

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/236634112>

Speech to Sign Language Interpreter System (SSLIS)

Conference Paper · January 2006

CITATIONS

0

READS

1,046

4 authors, including:



Khalid El-Darymli

MacDonald, Dettwiler and Associates Ltd

29 PUBLICATIONS 379 CITATIONS

[SEE PROFILE](#)



Othman Omran Khalifa

International Islamic University Malaysia

446 PUBLICATIONS 2,366 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Feature Parallelism Model for Enhanced Visual Recognition [View project](#)



Sense less Mobile Network [View project](#)

The citation of this paper is as follows:

Khalid El-Darymli, Othman O. Khalifa, and Hassan Enemosah, "**Speech to Sign Language Interpreter System (SSLIS)**", *the IEEE International Conference of Computer and Communication Engineering (ICCCE'06)*, in Proceedings of, Kuala Lumpur, Malaysia, 2006."

Speech to Sign Language Interpreter System (SSLIS)

Khalid Khalil El-Darymli¹, Othman O. Khalifa¹ and Hassan Enemosah¹

¹International Islamic University Malaysia, ECE Dept., Faculty of Engineering
kkkseld@yahoo.com, khalifa@iiu.edu.my and h_enemosah@yahoo.com

Abstract

The deaf and hearing-impaired make up a sizable community with specific needs that operators and technology have only recently begun to target. There is no such freely available software, let alone a single one with a reasonable price to translate uttered speech into sign language in real time. In this paper, this problem was tackled through presenting the "Speech to Sign Language Interpreter System (SSLIS)" to translate uttered English speech into video American Sign Language (ASL) in live mode. In addition to its main task, other interesting features were added to the SSLIS to make it even more comprehensive and beneficial. The Sphinx 3.5 was manipulated as the speech recognition engine for the SSLIS and for translation, ASL syntax was not followed, but rather the Signed English (SE) manual was employed as a manual parallel to English. Rule of operation, parameters optimization and accuracy measurements and snapshots of SSLIS are amongst the topics approached throughout this paper. We believe that SSLIS would facilitate the acquisition of English as a second language for deaf people and help to fill the gap between deaf and nondeaf communities.

1. Introduction

In the US, the number of deaf and hard of hearing people is estimated to be more than 8.6% out of the whole population wherein 5.6% out of them are in the age vicinity of 3 to 34 years old [7]. The commercial market was and still working on developing software that could fill the gap between deaf and nondeaf communities in the sense that it facilitates the communication amongst them and helps deaf people to improve their quality of life through translating the spoken speech to text and sign language. In this context there is only one such software that is already commercially available in the market; however, it poses a lot of burden on the deaf people since they have to pay a sizeable amount of money to purchase it apart from the fact that they would always be restricted to the developer to pay extra money for any updates [10]. Accordingly, the main motive for developing our software is to tackle this very problem and to attract the

researches attention to this area for the benefit of deaf people.

Figure 1 depicts the basic structure of our software. Live uttered input speech is captured through microphone then it is translated to text through some speech recognition engine. The speech engine we manipulated for this purpose is the Sphinx 3.5. The recognized text will be input to an ASL database on a word basis looking for a match. The database contains a certain number of prerecorded video signs where mainly there is one video clip per each basic word. If match occurred, the equivalent ASL translation will be displayed following the Signed English (SE) manual as a parallel to English rather than following the ASL syntax. Otherwise, the word will be fingerspelled. Finally, both recognized text and ASL translation will be displayed concurrently as a final output.

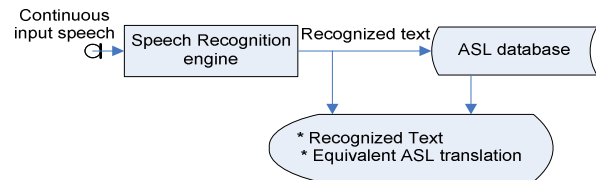


Figure 1: Basic structure of the SSLIS

2. Automatic speech recognition

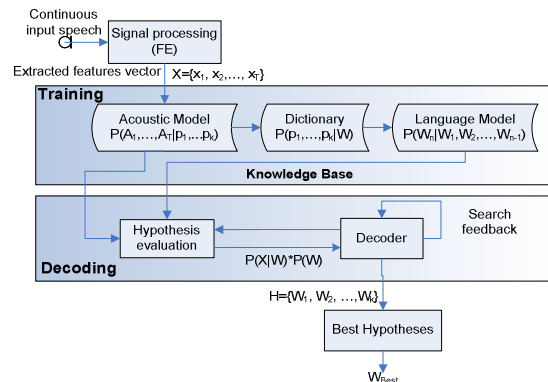


Figure 2: General structure of LVCSR

Automatic Speech Recognition (ASR) is the process of converting an acoustic signal to a textual message. The expected task of our software entails using a large

vocabulary, speaker independent and continuous speech recognizer. Figure 2 shows a simplified structure of large vocabulary continuous speech recognizer (LVCSR) [3, 4, 8, 9, 11, 12, 15, 16, and 17].

2.1. The Sphinx 3.5

This is the name of the SR engine we manipulated for developing the SSLIS. Sphinx originally started at CMU (Carnegie Mellon University) and then it has been released as open source software [6]. Sphinx is a fairly large program that offers a lot of tools and information. It is still in development but already includes trainers, recognizers, acoustic models, language models and some limited documentation. Sphinx 3 works best on continuous speech and large vocabulary and has been tested during the NIST (National Institute of Standards and Technology) evaluation campaigns [13]. Sphinx 3 does not provide any interface in order to make the integration of all components easier. In other words, it is a collection of tools and resources that enables developers/researchers to build successful speech recognizers. And actually this was our job. The most recent version of Sphinx 3 is Sphinx 3.5 [18, 19 and 20].

3. Sign language

Sign Language is a communication system using gestures that are interpreted visually. Many people in deaf communities around the world use sign languages as their primary means of communication. These communities include both deaf and hearing people who converse in sign language. But for many deaf people, sign language serves as their primary, or native, language, creating a strong sense of social and cultural identity. Languages can be conveyed in different ways known as modalities. Such modalities include speech and sign. To clarify, English and Malay languages share the same modality, speech, though they are totally different languages. Similarly, American Sign Language (ASL) and British Sign Language (BSL) share the same modality, sign, but they are totally different languages. This explains the fact that sign language is different from country to country and even it could have dialects that are varying from region to region within the same country [1 and 21].

3.1. American Sign Language (ASL)

ASL is the dominant sign language in the United States, anglophone Canada and parts of Mexico. Currently, approximately 450,000 deaf people in the United States use ASL as their primary language [14]. ASL signs follow a certain order, just as words do in spoken English. However, in ASL one sign can express meaning that would necessitate the use of several words in speech. For example, the words in the statement “I

stared at it for a long time” each contain a unit of meaning. In ASL, this same sentence would be expressed as a single sign. The signer forms “look at” by making a V under the eyes with the first and middle fingers of the right hand. The hand moves out toward the object being looked at, repeatedly tracing an oval to indicate “over a long time” [1].

The grammar of ASL uses spatial locations, motion, and context to indicate syntax. For example:

- If a signer signs a noun and then points to a certain spot, he or she can refer back to that noun by pointing again to the same spot.
- To intensify the meaning a verb or adjective (e.g., to say “very calm” instead of “calm”), the signer modulates the way it is expressed, first holding his or her hands rigid and then making the rest of the sign more quickly than usual.
- Raised eyebrows can indicate a yes-or-no question, while lowered eyebrows indicate a wh-question [1].

3.2. American sign language alphabets

The Sign Language alphabet is a manual alphabet. Manual alphabet is a system of representing all the letters of an alphabet, using only the hands. Making words using a manual alphabet is called *fingerspelling*. Manual alphabets are a part of sign languages. For ASL, the one-handed manual alphabet is used. Fingerspelling is used to complement the vocabulary of ASL when spelling individual letters of a word is the preferred or only option, such as with proper names or the titles of works. Letters should be signed with the dominant hand and in most cases, with palm facing the viewer. The American manual alphabet is depicted in Figure 3 [2].

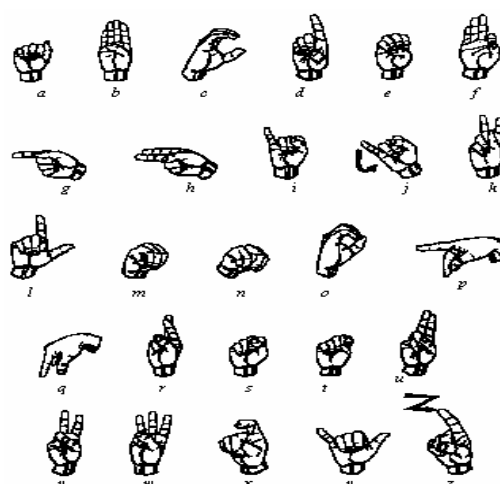


Figure 3: The ASL manual alphabet [2]

3.3. Signed English (SE)

SE is a reasonable manual parallel to English. It is an educational tool meant to be used while you speak and thereby help you communicate with deaf people

and normal hearing individuals who, for a variety of reasons, are experiencing a difficulty in development of spoken language. The idea behind Signed English and other signing systems parallel to English like Signing Exact English is that Deaf people will learn English better if they are exposed, visually through signs, to the grammatical features of English. SE uses two kinds of gestures or signs: *sign words* and *sign markers*. Each sign word stands for a separate entry in a Standard English dictionary. The sign words are signed in the same order as words appear in an English sentence. Sign words are presented in singular, nonpast form. Sign markers are added to these basic signs to show, for example, that you are talking about more than one thing or that some thing has happened in the past. The fourteen SE markers are depicted in Figure 4 [5].







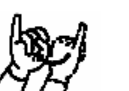








 Regular past verbs: -ed talked, wanted, learned	 Regular plural nouns: s bears, houses	 3rd person singular: -s walks, eats, s
 Irregular past verbs: (sweep RH open B, tips out, to the right) saw, heard, blew	 Irregular plural nouns: (sign the word twice) children, sheep, mice	 Possessive: -s cat's, daddy's, chair's
 verb form: -ing climbing, playing, running	 Adjective: -y Sleepy, sunny, cloudy	 Adverb: -ly Beautifully, happily, nicely
 Participle: Fallen, gone, grown	 Comparative: -er smaller, faster, longer	 Superlative: -est Smallest, fastest, longest
 Opposite of un-, im-, in-, etc. (made before the sign word, as a prefix) unhappy, impatient, inconsiderate	  Agent (person) Agent (thing) Agent (person): (sign made near the body) teacher, actor, artist Agent (thing): (sign made away from the body) washer, dryer, planter	

Figure 4: The fourteen SE markers [5]

In Signed English one can use either a sign word alone or a sign word and one sign marker to represent a given English word. When this does not represent the word in mind, the manual alphabet can be used to fingerspell the word. Most of signs in Signed English are taken from the American Sign Language. But these

signs are now used in the same order as English words and with the same meaning [5].

4. Demonstration of the ASL in SSLIS

For manipulating the ASL in our software, we are following the Signed English (SE) model as it is described in [5]. Simply, the recognized output text of the speech recognition engine will be inputted to the ASL database.

The ASL database includes the following:

- A number of 2,600 ASL prerecorded video clips where each single video clip represents the corresponding sign of a single basic word of the English vocabulary.
- The single-handed American Manual Alphabet.

As Figure 5 depicts, the constituents of the recognized text will be processed separately, i.e. single words and single alphabets if any. First of all, the word will be checked to find out whether it is a basic word or not, i.e. nonbasic words include adverbs, plurals ...etc. Subsequently there are two options; first, if the word is basic, it will be checked against the prerecorded ASL video clips database. If match occurred, the output will be that matched clip. Otherwise, the word will be fingerspelled through applying it to the American Manual Alphabet database. Secondly, if the word is nonbasic one then the basic word will be extracted out of it and checked against the ASL videos database. If it has an equivalent it will be displayed but after appending a suitable marker whether as a prefix or a suffix, this is specified by the Signed English manual. If this basic word was out of the ASL database vocabulary then simply, the original word before extraction will be entirely fingerspelled. Note that the final output depicted in Figure 5 below could be a mixture of prerecorded video clips, markers and fingerspelling.

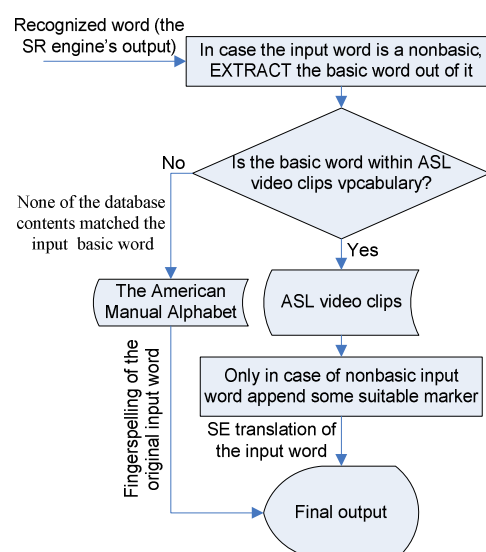


Figure 5: Demonstration of ASL translation in the SSLIS

5. Structure and flow of SSLIS

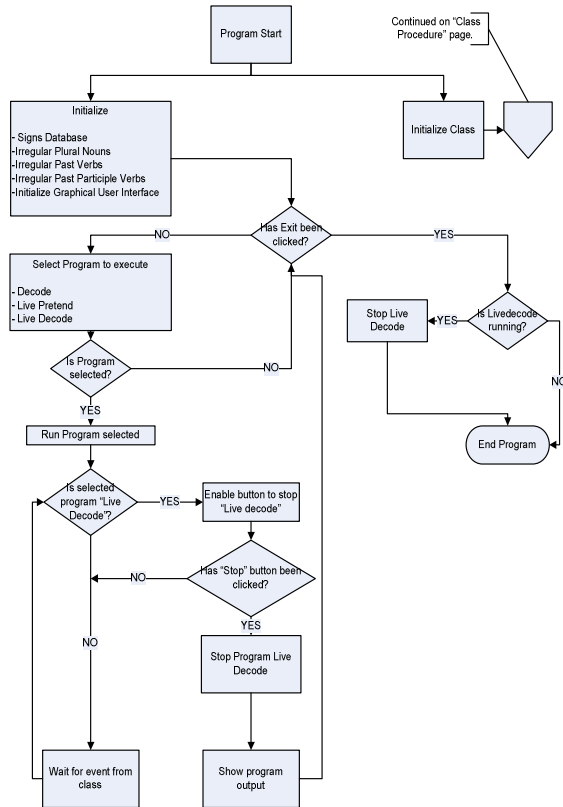


Figure 6: Flowchart of the main program

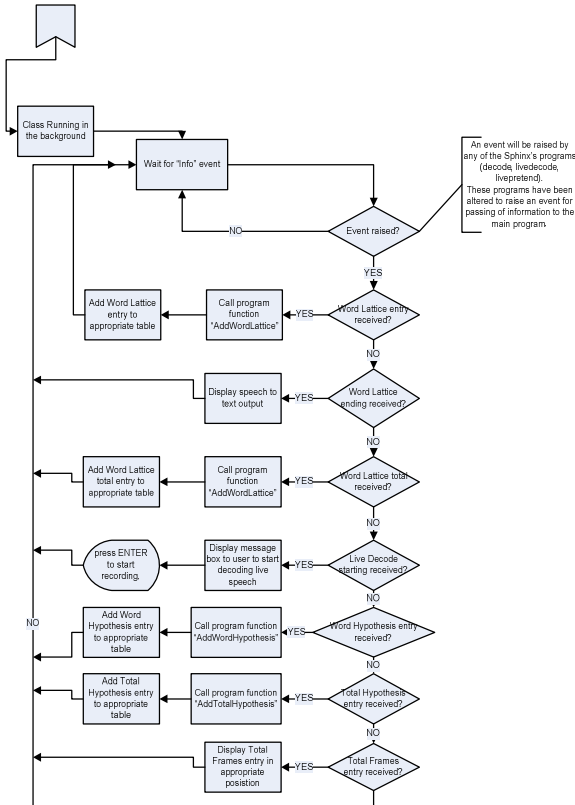


Figure 7: Flowchart of the class procedure

The overall structure of the SSLIS can be divided into two distinctive parts, the main program and the class procedure. The flowcharts for both of them are depicted in Figure 6 and Figure 7 respectively.

6. Parameters optimization & accuracy measurements

A dictionary and LM was generated then a test set follows their context has been recorded and applied to the “Live Pretend” program. There are mainly two groups of parameters pertaining to *pruning* and LM. As it is explained in the sketches hereinafter, per each session the parameter of interest was tuned then WER was calculated [6 and 18].

6.1. Tuning parameters pertaining to pruning behavior

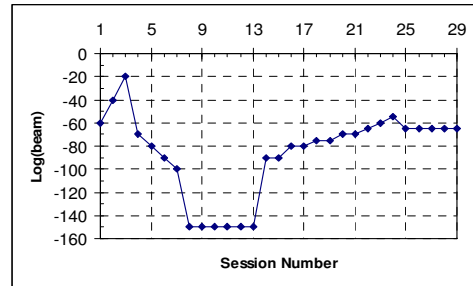


Figure 8: Log(beam) vs. Session Number

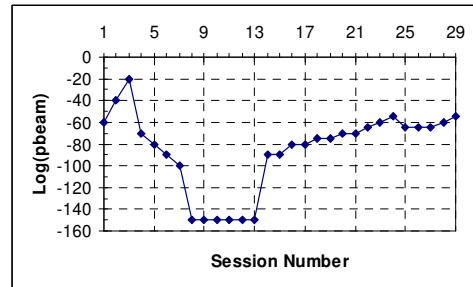


Figure 9: Log (pbeam) vs. Session Number

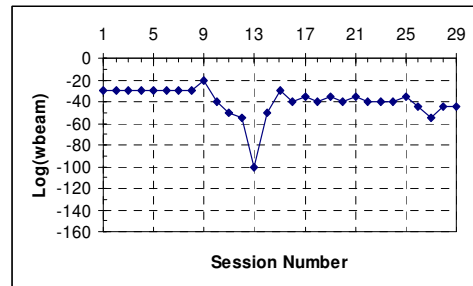


Figure 10: Log(wbeam) vs. Session Number

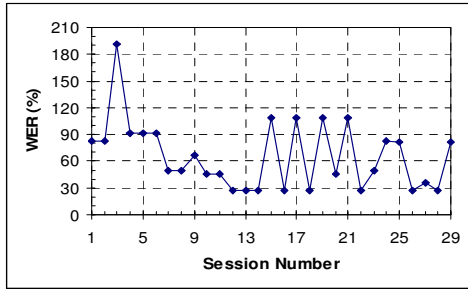


Figure 11: WER vs. Session Number

From the graphs depicted above, the least WER that was lastly obtained is 27.27% at session number 28 which corresponds to the following parameter values: -beam=1e-65, -pbeam=1e-60 and -wbeam=1e-45.

6.2. Tuning parameters pertaining to LW

To tune the language model related parameters, first we started with **-lw** keeping **-wip** to its default value.

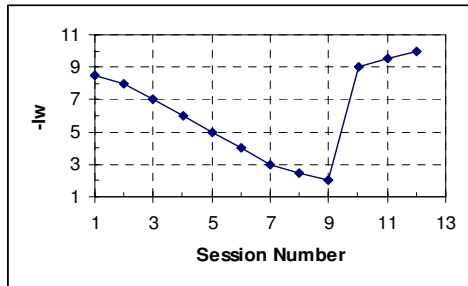


Figure 12: -lw vs. Session Number

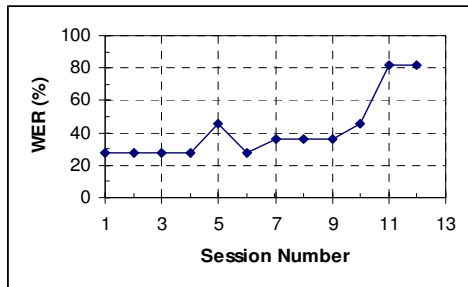


Figure 13: WER vs. Session Number

Comparing the two sketches above reveals that best recognition accuracy is in sessions 1, 2, 3 and 4. Through trying their corresponding **-lw** values on a different test set we found out that **-lw=7** works best. Using **-lw=7**, **-wip** parameter can be tuned as shown in Figure 14.

As it is shown throughout the parameter tuning procedures, the best WER obtained was 27.27%. Such degraded value was obtained even though we have replaced the language model that was originally provided along with the Sphinx package with our own language model where we have followed its context to dictate speech to the SR engine. The reason for such poor recognition accuracy is the fact that the acoustic

model we are using which was originally distributed along with the Sphinx package and which was generated from hub4 (96 and 97) broadcast news database was not meant to be used for dictation tasks.

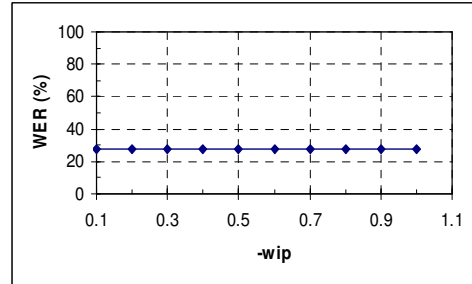


Figure 14: WER vs. -wip

7. SSLIS demonstration

Upon running the SSLIS, the program will start loading as depicted in figure below.

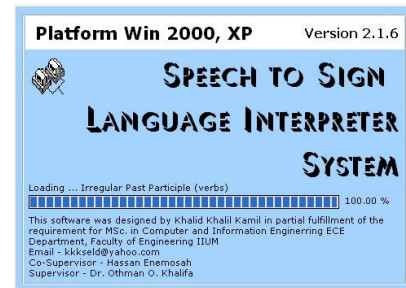


Figure 15: SSLIS is loading upon running it

Then the next window will be displayed as the default window which in turn enables the user to translate live speech into both text and sign language. Note that there are five tabs in the top of the displayed window. Hereinafter, a description for each of them will be approached.

7.1. Tab 1: Speech to Text to Sign Language

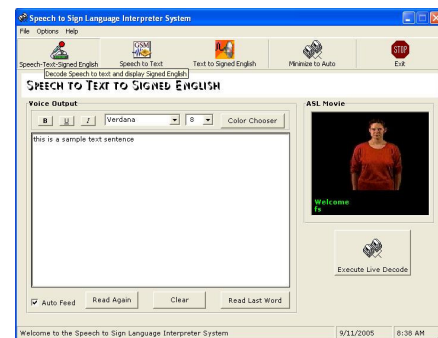


Figure 16: Default window of SSLIS, user can enter their voice in live mode to be translated to text and video sign language concurrently.

The captions shown under each of the following snapshots demonstrate the capabilities of SSLIS.

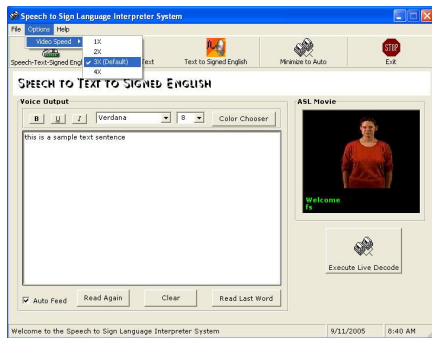


Figure 17: Selection of ASL movies' display speed

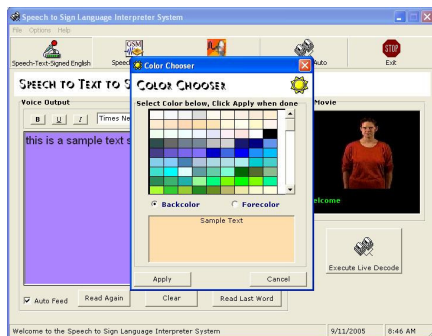


Figure 18: Controlling displayed text format and background color

7.2. Tab 2: Speech to Text

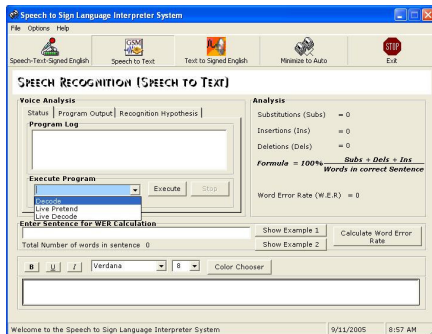


Figure 19: User can select one of the three included programs to convert speech into text. Both batch mode and live mode are supported

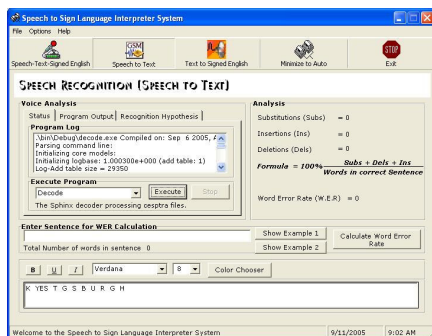


Figure 20: Demonstration of the batch mode "Decode" program

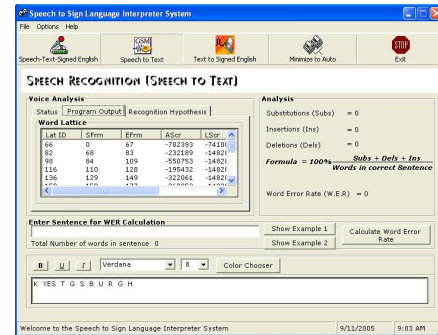


Figure 21: Program Output displays more information about "Decode"

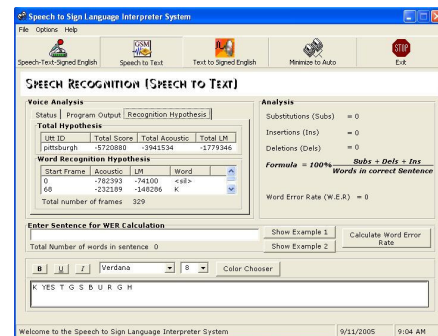


Figure 22: Recognition hypothesis for "Decode"

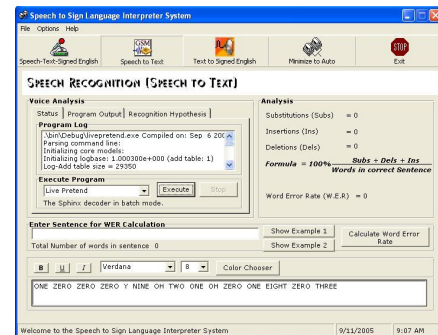


Figure 23: Demonstration of "Live Pretend" program. Live Pretend enables recognition of prerecorded speech

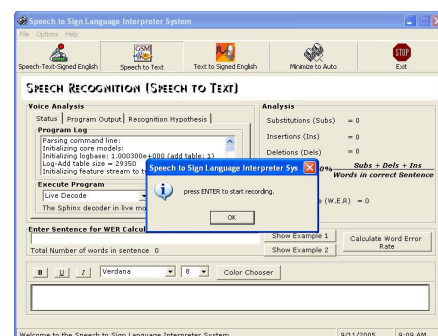


Figure 24: Opting for "live decode" (Speech Recognition in live mode)



Figure 25: This utility was designed for calculating WER. Two explanatory examples are provided

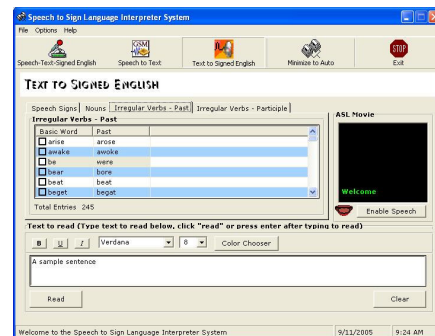


Figure 29: List of the past irregular verbs that SSLIS capable to recognize

7.3. Tab 3: Text to Signed English

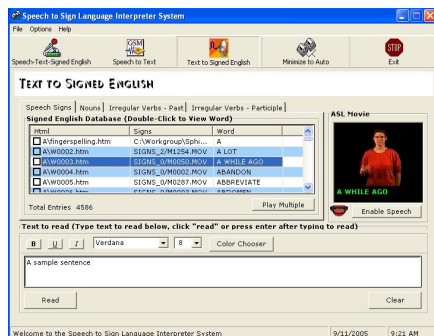


Figure 26: Demonstration of the sign language database constituents. User can type text and convert it to sign language and voice. Lips placed under the woman's picture are moving in sync with the computer's generated voice

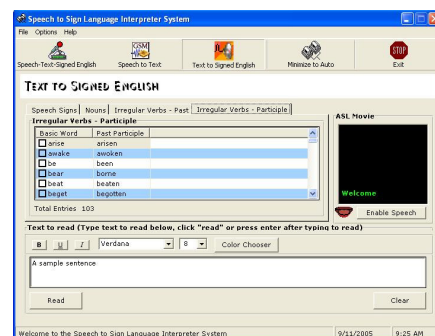


Figure 30: List of the past participle irregular verbs that SSLIS capable to recognize

7.4. Tab 4: Minimize to Auto



Figure 31: Clicking on "Minimize to Auto" places a small icon at the bottommost right hand side corner of the desktop

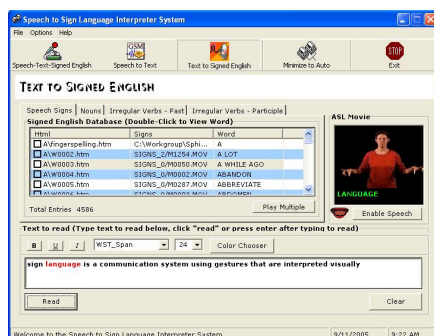


Figure 27: Demonstration of typing text and clicking on Read

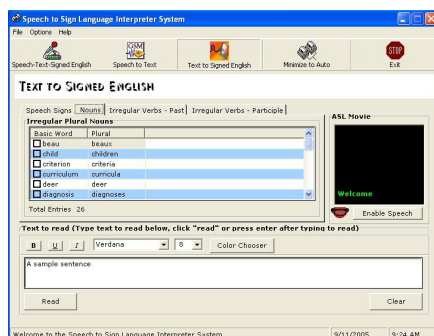


Figure 28: Irregular nouns defined in the SSLIS SE database



Figure 32: To translate text to sign language in "minimize to Auto" option, simply select, drag and drop the text of concern on the icon placed at the bottommost right hand side corner of the desktop

8. Conclusions

This paper has presented state-of-the-art GUI software to convert input uttered English speech in live mode to its equivalent text and video signed English concurrently. At the heart of it, Sphinx 3.5 was manipulated as the speech recognition engine. Rule of operation, parameters optimization and accuracy measurements and SSLIS capabilities have been described throughout the paper. The overall process performed by Sphinx 3.5 can be shortly described as follows: upon speaking in the microphone the speech signal will be captured, digitized and applied to FE signal processor wherein a certain number of MFCC features will be extracted per each input frame. The static structure already residing as a knowledge base comprises AM, dictionary and LM. Using the observations (extracted features) a search graph will be constructed dynamically and the final recognition output will be obtained through the Viterbi beam search. The process of translating the engine's output text to video signed English was further described. This has been achieved through including a certain number of ASL prerecorded movies, wherein each single movie is equivalent to a single basic word, markers and the ASL manual alphabet in the SSLIS sign language database and implementing them all together following the Signed English (SE) manual. In terms of accuracy, it was shown throughout the paper that the accuracy is quite degraded. This is due to the fact that the acoustic model used in SSLIS was generated from 96 and 97 hub 4 broadcast news database and was not generated from such similar dictation tasks. This problem can be tackled through constructing such dictation tasks database and generating an acoustic model from it which in turn entails a lot of work. In addition to SSLIS main task, other added capabilities were demonstrated as well. These are: speech to text conversion in both, batch and live mode, demonstration of the SR process, automatic calculation of WER, text to computer generated voice conversion, ability to directly sign text from web browser or any text editor through just selecting, dragging and dropping it, text display format control and ASL movie display speed control. We hope that the SSLIS will attract the researches attention to this area and help to fill the gap between deaf and nondeaf communities.

References

- [1] *American Sign Language Video Dictionary and Inflection Guide*. (2000). [CD-ROM]. New York: US. National Technical Institute for the Deaf, Rochester Institute of technology. ISBN: 0-9720942-0-2.
- [2] ASL University. *Fingerspelling: Introduction*. <http://www.lifepoint.com/asl101/fingerspelling/fingerspelling.htm>
- [3] Baker, J.K. (1975). *The DRAGON System-An Overview*. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-23(1). pp.24-29.
- [4] Becchetti, C., & Ricotti, L. R. (1999). *Speech Recognition Theory and C++ Implementation*. England: Wiley.
- [5] Bornstein, H., Saulnier, K.L. & Hamilton, L.B. (1992). *The Comprehensive Signed English Dictionary* (Sixth printing). USA: Washington DC, The Signed English series. Clerc Books, Gallaudet University Pres.
- [6] Gouvêa, E. *The CMU Sphinx Group Open Source Speech Recognition Engines*. <http://www.speech.cs.cmu.edu/sphinx/>
- [7] Harrington, T. (July, 2004). *Statistics: Deaf Population of the US*. <http://library.gallaudet.edu/dr/faq-statistics-deaf-us.html>
- [8] Huang, X., Acero, A., Hon, H-W., & Reddy, R. (2001). *Spoken Language processing, a Guide to Theory, Algorithm and System Development*. Prentice Hall PTR
- [9] Hwang, Mei-Yuh. (1993). *Subphonetic Acoustic Modeling for Speaker Independent Continuous Speech Recognition*. Ph.D. thesis, Computer Science Department, Carnegie Mellon University. Tech Report No. CMU-CS-93-230
- [10] iCommunicator™ pricing (2003). <http://www.mycommunicator.com/?action=pricing>
- [11] Jelinek, F. (Apr. 1976). *Continuous Speech Recognition by Statistical Methods*. Proceedings of the IEEE, Vol. 64, No. 4. pp. 532-556.
- [12] Lin, E. (May 2003). *A First Generation Hardware Reference Model for a Speech Recognition Engine*. Master Thesis, Computer Science Department, Carnegie Mellon University.
- [13] National Institute of Standards and Technology. <http://www.nist.gov/>
- [14] *Personal Communicator*. [CD-ROM]. version 2.4. (2001). Michigan: US. Communication Technology Laboratory, Michigan State University.
- [15] Rabiner, L. R. (Feb 1994). *Applications of Voice Processing to Telecommunications*. Proceedings of the IEEE, Vol. 82, No. 2, pp. 199-228.
- [16] Rabiner, L., & Juang, B-H. (1993). *Fundamentals of Speech Recognition*. New Jersey: Prentice Hall international.
- [17] Ravishankar, M. (May 1996) *Efficient Algorithms for Speech Recognition*. Ph.D. dissertation, Carnegie Mellon University. Tech Report. No. CMU-CS-96 143.
- [18] Ravishankar, M. K. (2004). *Sphinx-3 s3.X Decoder (X=5)*. Sphinx Speech Group. School of Computer Science, CMU. <http://cmusphinx.sourceforge.net/sphinx3/>
- [19] Rosenfeld, R. *The CMU Statistical Language Modeling (SLM) Toolkit*, http://www.speech.cs.cmu.edu/SLM_info.html
- [20] Seltzer, M. (1999). *Sphinx III Signal Processing Front End Specifications*. CMU Speech Group, www.cs.cmu.edu/~mseltzer/sphinxman/s3_fe_spec.pdf
- [21] Wilcox, S. (2005). *Sign Language*. The Microsoft Encarta Reference Library.