手話認識機能を搭載したオンライン手話辞書システムの開発と公開

研究代表者　　木村　勉　　豊田工業高等専門学校　情報工学科　教授
共同研究者　　神田　和幸　　国立民族学博物館　人類基礎研究部　外来研究員

1 背景・目的

豊田高専　情報工学科　木村研究室ではこれまで手話から日本語を検索することのできる「日本手話・日本語辞書システム」を開発してきた．開発の大きな理由として，日本語から手話を調べるという日本語―手話辞書が多く存在するのに対し，手話からそれに対応する日本語を引き当てる辞書（日本手話・日本語辞書）の数が少ないためである．図 1 に開発した日本手話・日本語辞書システムの GUI を示す．

図 1　日本手話・日本語辞書システムの GUI

この日本手話・日本語辞書システムは，手話表現において 6 つに分類した手話音素（手型，位置，動きなど）から調べたい手話に当てはまるものを選んでいき，目的の手話を検索するという仕組みである．しかし調べるためには手話音素について理解しておく必要があり，手話の初心者にとっては使いにくいシステムであった．また先天性聴覚障がい者の場合，多くは書記日本語に明るくない．そのため知らない手話を調べようとしても，GUI に書かれている選択肢の日本語の理解が難しいという問題もある．

これらの問題を解決する方法として，手話認識システムを導入する．利用者がカメラの前で手話単語を表現すれば，システムが手話を認識し，候補の一覧を画面に表示する．利用者は，その一覧の中から調べたい手話を探すことができるようにする．

2 オンライン手話辞書システムの概要

オンライン手話辞書システム（以下辞書システム）の概要図を図 2 に示す．辞書システムはオンライン上に構築し，インターネットを通じて使用することを前提としており，GUI はブラウザ上で実装する．辞書システムとしての動作は次のような流れとなる．

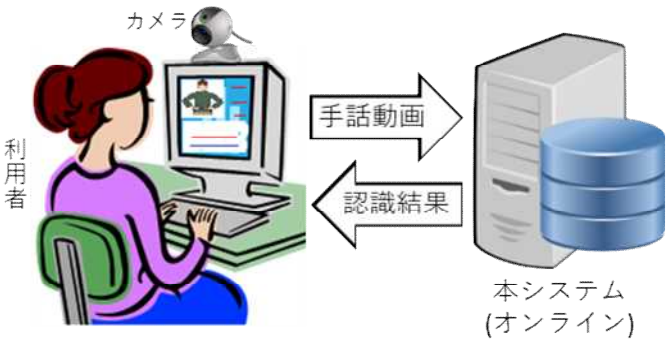

図 2 システム概要図

1) 利用者はUSBカメラなどを用いて手話を録画し，辞書システム（サーバ）に送信する．2) 手話認識エンジンを使用して動画を解析し，一致率の高い順に単語の一覧を利用者に掲示する．3) 利用者は一覧に提示された単語名をクリックすると動画が再生され，調べたい単語であるか確認する．

3 手話単語認識エンジンの開発

3-1 認識方法

本研究で行っている手話認識は，姿勢推定を行い，関節などのキーポイントを特徴量として学習させる．

これにより，背景の変化や衣服による影響を少なくすることができるという利点がある．本研究では，姿勢推定アプリとしてGoogleのMediaPipeを用いる．このアプリから関節点などの座標を取得し，これを1フレームごとにJSON形式のファイルに出力し，さらにこれを動画ごとに1つのCSV形式の動作データにまとめる．これらを学習データとして機械学習を行って，学習済みモデルを取得する．

3-2 作成方法

本研究では，学習モデル Conformer を採用する．これは自然言語処理で飛躍的な性能をもっているTransformerとCNNと組み合わせた学習モデルである．Transformerとは，2017年に発表されたエンコーダ・デコーダ型の深層学習モデルである．その特徴として，RNNの再帰やCNNの畳み込みなど採用しておらず，Attentionのみを用いている．そうすることで，並列計算が可能になり，訓練の時間が短縮される．その上，翻訳では精度が高いモデルの多くがTransformerをベースとしている．さらに，Transformerは自然言語処理だけでなく画像処理や音声認識などにも応用されている．その音声認識に応用するために考案されたモデルがConformerである．このモデルは，音声認識において大きな成果を出している．

対象とする手話単語は手話技能検定試験6級の101語および5級148語のうちの78語とした.手話動画は1語につき100個のデータを用意し，MediaPipeを用いて手の21か所，肘と手首2か所，および左右分として計46か所の3次元座標を取得し，それをCSV形式でまとめる．つまり，特徴量を138個とり学習させる．各単語の学習データのうち80%を訓練データ，20%をテストデータとする．また，前処理として取得した座標を，肩幅の倍を画面全体の大きさとし，さらに人物が中央に来るように正規化する．これによって辞書システムにおいて，人の位置やカメラとの距離によって，画面上での相対位置や大きさが異なることによる認識率が下がることを防ぐ．このデータを用いて学習を行う．

3-3 学習結果

Conformerを用いて学習させた結果，単語ごとの認識率にばらつきがあるものの，約97.4%の認識率が得られた．今回の学習で，テストデータの認識において20個中3つ以上誤認識した単語を表1に示す．

表1　誤認識した主な単語

単語	主に誤認識された単語
暑い	南
南	暑い
一昨日	昨日
コーヒー1	コーヒー2
金曜日	OK
辛い	甘い
売る	買う
数	いくつ？

まず表1より，「暑い」と「南」の2つの認識がうまくいっていないことがわかる．これは，同じ手話動作で異なる意味を表す「同形異義語」であるからである．図3に示した手話は，利き手を顔の近くで仰ぐ動作であるが，この手話は「暑い」,「南」,「夏」,「うちわ」の4つの意味を持ち，それを区別するためには文脈か，口の動きでそれぞれの言葉を表現する必要がある．

図3 「暑い」手話

「昨日」と「一昨日」の手話を図4に示す．この2つの手話はどちらも手を前から後ろに動かすものだが，指を一本立てるか，二本立てるかという違いがある．

「金曜日」と「OK」の手話を図5に示す．これはどちらも人差し指と親指で輪をつくるという同じ手型であるが,「金曜日」は手を振る動作が入り,「OK」は手型を提示するだけである．

また,「コーヒー1」と「コーヒー2」の手話を図6に示す．意味は同じであるが，手話表現が異なるものである．この手話は，利き手（図6では右手）の動作（スプーンでかき混ぜる動作）が同じで，非利き手の手型が異なる（カップの形状とカップを持つ仕草）だけである．

「辛い」と「甘い」の手話を図 7 に示す．この手話は，手型は異なるが，動きは同じ（口の周りで手を回転させる）である．

他にも「数」と「いくつ」，「売る」と「買う」といった手話を誤認識している．これらも手型もしくは動きがよく似ている手話である．

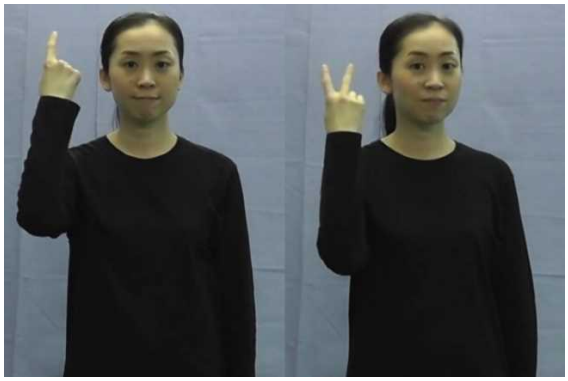

図 4「昨日」（左）の手話と「一昨日」（右）

図 5「金曜日」（左）と「OK」（右）の手話

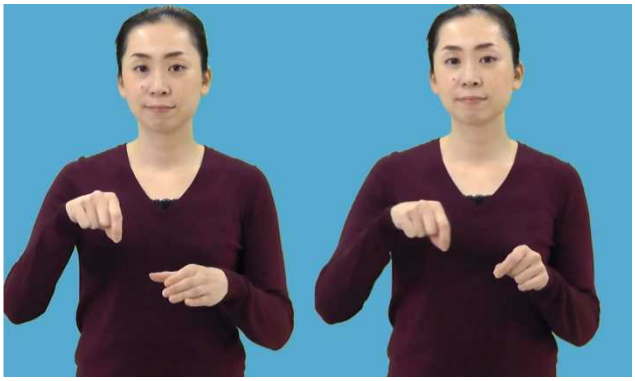

図 6　2 種類の「コーヒー」の手話

図 7「辛い」（左）の手話と「甘い」（右）の手話

4 オンライン手話辞書システムの開発

4-1 実装

2 章で述べた辞書システムを開発する．今回は AWS（Amazon Web Service）を用いて実装した．

ここではシステムの内部仕様についての概略を述べる．図 8 に各要素の通信概要図を，図 9 に内部仕様の概要図を示す．

図 8 各要素の通信

今回開発するシステムはオンライン上で動作する Web アプリケーションを想定している．そのため大きく分けて，図 8 のようにクライアントとサーバーの 2 つのサブシステムとそれらを繋ぐ通信から成り立つ．

クライアントでは，手話動画の撮影とサーバーへの送信を行い，サーバーからの認識結果をユーザーに提示する．

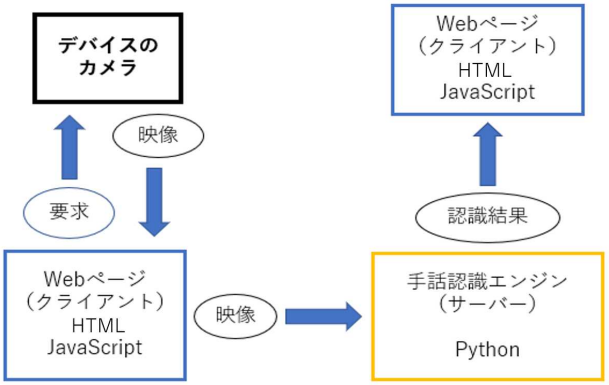

図 9 内部仕様の概要図

サーバーでは，クライアントから送信されてきた手話動画認識するが，最初に 3-2 節で述べた処理と同様の方法でデータを作成する．つまり姿勢推定アプリ MediaPipe を使ってキーポイントを推定し，前処理として人物の正規化を行う．これを 1 つの CSV 形式のデータとする．手話認識エンジンにこのデータを入力し認識を行い，その結果をクライアントに返す．

クライアント側は，図 9 のように Web ページとして HTML と JavaScript を用いて実装する．

サーバーは Amazon EC2 サービスから提供される Linux 搭載のサーバーに Python API である Flask を用いて実装する．サーバーが処理を行うのは，図 9 の動画送信から認識結果の表示ページへ遷移する部分である．クライアントから送信された手話動画をサーバーに備えた手話認識エンジンによって認識する．手話認識には 3 章で述べた手話認識エンジンを用いる．手話認識エンジンによって一致率の高い 10 個の単語を見本の手話動画の URL とともにクライアント側に返す．クライアント側では 10 個の単語を一覧として表示する．表示した単語をクリックすると見本動画を表示する．

4-2 実行結果

AWS 上に実装し，その評価を行う．このシステムでは，基本的に調べたい手話を録画して，送信するまでの操作を一人で行う．そのため，開始ボタンを押してから録画を始めるまでの時間および録画時間をあらかじめ設定しておく．（図 10）

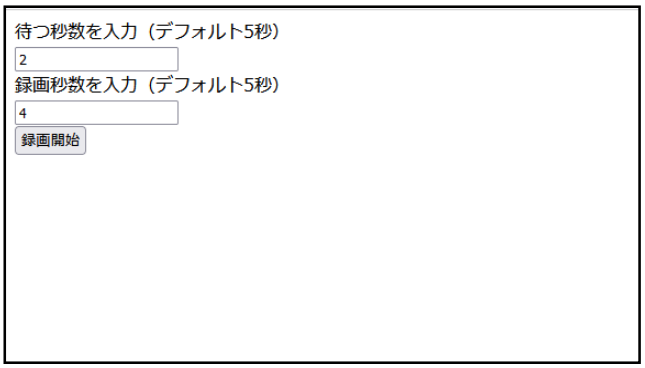

図 10 設定画面

図 11 に実行画面を示す．録画が開始されたら調べたい手話を表現する．録画が終了すると動画がサーバーに送信され，認識される．認識結果を図 12 に示す．ここで一致率の高い 10 個の単語が示される．図 11 では「涼しい」と表現したが，認識結果としては「すずしい」，「あお」，「あたたかい」，「わたし」，「いえ」，「おなじ」，「くろ」，「あう」，「よる」が提示されている．それぞれの単語をクリックすることで，見本の動画が再生される（図 12 下部）．これにより調べたい手話を確認することができる．

このシステムを利用して検索した時間は，3 秒の手話動画を送信してから検索結果が出るまでに約 7.44 秒であった．表 4 に示すように他の秒数の動画でも認識結果が出るまでの時間は，手話動画時間のほぼ倍程度であった．

図 11 手話動画入力画面

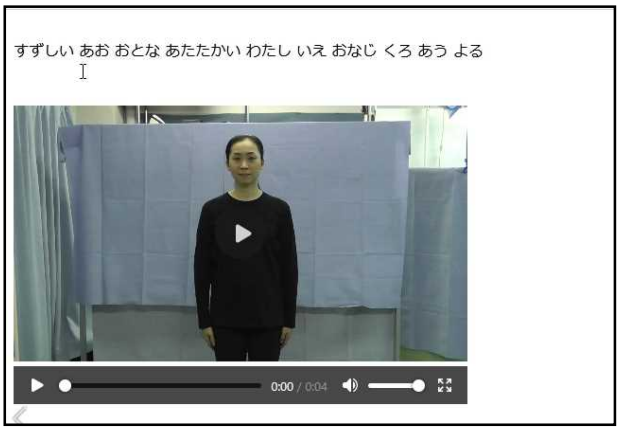

図 12 認識結果表示の画面

表 1：動画の秒数と認識結果が出るまでの時間

動画時間[s]	認識時間[s]
3	7.44
4	9.74
5	10.32

5 手話文の認識

5-1 研究概要

現在，手話単語だけでなく，手話文の認識も行うことができるように並行して研究を進めている．手話文は単語と単語の組み合わせで構成されており，それぞれの単語動作を連続で表現することで手話文として表される．このとき，手話単語の動作には含まれなかった「遷移」と呼ばれる，一つの単語の動作を終えた後に次の単語の動作を行う姿勢に移行するための補間の動作が入る．図 13 は「わたしの名前は佐藤です」と表現する手話文あるが，これは「私」,「名前」,「佐藤」,「です」の 4 つの単語から構成される（助詞は手話では用いない）．それぞれの単語間に遷移が入る．この遷移は意味を持たない動作であり，前後の単語間で最小の動きとなるような動作となる．この遷移の存在によって，手話文の認識は手話単語の認識と比べて困難である．

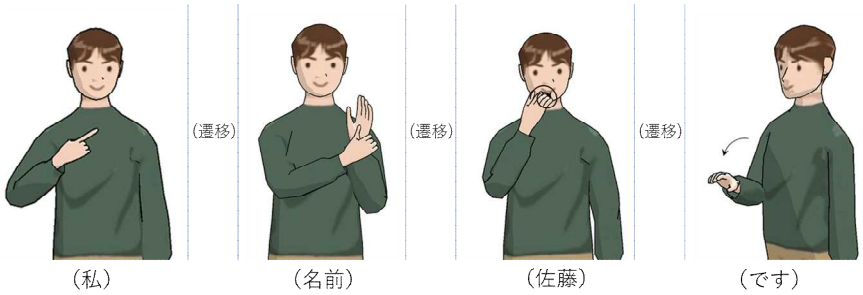

図 13 手話動画入力画面

そこで，音声認識の分野で使用される Connectionist Temporal Classification（以下 CTC）や Conformer を応用し，遷移動作にも対応することができる手話認識モデルを実装し，手話文を学習させて学習済みモデルを作成する．

本研究では，手話文中からの単語の認識を目的とする．CTC と Conformer による音声認識の研究を参考に，本研究でも手話文を学習データとして機械学習を行う．そしてこの両者での比較を行う．

5-2 研究方法

CTC は音声認識の際に，個人差による音素の長さや音素間に発生する意味を持たない要素であるブランクといった曖昧な部分を浮き彫りにしてから縮約することで,認識したデータの情報を簡潔にする特性を持つ．手話認識において CTC を利用することで， CTC を活用することで，音声で発生するブランクと同じように手話文中に発生する曖昧な動作である遷移の縮約が可能ならば，遷移に影響されずに手話単語の認識を行うことができると考え，実装を行った．

一方 Conformer はデータに畳み込み処理を行うことでデータを細分化し，局所的な情報を抽出できるようにしている．Conformer は自然言語処理，特に音声認識において従来の性能を大幅に上回る結果を出している．手話認識の過程は音声認識と似通っており，これまでの研究で，音声認識で使用される CTC 手法が手話認識においても動作することが確認できている．そのため，本研究では Transformer を使用したモデルの中でも，音声認識に適した Conformer なら手話認識への応用も可能であると考えて Conformer 手法の実装を行った．

本研究ではで示した CTC と Conformer の 2 つの手法で実験を行う．手話技能検定試験 5,6 級で指定されている単語のうち，99 語を用いて 208 種類の例文を作成して撮影を行い，全体で 4,735 文の手話文動画を作成した．この手話文動画に対して姿勢推定を行い，身体の関節点の情報を抽出し，手話文のデータセットを作成した．実験の際にはこのデータセットを学習データ 90%，テストデータ 10%に分割して実験を行った．本実験では学習済みモデルを評価する際に，誤認識の割合を示すエラー率を使用する．エラー率の算出には WER を使用している．

5-3 結果と考察

CTC 手法による機械学習での学習曲線は図 14 のようになった．学習データの損失値が減少するとテストデータの損失値が減少しているため，学習は正常に行われているといえる．作成した学習モデルでテストデータを評価したところ，エラー率（WER）は約 33%となった．

Conformer 手法による機械学習での学習曲線は図 15 のようになった．学習曲線のグラフより，テストデー

タの損失値は学習が始まってしばらくしたら増加しているため，テストデータに対応できるような学習をモデルが行えていないことになる．作成したモデルでテストデータを評価したところ，エラー率（WER）は約72%となった．以上の結果より，本実験においてはConformerを用いた手法よりもCTCを用いた手法のほうが優れているという結果となった．

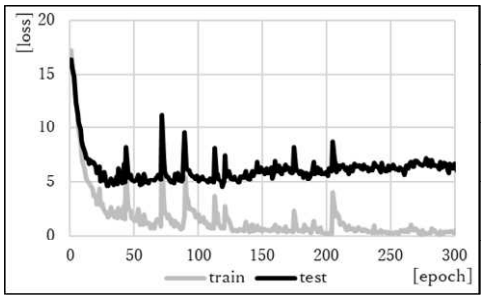

図14 CTC手法による機械学習の学習曲線

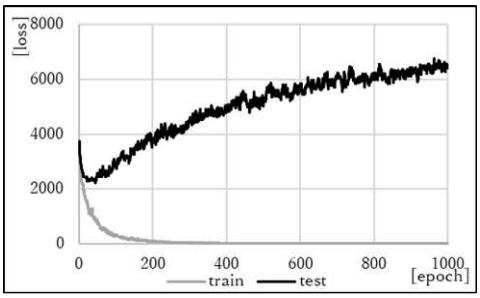

図15 Conformer手法による機械学習の学習曲線

CTC 手法における認識結果の傾向として，文から認識した単語の誤認識よりも文で使用されている十分な数の単語を認識できていないことによる認識率の低下が多かった．この理由として，フレームごとの解析を行うCTC手法ではデータ全体の時系列を考慮して解析することが難しく，その結果，必要数の単語を認識することが困難なのではないかと考えた．

Conformer 手法における認識結果の傾向として，学習データの内容をそのまま出力することが多かった．本研究ではデータセットを学習データとテストデータに分割する際に，同じ内容の手話文が入らないようにしている．しかし，Conformer 手法での認識のほとんどが学習データをそのまま出力しているため，誤認識が多発していた．これは過学習のときによく見られる現象である．これは，Conformer が学習データに対して各単語の依存関係に特化した学習を行う際に，単語の順序に関する要素を学習しすぎてしまい，順序をそのまま記憶してしまっていることが原因だと考えている．

6 今後の課題

ここでは現在進行中の研究を含め，今後の課題について述べる．

6-1 手話学習教材の提供

オンライン手話辞書システムの応用として，手話学習サイトの構築が挙げられる．手話に関する問題を提示し，それを利用者が手話で解答するというものである．解答は本システムの手話認識エンジンを使って認識し，正解かどうかを判断する．手話問題作成には，手話技能検定協会の協力を得て，準備を進めている．すでにコンテンツは用意できており，今後は実装を進める．

6-2 指文字認識

これまでの研究で手話単語の認識については成果が出ている．しかし，指文字への対応ができていない．ただ，これまでの研究において，手話文から単語を抽出することは可能となってきている．この研究成果を応用し，指文字認識を行う．単語認識との違い，指文字認識が困難である理由は以下の通りである．

1）手型の推定や識別

カメラのシャッタースピードにより手がぶれてしまい，手型がはっきりしない動画が多いことに加えて，指のみで表現するため，動きが小さく識別が難しい．

2）奥行き方向への動きの推定

指文字には手を手前に引いて表現するものがある．（拗音（ゃゅょ）/促音（っ））これにより，縦と横の動きだけでなく手の奥行き情報も必要である．

3）指文字間の区切りの判別

指文字は50音に対応しているが，それらの連続した動作で単語を表現するため，1つの指文字から別の指文字に遷移する際の区切りの判別が困難である．

4）指文字独特の表現

指文字表現には「手型」のみのものと「動き」を伴うものの2種類がある．「の」や「り」，「ん」，そして濁音，半濁音などは動きを伴うため，遷移との切り分けが困難である．

これらの解決のための方法であるが．1）および 2）については，MediaPipe を利用することで改善を図ることができるとわかっている．以前の研究では， OpenPose を利用していたが，MediaPipe であれば，取得できる関節点の数が多く，より詳細に手型の情報を得ることができる．また，奥行き情報も含めた手型の推定が可能である．3）および 4）については，これまでの研究を参考に，CTC や Conformer を用いて遷移の除去を行う予定である．

6-3 同形異義語の認識

3-3 節で述べたように，同じ形で別の意味を持つ「同形異義語」の区別ができないという問題がある．現在の手話単語認識システムでは，手話翻訳に「手型」「位置」「腕の動き」という手指動作のみを情報として使用しているが，「同形異義語」の認識には「口の動き」「頷き」「眉上げ」といった非手指動作を考慮する必要がある．

それを解決するために非手指動作である「口の動き」から母音を抽出し，「同形異義語」の判別を行うことが必要となる．機械読唇と音声認識の手法に従って母音読唇ネットワークの構築を行い，これを應用することで解決を図る．

まず動画から唇の特徴点を検出するために，MediaPipe の FaceMesh を使用する．機械読唇の手法は動画を学習させるケースがほとんどである．一方，木村研究室で研究している手話辞書システムや手話翻訳システムでは，手話認識は身体の関節点などから行っている．同形異義語の認識でも同様に口の特徴点のみでの母音抽出を試みる.そして,時系列データを扱うために RNN(Recurrent Neural Network)の一種である GRU(Gated Recurrent Unit)を使用する．さらに，音声認識や手話文認識の様に読唇においてもどの音素でも無いブランクや音素と音素のつながりによる連音が発生する．これを解決するために音声認識分野において主要な手法である CTC か Conformer を応用する予定である．

6-4 手話文の認識

手話文の認識率を上げるためには，データセットの増加が挙げられる．本実験で使用している手話文データセットは種類，語彙数共に十分であるとはいえない．また，Conformer 手法の適応方法を改善する必要がある．現在，改善を試みたところ，90%を超える一致率を得ている．データ数が少ないため，検証は必要であるが，良好な結果が得られているので，今後の研究でさらに進める予定である．

【参考文献】

T. Kimura and K. Kanda, "Sign Language Recognition through Machine Learning by a New Linguistic Framework", Association for the Advancement of Assistive Technology in Europe 2019, Proceedings S144-S145, 2019.

磯谷光, 木村勉, 神田和幸, ” ディープ・ラーニングを用いた手話認識に関する研究”, 信学技報, vol1.120, no.419, WIT2020-38, pp.47-52, 2021 年.

木村 勉他, ”手話認識機能を備えた手話辞書システムの開発”, 信学技報, vol. 120, no. 419, WIT2020-39, pp. 53-58 2021 年.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang, ”Conformer:Convolution-augmented Transformer for Speech Recognition, ”, INTERSPEECH 2020, no. 10. 21437 , pp. 5036-5040, Shanghai, China, 2020.

上乃 聖 他, ” CTC による文字単位のモデルを併用した Attention による単語単位の End-to-End 音声認識”, ”情報処理学会研究報告”, vol.2018-SLP-120, no.16, pp.1-6, 2018.

高 山 夏 樹 , Gibran BENITEZ-GARCIA, 高 橋 裕 樹 , “Spatial-Temporal Graph Convolution-Transformer に基づく手話認識”, 精密工学会誌, 87 巻, 12 号, p.1028-1035, 2021.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
ディープ・ラーニングを用いた手話認識に関する研究 ～ CTC と Conformer の比較～	信学技報, vol. 121, no. 418, WIT2021-48, pp. 29-34	2022 年 3 月
Conformer を用いた手話単語認識に関する研究	手話コミュニケーション研究会論文集 2022 pp.1-8	2022 年 5 月
手話認識システムの小型デバイスへの実装と開発ツール群の作成	手話コミュニケーション研究会論文集 2022 pp.9-16	2022 年 5 月
手型推定による指文字認識の研究	手話コミュニケーション研究会論文集 2022 pp.17-24	2022 年 5 月
手話表現中における読話認識に関する研究	手話コミュニケーション研究会論文集 2022 pp.25-32	2022 年 5 月
オンライン手話辞書システムの開発	手話コミュニケーション研究会論文集 2022 pp.33-40	2022 年 5 月