K-033

モバイルアプリケーション上で動作する構音障害者向け音声認識の検討
Investigation of Speech Recognition for Dysarthric People Running on Mobile Applications

山本 堅心† 高須 俊輔† 中村 直人†
Kenshin Yamamoto Shunsuke Takasu Naoto Nakamura

1 まえがき

構音障害は，脳血管疾患や運動機能障害に起因し，言葉を正常に発音する能力が損なわれる障害である．構音障害者の発話は，特に初対面の相手や介護者にとって聞き取りが困難な場合が多く，円滑なコミュニケーションの障壁となっている．

この課題に対し，構音障害者のコミュニケーションを支援するさまざまなアプローチが検討されてきた．その一つとして，著者らはリハビリテーションで用いられる絵カードに着想を得て，定型文を選択・発声するアプリケーションを開発した [1]．しかし，事前に登録された語彙しか利用できず，表現が限定される点や，四肢に障害を併発する患者にとっては操作自体が困難であるという課題があった．

一方，構音障害者の音声を対象とした認識技術として，土師らは大規模音声認識モデル Whisper を構音障害者の音声でファインチューニングする手法を提案し，特定話者に対する認識精度が向上することを示した [2]．しかし，この手法は実用的なシステムとして実装されておらず，誰もが容易に利用できる段階には至っていない．

そこで本研究では，構音障害者の発話をモバイルアプリケーション上で認識し，即座に合成音声として出力するコミュニケーション支援手法を提案する．

2 提案アプリケーション

本研究で提案するアプリケーションは，利用者がスマートフォンに向かって発話すると，デバイス内に実装された音声認識モデルがその内容をテキストに変換し，即座に音声合成として読み上げる．これにより，聞き手は発話内容を正確に理解することが可能となる．

本アプリケーションの全体像を図 1 に示す．最大の特長は，音声認識から音声合成までの一連の処理を，外部サーバーとの通信を介さず，すべてデバイス上で完結させる点にある．これにより，病院内や災害時など，通信環境が不安定または利用不可能な状況においても，安定したリアルタイムのコミュニケーション支援を提供できる．

3 提案手法

本研究では，モバイルアプリケーション上で構音障害者の発話を高精度に認識し，円滑なコミュニケーションを支援することを目的とする．そのために，先行研究 [2] に基づき，Whisper を構音障害者の音声でファインチューニングし，得られたモデルをモバイルアプリケーション上で効率的に動作させる手法を提案する．本章では，学習済みモデルをモバイルアプリケーション上で動作させるための変換プロセスと実装方法について後述する．

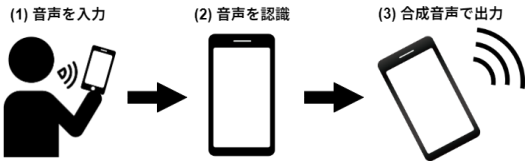

図 1 アプリケーションの全体像

3.1 モデルの変換と実装

ファインチューニングにより得られた構音障害者音声認識モデルを，iOS デバイス上で動作するモバイルアプリケーションに実装するため，モデルの変換を行う．

3.2 WhisperKit Tools

WhisperKit Tools は，Hugging Face Transformers などで学習・ファインチューニングされた PyTorch ベースの Whisper モデルを，WhisperKit が利用可能な Core ML モデル形式に変換する機能を提供するツールキットである．本研究では，このツールキットを用いて，ファインチューニング済みの構音障害者音声認識モデルを変換する．

3.3 WhisperKit

WhisperKit は，Apple の Core ML フレームワークと統合し，iPhone や iPad などの Apple 製ハードウェア上で Whisper モデルの効率的なローカル推論を実現するために設計された音声認識パッケージである．

CPU および Neural Engine を最大限に活用することで，低遅延かつ省電力なオンデバイス音声認識を可能にする．また，WhisperKit はオンデバイスですべての処理を行うため，外部サーバーとの通信を必要とせず，通信環境の制約を受けずに安定した利用が可能である．

4 予備実験

提案手法の有効性を検証するため，認識精度の定量的評価を行う予備実験を実施した．実験には，実験に参加していただいた構音障害のある話者 2 名 (以下，話者 A，話者 B) の発話録音データを用いる．単語は，著者らのプロトタイプモデル [1] で使用されたものを参考に 6 つの単語を選定した．モバイルアプリケーション上での実利用を想定し，データはスマートフォンの内蔵マイクで収録した．

† 千葉工業大学大学院 Chiba Institute of Technology

4.1 実験条件

ファインチューニングに用いた 6 つの単語と，各話者の収録数を表 1 に示す．

表 1: 実験に用いた発話データ

単語	話者 A	話者 B
ありがとう	35	45
寒い	44	40
飲みたい	42	40
トイレ	46	40
取って	53	40
つらい	43	40

音声認識モデルには，モバイルアプリケーション上での動作負荷を考慮し，軽量な Whisper-base モデルを採用した．本実験では話者 A の認識精度に着目し，以下の 2 つの条件でモデルを学習させ，性能を比較した．

- **条件 1:** 話者 A の音声のみでファインチューニング
- **条件 2:** 話者 A と話者 B の両者の音声でファインチューニング

いずれの条件でも，データを学習用に 80%，検証用に 10%，テスト用に 10%の割合で分割した．

4.2 実験結果

話者 A のテストデータに対する各条件での実験結果を表 2 に示す．評価指標には単語誤り率 (Word Error Rate: WER) と文字誤り率 (Character Error Rate: CER) を用いた．

この結果から，ファインチューニングによって認識精度が大幅に向上することが確認できる．さらに，条件 2 が条件 1 を上回る精度を示したことから，多様な構音障害者のデータを用いて学習することが，特定話者に対する認識性能の向上にも寄与する可能性が示唆された．これは，モデルがより頑健な音響的特徴を獲得したためと考えられる．

表 2: 話者 A のテストデータに対する認識精度

	WER	CER
条件 1	10.5%	6.7%
条件 2	3.2%	1.8%

5 評価実験

予備実験で学習した条件 1 および条件 2 のモデルをモバイルアプリケーションに実装し，実環境における認識性能を定性的に評価した．この評価実験は，ファインチューニング済みモデルが，実際のアプリケーション環境でどの程度の性能を発揮するかを確認することを目的とする．

5.1 実験環境

実験には，ファインチューニング済みモデルを実装した iPhone 12 を用いた．発話は静かな室内環境で行い，話者とスマートフォンのマイクとの距離は約 30cm に設定した．

5.2 実験結果

話者 A による認識結果の一例を図 2 に示す．ファインチューニング前のモデルでは，「寒い」という発話が「サムル」と誤認識された (同図左)．一方，ファインチューニング後のモデルでは正しく認識された (同図右)．実験で用いた他の 5 単語についても，ファインチューニング後のモデルは全て正確に認識することを確認し，実用レベルの性能が示された．

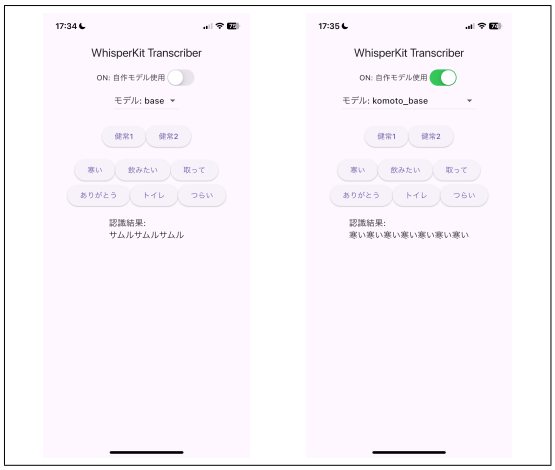

図 2 アプリケーション上での認識結果の比較 (左: FT 前, 右: FT 後)

6 おわりに

本研究では，構音障害者の円滑なコミュニケーションを支援するため，ファインチューニングした音声認識モデルを搭載したモバイルアプリケーションを提案し，その有効性を検証した．構音障害者の音声でファインチューニングした Whisper モデルをモバイルアプリケーションへ実装することにより，オフライン環境でもリアルタイムで高精度な音声認識が可能であることを確認した．

本稿で報告したのは，音声認識機能のモバイルアプリケーションへの搭載と，その基本性能の検証までであるが，この成果は，著者らが提案したボタン選択式アプリケーション [1] が抱える語彙の制約や操作性の課題を克服し，より実用的な支援の実現に向けた重要な一歩である．

今後の展望として，収録データ数を拡充し，より多様な発話や語彙に対応することで認識精度をさらに向上させることが挙げられる．また，日常会話のような連続音声や，さまざまな実環境下でのコミュニケーションを想定した実用性の評価を進めていく．

参考文献

[1] 中村直人, 佐野秀憲, 小林由美, ” 構音障害者支援スマホアプリの開発”, 2023 年電子情報通信学会総合大会, p.185
[2] 土師梧刀, 高島遼一, 滝口哲也, ” 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”, 神戸大学都市安全研究センター研究報告, 28:13-18