SIGN LANGUAGE CONVERTER

Taner Arsan and Oğuz Ülgen

Department of Computer Engineering, Kadir Has University, Istanbul, Turkey

ABSTRACT

The aim of this paper is to design a convenient system that is helpful for the people who have hearing difficulties and in general who use very simple and effective method; sign language. This system can be used for converting sign language to voice and also voice to sign language. A motion capture system is used for sign language conversion and a voice recognition system for voice conversion. It captures the signs and dictates on the screen as writing. It also captures the voice and displays the sign language meaning on the screen as motioned image or video.

KEYWORDS

Motion Capture, Motioned Image, Sign Language Converter, Voice Recognition.

1. INTRODUCTION

The aim of this paper is to improve the communication with the people who has hearing difficulties and using any sign language to express themselves. At the first sight, as an idea, how difficult could make a sign languages converter. After detailed research about sign language linguistics, it is figured out about 240 sign languages have exist for spoken languages in the world. To show how tough to working with any sign language, the general information about sign languages is given briefly.

After have an idea about sign language linguistics, Microsoft Kinect Sensor XBOX 360 is decided to use for capturing abilities and technical features to the motion capture of sign to voice conversion. Google Voice Recognition is used for the voice to sign conversion. Google Voice Recognition is available only on android based programs. Eventually, the voice recognition program CMU Sphinx is chosen. This allows us to combine both components in Java. Conversion program is also designed and written in Java.

Finally, Java based program is produced which can make voice recognition, motion capture and convert both of them to each other. So a deaf person easily speaks to in sign language in front of motion sensor, the person behind the screen can understand easily without ability to speak sign language and vice versa.

2. INFRASTRUCTURE AND IMPLEMENTATION

Infrastructure of a sign language system consists of three main branches as Sign Language, Speech Recognition and Implementation with MS Kinect XBOX 360TM. These are the main motivations of implementing such a system. The following sections are describing each term in details and giving necessary information.

DOI:10.5121/ijcses.2015.6403

2.1. Sign Language

It is easy to find a wide number of sign languages all over the world and almost every spoken language has its respective sign language, so there are about more than 200 languages available. There are several sign languages available such as American, British, German, French, Italian, and Turkish Sign Language. American Sign Language (ASL) is well-known and the best studied sign language in the world. The grammar of ASL has been applied to other sign languages especially as in British Sign Language (BSL). BSL is not closely related to ASL, so the differences between BSL and ASL are shown in Figure 1. This section is not going to go further with details of a single sign language because each sign language has its own rules. The next section will aim to give a general description of the shared or common characteristics between the different sign languages: origin, phonology, and syntax. Design a sign language translator is not an easy task.

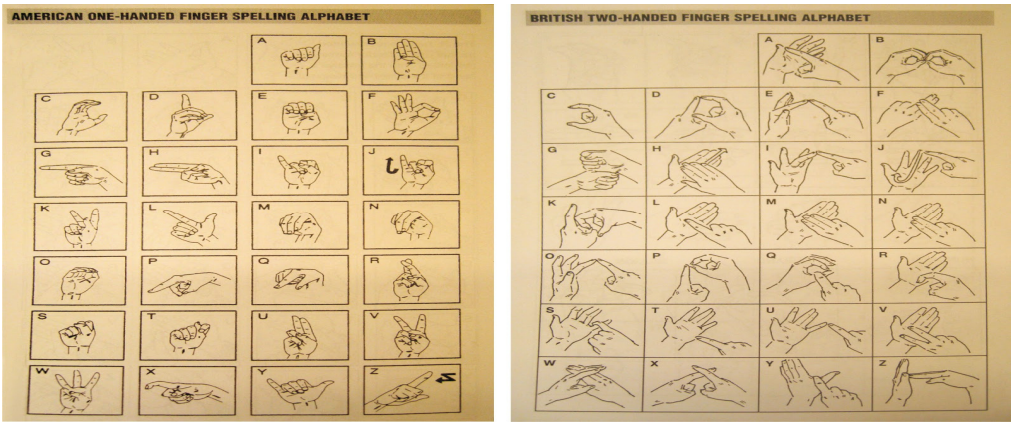

Figure 1. Differences between American Alphabet and British Alphabet.

2.1.1. Origin of Sign Language

Deaf people need sign language to communicate with each other and other deaf people. Moreover, several ethnic groups that use completely different phonologies (e.g. Plain Indians Sign Language, Plateau Sign Language) have used sign languages to communicate with other ethnic groups. The origin of the sign language is mainly related to the beginning of the history. The book of Juan Pablo Bonet called "Reduccion de las letras y Arte para ensenar a hablar los Mudos (Reduction of letters and art for teaching mute people to speak) is published in Madrid in 1620 [1]. This is accepted as the first modern treatise of phonetics, arranged a method of oral education for deaf people by using the manual signs, as shown in Figure 2, of manual alphabet to improve their communication. However, this manual alphabet was not good, but just a way to make communication possible.

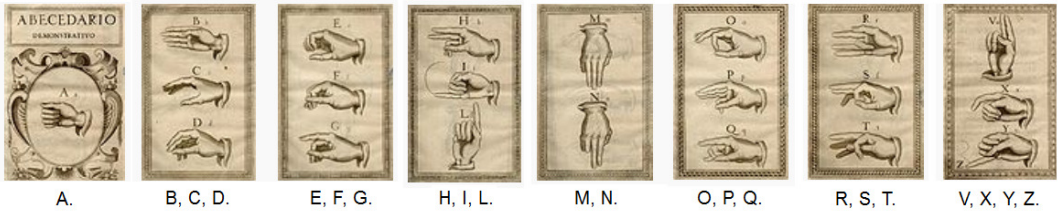

Figure 2. Manual Signs of Alphabet

The first real study of sign languages is achieved in 1960s. Dr. William C. Stokoe published the monograph Sign Language Structure [2] in 1960. Some of his deaf students from the University of Gallaudet help him to propose the signs. Then he published the first American Sign Language dictionary [3]. In this first dictionary, Dr. Stokoe organized the signs considering the position of the shape and motion. He did not consider on its English translation. This is a cornerstone and give a start for research about the Sign Language linguistics.

2.1.2. Phonology

The phonology refers to the study of physical sounds present in human speech. The phonology of sign language can be defined. Instead of sounds, the phonemes are considered as the different signs present in a row of hand signs. They are taking into account the following parameters:

1. Configuration: Hand shape when doing the sign.
2. Orientation of the hand: Where the palm is pointing to.
3. Position: Where the sign is completed.
4. Motion: Movement of the hand when doing the sign (straight, swaying, circularly) [2].
5. Contact point: Which part of the hand touch the body.
6. Plane: The sign is depending on the distance to the body.
7. Non-manual components: Information provided by the body. For example, when the body leans front, it expresses future tense.

2.1.3. Morphology

Spoken languages have inflectional morphology and also derivational morphology. The inflectional morphology refers to the modification of words. The derivational morphology is the process of forming a new word on the basis of an existing word. Sign languages have only derivational morphology because there are no injections for tense, number or person. The most important parameters regarding morphology are represented as:

1.Degree: Mouthing.
2.Reduplication: Repeating the same sign several times.
3.Compounds: Fusion of two different words.
4.Verbal Aspect: Expressing verbs in different ways. Several of these involve reduplication [2].
5.Verbal number: To express plural or singular verbs. Reduplication is also used to express it.

2.1.4. Syntax

It is primarily brought through a combination of word order and non-manual features. It is described by:

1. Word order: A full structure as [topic] [subject] [verb] [object] [subject-pronoun-tag].
2. Topic and main clauses: Background information sets.
3. Negation: Negated clauses can be mentioned by shaking the head during the entire clause.
4. Questions: The questions are mentioned by lowering the eyebrows.
5. Conjunctions: Separate sign in ASL is a short pause.

2.1.5. Concepts of Sign Languages

Briefly the basic knowledge about sign languages has been represented. The main linguistic characteristics used by the system are part of the Phonology section. The following parameters are considered:

1. Position: The position that the sign is occurred.
2. Motion: Movement of the hand when doing the sign (straight, swaying, circularly).
3. Plane: The distance with respect to the body.

The different basic signs from the ASL dictionary have been taken and the parameters assigned to these characteristics as respect to the position, motion and plane.

3. IMPLEMENTATION

In this section, the implementation steps of sign languages converter are represented. Firstly, the sensors are considered and then software and middleware specifications are mentioned.

3.1 Microsoft Kinect XBOX 360TM

Microsoft released new sensors which can establish watershed as RGB-D cameras can understand for using it with in gaming console. There are 3 sensors such as RGB, audio and depth. Both of them have serious roles like to obtain to detect movements, to allow the users to play games with their own bodies as a controller and to identify players' sounds. This Microsoft Kinect development also helped other useful applications in computer vision area such as robotics and action recognition. Microsoft Kinect Sensor and its components are shown in Figure 3.

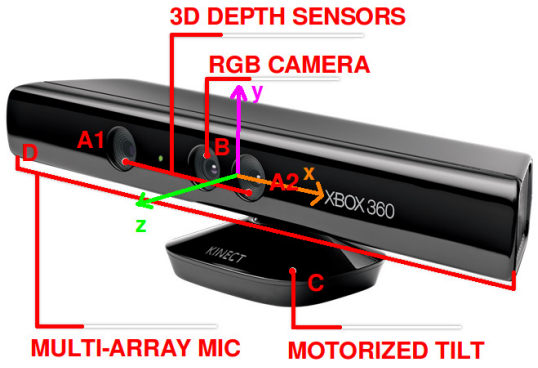

Figure 3. Microsoft Kinect Sensor

3.1.1 Components and Features

Part A is a depth sensor or called as 3D sensor too. It is a combine of infrared laser projector with a CMOS sensor to let the Kinect sensor to process 3D scenes in any environmental light conditions. BY using a grip of infrared light from projector on an area of any view, sensor receives from reflections of objects in the scene. Distance of object surfaces from the visibility point of the camera specifies by the depth map. This system called as A Time of Flight because it sets the depth map of the scene by considering the time which is the light takes to come back to the source after the jumping objects in the sensor's view. The optimal depth range of sensor is from 1.2 to 2.5 meters.

Part B is RGB camera which has 32 bits high color resolution. It can use 2 dimensional color video of the scene.

Part C is motorized tilt. It concerned with the field of view.

Part D contains an array of 4 microphones which is located along the horizontal bar. It is useful for speech recognition, ambient noise suppression and echo cancellation.

3.1.2. Drivers and Software Development Kits (SDKs)

When Microsoft used to connect the USB port to their game consoles, some of the drivers such as open source drivers, SDKs and APIs could have been useful in this system. In Table 1, there is the comparison of open source drivers which are available today.

Table 1. Drivers and SDKs of Kinect.

Name	Languages	Platforms	Features
OpenKinect/ libfreenect	C, Python, actionscript, C#, C++, Java JNI and JNA, Javascript, CommonLisp	Linux, Windows, Mac OS X	Color and Depth images. Accelerometer data. Motor and LED control. Fakenect Kinect Simulator. Record color, depth and accelerometer data in a file.
CL NUI SDK [4] and Driver	C, C++,WPF/C#	Windows	Color and Depth images. Accelerometer data. Motor and LED control.
Robot Operating System (ROS) [5]	Python, C++	UNIX	Color and Depth images. Motor and LED control.
OpenNI / NITE Middleware	C, C++,	Windows, Linux, Ubuntu	User identification Feature detection. Gesture recognition. Joint tracking, Color and Depth images. Record color and depth data in file.

Sign Language Translator is based on tracking of joints. It means adding the restriction of using Windows as Open NI / Nite Middleware can be the most suitable choice from the option in the list above.

3.1.3 Open NI Middleware

Open NI [6, 7] is a multi-language that defines APIs for writing applications as a cross platform framework. Main purpose of Open NI is to form a standard API that enables communication with the sensors in the system such as vision and audio sensors. Open NI standard API let the natural interaction application developers to 3D scenes by utilizing data types like array of the pixels in a depth map.

Open NI has 3 layers. Bottom layer contains devices that collect visual and audio data from the real world. Second layer contains Middleware components to analyze these data which is collected from the real world. The top layer contains software which implements natural application such as Sign Language Translator.

Open NI has 2 different categories in its production nodes. These are sensor related nodes and Middleware related nodes. The model relates the nodes that obtain the data directly from the device. In our project there are such functionalities such as accessing to raw color and depth images for Kinect sensors.

3.1.4 Calibration

User Generator is the mostly useful node in our project. Because it is one that allows accessing the joint tracking function and provides it to do its work and calibrate the pose. Kinect camera cannot be understood without a camera calibration step before doing anything else. By any depth and color cameras, every single pixels of the scene which will have the corresponding RGB-D values, is aligned.

3.1.5 Joint Positions

Joint tracking is the most useful point of our project. As shown in Figure 4, the implementation allows getting the position of the 15 joints and these 3D coordinates is called X, Y, Z coordinates which are given in millimeters from the device. X is horizontal plane, Y is the vertical plane and Z is to be normal to the depth sensor.

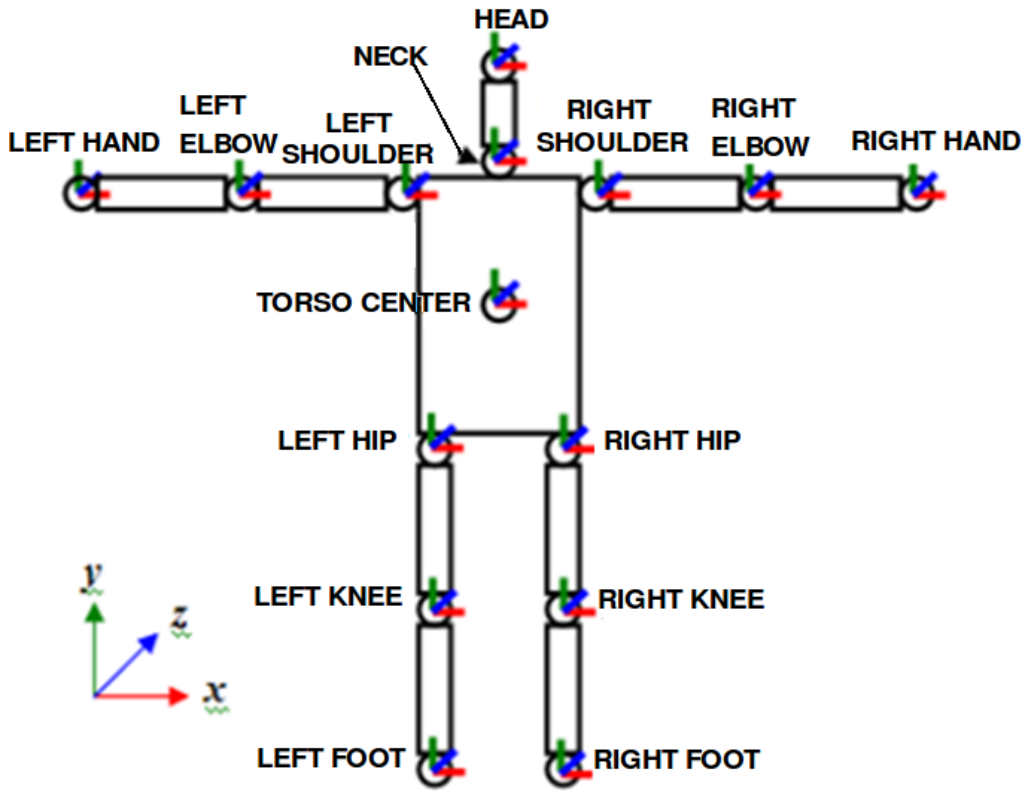

Figure 4. Joints that can be detected by Kinect

3.2 SPEECH RECOGNITION

3.2.1 Types of Automatic Speech Recognition (ASR)

ASR products have existed since the 1970s. However, early systems were very expensive hardware devices. These devices were not very reliable and user-friendly that could only recognize a few isolated words means that user stops each word and needed to be trained by users repeating each of the words several times. The 1980s and 90s had a substantial improvement in ASR algorithms and products and also the technology developed. In the late 1990s, software for

desktop dictation became available for only a few tens of dollars. Elimination with a technological perspective it is possible to characterize between two broad types of ASR: “direct voice input” (DVI) and “large vocabulary continuous speech recognition” (LVCSR). DVI devices are primarily aimed at voice command-and-control, whereas LVCSR systems are used for filling the forms or voice-based document creation. DVI systems are typically configured for small to medium sized vocabularies and might employ word or phrase spotting techniques. In both cases the underlying technology is more or less the same.

3.2.2. Speech Recognition Techniques

Speech recognition techniques are summarized in details in this section.

3.2.2.1. Template Based Approaches Matching

Unknown speech is compared against a set of pre-recorded words (templates) in order to find the best match.

3.2.2.2. Knowledge Based Approaches

An expert knowledge about variations in speech is hand coded into a system.

3.2.2.3. Statistical Based Approaches

In which variations in speech are modelled statistically, using automatic, statistical learning procedure, typically the Hidden Markov Models, or HMM.

3.2.2.4. Learning Based Approaches

To overcome the disadvantage of the HMMs machine learning methods could be introduced such as neural networks and genetic algorithm/programming.

3.2.2.5. The Artificial Intelligence Approach

The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features.

3.2.3. CMU Sphinx

CMU Sphinx, is also called as Sphinx in short, is the general name of speech recognition ystems which are developed at Carnegie Mellon University. There are three speech recognizer from Sphinx 2 to 4, and an acoustic model trainer which is SphinxTrain. In this project, Sphinx 4 is used.

4. METHODOLOGY

There are three parts of methodology:

1-Database
2-Voice Recognition Procedure
3-Motion Capture Procedure

4.1. Database

Words for Speech Recognition, .gif images and Motions together create the database.

4.1.1 Words for Speech Recognition

For the speech recognition [8-13], there are fifty words has chosen as shown in Table 2. There are 13 personal pronouns, 14 verbs, 5 adjectives, 12 nouns, 3 question words, and 3 yes, no and not for yes/no statements. With the flexibility of Sphinx it is possible to add new words to the system.

Table 2. Words for Speech Recognition

PERSONAL PRON.			**VERBS**			**ADJECTIVES**	**NOUNS**	
I	She	They	Am/Is/Are	Like	Thank	Good	Tea	Sign
You	It	Her	Feel	Go	Help	Bad	Student	Language
He	We	Him	See	Come		Sick	Teacher	Father
His	Me	My	Make	Can		Fine	Doctor	Mother
Us			Do	Eat		Great	School	Brother
			Drink	Love			Fruit	Sister
QUESTIONS			**YES AND NO**					
How			Yes					
Where			No					
Who			Not					

4.1.2 Images (.gif format)

In the Speech Recognition part, the program uses the .gif images to show the proper meaning for the recognized speech. Each words or word groups have a meaning on the Sign Language. For example, if the speaker says: “I am a doctor”, the program will show the following .gif images consecutively as shown in Figure 5 and Figure 6 sequentially for user to understand the Sign Language meaning of the sentence.

Figure 5. *“I AM”*

Figure 6. *"A DOCTOR"*

4.1.3. Motions

The program has the capability of capturing 12 motions and interprets them to the text. As an example, if user tries to say "I am good", the motion is as in Figure 7 and Figure 8.

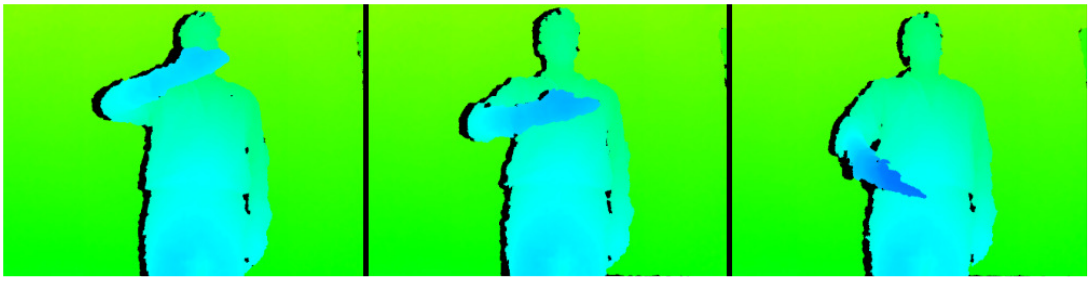

Figure 7. "I AM"

Figure 8. "GOOD"

4.1.4. Procedures

As the project has two parts with the sign language to voice, and voice to sign language translations, there are two procedures in the project. These procedures are working with respect the choice of the user. Once a procedure starts to work, it finishes its part and then give the user another choice for program. Main Menu, Voice to Sign and Sign to Voice Menu of the Program are shown in Figure 9.

Sign to Voice procedure can be said as working opposite as Voice to Sign Procedure. Sign to Voice procedure has two works to do, as to record the move, then find the proper text meaning

for the move, and within the program converts it to the Acoustic Signal. While, Voice to Sign procedure is, recording the acoustic signal and converts it to the text and then to the .gif files.

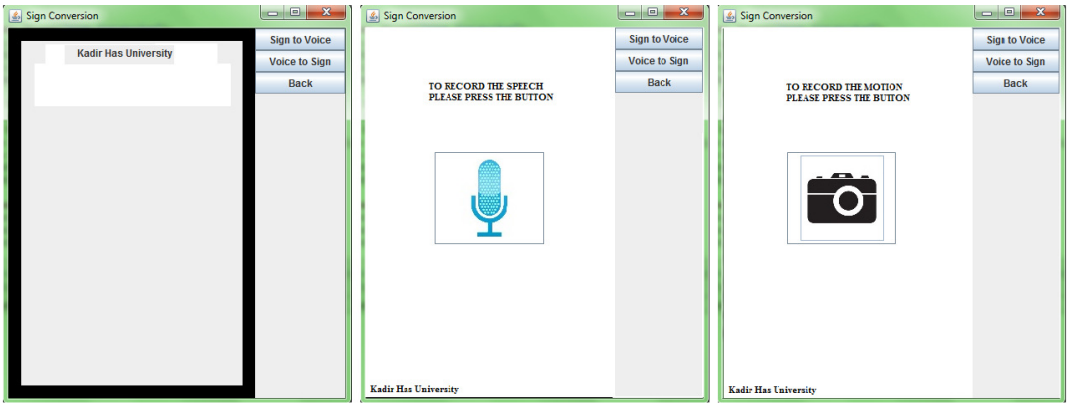

Figure 9. Main Menu, Voice to Sign and Sign to Voice Menu of the Program

4.2. Voice Recognition Procedure

Speech processing is the field which works on the speech signals and the processing of them. The signals are usually processed in a digital representation, although the signals are analog. Speech processing is interested in to gather, store, manipulate, transfer speech signals. It is faster to communicate with the voice than text, therefore with the translation of voice to the image will give healthy people to communicate with the people with the hearing disorders. Once the user press the button to record the speech, computer's microphone starts to listen, and after catching the voice with the help of CMU Sphinx, it finds the meaning as the text. Then in Java it is matched with the proper .gif image, so that the other user will understand. The diagram of the Voice Recognition Procedure is given in Figure 10.

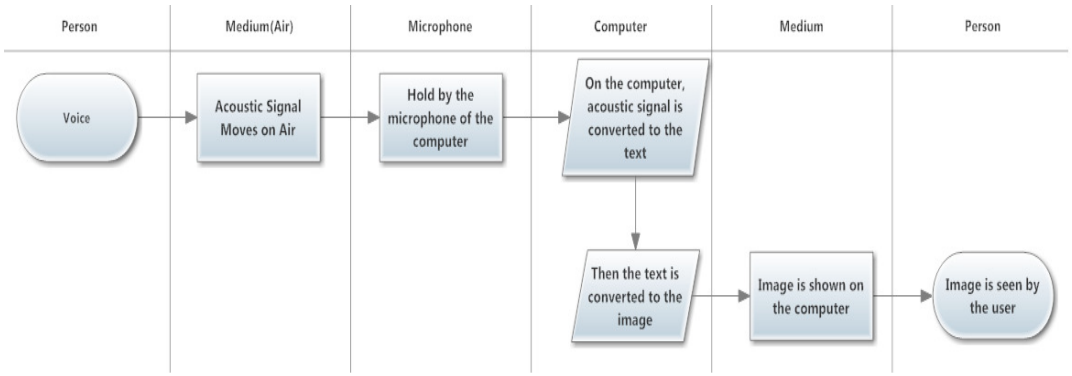

Figure 10. The diagram of the Voice Recognition Procedure

4.3. Motion Capture Procedure

In this procedure, image processing is really important. Image processing is used commonly in our life recently, and it seems that future will bring much more than that. The diagram of the motion capture procedure is given in Figure 11.

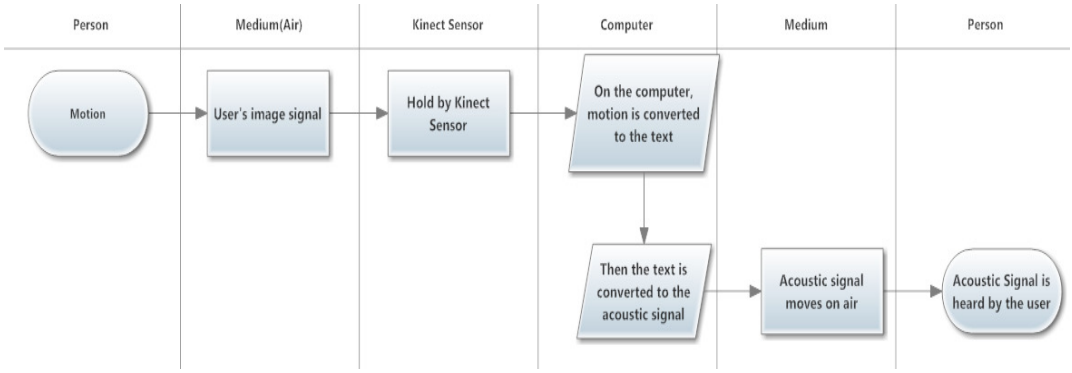

Figure 11. Motion Capture Diagram

One of the developed image processing sensors is Microsoft's Kinect Sensor [4-7, 14]. As it can be called the second part of the project, the motion capturing is the part where Kinect Sensor is used. Once the user press the button to record the motion, Kinect sensor starts to capture motions, but to start to record the sign motions it starts a specific motion, which is shown in the Figure 12.

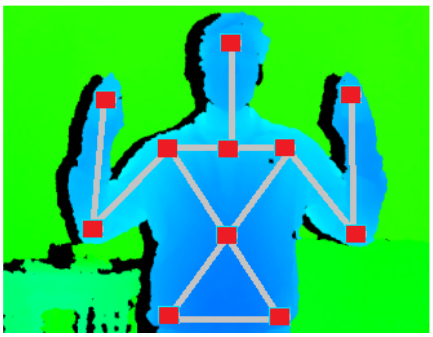

Figure 12. "Starting Motion"

After the "Starting Motion", Kinect captures the motions and it converts them to the text. On the computer, this text is converted to the voice and then the other user can hear the meaning of the sign. Flow chart of the sign language converter program is given in Figure 13.

5. CONCLUSIONS

This paper is about a system can support the communication between deaf and ordinary people. The aim of the study is to provide a complete dialog without knowing sign language. The program has two parts. Firstly, the voice recognition part uses speech processing methods. It takes the acoustic voice signal and converts it to a digital signal in computer and then show to the user the .gif images as outcome. Secondly, the motion recognition part uses image processing methods. It uses Microsoft Kinect sensor and then give to the user the outcome as voice.

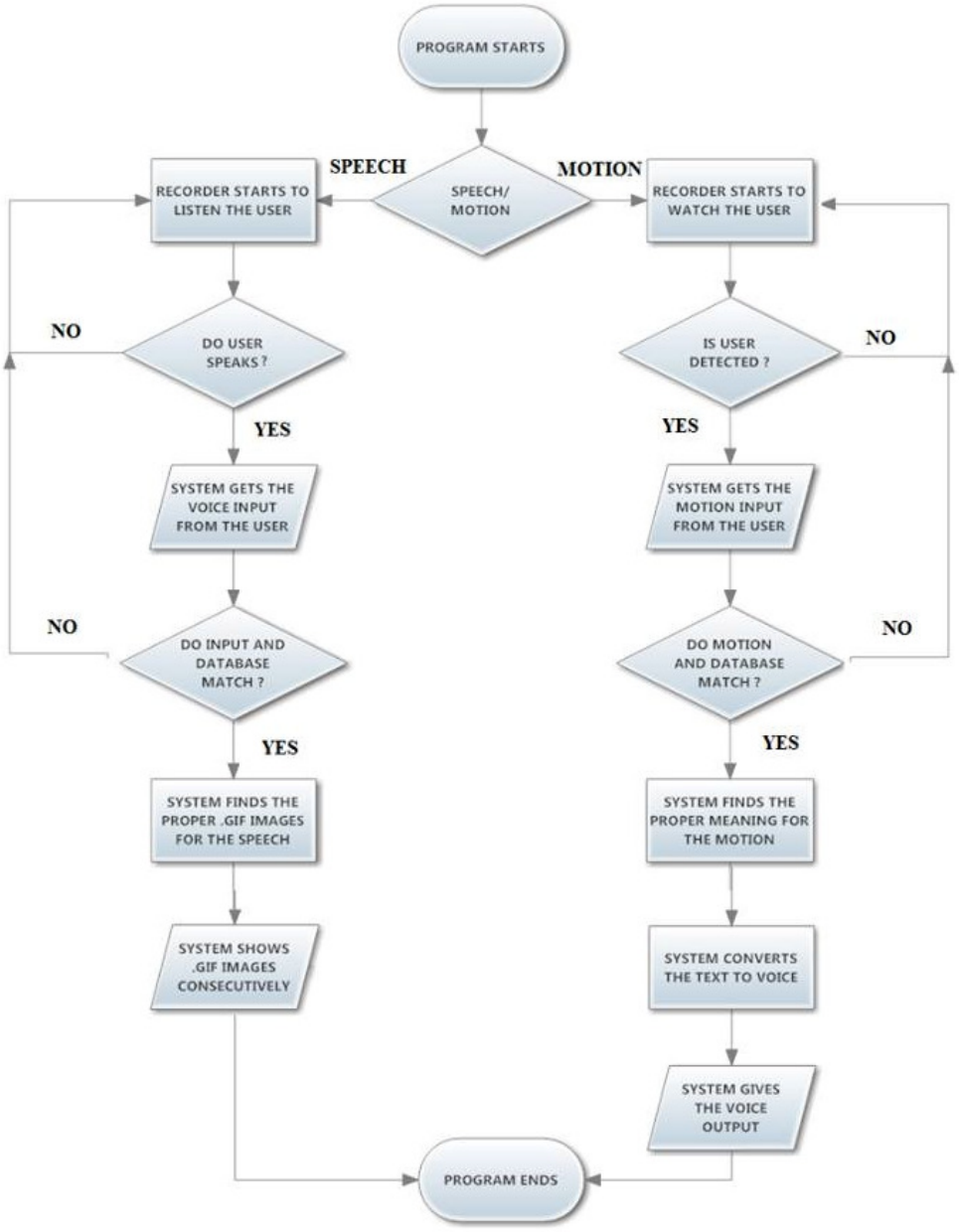

Figure 13. Main Program Flow Chart

The project gives us the many advantages of usage area of sign language. After this system, it is an opportunity to use this type of system in any places such as schools, doctor offices, colleges, universities, airports, social services agencies, community service agencies and courts, briefly almost everywhere.

One of the most important demonstrations of the ability for communication to help sign language users communicate with each other occurred. Sign languages can be used everywhere when it is needed and it would reach various local areas. The future works are about developing mobile application of such system that enables everyone be able to speak with deaf people.

REFERENCES

[1] J.P. Bonet. "Reducci_on de las letras y arte para ense~nar a hablar a los mudos", Coleccion Cl_asicos Pepe. C.E.P.E., 1992.

[2] William C. Stokoe. Sign Language Structure [microform] / William C. Stokoe. Distributed by ERIC Clearinghouse, [Washington, D.C.], 1978.

[3] William C. Stokoe, Dorothy C Casterline, and Carl G Croneberg. "A Dictionary of American Sign Language on Linguistic Principles" Linstok Press, [Silver Spring, Md.], New Edition, 1976.

[4] Code Laboratories. CL NUI Platform. http://codelaboratories.com/ kb/nui

[5] The Robot Operating System (ROS), http://www.ros.org/wiki/ kinect.

[6] Open Kinect Project, http://openkinect.org/wiki/Main_Page.

[7] Open NI API Reference. http://openni.org/Documentation/Reference/ index.html.

[8] Bridle, J., Deng, L., Picone, J., Richards, H., Ma, J., Kamm, T., Schuster, M., Pike, S., Reagan, R., "An Investigation of Segmental Hidden Dynamic Models of Speech co-articulation for Automatic Speech Recognition.", Final Report for the 1998 Workshop on Language Engineering, Center for Language and Speech Processing at Johns Hopkins University, pp. 161, 1998.

[9] Ma, J., Deng, L., "Target-directed Mixture Linear Dynamic Models for Spontaneous Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol. 12, No. 1, January 2004.

[10] Ma, J., Deng, L., "A Mixed-level Switching Dynamic System for Continuous Speech Recognition", Elsevier Computer Speech and Language 18 (2004) 4965, 2004.

[11] Mori R.D, Lam L., Gilloux M., "Learning & Plan Refinement in a Knowledge Based System for Automatic Speech Recognition", IEEE Tra. on Pattern Analysis Machine Int., 9(2):289-305, 1987.

[12] Rabiner, L., R., and Wilpon, J. G., "Considerations in Applying Clustering Techniques to Speaker-independent Word Recognition", Journal of Acoustic Society of America, 66 (3):663-673, 1979.

[13] Tolba, H., and O'Shaughnessy, D., "Speech Recognition by Intelligent Machines", IEEE Canadian Review (38), 2001.

[14] Kathryn LaBelle, "Kinect Rehabilitation Project", http://netscale.cse.nd.edu/twiki/bin/ view/Edu/KinectRehabilitation, June 2009.