Chapter 31
Malay Cued Speech Recognition Using Image Analysis: A Review

Muhammad Ghazali Twahir, Zulkhairi Mohd Yusof, and Izanoordina Ahmad

Abstract Automatic real-time translation of gestured languages for hearing-impaired would be a major advancement on disabled integration path. Cued speech (CS) is a specific visual hand gesture that complements oral languages lip-reading. Cued speech in Bahasa Malaysia (CSBM) is an adaptation of cued speech for use in the Malay language. The cued speech recognition system is capable of detecting all necessary parameters of CS (handshape, hand position, and hand movement) and translate it to text equivalent. The aim is to help the deaf learn and practice the basic of cued speech consonant and vowel using hand gesture. This paper looks into the existing researches involved in this area and also the sensors and methods they used. Due to the limited number of researches for cued speech, related researches such as sign language (SL) translator systems and hand gesture recognition are also reviewed. This paper gives a general overview on the implementation of cued speech recognition system that automatically recognize a succession of cued speech hand gestures in real time. A Malay cued speech recognition system using image analysis is proposed.

Keywords Cued speech · Machine learning · Transliterate system · Image analysis · Hand gesture

M. G. Twahir (✉)
R4R Research Cluster, Communication Technology Section, Universiti Kuala Lumpur British Malaysian Institute, Batu 8, Jalan Sungai Pusu, 53100 Gombak, Selangor, Malaysia
e-mail: ghazali@unikl.edu.my

Z. M. Yusof · I. Ahmad
R4R Research Cluster, Electronics Technology Section, Universiti Kuala Lumpur British Malaysian Institute, Batu 8, Jalan Sungai Pusu, 53100 Gombak, Selangor, Malaysia
e-mail: zulkhairi@unikl.edu.my

I. Ahmad
e-mail: izanoordina@unikl.edu.my

A. Ismail et al. (eds.), *Advanced Materials and Engineering Technologies*,
Advanced Structured Materials 162,
https://doi.org/10.1007/978-3-030-92964-0_31

31.1 Introduction

As someone speaks, a hearing-impaired can try to guess the oral message by lip-reading. This is a difficult task for different phonemes which correspond to identical mouth shapes. Therefore, Dr. Cornett developed the cued speech in order to improve the lip-reading efficiency [1]. The manual gestures to lip shapes are proposed to ensure that each sound has an original visual aspect. Thus, the “hand and lip-reading” becomes as meaningful as the oral message.

A significant difference between these two communication systems is that SL is a complete language, while CS is not a language at all. For example, cued speech in Bahasa Malaysia (CSBM) is a visual representation of the spoken Malay language itself. Due to that, CS may have an advantage over SL in an environment where a translation is being made [2]. The deaf will also learn to actually speak with their mouth while learning cued speech with their hand. However, it will take a lot of time and practice for them to be able to speak proficiently and fluently. In order to improve the communication of deaf cure, there is a need of an automatic system that recognizes the CS and translates it to text equivalent.

A new approach called CSBM is based on a syllabic decomposition: The message is formatted into a list of “consonant–vowel syllable” (a CV list). Each CV is coded with a specific gesture, which is combined to its lip shape, so that the whole looks unique and understandable. A gesture contains two pieces of information which is the area handshape (for the consonant coding as shown in Fig. 31.1) and a location around the face (for the vowel as shown in Fig. 31.2). Hand coding brings the same quantity of information than the lips movement. It is as difficult to lip-read without gestures as to understand the hand coding without lips movements. This symmetry explains why a single gesture codes of numerous phonemes, which correlate to different lip shapes. Thus, there are only eight handshapes and four positions for a combination of 32 CV-gestures.

The contribution of this research is to review methods of automatically recognize a succession of cued speech hand gestures in real time. In the future, a complete hearing-impaired translator could be feasible by coupling such a device with an automatic lip-reading module and others various automates.

31.2 Literature Review on Cued Speech

A cued speech recognition system is a system that automatically recognizes the hand shape, movement, and position and then display the text equivalent in the computer screen. In order for the system to do the recognition, a camera that can capture necessary parameters of CS such as handshape, hand position, and hand movement is needed. All these parameters will be processed by certain method in order for the gesture to be accurately determined. Some of the methods have been developed for

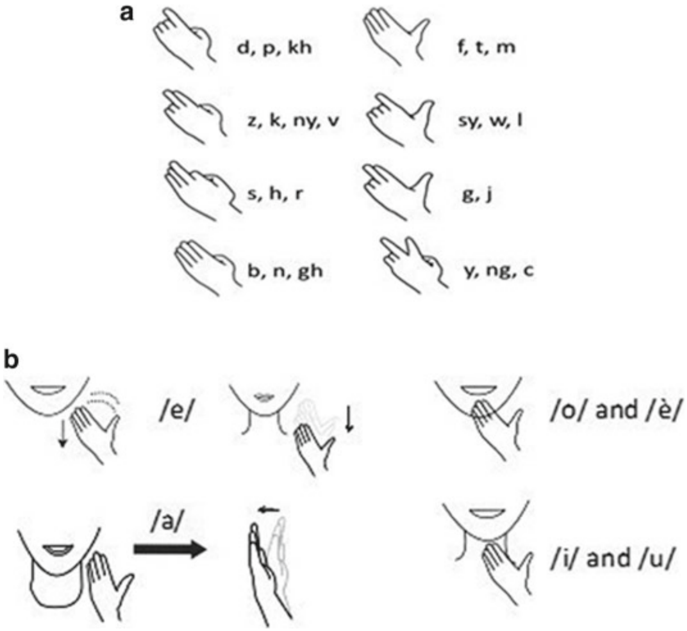

Fig. 31.1 **a** Eight handshapes for the 24 Malay consonants. **b** Four position with respect to the face for the six Malay vowels

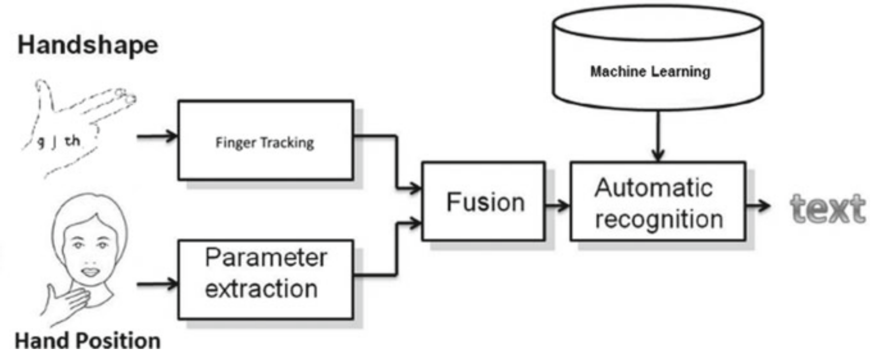

Fig. 31.2 System diagram of Malay cued speech recognition

recognizing the hand gesture such as using computer vision, depth cameras, glove-based system, hidden Markov models, adaptive boosting (AdaBoost), random forest regression (RFR) Progress, finger tracking algorithm, and glove-based system. These methods will be discussed in the following sub-section.

31.2.1 Existing Cued Speech Transliterate System

There are a limited number of researches for an automated cued speech recognition system. Some of the earliest explorations into CS recognition were conducted using computer vision techniques. Aboutabit et al. [3] focused on the identification of vowels by merging CS hand positions and lips information. Hand position was conducted using the Gaussian classifier which took the 2D hand positions as input. The vowel recognition used the merged features of the lips and hand position, and obtained 77.6% identification correctness.

Given the success that hidden Markov models (HMM) have had in the field of automatic speech recognition, Heracleous et al. [4] used the context-independent HMM-GMMs to decode a set of isolated phonemes extracted from CS sentences, i.e., the temporal boundaries of each phoneme to be recognized in the video was given at the test stage. In fact, the audio-based temporal segmentation was used for the temporal alignments of the lips, hand position, and shape. The corpus was derived from a video recording of the CS speaker with blue colors on lips and hand pronouncing and coding a set of 262 French sentences. The experiments concerning the vowel, consonant, and phoneme recognitions were presented and obtain an accuracy of 78.9%.

In the previous work on CS recognition, the video images were recorded with artifices applied to the CS speaker before the recording (blue sticks on the lips, blue marks on the hand and forehead) in order to mark the pertinent information and make their further extraction easier. However, recent researches are struggled to provide a robust and real-time solution that can adequately track handshapes against varying backgrounds and occlusions.

31.2.2 Depth Sensor Camera

Over the past decade, there has been significant exploration into using depth cameras such as Microsoft Kinect for tracking the hands gesture. The device itself features an RGB camera, a depth sensor, and a multi-array microphone and is capable of tracking the users' body movement. Since Kinect is able to track the user's full body, it seems natural to build a framework for sign language recognition. Much of the work has focused on generalized hand tracking with a priority on real-time processing and arbitrary camera angles. Ahmed et al. [5] develop Deaftalk, a sign language interpreter using Microsoft's Kinect depth camera that provides 84% accuracy detection.

The gestures recognition technologies in Kinect visual gesture builder implemented are the AdaBoost and RFRProgress. AdaBoost is a trigger which gives us a true Boolean value, while the person is performing a particular gesture; it uses the adaptive boosting machine learning algorithm. RFRProgress on the other hand produces continuous results giving us an analog data of progress. The user is performing the gesture, thus, enabling the system to detect how much of the gesture is completed and how much is the hit rate at the particular frame of the gesture. This approach uses the random forest regression machine learning algorithm. Although it facilitates body and hands tracking and creates the depth image directly, it does not support hand shape recognition. Since sign language and CS generally features different hand shapes, similar signs cannot be distinguished.

31.2.3 Finger Tracking

A number of approaches have been explored to get around the problem of hand shapes detection, but no clear consensus as to which holds the most promise has formed. Some of the approaches, such as finger counting using convex hull algorithm were shown to work in a particular case. Gurav et al. [6] developed a method using background subtraction and HSV segmentation together to create a mask. After the hand is segmented, the number of fingers raised could be detected. The largest contour in the image which is assumed to be the hand is then found. Then, the convex hull and convexity defects which are most probably the space between fingers is classified. All this algorithm such as convex hull, background subtraction, and HSV segmentation are provided in the open computer vision (OpenCV) library. OpenCV is used in HCI, robotics, biometrics, image processing, and other areas where visualization is important and includes an implementation of Haar classifier detection and training [7]. Its finger counting algorithm achieved 92% accuracy with convex hull technique. This is a manual way of finding the number of fingers and a necessary step to identify each different finger.

31.2.4 Glove-Based

Glove-based systems have achieved the most impressive SL results in terms of vocabulary size, with over 90% accuracy being obtained in continuous sign detection across more than 5000 Chinese signs [8]. However, such systems are both expensive and require the user to wear unnatural devices.

31.3 Requirement of Cued Speech Recognition

31.3.1 System Overview

The ultimate goal of this research is to develop a Malay cued speech hand gesture recognition system. The system diagram in Fig. 31.2 shows the basic flow of the proposed system. There are a number of subsystems within the proposed system that will be explained in more detail in the section B.

In this proposed system, the user will perform desired hand gesture; thus, the Kinect camera will capture this hand shape and hand position. The handshape image is processed by the finger-tracking algorithm to determine how many fingers are engaged. Then, each finger could be classified into its respective consonant. Similarly, hand position and movement video frame are captured in parameter extraction module where skin, depth, contour points, and movement are detected. The images or video streams are next sent to the machine learning functions where they are matched and compared with stored images templates to identify the correct consonant. Lastly, the consonant is fused with the vowel to produce text of Malay syllable in the computer screen.

31.3.2 Proposed Method

Hand shape recognition by using finger gesture tracking is proposed for this research. The hand shape will be captured by using the Kinect camera which has sensors of both RGB and depth data. The hand data will be processed by convex hull methods. This method will count the number of engaged fingers, and then, each finger name is classified using fingertip tracking methods. This will be made possible by implementing the OpenCV library in the system. In result, the algorithm will be able to translate eight different hand shapes into specific consonant.

Microsoft Kinect provides Visual Gesture Builder software to detect hand position and movement using machine learning methods. VGB has built in detection technology such as discrete gesture and continuous gesture. Hand position and movement recognition can be achieved by using the AdaboostTrigger and RFRProgress algorithm. In result, the system will be able to translate four different hand positions and movements into specific vowel.

The cued speech recognition software is build using C# in MS Visual Studio. This software will be used to integrate the finger tracking algorithm and machine learning method that could translate the consonant and vowel. This method will produce a syllable which resulting a Malay words.

31.3.3 Dataset, Test, and Evaluation

In order to measure the accuracy of the system, the average hit rates of particular gesture from test runs by different users will be compared. In addition, the data validation will be counted based on false positive and false negative against the correct result. A confusion matrix will summarize the result of the testing algorithm for further inspection.

31.4 Conclusion

After the survey on the approaches used in various CS and SL recognition systems, the methodologies and algorithms involved in this research could be justified. Most of the times, a combination of different methods and algorithms has to be used to achieve a moderate to acceptable rate of recognition. This will allow the system to offer complete phonetic representation of cued speech hand gesture recognition. In the future, a complete hearing-impaired translator could be feasible by coupling such a device with an automatic lip-reading module and others various automates. From a technical point of view, there is a vast scope for the future research and implementation in this very field. The ultimate gain of the proposed study is enormous.

References

1. Cornett RO (1994) Adapting cued speech to additional languages: procedures followed in assigning phonemes to cue groups in the development of cued speech in its original form. Cued Speech J 19–29
2. Nicholls GH, Ling D (1982) Cued speech and the reception of spoken language. J Speech Hear Res 25:262–269
3. Heracleous P, Aboutabit N, Beautemps D (2009) Lip shape and hand position fusion for automatic vowel recognition in cued speech for French. IEEE Signal Process Lett 16:339–342
4. Heracleous P, Hagita N, Beautemps D (2010) Gestures and lip shape integration for cued speech recognition. In: 2010 20th International conference on pattern recognition, pp 2238–2241
5. Ahmed M, Idrees M, Abideen ZUl, Mumtaz R, Khalique S (2016) Deaf talk using 3D animated sign language: a sign language interpreter using Microsoft's kinect v2. In: 2016 SAI Computing Conference (SAI), pp 330–335
6. Gurav RM, Kadbe PK (2015) Real time finger tracking and contour detection for gesture recognition using OpenCV. In: 2015 International Conference on Industrial Instrumentation and Control (ICIC), pp 974–977
7. Perimal M, Basah SN, Safar MJA, Yazid H (2018) Hand-gesture recognition-algorithm based on finger counting. J Telecommun Electron Comput Eng 10:19–24
8. Farooq U, Asmat A, Rahim MSBM, Khan NS, Abid A (2019) A comparison of hardware based approaches for sign language gesture recognition systems. In: 2019 3rd International Conference on Innovative Computing (ICIC). https://doi.org/10.1109/ICIC48496.2019.8966714