UNSYS digital

International Journal of Robotics and Mechatronics

Survey on Language Processing Approaches for Sign Language Synthesizers

Haris Al Qodri Maarif, Rini Akmeliawati, Teddy Surya Gunawan

Faculty of Engineering, International Islamic University Malaysia

Abstract—Hearing and Speech Impaired (HSI) community is often disadvantaged due to their difficulties in communicating with the rest of the world. Although sign languages (SL) have been used as a means to convey their message, problems still exist as there are not many people familiar with this type of languages. Besides, the number of available translators is insufficient to resolve the issue. This has motivated various research described in this paper. In general, we can divide the main contribution in solving the communication problem of the HSI community into two types, the first involves the development of automatic SL translators which allow the non-HSI communities to understand the messages conveyed by the HSI people, and the second relates to the development of SL synthesizers which allow the non-HSI community to pass their messages to HSI people. Thus, when both systems are integrated, two way communication can be established. In this paper, we provide a comprehensive review on the work in SL synthesizer development. SL synthesizer is a tool that synthesizes or constructs series of signs of a particular SL based on the speech input given to the system. The process involves adjusting speech input that can fit the sign language grammatical rule in order to provide an understandable output to the HSI people. Some algorithms, such as Natural Language Processing, are explored. In this paper, the details of each methods have been identified and discussed. In conclusion, we found that the Natural Language Processing (NLP) has provided relatively more efficient process and better results for SL synthesizers.

Keywords—HSI People, Sign Language, Natural Language Processing, Sign Language Synthesizer.

Copyright© 2017. Published by UNSYSdigital. All rights reserved.
DOI: 10.21535/ijrm.v4i2.1001

I. Introduction

HEARING and Speech Impaired (HSI) people face difficulties to communicate with non-HSI People. Sign languages have been one of the tools for HSI people to communicate between them. The HSI people are often required to learn Sign Language in order to be able to gain communication among themselves, but the non-HSI people are not required to learn the language unless there is an essential for them to do so. This creates the communication gap between HSI people and non-HSI people where the translators or interpreters (non-HSI people who understand sign language) are often needed to help the communication gap. The number of translators is very limited and it is not adequate to accommodate the demand in any country.

Corresponding author: Rini Akmeliawati (e-mail: rakmelia@ieee.org)
This paper was submitted on November 30, 2017,
and accepted on December 30, 2017.

There are almost 70 million HSI in the whole world, as reported by The World Federation. There are around 138 types of sign languages according to the Ethnologue catalogue (Joy & Balakrishnan, 2014). It has been a silent language for spoken by millions of HSI people all over the world. The other report was announced from a census held on 2001 which mentioned that in India, the ratio of HSI people and the interpreter is only less than 0.02 % from the population of the HSI people, or in simple words, it is only one interpreter for every 72,000 people (Dasgupta & Basu, 2008).

Referring to the report produced by Malaysian Federation of the Deaf (MFD), the number of qualified interpreters are 60 where they have to serve more than 55,000 people who are deaf in Malaysia. Sign languages have been used as a means to communicate besides other hearing aids, which are often very costly (Murad, 2013).

Sign language is a unique language. It does not have any written form. It involves hand movement, body gesture and face expression to express the words or sentences. Currently, human intervention is needed for converting text or audio input to sign language and sign language to text or audio. This refers to the automatic sign language synthesizer and translator, respectively. The implementation of Information Technology (IT) is an effective solution for providing solution of Sign Language synthesizer and translator (Joy & Balakrishnan, 2014).

Sign language is also known as finger-spelling, which is used to spell words (name of person/thing, rate words, and unknown signs) letter-by-letter. The type of finger-spelling system relies on the structure of a letter of a particular country. For instance, there are finger spellings that use one hand (one-handed) and two hands (two-handed) for finger spelling alphabets. The one-handed finger spelling is used in USA, France, and Russia. The two handed finger spelling is used in the Czech Republic and United Kingdom (Karpov, Kipyatkova, & Zelezny, 2016).

The 10 most popular SLs according to the Ethnologue catalogue is listed based on number of native signers (as first language) (Karpov *et al*., 2016) are as follows:

- Chinese SL – 20 million signers
- Brazilian SL – 3 million signers
- Indo-Pakistani SL – 2.7 million signers
- American SL (ASL) – 500,000 signers

- Hungarian SL – 350,000 signers
- Kenyan SL – 340,000 signers
- Japanese SL – 320,000 signers
- Ecuadorian SL – 188,000 signers
- Norwegian-Malagasy SL – 185,000 signers
- British SL (BSL) – 125,000 signers

In this paper, a comprehensive review of the existing works on sign language synthesizers is presented. Various methods such as Natural Language Processing, statistical methods, and speech recognition are discussed. The aim is to provide a platform for investigating the strength and limitations in developing SL synthesizer for further research to improve the existing methods. The organization of the paper is as follows. Section 1 provides the introduction of the papers, where the objective and basic ideas are presented. Section 2 discusses the general concept of SL synthesizer. In this section, certain main algorithms are also described. Section 3 elaborates the algorithms for SL synthesizers where NLP and Speech processing algorithm are involved. Section 4 and 5 show the analysis and conclusion, respectively.

II. Sign Language Synthesizers

The development of sign language synthesizer based on animated avatar has been done for some sign languages from all over the world. There were nine projects on SL that can deliver sign animation from the text (Pyfers, 2011). Such system was applied in Italy for Italian Sign Language (Atlas, 2011), Australia for Australian Sign Language (Auslan) (Wong, 2004), USA for American Sign Language (ASL) (Wolfe, McDonald, & Schnepp, 2011), French for French Sign Language (FSL) (Duarte, Kyle, & Gibet, 2011), South Africa for South African Sign Language (SASL) (Zijl & Olivrin, 2008), United Kingdom (Murph, 2007) and Greece (Eleni Efthimiou, 2009). The subsequent sections discuss those systems.

A. Automatic Translation into Sign Language (ATLAS)

Automatic translation to sign language - ATLAS is a sign language synthesizer for Italian Sign Language (LIS). The project involved learning and realizing the system which is capable of translating phrases from spoken Italian to sign language. This operation requires language analysis in linguistics, structure and semantics. It also includes the form of development - intermediate LIS written variants and the creation of virtual actors who will act as translators. The development of modules and full system maintenance modules will enable including the ATLAS translation system in various contexts, which also cite various communication channels. The system should pay attention to the maintenance of data, process and ensure it can be accessed in various evacuation areas so that the signal is described by the user (Atlas, 2011).

B. Auslan Tuition System

This system has been developed to perform Australian Sign Language (Auslan) tutorial. It provides a user-friendly system that provides interactive tutorial for learning the Australian Sign Language (Auslan). Such tutorial program (shown in Figure 1) allows the user to learn the Auslan, and the sign editor program allows the user to input the new database of sign language (Wong, 2004).

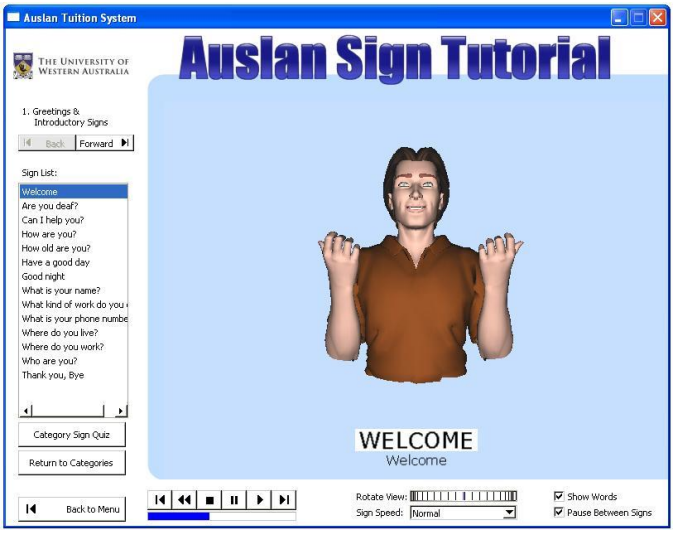

Figure 1 Auslan Tutorial Program – Tutorial Mode (Wong, 2004)

The Auslan tutorial system provides the interactive learning process to understand:

- The Auslan finger spelling
- Number
- Dialogue

The main features in the sign editor program include

- Personalize avatar poses
- Personalize avatar hand shapes
- Personalize avatar animations

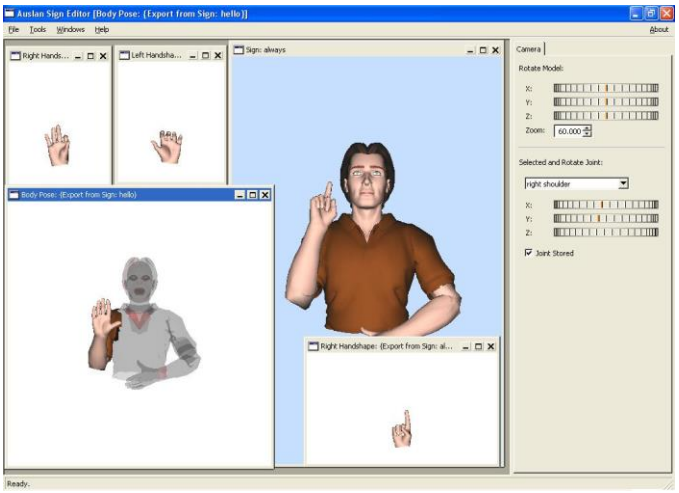

Figure 2 Auslan Sign Editor (Wong, 2004)

The Auslan Sign Editor (shown in Figure 2) system uses the Human Modeling System to provide the 3D animation of the human body. It consists of the Human Modeling Module that provides input and output XML definition which describes the model hierarchical object, Rendering Module displays the position together with the kinematic tree graph, while the Model Interpolation Module provides input levels, outputs and high level animation controls between one set of key frames Model (Yeates, Holden, & Owens, 2003).

C. DePaul ASL Synthesizer Project

The system was developed to synthesize voice input in English language into American Sign Language (ASL). The structure of ASL and its grammar are different from English (British Sign Language – BSL). The system is aimed to provide hearing/speech impaired people greater access to the hearing world.

The system allows the animation to adapt visual or gesture language by using hand shape, hand position, palm orientation, and non-manual signal. The main objective is to produce the animation that grammatically understandable and natural. The main idea was designed due to two critical reasons. First, imperfections in speech output distracted from the message that the signer showed. Poor visual quality adds to the pressure on viewers, and viewers easily get tired and frustrated (Wolfe, Cook, McDonald, & Schnepp, 2011).

Figure 3 The Screenshot of Simulated Avatar.

The system utilizes motion capture that has been used recently as an alternative way to exhaustive of hand animation. The system captures the movement of humans in real-time and the data would be saved in order to be used on the 3D animation. The process of using the 3D animation data is known as retargeting. This process is developed to adjust the position and movement into different artist of the animated avatar.

The system performs several tasks to achieve the naturally understandable sign language (as shown in Figure 3). Such tasks are:

- The articulator module allows the precision to point the position.
- The verb agreement allows the system to have similar signing method from the exemplars.
- Retargeting allows the adjustment of 3D animation data into some different object of the signing avatar.
- Face Rig and Simulated Wrinkles allow the movement and animation of face expression.
- Co-occurring Non Manual Signal.

Furthermore, the on-progress tasks are now phonetics, and timing. These processes are expected to provide better 3D animation of ASL and also the more understandable sign language animation.

D. DictaSign

DictaSign is based on the development of Web 2.0 where interaction of the people may be made constantly. The interaction is made by written input text, hence it is unfriendly for some users. The SL users are the groups who are affected by this model.

The sign language video cannot provide the interaction of the people with some disabilities, since the video has two problems. The video is accessible publicly, it means that the contributors can be recognized by public which may hold people back to contribute. Furthermore, since the video is recorded, editing is not possible to a video that has been produced (Eleni Efthimiou, 2009).

The Dicta-Sign provides an animated avatar as a solution. The animated avatar is anonymous. The avatar signing is uniform in terms of signing style which relates to the flexibility altered and expanded upon by any sign language user. Three proof-of-concepts are evaluated by Dicta-Sign (E. Efthimiou *et al*., 2010):

- A Search-by-Example system integrated isolated sign with interfaces for searching an existing lexical database.
- An SL-to-SL translation prototype.
- A Sign-Wiki will be developed for providing the same service as a traditional Wiki but using sign language.

E. SASL-MT

The South Africa Sign Language Machine Translation is the system used as practical tool for English text to a range of languages. This system to increase the sign language literacy rate in the underprivileged South African communities by utilizing the matching translation which has many diverse research areas (Van Zijl, 2006).

The SASL-MT project was done and some of the completed works are as follows (Zijl & Olivrin, 2008).

- The creation of graphical signing avatar.
- Linguistics aspects of SASL.
- Textual description of sign language.
- Data building management and data entry interface.

F. Say it Sign it (SiSi – IBM)

The Say It Sign It (SISI) is the product delivered by IBM Computer. It creates the speech to sign language translator in British Sign Language. The translation uses the text and translates it into the gestures used in sign language and animates an avatar that performs the signing. (Murph, 2007). SiSi unites a number of computer technologies. The voice recognition module changes the spoken word into text, which then interprets SiSi to be the signal used to turn an avatar that is signed on BSL.

Figure 4 An IBM Avatar in SiSi IBM System

Figure 4 shows the signing avatar of IBM. The avatar in this system is customizable, which means that users are allowed to select the appearance and size of the avatar.

In this section, the available system of sign language synthesizer has been delivered. It can be concluded that from the six currently established and ongoing sign language generator system, real time and natural signing can be achieved. However, the portability and simplicity have not reached the satisfactory level. The summary of literature review for prior sign language generator system is presented in Table 1.

Table 1 Summary of Available SL Synthesizer

Name	Mechanism	Output	Advangtages	Limitation	Number Database
ATLAS (Atlas, 2011)	➢ AEWLIS structure for linguistic input ➢ Animation track for animation system	Signing Avatar	➢ Real-time signing ➢ Natural sign	➢ Only for Italian ➢ Not mobile	Limited for Italian Sign Language
Auslan Tuition System (Wong, 2004)	➢ Human modeling system ➢ Use OpenGL to render	Signing Avatar	➢ User-defined input ➢ Able to update the database ➢ Has 2 components: Tutorial and Sign Editor Program	➢ Only for tutorial ➢ Not able to translate automatically ➢ Commercially released ➢ Limited database, based on package	Limited Database Number, based on the released version
DePaul ASL Synthesizer ("American Sign Language")	➢ Motion capture ➢ Retargeting ➢ Wrinkle simulation	Signing Avatar	➢ Real-time signing ➢ Natural sign ➢ Face expression	➢ Not mobile ➢ Problem with timing and phonetics	N/A (On-going project)
DictaSign (Eleni Efthimiou, 2009)	➢ Use SiGML ➢ Linguistic modeling ➢ AnCoL in Annotation Tools ➢ Sign Language Corpora	Signing Avatar	➢ Multi-sign Language ➢ Natural signing ➢ Non-manual signing	➢ Not mobile	1500 per SL
SASL-MT (Zijl & Olivrin, 2008)	➢ Graphical signing avatar ➢ Linguistic aspects ➢ Data building and data entry tools ➢ H-Anim standard	Signing Avatar	➢ Pluggable avatar ➢ Reusable avatar	➢ Under development	N/A (On-going project)
SiSi-IBM (Murph, 2007)	➢ Voice-to-Sign translation	Signing Avatar	➢ Real-time translation	➢ Not mobile ➢ Only for BSL ➢ Commercially released	Unlimited

III. General Approaches for Language Processing

The SL synthesizer converts the voice input into the SL motions. In general, the process is as shown in Figure 5. It consists of three stages, the voice input acquisition, processing stage, and output processing (recorded video or animated avatar). The first stage is the voice input acquisition in which the speech or audio input is captured (see Figure 5). The second stage is the speech processing stage. This part can be done using two different approaches, involving machine learning and voice processing (see Figure 6 and Figure 7).

Figure 5 General Processing of Sign Language Synthesizer (involving NLP)

The first approach implements machine learning to process the input language, this process allows the identification of the words of interest, then, reorganize them into a particular order (Figure 6). The second approach involves the pre-recorded videos that only includes the speech recognition module and the pre-recorded videos stored in the database. The second approach is limited to particular sentences and available only for the sentences recorded at the database (Figure 7).

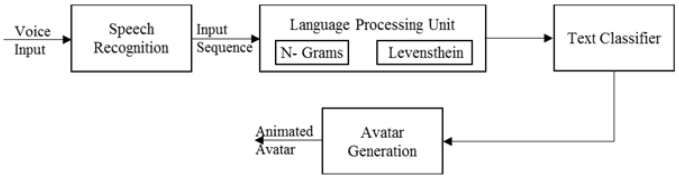

Figure 6 General Processing of Sign Language Synthesizer (involving NLP only)

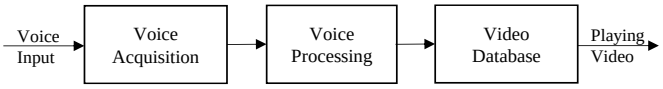

Figure 7 General Processing of Sign Language Synthesizer (involving Voice Processing and Pre-recorded video)

Figure 6 shows the general processing of SL synthesizer involving NLP only. The language processing unit stage plays an important role in processing the input sequence (in text form) into SL grammatical form. This process allows identification of the selected words in the input sequence, since the SL grammatical form only needs the basic form of the words from the input signs, *i.e.*, no tenses required, for examples, "*makan*" (eat), "*minum*" (drink).

Meanwhile, Figure 7 shows the general approach of SL synthesizer using voice processing and pre-recorded video. This process implements voice processing which allows the recognition and processing from the voice input and matched with the database. The main focus of the approach is located in the voice processing and database matching.

The speech processing approach in SL Synthesizer affects the computational load of the SL synthesizer system, processing time, and successful rate of sign language synthesizing. Characteristics of various Sign Languages also affect the complexity of the SL synthesizer. It is important to translate from simple to complex spoken words to signs in real time.

The third step of SL synthesizer is displaying the signs through video stream/animation. It is the final step which allows the HSI people to understand the meaning of the speaker. In this step, there are two approaches that have been

implemented on the system. The first approach is playing the animated avatar where the avatar plays the sequence given by the system and allows the dynamic animation (i.e. allows different input sequences, as long as the input is listed in the database). The second approach is playing the pre-recorded video where the stored video will be played whenever the matching signs are found in the database and it is only applicable for the static input sequence (as the SL video is pre-recorded only for particular sentences).

The animated avatars are the efficient approach for analysis of sign language and finger spelling. Some research has been conducted to develop animated avatar and sign language machine translation. It was conducted in SIGNSPEAK EU project for the Europe and in Dicta-Sign EU project for the USA. The other projects are SignCom EU project, DePaul ASL Synthesizer and SiSi (Say It Sign It) for British Sign Langauge. (Karpov *et al*., 2016)

IV. The Algorithm for Sign Language Synthesizer

The main idea of SL Synthesizer is identified by the main components, *i.e.*, speech recognition and processing. There are numerous approaches existed for SL synthesizers which will be considered in this paper. The selection of approaches reviewed is based on the complexity and algorithm used in the developed systems. Those methods were developed by using speech recognition with pre-recorded video or more complex algorithm which implemented natural language translation algorithm.

There are two main approaches of SL synthesizers reviewed in this paper, *i.e.*, those which implement Natural Language Processing (NLP) and those with direct Speech Recognition Algorithm. The first approaches which involve Natural Language Processing (NLP) include NLP basic process, NLP with gloss based approach, and NLP with the rule based translation and statistical translation (Section IV.A-IV.E). The second approaches implement speech recognition algorithm and involve the prerecorded video (section IV.F). Analysis and summary of the above mentioned approaches are described together with the advantages of each method.

A. NLP Basic Processing

The basic process of NLP that can be used for SL synthesizer is the simplest algorithm which has been implemented. This step only involves three basic processes, *i.e.* Part of Speech (POS) Tagger, optimizer, and stemming. Figure 8 shows the step of the processing (Joy & Balakrishnan, 2014).

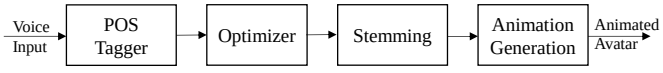

Figure 8 Algorithm of NLP Basic Processing (Joy & Balakrishnan, 2014)

The first stage is the POS tagger which involves the morphological analysis. The POS tagger reads the text input and tagging it into their grammatical type, such as noun, verb, adjective, and etc. As the output forms the POS tagger, the optimizer takes part to remove particles of the sentences and unnecessary words. Furthermore, the stemming process is applied to find the basic form of the words. SL only needs the basic form of the words and in the present tense form. Finally, the animation generation works as an agent to form the animated avatar (Hanke, 2004; Kaur & Kumar, 2016; Koller, Bowden, & Ney, 2016).

B. NLP with Gloss Based Approach

The gloss based approach is a method that associates the words and the meanings in a dictionary. The level of the glosses is defined by the order of the language grammar. In a report by Almeida, Coheur, and Candeias (2015), the order of blocks (or glosses) is calculated according to Portugese Sign Language (LGP) grammar. As the final step, the order of blocks will be converted into sign order.

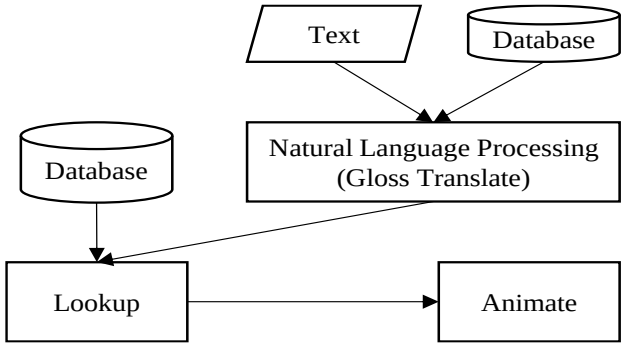

Figure 9 Gloss Based Approach Algorithm (Almeida *et al.*, 2015)

Figure 9 shows the gloss based approach by Almeida *et al*. (2015). The text will be associated with the dictionary to be processed in natural language processing block, where the output will be translated to sequence of glosses and actions.

Figure 10 shows the input to the NLP is a sequence of words/text. The input is split into words or token. Then, possible orthographic errors are revised, and a basic approach consults with the dictionaries to find the words that are translated, and returns the corresponding actions to the signing avatar.

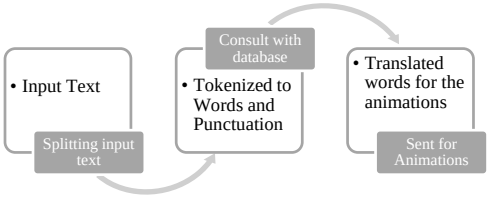

Figure 10 Step of Natural Language Processing in the NLP-gloss based approach

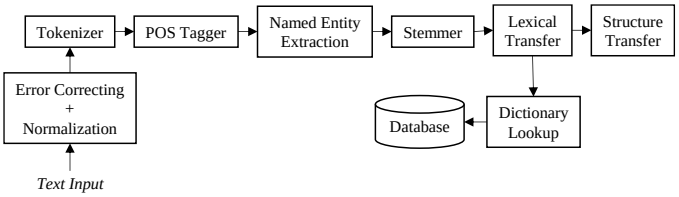

Figure 11 NLP with Gloss Approach (Almeida *et al.*, 2015)

Figure 11 shows the NLP with Gloss Approach. It is possible to get better translation where some additional steps are

available to gain better output. The processing involves Stemmer, POS Tagger, and Name Entry organizer. Stemmer is used for identification of proper stem and suffixes (and estimates). It was used to find possibility in conclusion of status of words, for example the gender and the number of words given. A Part-of-Speech (POS) Tagger may also contribute to translation process by pairing to a stemmer in the identification of different types of affixes. POS tagger typically performs further processing, such as entity identifiers and syntactic analyzers. Named Entity Recognizer enables identification of the person's name and Analog Syntactic if it is important to identify the components of the syntax of the sentence, such as subject and object.

C. NLP with Rule-Based Translation

The rule-based translation provides an analysis from input the word and group of words (sentence). The translation is processed by looking for a particular words and/or signs (or blocks). It starts from individual word search and extends to neighborhood context words (words or word combination) or already-formed signs (blocks) (Rayner *et al*., 2016).

The rule-based translation generates the final translation of the text input. The output depends on the scope of the block relations defined by the rules. The possibility to achieve different compromises is based on the reliability of the translated sign and the robustness against recognition errors.

The translation process requires two steps to deliver the output. The first step is mapping input words into one or more tags (syntactic/pragmatic tags). The tagged words are fed into an algorithm where rules are applied for altering tagged words into signs. It was made by grouping the concepts or signs (or blocks) and defining new signs. The applied rule is developed to cater short and longer sentences where scope relationship between concept and marks are involved. Furthermore, the block order from applied rule is expected to be matched with database. The rule-based translation module provides the translation rules for translation process, for instance, San-Segundo *et al*. (2008) utilizes the algorithm which contains 153 translation rules, and Inurrieta, Aduriz, de Ilarraza, Labaka, and Sarasola (2017) observed the algorithm with 89 translation rules.

There are four general performance measurement tools which are commonly used: SER (Sign Error Rate), PER (Position Independent SER), BLEU (BiLingual Evaluation Understudy). The SER and PER are error measures whereas BLEU is the accuracy measure (Inurrieta *et al*., 2017; Rayner *et al*., 2016; San-Segundo *et al*.,2008; San Segundo *et al*., 2010).

Table 2 Results Obtained with Rube Based Translation Module (San-Segundo *et al.*, 2008)

	SER	PER	BLEU	NIST
Experiment 1 (SR)	31.60	27.02	0.5870	7.0945
Experiment 2 (TR)	24.94	20.21	0.6143	7.8345
Experiment 3 (SR)	18.23	14.87	0.7072	8.4961
REF (TR)	16.75	13.17	0.7217	8.5992

SR = Speech Recognizer TR = Transcribed Sentences

The samples of final results reported by San-Segundo *et al*. (2008) are presented in Table 2. SER is higher when voice recognition output is used instead of transcribed sentences. The reason is that speech recognition introduces recognition errors resulting in more translation mistakes: improvements in incorrect markings and BLEU decreases.

San-Segundo *et al*. (2008) reports that the error obtained during the experiment happened due to the omitting of subject in conversational sentence where in Sign Language is compulsory, several possible translations, and a large number of unknown words in database which generates a significant number of errors.

D. NLP with Statistical-based Translation

Statistical-based translation approach is an approach which calculates the probability between the word sequence and sign sequence stored in a database as the reference. One of methods in statistical translation is Phrase-Based translation. Figure 12 shows the diagram of phrase based translation module San-Segundo *et al*. (2008).

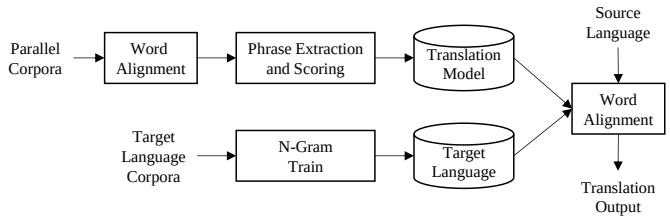

Figure 12 Diagram of the phrase-based translation module (San-Segundo *et al.*, 2008)

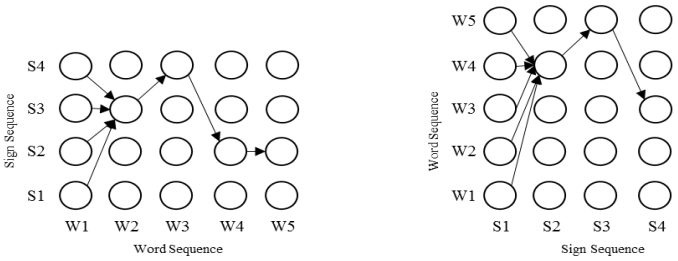

Figure 13 Alignments in Both Directions: words-signs and signs-words (San-Segundo *et al.*, 2008)

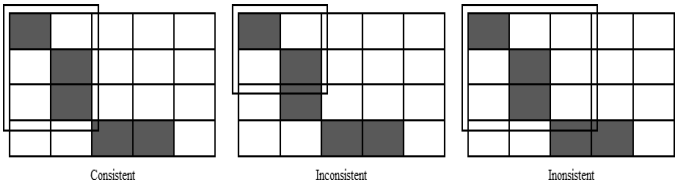

Figure 14 Examples of Phrase Extraction

The translation process uses a translation model based on phrases and a target language model. As reported in San Segundo *et al*. (2006) and San Segundo *et al*. (2007), the GIZA++ software (Och & Ney, 2000) has been used to calculate the alignments between words and signs:

- GIZA++ provides a lexical translation model and the translation probability between every word and sign. (word-sign relationship)
- Koehn, Och, and Marcu (2003) reports on phrase extraction algorithm. The collection of all phrase pairs is made if there are consistencies with the word alignment. (Figure 14).

- Phrase scoring is the computation of the translation probabilities for all phrase pairs. The probabilities are calculated: forward and backward.

Table 3 shows the result from statistical translation method. As reported by San-Segundo *et al*. (2008), the statistical translation shows the worst result was from the rule based strategy. This condition is due to its restricted domain and it has been possible to develop a complete set of rules with a reasonable effort.

Table 3 Results Obtained with Rule Statistical Translation Module (San-Segundo *et al*., 2008)

	SER	PER	BLEU	NIST
Experiment 1 (SR)	38.72	34.35	0.4941	6.4123
Experiment 2 (TR)	36.08	32.00	0.4998	6.4865
Experiment 3 (SR)	34.22	30.04	0.5046	6.5596
REF (TR)	33.74	29.14	0.5152	6.6505

SR = Speech Recognizer TR = Transcribed Sentences

E. The Multimodal Sign Language Synthesizer

The Multimodal SL synthesizer takes an input text and translates the input text into sign language for the HSI people and audio-visual speech for the non-HSI. The components available in multi-modal synthesizer system are listed below:

- Text Processing Model.
- Control selection of HamNoSys codes.
- TTS (text-to-speech) systems for Czech (Tihelka, Kala, & Matoušek, 2010) and Russian (Hoffmann *et al*., 2007).
- 3D model for Head and The Upper Body. (Železný, Krňoul, Císař, & Matoušek, 2006) (Krňoul, Kanis, Železný, & Müller, 2007).
- The controller for the audio-visual talking head which is made in order to synchronize lips movements with the input speech signal (Karpov *et al*., 2009; Krňoul, Železný, Müller, & Kanis, 2006).
- The user interface. This interface integrates the components in the signing avatar (Karpov, Krnoul, Zelezny, & Ronzhin, 2013). The integrated components are automatic generation of SL gestures, auditory, and visual speech.

The methodology approach in this system is a 3D model of talking head. The text-based system and controlled visual processing are considering output of processed input text and speech input with integration with asynchronous modal modality. It is based on the parametric-controlled 3D model of the human head where the moving part is controlled by control points (Železný *et al*., 2006).

Figure 15 shows the model of talking head represented by virtual space uniform. It is connected by edges in order to build solid triangle. The retrieved 3D data is processed to add other 3D face model. This is saved as virtual reality modelling language (VRML). The 3D head model is illustrated by many knots but less number of active knots which is only controllable by software. The parametric components have been added to complete the 3D Talking head. The additional components are eyes, mouth, and tongue. They are developed based on anthropological physiological knowledge and controlled by the software.

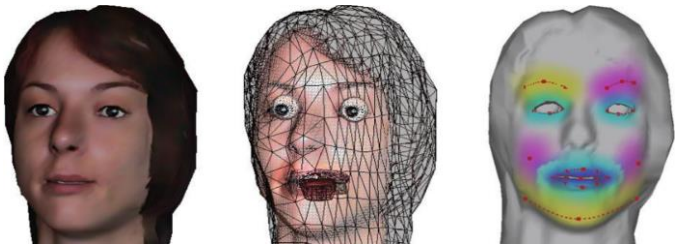

Figure 15 Talking Head (Karpov *et al*., 2013)

The purpose of automatic signal language synthesizer is to simulate human behavior when signing. Signal language syntheses are implemented in several levels. First, the input speech is translated into the appropriate order. Then the relevant signs are combined to form continuous speech. The synthesis module combines the conversion algorithm for Hamburg Notation System (HamNoSys) (Hanke, 2004) to create required SL and finger spelling gestures (Krňoul *et al*., 2007).

HamNoSys allows illustrating manual gestures using only four components for both hands: (1) hand shape; (2) hand/palm orientation; (3) location of hand; (4) type of movement. For each of these components, there is a symbol of its own written. Figure 16 presents an example of an official handwritten signature by HamNoSys (Russian "C" letter used in fingerspelling). The algorithm automatically changes the HamNoSys code to control the path and receives the most legitimate combination of symbols. The final animation frame is an input to the animated model (Karpov *et al*., 2016).

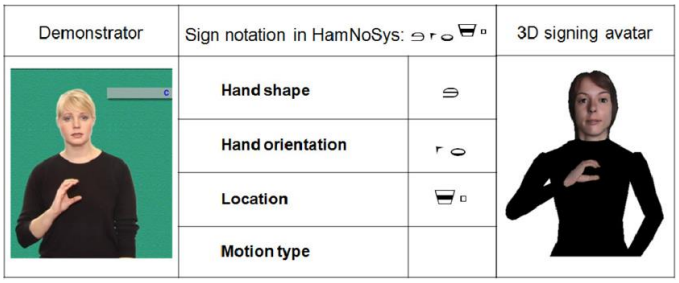

Figure 16 Manual Gestures by the HamNoSys Notation

The signing avatar provides animation of 3D model of the top part of human body. The baseline system incorporates 3D articulator models that resemble the human skin surface of a polygonal network. The mesh is divided into body segments depicting the arms, forearms, palms, abdominal bone and part of the talking head model. The system is designed to provide manual and non-manual signing components.

The manual component refers to the expression of hand motions to do signing by moving body segment rotation. The body segments are connected by joints and arranged in a hierarchy into the tree structure (i.e. approximate body frame). Each joint is attached to at least one body segment. Thus, the rotation of one body segment causes the rotation of the other

body segments to be in the lower hierarchy (Krňoul *et al*., 2007). In addition, the synthesis of non-manual components uses the second control through a more ordinary head-to-head model or target goal. Therefore, joints ensure shoulder, neck, skull, eyebrows and jaw movements (Krňoul *et al*., 2007).

A qualitative user rating system has been recommended with the help of several teachers, and they positively estimate the intelligence and articulation of the handwriting lips and the manual handwriting intelligence of the signatory avatar. Avatar reviewers of future researchers can be used in a variety of technology aids for human-computer interaction (Karpov *et al*., 2011), in multimodal information kiosks, interactive dialog systems, etc.

F. Voice to Text using Voice Recognition Method

This method implements the voice recognition algorithm. In general, the algorithm will detect the voice and match it with that in the database. The database contains the pre-recorded signing video, and the video is played once the data is matched with the database listed.

Figure 17 shows the block diagram of voice to text algorithm in voice recognition approach. Foong, Low, and La (2009) proposes an algorithm for speech to sign translator. It uses pre-recorded sign. The system comprises three main methods, *i.e.* the main component is the sound recording module (with training algorithm), MFCC for digital signal processing, and vector quantization.

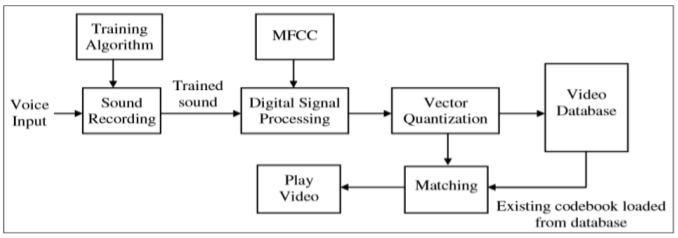

Figure 17 Voice to Text Algorithm (Foong *et al*., 2009)

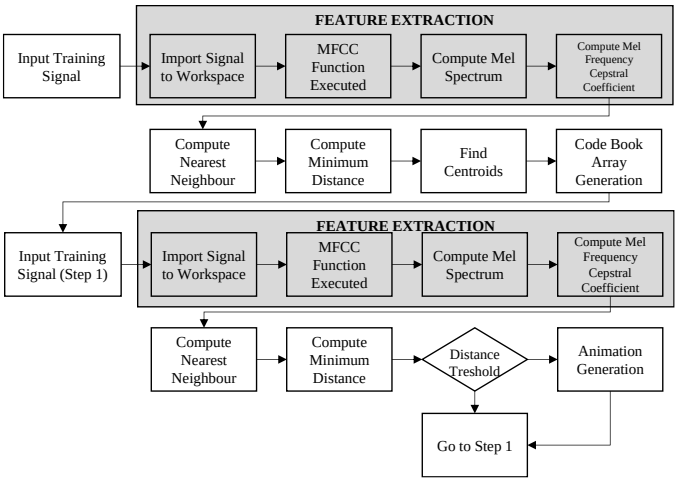

Figure 18 Synthesizer algorithm by Kasaei and Kasaei (2011)

Kasaei and Kasaei (2011) proposes another algorithm for voice to sign using voice recognition method as illustrated by Figure 18. It uses Feature Extraction and Feature Matching. Those features are used to find the characteristics of the voice input. The characteristic extracted and matched provides the information of the voice and links to the pre-recorded video in database.

Sound recording is a process to record sound which is captured by microphone. The quality of the recorded sound depends on the recording software and device. The quality can be enhanced by applying proper noise filter to remove some noise during recording. The process of recording allows some repetitions so that captured voice is a good quality recorded signal. It is very important to have good recorded sound/voice in order to determine the accuracy of voice recognition process.

The recorded sound is then fed to the Digital Signal Processing (DSP). The DSP module has the role to transform the input from time domain to frequency domain. This is required since the feature extraction of recorded sound is effectively extracted on frequency domain where it can match with the words used. The MFCC algorithm (Mail Frequency Cepstral Algorithm) is used in the extraction process.

Vector quantization is used to perform speech recognition. The concept of vector quantization is to compress each vector of sound/vector characteristics into a scalar vector. This helps the process to reduce space for feature storage and increases the efficiency of matching process where it compares to only one value. Training process is where the trained output will be stored in a codebook in the database where one code book is valid only for trained sound.

Code book data in database is used to compare the input sound for recognition process. The input (voice/voice) is compared to a codebook in the database to capture the corresponding video of sign language where resulting on the value of input. The value of each codebook and voice input is represented by a matrix. The average value of codebook and voice input is calculated. Then, each code book will be compared with the input signal (trained signal).

The fundamental process of this system is to synthesize voice into sign language. This system matches the sound captured with the sign language video recorded in the database to display the appropriate signs to provide alternative interactive communication between ordinary people and the deaf. At present, the prototype allows oral English translation into sign language in the Malaysian context. The accuracy of the system depends on the number of trainings performed to identify all trained instructions or words and carry out correct translation.

For the current performance, the system has the accuracy up to 80.3% (Kasaei & Kasaei, 2011). It is considered as quite good number for the performance analysis where the automated Sign Language through pre-recorded video. Moreover, the system is sufficient to help non-HSI people who are willing to learn Sign Language and to ease communication to HSI people more effectively.

V. Analysis

Various approaches in SL synthesizer have been presented in Section IV. The approaches can be categorized as active and passive approaches. The active approach is referred to the approach which can generate the sign animation based on the input given. It involves the Natural Language Processing as an engine to generate sign stream which will be sent for animation.

The NLP processes the input words (from speech recognition module) by removing unwanted words and gaining the important words for providing the sign stream.

The stages of involving NLP are started from the basic NLP structures to more complicated NLP systems. The implementation of each NLP method affects the efficiency and successful rate of sequence input word processing. Better results were obtained by the approach where NLP methods were implemented and the animated avatar was the compliment of the systems. On the other hand, the cost of computational works affects the performance of the systems. Based on the provided description in the previous section, the more NLP structures involved on the systems, the higher computational works become, and so does the computational time.

In Joy and Balakrishnan (2014), a 3D computer generated model is used for animation. The animation was generated using Linux, which was modeled as a sequence of key frames. The developed systems reduce the complexity of the SL synthesizer system since the process involves direct conversion from input to the SL animation. As the complexity being reduced, the systems can be ported to mobile handset, tablets, and computer.

The Gloss based Approach has been reported by Almeida *et al*. (2015). The prototype that combines NLP modules and animation have been presented to synthesizer as signing avatar of LGP (Portuguese Sign Language) utterances, given a text input in European/Portuguese. Preliminary evaluation has been conducted with the deaf community from which a positive response has been obtained.

San-Segundo *et al*. (2008) reports two approaches for sign language synthesizers. The rule-based translation module has provided a result which is reaching a 31.60% for SER (Sign Error Rate) and a 0.5780 for BLEU (BiLingual Evaluation Understudy). The second approach involves a statistical translation, where parallel corpora were used for training process, and has provided configuration which results in 38.72% of SER and 0.4941 of BLEU.

The passive approach can be described as the approach that only matches the input with the database and runs the pre-recorded video. This approach allows limited implementation of the static synthesis of sign language. It involves sophisticated speech recognition algorithm so that the intended database can be run based on the input given from the audio source. It involves MFCC and Vector Quantization method which help the process of input matching. These methods may cost the computational time of the system.

Kasaei and Kasaei (2011) provides the summary of the SL synthesizers which translate human voice to Sign Language. The captured video is matched to the pre-recorded Sign Language videos stored in the database. The database provides the SL animation video with the accuracy of 80.3%. This system can be utilized by many who wish to learn sign language independently and to help communication process with hearing/speech impaired people.

Based on the analysis of six existing approaches, we can conclude that the approaches can be categorized into active and passive category. These two categories have their own benefits and limitations. Moreover, there is always a tradeoff between the computational load and the processing time. On the other hand, higher computational load provides better output of SL synthesizer. The summary of the approaches is presented in Table 4.

Table 4 Summary of Approaches for SL Synthesizer

SL Synthesizer	Output	Advantages	Limitation
NLP Basic Processing	Animated Avatar	➢ Omits certain words (particle and conjunction word ➢ Sequence of Words ➢ Animation based on the output (sequence of words)	➢ Not able to differentiate a clause ➢ Higher Computational Works
NLP With Gloss Based	Animated Avatar	➢ Understand the clause (two words or more) ➢ Animation based on the output (sequence of words)	➢ Limited clause database ➢ Higher Computational Loads
NLP With Rule Based Translation	Animated Avatar	➢ Finding specific combinations of words / sentences ➢ Generates the final translation	➢ Higher Computational Works
NLP With Statistical Translation	Animated Avatar	➢ Finding specific meaning of words / sentences by comparing the similar database ➢ Generates the final translation	➢ Higher Computational Works
The Multimodal	Animated Avatar	➢ Multimodal Approach ➢ Wider Implementation ➢ More words and sentences	➢ Complicated Systems ➢ Not an integrated systems
Voice Recognition Approach	Playing Recorded Video	➢ Direct Translation ➢ Low Computational Load	➢ Limited Database ➢ Applied into some particular words / sentences

VI. Conclusion

The computer-based research and development in the area of SL synthesizer has been presented. The available SL synthesizers were developed in Australia, Europe, South Africa, and United Kingdom. The ATLAS is used for Italian Sign Language, Auslan is for Australian, DePaul translates the American Sign Language (ASL), SASL is for South African, and SiSi - IBM works for British Sign Language (BSL). Those available SL synthesizers have been limited to their own SL and to some particular application and with limited mobility (if any). The benefits of those SL synthesizers are various, but the common advantages of the available systems is the ability to synthesize dynamic SL by providing animated avatar.

The general approach of SL synthesizer was categorized into two categories, e.g. active and passive category. The active category refers to an approach where the dynamic SL is synthesized and represented by an animated avatar. This involves the NLP as an integral part for the synthesizer. This category requires higher computational load and longer processing time. Meanwhile, the passive category is simpler and provides direct translation as it recognizes the matched input with the database and run the pre-recorded video. This refers to the static synthesis of sign language and is applicable to some particular implementation. This category also provides lower computational load, but limited to lower number of translation or smaller database.

REFERENCES

[1] Almeida, I., Coheur, L., & Candeias, S. (2015). Coupling natural language processing and animation synthesis in portuguese sign language translation. Vision and Language.

[2] American Sign Language.). Retrieved 23 March, 2012, from http://asl.cs.depaul.edu/

[3] Atlas. (2011). ATLAS - Automatic Translation into Sign Language. Retrieved 23 March, 2012, from http://www.atlas.polito.it/index.php

[4] Dasgupta, T., & Basu, A. (2008). Prototype machine translation system from text-to-Indian sign language. Paper presented at the Proceedings of the 13th international conference on Intelligent user interfaces.

[5] Duarte, Kyle, & Gibet, S. (2011, 10-11 January). Presentation of the SignCom Project. Paper presented at the Proceedings of the First International Workshop on Sign Language Translation and Avatar Technology, Berlin, Germany.

[6] Efthimiou, E. (2009). DictaSign. Retrieved 23 March, 2012, from http://www.dictasign.eu

[7] Efthimiou, E., Fotinea, S.-E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., . . . Goudenove, F. (2010, 17 - 23 May). Dicta-Sign: Sign Language Recognition, Generation, and Modelling: A Research Effort with Applications in Deaf Communication Paper presented at the the Language Resources and Evaluation Conference Workshop on the Representation and Processing of Sign Languages : Corpora and Sign Languages Technologies, Valetta, Malta

[8] Foong, O. M., Low, T. J., & La, W. W. (2009). V2S: Voice to Sign Language Translation System for Malaysian Deaf People Visual Informatics: Bridging Research and Practice (pp. 868-876): Springer

[9] Hanke, T. (2004). HamNoSys-representing sign language data in language resources and language processing contexts. Paper presented at the LREC.

[10] Hoffmann, R., Jokisch, O., Lobanov, B., Tsirulnik, L., Shpilewsky, E., Piurkowska, B., . . . Karpov, A. (2007). Slavonic TTS and SST Conversion for Let's Fly Dialogue System. Paper presented at the 12th international conference on speech and computer SPECOM, Moscow, Russia.

[11] Inurrieta, U., Aduriz, I., de Ilarraza, A. D., Labaka, G., & Sarasola, K. (2017). Rule-Based Translation of Spanish Verb+ Noun Combinations into Basque. MWE 2017, 149.

[12] Joy, J., & Balakrishnan, K. (2014). A prototype Malayalam to Sign Language Automatic Translator. arXiv preprint arXiv:1412.7415.

[13] Karpov, A., Kipyatkova, I., & Zelezny, M. (2016). Automatic Technologies for Processing Spoken Sign Languages. Procedia Computer Science, 81, 201-207.

[14] Karpov, A., Krnoul, Z., Zelezny, M., & Ronzhin, A. (2013). Multimodal synthesizer for Russian and Czech sign languages and audio-visual speech. Paper presented at the International Conference on Universal Access in Human-Computer Interaction.

[15] Karpov, A., Ronzhin, A., & Kipyatkova, I. (2011). An assistive bi-modal user interface integrating multi-channel speech recognition and computer vision. Paper presented at the International Conference on Human-Computer Interaction.

[16] Karpov, A., Tsirulnik, L., Krňoul, Z., Ronzhin, A., Lobanov, B., & Železný, M. (2009). Audio-visual speech asynchrony modeling in a talking head.

[17] Kasaei, S. H. M., & Kasaei, S. M. M. (2011). Development an Automatic Translate Voice to Sign LanguageAnimation Based-on MFCC and Vector Quantization Method. International Journal of Computer and Electrical Engineering, 3(5), 629.

[18] Kaur, K., & Kumar, P. (2016). HamNoSys to SiGML Conversion System for Sign Language Automation. Procedia Computer Science, 89, 794-803.

[19] Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.

[20] Koller, O., Bowden, R., & Ney, H. (2016). Automatic Alignment of HamNoSys Subunits for Continuous Sign Language Recognition. LREC 2016 Proceedings, 121-128.

[21] Krňoul, Z., Kanis, J., Železný, M., & Müller, L. (2007). Czech text-to-sign speech synthesizer. Paper presented at the International Workshop on Machine Learning for Multimodal Interaction.

[22] Krňoul, Z., Železný, M., Müller, L., & Kanis, J. (2006). Training of coarticulation models using dominance functions and visual unit selection methods for audio-visual speech synthesis. Paper presented at the Interspeech.

[23] Murad, D. (2013, 20 September 2013). MFD: Massive shortage of sign language interpreters, The Star Online.

[24] Murph, D. (2007). IBM's SiSi virtually translates speech to sign language. Retrieved 2 November, 2011, from http://www.engadget.com/2007/09/13/ibms-sisi-virtually-translates-speech-to-sign-language/

[25] Och, F. J., & Ney, H. (2000). Improved statistical alignment models. Paper presented at the Proceedings of the 38th Annual Meeting on Association for Computational Linguistics.

[26] Pyfers, L. (2011). Open Sign from http://www.opensign.org/

[27] Rayner, E., Bouillon, P., Gerlach, J., Strasly, I., Tsourakis, N., & Ebling, S. (2016). An OpenWeb Platform for Rule-Based Speech-to-Sign Translation.

[28] San-Segundo, R., Barra, R., Córdoba, R., D'Haro, L., Fernández, F., Ferreiros, J., . . . Pardo, J. M. (2008). Speech to sign language translation system for Spanish. Speech Communication, 50(11), 1009-1020.

[29] San Segundo, R., Barra-Chicote, R., Luis Fernando, D. H., Montero, J. M., de Córdoba, R., & Ferreiros, J. (2006). A Spanish speech to sign language translation system for assisting deaf-mute people. Paper presented at the INTERSPEECH.

[30] San Segundo, R., López-Ludeña, V., Martín, R., Lutfi, S. L., Ferreiros, J., de Córdoba, R., & Pardo, J. M. (2010). Advanced speech communication system for deaf people. Paper presented at the INTERSPEECH.

[31] San Segundo, R., Pérez, A., Ortiz, D., Luis Fernando, D. H., Torres, M. I., & Casacuberta, F. (2007). Evaluation of alternatives on speech to sign language translation. Paper presented at the INTERSPEECH.

[32] Tihelka, D., Kala, J., & Matoušek, J. (2010). Enhancements of Viterbi search for fast unit selection synthesis. Paper presented at the Proceedings of Int. Conf. Interspeech 2010.

[33] Tomasco, S. (2007). IBM Research Demonstrates Innovative 'Speech to Sign Language' Translation System. Retrieved 23 March, 2012, from http://www-03.ibm.com/press/us/en/pressrelease/22316.wss

[34] Van Zijl, L. (2006). South African sign language machine translation project. Paper presented at the Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility.

[35] Wolfe, R., Cook, P., McDonald, J. C., & Schnepp, J. (2011). Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in American Sign Language. Nonmanuals in Sign Language Special issue of Sign Language & Linguistics, 14(1).

[36] Wolfe, R., McDonald, J., & Schnepp, J. (2011, 10 - 11 Januari). An Avatar to Depict Sign Language: Building from Reusable Hand Animation. Paper presented at the International Workshop on Sign Language Translation and Avatar Technology (SLTAT) Federal Ministry of Labour and Social Affairs, Berlin, Germany.

[37] Wong, J. C. (March 25, 2004). The Auslan Tuition System. Retrieved 22 March, 2012, from http://auslantuition.csse.uwa.edu.au/index.html

[38] Yeates, S., Holden, E.-J., & Owens, R. (2003). An Animated Auslan Tuition System. International Journal of Machine Graphics and Vision, 12(2), 203-214.

[39] Železný, M., Krňoul, Z., Císař, P., & Matoušek, J. (2006). Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. Signal Processing, 86(12), 3657-3673.

[40] Zijl, L. v., & Olivrin, G. (2008, April). South African Sign Language Assistive Translation. Paper presented at the IASTED International Conference on Assistive Technologies, Baltimore, USA.