事業成果報告書

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
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「FY25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．読唇および手話認識方式に関する技術開発

２－１．研究開発の実施内容

（１） 日本手話に対応するためのNMM（非手指情報）データセット整備

日本手話では、手指の形や動きに加えて、眉・目・顎・視線など顔全体の動き（Non Manual Marker：NMM）が意味の識別に関与する。表１に例を示す。

表 1　NMMの例

意味	NMM
Yes-No 疑問文	眉上げ、目の見開き、頭を前に突き出す
Wh疑問文	目の細め又は見開き、細かい横の首振り
肯定文	発言の後のうなずき

今年度は、NMMを学習・評価に利用できる形で整理したデータセットを構築するための前提整備として、アノテーション作業を成立させるための手順設計と、必要機能の要件整理を行った。

（２） 手話認識エンジンの改良（映像ベース＋特徴点ベースの統合）

今年度も、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。前年度は特徴点ベースの End-to-End モデルとして、

【パターン 1】全身骨格情報を一つの入力とするモデル

【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

の２方式に取り組んだ。しかし、特徴点ベースでは、表情の微細な変化といった情報を十分に扱えないことが精度低下の一因になっていると考えた。そこでこの問題に対処するため、映像をそのまま用いて表情の微細な変化を直接捉えるアプローチを採用し、精度向上を図る方針で開発を進めた。

２－２　研究開発の成果

(1)日本手話に対応するためのNMM（非手指情報）データセット整備

公開手話映像（YouTube等）を対象にアノテーションを試みた結果、NMM付与には日本手話の専門知識が必要で、非専門者による作業は困難であると判明した。そこで、ろう者などの手話知識者がGUI操作のみで扱えるアノテーションツールを試作した。動画区間の指定、NMMカテゴリ選択、結果保存まで一連で実施でき、日本語文を分割してグロス登録し、文種（肯定・疑問・否定）や眉・目・顎などのNMMも付与可能である。結果は再利用可能な形式で保存・読み込みでき、NMM付きデータを継続的に蓄積・整理する基盤を整備した。

(2) 手話認識エンジンの改良（映像ベース＋特徴点ベースの統合）

映像情報を直接活用する方式として、以下2パターンの映像特徴（I3D：動画の時間的な変化も含めて特徴を捉える深層学習モデル）を用いたEnd-to-Endモデルアーキテクチャを開発した。

【パターン1】　手話映像のみを入力とするモデルアーキテクチャ

手話映像（RGB）からI3D（Inflated 3D ConvNet：連続するフレームの動きも含めて特徴を抽出するモデル）により時空間特徴を抽出し、その特徴系列を単一の入力データとして扱い、映像から表情の微細な変化や口形変化など、特徴点ベースでは捉えにくい情報を解析するような特徴抽出層とfairseqの言語モデルを組み合わせて、テキストを直接生成するEnd-to-Endモデルアーキテクチャ

【パターン2】　特徴点ベース＋映像を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と、手話映像から抽出した時空間特徴を別々の入力として取り扱い、特徴点系列からは手指動作や身体動作の情報を、映像特徴系列からは表情や見えの揺らぎ（ブレ・遮蔽等）をそれぞれ捉え、両者を統合した出力を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。

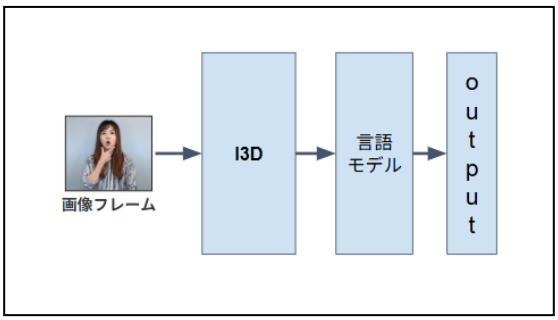

図 1 パターン１

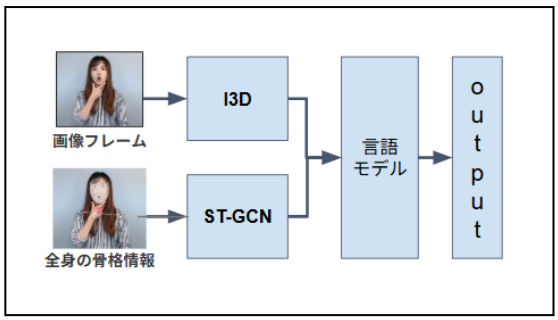

図 2 パターン２

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。

【テストデータに対するWER評価結果】
・pose（骨格情報）のみ：WER 62.12%
・I3D（映像時空間特徴）のみ：WER 16.67%
・骨格情報＋I3D（両方）：WER 10.61%

骨格情報のみでは認識誤り率が高く、手話認識に必要な情報を十分に捉えられないことが明らかとなった。一方、映像から抽出した時空間特徴（I3D）を用いることで、表情や口形、微細な動作など骨格では表現しきれない情報を捉え、認識精度が大きく向上した。さらに、骨格情報と映像特徴を統合した場合に最も低いWER（10.61%）を達成し、両者が相補的に機能することが確認された。すなわち、構造的情報（骨格）と外観的・時空間的情報（映像特徴）の両方の活用が日本手話認識に重要であることが示された。本結果は、骨格情報単独では不十分であり、映像特徴の併用が有効であることを定量的に示している。

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) 動画ラベリングツール開発

学習データの拡充にあたり、YouTube等の公開映像だけでなく、Zoom録画などの非公開映像も含めて効率よく学習モデルへ反映できるデータ収集基盤の整備が求められる。そこで今年度は、データ収集と並行して収集システムの改良を進め、公開／非公開データを同一の手順で取り扱えるようシステム強化を行った。

(2) オフライン手話認識の開発とスタンドアローン（オフライン）型アプリの試作

オフライン手話認識の開発とスタンドアローン型アプリの試作として、通信インフラが使えない災害時でも端末上で処理を完結する仕組みの実現を目指した。前年度は次世代手話認識エンジンを構築し、指文字では高性能（BLEU 86.2）を確認した一方、両手手話では性能不足（BLEU 5.3）が課題となった。今年度は、映像特徴と特徴点の統合によるエンジン改良に加え、モデル常駐化や前処理簡素化など推論の効率化を実施し、オフライン動作の実現性を検証した。さらに、推論エンジンを組み込んだスタンドアローン型アプリを試作し、動画入力から結果提示までをローカル環境で完結できることを確認した。

３－２．研究内容の成果

(1) 動画ラベリングツール開発

YouTube映像に加え、Zoom録画等の非公開動画ファイルも取り込み可能な動画ラベリングツールを開発した。本ツールは、動画から骨格データ（人物の関節位置）を取得したうえで発話内容等のラベル付けを行えるほか、複数人が同時に映る動画にも対応し、「誰が」「いつ」「何を発話したか」を人物単位・時間帯単位で登録できる。さらに、登録済みの動画・骨格データ・ラベルを一覧管理し、対象動画を選択して再編集できる運用フローを実装した。これにより、動画データを一貫して取り扱える基盤（アップロード→骨格取得→ラベル付け→管理）を構築し、対話形式など複数人動画においても人物別に骨格データを抽出・管理で

きるようになった。結果として、学習・分析に利用可能な構造化データ（骨格＋時間情報付きラベル）を作成できる環境を整備した。ツールの画面を図3に示す。

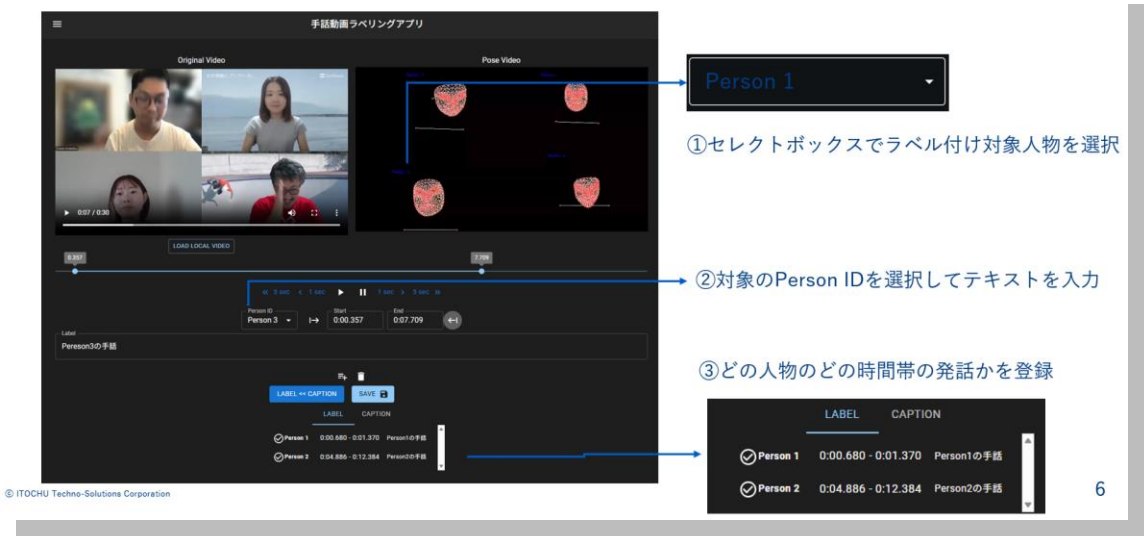

図 3 動画ラベリングツールの画面

(2) オフライン手話認識の開発とスタンドアローン（オフライン）型アプリの試作

災害時の利用を想定した検証のため、推論エンジンを組み込んだスタンドアローン（オフライン）型アプリを試作し評価を実施した。FastAPIによる推論サーバをローカルPC上で常駐させ（モデル常駐）、Webアプリ（ブラウザ）→Node（動画WebM→MP4変換）→FastAPI（推論）→結果返却、という一連の導線を構築し、ローカル環境で動画入力から推論結果提示までが完結することを確認した。検証端末としてMacBook Pro 14インチ M4 Pro（12CPU/16GPU、メモリ24GB、SSD 512GB、macOS、2024年10月モデル）を用い、オフライン環境での動作を確認した。

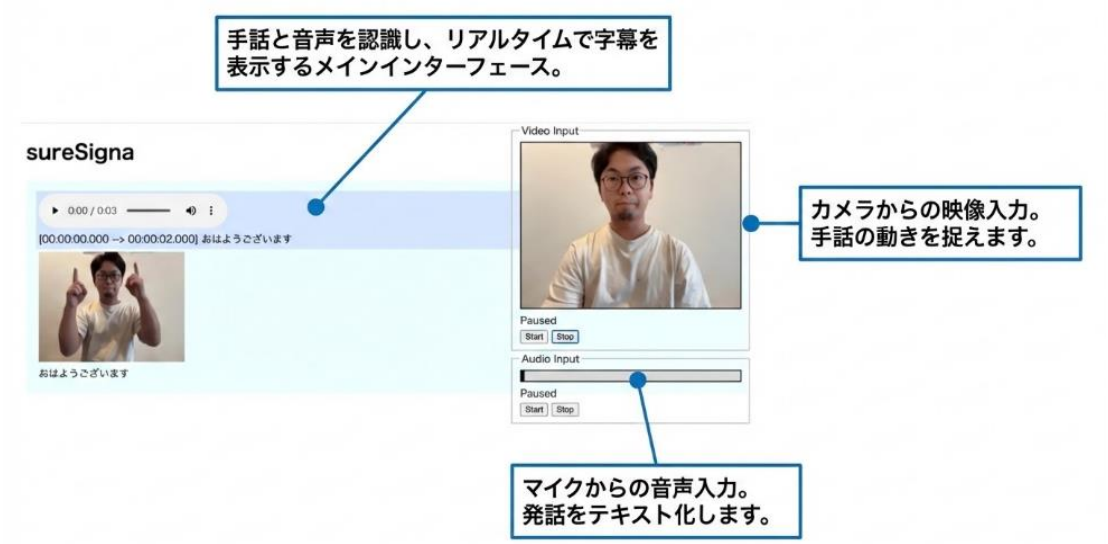

図 4 スタンドアローン（オフライン）型アプリ

災害時など通信インフラが利用できない環境でも使えるよう、手話認識を端末（ローカル環境）だけで完結させる構成を整備した。オフライン動作では計算量の削減が重要なため、動画（RGB）をそのまま処理するのではなく、軽量な特徴量へ変換してから認識する設計を採用した。具体的には、映像はI3Dにより16フレーム単位で特徴抽出し、骨格はMediaPipeにより手や上半身の関節点を取得する「映像＋骨格」の二系統構成とした。これにより計算負荷を抑えつつ効率的な学習・推論を実現した。さらに、両特徴は加算により統合し、追加計算を最小限に抑えた。加えて、モデル自体も層数・サイズを抑え、オフライン推論に適した軽量設計とした。

最後に、オンライン環境およびオフライン環境における推論応答時間の比較評価を実施した。平均応答時間は以下の通りである。

【オンライン環境】
平均応答時間：2.1秒
【オフライン環境】
・pose（骨格）のみ：2.75秒
・I3D（映像特徴）のみ：4.86秒
・骨格＋I3D（両方）：5.89秒

検証モデルを評価した結果、骨格情報のみを用いる構成が最も軽量であり、オフライン環境においても比較的短時間での応答が可能であることが分かった。一方で、I3Dを用いる構成では動画から時空間特徴を抽出する処理の計算負荷が大きく、特に動画特徴抽出処理が最も重いボトルネックであることが確認された。また、認識精度評価では「骨格＋I3D」の統合モデルが最も高精度であった一方、推論時間は最も長くなる結果となり、精度と処理速度の間に明確

なトレードオフ関係が存在することが分かった。
　この課題に対する今後の軽量化方策として、出力部分（文章を生成する部分）の仕組みを見直すことが有効であると考えられる。
具体的には、AIが認識結果の「候補文」を複数すばやく作り、その中から文章の自然さを判定する小さなモデルが最終結果を1つ選ぶ二段階方式が有望である。構成イメージは以下の通りである。

---

video → 特徴抽出（必要情報を取り出す） → 候補文を複数生成 → 小さなモデルが最終文を選択

本方式により、精度を維持しながら計算量を抑え、オフライン環境でのさらなる高速化と実用性向上が期待される。