構音障害者音声認識におけるテキスト音声合成によるデータ拡張の検討*

☆松坂勇樹, 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

構音障害とは，言葉を理解しているにもかかわらず，発声器官や神経などの異常によって，言葉を正しく発話できない状態である．構音障害となる原因や病気はいくつか存在し，口唇口蓋裂，脳性麻痺などがある．本研究では，脊髄性筋萎縮症 (spinal muscular atrophy; SMA) に起因する構音障害を対象とする．

脊髄性筋萎縮症者の多くは身体を動かすことが困難であるため，日常生活に支障が出てしまう．そのため AI スピーカーのような音声認識 (automatic speech recognition; ASR) を活用したアシスタント機能が，身体障害者の支援システムとして期待されている．しかし，構音障害者の発話は健常者の発話とは特徴が大きく異なるため，健常者の音声を用いて学習させた既存の ASR モデルで構音障害者の音声を認識することは困難である．そのため，構音障害者の音声を正確に認識するためには，構音障害者本人の音声を用いて ASR モデルを学習する必要がある．しかし，構音障害者にとって音声の収録は負担が大きいため，学習用音声を少量しか用意できないという問題がある．

この学習データ不足の問題を解決する方法の一つとして，事前に大量の健常者音声で ASR モデルを学習し，その後少量の構音障害者音声を用いてファインチューニングを行う，モデル適応のアプローチが研究されている [1]．また，学習データを人工的に生成するデータ拡張のアプローチも研究されている [2]．本研究では，テキスト音声合成 (text-to-speech; TTS) によって生成した構音障害者合成音声を用いたデータ拡張を検討する．この方法では，構音障害者の収録音声で TTS モデルを学習し，TTS によって ASR モデル学習用の構音障害者合成音声を作成する．脊髄性筋萎縮症者を対象とした音声認識実験により，提案手法の有効性を評価する．

2　テキスト音声合成によるデータ拡張

本研究における，ASR モデルの学習方法を Fig. 1 に示す．TTS を用いて学習用合成音声を生成するフェーズ (TTS フェーズ) と，収録音声と生成した合成音声を用いて ASR モデルを学習するフェーズ (ASR フェーズ) で構成されている．TTS フェーズでは，まず収録した構音障害者音声及びそれらと対になるテキストデータ (フルコンテキストラベル) を使用して，構音障害者音声合成用の TTS モデルの学習を行う．次に，学習済みの TTS モデルと生成したい音声のテキストを使用し，構音障害者合成音声を生成する．ASR フェーズでは，構音障害者収録音声に加えて生成した合成音声も用いることでデータ拡張を行い，それらと対になるテキストデータ (正解ラベル) とともに ASR モデルの学習を行う．

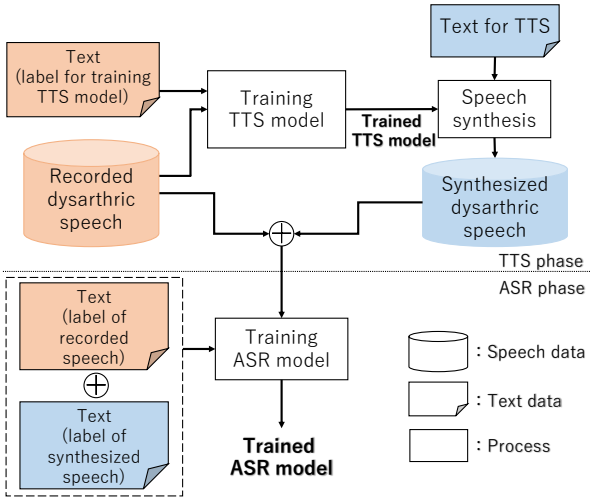

Fig. 1　Procedure for training ASR model using synthesized dysarthric speech.

3　DNN パラメトリック音声合成

本研究では，合成音声を生成するための TTS モデルとして，DNN パラメトリック音声合成 [3] の構造を使用する．近年では，Tacotron2 [4] などの End-to-End の TTS モデルが研究されており，自然性の高い音声が合成可能であることが知られている．しかし，一般に End-to-End の TTS モデルは多くの学習データが必要である．一方，DNN パラメトリック音声合成のモデルは End-to-End の TTS モデルと比べて少量の音声データでも学習が可能である．そのため，学習用音声データの不足が問題となる構音障害者の音声合成に適していると考え，本研究では DNN パラメトリック音声合成モデルを採用した．

*Data augmentation using text-to-speech for dysarthric speech recognition. by Yuki Matsuzaka, Ryoichi Takashima (Kobe University), Chiho Sasaki (Kumamoto Health Science University), Tetsuya Takiguchi (Kobe University)

4　音声認識タスクの概要

本研究では，単語音声を用いて，二種類の音声認識タスクにより提案法の有効性を評価した．また，合成音声生成時に使用するテキストもタスクに合わせて変更している．

1 つ目は既知単語認識タスク，すなわち音声認識の評価用音声の発話単語が既知の場合を想定したタスクである．このタスクでは，ASR モデルの学習データと評価データに含まれる発話内容が同じ単語でそれぞれ構成される．そのため，少量の学習データを用いて学習を行っても比較的高精度に認識が可能と考えられる．しかし，構音障害者の発話は健常者と比較して安定しないため，同じ発話内容であっても発話ごとに特徴がばらつきやすいという問題がある．そこで，合成音声の生成時に使用するテキストには学習および評価データと同じ単語を用いることで，同一単語内における学習データのバリエーションを増やし，それにより前述のばらつきの問題を軽減することが期待される．

2 つ目は未知単語認識タスク，すなわち評価用音声の発話単語が未知の場合を想定した認識タスクである．このタスクでは，ASR モデルの学習データと評価データに含まれる発話内容がそれぞれ異なる単語で構成される．そのため，既知単語認識タスクと比較して認識が難しいタスクである．そこで，合成音声の生成時に使用するテキストには学習用収録音声と異なる単語を使用することで，学習データ内の発話内容のバリエーションを増やし，それにより学習データと評価データの発話内容のミスマッチによる悪影響を軽減することが期待される．

5　評価実験

5.1　データ設定

本実験では，脊髄性筋萎縮症者 1 名による収録音声を使用する．ATR デジタル音声データベース [5] に含まれる音素バランス単語 216 語を，各単語につき 5 回発話した音声を収録した (216 単語× 5 発話=1,080 発話)．収録音声のサンプリング周波数は 16 kHz である．

既知単語認識タスクにおいて，音声認識モデルにおける収録音声データの振り分けとしては 5 発話の内の発話ごとに分割し，216 単語× 3 発話を学習データ，216 単語× 1 発話を開発データ，216 単語× 1 発話を評価データとした．また，このタスクにおける TTS モデルでは，音声認識モデルの学習データである 216 単語× 3 発話の中から選択した 2 発話 (216 単語× 2 発話) を学習データとし，3 通りの選び方があることから計 3 つの TTS モデルをそれぞれ学習した．学習した 3 つの TTS モデルそれぞれにおいて，収録音声と同じ単語である 216 単語の合成音声を生成することで，計 216 単語× 3 発話の合成音声を生成した．音声認識モデル学習時には，学習用収録音声である 216 単語× 3 発話に加えて，生成した同じ単語の 216 単語に関する合成音声を使用した．このとき，各単語につき 1 発話，2 発話，3 発話 (× 216 単語) 追加した場合で評価を行い，性能を比較した．

未知単語認識タスクにおいては，音声認識モデルにおける収録音声データの振り分けとしては 216 単語の単語ごとに分割し，150 単語× 5 発話を学習データ，30 単語× 5 発話を開発データ，36 単語× 5 発話を評価データとした．また，このタスクにおける TTS モデルでは，音声認識モデルの学習データと開発データを合わせた 180 単語× 5 発話を学習データとした．このとき，TTS モデルの学習に使用する発話内容のバリエーションを増やすために，音声認識モデルの開発データも含めている．学習した TTS モデルによって音声を合成する際は，収録音声とは異なる発話単語として，ATR デジタル音声データベースの重要単語 5240 語の内の 750 単語を使用して合成音声を生成した．音声認識モデル学習時には，学習用収録音声である 150 単語× 5 発話に加えて，750 単語× 1 発話の合成音声を使用した．また，未知単語認識タスクでは，音声認識モデルの学習の際には，健常者音声データによる事前学習を行った．健常者音声データとしては，JSUT コーパス [6] の basic5000 を使用した．既知単語認識タスクでは，健常者事前学習による性能改善が見られなかったため，行っていない．

5.2　モデルとパラメータの設定

TTS モデルには，文献 [7] を参考にした上で DNN パラメトリック音声合成のモデルを使用した．継続長モデルは中間層が 3 層の全結合で構成され，入力には 325 次元の音素単位の言語特徴量，出力は各音素の継続長とした．音響モデルに関しても同様に中間層が 3 層の全結合で構成され，入力は 329 次元のフレーム単位の言語特徴量とし，出力はメルケプストラム 40 次元，連続対数基本周波数 1 次元，有声/無声フラグ 1 次元，帯域非周期性指標 1 次元，加えて 1 次, 2 次の動的特徴量 (有声/無声フラグは除く) を含めた計 127 次元の音響特徴量とした．また，音声波形の生成や，音響モデルの教師データに使用する音響特徴量の取得などには WORLD [8, 9] を使用した．

ASR モデルには，文献 [10] を参考にした上で CTC [11] のモデルを使用した．入力特徴量には 40 次元の対数メルフィルタバンク特徴量を使用した．モデルの中間層は 5 層の双方向 GRU で構成され，各

Table 1　PER [%] of each training dataset in ASR models (known word recognition task).

Training data set	PER[%]
recoded data	13.84
+synthetic data(216words × 1utt.)	12.16
+synthetic data(216words × 2utts.)	11.79
+synthetic data(216words × 3utts.)	11.51

Table 2　PER [%] of each training dataset in ASR model (unknown word recognition task).

Training data set	PER[%]
recoded data (w/o pre-training)	60.03
recoded data (w/ pre-training)	49.58
+synthetic data	46.22

GRU 層の出力後に次元削減のための線形層を 1 層ずつ追加している．また，GRU 層と線形層の間においてフレーム数のサブサンプリング処理も一部の層で行い，2 層目と 3 層目でそれぞれ入力フレーム数を 2 分の 1 に削減している (最終的にフレーム数を 4 分の 1 に削減)．5 層の GRU 層 (+サブサンプリング+線形層) の後には出力層として線形層 1 層があり，音素単位での認識を行うため，出力は各音素ごとに出力ノードを持つように設定した．バッチサイズは 10，オプティマイザには AdaDelta[12] を使用した．学習率の設定として初期学習率は 1.0 に設定し，エポックごとに開発データによる評価を行うことで学習率減衰の処理も行った．

5.3　実験結果

既知単語認識タスクにおける結果を Table 1 に示す．ASR モデルにおける学習データの構成による認識性能を比較するために，音素誤り率 (Phoneme Error Rate; PER) で評価を行った．比較する学習データセットとしては，ベースラインである収録音声 (216 単語 × 3 発話) のみで学習した場合，そして収録音声に加えて合成音声を 1, 2, 3 発話 (216 単語× 1, 2, 3 発話) 使用したときで 3 つの場合があり，計 4 つの場合それぞれで ASR モデルの学習・評価を 10 回行った平均の音素誤り率を示している．結果より，合成音声も学習に加えることで音素誤り率が改善していることがわかる．ただし，合成音声を 2 発話以上に増やしても，0 から 1 発話に増やしたとき程の改善は見られなかった．

未知単語認識タスクにおける結果を Table 2 に示す．Table 1 と同様に，ASR モデルにおける学習データの構成それぞれにおいて，ASR モデルの学習・評価を 10 回行った平均の音素誤り率を示している．比較では学習データセットとして，収録音声 (150 単語 × 5 発話) のみで学習する場合，収録音声に加えて合成音声 (750 単語× 1 発話) も使用する場合を比較している (ただしどちらも健常者事前学習を行った)．また，健常者事前学習の効果を確認するために，収録音声のみで学習する場合において事前学習の有無による認識性能も比較した．結果の値より，まず収録音声のみで学習する場合において，健常者事前学習を行うことで音素誤り率が大きく改善していることから，健常者事前学習の有効性が確認できる．そして収録音声に加えて合成音声も学習に使用することで，さらに音素誤り率が改善している．しかし，合成音声によるデータ拡張を行った場合においても音素誤り率は 40%以上と依然に高い値であった．

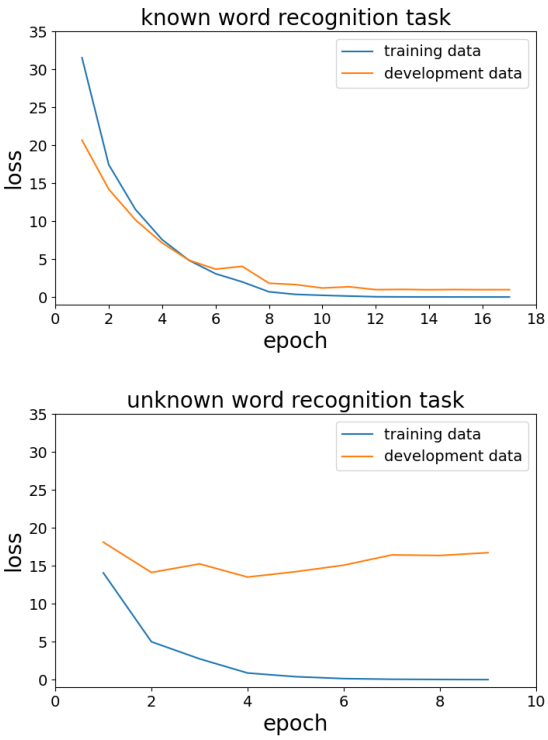

Fig. 2　Training curves of ASR models using synthetic speech on the known word recognition task (216words × 3utts.) and the unknown word recognition task.

両タスクにおいて，合成音声を学習データに加えて ASR モデルの学習を行った際の，学習・開発データのエポックごとの損失値の変化の一例を Fig. 2 に示す．既知単語認識タスクでは学習データに合成音声を 216 単語× 3 発話加えた場合を示している．図より，既知単語認識タスクでは学習・開発データともに

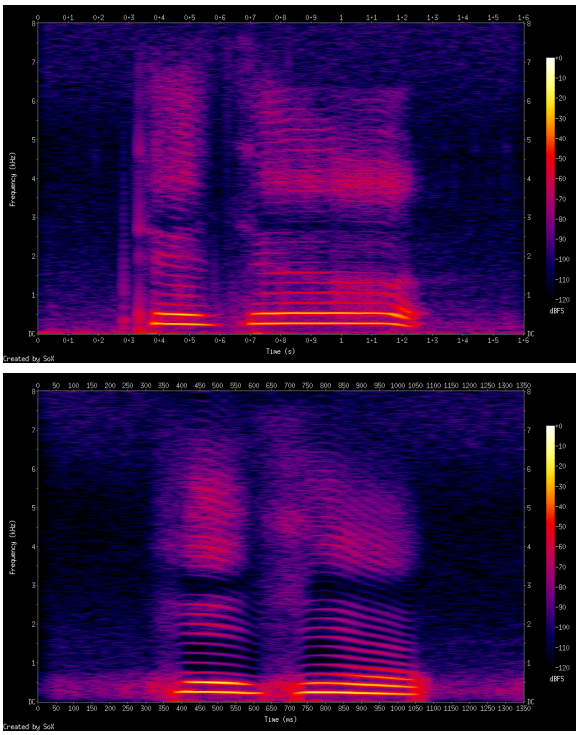

Fig. 3　Spectrograms of the recorded speech (top) and the synthesized speech (bottom) for the word “t e ch o o”.

損失が下がっているのに対して，未知単語認識タスクでは開発データの損失が下がっていないことが確認できる．これは，合成音声を使用しない場合においても同様の傾向が見られており，学習データと開発データの発話内容が異なることに起因するものと考えられる．未知単語認識タスクでは，音声の発話内容の違いに対して頑健に ASR モデルを学習するために様々な発話内容の合成音声を学習データとして追加したが，合成音声の音質の低いために頑健性の改善が効果的にされなかったと考えられる．「手帳」と発話した際の収録音声と，合成音声をスペクトログラムで比較したものを Fig. 3 に示す．合成音声では収録音声と比較して，スペクトル構造の再現が不十分であり，TTS モデルの学習データが少量であったことが原因の一つと考えられる．このようなスペクトル構造のミスマッチが，認識性能向上を妨げる原因であったと考えられる．

6　おわりに

本研究では，構音障害者音声認識における学習データ不足の問題を改善するために，テキスト音声合成を用いたデータ拡張を検討した．二つの単語認識タスクの実験結果に基づき，TTS モデルで生成した構音障害者合成音声を音声認識モデルの追加の学習データとして使用することで，構音障害者音声の認識精度向上が確認できた．しかし，合成音声の音質に関して課題があることも確認し，合成音声の音質を改善することができれば，さらなる認識精度向上も期待できる．また，本研究では単語音声を取り扱ったが，文章音声において合成音声によるデータ拡張を行うことも検討している．

参考文献

[1] R. Takashima, T. Takiguchi, and Y. Ariki, “Two-step acoustic model adaptation for dysarthric speech recognition,” Proc. ICASSP, pp. 6104-6108, 2020.

[2] K. Fujiwara *et al.*, “Data augmentation based on frequency warping for recognition of cleft palate speech,” Proc. APSIPA, 2021.

[3] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” Proc. ICASSP, pp. 7962-7966, 2013.

[4] J. Shen *et al.*, “Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,” Proc. ICASSP, pp. 4779-4783, 2018.

[5] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, 9, pp. 357-363, 1990.

[6] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” arXiv preprint arXiv:1711.00354, 2017.

[7] 山本龍一, 高道慎之介 , “Python で学ぶ音声合成,” インプレス, 2021.

[8] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE transactions on information and systems, E99-D (7), pp. 1877-1884, 2016.

[9] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, 84, pp. 57-65, 2016.

[10] 高島遼一, “Python で学ぶ音声認識,” インプレス, 2021.

[11] A. Graves *et al.*, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” ICML, pp.369-376, 2006.

[12] M. D. Zeiler, “ADADELTA: An adaptive learning rate method,” arXiv preprint arXiv:1212.5701, 2012.