

Deep Multimodal-based Finger Spelling Recognition for Thai Sign Language: A New Benchmark and Model Composition

Wuttichai VIJITKUNSAWAT
2120006

NGUYEN Le Minh Laboratory

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science

June 2024

Abstract

Video-based sign language recognition is vital for improving communication for the deaf and hard of hearing. However, due to a lack of resources, creating and maintaining the quality of Thai sign language video datasets is challenging. To address this issue, we assess multiple models with a novel dataset of 90 signs, covering the full letters of alphabets, vowels, intonation marks, and numbers, as demonstrated by 43 signers. We investigate seven deep learning models with three distinct modalities for our analysis: video-only methods (including RGB-sequencing-based CNN-LSTM and VGG-LSTM), human body joint coordinate sequences (processed by LSTM, Bi-LSTM, GRU, and Transformer models), and skeleton analysis (using TGCN with graph-structured skeleton representation). A thorough assessment of these models is conducted across seven circumstances, encompassing single-hand postures, single-hand motions with one, two, and three strokes, and two-hand postures with static and dynamic point-on-hand interactions. The research highlights that the TGCN model is the optimal lightweight model in all scenarios. In single-hand pose cases, a combination of the Transformer and TGCN models of two modalities delivers outstanding performance, excelling in four particular conditions: single-hand poses, single-hand poses requiring one, two, and three strokes. In contrast, two-hand poses with static or dynamic point-on-hand interactions present substantial challenges, as the data from joint coordinates is inadequate due to hand obstructions stemming from insufficient coordinate sequence data and the lack of a detailed skeletal graph structure. The study recommends integrating RGB-sequencing with visual modality to enhance the accuracy of two-handed sign language gestures. Moreover, experimental results on our dataset show that our method outperforms previous state-of-the-art methods significantly in five out of seven conditional hand pose experiments, especially two-hand poses.

Keywords: Thai Finger Spelling, Sign Language Recognition, Deep Learning, Multimodal Learning, Benchmark Dataset

Research Content

Sign language plays a critical role as a communication medium, predominantly within the communities of individuals who are deaf or hard of hearing. This form of communication comprises a comprehensive range of hand signs, facial expressions, and body movements, all of which work together to convey meaning independently of spoken languages. Over the years, sign language has gained immense recognition as a valid linguistic system, a development largely attributed to extensive research and strong advocacy. Sign languages are diverse, with various forms developed across different regions and communities. American Sign Language (ASL), British Sign Language (BSL), Thai Sign Language (TSL) and French Sign Language (LSF) are among the many distinct sign languages in use today, each with its unique grammar, vocabulary, and syntax. These languages are not universally interchangeable, even between countries that share the same spoken language, highlighting their unique evolution and cultural significance.

This research primarily focuses on improving communication and accessibility for the deaf and hard-of-hearing communities in Thailand. However, developing and maintaining high-quality sign language datasets from video inputs is challenging due to the absence of a standard Thai finger spelling video dataset. To address this challenge, our research aims to accumulate a comprehensive dataset featuring 90 primary elements of Thai Finger Spelling, which includes alphabets, vowels, intonation marks, and numbers, represented by 43 signers from diverse backgrounds, genders, and appearances. We employ seven deep learning-based architectures across three modalities: RGB-sequencing-based CNN-LSTM and VGG-LSTM for the video-only modality; a sequence of coordinates of human body joints using LSTM, Bi-LSTM, GRU, and Transformer models; and the structure of human joints modality using TGCN, as well as various combinations of these modalities. Furthermore, we continue to compare our method with numerous baseline and state-of-the-art models to assess the performance benchmark of our dataset and approach.

Table: In-Sample and Out-of-Sample performance benchmarks for total two-hand poses.

No.	Model	App	Pose	In-Sample Testing				Out-of-Sample Testing			
				Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
1	I3D	✓		43.52	46.34	33.13	38.64	9.41	10.5	3.65	5.42
2	Fusion-3	✓		72.94	75.49	68.76	71.97	34.12	40.68	29.23	34.02
3	MEMP	✓		85.88	87.73	84.22	85.94	32.55	38.05	24.45	29.77
4	DeepSign-CNN	✓		83.52	85.84	81.47	83.6	45.88	60.67	40.18	48.34
5	Pose-GRU		✓	92.94	93.62	92.03	92.82	47.45	53.62	43.38	47.96
6	Pose-TGCN		✓	80.58	81.06	79.01	80.02	22.74	24.04	17.02	19.93
7	SPOTER		✓	61.17	67.49	51	58.1	32.54	35.55	27.81	31.21
8	Bi-RNN		✓	85.29	86.46	83.09	84.74	32.94	36.68	28.41	32.02
9	FNN-LSTM		✓	79.41	84.16	73.78	78.63	29.8	34.12	22.67	27.24
10	ChatGPT4	✓		86	82.14	86.07	83.39	n/a	n/a	n/a	n/a
11	Ours (T+TG)	✓	✓	77.6	78.2	74.11	76.1	53.3	58.3	47.92	52.6

Where: App = Appearance representation, Pose = Pose-based representation, Acc = Accuracy, Pre = Precision, Rec = Recall and F1 = F1-score

The table, which is a part of our results, provides a comprehensive analysis of 11 notable models, focusing on a total two-hand pose with 17 letters. The table details total parameter usage and a variety of evaluation metrics such as accuracy, precision, recall and F1-score, all assessed both in-sample and out-of-sample testing.

Table shows the efficacy of various models in 17 letters of total two-hand poses recognition task, with the Pose-GRU model exhibiting superior in-sample performance, reflected by high accuracy 92.94%, precision 93.62%, recall 92.03%, and F1 score 92.82%. This suggests that the Pose-GRU effectively learns and captures the nuances of total two-hand poses within the training dataset. On the other hand, the "Our(C+G)" model, which is an integration of CNN-LSTM (C) model, which offer feature extraction both the visual and structural data, and GRU (G) model, contributing a robust mechanism for understanding temporal sequences and predicting future states based on both past and anticipated information, and utilizes both appearance and pose-based representations, outstrips its counterparts in out-of-sample performance with an accuracy of 53.3% and an F1-Score of 52.6%. In part of the out-of-sample of ChatGPT4, it cannot have the capability to interpret or generate pose estimations, making it incompatible with the specific tasks these metrics are intended to measure, or the model was not designed for this type of evaluation. As a result, these metrics are not applicable to ChatGPT4, and no data.

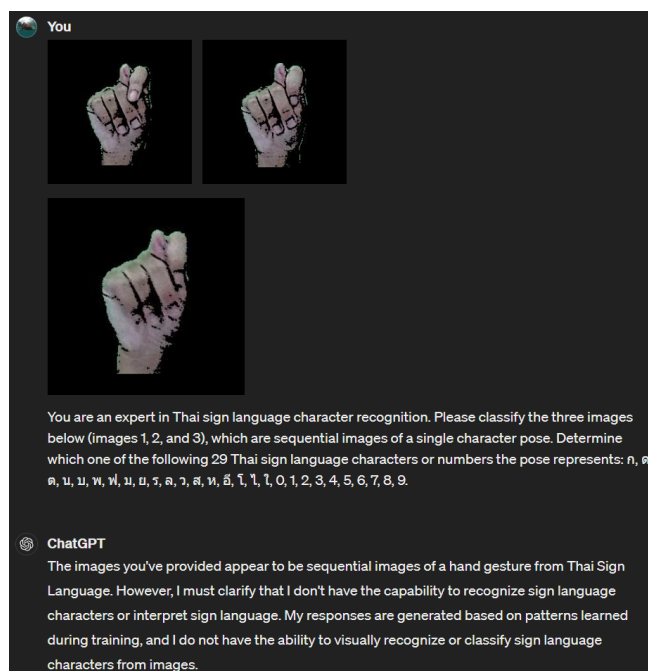


Figure 1. Example result of ChatGPT4 in out-of-sample testing

Our experimental results demonstrate that our combination modalities can help the one-hand and two-hand pose problems in case of insufficient data from the obscured hand's joints and images. Therefore, integrating of many modalities is suitable for scenarios. This method is also compatible with one-hand and two-hand poses.

Research Significance

The contribution of our work mainly lies in the following three aspects. Firstly, we have developed a comprehensive video database for Thai Finger Spelling (TFS) in sign language, featuring 10,467 videos of 90 unique letters demonstrated in different poses with one or both hands, contributed to by 43 diverse signers, appearances and backgrounds. Furthermore, our dataset comprehensively covers all aspects of TFS and nearly achieves complete balance, with 90% of the dataset obtained through direct video recording and 10% from internet sources. In this research, our video dataset stands as the largest in the TFS domain, marking a significant

milestone. It is the first to comprehensively cover primary letter finger spelling, catering specifically to the needs of the Thai sign language research community.

Secondly, we perform comprehensive research on designing and developing a finger spelling recognizer for TFS based on our collected dataset. In particular, our recognizer is analyzed based on extensive experiments in three modalities and different representation learning techniques: RGB-sequencing-based modality on CNN-LSTM and VGG-LSTM models, The coordinate sequence of joint structure modality in the human body with LSTM, Bi-LSTM, GRU and Transformer models, and the graph structure on the skeleton modality using TGCN. We have designed seven important experiments to meticulously evaluate our framework, focusing on distinct hand poses and gestures. The experiments cover: static single-hand poses with single-stroke, dynamic single-hand poses that require two or three strokes, two-hand poses with a static point-on-hand, and two-hand poses with dynamic point-on-hand, total two-hand poses. To measure the performance across these various scenarios, we use evaluation metrics such as accuracy (Top-1), Top-3, Top-5, recall, precision, and F1-score, testing 29 experimental models that include single-based, dual, and triple modalities. Upon obtaining the optimal model, it will be compared against baseline and state-of-the-art models to benchmark its performance.

Thirdly, we conduct comprehensive statistical tests, including both in-sample and out-of-sample evaluations, to rigorously identify the model that demonstrates the highest efficiency. This meticulous approach ensures that we are able to recommend a model that is most suitable for practical, real-world applications, guaranteeing its reliability and effectiveness in various situations.

Research Accomplishment

Journal

[1] Wuttichai Vijitkunsawat}, Teeradaj Racharak, and Nguyen Le Minh. “Deep Multimodal-based Finger Spelling Recognition for Thai Sign Language: A New Benchmark and Model Composition”, *Machine Vision and Applications*, (peer reviewed journal).

International Conference papers

[2] Wuttichai Vijitkunsawat, Teeradaj Racharak, Chau Nguyen, Nguyen Le Minh, “Video-Based Sign Language Digit Recognition for the Thai Language: A New Dataset and Method Comparisons”, *12th International Conference on Pattern Recognition Applications and Methods (ICPRAM2023)*, pages 775-782, 2023. (Presentation at the conference with peer reviewed)

[3] Wuttichai Vijitkunsawat, Teeradaj Racharak, and Nguyen Le Minh. “Deep Multimodal-based Number Finger Spelling Recognizer for Thai Sign Language”, *2023 International Symposium on Communications and Information Technologies (ISCIT2023)*, pages 105-110, 2023. (Presentation at the conference with peer reviewed)

論文題目：聴覚末梢モデルを利用した音質評価指標の計算モデルの構築

希望取得学位：情報科学

研究室名：鶴木研究室

学生番号：1820402

氏名：磯山拓都

第一部：研究内容

ラウドネス、シャープネス、ラフネス、変動強度などの音質評価指標（SQMs）は、感覚的快さのモデル化や煩わしさのモデル化、製品のサウンドデザイン、サウンドスケープ分析、製品の故障判断などに利用されている。不快音を快音に変換させる研究について考えると、入力と出力だけではなく、その途中にあるブラックボックスつまりは、ヒトの聴覚メカニズムの解明が鍵となる。もし、SQMs に対する聴覚メカニズムが明らかになれば、不快音を快音に変換できるかもしれない。

ラウドネス、シャープネス、ラフネス、変動強度はラウドネス密度をもとに構築されている。特に、ラフネスと変動強度はラウドネス密度の時間変動から求めるため、時間領域で計算するラウドネス密度を計算するモデルが必要である。

Moore and Glasberg は、「ヒトの知覚経路に沿ってラウドネスを計算すべき」という立場で、等価矩形帯域（ERB）によって特徴づけされた聴覚フィルタとして roex 聴覚フィルタを用いてラウドネスモデル（Moore-Glasberg モデル）を提案した。この roex 聴覚フィルタは、音圧レベルに依存した聴覚フィルタのフィルタ形状の非対称性を考慮している。Moore-Glasberg モデルは、ヒトの聴覚末梢系の機能とよく対応のとれたモデルであり、聴覚末梢系での情報表現とラウドネスとの関係が明らかとなった。しかし、Moore-Glasberg モデルで使われている roex 聴覚フィルタはインパルス応答を持たない周波数領域で定義されたフィルタであるため、ラウドネス密度の時間変動を計算することが難しい。そのため、SQMs の計算モデルを構築するためには、時間領域で定義された聴覚フィルタを用いてラウドネスモデルを構築する必要がある。

本研究は、精緻化された聴覚末梢系モデルを用いて算出されたラウドネス密度を中心に、入力信号の音響特徴（物理量）と SQMs との関係について取り扱う。そのために、(1) 時間領域の聴覚フィルタを用いたラウドネスモデルで Moore-Glasberg モデルと同様のラウドネスを計算できるか、(2) ERB 尺度や、聴覚フィルタのフィルタ形状の非対称性が SQM の計算に与える影響は何かを焦点を当てて検討する。

はじめに、時間領域で定義された聴覚フィルタバンクとして、ガンマトーン聴覚フィル

タバンク (GTFB) 及びガンマチャープ聴覚フィルタバンク (GCFB) を用いてラウドネス密度の計算モデルを構築した。GTFB と GCFB は roex 聴覚フィルタと同様に聴覚フィルタ形状の推定に使われたものであり、音圧レベルが低い場合は、GTFB と GCFB 及び roex 聴覚フィルタは同程度に聴覚フィルタを推定することができる。一方、音圧レベルが高いときは、音圧依存の聴覚フィルタ形状の非対称性を考慮した GCFB は音圧依存の聴覚フィルタ形状の非対称性を考慮した roex 聴覚フィルタと同様にヒトのデータをよく説明できることが知られている。このことから Moore-Glasberg モデルで使われている roex 聴覚フィルタの代わりに GTFB や GCFB を使用して、Moore-Glasberg モデルと同程度にラウドネス密度を計算することができると考えられる。聴覚フィルタのフィルタ形状の非対称性を考慮していない GTFB と聴覚フィルタのフィルタ形状が非対称性を考慮した GCFB を用いてラウドネス密度を計算することには、音質評価指標の推定に聴覚フィルタのフィルタ形状の非対称性がどのように影響しているか調べるという狙いがある。

上記で構築したラウドネス密度の計算モデルを用いて、音質評価指標の計算モデルを構築した。従来のモデルを参考に、構築したラウドネス密度の計算モデルより算出したラウドネス密度に合わせたパラメータの設定を行った。この時のパラメータ設定は、Fastl and Zwicker の検討により明らかになった音質評価指標に関する主観評価結果を元に行った。構築した音質評価指標の計算モデルの出力及び従来の音質評価指標の計算モデルの出力と音質評価指標に関する主観評価結果を比較することで、構築した音質評価指標の計算モデルの有効性を評価した。

時間領域で定義される聴覚フィルタを用いて構築したラウドネスモデルによって計算されたラウドネスと Moore-Glasberg モデルによって計算されたラウドネスは同様であったことから、構築したラウドネスモデルは Moore-Glasberg モデルの時間領域版として扱うことができることが分かった。また、シャープネス、ラフネス、変動強度の計算モデルの検討から、精緻化された聴覚末梢系モデルを使うことは、音質評価指標の推定誤差を小さくすることに寄与することが分かった。

以上のことから、時間領域での聴覚フィルタバンクの利用が Moore-Glasberg モデルと同等の性能を発揮することが可能であり、聴覚末梢系モデルの精緻化が音質評価の精度を高める重要な要素であることを明らかにした。これにより、音質評価指標の計算モデルの開発において、ERB 尺度や聴覚フィルタのフィルタ形状の非対称性が重要な役割を果たすことが明らかとなった。

第二部：研究の意義

本研究は、聴覚末梢での聴覚特性が音質評価指標の推定に影響を与えるという仮説の元で、仮説検証型の科学的方法に基づいて行われた。検証方法として、音質評価指標に関する聴取実験の結果をどれくらい構築したモデルが説明できるか評価を行った。本研究の成果は Q1 ジャーナルに採択されたことから、本研究は高い学術的水準を有するといえる。

本研究の新規性は音質評価指標の計算において、従来の Bark 尺度フィルタバンクを用いた Zwicker モデルの代わりに、ERB 尺度を使用し、聴覚フィルタの非対称性を考慮したラウドネスの計算モデルを中心としたアプローチにある。また、この研究の独創性は、時間領域で定義された聴覚モデルを使用して人間が音の質感をどのように知覚しているかを解明しようとする点にある。

今回使用した聴覚フィルタバンクでは、聴覚フィルタの非対称性のみを考慮したが、聴覚抹消での圧縮特性を考慮した聴覚フィルタバンクも入野らによって提案されている。このような聴覚フィルタバンクを使用することで、ヒトの聴知覚により近い音質評価指標の計算モデルを提案することができる。これによって、聴覚研究の発展に貢献できるところに本研究の発展性を有する。

第三部：研究業績

ジャーナル論文

[1] Takuto Isoyama, Shunsuke Kidani, Masashi Unoki, “Computational models of sharpness and fluctuation strength using loudness models composed of gammatone and gammachirp auditory filterbanks,” *Journal of Signal Processing*, Vol. 25, No. 4, pp. 141-144, 2021, doi: 10.2299/jsp.25.141.

[2] Takuto Isoyama, Shunsuke Kidani, Masashi Unoki, “Computational models of auditory sensations important for sound quality on basis of either gammatone or gammachirp auditory filterbank,” *Applied Acoustics*, Vol. 218, No. 15, 2024, doi: 10.1016/j.apacoust.2024.109914.

国際会議

(査読有・口頭発表)

[3] Takuto Isoyama, Shunsuke Kidani, Masashi Unoki, “Computational models of sharpness and fluctuation strength using loudness models composed of gammatone and gammachirp auditory filterbanks,” *Proc. RISP International Workshop in Nonlinear Circuits, Communications and Signal Processing 2021 (NCSP'21)*, Online, Mar. 2021.

(査読有・ポスター発表)

[4] Takuto Isoyama, Shunsuke Kidani, Masashi Unoki, "Modeling of sound quality metrics using gammatone and gammachirp filterbanks," Proc. Forum Acusticum 2020, pp. 2731-2735, Online, Dec. 2020.

[5] Takuto Isoyama, Shunsuke Kidani, Masashi Unoki, "Computational model for predicting sound quality metrics using loudness model based on gammatone/gammachirp auditory filterbank and its applications," Proc. Inter Noise 2023, 2-P-48, Chiba, Japan, Aug. 2023.

国内発表

研究会

[6] 磯山拓都, 木谷俊介, 鶴木祐史, "聴覚フィルタバンクを利用した定常音に対するラウドネスモデルの構築," 電子情報通信学会技術研究報告, Vol. 119, No. 440, pp. 273-278, 沖縄, 2020年3月.

[7] 磯山拓都, 木谷俊介, 鶴木祐史, "聴覚フィルタバンクを用いたラウドネスモデルとシャープネスモデルの構築," 日本音響学会聴覚研究会資料, Vol. 50, No. 20, pp. 129-124, オンライン, 2020年5月.

[8] 磯山拓都, 木谷俊介, 鶴木祐史, "聴覚フィルタバンクを利用した音質評価指標の計算モデルの検討," 日本音響学会聴覚研究会資料, Vol. 52, No. 6, pp. 441-446, 宮城, 2022年8月.

[9] 磯山拓都, 木谷俊介, 鶴木祐史, "聴覚フィルタバンクを用いた時変動音のラウドネス計算法の検討," 日本音響学会聴覚研究会資料, Vol. 53, No. 6, pp. 283-288, 宮城, 2023年9月.

国内学会

[10] 磯山拓都, 木谷俊介, 鶴木祐史, "ガンマトーンフィルタバンクを用いたラウドネスモデルの構築," 日本音響学会 2020年春季研究発表会講演論文集, 1-Q-1, 埼玉, 2020年3月.

[11] 磯山拓都, 木谷俊介, 鶴木祐史, "聴覚モデルを用いたシャープネスと変動強度の計算モデルの構築," 日本音響学会 2021年春季研究発表会講演論文集, 1-4P-2, オンライン, 2021年3月.

[12] 磯山拓都, 木谷俊介, 鶴木祐史, "聴覚フィルタバンクを用いたラフネスモデルの構築," 日本音響学会 2022年秋季研究発表会講演論文集, 1-P-13, 北海道, 2022年9月.

[13] 磯山拓都, 木谷俊介, 鶴木祐史, "聴覚フィルタバンクを用いた時変動音のラウドネス計算法の構築," 日本音響学会 2023年秋季研究発表会講演論文集, 1-R-5, 愛知, 2023年9月.