている。

音素継続長の修正も同様の考え方に基づき，発話のテンポ（各音素の長さの比率）に健常者の値を用いて，平均音素継続長（各音素継続長）には話者性が多く含まれているとして重度難聴者の値に線形変換する [6]。つまり障害者の発話スピードが，全体的に速い発話であれば，その平均スピードを速くし，平均発話スピードが遅ければ，平均スピードを遅くなるようにして，話者性を保ちながら，より聞き取り易い合成音を作成している。

4. おわりに

本解説では，構音障害のうち脳性麻痺者，重度難聴者に対して，話者性を維持しながら聞きとり易い声に変換する声質変換とテキスト音声合成を

紹介した。聞き取りの難しい構音障害者の発話において，母音，子音などの音素（クラス）ごとに補正が必要である。しかし声質変換では発話内容（音素系列）が正確には分からないため，構音障害者発話に対して細かい補正を行うことが難しい。

一方，テキスト音声合成では入力テキストを与えられるため，細かい補正をし易くなる。本解説では話者性を維持しながら聞き取り易い声を生成する手法として，スペクトルだけでなく，f_0，継続長の補正において，健常者の合成モデルと障害者の合成モデルの統合により実現する方法を紹介した。ただしテキスト音声合成では，誰かが事前にテキストを入力する必要がある。特に脳性麻痺者で手足に麻痺がある場合，個人でテキスト入力を行うことが困難な場合もあり，アプリケーションとしては制約がある。とはいえ，日常よく使う発話内容については，事前にテキスト音声合成を使い，声を作成しておくことで自立に向けた生活支援が期待できる。

近年，声質変換，テキスト音声合成ともに，Deep neural networks 等の深層学習に基づいたアプローチが注目されている。しかし深層学習では，一般的に多くの学習データが必要となる。また，構音障害者の苦手な発話音素，音素の欠落などの解析についても，多くの発話データが必要となる。脳性麻痺者は，発話時においても筋肉の緊張が非常に大きくなり，健常者と比べて身体への負担が大きい。そのため，多量の学習データを収集するのは容易なことではない。今後，少量学習データによる話者性を維持した音声変換手法の研究のみならず，様々な症状の構音障害者の発話に関して解析を進めていく必要がある。

文　献

[1] 日本音響学会編, 新版 音響用語辞典（コロナ社, 東京, 2003）.

[2] K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” *Speech Commun.*, **54**, 134–146 (2012).

[3] K. Tanaka, T. Toda and S. Nakamura, “A vibration control method of an electrolarynx based on statistical F0 pattern prediction,” *IEICE Trans. Inf. Syst.*, **E100-D**, 2165–2173 (2017).

[4] J. Yamagishi, C. Veaux, S. King and S. Renals, “Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction,” *Acoust. Sci. & Tech.*, **33**, 1–5 (2012).

[5] H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li and T. Nakabayashi, “PCA-based feature extraction for fluctuation in speaking style of articulation disorders,” *Interspeech 2007*, pp. 1150–1153 (2007).

[6] T. Kitamura, T. Takiguchi, Y. Ariki and K. Omori, “Individuality-preserving speech synthesis system for hearing loss using deep neural networks,” *Int. Workshop Challenges in Hearing Assistive Technology* (*CHAT*), pp. 95–99 (2007).

[7] R. Aihara, R. Takashima, T. Takiguchi and Y. Ariki, “A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary,” *EURASIP J. Audio Speech Music Process.*, doi:10.1186/1687-4722-2014-5 (2014).

[8] R. Aihara, T. Takiguchi and Y. Ariki, “Individuality-preserving voice conversion for articulation disorders using phoneme-categorized exemplars,” *ACM Trans. Access. Comput.*, **6**(4), Article No. 13, pp. 1–17 (2015).

[9] R. Ueda, T. Takiguchi and Y. Ariki, “Individuality-preserving voice reconstruction for articulation disorders using text-to-speech synthesis,” *ACM ICMI*, pp. 343–346 (2015).

滝口　哲也

1999 年奈良先端科学技術大学院大学博士後期課程修了。1999 年日本アイ・ビー・エム東京基礎研究所。2004 年神戸大学都市安全研究センター講師，2009 年同准教授，2017 年同教授。博士（工学）。2008 年ワシントン大学客員研究員。2013 年リヨン工科大学客員研究員。日本音響学会，電子情報通信学会，情報処理学会，日本小児科学会，日本小児神経学会，日本小児精神神経学会，IEEE 各会員。