

健常者の子音との組み合わせを用いた構音障がい者音声合成*

☆長久保 諒¹, 高島 遼一^{1,2}, 佐々木 千穂³, 滝口 哲也¹

(¹神戸大学, ²JST さきがけ, ³熊本保険科学大学)

1 はじめに

発話障がい的一种として構音障がいが存在する。これは、言語の理解に問題は見られず発話内容は明確であるが、発声発話器官及びその動作に何らかの異常が見られるために思うように発話ができず不明瞭な発話になってしまう障がいである。本研究では構音障がいの中でも脊髄性筋萎縮症 (Spinal Muscular Atrophy: SMA) と脳性麻痺 (Cerebral Palsy: CP) に伴うものを対象とする。前者は脊髄の運動神経細胞の病変によって起こる神経原性の筋萎縮症であり、後者は脳への損傷によって起こる運動機能の障がいである。いずれも構音障がいを引き起こし、コミュニケーションを困難にしている。発話障がい者のコミュニケーションを支援する技術として、テキスト音声合成 (Text-to-Speech: TTS) システムが注目されている。しかし、一般的な TTS システムは明瞭な音声を合成可能である一方で、使用者本人ではない他者の声質で合成されるため、自身の声でコミュニケーションを取りたいという需要を満たせない。

本研究では、構音障がい者本人の声質で明瞭な音声を合成可能な TTS システムの作成を目的とする。まず、構音障がい者本人の声質で合成可能にするためには本人の音声でモデルの学習を行う必要があるが、構音障がい者にとって音声収録は健常者以上に大きな負担となるため、大量の音声データを収録することが難しい。そこで本研究ではその補完を目的として、大量の健常者音声を用いて学習した事前学習モデルに対して少量の構音障がい者音声を用いてファインチューニングを行う。これによって構音障がい者の声質は反映されるが、同時に構音障がい者音声の発話特徴も反映され不明瞭な音声が合成されてしまう。この問題を解決するために、本研究では健常者音声特徴との組み合わせを用いることで構音障がい者音声特徴の中でも特に不明瞭な子音の代替を行う手法を提案する。

2 提案手法

調査の結果、構音障がい者の発話においては母音と比較して子音の発音が著しく不明瞭な傾向にある事がわかった。これを改善するために、提案手法では、

音声の話者性は子音よりも母音に強く現れると仮定し、不明瞭な子音の音声特徴を健常者の音声特徴で代替することを目指す。

提案手法の概要を Fig. 1 に示す。通常の音声合成システムでは、対象となる一名の話者の発話音声データとその発話内容に対応する音素単位をトークンとしたテキストラベルを用いて学習を行い、合成の際には入力テキストを音素単位のトークンに変換して合成を行うことで、対象話者の声質で音声合成を行う。一方、提案手法では、健常者と構音障がい者の音声にそれぞれ別の種類としてトークンを割り当てたうえで同時に2話者の学習を行い、合成の際には子音は健常者音声に割り当てたトークンを、母音は構音障がい者音声に割り当てたトークンを用いて合成することで、健常者子音と構音障がい者母音を組み合わせる。

3 実験

3.1 実験条件

本実験で用いる構音障がい者の音声データには脳性麻痺者男性2名 (CP1, CP2) 及び脊髄性筋萎縮症者女性1名 (SMA) が ATR コーパス [1] に含まれる音素バランス文 503 文を朗読した録音音声を使用する。健常者の音声データには同じく ATR コーパスに含まれる音素バランス文 503 文を朗読した健常者男性音声 (MHO) 及び健常者女性音声 (FTK) のそれぞれ1名ずつを、対象となる構音障がい者の性別に合わせて用いた。音声合成モデルは JSUT コーパス [2] を用いて学習した VITS [3] を事前学習モデルとして使用し、このモデルを構音障がい者を用いて従来手法と提案手法でそれぞれファインチューニングを行ったモデルで学習データに含まれていない50発話のサンプルを合成し、これを用いて明瞭性及び話者性の2点において比較する。

明瞭性の評価には、合成音声に対して音声認識モデルで音声認識を行い、その文字誤り率 (Character Error Rate: CER) を用いる。音声認識モデルとして、Transformer [4] ベースのモデルを日本語話し言葉コーパス (CSJ) [5] を用いて学習したものを用いた。話者性の評価には x-vector [6] に基づく話者埋め

*Speech synthesis for a person with dysarthria using combinations with consonants of physically unimpaired speaker. by Ryo Nagakubo¹, Ryoichi Takashima^{1, 2}, Chiho Sasaki³, Tetsuya Takiguchi¹ (¹Kobe University, ²JST PRESTO, ³Kumamoto Health Science University.)

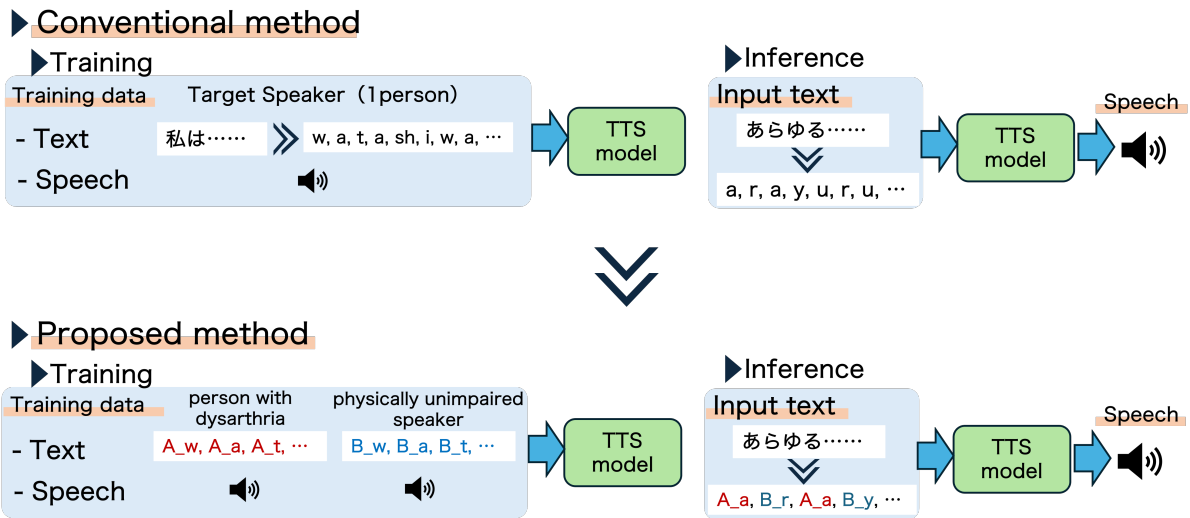


Fig. 1 Overview of the proposed method using combinations with consonants of physically unimpaired speaker.

込みベクトルのコサイン類似度 (Speaker Embedding Cosine Similarity: SECS) を用いる。

3.2 提案手法の明瞭性評価

収録音声及び合成音声に対する文字誤り率を Fig. 2 に示す。提案手法 (proposed method (vowel only)) は従来手法 (conventional method) や収録音声 (recorded speech) と比較して文字誤り率が低下しており、特に CP1 においては健常者のトークンを用いて音声合成した結果 (synthesized speech of physically unpaired speaker) と同等の文字誤り率を達成している。このことから、提案手法によって明瞭性を改善できていると言える。

3.3 提案手法の話者性評価

構音障がい者の収録音声と合成音声との間の話者埋め込みベクトルのコサイン類似度を Fig. 3 に示す。類似度は発話ごとに計算しており、図中の数値はその平均値を、エラーバーは最大値及び最小値をそれぞれ表している。図より、提案手法 (proposed method (vowel only)) による類似度は従来手法 (conventional method) よりも低下していることが確認された。この結果は、提案手法で生成された音声の話者性が従来手法と比較して元の音声から離れていることを示している。一方、健常者の収録音声と合成音声との間の話者埋め込みベクトルのコサイン類似度を Fig. 4 に示す。図より、提案手法では従来手法よりも類似度が上昇していることがわかった。これは、提案手法による合成音声に健常者音声の話者性が一部反映されていることを示している。特に、提案手法では、構音障がい者音声との類似度の最小値と健常者音声との類似度の最大値が平均値から大きく外れていることから、一部の合成音声は健常者の音声に非常に近い特

性を持つことが確認でき、これが話者性が大きく低下した要因であると考えられる。しかし、CP1 および SMA の結果では、提案手法における構音障がい者音声との類似度が健常者音声との類似度を大幅に上回っていることから、話者性の維持が一定程度達成されていると考えられる。

3.4 特定子音の構音障がい者音素への置換

前節では発話の一部において健常者の話者性が見られることが確認された。そこで、本節では音声合成時に母音に加えて特定の子音の音素についても構音障がい者のトークンを用いる実験を行う。主な手順として、子音を健常者のトークン、母音を構音障がい者のトークンを用いて数発話分音声合成を行い、その音声において健常者者の話者性が出現した箇所の先頭に対応する子音の音素を対象として、その音素及び関連する調音方式 [7] の音素を構音障がい者のトークンに置換して合成を行う。具体的には CP1 は {m,k,b,g,h,d,sh} (鼻音, 破裂音, 摩擦音), CP2 は {r,y,b,cl,ky,f,p} (半母音, 破裂音, 促音, 拗音, 摩擦音), SMA は {b,k,h,m,w} (破裂音, 摩擦音, 鼻音, 半母音) の音素を対象とし、母音に追加して、対象子音音素が含まれる調音方式グループを単位として構音障がい者トークンにした場合と対象子音音素のみを構音障がい者トークンにした場合の2パターンで評価を行った。明瞭性の評価のために、この手法における文字誤り率の結果を Fig. 2 に前者を proposed method (vowel + specific articulation groups) として、後者を proposed method (vowel + specific consonants) として示す。また、話者性の評価のために、この手法における構音障がい者音声及び健常者音声との類似度を Fig. 3 及び Fig. 4 に前者を proposed method (vowel + specific articulation groups) として、後者

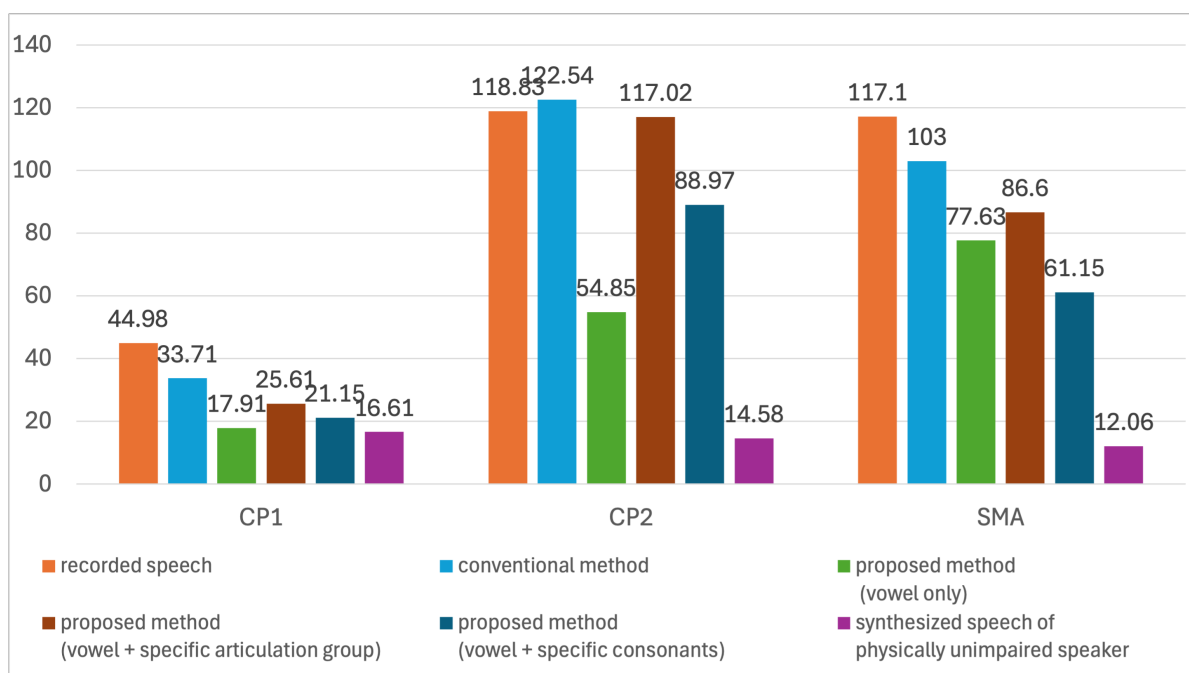


Fig. 2 Character error rates [%] on speech recognition experiments.

を proposed method (vowel + specific consonants) として示す。

まず、調音方式グループを単位として子音音素を置換した場合の話者性の結果について、構音障がい者音声との類似度の向上、並びに健常者音声との類似度の低下が確認された。この結果は、健常者の話者性を取り除き構音障がい者の話者性により近い音声を合成することに成功したことを示している。一方、明瞭性の結果については悪化が見られ、特に CP2 は収録音声や従来手法と同等の誤り率を示してしまった。この原因は子音の大半を構音障がい者のものに置換したことで、不明瞭な発話特徴の大半も反映されてしまったことであると考えられる。続いて、対象子音音素のみを置換した場合の話者性の結果について、CP1 と SMA については調音方式グループを単位として子音音素を置換した場合と同等の話者性のまま明瞭性の改善に成功し、母音のみを構音障がい者トークンに置換する手法と同等の誤り率を示した。一方、CP2 においては話者性が損なわれてしまったものの、健常者よりも構音障がい者に話者性の類似度が高い状態で明瞭性の改善に成功した。

まとめると、母音に加えて特定子音の構音障がい者音素への置換を行うことで、構音障がい者音声への類似度を大幅に向上させることができた。これより、母音以外の音素にも話者性への影響がある程度見られることや、明瞭性への影響が小さく話者性を向上できる子音音素の存在もみられることがわかった。一方で、子音の置換を大量に行うと大幅に明瞭性が損なわれるため。適切な子音音素を特定する手法を

検討する必要がある。

4 まとめ

本研究では、構音障がい者のコミュニケーション支援を目的として、健常者の子音との組み合わせを用いた手法によって構音障がい者音声合成における明瞭性の改善を目指した。結果について、従来手法からの明瞭性の改善に成功した一方で、話者性については一部に学習に用いた健常者の特徴が表れてしまう問題が見られ、特に対象話者によっては構音障がい者よりも健常者への類似度が高くなる結果となってしまう。この問題に対して、特定子音の構音障がい者音素への置換を行うことで話者性の向上には成功したが、同様に対象話者によっては明瞭性が大幅に損なわれるため依然として完全な解決には至っていない。今後は置換を行う特定子音の選出手法など、話者性に関する問題の解決手法を検討することによって、構音障がい者音声合成における話者性及び明瞭性のさらなる向上を目指す。

謝辞 本研究の一部は、JST さきがけ JPMJPR23I7 および JSPS 科研費 JP23K20733, JP22K12168 の支援を受けたものである。

参考文献

- [1] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and syn-

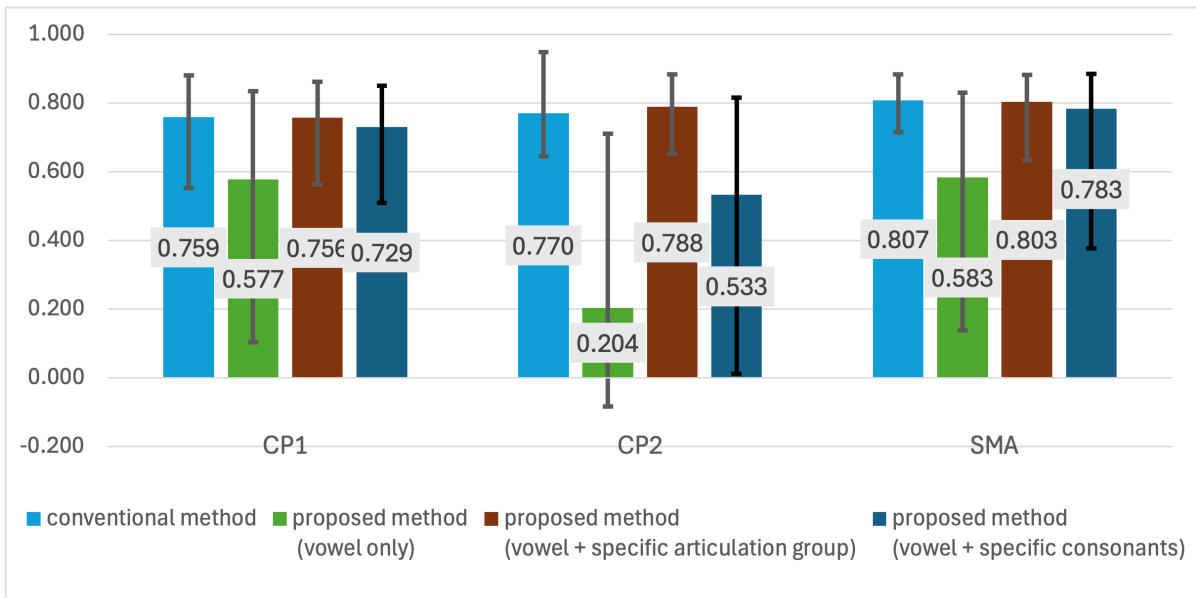


Fig. 3 Speaker embedding cosine similarity to speech of a person with dysarthria.

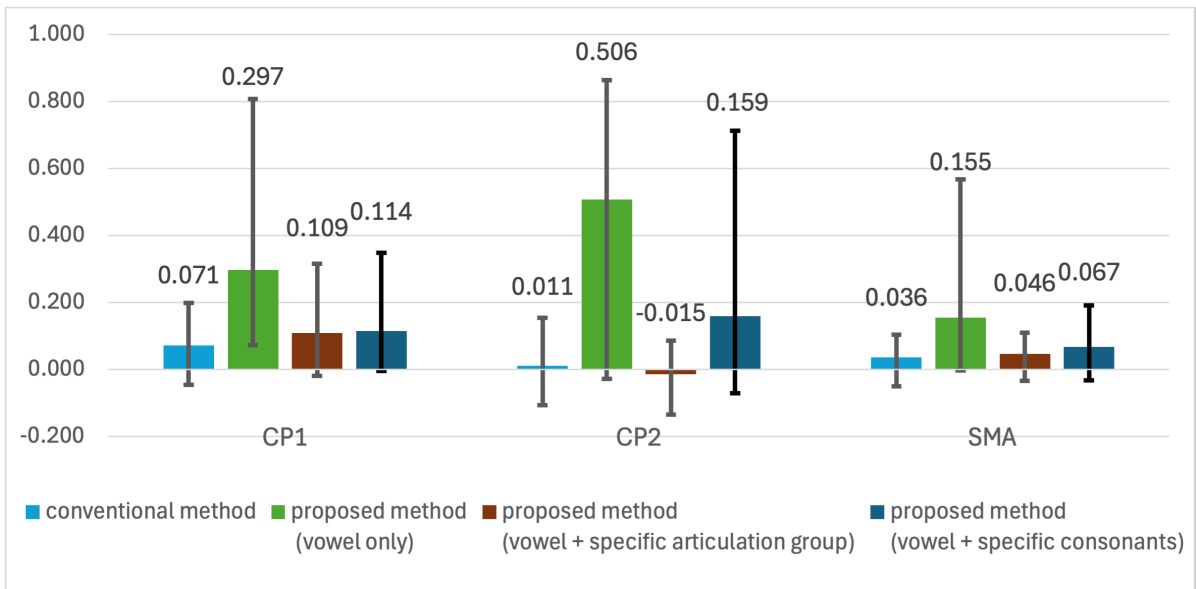


Fig. 4 Speaker embedding cosine similarity to speech of a physically unimpaired speaker.

thesis,” *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

- [2] R. Sonobe *et al.*, “JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis,” *arXiv preprint arXiv:1711.00354*, 2017.
- [3] J. Kim *et al.*, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.
- [4] A. Vaswani *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

tems, vol. 30, 2017.

- [5] K. Maekawa, “Corpus of spontaneous japanese: Its design and evaluation,” in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.
- [6] D. Snyder *et al.*, “X-vectors: Robust dnn embeddings for speaker recognition,” in *IEEE international conference on acoustics, speech and signal processing*. IEEE, 2018, pp. 5329–5333.
- [7] 鹿野 清宏 他, “音声・音情報のデジタル信号処理” 昭晃堂, デジタル信号処理シリーズ, 1997.