MULTICOLLAB-ASL: Towards Affective Computing for the Deaf Community

Hayden Orr
Rochester Institute of Technology
Rochester, New York, USA
hro4957@rit.edu

Michael Peechatt
Rochester Institute of Technology
Rochester, New York, USA
mp6510@rit.edu

Cecilia O. Alm
Rochester Institute of Technology
Rochester, New York, USA
coagla@rit.edu

Abstract

In American Sign Language (ASL), a prominent resource gap exists for affective computing datasets. This manuscript explores preliminary findings from an ongoing multimodal ASL corpus collection and analysis study, focusing on human-generated modalities (e.g., eye tracking, facial expression, head movement) and the expression of frustration and confusion among deaf and hard of hearing study participants. These affective states can be important for understanding user experiences in human-computer interaction or for offering system feedback towards enhancing AI-human collaboration. Expanding a data collection methodology from prior work involving English-speaking participants, this exploratory study seeks to discern characteristics associated with confused or frustrated affect states in collected signed language interactions. Such insights have the potential to facilitate the development of models capable of recognizing emotional expressions. Initial results reveal distinctions in the characteristics of self-annotated instances of participant frustration and confusion, with certain features showing some divergence between the two emotions.

CCS Concepts

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in accessibility**; **Accessibility design and evaluation methods**.

Keywords

ASL, deaf and hard of hearing, multimodalilty, affective computing

ACM Reference Format:
Hayden Orr, Michael Peechatt, and Cecilia O. Alm. 2024. MULTICOLLAB-ASL: Towards Affective Computing for the Deaf Community. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24), October 27–30, 2024, St. John's, NL, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3663548.3688500

1 Introduction

Automatically inferring emotional states such as frustration in interactions can enable AI-based systems to facilitate the analysis and study of interactions between interlocutors. Affective computing modeling can also help enhance and adjust a system's interactions with users in AI-human teaming contexts, better adapt to enhance user experiences, or even personalize such experiences [17]. There is a substantial gap in American Sign Language (ASL) technologies capable of making such inferences, largely due to the lack of emotional language corpora focused on signed languages [3].

In this study, we report on the early findings of an ongoing multimodal ASL corpus collection and analysis study, focused on studying the expression of frustration and confusion in task-based ASL interactions. Work in this area has potential to contribute to advancing the field of sign language technologies, enabling new insights and resources for the study and modeling of emotional expression in ASL interactions. It may contribute to the development of more robust multimodal machine-learning models that can enhance accessibility for deaf and hard of hearing users, including students in online education. For example, identifying confusion may help bridge communication barriers and facilitate effective communication between instructors and students.

The study uses a multimodal data collection methodology from prior work with task-based dialogue interactions [16]. In contrast to the previous work, which involved English, the present study focuses exclusively on task-based multimodal dialogues in ASL, with participants who are deaf or hard of hearing. We investigate parallels and distinctions related to confusion and frustration in dialogues, in terms of their multimodal expressions.

ASL incorporates layers of meaning functions, including expressive dimensions. For example, signing velocity and space, facial expressions, body tilt, and mouthing may convey expressive and grammatical or lexical/sentential meaning functions. By studying expressions of frustration and confusion in signers, we can gain insights towards developing models to recognize affect expressions, including from non-manual signals. In the future, access to expressive corpus data for training models may also benefit sign language processing models that are robust to emotional variation.

2 Related Work

There is a body of research on multimodal language data involving speech and spoken interactions, including work that explores emotional expressions [1]. In contrast, there is a need for affective computing study centered on sign language in general and ASL specifically. Previous work has pointed out challenges in computer vision-centered sign language research, involving spatial coherence and written transcription [18]. Prior insights indicate that addressing this research gap requires collaborative research and guidance from Deaf communities, a practice that has not always been followed in the past [7, 18].

Sign Language Recognition (SLR) has been studied in terms of *isolated* and *continuous* SLR. The former aims to identify isolated

ASSETS '24, October 27–30, 2024, St. John's, NL, Canada

ACM ISBN 979-8-4007-0677-6/24/10
https://doi.org/10.1145/3663548.3688500

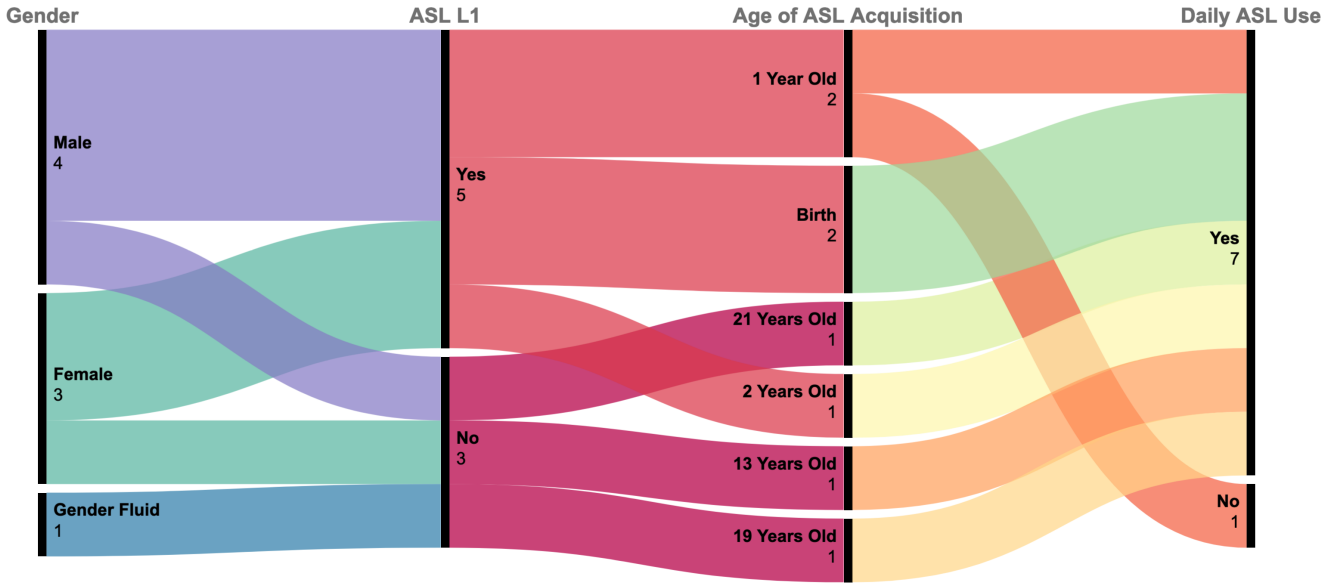

Figure 1: The participants' age range so far, based on the current set of completed study sessions, was mostly between 18 to 29 years old. Three identified as Hispanic or Latino/a. The study is ongoing and these demographics are preliminary.

signs and tends to involve a limited pool of signs [5, 13] or letter units [4, 10, 19]. Continuous SLR seeks to identify where signs begin, continue, and end, and longer linguistic structures [6, 11, 12], yet often resulting in lower performance and they requires large datasets and resources, This highlights the need to create more resources in ASL and other sign languages.

ASL grammar and expression also prominently rely on non-manual cues, such as characteristics related to facial expressions and body movements. Some studies on Sign Language Recognition have successfully incorporated human-generated modalities with an increase in accuracy to some degree [12]. However, other work suggests that model speed might degrade with the inclusion of human-generated modalities despite the increase in accuracy, as their models generally perform faster with fewer landmarks [15], or indicated a performance improvement when modeling did not use facial information [14]. There is a need for studying the adequate integration of non-manual signals into automated modeling for sign language technologies.

3 Methodology

3.1 Participants

We report on our initial data collection, which exclusively involved deaf or hard of hearing ASL users ($n = 8$). Following the existing methodology [16], each participant in the study was assigned a role of *builder* or *instructor*. Figure 1 shows some demographic information about the participants at this early study stage. The ASL background of the participants was as follows: 62.5% had American Sign Language as their native language (L1), 25% were fluent in American Sign Language as their second language (L2), and 12.5% had conversational proficiency or were inexperienced in ASL, with 1 of the 4 paired groups being a mixture of L1 and L2 interactions.

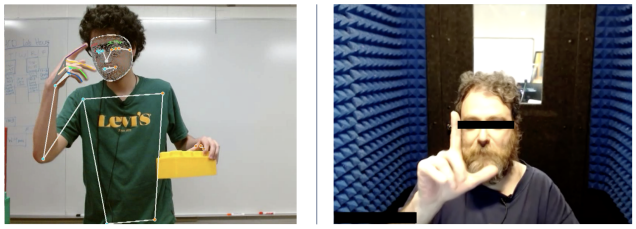

Figure 2: The builder (left) receives instructions about what to build from the instructor (right), who is in their own space and connected through Zoom. The builder wears an eye-tracker and their skeletal pose is tracked. The instructor's eye movements are also tracked using a non-wearable remote eye-tracker. The instructor signs with their dominant hand, while their non-dominant hand is connected to a hand-worn Galvanic Skin Response sensor (not seen in the picture).

3.2 Materials and Procedure

After roles (builder and instructor) were assigned, the paired participants collaborated in their individual space and communicated with each other through a Zoom video session, as shown in Figure 2. The instructor's responsibility was to direct the builder in constructing a structure consisting of blocks and pieces by using an image which was visible only to them. Communication between the instructor and the builder occurred through sign language. The builder had full use of both hands to communicate, while the instructor was restricted to using only their dominant hand, and their other hand remained still with a worn Galvanic Skin Response (GSR) sensor. The builder's task was to recreate the structure to the best of their ability based on the instructor's directions.

3.3 Post-experimental Data Processing

Each group session recorded time series sensor data both for the instructor and the builder. The sensors had disparate polling rates and different recording start times. To synchronize the signal across these modalities, a movie director clapboard was used to time the signal at the beginning and end of each task. This audiovisual signal was captured across all sensors and provided an anchor, a common starting point, from which all data points could be aligned. The timestamp of the clapboard signal was manually determined in post-data processing. The timestamp for each modality was then used to associate participant annotations with the sensor data for the beginning of each task. The Zoom video recording was also used to collect the participants' self-annotated instances of frustration and confusion, both for the builder and the instructor.

After the different modalities had been aligned, it was necessary to address the different sampling rates among the sensors. This involved sampling sensor data surrounding a specified time window from annotations and then dividing the signals into time steps. Cubic spline interpolation was used for upsampling modalities with lower polling rates [9], including for the instructor's facial action units and eye tracking data extracted using iMotions [8]. Downsampling averaging was used when the number of time steps was less than the number of samples collected within the specified time window, for signals with high sampling rates, including the GSR sensor and the Freemocap skeletal pose recordings [2]. All signals were Z-score normalized per participant for comparisons in the diverse groups of subjects.

4 Results

After processing the data, we proceeded to perform analysis on the sensor data surrounding the timestamps of the self-annotated instances of frustration and confusion from the participants. The annotation tool allowed participants to rate their emotion on a discrete ordinal scale of four values: *Not At All*, *Slightly*, *Very*, and *Extremely*. To convert the collected data into binary labels, we combined the labeled instances of *Very and Extremely* to form the positive class, and we used *Not At All* instances as the negative class. These values were then used to calculate the M_{rank} metric for features [16], shown in Table 1. These values were calculated with the following equation:

$$M_{rank} = \left| \sum_{t=1}^{t_s} \overline{\mathbf{m}_t^+} - \sum_{t=1}^{t_s} \overline{\mathbf{m}_t^-} \right| \tag{1}$$

where $\overline{\mathbf{m}_t^+}$ represented the average value for modality m at time step t for the positive class, and $\overline{\mathbf{m}_t^-}$ represented the negative class. The higher the value, the larger the difference was between the average curves of time steps for the modality surrounding self-annotated instances from participants. Moreover, Figure 3 illustrates the difference in the distribution of builder head velocities for frustration.

5 Discussion

A key finding of this exploratory study is that head velocity tends to increase when participants indicated that they perceived themselves frustrated compared to when they were not frustrated. The data indicate that incorporating signals such as head movement measures might support distinguishing emotional expressions. This is evidenced by the spread of the data points shown in Figure 3. Although there were more confusion annotations (n = 157) than frustration annotations (n = 134), this early study confirms that non-manual data seems important for emotion modeling in ASL interactions. ASL integrates cues from facial expressions and body movements to convey nuances, emphasis, and emotion, in addition to other information. Frustration may stem from the recognition that the signer's message is not being understood by the recipient, while confusion may arise from the signer's difficulty in understanding the message being conveyed to them. The study suggests, for instance, that there are important distinctions between states of frustration and non-frustration, underscoring the need for more advanced models to analyze emotional multimodal data further, towards enhancing detection capabilities.

The limitations of this early study include a relatively small sample size that predominantly involved a younger adult age group. In addition, many participants were familiar with each other, which could have reduced the likelihood of experiencing or expressing frustration or confusion.

Table 1: Ranked instructor frustration and confusion features by quantifying separation of the average curves between both positive and negative classes. Bold features show a difference in ranking between the two emotions (frustration or confusion), suggesting that each emotion is expressed differently. Certain features, such as Lip Corner Depressor, are perceived to have an increased difference between positive and negative class instances when expressing confusion rather than frustration.

Frustration: Features	M_{rank}
Saccade Duration	8.13
Brow Furrow	4.86
Lip Corner Depressor	4.24
Fixation Duration	2.53
Gaze Velocity	2.40
Saccade Peak Velocity	2.22
Lid Tighten	2.16
Chin Raise	1.35
Fixation Dispersion	0.74

Confusion: Features	M_{rank}
Lip Corner Depressor	6.67
Brow Furrow	4.30
Fixation Dispersion	3.44
Fixation Duration	3.32
Lid Tighten	3.18
Saccade Peak Velocity	2.23
Saccade Duration	0.89
Chin Raise	0.82
Gaze Velocity	0.73

6 Conclusion

In conclusion, this study underscores the importance of considering multimodal data when analyzing emotional sign language resources. Additionally, collecting such resources can be useful for developing

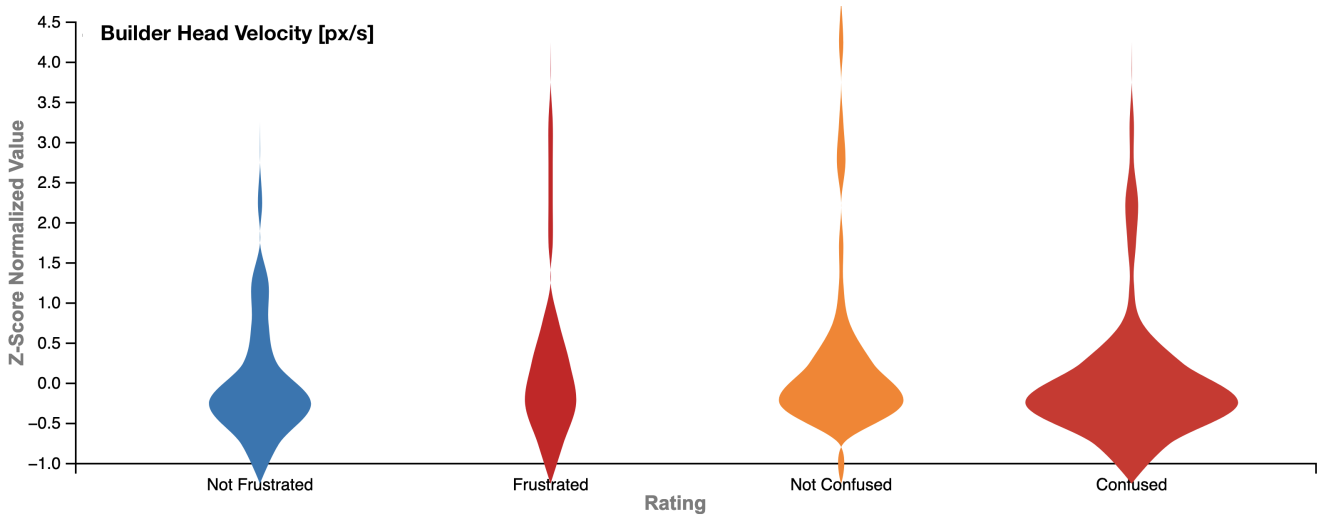

Figure 3: Frustration annotations from builders indicated a faster head movement velocity when their self-annotation marked that they perceived they were frustrated, as shown by the distribution of instances above the baseline. Confusion annotations for both positive and negative classes show head velocities clustered around the baseline, indicating a more stable movement when the builder was confused.

robust sign language technologies. By focusing on the expressions of frustration and confusion, this study provides insight that can contribute to technical advances in affective computing, e.g., for improving user experiences for deaf and hard of hearing individuals.

Although the current study is in progress, preliminary findings indicate a wealth of potential information to explore. We will continue to collect data and include participants with diverse levels of sign language proficiency and linguistic backgrounds to broaden representation in the collected dataset. We plan to study the ambiguity in how non-manual cues such as facial expressions convey emotions in addition to grammatical functions.

7 Ethics Statement

The study received approval from the Institutional Review Board (IRB) and involved the acquisition of informed consent from all participants, each being compensated $40 USD for participation. The collected biophysical signal data were anonymized. This study was conducted by an interdisciplinary team composed of deaf and hearing members, with one member having primary fluency in ASL.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation* 42 (2008), 335–359.

[2] Aaron Cherian, Philip Queen, Wirth Trent, Idehen Endurance, and Jonathan Samir Matthis. [n. d.]. *FreeMoCap: A Free, Open Source Markerless Motion Capture System.* https://doi.org/10.5281/zenodo.7233714

[3] Emely Pujólli da Silva, Paula Dornhofer Paro Costa, Kate Mamhy Oliveira Kumada, José Mario De Martino, and Gabriela Araújo Florentino. 2020. Recognition of Affective and Grammatical Facial Expressions: A Study for Brazilian Sign Language. In *Computer Vision – ECCV 2020 Workshops*, Adrien Bartoli and Andrea Fusiello (Eds.). Springer International Publishing, Cham, 218–236.

[4] Ahmad Yahya Dawod and Nopasit Chakpitak. 2019. Novel Technique for Isolated Sign Language Based on Fingerspelling Recognition. In *2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*. 1–8. https://doi.org/10.1109/SKIMA47702.2019.8982452

[5] Rabeet Fatmi, Sherif Rashad, and Ryan Integlia. 2019. Comparing ANN, SVM, and HMM based Machine Learning Methods for American Sign Language Recognition using Wearable Motion Sensors. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*. 0290–0297. https://doi.org/10.1109/CCWC.2019.8666491

[6] Sevgi Z. Gurbuz, Ali Cafer Gurbuz, Evie A. Malaia, Darrin J. Griffin, Chris S. Crawford, Mohammad Mahbubur Rahman, Emre Kurtoglu, Ridvan Aksu, Trevor Macks, and Robiulhossain Mdrafi. 2021. American Sign Language Recognition Using RF Sensing. *IEEE Sensors Journal* 21, 3 (Feb 2021), 3763–3775. https://doi.org/10.1109/JSEN.2020.3022376

[7] Raychelle Harris, Heidi M. Holmes, and Donna M. Mertens. 2009. Research Ethics in Sign Language Communities. *Sign Language Studies* 9, 2 (2009), 104–131.

[8] iMotions. 2023. Powering Human Insights - Biometric Research. https://imotions.com Accessed on October 13, 2023.

[9] Ibtissem Khouaja, Ibtihel Nouira, M. Hedi Bedoui, and Mohamed Akil. 2016. Enhancing EEG Surface Resolution by Using a Combination of Kalman Filter and Interpolation Method. In *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*. 353–357. https://doi.org/10.1109/CGiV.2016.74

[10] Karly Kudrinko, Emile Flavin, Xiaodan Zhu, and Qingguo Li. 2021. Wearable Sensor-Based Sign Language Recognition: A Comprehensive Review. *IEEE Reviews in Biomedical Engineering* 14 (2021), 82–97. https://doi.org/10.1109/RBME.2020.3019769

[11] David Laines, Miguel Gonzalez-Mendoza, Gilberto Ochoa-Ruiz, and Gissella Bejarano. 2023. Isolated Sign Language Recognition based on Tree Structure Skeleton Images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 276–284. https://doi.org/10.1109/CVPRW59228.2023.00033

[12] Taeryung Lee, Yeonguk Oh, and Kyoung Mu Lee. 2023. Human Part-wise 3D Motion Context Learning for Sign Language Recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 20683–20693. https://doi.

org/10.1109/ICCV51070.2023.01896

[13] Peter B. Shull Ling Li, Shuo Jiang and Guoying Gu. 2018. SkinGest: Artificial Skin for Gesture Recognition via Filmy Stretchable Strain Sensors*. *Advanced Robotics* 32, 21 (2018), 1112–1121. https://doi.org/10.1080/01691864.2018.1490666 arXiv:https://doi.org/10.1080/01691864.2018.1490666

[14] Boris Mocialov, Graham Turner, Katrin Lohan, and Helen Hastie. 2017. Towards Continuous Sign Language Recognition with Deep Learning. In *Proceedings of the Workshop on Creating Meaning with Robot Assistants: The Gap Left by Smart Devices*, Vol. 5525834.

[15] Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. 2020. Real-Time Sign Language Detection Using Human Pose Estimation. In *Computer Vision – ECCV 2020 Workshops*, Adrien Bartoli and Andrea Fusiello (Eds.). Springer International Publishing, Cham, 237–248.

[16] Michael Peechatt, Cecilia O. Alm, and Reynold Bailey. 2024. MULTICOLLAB: A Multimodal Corpus of Dialogues for Analyzing Collaboration and Frustration in Language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italy, 11713–11722. https://aclanthology.org/2024.lrec-main.1023

[17] Carson Reynolds and Rosalind W Picard. 2001. Designing for Affective Interactions. In *Proceedings from the 9th International Conference on Human-Computer Interaction*. 6.

[18] Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 7347–7360. https://doi.org/10.18653/v1/2021.acl-long.570

[19] Guan Yuan, Xiao Liu, Qiuyan Yan, Shaojie Qiao, Zhixiao Wang, and Li Yuan. 2021. Hand Gesture Recognition Using Deep Feature Fusion Network Based on Wearable Sensors. *IEEE Sensors Journal* 21, 1 (Jan 2021), 539–547. https://doi.org/10.1109/JSEN.2020.3014276