脳性麻痺音声認識のための日本語および英語障害者音声を用いた音響モデルの学習

土師, 梧刀
高島, 遼一
滝口, 哲也

(Citation)
神戸大学都市安全研究センター研究報告,28:13-18

(Issue Date)
2024-03

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/0100490305

(URL)
https://hdl.handle.net/20.500.14094/0100490305

神戸大学都市安全研究センター
研究報告，第28号，令和６年３月

脳性麻痺音声認識のための日本語および英語障害者音声を用いた音響モデルの学習

Acoustic model training using Japanese and English dysarthric speech for speech recognition of a person with cerebral palsy

土師 梧刀 [1)]
Kirito Haze
高島 遼一 [2)]
Ryoichi Takashima
滝口 哲也 [3)]
Tetsuya Takiguchi

概要：本研究では構音障害者、特に脳性麻痺者のための音声認識技術の開発を目的とする。脳性麻痺者は手足の動作が不自由である場合が多いため、ハンズフリーで操作できる音声認識デバイスは魅力的である。しかし、脳性麻痺者の発話スタイルは健常者と大きく異なっており、既存のモデルでは音声認識が困難である場合が多い。そのため、本研究ではOpenAI社が公開している多言語音声認識モデルWhisperに着目する。Whisperは日本語音声に対して高い認識性能を示しているが、構音障害者音声に対する認識精度は低下する課題がある。本研究ではWhisperを日本人の構音障害者音声でFine tuningすることによって認識率の向上を目指した。その結果、学習前と比べて認識率が大幅に上昇することが確認された。また、構音障害者は発話の際の心身への負担から音声を大量に収録することが難しいという課題があり、それに対処するため、英語の構音障害者音声も学習させる実験も行った。その結果、日本語障害者音声のみを学習した場合と比べて、さらに1%程度誤り率が改善した。

キーワード：構音障害、脳性麻痺、音声認識、Fine tuning

1. はじめに

近年、音声認識技術の普及により、日常生活などの様々な環境下で音声認識システムの利用が期待されている。本研究で対象にする脳性麻痺者は手足の動作が不自由である場合が多いため、ハンズフリーで操作できる音声認識デバイスは魅力的である。しかし、脳性麻痺者は自分の思い通りの発話ができない構音障害を持っていることが多い。構音障害者をはじめとする発話障害者の発話スタイルは健常者の発話スタイルとは異なっており、健常者を対象とした既存の音声認識モデルでは、構音障害者の音声認識が困難である場合が多い。また、構音障害者は発話の際に心身への負担が大きいため、大量の音声を収録することができないという課題もある。すなわち、少量の学習データのみで構音障害者の発話スタイルに適応した音声認識システムを開発することが必要とされている。もし構音障害者の音声を認識できるシステムがあれば、健常者もその発話内容を理解することができ、円滑なコミュニケーションを行うことができる。したがって、構音障害者を対象にした音声認識技術の開発は必要不可欠であり、それにより健常者と障害者の社会活動における差別の解消も期待される。

2. 提案手法

(1) モデル構造

本研究では、2022 年 9 月に OpenAI 社によって公開された大規模音声認識モデル Whisper[1] を用いて構音障害者の音声認識を行う。Whisper とはインターネット上から収集された 680,000 時間の多言語データを用いて学習された大規模な音声認識モデルである。単一のモデルで、多言語の音声認識、言語特定、英語への翻訳ができることが特徴である。多言語の音声認識においては、音声の言語を自動で特定し認識結果を出力することができるが、認識する際の言語を指定することも可能である。大規模でかつ多様なデータで学習されたため、アクセント、背景雑音、専門用語に対する堅牢度が高いとされている。図 1 に Whisper のモデル構造を、図 2 に特別なトークンを使用したデコーダーの出力フォーマットを示す。

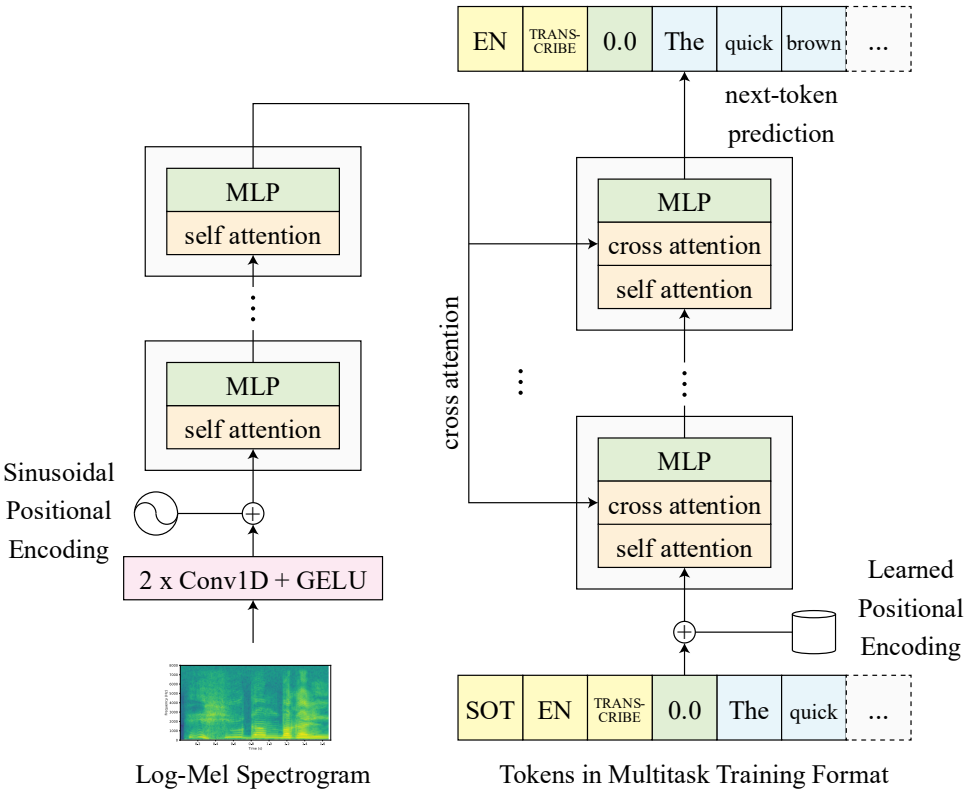

図 1　Whisper のモデル構造。

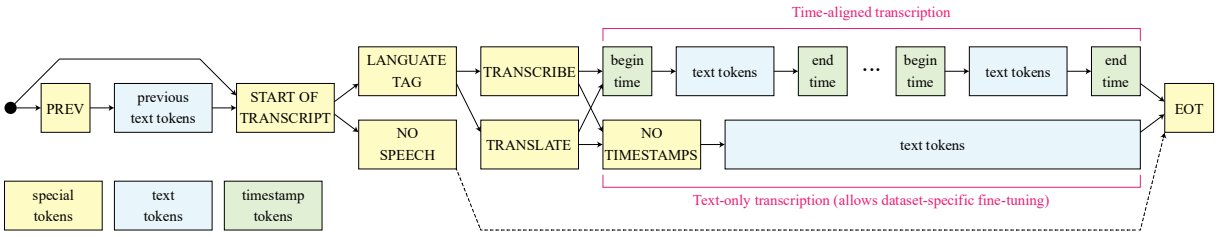

図 2　マルチタスク学習フォーマット。

Whisper は Encoder-Decoder ベースの Transformer[2] として実装されている。入力音声はリサンプリングによってサンプリング周波数が 16,000Hz に変換され、30 秒ごとのセグメントに分割される。それから、25 ミリ秒窓、10 ミリ秒ストライドで 80 次元のログメルスペクトルグラムに変換され、

Input Embedding として、2 層の畳み込み層を通過する。畳み込み層はフィルターサイズ 3、活性化関数は GELU であり、2 層目はストライドを 2 とする。次に、Sin 関数を使用した Positional Encoding が行われ、それから Encoder によって潜在表現が抽出される。抽出された情報は Decoder に渡され、入力した音声に対応する認識結果と、そこに付与された特別なトークン(言語情報やタスク情報、タイムスタンプなど)が出力される。

(2) Whisper を用いた構音障害者音声認識モデルの学習

Whisper は日本語音声に対して高い認識性能を示すが、Whisper の学習に用いられた音声のほとんどは健常者の音声であると考えられるため、健常者の音声に対しての認識性能が高くても、構音障害者の音声に対する認識性能が高いとは限らない。そのため本研究では、Whisper モデルを障害者音声で Fine tuning することにより性能の向上を検討する。まず Whisper を日本語障害者音声データセットのみで Fine tuning しその性能を評価する。また、構音障害者の音声が少量しか収録できないという課題に対処するため、Whisper を日本語障害者音声と英語障害者音声を混ぜたデータセットで Fine tuning しその性能を評価することも行う。 ただし、評価時は日本語構音障害者音声のみを認識させ評価する。

3. 評価実験

(1) 使用するデータ

日本語構音障害者音声は、アテトーゼ型脳性麻痺者男性 2 名(J1, J2)について、ATR 研究用日本語音声データベース [3]に含まれる音素バランス文を、J1 は 429 文、J2 は 501 文収録した。 日本語発話データのうち、50 文を評価データ、残りを訓練データとして使用した。英語構音障害者音声は、TORGO データベース [4]で提供される構音障害者 8 名(E1, E2, …, E8)の音声を使用した。

(2)モデル設定及び学習手法

提案手法は文字認識タスクで評価を実施した。Whisper にはサイズの異なる 5 つのモデルが用意されており、小さいものから tiny、base、small、medium、large である。本研究では Whisper の small モデルを使用した。また、学習率は 1e-5、学習エポック数は 3 とした。

学習前の事前準備として、データセットの音声それぞれに言語情報(日本語あるいは英語)を付与した。実験ごとに学習データや評価データの異なるデータセットを複数作成し、学習データは作成した段階で一度だけ順番をランダムに並び替えた。また、モデルが認識結果を出力する際の言語として、日本語障害者音声のみで学習する際は、学習時及び評価時共に出力言語として日本語を指定した。一方、日本語障害者音声と英語障害者音声の両方を学習する際は、学習時には出力言語を指定せず、評価時のみ出力言語を日本語に設定した。また、評価時にモデルの認識結果及び正解ラベルに対して正規化を行った。正規化とは具体的に述べると、句読点の削除、クエスチョンマーク及び空白の削除、数字及びアルファベットの全角文字への統一である。

(3) 実験結果

a) 日本語障害者の音声を用いた学習結果

日本語障害者音声のみを用いて Whisper モデルを学習させた場合の文字誤り率(character error rate; CER)を図 3 に示す。一般に誤り率の値が小さい方が認識精度は高い。ここで、Not fine-tuned は Fine tuning を行う前のモデルでの認識結果であり、これをベースラインとする。また、Epoch 1、2、3 はそれぞれ 1、2、3 エポック学習時点での認識結果の誤り率である。2 人の障害者 J1 及び J2 のいずれにおいても、Fine tuning 前は文字誤り率が大きいが、Fine tuning によってその誤り率が大幅に低下しており、障害者音声に対する認識精度が高くなったことがわかる。このことから、Fine tuning によって Whisper が日本語構音障害者音声に対して適応することができたと考えられる。

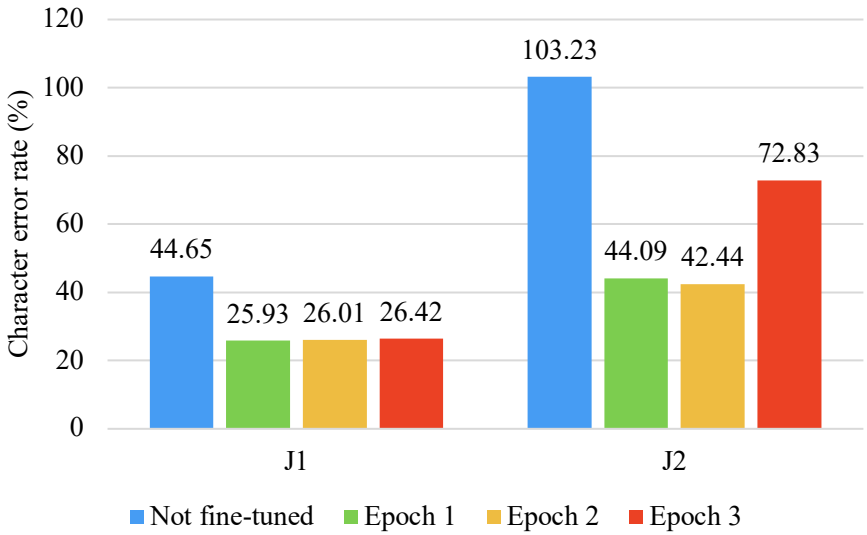

図 3　日本語構音障害者音声での学習結果。

b) 日本語障害者 1 名と英語障害者 1 名ずつの音声を用いた学習結果

日本語障害者 J1 の音声と英語障害者 8 名(E1, …, E8)のうちのそれぞれ 1 名ずつの音声を混合したデータセットで Fine tuning を行った場合の、3 エポック分のモデルによる認識結果の誤り率を表 1 に示す。ただし、評価は J1 の評価データのみで行った。

表 1　日本語障害者 1 名と英語障害者 1 名の音声を用いた学習結果。

Datasets	Epoch 1	Epoch 2	Epoch 3	Mean (%)	Min (%)
J1	25.93	26.01	26.42	26.12	25.93
J1+E1	26.01	25.77	26.01	25.93	25.77
J1+E2	29.74	26.09	26.74	27.52	26.09
J1+E3	28.44	26.09	25.45	26.66	25.45
J1+E4	28.04	27.15	25.77	26.99	25.77
J1+E5	27.55	26.01	26.34	26.63	26.01
J1+E6	26.74	26.01	25.53	26.09	25.53
J1+E7	27.39	25.85	26.09	26.44	25.85
J1+E8	28.93	25.69	26.01	26.88	25.69

Mean 及び Min は、データセットごとの 3 エポック分の誤り率の平均値及び最小値をそれぞれ意味する。また、Datasets における例えば J1+E1 とは、J1 の学習データと E1 の学習データをランダムに混合したデータセットで学習したことを意味する。また、Datasets における J1 は、J1 のみの学習データで学習したことを意味する。3 エポック分の誤り率の最小値に注目すると、J1 のみで学習した場合は 25.93%なのに対し、J1+E3 では 25.45%、J1+E6 では 25.53%などと、それよりも低い値となっている。すなわち、日本語障害者音声のみで学習した場合と比べて、日本語と英語の両方の障害者音声で学習した方が、誤り率が小さくなったものが複数存在する。

c) 日本語障害者 1 名と英語障害者複数名の音声を用いた学習結果

次に、日本語障害者 J1 の音声と英語障害者複数名の音声を混合したデータセットで Fine tuning を行った場合の、3 エポック分のモデルによる認識結果の誤り率及びその平均値と最小値を表 2 に示す。ただし、評価は J1 の評価データのみで行った。

表２　日本語障害者１名と英語障害者複数名の音声を用いた学習結果。

Datasets	Epoch 1	Epoch 2	Epoch 3	Mean (%)	Min (%)
J1	25.93	26.01	26.42	26.12	25.93
J1+E3+E6	27.55	25.28	25.36	26.06	25.28
J1+E3+E6+E8	28.77	27.23	25.69	27.23	25.69
J1+E1+E3+E6+E8	27.71	26.42	24.96	26.36	24.96
J1+E1+E3+E4+E6+E8	26.50	26.74	25.53	26.26	25.53
J1+E1+E3+E4+E6+E7+E8	25.61	27.31	26.26	26.39	25.61
J1+E1+E3+E4+E5+E6+E7+E8	29.17	26.14	26.18	27.16	26.14
J1+E1+E2+E3+E4+E5+E6+E7+E8	28.20	25.93	26.18	26.77	25.93

本実験においては、表１の誤り率の最小値が低い順に、そのデータセットに含まれる英語障害者を新たなデータセットに追加した。すなわち、E3、E6、E8、E1、E4、E7、E5、E2 の順で英語障害者音声を追加した。その結果、英語障害者１名の音声しか使用しなかった場合と比べて、英語障害者複数名の音声を使用した方が、誤り率が小さくなったものが複数存在した。最も値の低い誤り率は J1 及び E1、E3、E6、E8 の音声で学習されたモデルによる 24.96%であり、これは日本語障害者音声のみで学習された結果の最小値である 25.93%を 1%程度下回る結果となった。したがって、日本語障害者音声だけでなく英語障害者音声も学習させることによって日本語障害者音声の認識精度を上昇させることができるといえる。

4. まとめ

本研究では、音声認識モデル Whisper を用いた構音障害者の音声認識について検討した。Whisper モデルを日本人構音障害者の音声で Fine tuning させることで、構音障害者音声の特定話者に対する認識性能が向上することを示した。加えて、英語障害者の音声も Fine tuning に使用することでさらなる認識性能の向上が見られることを示した。今後はより大きな medium や large といったモデルを学習させ、さらなる認識性能の向上を検討していく。

参考文献

１）Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I., Robust speech recognition via large-scale weak supervision, *International Conference on Machine Learning*, 28492-28518, 2022.

２）Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., Attention is all you need, *Advances in neural information processing systems 30*, 5998-6008, 2017.

３）Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., and Shikano, K., ATR Japanese speech database as a tool of speech recognition and synthesis, *Speech Communication*, 9, 4, 357-363, 1990.

４）Rudzicz, F., Namasivayam, A. K., and Wolff, T., The TORGO database of acoustic and articulatory speech from speakers with dysarthria, *Language Resources and Evaluation*, 46, 4, 523-541, 2012.

筆者：１）工学部情報知能工学科、学生；２）都市安全研究センター、准教授；３）都市安全研究センター、教授

Acoustic model training using Japanese and English dysarthric speech for speech recognition of a person with cerebral palsy

Kirito Haze
Ryoichi Takashima
Tetsuya Takiguchi

Abstract

In this study, we aim to develop speech recognition technology to facilitate smooth communication for people with articulation disorders, especially those with cerebral palsy. Since people with cerebral palsy often have difficulty with limb movements, speech recognition devices that can be operated hands-free are attractive and expected to be utilized. However, the speech style of people with cerebral palsy differs significantly from that of unimpaired people, and existing speech recognition models often have difficulty in recognizing it.

Therefore, this study focuses on the speech recognition model Whisper published by OpenAI, which is a multilingual speech recognition model trained on 680,000 hours of multilingual data collected from the internet. It boasts high recognition performance for Japanese speech, but its recognition accuracy for dysarthric speech is not good. Thus, in this study, we investigate fine-tuning of Whisper with Japanese dysarthric speech. As a result, a significant increase in recognition rates was confirmed compared to before the training. In addition, since people with cerebral palsy have difficulty in recording many voices due to the physical and mental strain of speech, we also conducted an experiment in which English dysarthric speech was also learned. As a result, the error rate was improved by about 1% compared to the case where only Japanese dysarthric speech was learned.