# 筋萎縮性側索硬化症者の音声合成を目的とした モデル適応と声質変換の比較評価*

☆吉本拓真, 高島遼一 (神戸大), △佐々木千穂 (熊本保健科学大), 滝口哲也 (神戸大)

## 1 はじめに

近年日本国内の障害者の人数は増加傾向にあり，内閣府の調査によると，国内には身体障害者が 436 万人，知的障害者が 109.4 万人，精神障害者が 419.3 万人いるとされている [1]．また，身体障害者の中でも，在宅の聴覚・言語障害者は 34.1 万人いるとされている [2]．このような障害はコミュニケーションに対する大きな障壁となるため，バリアフリー社会の実現には円滑なコミュニケーションを行うための支援が不可欠である．

言語障害の一つに，ことばは正確に理解しており話したい言葉は明確であるものの，声を作る器官やその動きに問題があるためにうまく発話できない構音障害というものがある．また構音障害にも様々な種類があり，構音器官そのものに起こる形態的異常による器質性構音障害，構音器官の運動を制御する神経筋系の異常による運動障害性構音障害などが存在する [3]．本研究では，筋萎縮性側索硬化症（amyotrophic lateral sclerosis; ALS）者を対象とする．この病気は身体を動かすための神経系である運動ニューロンが変性することで生ずるもので，神経の命令が伝わらなくなることで筋肉が徐々に縮み，力がなくなっていく症状が現れる．したがって ALS 者は運動障害性構音障害を持つこととなる．

このような構音障害者のコミュニケーションを支援するために，近年ではスマートフォンやタブレットを用いたテキスト音声合成（text-to-speech; TTS）アプリケーションが開発されている．しかし，一般的な TTS アプリケーションでは，学習時に用いられた健常者の声をもとに合成音声が生成されるため，実際の使用者とは異なる声となる．一方で，ALS 者は話せなくなる前の声を再現してコミュニケーションを図りたいという願望がある．この問題を解決するための研究は過去にもされており，代表的なものとしてはボイスバンクプロジェクト [4] が挙げられる．本研究では，ALS 者の音声を生成するディープニューラルネットワーク（DNN）を学習することを目的とする．

ALS は進行性の病気であるため，構音障害を発症する前の音声を収録しておくことで，ALS 者本人の音声合成モデルを学習することは可能である．このとき，高品質な学習データを得るために，ALS の症状が軽い段階で収録を行うことが理想である．しかし現実的には既に症状が進行した段階で音声収録を行うケースも多く，この場合以下に挙げるような，収録音声品質上の問題が生じる．第一に，外部雑音の問題である．本来音声合成モデルの学習データは，防音室のような雑音の混入しにくい環境で収録されることが望ましい．しかし ALS が進行している場合，病室での収録や，人工呼吸器のような音を発する医療器具を装着された状態での収録が必要となり，結果として雑音が多く混入した音声が収録されることとなる．第二に，収録音声の量の問題である．健常者のデータであれば，TTS モデルの学習に十分な量のデータが比較的容易に収集できる．一方 ALS 者の場合，そもそも身体的負担から大量の収録が困難なことに加えて，症状の進行によっては収録可能な期間が限られるケースもあるため，多くのデータを収録することが困難である．第三に，収録音声の明瞭性の問題である．症状の進行度合いによっては，収録する時点で，既に健常者レベルで明瞭な音声を発することが困難となっているケースも存在する．

雑音の混入や明瞭度の低下により品質が劣化した音声を用いて音声合成モデルを学習することは，合成音声の品質劣化にも繋がり，また収録音声量が少ないことは，モデルの学習が不十分となることから，やはり合成品質の劣化の原因となる．これらの問題に対処するため，本研究では，雑音の問題に対しては音声強調を，少量データおよび明瞭度の問題に対しては，声質変換あるいは TTS モデルの適応を行うアプローチについて検討し，各手法による音声合成品質を比較評価する．

## 2 ALS 者に向けた音声合成

### 2.1 音声強調

外部雑音を含んだ ALS 者の収録音声をそのまま学習に用いると，合成音声がノイズを多く含んだものになったり，ノイズの影響でうまく学習ができなくなったりする可能性がある．そのため，本研究では，音声強調（speech enhancement; SE）処理によって収録音声のノイズを抑制したうえで学習に用いる．音声強調手法には，時間周波数マスキング [5] を用いており，本研究では学習済みモデル（`https://zenodo.org/record/4923697`）を使用する．

### 2.2 声質変換

構音障害者の音声を生成する際に考えられる手法として，健常者の声を障害者の声に変換する声質変換（voice conversion; VC）が挙げられる [6, 7]．本研究では，MaskCycleGAN-VC [8] による声質変換を行う．この手法は，CycleGAN-VC2 [9] の拡張であり，filling in frames （FIF）と呼ばれる補助タスクを用いて学習が行われる．FIF では入力スペクトログラ

*Comparative Evaluation of Model Adaptation and Voice Conversion on Speech Synthesis for a Person with Amyotrophic Lateral Sclerosis. by YOSHIMOTO, Takuma, TAKASHIMA, Ryoichi (Kobe Univ.), SASAKI, Chiho (Kumamoto Health Science Univ.), TAKIGUCHI, Tetsuya (Kobe Univ.)

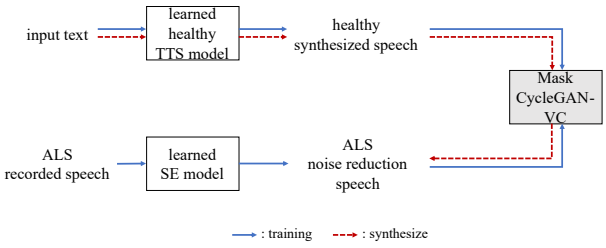


Fig. 1　The procedure of the TTS system using voice conversion model.

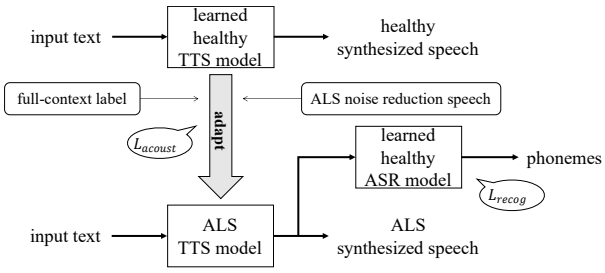


Fig. 2　Training procedure of the TTS system using model adaptation.

ムに対して部分的に時間マスクをかけ，その欠損したフレームを生成器が復元するように学習が進められる．ただしマスキングは入力スペクトログラムに対してのみであり，逆変換の入力にはマスキングを行わない．MaskCycleGAN-VC はノンパラレル声質変換であるため，ソース音声とターゲット音声のデータを合わせる必要がない．また，ソースおよびターゲット音声は 5 分程度の少量のデータで学習が行えることが報告されているため [8]，今回の ALS 者の音声合成に適したモデルであると期待される．

概要を Fig. 1 に示す．学習時，健常者音声については事前に学習しておいた TTS モデルを用いて合成音声を生成し，その音声をソース音声として使用する．このようにすることで，音声からではなくテキストから任意の合成音声を作れるようになる．ALS 者音声については 2.1 節で触れたとおり，学習済みの SE モデルに通して得られたノイズ抑制音声をターゲット音声として使用する．また，本 VC モデルでは 22.05 kHz のサンプリング周波数音声を扱うため，VC モデルへの入力の際は SoX を用いて 16 kHz よりアップサンプリングを行っている．合成時は，テキストから健常者合成音声を生成し，MaskCycleGAN-VC に通して ALS 者合成音声を得る．ただしその際のサンプリング周波数は 16 kHz にダウンサンプリングする．

### 2.3　モデル適応

声質変換のほかに構音障害者の音声合成について考えられる手法として，健常者のデータで学習した TTS モデルを ALS 者のデータを用いてファインチューニングするというモデル適応（model adaptation; MA）が挙げられる．健常者のデータは十分に確保できるため，健常者 TTS モデルの学習は十分に行うことができ，適応の際はスクラッチからの学習に比べてデータ量が少量で済むため，ALS 者のデータは少量でもうまく学習することができる．

本研究では，先行研究 [10] で行った適応のアプローチを使用する．Fig. 2 にその概要を示す．このシステムでは，音素レベルの言語特徴量から各音素の長さを推定する継続長モデル，フレームレベルの言語特徴量から波形を生成するための音響特徴量を推定する音響モデルの 2 つからなる TTS モデルを用いる．学習では初めに健常者音声データとそのラベルを用いて健常者 TTS モデルを作成する．これは 2.2 節で用いる TTS モデルと同一である．また，同様に健常者データを用いて音響特徴量からフレームごとの音素を推定する音声認識（automatic speech recognition; ASR）モデルを学習しておく．TTS モデル，ASR モデルにはどちらも 3 層の双方向 LSTM（Bidirectional LSTM）[11] を用いる．次に，学習された音響モデルに対して ALS 者音声データおよびそのラベルを用いてモデル適応を行う．但し，ALS 者音声データについては 2.1 節で触れた SE モデルを使用したノイズ抑制音声を使用する．その際，TTS モデルの出力を健常者データで学習しておいた先述の ASR モデルに入力し，それから得られる出力（音素）と正解ラベルとの損失を考慮する．すなわち，適応の際の全体の損失 $L$ は次の式のように表せる．

$$L = L_{acoust} + \alpha \cdot L_{recog} \tag{1}$$

ここで，$L_{acoust}$ はモデル適応した音響モデルの出力と実際の音響特徴量との平均二乗誤差，$L_{recog}$ は音響モデルで推定した音響特徴量を ASR モデルに入力した際の出力と実際の音素ラベルとのクロスエントロピー損失である．このように健常者データで学習した音声認識モデルによる損失を加えることで，健常者よりも明瞭性が比較的低い ALS 者の音声を用いて適応を行う際に明瞭度の低下を抑える効果が期待される．

合成においては，はじめに健常者データで学習した継続長モデルを用いて，入力されたテキストに対応する音素列の継続長をそれぞれ推定する．ただし ALS 者の話速などの本人性を反映させるため，継続長モデルから出力される正規化された音素継続長について次の処理を加える．

$$d_{(syn)} = d_{(norm)} \cdot s_{un} + \bar{d}_{dys} \tag{2}$$

ここで $d_{(norm)}$ は平均 0，分散 1 で正規化された音素継続長，$s_{un}$ は健常者の音素継続長の標準偏差，$\bar{d}_{dys}$ は ALS 者の音素継続長の平均をそれぞれ表している．式 (2) で得られる $d_{(syn)}$ を用いてフレームレベルの言語特徴量を作成する．次に得られた特徴量をモデル適応した音響モデルに入力することで音響特徴量が推定され，その特徴量から音声を合成する．

## 3　実験

### 3.1　実験条件

ALS 者の音声データには，女性 1 名が ATR デジタル音声データベース（ATR コーパス）[12] に含まれる音素バランス 216 単語を 1 単語当たり 5 回発話したものを使用する．ただし，一部収録の取りこぼしがある．また，1 回目の発話はほかの発話と比べて比較的不安定な場合があるため，実際の学習には 2 回目以降のデータを使用しており，学習には計 634 個（14 分）の音声を使用している．音声に対する音素セグメンテーション（音素とその開始・終了時間の対応付け）は強制アライメントを行い獲得した情報をもとに著者が修正を加えて作成した．声質変換およびモデル適応に用いた健常者 TTS は，ATR コーパスに含まれる女性 1 名の音素バランス文 503 文のデータを用いた．声質変換のソース音声には ATR コーパスに含まれる音素バランス単語 216 語のデータで合成した音声を使用し，モデル適応の際に用いた健常者 ASR モデルの学習には，ATR 音素バランス文 503 文のデータを男女合わせて 10 名分使用した．健常者，ALS 者とも，音声のサンプリング周波数は 16 kHz である．

音声強調モデルは，ESPnet2 [13, 14] で学習済みの SE モデルを使用した [15]．声質変換モデルとして使用した MaskCycleGAN-VC の各パラメータは文献 [8] と同様であり，エポック数は 1,000 とした．また，音声からメルスペクトログラムへの変換およびボコーダに関しては MelGAN [16] を使用した．健常者 TTS モデルの学習は先行研究 [10] と同様で，最適化には Adam を用いており学習率は 1e-3 である．モデル適応の際にも最適化には Adam を用いており，学習率は 1e-4 とした．また，式 (1) における $\alpha$ の値は 2.0 としており，音響特徴量抽出およびボコーダには WORLD [17, 18] を用いた．ここでの音響特徴量は，メルケプストラム 60 次元，帯域非周期性指標，対数基本周波数，有声/無声フラグで構成され，有声/無声フラグ以外に関しては 2 次までの動的特徴量を含んだ計 187 次元からなり，次元ごとに平均 0 分散 1 となるよう正規化（標準化）を行っている．TTS モデルの学習及び適応時に用いるフルコンテキストラベルは，Open JTalk [19] のフロントエンド部を利用して生成した 38 種類の音素（空白を含む）からなる HTS 形式のものを使用した．

### 3.2　実験結果

#### 3.2.1　スペクトログラムの変化

「勢い/i k i o i/」という単語について，ALS 者の収録音声およびノイズ抑制音声，声質変換手法（VC）によって生成した合成音声，モデル適応手法（MA）によって生成した合成音声のスペクトログラムを Fig. 3 に示す．収録音声ではあらゆる領域に一定以上のパワーがあることから音声全体にノイズがかかっていることが確認できるが，音声強調モデルに通した後の音声は不要な部分のパワーが減少していることから

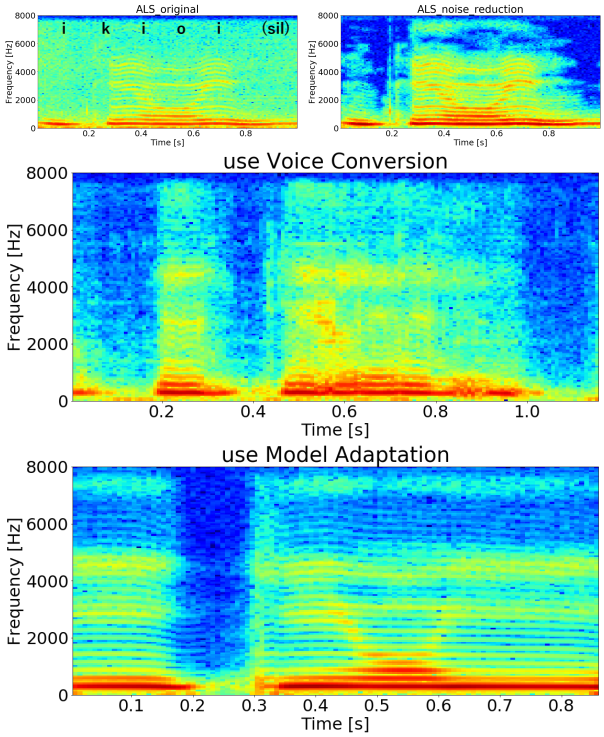


Fig. 3　Spectrogram of recorded speech (upper left), noise reduction speech (upper right), and synthesized speech using VC (middle) and using MA (lower).

ノイズが低減されていることが分かり，また 3 つ目の音素から 5 つ目の音素までの/i/，/o/，/i/における第 2 フォルマントの変化 [20] がより明確になっている．合成音声を比較すると，声質変換による合成音声は学習にノイズ抑制音声を用いたにも関わらずノイズが多く含まれた音声が合成されており，無音区間などにパワーがまだらに分布している部分があることで確認できる．また 4 つ目の音素/o/以降の第 2 フォルマントが不明瞭になっており，発話内容も聞き取りづらいことが分かる．これはモデルの学習の際にマスクされた領域を周囲のフレームから予測するため，ノイズ抑制音声に残されていたノイズを復元してしまい声質として学習したことが要因だと考えられる．一方，モデル適応による合成音声は無音区間でのパワーは弱くなっており，また第 2 フォルマントの変化もノイズ抑制音声の場合と同じような変化が視認できることから，ALS 者の音声に似た明瞭な音声が生成できていることが分かる．無音区間を示す sil も含めたラベル付きのデータで適応を行っているため，各音素における話者の特徴を効率よく学習できたものと考えられる．また，健常者 ASR モデルの損失も考慮していることから，より聞き取りやすい音声になるように適応が進められたと考えられる．

### 3.3　主観評価実験

音声の品質を評価するために，平均オピニオン評点（mean opinion score; MOS）を用いた聴取実験を

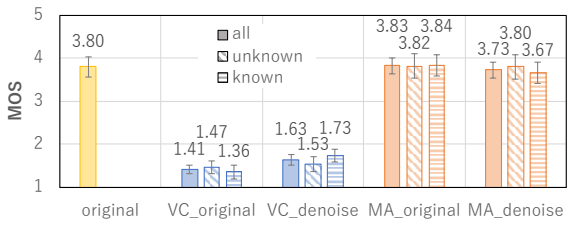


Fig. 4　Results of MOS test with nine listening subjects. Error bars represent 95% confidence intervals.

行った．10 種類の単語それぞれについて，収録音声および 4 モデルでの音声の計 5 パターンを用意し，9 名の参加者がヘッドフォン聴取により評価した．ただし，学習時に用いなかった単語（未知単語）での合成音声についても評価を行うため，10 種類のうち 5 種類の単語については収録音声が存在しない未知単語であり，合計 45 個の音声を評価した．Fig. 4 に実験の結果を示す．はじめに，VC による合成音声は収録音声（original）に対して品質が悪いという結果となった．ただし，音声強調によるノイズ抑制音声を用いて学習したモデル（VC_denoise）の方が収録音声をそのまま用いて学習したモデル（VC_original）よりも品質が優れていることが分かった．ノイズが抑制された音声を学習に用いたことで，合成音声に含まれるノイズを低減することができ音質が向上したと考えられる．なお，既知単語（known）と未知単語（unknown）での有意な差は見られなかった．つぎに，MA による合成音声は，既知単語においても未知単語においても収録音声と同等の品質であるという結果となった．また，収録音声をそのまま学習データとして用いたモデル（MA_original）とノイズ抑制音声を学習データとして用いたモデル（MA_denoise）とでは有意な差は見られなかった．この結果からも，MA は収録音声のノイズに対して頑健であることが分かる．

## 4　おわりに

本研究では，ALS 者を対象とした音声合成について，データ量が少なくても学習が行える声質変換とモデル適応の 2 つのアプローチの比較評価を行った．また，ALS 者の収録音声にはノイズが多く含まれているため，学習済みの音声強調モデルで雑音除去を行い，そのデータで各モデルの学習を実行した．実験の結果，ALS 者を対象としているという条件下では，声質変換による手法よりもモデル適応の手法の方が品質の高い音声を合成できることが分かった．

声質変換モデルについては今回一話者対一話者の変換モデルを使用したが，声質変換モデルのパラメータチューニングや，他の声質変換モデルを検討することで，改善の余地がある．モデル適応の場合，部分的にイントネーションが平坦になるなどの不自然さが残ることがあるため，今後はより自然な音声を生成するための方法について模索する．また，適応時に考慮した ASR モデルの損失について，明瞭性を担保するためのより効果的な手法についても検討していく．

**謝辞**　本研究の一部は，JSPS 科研費 JP21H00906, JP22K12168 の支援を受けたものである．

## 参考文献

[1] 内閣府, “令和 3 年版 障害者白書,” 2021.
[2] 厚生労働省, “平成 30 年版 厚生労働白書,” 2019.
[3] 菊谷武 他, “歯科医師のための構音障害ガイドブック,” 医歯薬出版, 2019.
[4] J. Yamagishi *et al.*, “Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction,” *Acoustical Science and Technology*, 33 (1), 1-5, 2012.
[5] D. Yu *et al.*, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” *ICASSP*, 241-245, 2017.
[6] L. W. Chen *et al.*, “Generative adversarial networks for unpaired voice transformation on impaired speech,” *Interspeech*, 719–723, 2019.
[7] 松原圭亮 他, “CycleVAE 型声質変換を用いた構音障害者のための高明瞭度音声合成,” 音講論 (春), 783-786, 2021.
[8] T. Kaneko *et al.*, “Maskcyclegan-VC: Learning Non-Parallel Voice Conversion with Filling in Frames,” *ICASSP*, 5919-5923, 2021.
[9] T. Kaneko *et al.*, “Cyclegan-VC2: Improved Cyclegan-based Non-parallel Voice Conversion,” *ICASSP*, 6820-6824, 2019.
[10] 吉本拓真 他, “音響モデルの話者適応に基づく脊髄性筋萎縮症者の音声明瞭化の検討,” 音講論 (秋), 1053-1056, 2021.
[11] M. Schuster, K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, 45 (11), 2673-2681, 1997.
[12] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech Communication*, 9 (4), 357-363, 1990.
[13] T. Hayashi *et al.*, “Espnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit,” *ICASSP*, 7654-7658, 2020.
[14] T. Hayashi *et al.*, “ESPnet2-TTS: Extending the Edge of TTS Research,” arXiv, 2021.
[15] C. Li *et al.*, “ESPnet-SE: End-To-End Speech Enhancement and Separation Toolkit Designed for ASR Integration,” *IEEE Spoken Language Technology Workshop (SLT)*, 785-792, 2021.
[16] K. Kumar *et al.*, “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis,” *NeurIPS*, 32, 2019.
[17] M. Morise *et al.*, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE transactions on information and systems*, E99-D (7), 1877-1884, 2016.
[18] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, 84, 57-65, 2016.
[19] “Open JTalk,” `http://open-jtalk.sourceforge.net/`
[20] T. Hirahara, R. Akahane − Yamada, “Acoustic Characteristics of Japanese Vowels,” *Proc. ICA*, 3287-3290, 2004.