Sign Language Interpreter

Sanjay Kumar Suman, Himanshu Shekhar, Chandra Bhushan Mahto, D. Gururaj, L. Bhagyalakshmi, and P. Santosh Kumar Patra

Abstract This article addresses a design of an apposite system which provides a supportive hand for hearing and speaking challenged person to expediently communicate with normal people. Normally, a sign language is adopted by them for their communication which needs an interpreter to convert into user's understandable language. The proposed system is used for converting the sign language into voice and text and vice versa. The idea of the proposed project is to come up with a device that captures the gestures and converts it to voice output as well as in text output and also to capture the voice by speech recognition module and convert it to corresponding sign language by displaying on a screen with the help of various elements like microphone, camera, sign language database and display unit. For the general-purpose indoor implementation, a facial expression recognition system can also be additionally included.

Keywords Sign language · Interpreter · Speech recognition · Communicate · Gesture recognition · Facial expression recognition

1 Introduction

Earlier, it was very difficult for the deaf/dumb to communicate with a normal person because of the lack of a proper sign language and ease of understanding. But after the advent of sign language, the deaf/dumb now, are able to communicate not only with similarly placed, but also with normal people. At times, it is difficult to communicate

S. K. Suman · P. S. K. Patra
St. Martin's Engineering College, Secunderabad, Telangana, India

H. Shekhar
Hindustan Institute of Technology and Science, Chennai, TN, India

C. B. Mahto
Department Electrical Engineering, MIT Muzaffarpur, Muzaffarpur, Bihar, India

D. Gururaj · L. Bhagyalakshmi (✉)
Department of ECE, Rajalakshmi Engineering College, Chennai, TN, India
e-mail: Prof.Dr.L.Bhagyalakshmi@gmail.com

A. Kumar et al. (eds.), *Advances in Cognitive Science and Communications*,
Cognitive Science and Technology, https://doi.org/10.1007/978-981-19-8086-2_96

with normal people since, and it is not necessary that all the people whom they come across is acquainted with the sign language to understand what the deaf/dumb has to say. This is called as unintentional misunderstanding [1, 2]. In such cases, they have to hire an interpreter who can interpret their sign language and convert into speech for normal person to understand and vice versa. Still, there are some fallacies occurring in sign interpretation, predominantly in the field of business and transactions. To overcome this, we have an electronic interpretation device to stand by the deaf/dumb, so that they can communicate with ease. This would go a long way in establishing effective and reliable communication with the deaf/dumb and normal person without having to approach an interpreter.

Any movement of the hand or change in face that expresses a thought, emotion, feeling or reaction can be defined as a gesture such as: eyebrows movement and raising soldier are normal behavior used in our daily life. Sign language is a systematic and defined communication method in which each word or letter is assigned a specific gesture [3]. Here, a motion capture system is used for sign language conversion and a speech recognition system is used for speech conversion [4].

This idea can be executed using two different implementations, namely indoor and outdoor. Indoor module consists of facial expression recognition system. The only major issue will be collecting the list of all the words with their sign language. Creating the database is the most difficult task. Since there are many ways to interpret sign language, different possibilities can be used to design the system.

Facial expression recognition or emotion detection systems include three steps: face detection, feature extraction and facial expression recognition [5, 6]. The face detection algorithm for this system is based on the work of Viola and Jones. They proposed a face detection framework that can process images very fast and achieve high detection rates [7]. The database used is the most comprehensively tested Cohn Kanade database for a comparative study of the facial expression and emotion database [8]. Also, a local binary pattern is used for analyzing attitude emotions and textures [9, 10]. In addition, Microsoft's Kinect sensor XBOX 360 includes motion capture technology that can convert signatures to voice, and the camera decides to use it for scene capture.

2 Earlier and Current Issues

Earlier project of text-to-sign language conversion was limited the output to the PC base module and no portability [8]. A fact to be known is that sign language interpreters have cognitive abilities, perceptual skills and other characteristics that make them unique from others [11]. Further, there are about 12.3 million people having moderate to complete hearing loss in India, but we have only around 25,000 interpreters. Only a part of these deaf people (about 4.5 million) would not be able to succeed in a school for hearing people, whereas they can obtain education in special schools for deaf. They would then be introduced to sign language which might become part of life of the deaf and dumb community. A statistical analysis

reveals that around 478 govt. running school and 372 private schools receive fund from govt. agency for development of deaf children in India using oral approach rather than sign language [12].

Worse situation found in rural part of developing country is where the CODA does not receive an education due the distance. They have to struggle to attend the school and also to the need to work at home. Many deaf and mute people are talented at many fields in spite of their disability to speak or hear. The various fields include business, biology, psychology, arts, science, mathematics and computer, etc. Presently, there is a need for qualified interpreters in medical fields, businesses and offices for making the language translation easier.

Children of deaf adults, called as CODAs, often serve as interpreters in most parts of the world. However, in many developing countries, CODAs are either not qualified or reluctant to work as interpreters due to their inevitability. Another issue is dependency of the deaf parents on their parent or on relative, for nurturing care and education of their children. In this case, the children do not get proper education and even fundamental requirements. Also, many CODAs do not admit that they possess deaf parents due to the fear of discrimination and uncertain problems that might arise upon revelation [12].

3 Proposed Work

This paper describes two implementations namely indoor and outdoor. The indoor module contains additional features such as facial expression recognition and lip reading (optional) for more accurate results along with camera, microphone, speakers and display unit. By introducing this device for sign language interpretation, we can overcome the discrimination and difficulties that a deaf or mute person faces in the society while communicating with a hearing person. The deaf and mute community calls a normal person as "hearing person." The device would consist of a database of sign language visuals like animation [13] and the corresponding word display. The intermediate module would be the converter depending upon the process that is to be carried out.

The database of the sign language can be bifurcated to be used by different kinds of people, like for kids, smart class and schools for learning purposes with the help of smart screens and projectors. They can get educated along with normal children in normal schools and avoid going to special schools where people only of similar kind are enrolled. A universal language for example English can be used at an initial level. Depending upon the usage and purpose of application, it can be developed in other languages for better understanding. Regular updates by means of apps can be launched for updated and modified version of the sign language conversion tool.

4 System Model

This section presents the system model which comprises: voice input to sign input, sign input to voice input and facial expression recognition.

- **Module I**

  The first module, as shown in Fig. 1, consists of a speech recognition device, the database containing the video content to various actions in sign language, the matching device that converts the corresponding audio to respective video (sign language) and the output display unit. The sign language animations and symbols are loaded to the database of the device. When the mic captures the audio of a normal person, it gets converted to text internally and that will be matched with the corresponding interpreted sign output in the database. This converted output, i.e., the matched output, will be displayed on the display screen. An adaptive noise canceling microphone system is used here to capture the voice.

Speech Recognition

We can either make use the design of Kaldi which is a free open-source toolkit for the purpose of speech recognition [14, 15]. Kaldi gives us a speech recognition system which is based on finite-state transducers (using the freely available OpenFst), together along with the detailed documentation and scripts for creating a complete recognition system. The only issue is that it gives only a considerable level of product satisfaction. Another idea is to use google speech recognition system which is comparatively quick and also easy which makes use of the IOT concept. There are various approaches for speech recognition as follows:

- *Template*: An unknown speech is compared with a set of pre-recorded words and alphabets (templates) to find the best match.
- *Knowledge*: A robust knowledge about variations in speech is hand coded into a system so that recognition is facilitated.
- *Statistical*: This is the method by which variations in speech are modeled statistically, using automatic, statistical learning procedure, typically the hidden Markov models, or HMM. This method is usually tedious and not up to the level of satisfaction.

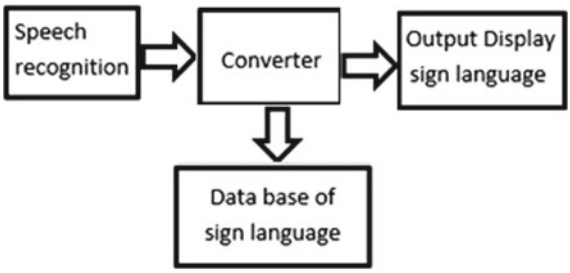

Fig. 1 Speech-to-sign conversion module

- *Learning*: Machine learning methods could be introduced such as neural networks and genetic algorithm/programming in order to overcome the disadvantage of the HMMs.
- *Artificial Intelligence*: The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person is applying his/her intelligence in visualizing, analyzing and finally making a decision on the *measured* acoustic features or data.
- *CMU Sphinx*: CMU Sphinx, is also called as Sphinx in short, is the general name of speech recognition systems which are developed at Carnegie Mellon University. There are three speech recognizers from Sphinx 2 to 4 and an acoustic model trainer which is Sphinx Train. In this project, Sphinx 4 can be used. It purely depends on the quality of output required, that we make the choice of the appropriate speech recognition tool. There are various sub-modules of the Sphinx.

Database

Words for speech recognition, images and motions (videos) of sign language all together create the database for the product.

Display Unit

The display unit is an ordinary screen like led display or a normal smart device like smart phone or iPad. The output this module will be displayed as animated video or still images demonstrated in Fig. 2 and Fig. 3, respectively.

Fig. 2 Animated sign language output

Fig. 3 Animated sign converted output

- **Module II**

The second module, shown in Fig. 4, the same device consists of depth camera, the database of speech audios matched to a corresponding sign or gesture of hands or body and finally an output speaker. The IR depth camera and its associated gesture recognition camera are depicted in Fig. 5 and Fig. 6, respectively. The sign gestures of the dumb person is captured by the depth camera and matched with the available database of voice audios and when the match to the action is found, the respective audio is played in the speakers.

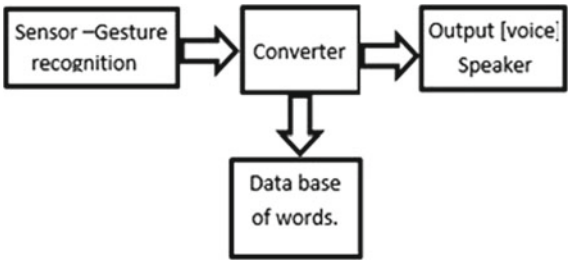

Fig. 4 Sign-to-speech conversion module

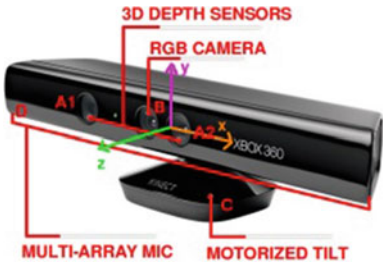

Fig. 5 IR depth camera

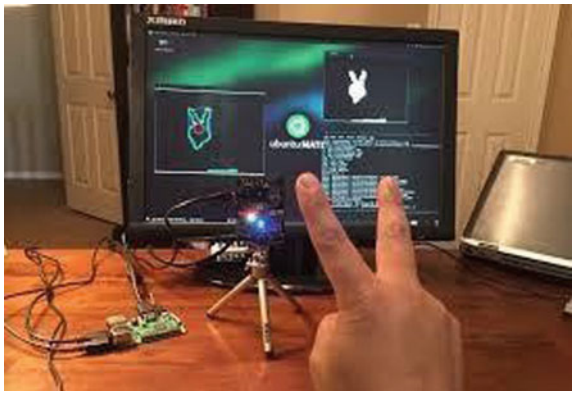

Fig. 6 Gesture recognition using camera

Camera

The Microsoft Kinect Sensor XBOX 360 was chosen to capture the technical and motion capture capabilities of converting signals to voice. Google Speech Recognition is used to convert speech into signatures. For android-based programs, only Google voice recognition is available.

Finally, you can combine the two components in Java by choosing the speech recognition program CMU Sphinx. The converter can also be configured and written in Java. Finally, a Java-based program can be written that are capable of speech recognition and motion capture. One can use this program to convert the two to each other. As a result, hearing-impaired people can easily talk to ordinary people in sign language in front of a suitable camera, and people behind the screen can easily understand it even if they cannot sign language. The reverse is also true.

Microsoft Kinect XBOX 360TM was released by Microsoft with various sensors within. There are three sensors such as depth, audio and RGB as shown in Fig. 5. The various sensors are engaged to detect movements and recognize bodily gesture and sound. This is also widely used in robotics and action recognition for creative designing in games [3].

Features

Figure 6 illustrates the gesture recognition using camera. The features of IR depth camera are divided into four parts:

- Part A is also called a depth sensor or 3D sensor. The combination of an infrared laser projector and CMOS sensor allows the Kinect sensor to process 3D scenes in ambient lighting conditions. Using infrared light from the projector in the area of consideration, the sensor receives reflections from various objects in the scene. The depth map correctly specifies the distance between the object's surface and the point visible to the camera. This is called time-of-flight because it sets up the depth map of the scene, taking into account the amount of time it takes light

reflected off an object from the sensor view to return to the light source. The optimal depth range for the sensor is 1.2–2.5 m.
- Part B is a 32-bit and high-resolution RGB camera. It has the ability to create a two-dimensional color video of a scene.
- Part C is called motor tilt, which is primarily related to the field of view.
- Part D includes a microphone. It is on a horizontal bar. This is the 4 microphone array. It is useful for environmental noise suppression, correct speech recognition and echo cancelation.

Voice Output

Miniaturized speakers are used so that it becomes handy and portable for the person and gives a level of comfort. These speakers are manufactured in smaller sizes than a normal loud speaker yet providing louder output tone.

- **Module III**

The module 3, shown in Fig. 7 comprises of facial expression recognition system. The interaction between human and computer can be made effective if the computer can recognize the emotional state of a human being. Information regarding a person's emotion is expressed in terms of facial expression. Hence, recognizing the facial expressions will let us know something about the emotional state of the person. However, it is hard to categorize facial expressions from normal images. In this problem neural network may be suitable because it can be used to improve its performance. Moreover, it is not necessary to know much about the features of the facial expressions to build the system.

An image containing a human face with an expression in the size of 96×72 pixels can be used as the input to the system. On an average there are 6 outputs representing each facial expression (with further advancements more expressions can be loaded depending upon usage). Each number represents a facial expression (smile, angry,

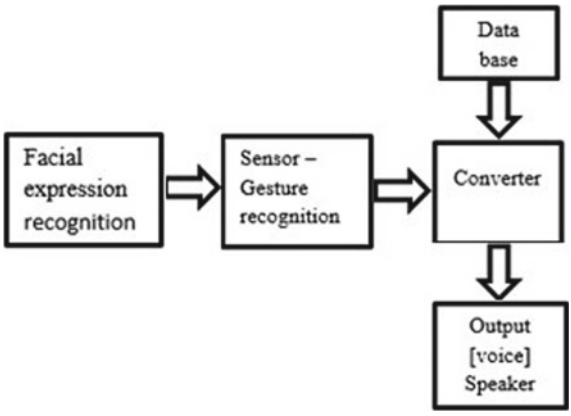

Fig. 7 Facial expression recognition module

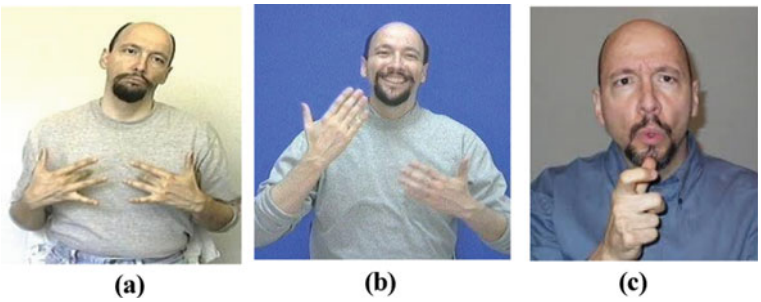

Fig. 8 The facial expression for: **a** sadness, **b** happiness and **c** reaction "who?"

fear, disgust, sadness, surprise). If that facial expression is present in the memory, then the number is 1 (one), and if not, it is 0 (zero) [6].

Facial Expression Recognition

A few examples of various facial expressions are explained as follows:

- *Sadness*: In Fig. 8a, this expression is to slightly lower the corners of the lips while raising the inner eyebrows. Darwin described this expression with a look that did not want to cry. The lower lip is lowered because the upper lip control is larger than the lower lip control. When a person cries and screams, close their eyes to protect them from the build-up of blood pressure in their face. So when I have the urge to cry and want to stop, I try to raise my eyebrows without closing my eyes.
- *Happiness*: This expression usually involves a smile: both corner of the mouth rising, shown in Fig. 8b.
- *A question such as-Who?*: This expression, Fig. 8c, is obtained as a result of inverted v-shaped eyebrows and curvy mouth with hand in the shape as shown in the picture

5 Implementation

(1) *Implementation 1*

The first method is to make a handy device like a wearable chain where the locket consists of the miniaturized camera, microphone, speaker and a memory device for the database. The output unit will be a display unit in the form screen either a smart phone or an iPad or a unique special purpose display screen. This display unit is also miniaturized one for portability and easy handling. The person wears the chain and activates the device, the device starts sensing the gesture made by the deaf or dumb person's hand (in front of the camera), and the gesture recognition system comes to work and translates the sign to corresponding voice output. Thus, this establishes a conversation between the deaf or mute person and a hearing person so that the sign is translated to voice.

Now, if the deaf or mute person has to understand the normal person, he/she activated the device for speech recognition.

(2) *Implementation 2*

The second method is for indoor purpose where the device along with facial expression recognition system is installed at a particular location inside the room (e.g., classroom). This method is exclusively for the classroom purpose.

6 Conclusion

This article presented the interpreter which can be used in a closed room or outside. Also, this device can be used for smart classes, library and public utility services like airport, bus stations, railway stations, hospitals, Internet hubs, hotels, restaurants, malls and offices. This could be more beneficial for communication at schools for the deaf/dumb so that they can feel themselves on par with normal person. Deaf and mute children are prone to be looked down upon by normal children, thereby creating an inferiority complex among themselves. This can be avoided by using the interpretational device which will remove the barrier of emotional differences between them and a normal child. Even though they lack the power to hear and speak, they are multiskilled personalities and excel in their own interests. Their potential and capability can be discovered to achieve greater heights in life.

References

1. Read MK (1977) Linguistic theory and the problem of Mutism: the contributions of Juan Pablo Bonet and Lorenzo Hervas Y Panduro. Historiographia linguistica 4(3):303–318. https://doi.org/10.1075/hl.4.3.03rea
2. Harvey MA (2003) Shielding yourself from the perils of empathy: the case of sign language interpreters. J Deaf Stud Deaf Educ 8(2):207–213. https://doi.org/10.1093/deafed/eng004
3. Arsan T, Ülgen O (2015) Sign language converter. Int J Comput Sci Eng Surv 6(4):39–51. https://doi.org/10.5121/ijcses.2015.6403
4. Kalsh A, Garewal NS (2013) Sign language recognition system. Int J Comput Eng Res 03(6). http://www.ijceronline.com/papers/Vol3_issue6/part%201/D0361015021.pdf
5. Piatkowska E (2010) Facial expression recognition system. Master thesis: technical reports. https://via.library.depaul.edu/cgi/viewcontent.cgi?article=1017&context=tr
6. Do C-T, Pastor D, Goalic A (2010) On the recognition of cochlear implant-like spectrally reduced speech with MFCC and HMM-based ASR. IEEE Trans Audio Speech Lang Process 18(5):1065–1068. https://doi.org/10.1109/TASL.2009.2032945
7. Viola P, Jones MJ (2004) Robust real-time object detection. Int J Comput Vision 57(2):137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb
8. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture with local binary patterns. IEEE Trans Pattern Anal Mach Intell 7(7):971–987. https://doi.org/10.1109/TPAMI.2002.1017623
9. Wallhoff F, Schuller B, Hawellek M, Rigoll G (2006) Efficient recognition of authentic dynamic facial expressions on the feedtum database. In: IEEE international conference on multimedia and expo. https://doi.org/10.1109/ICME.2006.262433

10. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proceedings of the 4th IEEE international conference on automatic face and gesture recognition, Grenoble, France, pp 46–53. https://doi.org/10.1109/AFGR.2000.840611
11. Seal BC (2015) Psychological testing of sign language interpreters. J Deaf Stud Deaf Educ 9(1):39–52. https://doi.org/10.1093/deafed/enh010
12. Sugandhi PK, Kaur S (2021) Indian sign language generation system. IEEE Mag Comput 54(3):37–46. https://doi.org/10.1109/MC.2020.2992237
13. Halawani SM (2008) Arabic sign language translation system on mobile devices. Int J Comput Sci Netw Secur 8(1):251–256 (King Abdulaziz University, Jeddah, Saudi Arabia). https://www.kau.edu.sa/Files/830/Researches/56041_26352.pdf
14. Povey D et al (2011) The Kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding, Hilton Waikoloa Village, Big Island, Hawaii, US. https://www.danielpovey.com/files/2011_asru_kaldi.pdf
15. Oliveira T, Escudeiro N, Escudeiro P, Rocha E, Barbosa FM (2019) The virtual sign channel for the communication between deaf and hearing users. IEEE revista iberoamericana de tecnologias del aprendizaje 14(4):188–195. https://doi.org/10.1109/RITA.2019.2952270