FSboard: Over 3 million characters of ASL fingerspelling collected via smartphones

Manfred Georg[1*], **Garrett Tanzer**[1*], **Esha Uboweja**[1], **Saad Hassan**[2‡], **Maximus Shengelia**[3‡], **Sam Sepah**[1], **Sean Forbes**[4†], **Thad Starner**[1*†]
[1]Google, [2]Tulane University, [3]Rochester Institute of Technology,
[4]Deaf Professional Arts Network
{mgeorg, thadstarner}@google.com

Abstract

Progress in machine understanding of sign languages has been slow and hampered by limited data. In this paper, we present FSboard, an American Sign Language fingerspelling dataset situated in a mobile text entry use case, collected from 147 paid and consenting Deaf signers using Pixel 4A selfie cameras in a variety of environments. Fingerspelling recognition is an incomplete solution that comprises only a small part of sign language translation, but it could provide some immediate benefit to Deaf/Hard of Hearing signers while more broadly capable technology develops. At >3 million characters in length and >250 hours in duration, FSboard is the largest fingerspelling recognition dataset to date by a factor of >10x. As a simple baseline, we finetune 30 Hz MediaPipe Holistic landmark inputs into ByT5-Small and achieve 11.1% Character Error Rate (CER) on a test set with unique phrases and signers. This quality degrades gracefully when decreasing frame rate and excluding face/body landmarks—plausible optimizations to help with on-device performance—but falls short of human performance measured at 2.2% CER.[1]

1. Introduction

The quality of sign language translation, particularly from American Sign Language (ASL) to English, has been steadily improving [14, 47, 53, 54], but it is still far from being usable in practice. A body of work on participatory methods for ML [9, 11] suggests dividing such an ambitious goal into intermediate milestones that can provide concrete and immediate benefit to the community (*i.e.*, to Deaf/Hard of Hearing signers). In this manner, the work starts addressing the community's needs immediately, and the community can drive the direction of future technology.

We focus on the intermediate goal of recognizing fingerspelling as an alternative to smartphone text entry. While full signing for text entry (the proper analogue of speech recognition) is ideal [24], fingerspelling may still be a valuable stopgap due to improved speed or convenience vs. typing on a keyboard. An analogy can be made to gesture keyboards [57], where the user swipes through the letters of a word as opposed to touching and releasing each letter's virtual key. Even though the system uses pattern recognition to determine which word the typist intended and sometimes returns an incorrect word, many smartphone users prefer such gesture-based keyboards as they feel they can enter text more quickly with the added benefit of requiring one hand instead of two [46]. Similarly, there is evidence that Deaf signers may find fingerspelling faster or more convenient than current smartphone text entry keyboards [24].

In this paper we present FSboard (**F**inger**s**pelling-**board**, as in "keyboard"), an ASL fingerspelling dataset situated in a mobile text entry use case. We collect FSboard by creating a domain-appropriate phrase distribution, recruiting 147 paid and consenting Deaf signers through the Deaf Professional Arts Network (DPAN), and having them record one-handed fingerspelled renditions of the phrases using Pixel 4A selfie cameras in a variety of environments. The videos were usually recorded at 1944x2592 pixels and 30 frames per second, though sometimes the resolution varied due to participants accidentally changing settings. At 3.2 million characters in length and 266 hours in duration, FSboard is the largest fingerspelling recognition dataset to date by a factor of >10x.

As a simple model baseline, we finetune 30 Hz MediaPipe Holistic landmark [18] inputs into ByT5-Small (300M) [55] and achieve 11.1% Character Error Rate (CER) on FSboard's test set, which features 15 unique signers and no train phrase overlap. We ablate our baseline across several factors like frame rate and exclusion of face/body landmarks which could be used to optimize

*equal contribution † equal advising ‡ work conducted at Google
[1]We publicly release FSboard at under CC BY 4.0.

MediaPipe Holistic's on-device performance, and find that these compromises (in moderation) cause minimal regressions. However, these results fall short of our measurement of human performance at 2.2% CER. Qualitatively, the baseline outputs are promising but should be evaluated end-to-end in realistic settings in future work as results improve.

We hope that FSboard will help develop text entry methods that start to give signers a more equitable experience with technology, as well as aid in longer term research towards full sign language understanding.

2. Background

According to the United Nations and the World Health Organization, there are over 70 million Deaf and Hard of Hearing people in the world [1, 2]. Many use one or more of around 150 sign languages to communicate [16]. For example, American Sign Language (ASL) is used by about 500,000 people as a primary language in United States alone [35].

Sign languages are complete, natural languages that can differ from one another significantly even across societies where the same spoken language is used [4, 16]. For example, American Sign Language differs from British Sign Language (BSL) significantly and is instead genetically related to French Sign Language [4]. However, almost all sign languages include a manual alphabet used to represent letters as hand shapes and movements. Often, fingerspelling is used for proper nouns or when introducing new concepts. Fingerspelled terms may also be adapted into the closed vocabulary of a sign language in a process called lexicalization [28, 56]. For example, the sign for "job" in ASL can be seen as a combination of the letters J and B, with the O abbreviated or elided, but the B is in a different orientation from typical fingerspelling. The amount of fingerspelling used in conversational discourse varies depending on the sign language. In ASL, fingerspelling is about 12%-35% of signing [29, 40].

Some fingerspelling systems are one-handed, such as ASL or Japanese Sign Language (JSL), while others, such as BSL, are two-handed [43]. Even though meanings may be different, similar hand shapes and movements can be seen across sign languages due to the physical constraints of the hand [34], which suggests that datasets collected for one sign language may offer some transfer to others, or to recognizing handshapes and movements in signing more broadly. This benefit may be especially true for ASL fingerspelling, which belongs to the largest cluster of manual alphabets, the "French-origin group" [43]. For example, many of the letters in the ASL manual alphabet have the same handshape in French, Italian, and German Sign Languages.

There are many works, both informal and academic, which claim to study fingerspelling recognition for ASL and other sign languages but operate on single images. Such efforts are in reality studying handshape classification, with exceptions for fingerspelled letters such as J and Z that incorporate movement. Ghanem et al. [17] provide a survey. Real-time demonstrations are often slow, with feedback on one letter at a time. These systems ignore the co-articulation effects that occur when recognizing fingerspelling at speed as well as the problem of determining where the space is between two fingerspelled words. In addition, when fingerspelling at speed signers often "bounce" or "slide" a handshape from inside to outside the body when a letter repeats in a word. We are only aware of the ChicagoFS series of datasets [6, 7, 30] that are directly comparable to the effort here for American Sign Language fingerspelling; we provide a detailed comparison to ChicagoFS in Section 3.4 below.

2.1. Fingerspelling for smartphone text entry

Historically, most sign language recognition systems have had little usefulness or usability for the Deaf community [9, 10, 15, 25]. Often the Deaf community is not consulted on the technology being created, nor formative or summative user studies performed. For this work, three of the authors are members of the Deaf community and were integral to the selection of the task, pilot studies and testing, and recruitment of the participants whose primary language is ASL.

Our efforts to create a fingerspelling dataset are motivated by Hassan et al. [24], a user study which established the potential benefits of text entry for smartphones based on fingerspelling. It compared an emulated fingerspelling keyboard to normal smartphone typing on Gboard (Android's default keyboard) for 12 Deaf participants and found that fingerspelling was faster than the smartphone keyboard (42.5 wpm vs. 31.9 wpm), had fewer errors (4.0% vs. 6.3%) and had higher throughput (14.2 bits/second vs. 10.9 bits/second). In post-study surveys, 50% of these Deaf participants preferred fingerspelling for text entry using the emulated recognition system.

Further adding support that fingerspelling may prove faster than smartphone virtual keyboards, we examined common MacKenzie phrases fingerspelled in the FSboard dataset presented below. Signers averaged around 65 wpm, with some maintaining over 100 wpm. This result is significantly faster than the average smartphone typist at around 36 wpm [41] and is consistent with the conversational fingerspelling rates reported in the literature [44].

Texting is often the first use case that comes to mind when thinking about text entry on a smartphone. However, members of the Deaf community have emphasized that fingerspelling to a smartphone may be best suited for entering names or addresses into specific smartphone applications like Google Maps. One can imagine a Deaf signer setting

Figure 1. **A sample of frames from FSboard.** Faces blurred here but not in the dataset.

MacKenzie [33]	prevailing wind from the east elephants are afraid of mice my favorite place to visit
URLs	http://datastudio.google.com si.wikipedia.org /dfinance/list.asp?id=418/
Addresses	9841 gritt hill 200ab lake charles 24 north 118th place
Phone Numbers	166-893-6320 +44-527-848-96-69-05 +678-92-00-9661
Names	mohammed kim gustavo ho clifford davenport

Figure 2. **A sample of phrases from each category of FSboard.** Addresses, phone numbers, and names are generated randomly; they are not real personally identifiable information (PII).

their default keyboard on their smartphone to one that shows both the on-screen keyboard as well as a selfie camera feed that could be used for fingerspelling. In this manner, the signer could easily switch between, or combine, input methods for all applications that require text entry.

2.2. Community-centered sign language datasets

PopSign and ASL Citizen are two prior works grounded in focused tasks that could benefit signing communities. Both are isolated sign recognition tasks (classifying which single sign is present in a given clip). PopSign [50] is an educational smartphone game intended to help hearing parents of deaf infants practice sign language (and avoid language deprivation [19–23, 26, 27, 38]).

In the game, the user signs one of five options to select a bubble of a particular color, and these five active options rotate among a library of 250 signs in a way that avoids recognizer confusion. Limiting the number of classes ensures >99% top-1 accuracy [8]. As with FSboard, the PopSign dataset is collected at high resolution (1944x2592) using Pixel 4A smartphone selfie cameras. The PopSign dataset includes over 200,000 clips and 128 hours of video.

ASL Citizen [13] grounds the sign recognition task in a dictionary retrieval setting, enabling signers to record a video clip of a sign to retrieve its dictionary entry. This task demands a much larger sign vocabulary/number of classes, but retrieval is relatively forgiving because a number of outputs can be returned (*i.e.*, top-N accuracy) from which the user can choose. The authors report a top-1 accuracy of 63% but a top-10 accuracy of 90% on a 2731 sign vocabulary. The dataset contains 83,912 videos taken from webcams, and the dataset resolution is often 640x480.

3. The FSboard Dataset

In this section we describe the set of phrases we elicited for FSboard and how the data was collected.

3.1. Phrases

Numbers should not necessarily be considered fingerspelling, but they are relatively formulaic and often appear in the same contexts as/alongside fingerspelling. There are special signs for various numbers (such as 11, 12, 23, 25, 33, and many more) and specific movements which are generally used while signing numbers, but for practical purposes, the systems for signing cardinal numbers and fingerspelling words are very related, so we choose to include numbers within the scope of our fingerspelling recognition system.

We construct our phrase set with a variety of domains: MacKenzie phrases, URLs, addresses, phone numbers, and names. See Figure 2 for examples from each of these categories.

First, we include the MacKenzie phrase set [33], a classic set of 500 phrases used to evaluate text entry systems. These phrases are intended to be collected multiple times by different signers, to serve as a closed vocabulary testbed for sanity checking methods. The rest of the domains are intended to be unique phrases.

Second, we include randomized URLs using URL parts from a web crawl in April 2022. The crawled URLs were

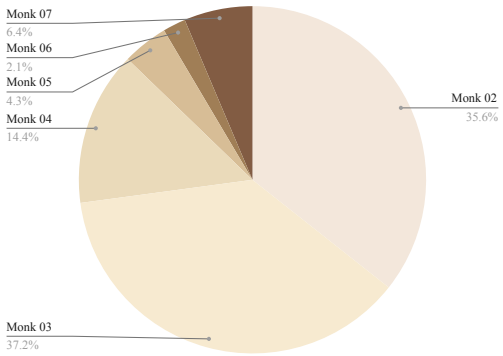

Figure 3. **Monk Skin Tone Scale ratings for FSboard participants**, annotated by majority vote of three human raters trained specifically for the skin tone task.

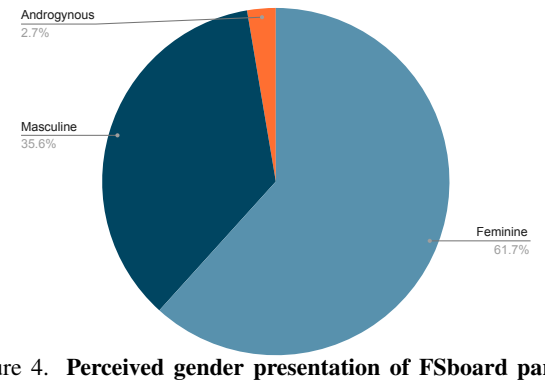

Figure 4. **Perceived gender presentation of FSboard participants**, annotated by human raters. Note that this is not equivalent to gender identity, because it is predicted from the videos rather than self-identified.

broken into unique domain name parts and directory parts. Gibberish parts were removed using a set of manual rules and a simple two character Markov chain trained to recognize URLs a human might want to fingerspell [39]. Finally, URLs were randomly generated from the parts, sometimes including the protocol identifier (such as `https://`) and sometimes removing the domain name entirely. Some URLs suggest explicit content, which we release as a separate collection of metadata to ensure they are not included in the dataset/models trained on it by default.

Third, we include randomly constructed street addresses. The street names were sampled from the US Census Bureau's 2019 TIGER release [12], filtered to remove repetitive, uncommon, and hard to fingerspell names (such as roads that are named by number "Co Rd 87" and "Cr-1601Q4"). Street numbers of 1 to 6 digits were randomly generated with a heavy bias towards 4 digit numbers. Some addresses had their standard abbreviations expanded ("Lane" for "LN", "Road" for "Rd") while others maintained the abbreviation.

Fourth, we include names randomly generated as combinations of the 1000 most common first and last names in the United States.

Fifth, we include random phone numbers. These include 10 digit US numbers, with and without the "+1", and semi-realistic non-US numbers. The non-US numbers were generated with a valid country code and realistic groupings, but no effort was made to create correct lengths for the country code used.

These categories are not exhaustive but form a solid basis for creating and evaluating a practical fingerspelling system. For future data collections, we suggest increasing the representation of symbols and diversity of formats to avoid overfitting. More care could also be taken around elements that should perhaps be explicitly controllable in text entry even if they are less salient in naturalistic fingerspelling, such as whether/how to represent spaces and capitalization.

3.2. Data Collection

The Deaf Professional Arts Network (DPAN) recruited a pool of Deaf signers who use ASL as their primary language to participate in the data collection, of whom 147 completed the task. DPAN shipped loaner Pixel 4A smartphones to them with a data collection app installed, based on the open source Record These Hands platform [24, 48].

The app showed a phrase as text. The participant would touch an on-screen button to begin the phrase, then fingerspell the phrase (with the other hand), and finally press another button to advance to the next phrase. There was also a button to be pressed if a mistake was made in fingerspelling the phrase. For most participants, this button did not work properly, leading to issues in data cleanup. A single video was recorded for the entire session with button presses merely recording a timestamp.

The participants were asked to record videos in various settings and circumstances, leading to many different views and backgrounds. Some people wear masks; sometimes the face or portions of the hands are out of frame; sometimes the field of view is at a strange angle recording videos from below. Many participants opted to place the phone down in some way leading to a large variability in distance to participant, motions required for button presses, and timing between button press and the beginning or end of fingerspelling.

3.3. Data Cleanup

Due to bugs in the data collection app and some user errors, the timespans recorded above frequently did not capture the actual phrase. In order to generate reasonable clips we used a bootstrapping method based on our baseline ByT5 model, described in Section 4. First the model was trained on a large corpus of YouTube videos with captions. Next, five models were finetuned to transcribe folds of the noisily bounded fingerspelling data. (Each fold was trained on

4/5 of the data and transcribed the remaining 1/5.) Since evaluation bias was irrelevant, the dataset splits were performed at the clip level such that every participant was in every split. Each model was then used to predict text for the remaining 1/5th that it had not yet seen. Where the model agreed with the clip boundaries and content, the clip was labeled as clean, otherwise the clip was labeled as noisy. The whole process was repeated two more times (starting each time with a fresh model) using only the clean clips from the previous round. A significant amount of manual editing and custom rules tailored to each participant were then used to further clean the clips. There are still some issues with the clip boundaries in the dataset, and it would benefit from further annotation.

3.4. Dataset Statistics

We divide FSboard into train, validation, and test splits with unique signers in each split (117, 15, and 15 signers respectively) and no overlap of phrases between splits.[2]

See Figure 5 for statistics about FSboard (and its splits) in comparison to prior fingerspelling recognition datasets. At 3.2M characters in phrase length and 266 hours in video duration, FSboard is more than 10x larger than ChicagoFSWild+, the largest prior fingerspelling recognition dataset. The number of sequences in FSboard is only about 3x that of ChicagoFS, reflecting that the average sequence length in FSboard (21.2 characters) is much longer than prior works (5.5 in ChicagoFSWild+) due to the domain. The number of unique signers is also lower (147 vs. 260), since FSboard's data is created from new participants rather than scraped from the web like ChicagoFSWild(+). FSboard is unique for fingerspelling datasets in that it is recorded from a one-handed smartphone perspective, like PopSign [50] does for isolated sign classification.

We used trained human annotators to give more visibility into FSboard's demographic fairness: see Figure 3 for Monk Skin Tone Scale [36] ratings and Figure 4 for classifications of perceived gender presentation. The dataset has a good amount of variation in skin tone but lacks the lightest and darkest ends of the scale. It is approximately as diverse as OpenASL [49] (though the classification system is different), and much more diverse than YouTube-SL-25 [51]. Masculine presentation is underrepresented in FSboard, with only 35.6% of signers. This statistic stands opposite ChicagoFSWild+ [6], which underrepresents feminine presentation to about the same degree; OpenASL and YouTube-SL-25 are essentially at parity.

[2]Due to a bug in the data collection app, sometimes up to 3 signers were prompted to fingerspell the same phrase when we intended to keep the phrases unique. We discarded some data in order to create splits without overlap in signers or phrases. The signers in each split were chosen so as to minimize the amount of data which needed to be discarded in order to keep each phrase only in one split. This resulted in a 29% reduction in the number of phrases in the non-MacKenzie portion of FSboard.

4. Baselines

We provide two sets of baselines, using trained models and human raters respectively. We report character error rate (CER), implemented as length-normalized Levenshtein distance [31] using TensorFlow's implementation [3], as well as top-1 accuracy (*i.e.*, the fraction of examples that are transcribed perfectly).

4.1. Model Baselines

We build our model baselines on ByT5 [55], a character-level encoder-decoder language model in the T5 family [45]. Following the YouTube-ASL baselines [54], we linearly project 85 3D MediaPipe Holistic [18, 32] landmarks into the encoder, with one soft token for each frame of input. Unlike YouTube-ASL's baselines, we feed the input frames in at full (30 Hz) frame rate by default, rather than half; we use up to 256 frames of input and 256 characters of output and decode greedily with a beam size of 5. We train each run using 32 TPUv3 with a batch size of 64 and Adafactor optimizer with base learning rate 0.001 for up to 200k steps per run (or convergence), which takes up to 16 hours. We select checkpoints based on CER on the validation set. In practice, the sampling-based metrics plateau without apparent overfitting, so checkpoint selection is not especially sensitive.

See Figure 6 for a full table of quantitative results; see Figure 7 for a qualitative sample of outputs. Our baseline achieves 11.1% CER and 52.9% top-1 accuracy on the FSboard test set. This baseline surpasses the best score (16.4% CER) set in a Kaggle competition that we hosted based on FSboard [5], though the participants were limited in terms of model size and runtime. See the final leaderboard for descriptions of many more baseline methods. The main difference is that, as far as we can tell, the top submissions did not use pretrained language models but rather trained models with new, custom architectures on FSboard only.

Our baseline is also substantially better than the 37.7% CER achieved by ChicagoFSWild+'s baselines [6].[3] Results obviously cannot be directly compared across test sets in different domains, but it speaks to a combination of our dataset's increased size and choice of domain (including the use of new footage from high-quality selfie cameras, rather than crops of potentially low-resolution web videos). Our baseline results are even better than ChicagoFSWild(+)'s references for human performance at 17.3% (Wild) and 13.9% (Wild+), respectively. This result is because, as they note, the datasets are mined as clips within longer signing videos, and the ground truth annotators have access to broader context but the recognizer being tested does not. While semantic context helps to decode fingerspelling in

[3]We are unable to evaluate our own models on ChicagoFSWild+ due to licensing.

name	lang	# seqs	# chars	# hrs	# signers	source
ChicagoFSVid [30]	ASL	4K	21K	<1	4	Lab
ChicagoFSWild [7]	ASL	7K	38K	2	160	Web
ChicagoFSWild+ [6]	ASL	55K	0.3M	14	**260**	Web
FSboard (ours)	ASL	**151K**	**3.2M**	**266**	147	Smartphone
train		*126K*	*2.8M*	*224*	*117*	
validation		*12K*	*0.2M*	*19*	*15*	
test		*13K*	*0.2M*	*23*	*15*	

Figure 5. **Summary statistics for fingerspelling recognition datasets.**

general, the problem is complicated by signers who systematically fingerspell faster and more sloppily when licensed by the discourse context [42]; to some degree attempting to transcribe isolated fingerspelled subclips may be doomed [52]. We sidestepped this issue by eliciting new data in the same (isolated) context as the desired task, rather than using clips from preexisting longform data.

We ablate several factors that contribute to our baseline's performance:

Pretraining. Finetuning from the pretrained ByT5 checkpoint rather than the randomly initialized architecture makes a massive difference in ultimate performance (11.1% vs. 33.8% CER), and also gives much faster convergence (most of the way by 30k steps, vs. at least 200k).

Model size. We try scaling from ByT5 Small (300M) to ByT5 Base (580M), but quality decreases. We assume that this dataset is not large enough to warrant the extra modeling capacity, and the model overfits to the target distribution too easily. Models even smaller than ByT5 Small might perform better, but it is the smallest available model in its family.

Frame rate. In practice, MediaPipe Holistic is the performance bottleneck for running sign language applications on device, especially for devices that are not quite cutting edge. We ablate frame rate to show that—as expected—quality degrades monotonically with reduced frame rate, but 15 Hz still performs pretty well (11.1% vs. 11.8% CER).

Holistic components. Likewise, we can improve on-device performance by removing some of the component models of MediaPipe Holistic (which consists of Hands, Pose, and Face). Unlike other aspects of sign language, the meaning of fingerspelling can be read purely from the shape of the hands, which makes this test more principled than it would be in other contexts. Removing the face causes slight degradation (11.1% to 12.0% CER), presumably due to the loss of lipreading cues (which different signers use to varying extents), and removing pose has seemingly no effect. It is possible that this gap could grow with more data, if FSboard is not large enough to learn the relevant features.

Model	CER (↓)	Top-1 Accuracy (↑)
Baseline	**11.1**	**52.9**
Pretrained	**(11.1)**	**(52.9)**
Scratch	33.8	17.9
ByT5 Small	**(11.1)**	**(52.9)**
ByT5 Base	13.3	49.1
30 Hz	**(11.1)**	**(52.9)**
30/2 Hz	11.8	51.8
30/3 Hz	13.4	48.2
30/4 Hz	14.6	45.1
30/6 Hz	20.0	33.4
30/8 Hz	27.1	22.2
30/16 Hz	64.0	0.9
30/32 Hz	88.7	0.0
Holistic	**(11.1)**	**(52.9)**
–Face	12.0	50.6
–Face –Pose	12.5	49.7

Figure 6. **Character error rate (CER, ↓) and top-1 accuracy (↑) for FSboard fingerspelling recognition model baselines.** We provide several ablations with respect to our best-performing baseline (which uses 30 Hz MediaPipe Holistic (Hands+Pose+Face) landmarks finetuned into ByT5 Small): the effect of using pretrained language knowledge vs. training the architecture from scratch on FSboard only, building off ByT5 Small (300M) vs. Base (580M), reducing frame rates, and removing Holistic components.

4.2. Human Baselines

In order to quantify how much room remains for improvement on FSboard given participant fingerspelling errors, data cleaning errors, and inherent ambiguities in the video recordings, we also provide a human baseline for the finger-

	Target	Prediction
Random	hubert avalos	elbert avalos
	135-433-9049	135-433-9049
	870055 sunset creek court	870055 sunset creek court
	+43-795-19-03-4208	+43-795-19-03-4208
	www.rehanfyzio.sk	www.rohanfyzio.sk
	a131003/iandrade82	a131003/iandrade82
	eugene or	eugene or
	331 super chief	331 super chief
	nashville tennessee	nashvelle tenssee
	/passeig_de_maragall	/passigo-de-marsgoll
	https://www.rauschenbach.de	https://www.rauschenbach.de
	5225 everette mcclerran	25225 everett mecclar road
	+95-40-860-061-646	+95-40-860-061-647
Failures	sparks nv	sparks.net
	www.tttvw.com/lemoyne-pa	www.tt26.com/lenoyne-p
	2786 lily xing	78 william g
	7806 skunk creek road	378606 skunk creek road
	+54-5828-275-06	7158 twp 2303
	newark new jersey	newton jussey

Figure 7. **Qualitative examples of our baseline's predictions on the FSboard validation set**. Note that these phrases, like all in the validation set, are unseen in the training set. "Random" examples are sampled without cherrypicking, and "failures" are a selection of those with the worst errors. For example, for "tttvw.com", the fingershapes for "vw" and "two six" (not twenty six) look identical.

spelling recognition task. DPAN recruited 2 Deaf signers whose primary communication is in ASL to annotate 100 random sequences from the FSboard test set. We described the task with two variants: first, one pass over the video in real-time at normal playback speed, typing a transcription as they go; and second, unlimited playback/scrubbing of the video at arbitrary speed, until they are satisfied with the transcription. We provided examples of phrases from each category in the train set in order to give a sense of the unconditional text distribution. Even so, the model may more easily learn idiosyncrasies of the train text distribution, such as particular formatting that is always used for phone numbers, so to make the comparison fairer we applied additional canonicalization postprocessing steps to all text before scoring: apply lowercase and remove whitespace, '+', & '-'. We scored using CER and top-1 accuracy as above.

See Figure 8 for quantitative results. Human 1 scores a CER of 2.2% and top-1 accuracy of 74%, trouncing the best model baseline, which scores 13.5% CER and 60% top-1 accuracy on this particular set. (Human 2 scores similarly to Human 1 but slightly worse.) The human one-pass performance is markedly worse, with a best of 22.4% CER and 19% top-1 accuracy. That is to say, the best baseline model performs between a human interpreting the content in real time and a human with arbitrary replays of the content. Qualitatively, this is especially pronounced on phone numbers and URLs, where it is challenging to process and remember/type the long, high-information density sequences in real time. This provides an early example of a use case where sign language technology could be meaningfully more performant than humans rather than just more economical.

Participant	CER (↓)	Top-1 Acc (↑)
Model	13.5	60
Human 1	**2.2**	**74**
One pass	24.8	21
Human 2	3.8	72
One pass	22.4	19

Figure 8. **Character error rate (CER, ↓) and top-1 accuracy (↑) for FSboard fingerspelling recognition with two human raters**, compared to the best model baseline, on a random set of 100 test sequences. For humans, we provide scores for both a) the first pass watching the clip in real time and b) unlimited additional passes with the ability to pause/replay at arbitrary speeds.

5. Limitations

Beyond the intentional limitations in scope of the problem (tackling American Sign Language fingerspelling recognition for a mobile keyboard use case), FSboard has a number

of areas for improvement in both data and modeling.

Our phrase set is a mixture of several relatively narrow domains, and even within those domains the phrases follow some patterns due to the way we synthetically generated them. An independent test set that is based on real queries, rather than being constructed from the same synthetic grammar as the training set (even if the phrases are unique) would give a more robust understanding of the dataset/model's performance and inform future data collection efforts. We have also observed some considerations specific to the mobile keyboard application that were underexplored in our collection because they are not typically important for fingerspelling. One example is capitalization. Signers only distinguish capital from lowercase letters in rare circumstances where it is contextually relevant, but capitalization is more important for text entry and should be elicited intentionally. Special characters and differences in punctuation like hyphens and underscores are also important but relatively rare/nonstandardized in signing generally.

In terms of modeling, while MediaPipe Holistic handles the vision aspects of fingerspelling recognition in a way that is performant on device, prior work has found limitations in the accuracy of current pose models [37] (though many of the failure modes relate to interaction between body parts, which is less relevant for fingerspelling). Future work should explore direct modeling of video input, as in B. Shi and Livescu [6, 7], but for the mobile on-device setting.

6. Conclusion

In this paper we introduced FSboard, the largest ASL fingerspelling dataset to date by a factor of >10x. Informed by a participatory approach that prioritizes real yet tractable needs of Deaf/Hard of Hearing users, FSboard focuses narrowly on a mobile text entry use case to enable signers to fingerspell short phrases and pieces of information as an analogue to faster text entry techniques (as gesture/swipe keyboards are to two thumb typing on smartphones) as opposed to speech recognition (for which full sign recognition would be the analogue). Our baseline achieves 11.1% CER on FSboard's test set (with unique phrases and signers) and degrades minimally with compromises to frame rate and body tracking that could help maintain real-time on-device performance. We hope that these results (or those achieved by future modeling work on the dataset) will prove high enough quality to be useful in practical applications, and serve as a stepping stone on the way to more generally capable sign language technology.

Ethics Statement

The signers who participated in the data collection for FSboard were each paid approximately $300 for providing 1000 fingerspelled phrases (which typically required 8-12 hours) and consented to their videos being published in a public dataset. Some participants were paid twice, once for the MacKenzie phrase set and once for the addresses, phone numbers, etc. phrase set. Our data collection procedure was reviewed by the relevant approval processes of our institution.

DPAN is a non-profit Deaf media company whose employees' primary language is ASL. They recruited and consented the contributors to the dataset. While names are not affiliated with any of the videos, it was clear in the consent process that participants' faces would be identifiable and that the dataset would be public.

While we release the underlying data with faces unblurred because mouthing can be important for fingerspelling recognition, we ask that dataset users blur the signers' faces when including examples in publications (as we do in this paper). Dataset users should not attempt to infer the signers' personal identities or use their likenesses to generate and publish other content (deepfakes).

Given the sociohistorical context surrounding sign language technology and perceptions of fingerspelling, it is important to emphasize that fingerspelling recognition/transcription *is not* sign language translation. Fingerspelling is an important part of ASL, but ultimately *just a part* of the language. Please do not exaggerate the scope of this dataset or task in any follow-up work.

Acknowledgements & Disclosure of Funding

We would like to thank all the signers who contributed their data to the project, and Michaela Jitaru and Nathan Qualls at DPAN for their contributions to data collection logistics. Thanks to the Kaggle team (Sohier Dane, Glenn Cameron, Mark Sherwood, Ashley Chow, and Phil Culliton) for all their advice on how to best create useful datasets. We also thank Chris Dyer and anonymous reviewers for giving feedback on drafts of this paper, as well as Caroline Pantofaru for institutional support. Funding for this dataset is from Google. DPAN is a non-profit.

References

[1] International day of sign languages, 2021. 2

[2] Deafness and hearing loss, 2022. 2

[3] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Watten-

berg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[4] Natasha Abner, Grégoire Clarté, Carlo Geraci, Robin J Ryder, Justine Mertz, Anah Salgat, and Shi Yu. Computational phylogenetics reveal histories of sign languages. *Science*, 383(6682):519–523, 2024. 2

[5] Manfred Georg Mark Sherwood Phil Culliton Sam Sepah Sohier Dane Thad Starner Ashley Chow, Glenn Cameron. Google - american sign language fingerspelling recognition, 2023. 5

[6] J. Keane D. Brentari G. Shakhnarovich B. Shi, A. Martinez Del Rio and K. Livescu. Fingerspelling recognition in the wild with iterative visual attention. *ICCV*, 2019. 2, 5, 6, 8

[7] J. Keane J. Michaux D. Brentari G. Shakhnarovich B. Shi, A. Martinez Del Rio and K. Livescu. American sign language fingerspelling recognition in the wild. *SLT*, 2018. 2, 6, 8

[8] Khushi Bhardwaj, David Martin, Rajandeep Singh, Gururaj Deshpande, Matthew So, William Neubauer, and Thad Starner. Popsignai: Using sign language recognition to improve american sign language learning in novice signers. *IMWUT (in submission)*, 2024. 3

[9] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, page 16–31, New York, NY, USA, 2019. Association for Computing Machinery. 1, 2

[10] Danielle Bragg, Naomi Caselli, Julie A Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E Ladner. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Transactions on Accessible Computing (TACCESS)*, 14(2):1–45, 2021. 2

[11] Danielle Bragg, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E. Ladner. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Transactions on Accessible Computing*, 14(2), 2021. 1

[12] United States Census Bureau. Tiger: Topologically integrated geographic encoding and referencing data (roads), 2019. 4

[13] Aashaka Desai, Lauren Berger, Fyodor O. Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E. Ladner, Hal Daumé III au2, Alex X. Lu, Naomi Caselli, and Danielle Bragg. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition, 2023. 3

[14] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro i Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language, 2021. 1

[15] Michael Erard. Why sign-language gloves don't help deaf people. *The Atlantic*, 2017. 2

[16] Jordan Fenlon and Erin Wilkinson. Sign languages in the world. *Sociolinguistics and deaf communities*, 1:5, 2015. 2

[17] Sakher Ghanem, Christopher Conly, and Vassilis Athitsos. A survey on sign language recognition using smartphones. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 171–176, 2017. 2

[18] Ivan Grishchenko and Valentin Bazarevsky. Mediapipe holistic - simultaneous face, hand and pose prediction, on device, 2020. 1, 5

[19] Sanjay Gulati. Language deprivation syndrome. In *Language deprivation and deaf mental health*, pages 24–53. Routledge, 2018. 3

[20] Matthew L. Hall, Inge-Marie Eigsti, Heather Bortfeld, and Diane Lillo-Martin. Auditory Deprivation Does Not Impair Executive Function, But Language Deprivation Might: Evidence From a Parent-Report Measure in Deaf Native Signing Children. *The Journal of Deaf Studies and Deaf Education*, 22(1):9–21, 2016. _eprint: https://academic.oup.com/jdsde/article-pdf/22/1/9/8675438/enw054.pdf.

[21] Matthew L Hall, Wyatte C Hall, and Naomi K Caselli. Deaf children need language, not (just) speech. *First Language*, 39(4):367–395, 2019.

[22] Wyatte Hall. What You Don't Know Can Hurt You: The Risk of Language Deprivation by Impairing Sign Language Development in Deaf Children. *Maternal and Child Health Journal*, 21, 2017.

[23] Wyatte Hall, Len Levin, and Melissa Anderson. Language deprivation syndrome: a possible neurodevelopmental disorder with sociocultural origins. *Social Psychiatry and Psychiatric Epidemiology*, 52, 2017. 3

[24] Saad Hassan, Abraham Glasser, Max Shengelia, Thad Starner, Sean Forbes, Nathan Qualls, and Sam S. Sepah. Tap to sign: Towards using american sign language for text entry on smartphones. *Proc. ACM Hum.-Comput. Interact.*, 7 (MHCI), 2023. 1, 2, 4

[25] Joseph Hill. Do deaf communities actually want sign language gloves? *Nature Electronics*, 3(9):512–513, 2020. 2

[26] Tom Humphries, Poorna Kushalnagar, Gaurav Mathur, Donna Jo Napoli, Carol Padden, Christian Rathmann, and Scott R Smith. Language acquisition for deaf children: Reducing the harms of zero tolerance to the use of alternative approaches. *Harm Reduction Journal*, 9(1):1–9, 2012. 3

[27] Tom Humphries, Poorna Kushalnagar, Gaurav Mathur, Donna Jo Napoli, Christian Rathmann, and Scott Smith. Support for parents of deaf children: Common questions and informed, evidence-based answers. *International journal of pediatric otorhinolaryngology*, 118:134–142, 2019. 3

[28] Trevor Johnston and Adam Schembri. Variation, lexicalization and grammaticalization in signed languages. *Langage et société*, 131(1):19–35, 2010. 2

[29] Jonathan Keane, Diane Brentari, and Jason Riggle. Coarticulation in asl fingerspelling. In *Proceedings of the North East Linguistic Society*, 2012. 2

[30] Taehwan Kim, Jonathan Keane, Weiran Wang, Hao Tang, Jason Riggle, Gregory Shakhnarovich, Diane Brentari, and

Karen Livescu. Lexicon-free fingerspelling recognition from video: Data, models, and signer adaptation, 2016. 2, 6

[31] Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals, 1966. 5

[32] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines, 2019. 5

[33] I. Scott MacKenzie and R. William Soukoreff. Phrase sets for evaluating text entry techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, page 754–755, New York, NY, USA, 2003. Association for Computing Machinery. 3

[34] Michele Miozzo and Francesca Peressotti. How the hand has shaped sign languages. *Scientific Reports*, 12(1):11980, 2022. 2

[35] Ross Mitchell, Travas Young, Bellamie Bachleda, and Michael Karchmer. How many people use asl in the united states? why estimates need updating. *Sign Language Studies*, 6, 2006. 2

[36] Ellis Monk. Monk skin tone scale, 2019. 5

[37] Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3434–3440, 2021. 8

[38] Donna Jo Napoli, Nancy K Mellon, John K Niparko, Christian Rathmann, Gaurav Mathur, Tom Humphries, Theresa Handley, Sasha Scambler, and John D Lantos. Should all deaf children learn sign language? *Pediatrics*, 136(1):170–176, 2015. 3

[39] Rob Neuhaus. Gibberish detector, 2014. 4

[40] Carol A Padden and Darline Clark Gunsauls. How the alphabet came to be used in a sign language. *Sign Language Studies*, pages 10–33, 2003. 2

[41] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. How do people type on mobile devices? observations from a study with 37,000 volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–12, 2019. 2

[42] Carol J Patrie and Robert E Johnson. *RSVP: Fingerspelled word recognition through rapid serial visual presentation*. 2011. 6

[43] Justin M Power, Guido W Grimm, and Johann-Mattis List. Evolutionary dynamics in the dispersal of sign languages. *Royal Society Open Science*, 7(1):191100, 2020. 2

[44] David Quinto-Pozos. Rates of fingerspelling in american sign language. In *Poster presented at 10th Theoretical Issues in Sign Language Research conference, West Lafayette, Indiana*, 2010. 2

[45] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. 5

[46] Shyam Reyal, Shumin Zhai, and Per Ola Kristensson. Performance and user experience of touchscreen and gesture keyboards in a lab setting and in the wild. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 679–688, 2015. 1

[47] Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgöz, and Jean Maillard. Towards privacy-aware sign language translation at scale, 2024. 1

[48] Sahir Shahryar, Manfred Georg, and Matthew So. Record these hands android app. 4

[49] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video, 2022. 5

[50] Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, Daksh Sehgal, Saad Hassan, Bill Neubauer, Sofia Anandi Vempala, Alec Tan, Jocelyn Heath, Unnathi Utpal Kumar, Priyanka Vijayaraghavan Mosur, Tavenner M. Hall, Rajandeep Singh, Christopher Zhang Cui, Glenn Cameron, Sohier Dane, and Garrett Tanzer. Popsign ASL v1.0: An isolated american sign language dataset collected via smartphones. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3, 5

[51] Garrett Tanzer and Biao Zhang. Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus, 2024. 5

[52] Garrett Tanzer, Maximus Shengelia, Ken Harrenstien, and David Uthus. Reconsidering sentence-level sign language translation, 2024. 6

[53] Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. Sign language translation from instructional videos, 2023. 1

[54] David Uthus, Garrett Tanzer, and Manfred Georg. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus, 2023. 1, 5

[55] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models, 2022. 1, 5

[56] Polina Yanovich, Carol Neidle, and Dimitris Metaxas. Detection of major ASL sign types in continuous signing for ASL recognition. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3067–3073, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). 2

[57] Shumin Zhai and Per Ola Kristensson. The word-gesture keyboard: reimagining keyboard interaction. *Communications of the ACM*, 55(9):91–101, 2012. 1