補助事業成果報告書

補助事業の名称	SureTalkの高度化によるきこえない人へのアクセシビリティ向上のための研究開発
補助事業の概要	きこえない人へのアクセシビリティ向上のため、AI を活用して SureTalk を高度化するための下記研究開発を行う。 １．手話認識精度向上のための読唇技術、正解ラベルの自動抽出機能開発 ２．利用者の利便性向上のための手話認識変換技術等の開発 ３．SureTalk機能向上のための音声⇒手話出力機能、手話⇒音声出力機能開発

【研究開発の実施内容と成果】

１．手話シーンの撮影

今年度研究開発にあたって、別紙の「FY24撮影実績（最終報告用）.xlsx」に記載されている通り、手話シーンの撮影を行った。

２．手指認識、読唇認識を統合したEnd-to-Endモデル構築

２－１．研究開発の実施内容

(1)発話シーンの収集

初年度で収集した49単語で読唇技術および簡易ハイブリッドモデルの評価を行ったが、さらに有用性の評価をするために、100単語を選定追加し新たに収集した。表１に例を示す。2単語は同じ手指動作で異なる意味をもつ単語を示す。手話単語の選定には、自治体でよく使われる単語リストから選定した。

申請書	補助金
申込用紙	支援金
窓口	住民
受付	在住者

(2)読唇技術を組み込んだEnd-to-Endモデル開発

今年度は口の動きも含めた全体の情報から手話テキストを正確に予測するタスクとし、手話シーンにおける口の動きを読み取る読唇技術を手話認識システムに統合するためのEnd-to-Endモデルの開発に取り組んだ。

End-to-Endモデルのアーキテクチャの設計にあたって、前年度に研究開発した手話の骨格情報からダイレクトに日本語テキストに翻訳できるAIアルゴリズム（fairseqモデル）をベースとして採用した。しかし、このベースモデルは全身骨格情報を一列に並べて線形層に入力するアプローチとなっており、隣接情報の関係（一つ一つのデータがどの部分と関連しているかの情報）が失われることが精度低下の一因となっていると考えた。そこでこの問題に対処するため、読唇技術として骨格情報の本来の形状や関係性をそのまま理解できるニューラルネットワーク（GNN）を導入したモデルアーキテクチャを採用することにより精度向上を図った。さらに後述するエッジリストやモデル学習方法の改善にも取り組んだ。

２－２　研究開発の成果

(1)発話シーンの収集

SB社員、筑波技術大学および九州工業大学の学生協力のもと12名の話者数分の手話単語791シーンを収集した

(2)End-to-Endモデル開発と認識実験

以下2パターンの読唇技術とfairseqの言語モデルを組み合わせたEnd-to-Endモデルアーキテクチャを開発した。

【パターン1】 全身骨格情報を一つの入力とするモデルアーキテクチャ

顔の情報を含めた全身の骨格情報を単一の入力データとして扱い、手話動作全体の動きを統一的に解析するような特徴点抽出層とfairseqの言語モデルを組み合わせたモデルアーキテクチャ

【パターン2】: 顔の骨格と身体骨格を分けて入力するモデルアーキテクチャ

顔の骨格情報と顔以外の骨格情報を別々の入力として取り扱い、顔の動き（読唇を含む）を解析するための特徴点抽出層と、体全体の動きを解析するための特徴点抽出層をそれぞれ設け、これら二つの出力をfairseqの言語モデルに入力するモデルアーキテクチャ

それぞれパターン1を図1、パターン2を図2に示す。

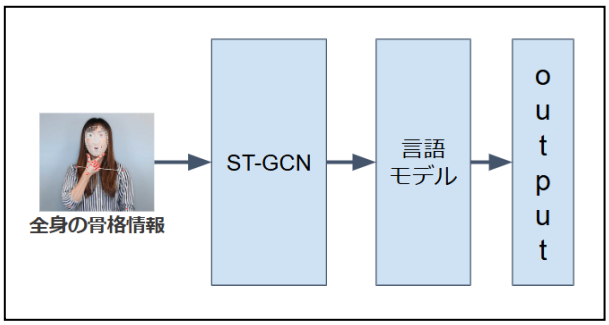


図１ パターン1

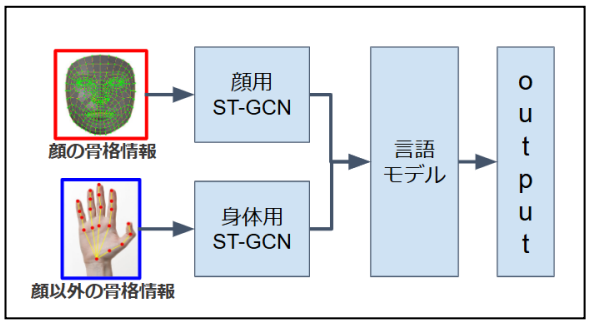


図２ パターン2

前年度、今年度で集めた発話シーンに加え、すべての収集したデータを用いて、認識試験を行った結果を以下に示す。それにより読唇技術を導入する有効性を確認した。

【テストデータに対するBLUE評価結果（日本語ベース）】

ベースモデルのBLEUスコア：68.1

パターン１BLEUスコア：68.7

パターン２BLEUスコア：69.3

また、これまでの読唇技術の結果に基づいて構築されていたエッジリスト(どの特徴点を使うか、どの点同士をつなげるかといった要素の集合体)についても、手話認識に最適化するために再検討し、再評価を行った。パターン2をベースにエッジリストを3パターンに分け、認識試験を行った結果を図３に示す。その結果から特徴点が40、エッジ数が80のエッジリストが3パターンの中でBLEUスコアが高い結果になった。

上記すべての認識試験は、発話シーン以外も含む全てデータを用いた結果であるため、発

話シーンに絞った詳細な認識試験は来年度にて取り組みを行う。

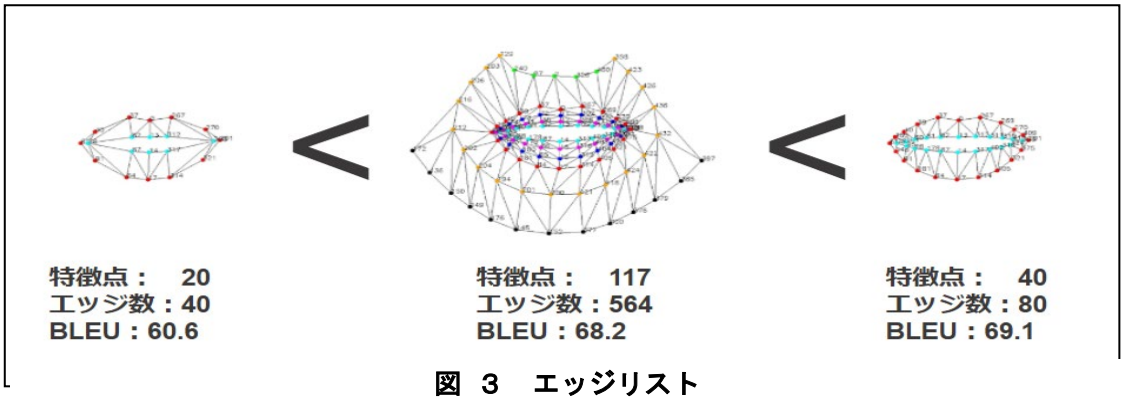


図 ３　エッジリスト

２－３　研究成果の考察と令和7年度計画の方針策定について

今年度は、手話認識システムに読唇モデルを組み込んだEnd-to-Endモデルの構築を行い、認識実験の結果、読唇技術を導入する有効性を確認できた。また、処理コストも商用展開しているfairseqシステムと同水準に抑制することができた。

一方で、本研究のEnd-to-Endモデルでは、読唇が口のみを見ているため日本語対応手話に限定されていることが課題として明らかになった。具体的には日本語対応手話は口の動き（読唇）で判断できるが、日本手話では眉や顎の動きなど顔全体の情報Non Manual Marker(以下、NMM)が重要であり、読唇だけでは対応できない。そのため、より広範な日本手話に対応するため読唇範囲を口から顔全体へ拡大し、NMMを考慮したモデルアーキテクチャの構築に取り組む。また更に認識精度をあげるために、特徴点のみを用いたアプローチだけでなく、映像をそのまま用いるアプローチも検討を進めていく。映像ベースの有効性については研究分担者の研究グループで提案済みである（荒金&齊藤、MIRU2023）。この成果を手話認識に利用することに取り組む。

３．話者適応モデルに関する技術開発

３－１．研究開発の実施内容

(1)手話データ収集に関するツールの開発

YouTubeには多くの地方手話の映像が公開されているため、これらを学習モデルに反映できるシステムの構築が求められる。今回は、CTC様に開発していただいたツールにより、以下の対応が可能となった。

- YouTubeのURLを利用して映像をブラウザに取り込む機能
- 手話映像に対する正解ラベルの追加（時系列単位）
- 手話映像に対する骨格情報データの生成およびダウンロード

(2)話者適応モデル構築/オフライン手話認識の検証

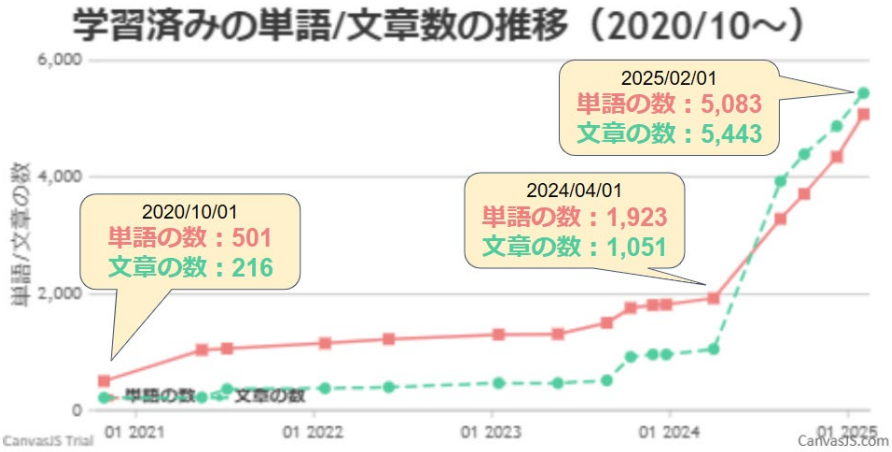
話者適応モデルの構築において、fairseqのファインチューニング機能を使用して少ない

データで高い認識精度を目指す。ファインチューニングとは、学習済みの汎用的なモデルに対して新たなデータを使って追加学習を行うことである。汎用的なモデルの性能向上はSureTalkの認識精度向上のために必要であることから、令和5年度に開発したfairseqによる学習モデル（今年度の７月にSureTalkアプリに実装済み）をベースに性能向上を図るため、下記の施策を行っている。

具体的な施策は以下の２点である。

① 汎用的な学習モデルの更新頻度について、昨年度までは１～２カ月に１度の更新頻度だったのがfairseq導入やデータフローの改良により**１週間に１回の更新が可能となり、更新の自動化も可能**とした。

② 汎化性をさらに高めた学習モデルを作るには多くの個人の手話表現のデータ収集が必要であるが、日本語の単語や文章の構成に依存しないDBの構築を行うことで、効率的なデータ収集を行うことができた。日本語の単語や文章の構成に依存しないDBとは、令和５年度の技術開発の一つである「**単語列情報を管理しないDBの検討と実装**」で実装したものである。これまで収集した単語/文章の推移は以下のグラフの通りである。



１カ月当たりの増加率は以下の通りで、データ収集の効率化実現により2024年夏ごろより**収集スピードが約３倍近く向上した。**

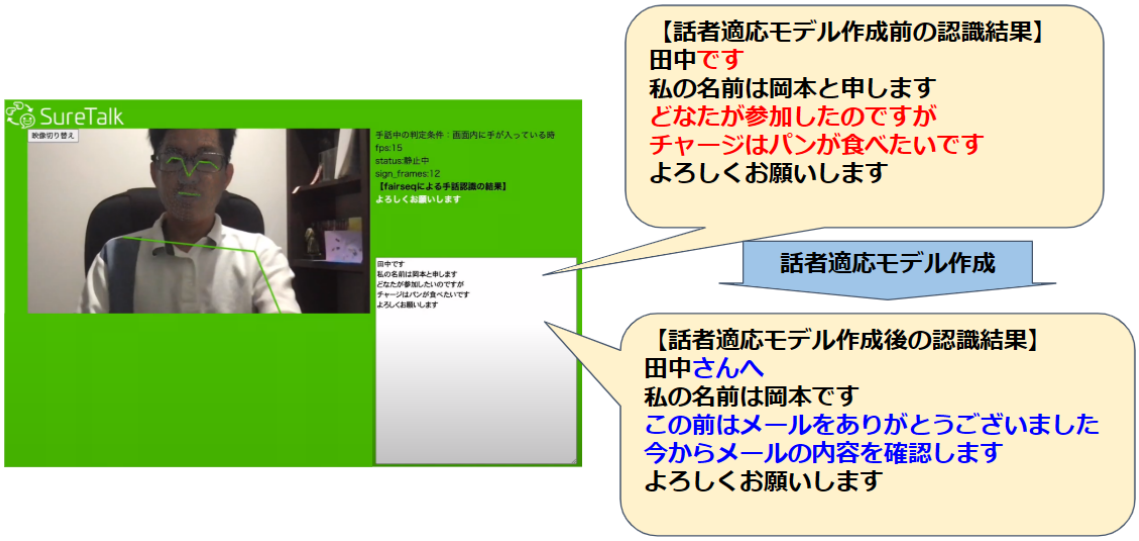
2020/10/01から2024/03/31まで（42カ月間）の１か月あたりの増加率は約3.31%

2024/04/01から2025/02/01まで（10カ月間）の１か月あたりの増加率は約9.63%

３－２．研究内容の成果

上記施策を行い汎用的な学習モデルの性能向上をしつつ、ファインチューニング機能を利用して、個人手話データに対して認識精度が上がるか検証を行った。今回は簡易的に稼働させることができるブラウザベースとし、話者適応モデル用の手話認識エンジンをPC上でサーバーとして立ち上げる仕様とすることで、端末側で処理を完結できる構成の構築を行った。また、オフライン手話認識も検討しているため、推論だけではなく、個人データの学習機能も入れている。オフライン手話認識に関しては、災害時におけるきこえない人とのコミュニ

ケーションの実現に有効な手段である。



未学習の個人手話データに対する認識結果はほとんど誤りであった。誤認識の５文章のデータに対してファインチューニングを用いて一般PC上で学習を行い、その場で作成された学習モデルを利用して検証したところ、認識結果が大きく改善されていることを確認した。ただし、機械学習に利用できるGPUが搭載されていない分、学習する手話データの増加に伴い学習時間がほぼ累乗的に増えてしまう問題がある。そのため、**学習処理も認識処理も効率化させた次世代手話認識エンジンを開発する**というアプローチを行うこととした。GPUを利用することで今の仕組みのままで話者適応モデルの学習時間が短縮できるが、 GPUは汎用性が低くオフライン向け手話認識システムにおいては実現が難しい。そのため、「手話モデル」と「言語モデル」に分割してそれぞれのAIモデルを作成することにより学習と認識の処理高速化を図る。まずは簡易的な指文字データを利用して検証した。指文字のみの学習モデルの結果は以下の通り。

・評価スコア・・・　BLEUスコア：86.2、WER：6.48＝認識率93.52%

次に両手の手話を利用した学習モデルの結果は以下の通り。

・評価スコア・・・　BLEUスコア：5.3

３－３．　研究成果の考察と令和7年度計画の方針策定について

CTC様に開発していただいたデータ収集ツールにおいて、**地方自治体などが公開している手話映像データが本ツールを介して汎用的な学習モデルに反映できるようになった**。そのため、令和７年度以降ではより多くの地方手話を簡単に取り込むことができる見通しである。今後はYouTubeだけでなく、Zoom映像などの非公開データについても本ツールを活用して学習モデルに反映できるよう、さらなる改良を進める予定である。

オフライン手話認識の実現を見据えた次世代手話認識エンジンの検証結果においては、簡易的な指文字データを学習したアルゴリズムは学習も推論も十分な性能向上が見られたが、両手の手話については更なる改善が必要である。評価スコアが低下した原因は様々考えられるが、今後は以下の対応を検討する。

・データセットの整備とデータ当たりの収集人数重視のデータセット構築

例：特定の50文章に絞って、１文章あたり10人分のデータセットを作るなど

・手指情報や動きの入力データの正規化の見直し

例：数値的に異常データが入っている場合、除外するなど

令和７年度で上記の対応により改善を行った上で**SureTalkのオフライン版アプリの導入を目指す**。また、災害時によく利用する単語や文章に絞ったデータ収集についても、データ収集システムの改良と並行してデータ収集していく予定である。