Universal Access in the Information Society
https://doi.org/10.1007/s10209-019-00668-9

LONG PAPER

Automated 3D sign language caption generation for video

Nayan Mehta[1] · **Suraj Pai**[2] · **Sanjay Singh**[1]

Abstract
Efforts to make online media accessible to a regional audience have picked up pace in recent years with multilingual captioning and keyboards. However, techniques to extend this access to people with hearing loss are limited. Further, owing to a lack of structure in the education of hearing impaired and regional differences, the issue of standardization of Indian Sign Language (ISL) has been left unaddressed, forcing educators to rely on the local language to support the ISL structure, thereby creating an array of correlations for each object, hindering the language building skills of a student. This paper aims to present a useful technology that can be used to leverage online resources and make them accessible to the hearing-impaired community in their primary mode of communication. Our tool presents an avenue for the early development of language learning and communication skills essential for the education of children with a profound hearing loss. With the proposed technology, we aim to provide a standardized teaching and learning medium to a classroom setting that can utilize and promote ISL. The goals of our proposed system involve reducing the burden of teachers to act as a valuable teaching aid. The system allows for easy translation of any online video and correlation with ISL captioning using a 3D cartoonish avatar aimed to reinforce classroom concepts during the critical period. First, the video gets converted to text via subtitles and speech processing methods. The generated text is understood through NLP algorithms and then mapped to avatar captions which are then rendered to form a cohesive video alongside the original content. We validated our results through a 6-month period and a consequent 2-month study, where we recorded a 37% and 70% increase in performance of students taught using Sign captioned videos against student taught with English captioned videos. We also recorded a 73.08% increase in vocabulary acquisition through signed aided videos.

Keywords Indian Sign Language (ISL) · Profound and severe hearing loss · Bilingual education · Oralism · Total communication (TC)

1 Introduction

India, a country with a population of 1.3 billion people, nearly a fifth of the world population [7], is estimated to have people with hearing loss of the order of 5 million [8]. According to the Government of India Disabled Persons Statistics Survey 2016 [35], 32.5% of this number is constituted of children. In the survey, for the age-group 5–9 years old, 209 of a sample set of 100,000 children and for the age-group of 10–14 years old, 212 of a set of 100,000 children have been found hearing impaired. A significant portion of this population, namely 32% of these children have a profound hearing loss, and 39% are diagnosed with severe hearing loss [35]. A child with hearing loss faces immense obstacles in the development of speech and language capabilities. Hearing loss limits the child's schooling, higher education and impacts future professional opportunities.

✉ Sanjay Singh
sanjay.singh@manipal.edu

Nayan Mehta
nayan.taurian@gmail.com

Suraj Pai
surajballambat@gmail.com

[1] Department of Information and Communication Technology, Manipal Institute of Technology, MAHE, Manipal 576104, India

[2] Department of Electronics and Communication Engineering, Manipal Institute of Technology, MAHE, Manipal 576104, India

Published online: 22 July 2019

Moreover, different methods of teaching used for the hearing impaired in India add to a lack of structure and approach with regard to overcoming this hurdle. Within India, there exist three prime methods of teaching the hearing impaired, namely Indian Sign Language, Oralism and Total Communication. Oralism [14] is the education of students with hearing loss through oral language by use of lip reading, speech and mimicking the mouth shapes and breathing patterns of speech. Total Communication (TC) [38] is an approach to the education of people with hearing loss that aims to make use of many modes of communication such as signed, oral, auditory, written and visual aids, depending on the particular needs and abilities of the child. Sign language, although a preferred mode of communication around the globe, is attributed to the least amount of usage of the three methods in India. Total Communication remains the most widely adopted methodology. The philosophy behind this technique is that it provides the child with multiple instances of modes to rely on. One of the drawbacks of TC is that it deprives the child of complex language learning (English or ISL) and combines both while teaching, which might attribute to confusion [38]. A majority of schools also follow the Oralism methodology that may not successfully aid education in cases of profound and severe hearing loss.

Research has established the childhood advantage for language acquisition is linked to efficient sign (word) recognition [20]. The learning of English among hearing-impaired children appears to benefit from the acquisition of even a moderate fluency in ASL [21]. Vocabulary is an essential part of the educational process which helps the students become proficient in English or American Sign Language. Skills in American Sign Language are linked to increased English literacy for children for whom access to spoken word is limited [9, 31]. There exists a direct correlation between English literacy and Sign language fluency for children with hearing impairment. When children repeatedly hear unfamiliar words throughout a story, their vocabulary recognition increases [18, 26]. However, Children with hearing loss are at a higher risk of decreased incidental vocabulary through stories being read aloud by adults [28]. In a field study carried out with pre-teaching and DVDs with ASL sign captions, a sufficient increase in early age vocabulary word acquisition in children with profound and severe hearing impairment was found [4]. A tool that can countermand manual repetition and move toward inculcating technology (digesting everyday media) for easy word–sign mapping and continuous repetition could create a powerful platform for students classroom/home learning.

Our team began with a 6-month empirical study in Padsad Karnabadhir Vidyalaya, Nashik, which led to an artifact contribution [39] through a sign language-based video captioning prototype. The prototype was validated at Sheila Kothavala Institute for the Deaf, Bangalore, where the expected impact of the artifact was consistent with our observations and inferences from the empirical study.

The initial research involved establishing an understanding of the sign language education scenario in India on the field. We interacted with stakeholders in the ecosystem—students, teachers, administration and executives— and delved into their pain points with the help of semi-structured interviews. After a comprehensive discussion using participatory design, we decided to pursue a tool which would act as a teaching aid that provides, textual and signage cues on videos and maps the understanding of students for the topics encountered. We then attempted to validate the need for such a system through A/B testing and carrying out an observational study [17] with the students where they were able to grasp better and retain information learned through the captioning offered by the system. The classroom environment was our target, so we conducted the field research at a school. Exploring other fields where technology can assist the hearing impaired might lead to more solutions in such spaces, which would eventually pave the way toward an all-inclusive solution.

The system thus operates in an educational setting and aims to improve communication and learning for the hearing impaired. Hearing-impaired children prefer to use signing as a method of communication with their teacher and peers as opposed to oralist methods or spoken language [12]. Our field research also bolstered this inference in studies conducted with teachers of children with hearing loss. Despite this, an inclination toward adopting oralism exists in schools, which might be owing to multiple reasons such as:

1. Hearing parents of children with hearing loss prefer oralist methods to assimilate children into their household;
2. Limited resources in sign language for the community. For example, ISL proficient teachers are limited; ISL resources are not widespread. Problems in mainstreaming owing to non-sign accommodating society infrastructure.

We propose a technology-supported solution to bridge the resource gap among people with hearing loss (in ISL—their preferred mode of communication). The proposed system leverages already existing educational videos online and provides sign captioning available during the run of the video. The purpose of our platform is to create an interface that serves content in the primary language of the hearing-impaired community that makes it easy to correlate mappings and collectively form an efficient system for learning and evaluation of young students during their language building years.

The system is built around a database of 3D generated signs that act as the sign captions for the video. Subtitles or speech processing is used to infer the audio content of the

video, and it is then sent to the Natural Language Processing module which has a Subject–Object–Verb rule-based grammar, and sentences are converted to this format; eventually, the video is overlayed with the 3D sign captions.

We further evaluated this system at a different school in Bangalore, Sheila Kothavala Institute for the Deaf. A phase-wise trial was conducted with ten children of profound and severe hearing impairment splitting them into two groups. A pre-evaluation test was held to gauge the knowledge of the students followed by a comparison between learning and understanding between the two groups. The evaluation metric used was percentage improvement in the control versus experimental group from the pre-evaluation results. We recorded a 37% and 70% increase in performance of students taught using Sign captioned videos against students taught with English captioned videos. We also recorded a 73.08% increase in vocabulary acquisition through signed aided videos.

We intend to freely distribute this tool to the schools and set up the environment for daily usage.

This research work is part of a validation study. For the validation study, we have chosen two schools for students with hearing loss, one in Nashik and another from Bangalore, India. We chose the school for the validation study so that we get a representative population for this study. Since this research work is based on a validation study, we have not sought any Institutional Ethical Committee (IEC) approval, which we will seek for the effectiveness study in the future.

2 Related work

2.1 Commercial products for people with hearing loss

Efforts to aid the hearing impaired have long since been focused on communication through translation of signage. A few leading commercially available translation tools from around the globe are discussed here. The 'HandTalk Translator' Application [34] converts Brazilian Portuguese audio to Brazilian Sign Language. This product is market ready and available on the Google Play Store. It uses an interactive avatar with facial expressions and fluidity. However, it aids only one part of the communication from the non-signer to the signer. MotionSavvy's 'Uni' [24] leverages leap motion technology to convert audio and vice versa to American Sign Language. Currently in R&D phase, its interactions are limited only to hand movement and do not take into account facial expressions as a part of signage.

Platforms and resources to ease communication with the hearing impaired have also been developed and maintained by various groups and organizations who promote Sign Language Learning. Within India, 'Talking Hands' [33] a web-based platform provides an extensive dictionary of Indian Sign Language. It is used for educational purposes and interactional videos to develop skill sets. However, this platform lacks responsiveness and contains a limited subset of commonly used signs for language development. 'Ramakrishna Mission' [25] provided the first and most widely adopted online resource for ISL signs but presented a limited user interaction and a lacking vocabulary set of signs. 'SignTalk' [32] acted as a relay service between the signer and the interpreter and was by far the most evolved system to serve people with hearing loss in India. The platform, however, required paid interpreters and necessitated the need for an interpreter to be present to relay the request. This service is no longer operational. Coming to the most straightforward mode of communication, to manually hire a sign language interpreter, due to the need and lack of ISL teachers still remains a constant struggle and unaffordable circumstance for most coming from impoverished neighborhoods.

2.2 Experimental/research work in sign language recognition

A large set of papers focus on sign language recognition for alphabets and numbers and consider them as isolated. One approach was to generate a depth and motion profile for each sign language gesture and use the feature matrix thus generated with a multi-class SVM classifier [1]. Agarwal and Thakur [1] used both depth and motion information to create their feature vectors which allowed to capture motion relationships as well. Another approach by Lang et al. [16] presents Hidden Markov Models to recognize signs based on Kinect input data. These approaches generally focus on higher-level features due to the use of Kinect which does not provide finger-level predictions. It would be far from ideal in the real world where signs are not isolated and have much contextual information where such methods would fail to perform.

Experimental- and prototype-based solutions such as glove-based wearables have been used quite extensively across the world. Wearable gloves that teach sign languages such as GyGSLA [30], gloves that help communicate sign language alphabets, and gloves that help with words and sentences exist. Much research has been done with glove-based and sensor-based methods, and they exist mostly in the prototype phase. However, glove-based gesture recognition requires that the user wears a cumbersome data glove to capture hand and finger movements. It hinders the convenience and naturalness of human–computer interaction. The limitation faced by this approach is the inability to obtain meaningful data complementary to gestures to give the full meaning of the conversation, such as facial expressions, eye movements, and lip-perusing [2]. Further, these

are superficial in accounts of the construct of Indian Sign Language as with an adaptation of regional dialect for grammar support.

Much experimental research has also been carried out in this field with deep learning methods breaking through previous accuracy barriers. Convolutional Neural Network, Hidden Markov Model-based approaches have found high precision levels on popular sign language datasets [42]. These approaches use RGB, RGB-D, and video stream data to identify sign language gestures without imposing the constraints from glove-based methods such as a 3DRCNN-based method proposed by Ye et al. [40] which uses three-dimensional CNN's along with Recurrent Nets to recognize sign language from RGB, depth and motion data. Research by Huenerfauth [13] features a novel project for the machine translation of English to ASL and identifies the need to produce classifier predicates in user interface applications for the deaf. Another interesting deep learning-based approach investigated by Ahmed [3] uses Sequence-to-Sequence LSTMs to generate captions from videos of American Sign Language, and another study evaluates the usability of Automatic Speech Recognition (ASR) technology to be used as a real-time captioning tool [15]. These methods offer a promising avenue for developing communication systems to bridge the gap between the hearing and hearing impaired.

From a large number of commercially available products and prototype-based solutions, we infer that much work usually tries to establish a single solution to solve the problem of communication between a hearing-impaired person and a hearing person. During our field research, we observed that such a solution in the current context might not obtain the desired levels of accuracy. Another factor that would render these approaches hard to implement is the difference in Sign Language modalities used by different schools across India. Some schools might use variations of ASL; some even prefer to use their signs based on other popular Sign languages. However, targeting specific environments where not just communication but also learning can be selectively improved by imposing environmental and contextual constraints might be a more feasible way to develop technology-based assistance systems. Specific environments would limit the sample space of possible interactions, and a system could be developed using a subset of sign language to be adept at assisting in that particular scenario.

2.3 Role of sign language in the development of children with hearing loss

A study by Mayberry and Eichen [20] found that language acquisition is not unique to speech alone and in hearing-impaired children is linked to efficient word-to-sign recognition. Further, the subjects of this study based on their age of acquisition had a higher and faster level of linguistic understanding particularly in grasping sentence meaning. Another study of the education of people with hearing loss in Italy [23], linked learning of sign language as an essential factor in the cognitive advancement of hearing-impaired children. A study to increase vocabulary using DVDs [4] was also found to be effective when accompanied by pre-teaching of signs from teachers, thereby proving to be a useful aid for teachers in a classroom setting. Further systems that utilize computer animated tutors recorded an increased level of vocabulary acquisition in hearing-impaired children [6, 19]. Vcom3D software that incorporates avatars as a part of educational lessons, recorded an increase in comprehension of a story, from 16 to 67% after watching Signs (vs. being read) for a study conducted at Florida School for the Deaf and Blind [37].

3 Materials and methods

The Indian population lacks awareness of education of people with hearing loss, activities, language and many other aspects. This is credited to the limited interaction between hearing people and people with hearing loss due to communication barriers, interpreter inefficiencies, and possibly cultural differences. To formulate our research statement, interaction with people from the community was integral. Culture and society among people with hearing loss vary across different regions of India, much like the hearing populous. Interactions and observations with the children in a hearing-impaired school led us to carve out our system. The following include the insights and gatherings of our in-person research work carried out over the six months with Padsad KarnaBadhir Vidyalaya, Nashik, Maharashtra, India.

3.1 Initial study design

Our initial fieldwork included regular visits, interactions and semi-structured interviews with hearing-impaired students, teachers, and parents at Padsad Karnabadhir School, Nashik, India. The team visited a list of hearing-impaired schools in the area and decided to choose Padsad as they were among the few schools that chose to use signage as a medium of instruction (with the Total Communication Method). Many schools relied on Oralism only for teaching, and, due to this, we were not able to involve these schools in our study.

Based on the previous research conducted on Sign Language communication systems, we offered to build an initial prototype in participatory design [17], a system that converts Indian Sign Language to audio and vice versa using Microsoft Kinect v2 that is, as a communication aid. However, following multiple visits to the school, we found that such a tool given its limitations such as the cost of having a Kinect device in every classroom along with processing

power for each device, might not be beneficial. The schools also acknowledged that such a system might not be among their priorities due to their focus on imparting lingual skills to the students over communication with hearing people.

In the multiple visits we:

1. Interacted one on one with the teachers discussing their routine preparation for teaching a particular topic/lesson to the students. The day-to-day routine of teachers started far before the actual classes, where they went through the entire syllabus as prescribed by the State Board and outlined the critical aspects needed to understand the lesson, after which teachers went back to the drawing board and noted down all this line by line. We noted the typical process of how a teacher prepared for a particular topic and compared it against how the same is done in a mainstream school;
2. Participated in classroom settings where we observed how students consumed information from the teachers. This allowed us to establish a proper understanding of how information flows in a classroom and different complications and intricacies involved. During class, teachers would point out word by word and further explain using gestures to the students in the class; compound words are broken down to first the spellings and more straightforward concepts more relatable to the class. Each class required manual repetition, sometimes as much as 30 times weekly for a five-worded sentence. During Classroom sittings, we collected data on how each student was responding to different topics, their proficiency with the Signing, answers to questions furnished by the teachers and finally overall grasp of a particular topic;
3. Observed and spoke to students and understood how they learn topics taught in a classroom. Students here were mostly from impoverished backgrounds, often found it hard to get back into sign-based communication, given their parents at home used different signage (some signs that were used to associate everyday activities and words). Most students' parents have menial jobs and found it hard to dedicate time to their child's language and cognitive development.

3.2 Observations of preliminary study

The observations were utterly different from what we had expected based on our reading of literature and limited knowledge of hearing-impaired education in India. One of the major surprises was the inconsistency in Sign Language usage and the idea of Total Communication which combines different methods and allows the students to work with what they can best understand. Exposure to this methodology and the current scenario of classroom teaching led us to re-evaluate the need for a Kinect-based communication system. Based on our classroom settings, we observed that the process of knowledge transfer between the teacher and student could be eased using technology. Although many tools exist to make learning and education more comfortable for the mainstream audience, these tools have not penetrated the space for people with hearing disabilities in India. A school teacher had to repeat the same topic a number of times for every student to have a complete understanding, which is effort and time-consuming. It has worsened by the fact that the student did not have another method to understand the topics on their own. The curriculum followed was the same as regular schools, and where students from mainstream schools have the opportunity to try and understand these lessons on their own, these students were not able to do this due to the lingual barrier.

3.3 Inferences of preliminary study

After gathering these observations, we sat down with senior teachers and decided on an interface that could reduce the workload on the teachers and students. An interface that would help children watch videos without the teacher having to interpret and convey context for every single sentence was the proposed target by the teachers. It was directly derived from our observations during classroom settings. During this interaction, we proposed a system that would overlay the video of signs from an openly accessible dataset and synchronize with the target video. We wanted to incorporate as many features of the daily classroom teaching as possible into the system, considering the number of regional languages in India, such a system in only one language would barely be useful. We identified multiple issues with adoption due to factors like costs, value addition, etc. Although the proposed system adopts some aspects of the total communication methodology, it promotes sign language primarily and makes it easier for classrooms to use signage. Currently, no such learning methods exist in India or elsewhere that convert videos to sign language automatically. Often, these require interpreters to add and record this signed content to videos online manually.

3.4 Procedure and participants of subsequent study

The next phase included in-person interactions with children with different ability levels of partial, severe and profound hearing loss with video content to evaluate the inclusion criteria [17] related to the severity of disability and performance. This phase was set up with the help of a professional audiologist, who guided us on setting the right environment and conditions for this test. We asked teachers to select the highest performing students in their class and observed them

while they were asked to understand and explain videos with subtitles. It helped us gain an insight into comprehension and knowledge gathering among the children, who followed lessons sentence-wise and thus content from the videos as well. Their speed of grasping information was a lot slower than the pace of the video, and they relied heavily on the subtitles and attempted to break all texts down to base signs for understanding. We recorded our observations and designed specific components of the system in accordance which is highlighted in Sects. 4.3 and 4.4.

4 System implementation

4.1 Implementation approach

The system proposed in this paper mimics the assimilation process of the people with hearing loss in consuming mainstream media outlets (television, events, speeches, conversation, etc.) as outlined in Fig. 1. Information exchange occurs through a sign interpreter acting as a medium between the two parties (two-way communication) or simultaneous sign translation for one-way communication. In our paper, we focus on the second case where every new source of media needs to be translated for assimilation. It renders a plethora of content inaccessible by sheer virtue of lack of translation medium. The necessity of human intervention in this process forms a barrier to learning and communication; automating this has witnessed multiple attempts, though they are by no means comprehensive. Such efforts need to incorporate multiple modalities to emulate a human interpreter.

The operational flow outlined in Fig. 1 relies heavily on the human interpreter, integral at each step. Due to its nascent state, Indian Sign Language (ISL) still lacks basic grammar constructs. To accommodate this, ISL is adopted across India by using ISL signs and local grammar structure. As a result, the interpreter is expected to be context and region aware in order to sign efficiently. There is also a need for sincerity and diligence on the part of the interpreter. Such a dependency on the interpreter who may be susceptible to errors may affect the quality of information received at the destination. The interpreter may also introduce bias which could lead to differences in information consumed at the endpoints.

The system proposed in this paper paves the way for the reduction in dependency on the interpreter and establishing standardization in the process of information exchange. Another significant advantage offered is the ability to consume mainstream media privately.

The pipeline implemented in our algorithm derives from the operational flow highlighted earlier.

As shown in Fig. 2, the input to the pipeline is in the form of YouTube videos or audio sources. The pipeline renders a smooth animated video at the output end. The pipeline can be broken down broadly into these sub-modules described below.

1. Information Entry: Information entry refers to the entry sub-module where the input source is given into the pipeline; this input can be speech and subtitles (from YouTube videos/audio).
2. Information Processing: Information from the entry sub-module is processed to gain textual information. Speech

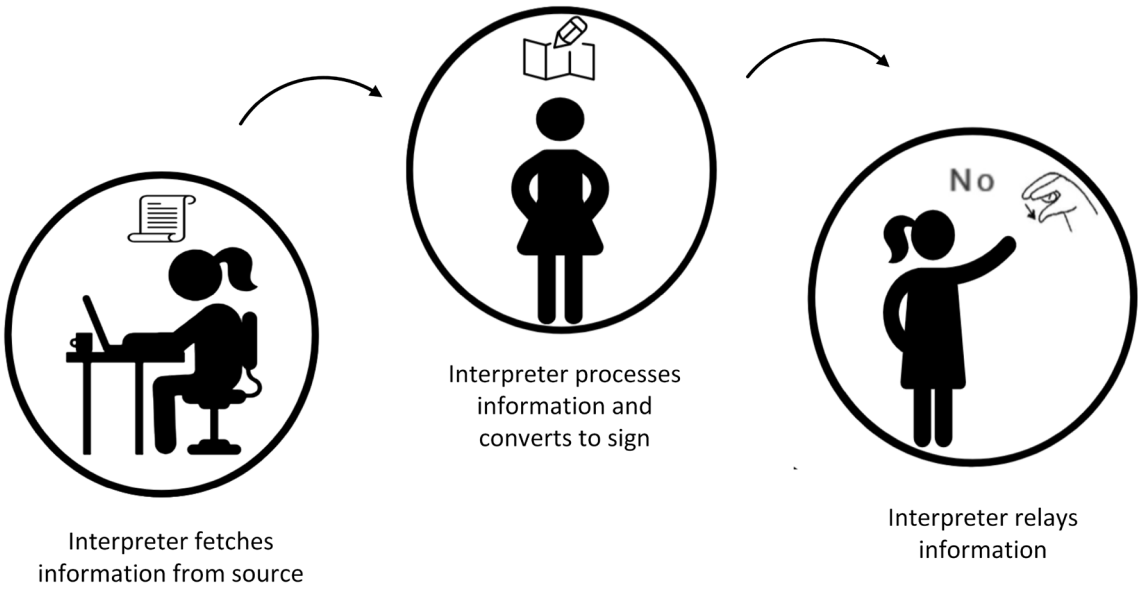

Fig. 1 Assimilation of information from mainstream media (Operational Flow)

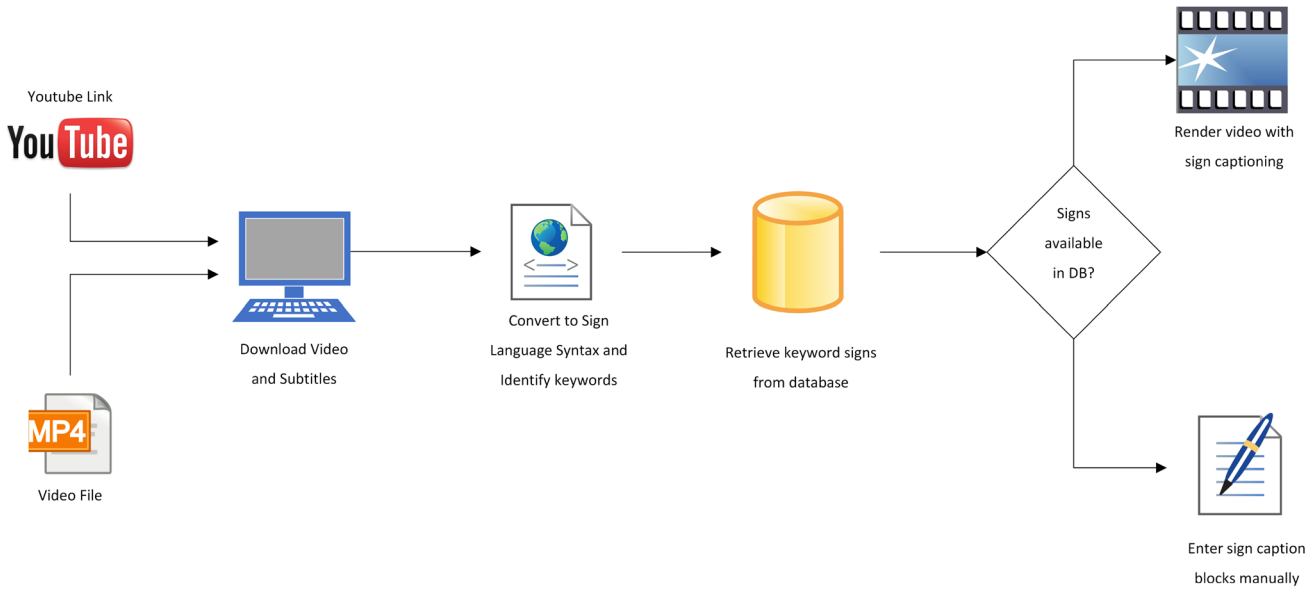

Fig. 2 Logical workflow of modules

input is converted to text and merged with subtitle information to obtain the best match representation of the data.

3. Information Understanding: The pipeline now possesses text representation of the input. At this point, we try to capture the context and message via Natural Language Processing. This understanding relies on the structure of the intended output media. In this case, we look at the Indian Sign Language structure and extract elements accordingly.
4. Mapping: Our information is resonant with sign language structure and can now be used to generate relevant output form. The mapping takes places between identified words and context to 3D avatar gestures stored in our database. This avatar gestures database is central to our application as it acts as an interface to exchange information.
5. Interaction: Interaction is the interface between the input entry and consumption of output. An overlay container on videos provides a closed captioning equivalent to the end user. This interaction container can be re-sized, paused or sped to provide maximum control of the method of understanding. This feature is based on a study we conducted on learners with hearing loss representing different modes of grasping information.

The pipeline provides a complete end-to-end solution to process video or audio input and is central to the avatar database dictionary. A full and populated database could provide a large output to bolster understanding among the end users. The interaction first starts with the user logging onto the portal and pasting the link of the desired video that requires translation; the system then consolidates a file of pre-recorded avatar-based signs to form a single file loaded from the back end. The web portal also allows teachers to record signs of words unrecorded from previous videos, which are translated later by the back-end team upon validation. An interactive application renders signs for the general educational purposes from the now-populated database of sign mappings.

4.2 System components

Each of the modules involved in the pipeline is mutually exclusive and shares only the previous output to the next input relationships.

4.2.1 Information entry module

YouTube videos or audio sources contain a plethora of information that with due processing can be made accessible to the hearing-impaired community. With YouTube gaining popularity, content is offered in a lot of different languages via community-contributed captions or auto-generated captions. These captions generally provide an accurate representation of oral or visual information in the media. This module looks first for user-captioning and, if unavailable, proceeds to store auto-generated captioning from YouTube (L1). This module continues to gather speech/audio information from the video (L2). In the case of non-YouTube media, only the second step applies. This priority for user-captioning helps increase the accuracy of later stages in the pipeline. Speech is stored in a standard format as required by most transcription APIs. Once the user has entered the

video link in the text entry area, we use the YouTube Data API to fetch the captions list and subsequently the captions in the required language through an API call. We prioritize user-entered captions and then fallback to auto-generated captions if the user-entered captions are unavailable for the video. We fetch the captions in the WebVTT format with relevant timing information. If the user-entered captioning is not found, we also pass the audio from the video to Google Speech to Text API and Microsoft's Cognitive Speech Services and create a timed transcription.

4.2.2 Information processing module

The processing module functions to create a consumable form of data for the Understanding module. A modality that can be processed by mainstream algorithms such as text is the target for this module. In the L1 phase of information entry, the auto-generated captioning is used as the text form. Consequently, in the L2 phase, the audio captured is converted to text using Transcription APIs. On completion of L1 and L2 processing, we converge the information contained in both. L1 captions and vice versa solve inconsistencies in the Speech API. A weighted procedure is used to make an appropriate target from L1 and L2 processing outputs. Sentence structure and words from L1 when user annotated are taken as high confidence, and L2 is skipped in this scenario. In a situation where L1 is auto-generated captioning, we use L2 to reinforce the confidence of assumptions made by L1. This reinforcement is handled through a check for intersection of words along different intervals of time. If a word occurs in both L1 and 2 APIs in the L2 mode at a particular time interval, it is considered high confidence with a score of 1. We assign a confidence score relative to a number of intersections. We thus create a parallel chain of words at each time interval with respective confidence scores through L1 and 2 APIs from L2. Figure 3 represents this chain with probabilities assigned to each word.

4.2.3 Information understanding module

Information is now in the form of text that can be interpreted and understood by NLU systems. We used Python's NLTK libraries to gather relevant subsets from the text corpus. These subsets depend on the structure of the target—Indian Sign Language. Indian Sign Language is highly context oriented and emphasizes the subject. Grammatical nuances of the English language such as conjunctions and pronouns are not highly used in ISL. ISL sentence structures are mostly SOV (Subject–Object–Verb) [41]. The reasoning and variations of this is another area of research in ISL. The SOV structure helps us gain an understanding of the lingual constructs that our Understanding module needs to be aware of. We use NLTK's POS taggers (Part of Speech taggers) from the Penn TreeBank Dataset [27] to infer different parts of speech from the chain of words. Our algorithm identifies what parts of speech are essential for the next module by following the SOV (Subject–Object–Verb) model. First, we use the different parallel word chains from Fig. 3 as tokens and perform POS tagging on these words. We extract the Subject, Object and Verb from these group of words through noun and verb forms and use them to form a sentence in the SOV form [11]. In case multiple subjects, objects or verbs are found in the parallel word chains, and we select the ones with the highest intersection scores. These SOV sentences are then sent to the mapping module to comprehend.

4.3 Mapping module

4.3.1 Modeling

Gestures are mapped onto a 3D modeled avatar, (see Fig. 4), that captures each word-to-sign correlation. Microsoft Kinect v2 Plugin on the iClone software was used to map all hand gestures and body postures. Finger intricacies and movement were captured manually. Avatars were chosen based upon cartoonized aspect to grasp the attention of

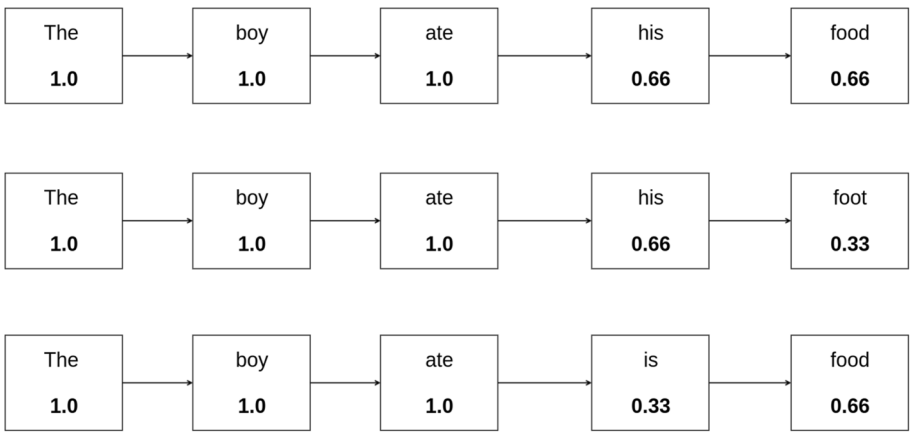

Fig. 3 Word chains for a sentence from L1 and 2 L2 APIs

Fig. 4 3D child avatar

children, big eyes, natural characteristics, i.e., someone that students can relate to as a guide. Each word is mapped to a gesture, and each gesture settles down to a base position making the transition between words seamless and each sign start afresh from the canvas. Gestures were recorded based on a Solar System video for grade 6, as presented in Fig. 5, Science lesson, based on the .srt file downloaded, and all words with mappable Signs were saved onto the database, thereby populating signs specific to translation.

We incorporated multiple features were into the UI based on classroom interaction and research. For instance, a replay button was added to repeat the last word and the previous sentence, and this was done to reinforce learning and also tackle the problem of repetition faced by numerous teachers. Subtitles were included in the regional vernacular language to directly correlate word to sign the sign appearing (Marathi). Further, the signed caption window was made detachable and re-sizable for convenience, and an easy to use interface was implemented to allow increased usability. The system also allows for a teacher module, to map unrecorded words and recording signs for these words, thereby, populating the database as and when needed.

4.4 Evolution of the system

Based on the observations from our field research, we started with assembling a video corpus which comprised of NCERT Class-wise video lessons used by teachers for a thorough understanding of the texts [22]. Post-collection of the corpus, we designed the architecture of the proposed system. To illustrate the proof of concept (Fig. 6) to the school authorities, we used pre-recorded signs and embedded it as per the word occurrence in a YouTube video of 'What a wonderful world by David Attenborough.' After a demo and comprehensive feedback session, we decided to add multiple features such as speed control, repeat sentence/sign and animated avatar-based sign captioning to boost attentiveness which led us to work toward our final proposed system.

5 Data collection procedure

A complete audio or video to sign caption pipeline allows the hearing-impaired user to understand and infer content from the input source. In the event of an important video where the user may not be familiar with the signs, content from the video allows mapping of signs to contextual visual information from the video, helping with sign education. For a video that may not possess a plethora of visual or sign infer-able cues, the sign captioning helps with understanding and learning video content. The collective performance of the system can be justified as learning and understanding.

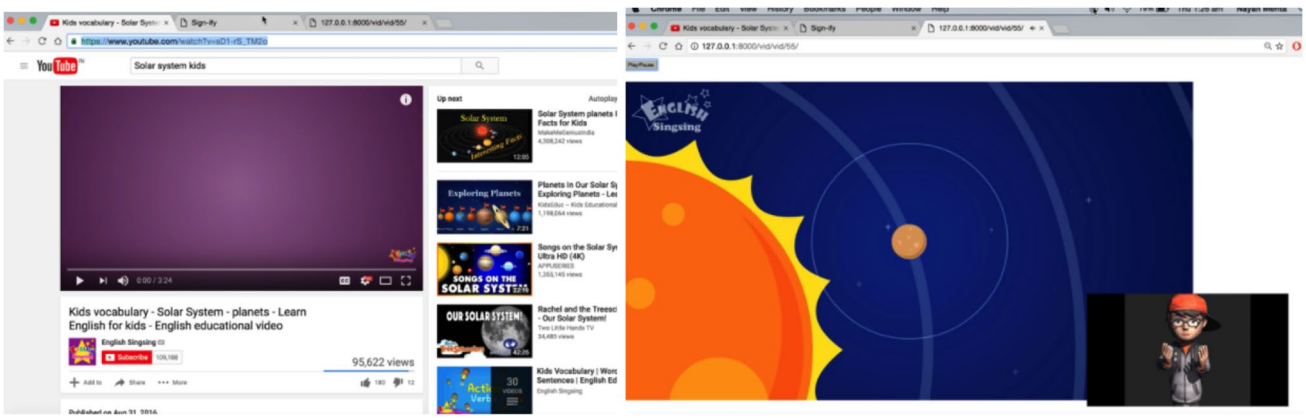

Fig. 5 WorkFlow, final translation of a solar system video with 3D signed avatar

Fig. 6 Translation of 'What a wonderful world by David Attenborough' (Rough Prototype)

Details on program implementation and measurement on the progress of the program are described in what follows.

Phase 1: Baseline Phase

1. Sit with the participant at a table and ask them to try to sign any words they recognize on the flash cards. Explain that if they do not know the word it is fine to guess or we can skip to the next card;
2. Show the participant one flash card. If the participant signs it correctly, place it in a pile to the right otherwise place it in a pile to the left;
3. At the end of the session, praise the participant for their participation. Then, allow them to return to class;
4. Record a check for the cards in a pile on the right onto the data collection checklist for that participant;
5. The purpose of this phase is to gauge the words known by the students.

Phase 2: Understanding of content through sign captions

1. Divide participants into two groups randomly based on the evaluation in the Baseline phase;
2. One group is shown video content with English captioning, and another is shown the same with added sign captions. (Video contains words whose signs students are familiar with, as gathered from the Baseline phase);
3. Video context-based questions are presented to the student, and answers are documented in an organized, predetermined approach;
4. At the end of the session, praise the participant for their participation;
5. This exercise is to be performed over multiple videos to compare understanding levels with/without sign captioning.

Phase 3: Learning of content through sign captions

1. All students are included for evaluation in the Baseline phase;
2. The students are taught signs for previously unknown words and taught new signs through illustrative videos with sign captioning;
3. Sign understanding-based questions are presented to the student, and answers are documented in an organized, predetermined approach;
4. At the end of the session, praise the participant for their participation;
5. This exercise is to be performed over multiple videos to compare learning levels of signs with/without sign captioning.

The Baseline phase is used to estimate the current learning level of a child: A primary researcher first records if the students can correctly identify signs A–Z alphabets, 1–10 numbers and A–Z words. In the next phase, Understanding phase, we estimate the understanding level of the above signs, to evaluate the effect of the tool in a similar set of signed captions versus English captions. The class was divided into two sections; one was shown English captions videos, and the other half was shown Sign captions videos for A–Z alphabet and 1–10 numbers videos. After that, the students were given an MCQ quiz for A–Z signs and 1–10 numbers to gauge their understanding of these signs; the

questions required knowledge of the order of alphabets and numbers, missing letter or number. The quiz consisted of Signed questions for the Signing group and simple English text for the English captioning group.

In the final phase, Learning phase, we estimate the level of active learning of vocabulary words by watching an A–Z words video with Signed captions; this was carried out for all ten students. Along with the video, the researcher also paused and pre-taught finger spelling for each word along with an emphasis on the associated sign. Then, a quiz featured identification of words by looking at the signs alone; the quiz also consisted of similarly spelled words in the options to ensure the accuracy of learning and a higher difficulty level. Researchers ensured no feedback was given to the participant to avoid any impact on the participant's performance that might lead them to exhibit demand characteristics [17].

All of the above data were recorded by the primary researcher using flash cards in 30-min sessions every consecutive day over two months. A second researcher simultaneously and independently carried out the same steps using Google Forms to correlate the data, which yielded a 100% accuracy of results thereby maintaining procedural fidelity.

6 Results

By using the proposed methodology, a phase-wise trial was conducted with the prep class at Sheila Kothavala Institute for the Deaf [29], with a sample group of ten research participants having a profound and severe hearing impairment. At the time of the study, Sheila Kothavala Institute for the Deaf followed the curriculum by the Karnataka School Board and taught signs using ASL and ISL. A pre-screening of the system was carried out by the school for a demo video to learn foods consumed by farm animals before the permission for our case study was granted. All the staff employed were trained in Deaf Education. A–Z alphabet, 1–10 numbers and A–Z words signed animations used for the study were referred from the National Association of the Deaf and issued database by Gallaudet University [5] and personally verified by teachers from the school. After the trial, graphs were plotted to record both pre- and post-assessment to evaluate the understanding and learning of signs.

Our evaluation was carried out in accordance with a previous study used to record the effectiveness of DVDs as a tool, presented in American Sign Language to increase the vocabulary recognition of hearing impaired or hard of hearing children [4]. The study was carried out over multiple baseline designs over three sets of vocabulary words. Figure 7 represents both the Baseline and Learning phases of the ten students for signs of A–Z Alphabets. A relative comparison of understanding levels of these signs can be gauged

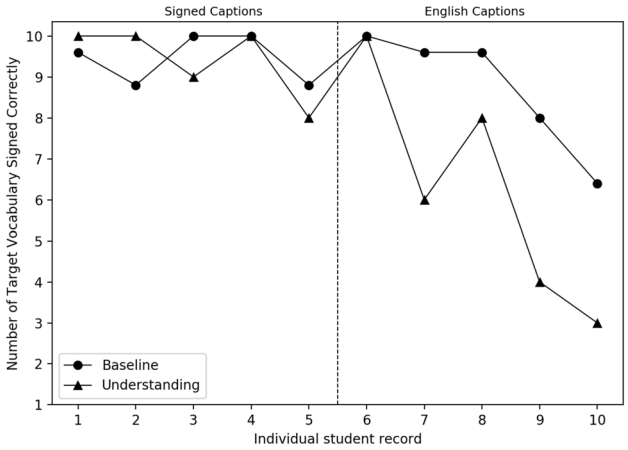

Fig. 7 Baseline and Understanding phases for A–Z alphabet signs

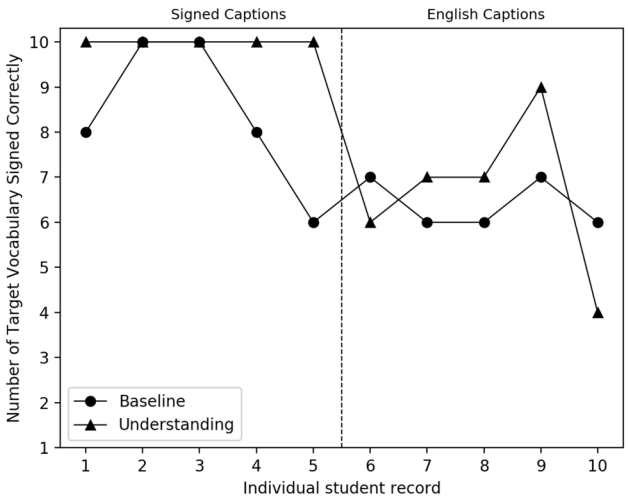

Fig. 8 Baseline and Understanding phases for numbers 1–10 signs

by the difference in baseline and learning data points. Children with signed captioned teaching videos for A–Z alphabet signs fared better than children with A–Z English captioned teaching videos during the assessment of the understanding level of signs. Children taught via signed video also recorded a higher accuracy and were much quicker during their assessments, whereas children taught through English captions would take time to decipher the text through signs and relatively struggled during evaluation. The baseline scores of 26 alphabet signs were normalized to a grade point of 10, which was also the number of questions asked in the quiz.

Figure 8 records the data entry points of Baseline and Learning phases for 1–10 number signs. A relative comparison again in the teaching methods pre-assessment reveals that children with signed captioned teaching videos fared better than children taught with English captioned 1–10 numbers video. The baseline score computed out of 10 for

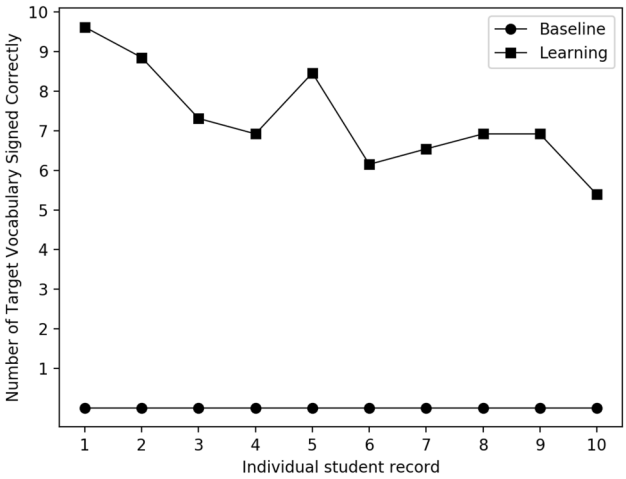

Fig. 9 Baseline and Learning phases for A–Z words signs

numbers (1–10), and the quiz of 15 questions was scored and normalized to 10.

Finally, Fig. 9 depicts a comparison of results from the Learning phase with the baseline, representing A–Z word–sign knowledge of all students of the class. The students had no prior knowledge of the signs, hence the flat-lined baseline. The learning curve represents the acquired vocabulary words in just 3–4 sessions of watching a signed captioned video of common A–Z words. After that, students were quizzed out of 26 questions normalized to a grade point of 10. All students see a significant increase in vocabulary acquisition on incidental exposure to signed captioned videos.

7 Discussion

The phase-wise trial conducted at the Sheila Kothavala Institute for the Deaf [29] and the various insights gained from Padsad Vidyalaya, Nashik [36], helped us identify that hearing-impaired children learn and understand better when sign language is involved. The trials centered around the ability to sign language to contribute to better learning and understanding outcomes in the children. As found in Justiceś work [18], language learning of children appears to benefit when novel words are introduced in a simple setting as opposed to complex prompts. Additionally, previous studies verify increased vocabulary recognition on repeated usage of unfamiliar words [18, 26]. As in other studies [4, 10, 18, 26], children in our study recorded an increase in their vocabulary recognition by participating in repeated probes in which the target vocabulary was repeated. For the understanding-based trials, the children were divided randomly into two groups. The first group was assisted with sign language captions, while the second group was exposed only to English captions. Table 1 shows the outcomes regarding score averages for the understanding tasks. *UAS denotes Understanding Average Score, and BAS represents a Baseline Average Score.* The BAS score gap gives us the baseline performance difference between the two groups, and the UAS score gap gives us the understanding of task difference between the two groups. The BAS score gap gives us the base difference between the performance of the two randomly selected groups regarding a standard test conducted to test their knowledge regarding understanding or learning. The UAS score gap gives the difference in the performance of the same two groups regarding the understanding tasks presented to them. A small BAS difference with a high UAS difference would suggest that the group with the higher UAS performed significantly better than the other group on a particular set of tasks. In our case, the particular set of tasks were performed using two different approaches, and a high UAS for one approach would give strength to that approach as students with similar baselines were able to perform better using one of the two approaches. In the A–Z task, for a small BAS difference, there is a huge UAS difference in favor of sign assistance. For the 1–10 task, the BAS difference is quite large between the two groups which may translate to a large difference in the UAS difference as well. However, all the students attain a perfect score of 10 with sign assistance, whereas the same improvement is not seen for the Pure English group.

In the task of understanding A–Z, the first group had a *Baseline Average Score* (*BAS*) of 9.44 and the second group had a baseline average of 8.72. The *Understanding Average Score* (*UAS*) of group 1 was 9.4 and for group 2 was 6.2. There was an understanding average score difference of 3.4 in favor of group 1 for a baseline difference of 0.58. Similarly, for the number understanding test, the BAS for group 1 was 8.4 and for group 2 was 6.4; the UAS for group 1 was 10, while the UAS for group 2 was 6.6. For a baseline difference of 2, there was a difference of 3.4 in favor of group 1.

From both these understanding trials we infer that students were able to understand better and answer questions related to A–Z and numbers between 1–10 when assisted by sign captions.

Table 1 Understanding outcomes

Task type	Group type	BAS	UAS
Understanding A–Z	Sign assistance	9.44	9.4
	Pure English	8.72	6.2
	Score gap	0.72	3.4
Understanding 1–10	Sign assistance	8.4	10
	Pure English	6.4	6.6
	Score gap	2.0	3.4

The latter part of the trial involved collecting data on how students could be aided in not just understanding but also learning itself using sign assistance. The students were initially unaware of the words with a BAS score of 0 which meant not knowing any of the words associated with each later was then able to identify the word based on the spellings when provided with sign cues. Their average learning score for this task improved by 7.308; in just a mere six sittings, we noticed a significant increase in vocabulary acquisition of students. Through this, we can infer that students can learn effectively using signed assistance in a classroom setting.

8 Conclusion

Our research puts forward the need for sign assistance to aid with understanding and learning for students with a hearing disability from an early age onward. From our various trials, we conclude that sign assistance helps the children learn, remember and understand the content better. It is also widely supported by the previous research which outlines the benefit of sign language for understanding and learning language and grammar. Based on these observations, we propose a system that can ensure life-long learning by providing sign assistance to the hearing-impaired students from consuming mass media such as the likes of YouTube. The system is scalable and easy to use in a classroom setting which can make a beneficial addition to boost the knowledge base of students who otherwise find it hard to understand and learn content outside their classrooms. Our research also highlights different modules that comprise the system and how each module was crafted based on student–teacher interactions and observations to ensure maximum engagement from the students. The system can become commonplace in schools across the country if aided by similar efforts in recording and maintaining a database of all possible signs and grammatical structures, something that is sparsely present as of today. Due to lack of awareness and resources in Indian Sign Language, the system may not be able to decode complex grammatical structures and interactions. Our future research work will be conducted on what scale these systems can be currently implemented and adopted as teaching aids with a limited sign language database at our disposal.

Acknowledgements The authors would like to acknowledge Mrs. Sucheta S Saundankar, Principal, Padsad Karanbadhir Vidyalaya Nashik, and Mrs. P Jessy, Principal, Sheila Kothavala Institute for the Deaf, Bangalore, India, for allowing us to conduct the validation study at their respective schools. We are grateful to the Principal, Jessy Samuel and Speech Therapist, Ms. Suvasini Isaac for personally verifying the signs used in the validation stage for accuracy and consistency. We thank the anonymous reviewers whose insightful comments and suggestions have significantly improved this paper.

References

1. Agarwal, A., Thakur, M.K.: Sign language recognition using microsoft kinect. In: 2013 Sixth International Conference on Contemporary Computing (IC3), pp. 181–185 (2013). https://doi.org/10.1109/IC3.2013.6612186
2. Ahmed, M.A., Zaidan, B.B., Zaidan, A.A., Salih, M.M., Lakulu, M.M.B.: A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017. Sensors **18**(7), 2208 (2018). https://doi.org/10.3390/s18072208
3. Ahmed, S.: Real time American sign language video captioning using deep neural networks [PowerPoint presentation]. http://on-demand.gputechconf.com/gtc/2017/presentation/s7346-syed-ahmed-real-time-american-sign-language-video-caption.pdf (2018)
4. Cannon, J.E., Fredrick, L.D., Easterbrooks, S.R.: Vocabulary instruction through books read in American sign language for English-language learners with hearing loss. Commun. Disord. Q. **31**(2), 98–112 (2010). https://doi.org/10.1177/1525740109332832
5. Center, L.C.N.D.E.: Learning american sign language: Books, media, and classes. http://www3.gallaudet.edu/clerc-center/info-to-go/asl/learning-asl-books_media_classes.html. [Online]
6. Chambers, B., Abrami, P.C., McWhaw, K., Therrien, M.C.: Developing a computer-assisted tutoring program to help children at risk learn to read. Educ. Res. Eval. **7**(2–3), 223–239 (2001). https://doi.org/10.1076/edre.7.2.223.3863
7. Chandramouli, C.: Census of India 2011, provisional population totals, government of India. http://censusindia.gov.in/2011-prov-results/paper2/data_files/india/paper2_1.pdf (2011). [Online]
8. Division, S.S.: Disabled persons in India, a statistical profile 2016. http://mospi.nic.in/sites/default/files/publication_reports/Disabled_persons_in_India_2016.pdf (2016). [Online]
9. Easterbrooks, S.R., Huston, S.G.: The signed reading fluency of students who are deaf/hard of hearing. J. Deaf Stud. Deaf Educ. **13**(1), 37–54 (2007). https://doi.org/10.1093/deafed/enm030
10. Fung, P.C., Chow, B.W.Y., McBride-Chang, C.: The impact of a dialogic reading program on deaf and hard-of-hearing kindergarten and early primary school-aged students in Hong Kong. J. Deaf Educ. Deaf Stud. **10**(1), 82–95 (2005)
11. Goyal, L., Goyal, V.: Automatic translation of English text to Indian sign language synthetic animations. In: Proceedings of the 13th International Conference on Natural Language Processing, pp. 144–153 (2016)
12. Hermans, D., Knoors, H., Ormel, E., Verhoeven, L.: The relationship between the reading and signing skills of deaf children in bilingual education programs. J. Deaf Stud. Deaf Educ. **13**(4), 518–530 (2008). https://doi.org/10.1093/deafed/enn009
13. Huenerfauth, M.: Generating American sign language animation: overcoming misconceptions and technical challenges. Univers. Access Inf. Soc. **6**(4), 419–434 (2008). https://doi.org/10.1007/s10209-007-0095-7
14. Journal, C.W.R.: Oralism and how it affects the development of the deaf child. http://www.csus.edu/wac/journal/2010/hood_research_paper.pdf (2010)
15. Kafle, S., Huenerfauth, M.: Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In: Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS'17, pp. 165–174. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3132525.3132542
16. Lang, S., Block, M., Rojas, R.: Sign language recognition using kinect. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) Artificial

Intelligence and Soft Computing, pp. 394–402. Springer Berlin Heidelberg, Berlin (2012)
17. Lazar, J., Feng, J.H., Hochheiser, H.: Research Methods in Human–Computer Interaction. Wiley, New York (2010)
18. Justice, L.M.: Word exposure conditions and preschoolers' novel word learning during shared storybook reading. Read. Psychol. **23**(2), 87–106 (2002). https://doi.org/10.1080/027027102760351016
19. Massaro, D.W., Light, J.: Improving the vocabulary of children with hearing loss. Volta Rev. **104**(3), 141–174 (2004)
20. Mayberry, R.I., Eichen, E.B.: The long-lasting advantage of learning sign language in childhood: another look at the critical period for language acquisition. J. Mem. Lang. **30**(4), 486–512 (1991). https://doi.org/10.1016/0749-596X(91)90018-F
21. Michael Strong, P.M.P.: A study of the relationship between American sign language and English literacy. J. Deaf Stud. Deaf Educ. **2**(1), 37–46 (1997)
22. NCERT: Chapter-wise video lessons (2018). http://www.ncert.nic.in/new_ncert/ncert/publication/publication_list/list_of_publication3.html. [Online]
23. Capirci, O., Cattani, A., Rossini, P., Volterra, V.: Teaching sign language to hearing children as a possible factor in cognitive enhancement. J. Deaf Stud. Deaf Educ. **3**(2), 135–142 (1998). https://doi.org/10.1093/oxfordjournals.deafed.a014343
24. Pettengill, R.: Motionsavvy uni (2011). http://www.motionsavvy.com/about.html. [Online]
25. RKMVERI: Isl dictionary (2004). http://www.indiansignlanguage.org/. [Online]
26. Robbins, C., Ehri, L.C.: Reading storybooks to kindergartners helps them learn new vocabulary words. J. Educ. Psychol. **86**(1), 54–64 (1994). https://doi.org/10.1037/0022-0663.86.1.54
27. Santorini, B.: Part-Of-Speech tagging guidelines for the Penn Treebank project (3rd revision, 2nd printing). Tech. rep., Department of Linguistics, University of Pennsylvania, Philadelphia, PA, USA (1990)
28. Schleper, D.R.: Read it again... and again and again. Perspect. Educ. Deaf. **14**(2), 16–19 (1995)
29. Society, D.A.: Sheila Kothavala Institute for the Deaf. http://deafaidsociety.in/ (2018)
30. Sousa, L., Rodrigues, J.M.F., Monteiro, J., Cardoso, P.J.S., Lam, R.: Gygsla: a portable glove system for learning sign language alphabet. In: Antona, M., Stephanidis, C. (eds.) Universal Access in Human-Computer Interaction. Users and Context Diversity, pp. 159–170. Springer, Cham (2016)
31. Strong, M., Prinz, P.: A study of the relationship between American sign language and English literacy. J. Deaf Stud. Deaf Educ. **2**(1), 37–46 (1997)
32. Talk, S.: Signtalk: The resource for sign language interpreters. http://signtalk.org/ (2018)
33. Talkinghands: Indian sign language dictionary (2013). http://www.talkinghands.co.in/. [Online]
34. Tenório, R.: Virtual interpreter for Brazillian sign language. https://www.handtalk.me/ (2012). [Online]
35. Verma, D., Dash, P., Bhaskar, S., Pal, R.P., Jain, K., Srivastava, R.P., Hansraj, N.: Disabled persons in India: a statistical profile 2016. http://mospi.nic.in/sites/default/files/publication_reports/Disabled_persons_in_India_2016.pdf (2017). [Online]
36. Karnabadhir Vidyalaya, P.: PADSAD-the unique school aiming to main stream every hearing impairment and divyang student. http://www.padsad.org/ (2018)
37. Wideman, C.: Vcom3d. http://www.vcom3d.com/ (2002)
38. Wiggins, E.: Total communication as an education philosophy. http://www.deaflinx.com/DeafEd/OptionsGuide/TC.html (1998). [Online]
39. Wobbrock, J.O., Kientz, J.A.: Research contributions in human–computer interaction. Interactions **23**(3), 38–44 (2016). https://doi.org/10.1145/2907069
40. Ye, Y., Tian, Y., Huenerfauth, M., Liu, J.: Recognizing American sign language gestures from within continuous videos. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2018)
41. Zeshan, U., Vasishta, M., Sethna, M.: Implementation of Indian sign language in educational setting. Asia Pacific Disab. Rehabil. J. **16**(1), 16–40 (2005)
42. Zheng, L., Liang, B., Jiang, A.: Recent advances of deep learning for sign language recognition. In: 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, Australia, pp. 1–7 (2017). https://doi.org/10.1109/DICTA.2017.8227483