Proc. of Int. Conf. on Advances in Computer Science, AETACS

ELSEVIER

A Video-based Approach for Translating Sign Language to Simple Sentence in English

Aradhana Kar[1] and Pinaki Sankar Chatterjee[2]
[1]Kalinga Institute of Industrial Technology, Bhubaneswar, Orissa, India-751024
aradhana140587@gmail.com
[2]Kalinga Institute of Industrial Technology, Bhubaneswar, Orissa, India-751024
pinaki.sankar.chatterjee@gmail.com

***Abstract*— Sign Language is the language of deaf. This paper discusses an approach for translating Sign Lan-guage videos to its corresponding simple sentence in English. There are different types of sign languages spread all over the world. American Sign Language (ASL) is one of the sign languages. ASL is used by deaf Americans. We have collected some ASL videos in which a person is signing a particular word which will be translated to words of English language. We have created a translator that converts the ASL video to grammatically correct textual simple sentence in English. This textual simple sentence is then converted to speech.**

***Index Terms*— American Sign Language, Sign Language, Simple Sentence, Natural Language**

I. INTRODUCTION

Deaf people are people who cannot hear[7]. Dr. Johnson has viewed deafness as a physical defect[7]. Deaf people have their own language which is not oral but signed[7]. Sign language is a language that uses visually transmitted patterns to convey speaker's thoughts. It does not use sound patterns. There are different types of sign languages such as, American Sign Language (ASL) for United States of America, British Sign Language (BSL) for Britain, Spanish Sign Language (SPL) for Spain, Japanese Sign Language for Japan, Korean Sign Language for Korea, Mexican Sign Language for Mexico, Brazilian Sign Language for Brazil etc.

Like spoken language, ASL has its own grammar. ASL grammar is not same as spoken language grammar.

A. ASL Grammar

ASL Grammar consists of ASL Phonology, ASL Morphology and ASL Tenses. ASL Phonology consists of structural units which have five parameters. Originally, William Stokoe identified four parameters. The four parameters are, handshape, palm orientation, movement, location [8]. Later, another parameter that is, non-manual markers (NMM) was added. NMM may be defined as facial expression, upper-body position and mouthing. Mouthing includes use of the lips, tongue, jaws, cheeks and breath.

ASL Morphology consists of verbs, nouns, adjectives and pronouns[9]. Adjectives generally follow nouns[9]. Verbs are closely related to nouns and may differ in reduplication of movement. For example, the signs for

"food" and "eat" are nearly same. The only difference is that the "food" (the noun) is performed twice and the "eat" (the verb) is performed once. Although most nouns do not have a verb that looks same, but few nouns need double motion. ASL is a pro-drop language. A pro-drop language is a language in which certain classes of pronouns are omitted when they are in some sense. ASL words are historically compounds. The elements of two different signs are fused to form compounds. For example, the verb "agree" is derived from two words "think" and "alike". The word-order of ASL is subject-verb-object. The full sentence structure in ASL is [topic] [subject] verb [object] [subject-pronoun-tag]. Topics and tags are expressed through NMM.

ASL tenses is one of the components of ASL syntax. In the past, researchers believed that ASL used only Time Adverbials to indicate time. However, in the present situation it is believed that ASL speakers use tenses to a large extent. To communicate tenses in ASL, ASL signers need their body and hands. For present tense, ASL signers sign close to their body, just like they normally do in a signed conversation [4]. For past tense, ASL signers sign the word "finish" at the beginning or end of the sentence to indicate that everything has already happened [4]. Most ASL signers sign "finish" at the beginning of the sentence and then they sign the whole sentence. For future tense, ASL signers sign the word "will" at the end of the sentence [4]. In English, we use time bound indicator morphemes to express the tense [4]. For example, we add "s" to indicate present third person and "ed" bound morphemes to indicate the past tense, but ASL does not use the time bound indicator.

ASL can be recognized by using two different approaches. The two approaches are: video-based approach and instrumented-based approach. In video-based approach, ASL videos are taken as input. In instrumented-based approach, different instruments like AcceleGlove, DataEntryGlove, CyberGlove are used for capturing hand movements.

We have adopted video-based approach for converting ASL videos to English language. ASL videos will be mapped into static pictures. These static pictures are taken from Sign Writing pictures. Sign Writing uses visual symbols to represent the handshapes, movements and facial expressions of signed languages [11]. Different magazines, books, newspapers and literature are written by using Sign Writing[10]. Sign Writing is used to teach signs and signed language grammar to beginner signers[10].

II. Related Work

Several works have been done in the field of translation of ASL to English language. Some researchers have adopted instrumented-based approach and some have adopted video-based approach.

In Ref. [1], Jose L. Hernandez-Rebollar, Nicholas Kyriakopoulos, Robert W. Lindeman have adopted an instrumented-based approach. They have proposed an approach for capturing and translating isolated gestures of American Sign Language into written text and sound. They have broken down the gestures of American Sign Language into unique sequence of phonemes called Poses and Movements. They have considered the following definitions:

(a) Pose is a static phoneme which is represented by vector P = [hand shape, palm orientation, hand location] [1].
(b) Posture is represented by a vector
Ps = [hand shape, palm orientation] [1].
(c) Movement is a dynamic phoneme. It is represented by a vector M = [direction, trajectory] [1].
(d) A manual gesture is a sequence of poses and movements, P-M-P [1].
(e) A manual gesture s is said to be sign if s ϵ L [1].

They have used a Lexicon of one-handed signs of the type Pose-Movement-Pose for recognition based on the framework set by these definitions. By doing so, the recognition system is divided into smaller systems trained to recognize a finite number of phonemes. Since any word is merely a new combination of the same phonemes, the individual systems do not need to be re-trained when new words are added to the lexicon.The system comprises of an AcceleGlove and a two-link arm skeleton. AcceleGlove is used to detect different hand sizes accurately. The two-link arm skeleton consists of three components: one dual-axis accelerometer and two resistive angular sensors. One axis of the accelerometer detects arm elevation (Θ_1) and the second axis detects arm rotation (Θ_2). One resistive angular sensor placed on the shoulder measures forearm rotation (Θ_4) and the second angular sensor placed on the elbow measures forearm flexion (Θ_3). The capturing system is provided by two push buttons pressed by the user to indicate the beginning and ending of a gesture. One byte per signal is sent via serial port at 9600 baud to a laptop think-pad IBM T-21 with a Pentium III running at 500 MHz. The program to read the signals and extract the features, discriminate postures, locations, movements and search for the most likely sign, was written in Pascal 1.5 for Windows. The micro controller

is connected to a speech synthesizer V8600 'DoubleTalk' from RC Systems which receives the ASCII string of the word corresponding to the recognized gesture. To classify complete signs, they have used conditional template matching, a variation of template matching. Conditional template matching compares the incoming vector of components (captured with the instrument) with a template (in the lexicon) component by component and stops the comparison when a condition is met.

In Ref. [2], Philippi Dreuw, Daniel Stein, Thomas Deselaers, David Rybach, Morteza Zahedi, Jan Bungeroth, and Hermann Ney have adopted a video-based approach. They have proposed a system that automatically recognizes sign language and converts it to the spoken language. They have used a video signal as input. Therefore, they have used a speech recognition system to obtain the textual representation of the sign language. This intermediate representation is then fed into the statistical machine translation system to translate it into spoken language. They have used an Automatic Speech Recognition (conversion of an acoustic signal into a sequence of written words) to create a sign language recognition system. The conversion of video signals (images) into a sequence of written words (texts) is called Automatic Sign Language Recognition (ASLR). According to them, in order to build a robust recognition system which can recognize continous sign language independently, they have to cope with various difficulties: (i) coarticulation : the appearance of the sign depends on the preceding and succeeding signs. (ii) Inter- and intrapersonal Variability : the appearance of a particular sign can vary significantly in different utterances of the same signer and in utterances of different signers. ASLR generates an intermediate output which is fed into the statistical machine translation system for generating spoken language. Automatic machine translation is the translation from a source language into a target language by means of either data-based or rule-based methods. For rule-based systems, a set of translation rules has to be created manually by bilingual language experts, while for data based approaches the machine has to derive the rules itself by extracting them from given examples (supervised learning), without any prior language or grammar knowledge involved.

In Ref. [3], Samit Bhattacharya has proposed an approach for generating natural language from icons for persons who suffer from severe speech and motor impairment (SSMI). The system is an Iconic Communication System where the user communicates by selecting icons. The system has three components: physical interface, a processing unit and a language set. The language set of the system is a set of unambiguous icons. They have used unambiguous icons for reducing the learning cost of the system. The icons are organized in the form of a hierarchy. The icons of the language set are displayed to the user through the physical interface. The user can select the icons from the interface in an interactive way. The interaction is performed by questioning the user and getting answers from the user. The user answers by selecting icons from the interface. The sequence of selections made by the user is converted to an instantiated representation which is a frame like structure. This intermediate representation is accepted by a natural language simple sentence generator that acts as the processing unit of the system. The language generator produces grammatically correct natural language simple sentences (in textual form) from the intermediate representation, which is the output of the system. The system is integrated with a text-to-speech synthesizer to “speak out” the textual output. The system can be operated with special access switches. The switches are required to make the system accessible to the people with severe motor disabilities.

In Ref.[5], Sumit Das, Anupam Basu, Sudeshna Sarkar have proposed an approach for generating Bengali compound sentences using identified constructs. Their objective is to propose an approach for automatically generating discourse marker to connect coherent spans and to perform syntactic aggregation on the text after the generation of the appropriate discourse marker. The input is the discourse structure tree, that is, the elementary text spans connected by the rhetorical relations. The semantic representation for the text spans chosen is a case-frame representation. The basic building block in semantic representation is sentence. A sentence contains clause-count and a clause frame. The clause-count denotes the number of simple clauses present in the sentence. The clause frame is a recursive structure that can contain clauses inside it, which helps in representing both simple and composite sentences. For simple sentence, the outer clause only contains one inner clause. On the other hand, for composite sentence the outer clause contains the constituent inner clauses along with the rhetorical relation (rhrel) and discourse marker (dm) realizing that rhetorical relation. Clause frame contains a predicate frame (pred) and list of argument frames (arg). The pred and arg frames contain the required functional information. They have focused on the paratactic constructs for syntactic aggregation of Bengali text. In this work, they have considered multi-nuclear rhetorical relations such as CONJUNCTION, DISJUNCTION, CONTRAST, and SEQUENCE as defined by the Rhetorical Structure Theory. In addition to these relations, they have also considered another multi-nuclear temporal coherence relation PARALLEL. According to them, two text spans are said to be related by PARALLEL relation if the actions or the events in those two text spans are occurring simultaneously. For performing text

aggregation they have conducted a corpus analysis to identify the prevalent syntactic aggregation constructs used in Bengali for generating compound sentences. In corpus analysis, they have randomly chosen 350 sentences from a corpus 600 Bengali compound sentences and segmented into constituent simple clauses. They have identified two types of frequently used syntactic aggregation constructs in Bengali: simple paratactic construct and elliptic construct. In simple paratactic construction, the two constituent simple clauses are simply connected by the conjunctive discourse marker and no word deletion is required. In elliptic construction, superfluous words from the surface form which are inferable from the entities in the remaining text are omitted. In syntactic aggregation, they have given two simple clauses, the rhetorical relation between them, and the discourse marker realizing that relation as inputs. These arguments are ordered in the constituent clauses. Then repeated entity is identified. Then the constituent clauses are ordered. At the end, superfluous words are deleted and non-finite verb are generated.

In Ref.[6], Vicki L. Hanson, Carol A. Padden have proposed an educational software named as HandsOn that uses a bilingual approach for developing the reading and writing skills of elementary-aged deaf children. This system is fully interactive, using touch (or mouse) responses. The system uses laser disc technology to present a real person on screen signing American Sign Language (ASL). English text stored on the computer is combined with the ASL video on the screen. Laser discs were created that contain stories for HandsOn. These stories cover topics such as Science, Social Studies, Conservation, Geography, and Literature. They have developed following activity options in HandsOn:

- Read a Story – With this option, the students see a printed text in English. Students can request the ASL translation of one of the English sentences by touching the sentence that they wish to see translated. Periodically, students must answer written questions about what they have just read. Students also can get ASL translations of the questions.
- Watch a story – With this option, the students see a story signed in ASL. They can get an English translation of an ASL sentence, can choose to play the signing in slow motion (forwards or backwards), at fast speed or can stop it.
- Caption a story – With this option, students write English translations for the ASL stories, one sentence at a time. These translations can be saved, edited, printed out, and later played back

III. System Architecture

Our system consists of three modules:

- Video Processing – In this module, ASL videos will be mapped into static pictures. These static pictures are taken from Sign Writing Pictures. We have extracted the frames of the collected videos which serve as our dictionary. We have also extracted frames of the input video. Then frames of the inputted video are compared with the frames present in the dictionary for obtaining a right match. This right match is mapped into a sign writing picture and this picture serves as input to the Natural Language Generation module.
- Natural Language Generation - In this module, the Sign Writing Pictures will be converted to grammatically correct textual simple sentence in English. We have prepared another dictionary of sign writing pictures in this module. The output of the Video Processing Module is compared with the sign writing pictures present in the dictionary for obtaining a right match. The right match is an ASL word expressed in sign writing picture. All the information about this word is extracted from an excel file. These information are used to fulfill the slots of the uninstantiated SSU frame for obtaining an instantiated SSU frame. These instantiated SSU frame is given as input to the sentence maker to form gramatically correct textual simple sentence in English.
- Text to Speech Conversion - In this module, the grammatically correct textual simple sentences in English are converted to speech.

The overall system architecture is shown in Fig. 1. The rectangles with round corners denote processes, the white page symbols denote output and input of various sub stages and the circles denote system resources.

A. Video Processing

In this module, we have collected some ASL videos from a website. In these collected ASL videos a person is signing some word of ASL. Frames are extracted from each collected ASL video. The frames of each collected ASL video are stored in a folder and the folder is named with the word which is signed by the person in the collected ASL video. By following this procedure, many folders consisting of frames extracted

from collected ASL videos are created and they are named in the same way. All these folders are now stored in the dictionary folder called *CollectedVideoFrames*.
Then, we have given an ASL video which has to be converted to speech as input to the system. Frames of the inputted ASL video are extracted and these frames are stored in a folder called *InputFrames*. This procedure is carried out for every inputted ASL video.

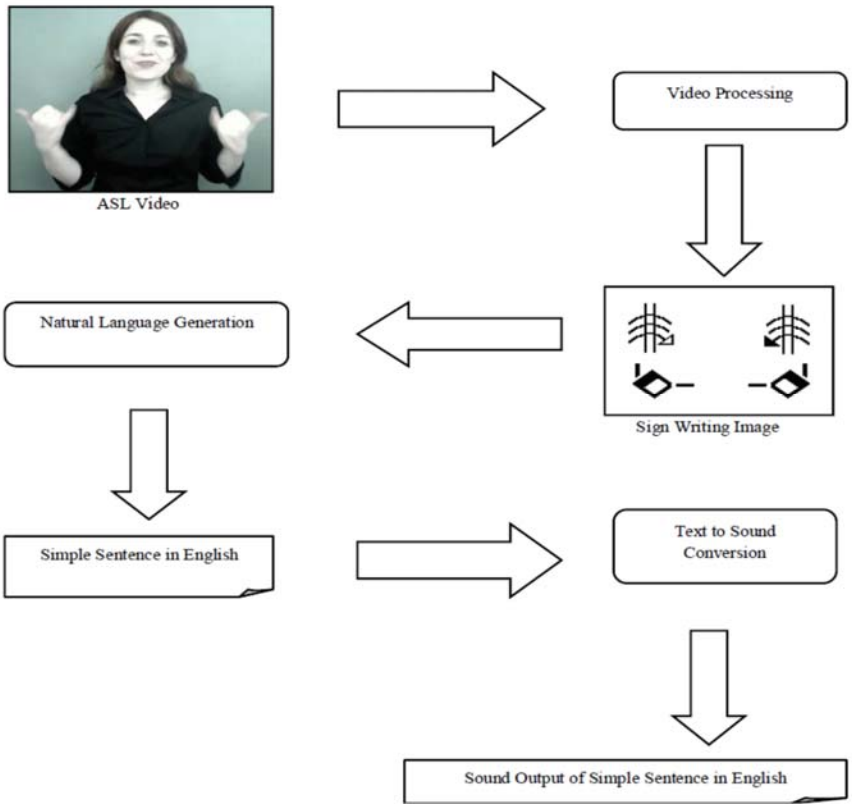

Fig. 1: Overview of System architecture

Now, the frames present in the folder *InputFrames* are compared with the frames present in the folders of *CollectedVideoFrames* using picture comparison technique. In the matching procedure, we have agreed upon a threshold value. Threshold value is the number of matched frames in a folder present in C*ollectedVideoFrames*. If the number of matched frames in a folder of *CollectedVideoFrames* is equal to or more than the threshold value then that folder is accepted as the right match.
The name of the matched folder is extracted and that name is searched in the excel file *VideoToImage.xls* and its corresponding Sign Writing image file name is retrieved from the second column. This excel file consists of the names of folders that are stored in the *CollectedVideoFrames* in one column and their corresponding Sign Writing image file name in another column. The above file name is searched in the *SignWritingDictionary.* The *SignWritingDictionary* is a dictionary that consists of Sign Writing images. Each image file in the *SignWritingDictionary* folder is named with the meaning of the sign in Sign Writing.
After successful searching, the Sign Writing file with matched file name from *SignWritingDictionary* is dumped into a folder called *SignWritingInput*. This folder serves as the input to the Natural Language Generation.
The architecture of video processing is shown in Fig. 2. The rectangles with round corners denote processes, the white page symbols denote output and input of various sub stages and the circles denote system resources.

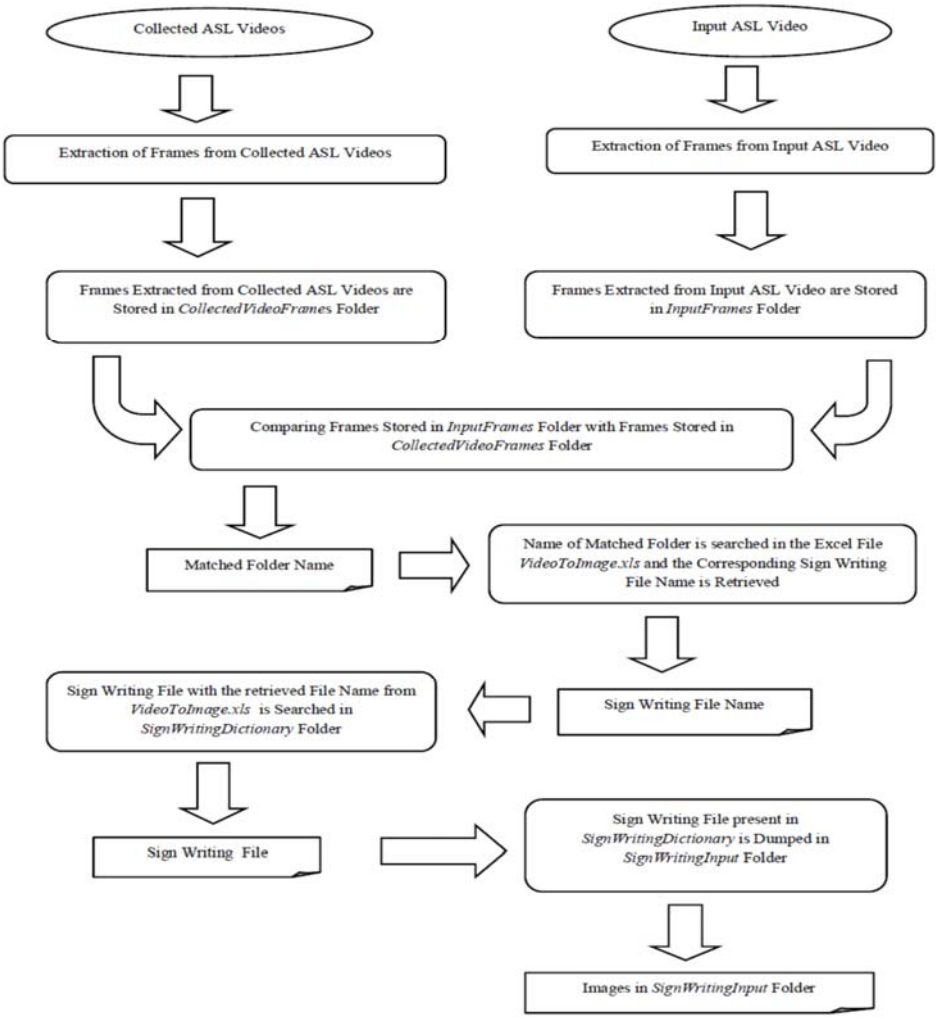

Fig. 2: Overview of Video Processing

B. Natural Language Generation

In this module, the *SignWritingInput* folder is given as input. The images of *SignWritingInput* are compared with the Sign Writing images present in *SignWritingDictionary*. The file name of the matched file is extracted.The extracted name of the file is searched in an excel file named as *SignWritingInfo.xls*. This excel file consists of name of the files present in the *SignWritingDictionary* folder, the corresponding Sign Writing word and identifications of the Sign Writing word (i.e, whether a noun or a pronoun or a verb) are stored in the first, second and third column respectively.

After successful searching, all information related to the right match is retrieved from *SignWritingInfo.xls*. These information are used to fill up the slots of the Semantico-Syntactic Unit (SSU) frame.

Consider an English sentence: "I read book". The information contained in this sentence can be obtained by asking a series of questions to the main verb (read). For this sentence the questions can be who and what. Each of these questions is called as *Semantico-Syntactic Unit or SSU*. Each of these questions correspond to a role. We have taken SSU and verb group as intermediate representation. Due to the presence of SSUs, the representation is called as *SSU frame*.

Initially uninstantiated SSU frames are kept in the system for each verb. In the uninstantiated SSU frame, we want the verb along with other information such as tense, aspect, polarity, theme about the verb. The first argument, Arg1 consists of a SSU frame and this slot of uninstantiated SSU frame is filled with a pronoun. Similarly, the second argument, Arg2 consists of another SSU frame and this slot of uninstantiated SSU

frame is filled by a noun. The filling up the slots of uninstantiated SSU frame is called the instantiation of the SSU frame. In our system, the slots of the uninstantiated SSU frame are filled with the extracted information from *SignWritingInfo.xls* to make an instantiated SSU frame. The instantiated SSU was then given as input to the Sentence Maker to produce simple sentence in English. In the Sentence Maker, the instantiated SSU frame is modified. During the modification of frame, the SSUs are ordered.
An example of instantiated SSU Frame for "I read book" has been shown in Fig.3.

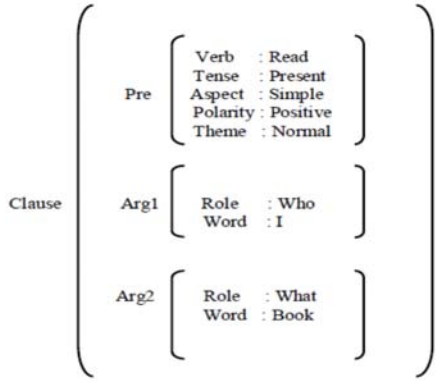

Fig. 3: Instantiated SSU Frame for "I read book"

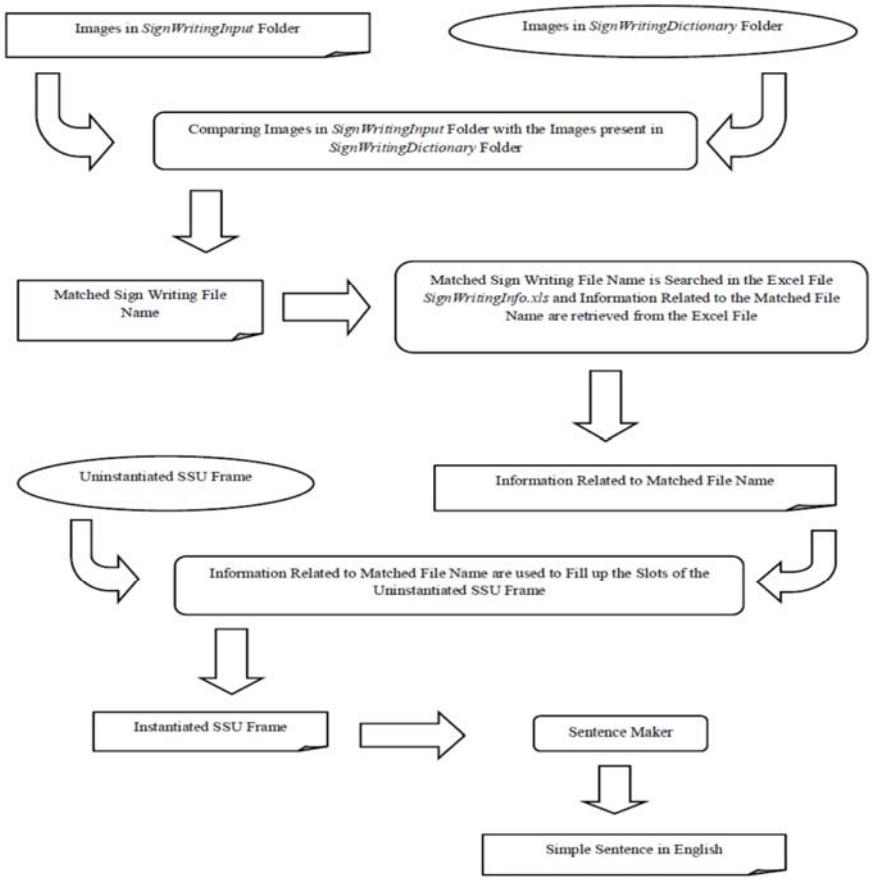

Fig. 4: Overview of Natural Language generation

The architecture of Natural Language Generation is shown in Fig. 4. The rectangles with round corners denote processes, the white page symbols denote output and input of various sub stages and the circles denote

system resources. The input is the Images in *SignWritingInput* folder which is shown inside a white page symbol and the output is the simple sentence in English which is also shown inside a white page symbol. The dictionary folder, *SignWritingDictionary* that consists of SignWriting Images is shown inside a circle. The comparing process is shown inside the rectangles with round corners.

C. Text to Speech Conversion

The output of the Natural Language Generation module, that is, the simple sentence in English is given as input to the text to sound conversion module which speak out the simple sentence in English.
The architecture of text to sound conversion is shown in Fig. 5. The rectangles with round corners denote processes, the white page symbols denote output and input of various sub stages and the circles denote system resources.

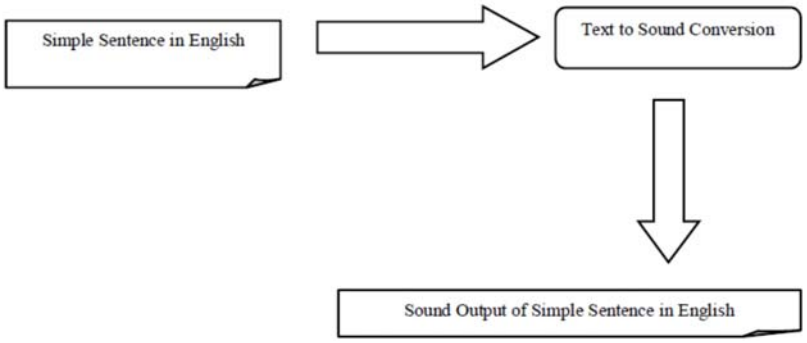

Fig. 5: Overview of Text to Sound Conversion

IV. TESTING AND EVALUATION

We have developed a system that translates ASL to simple sentence in English. We were not able to collect ASL signs from ASL signers (person those uses ASL for communication) as we did not have the scope to get connected with them. So we collected 100 ASL videos from a website and created our ASL dictionary. However, we have tested the simple sentences which were generated by the system as the output. We had given the simple sentences in English that are generated by the system to five human evaluators. Each human evaluator rates the simple sentence in a scale of 1 to 4 based on some factors. Depending upon their feedbacks, the working of the system is evaluated. We have concentrated on two aspects of the outputted simple sentence by the system:

- *Syntactically correct*:- An outputted simple sentence in English is said to be syntactically correct if it follows all the grammatical rules of English.
- *Semantically correct*:- An outputted simple sentence in English is said to be semantically correct if it preserves the correct meaning.

Syntactically correctness of the sentence is shown in Fig.6. Syntactic correctness of the sentence is represented in pie chart. The percentage labelled on the graph indicates the correctness and changes required in a sentence which was given for testing to the evaluators. We have taken four categories: syntactically correct, Syntactically correct but minor changes required, syntactically correct but major changes required and syntactically incorrect. According to the results of evaluation, 88 percent of the total sentences are syntactically correct, 6 percent are syntactically correct but minor changes required, 3 percent are syntactically correct but major changes required and 3 percent are syntactically incorrect.
Semantically correctness of the sentence is shown in Fig.7. Semantic correctness of the sentence is represented in the pie chart. The percentage labelled on the graph indicates the semantic correctness and changes required in a sentence which was given for testing to the evaluators. In this case, we have taken four categories: correct meaning, correct meaning but minor changes required, correct meaning but major changes required and incorrect meaning. According to the results of evaluation, 91 percent of the total sentences gives correct meaning, 7 percent gives correct meaning but minor changes required, 2 percent gives correct meaning but major changes required and 0 percent gives incorrect meaning.

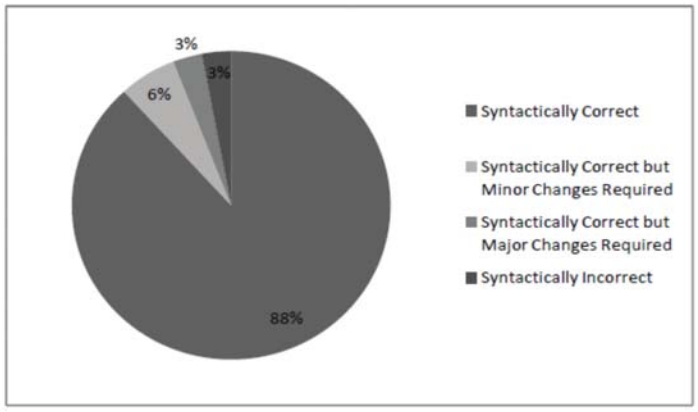

Fig. 6: Syntactically Correct Pie Chart

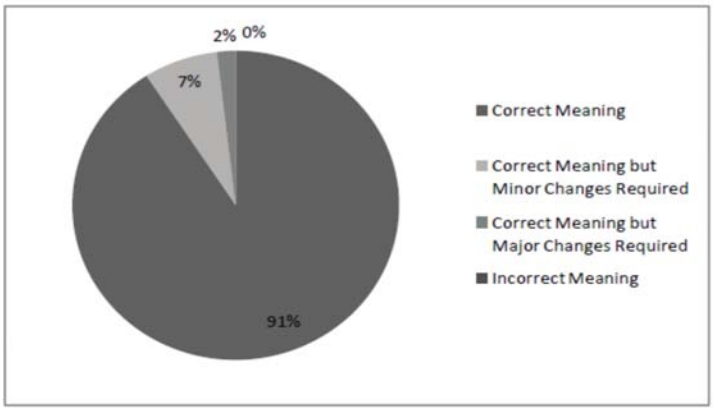

Fig.7 : Semantically Correct Pie Chart

V. Conclusion and Future Work

This paper discuss about the generation of simple sentences in English from the captured ASL videos. These simple sentences were converted into audio form. Frames of the inputted video were generated and they were compared with frames present in the dictionary. Based on the right match, the Sign Writing Image File is retrieved and stored in a folder. This folder served as the input to Natural Language Generation Module. The Sign Writing Images in this folder compared with the Sign Writing Images present in Sign Writing Dictionary and the right match retrieved. Based on the right match, all information related to the right match is retrieved from a excel file. This information is used to fill up the slots of uninstantiated SSU frame which results in an instantiated SSU frame. This instantiated SSU frame is given as input to sentence maker to make simple sentence in English. We have also converted the simple sentence in English to speech.

Till now, we have prepared a frame work that takes ASL videos as input gives simple sentence in English in sound form as output. Presently, we have collected a small number of ASL videos for our system. In future, we will add some other ASL videos to our dictionary for increasing the accuracy of the system.

References

[1] Jose L. Hernandez-Rebollar, Nicholas Kyriakopoulos, Robert W. Lindeman, "A New Instrumented Approach for Translating American Sign Language into Sound and Text", Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition, 2004.

[2] Philippi Dreuw, Daniel Stein, Thomas Deselaers, David Rybach, Morteza Zahedi, Jan Bungeroth, and Hermann Ney, "Spoken Language Processing Techniques for Sign Language Recognition and Translation", Human Language Technology and Pattern Recognition, Computer Science Department 6, RWTH Aachen University, Germany, 2008.

[3] Samit Bhattacharya, "Sanyog : An Iconic System for Multilingual Communication for People with Speech and Motor Impairments", M.S Thesis, IIT Kharagpur, Supervisor – Anupam Basu, Sudeshna Sarkar, 2004.

[4] Karen Alkoby, “A Survey of ASL Tenses”, Proceedings of the 2nd Annual CTI Research Symposium. Chicago, Illinois, November 4, 1999.
[5] Sumit Das, Anupam Basu, Sudeshna Sarkar, “Discourse Marker Generation and Syntactic Aggregation in Bengali Text Generation”, Proceedings of the 2010 IEEE Students’ Technology Symposium, 2010.
[6] Vicki L. Hanson, Carol A. Padden, “HandsOn : A Multi-media Program for Bilingual Language Instruction of Deaf Children”, IEEE 1992.
[7] David M. Perlmutter, “The Language of Deaf”, The New York Review of Books, 28th March 1991.
[8] William Stokoe, “Dictionary of American Sign Language on Linguistic Principles”, Linkstok Press 1976.
[9] Clayton Valli, “ Linguistics of American Sign Language: An Introduction”, 2005
[10] All the Sign Writing images are collected from http://www.signwriting.org/
[11] Information about Sign Writing can be obtained from http://www.omniglot.com/writing/signwriting.htm

Learn sentence structure using a topic-comment structure in American Sign Language (ASL).Â In an OSV sentence, the non-manual signal is raised eyebrows and tilt head forward at the beginning of the sentence when signing the object (O), then proceeding with the rest of the sentence (SV). See a few examples below. The gloss /\ depicts raised eyebrows. I'll start with a few simple sentences before giving a more complex sentence. Sign language videos visible in HTML5-based browsers. The sentence above, /\book/\, ix-me give-hir (where "hir" is a non-gendered pronoun for "her/him"). The signer raises her eyebrows and slightly tilted her head forward when signing &