V2S: Voice to Sign Language Translation System for Malaysian Deaf People

Conference Paper · November 2009

DOI: 10.1007/978-3-642-05036-7_82 · Source: DBLP

CITATIONS
4

READS
2,451

3 authors, including:

Oi-Mean Foong
Universiti Teknologi PETRONAS
47 PUBLICATIONS **140** CITATIONS

SEE PROFILE

Tang Jung Low
Universiti Teknologi PETRONAS
158 PUBLICATIONS **734** CITATIONS

SEE PROFILE

V2S: Voice to Sign Language Translation System for Malaysian Deaf People

Oi Mean Foong, Tang Jung Low, and Wai Wan La

Computer & Information Sciences Department,
Universiti Teknologi PETRONAS,
Bandar Sri Iskandar, 31750 Tronoh, Malaysia
{foongoimean,lowtanjung}@petronas.com.my, lawai85@yahoo.com

Abstract. The process of learning and understand the sign language may be cumbersome to some, and therefore, this paper proposes a solution to this problem by providing a voice (English Language) to sign language translation system using Speech and Image processing technique. Speech processing which includes Speech Recognition is the study of recognizing the words being spoken, regardless of whom the speaker is. This project uses template-based recognition as the main approach in which the V2S system first needs to be trained with speech pattern based on some generic spectral parameter set. These spectral parameter set will then be stored as template in a database. The system will perform the recognition process through matching the parameter set of the input speech with the stored templates to finally display the sign language in video format. Empirical results show that the system has 80.3% recognition rate.

Keywords: image processing, sign language, speech recognition, spectral parameter.

1 Introduction

There are at least 70 million people around the globe who suffer from speech and hearing disabilities, either at birth or by accident [1]. It is somehow difficult for us to interact with them because of the unfamiliar communication means used. Sign Language (SL) is a common form of communication which is widely used by the speech and hearing impaired. Thus, probably the only way of easier communication/ interaction with them is by learning their language - the sign language [2].

We may have friends or family members who have hearing or speech disabilities. Such disabilities may be from birth, or by accident. Surely it is difficult for us to communicate with them if we do not know their language – the Sign Language. It is also difficult for them as well to communicate with us since they have such disability. One may be interested to learn up this language; however it may be costly to attend tuition classes to learn this language. Furthermore, tuition classes exhibits time constraint, where one does not have the flexibility in time on whether or not to attend the tutorial. He/she may prefer to have a self tutorial, however there is no such inexpensive software that can self-taught them. These may contribute to the negligence of the

H. Badioze Zaman et al. (Eds.): IVIC 2009, LNCS 5857, pp. 868–876, 2009.

public to learn the Sign Language to better communicate with those with hearing or speech impairment.

There are campaigns of speeches and talks given to the public. However, these talks usually are not able to reach those with hearing disabilities. So far, only the news on RTM 1 uses the Sign Language extensively to present the daily news to them. The cause of the lack of programs using this technique may be the reason for the extra cost incurred in hiring the translator to translate the speech. Furthermore, there are lack of trained personnel in Malaysia who are able to translate these speeches to Sign Language. These had caused those with hearing disabilities to know less on the ongoing news around them. The objectives of this research are three folds:

1. To ease the communication between normal people and the hearing/ speech disabled.
2. To eliminate the need of attending costly Sign Language classes – it can be done at home.
3. To be able to reach more audience (hearing impaired) during speeches and campaigns.

1.1 Research Motivation

According to The Star Online on the 20th December 2006, "Radio Television Malaysia (RTM 1) will be incorporating more Sign Language in their news segments and dramas for the benefits of the hearing-impaired" and on 22nd July 2007, "there is an acute shortage of Sign Language interpreters because at present there are only 10 qualified interpreters cater to 24,000 registered deaf people nationwide, according to the Malaysian Federation of Deaf".

This simply says that media agencies such as RTM is offering opportunities for more Sign Language interpreters to join its organization. But according to The Star report, the opportunity is not taken up due to the shortage of people with such qualification in this country.

The development of this V2S system shall be a solution to reduce the cost of employing special skilled employees for media agencies such as RTM. They may not need to hire workers as Sign Language interpreters to interpret their news segments or talk shows. Hiring extra interpreters is, for certain, will increase the cost of salary payout. Not only in reducing cost incurred in hiring SL interpreters, the media agencies are in fact directly providing community services to the less fortunate audiences.

V2S system can be taken as an alternative means to SL interpreter thus to replace the old fashion way (a real person doing interpretation) of translating Sign Language. The interpretation process is made available by taking the advantages of modern ICT technology via easily affordable gadgets such as computers, hand phones or PDA as a mediator (translator). In some ways this system may solves the problem of Sign Language interpreters shortage.

2 Related Works

Voice recognition can be generally classified into speaker recognition and speech recognition categories. Speaker recognition is a way of recognizing people from their voices. Such systems extract features from speeches, modeled them and use them to

recognize the person from his/her voice. There is a difference between speaker recognition (recognizing who is speaking) and speech recognition (recognizing what is being said). Voice recognition is a synonym for speaker, and thus not speech recognition. Speaker recognition has a history dating back some four decades, where the output of several analog filters was averaged over time for matching. Speaker recognition uses the acoustic features of speech that was found to be different between individuals. These acoustic patterns reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioral patterns (e.g., voice pitch, speaking style). This incorporation of learned patterns into the voice templates (the latter called "voiceprints") has earned speaker recognition its classification as a "behavioral biometric" [3].

The fundamental task of speech recognition is the deriving of a sequence of words from a stream of acoustic information. A more general task is automatic speech understanding, which includes the extraction of meaning (for instance, a query to a database) or producing actions in response to speech. For many applications, interaction between system components devoted to semantics, dialog generation, etc., and the speech recognition subsystem can be critical [4].

Feature extraction is a critical element in speech recognition since it is the first step of recognition process and generate the parameters on which the recognition is based. It is well known that Mel Frequency Cepstral Coefficients are the most widely used features parameters. One of the step of MFCC is Mel-scaled filter bank processing. This step may result in some loss of information from the original signal, but it is widely accepted that such step is helpful in extraction information component from the signals [5], [6].

A vector quantizer is a system for mapping a sequence of continuous or discrete vectors into a digital sequence suitable for communication over or storage in a digital channel. The goal of such a system is data compression to reduce the bit rate so as to minimize communication channel capacity or digital storage memory requirements while maintaining the necessary fidelity of the data. The mapping for each vector may or may not have memory depending on past actions of the coder, just as in well established scalar techniques such as PCM. Even though information theory implies that one can always obtain better performance by coding vectors instead of scalars, scalar quantizers have remained by far the common data compression systems because of their simplicity and good performance when the communication rate is sufficiently large. In addition, relatively few design techniques have existed for vector quantizers [7]. Even though there are other technique for pattern matching but Vector quantization is considered as one of the best approach for its flexibility in training as well as recognizing.

Sign Language is used primarily by deaf people throughout the world. Sign Language differs from spoken languages in that it is visual rather than auditory, and is composed of precise hand shapes and movement. This language has evolved in a completely different medium, using the hands and face rather than the vocal and is perceived by the eyes rather than the ears.

Sign Language is not a universal language shared by deaf people of the world because there are many sign languages that have evolved independently of each other. Just as spoken languages differ in grammatical structure, rules and historical relationships, sign languages also differ along these parameters.

An important property of human sign language is that the form of words is generally arbitrary, and there are no indigenous sign languages that are simply a transformation of a spoken language to the hands. Sign language is also equipped with the

same expressive power that is inherent in spoken languages and it can express complicated, intricate concepts with the same degree of explicitness and eloquence as spoken language. Sign Language portrays the image, identity and culture of the country that the Deaf Community belongs to. In Malaysia, we have the Malaysian Sign Language (Bahasa Isyarat Malaysia—BIM).

BIM has many dialects, differing from state to state. American Sign Language (ASL) has had a strong influence on BIM, but the two are different enough to be considered separate languages. Other sign languages in use in Malaysia are Penang Sign Language (PSL), Selangor Sign Language (SSL or KLSL), and Kod Tangan Bahasa Malaysia (KTBM), and Chinese Sign Languages [8].

3 Proposed System

The proposed V2S system architecture is shown in Fig.1. As illustrated in the diagram, the main components are the sound recording component (with its supporting sound/voice training algorithm), digital signal processing component (with its supporting MFCC – Mel Frequency Cepstral Coefficients algorithm counterparts), and the vector quantization component (supported by its matching sub-component).

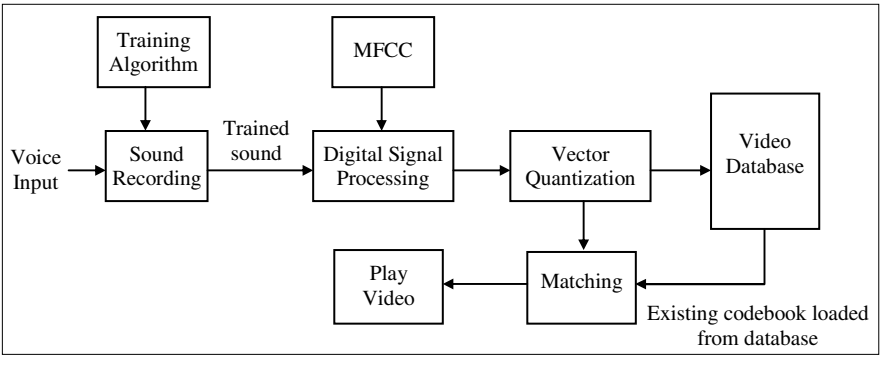

Fig. 1. The proposed V2S system architecture

3.1 Main System Components Description

A. Sound Recording – Sound recording process is responsible to capture and record the sound using microphone. Output of this process is the recorded sound which can be in *.wav* or *.midi* format. The quality of the recorded sound is highly dependent on the sound recording software used. However the quality of the recorded sound can be enhanced by applying proper noise filtering process. The sound training algorithm allows the user to "train" the system to capture the same sound/voice an appropriate number of times so as to produce a good quality recorded signal. The quality of the recorded sound/voice is an important factor in determining the accuracy of the system voice-to-signal translation.

B. Digital Signal Processing – The "trained" sound from the sound recording is then fed into the Digital Signal Processing(DSP) part. The DSP is the most important and a

difficult process implemented in this S2V system. The main task of DSP is to convert the recorded sound from its time domain to the frequency domain. This is a necessary process for extracting the features (formant) of the recorded sound so that the system can recognize words spoken to it. MFCC (Mel Frequency Cepstral Coefficients) algorithm is used in formant extraction process.

C. Vector Quantization – Vector quantization is used to perform speech recognition. In fact vector quantization is one of the most effective matching techniques that is popularly used for speech recognition. The basic concept of vector quantization is to compress any vector of a speech/voice feature into one scalar vector. By compressing the feature vector a lot of space for feature storage can be saved and helps to increase the matching process efficiency since we just have to compare a new feature with one value instead of many. In vector quantization we need to train the system first. Then the trained sound will be stored in a codebook in the database. Each trained sound will have its own codebook. During the recognition stage, the new input signal will be used to compare with all the stored codebooks and the codebook which has the closest distance will be chosen as the recognized word.

In general the input (sound/voice) is compared with the existing codebooks in the database for video retrieval. The value of each codebook and the voice input are represented using matrices. The respective average value of each codebook and the voice input would be computed. Then, each codebook will be compared with the input voice (the trained recorded signal) value. The confirmation of which video to be played is based on the closest input voice value (distance) to the codebooks stored in the database.

Fig.2(a) and (b) show samples of sound/voice for single word matching process. That is, the matching of calculated codebook (from the input voice) with the stored codebook. The samples of these sound/voice matching were processed using one of the utilities available in the MatLab® application.

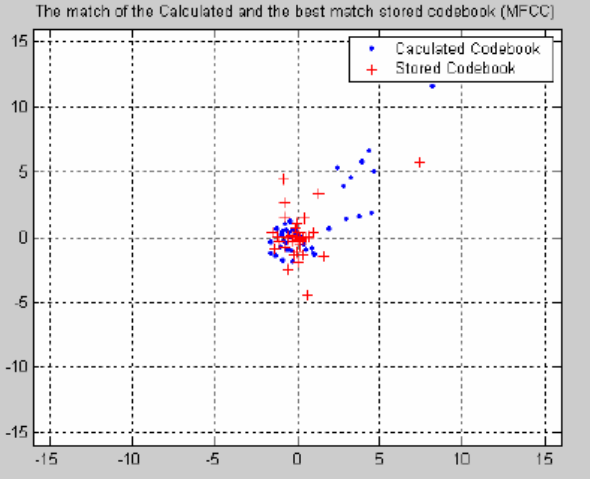

Fig. 2(a). Matching of the “me” voice with the stored codebook

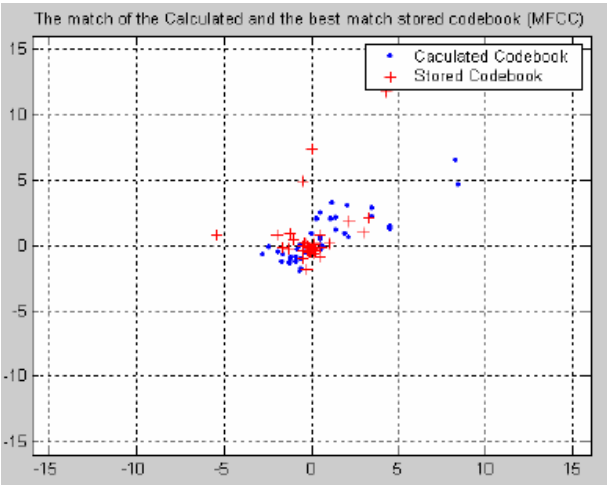

Fig. 2(b). Matching of the "us" voice with the stored codebook

4 Empirical Results and Findings

The interface of this project was designed to have few buttons and a display panel for simplicity purposes. By clicking on the "V2S" button on the screen the user is allowed to input raw voice i.e. recording of spoken voice into the system. There is a display panel to display the appropriate video output (Sign Language) which is the relevant translation of the spoken word(s).

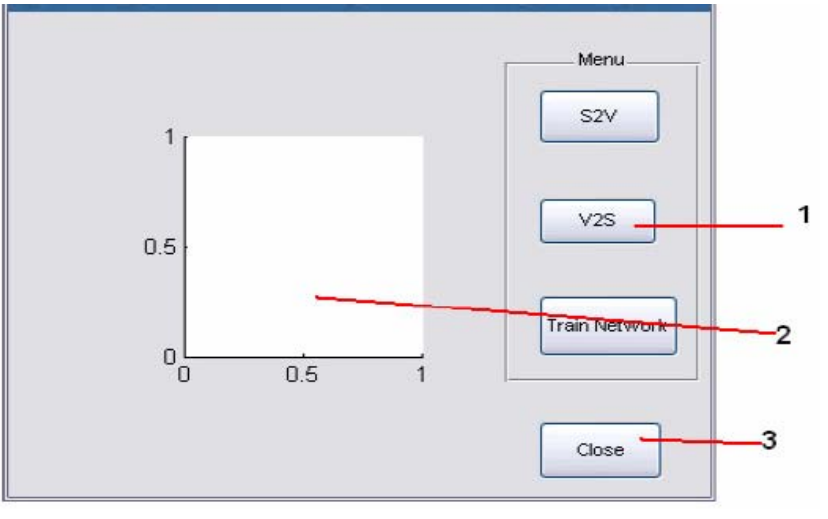

Fig. 3. The main user interface of the V2S system. (1) The V2S button – clicking it shall prompt the user for training the voice input. The system will then matches the voice to the corresponding video in the database. (2) The display panel – displays the relevant video for the translated voice. (3) Close button – exit the V2S system.

Fig.3 shows the main interface of the V2S prototype application. It is worth mentioning here that the prototype does include the S2V (Sign-to-Voice) system [9]. The S2V system was presented in WASET 2008 Congress in Singapore.

Fig.4(a) to (d) show some samples of the video output for the relevant translated words.

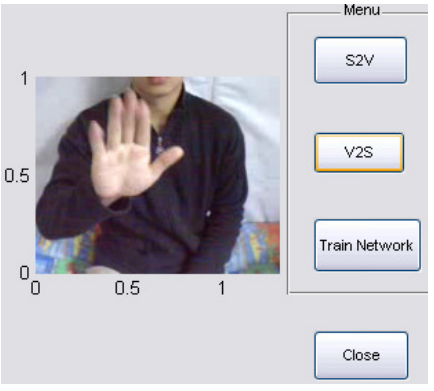

Fig. 4(a). Video for "come here"

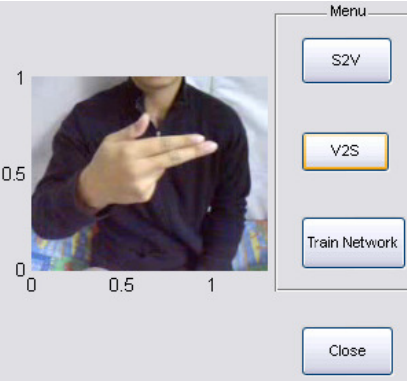

Fig. 4(b). Video for "turn left"

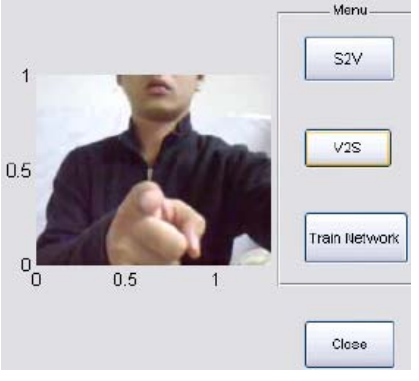

Fig. 4(c). Video for "you"

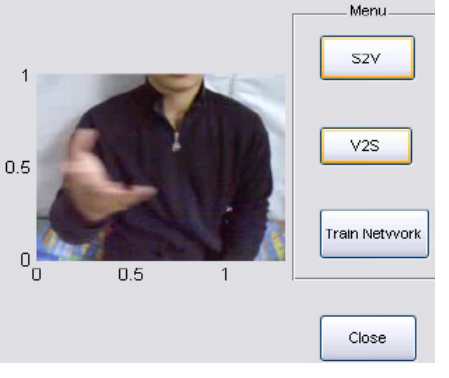

Fig. 4(d). Video for "us"

5 System Evaluation

We took a sample of 100 people comprised of elderly women and men, young male and female children, and middle age male and female to test on the accuracy of the V2S system. The test was monitored and conducted during the PECIPTA 2007 (Expositions of Research and Inventions of International Institution of Higher Learning) in Kuala Lumpur and CDC (Career Development Carnival) exhibitions in UTP. It was found that at least 80 out of the 100 people were able to callback the desired sign language video. That means the system is at least 80% accurate to display the correct sign language. It should be mentioned here that the words uttered by the test samples

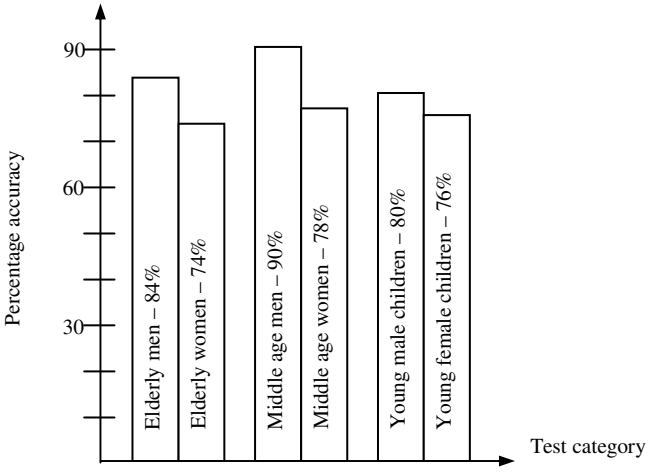

Fig. 5. Overall system accuracy test

were “you”, “us”, “come here”, and “turn left”. Fig.5 shows the overall statistic of the test conducted.

The graph shows the accuracy of each category by averaging the accuracy percentage of the 4 spoken words. The middle age man category shows the highest accuracy. This may be due to the training of the system via ONLY the middle age man. However, other categories do show high accurate responses. This implies that the more training the system gets for each word, the more accurate it would be.

6 Conclusion

Natural Language to SL translation is the main scope of this research. The fundamental idea of this system is to translate the human voice to SL. The system will match the captured voice with the pre-stored SL videos in the database to display the appropriate sign/gesture thus provide an alternative interactive way of communication between a normal person and a hearing impaired person.

The prototype allows translation of spoken English to SL in Malaysian context. The system accuracy depends on how much system training was conducted. With sufficient training, it will be able to recognize all the trained commands or words and execute the corresponding translation. Currently, the system has the accuracy of 80.3%. This system may be used by users who wish to learn SL and to help those who wish to communicate with the hearing disabled people more effectively. For future work, the system may be implemented in mobile devices with animated hand gestures [10] for deaf people.

References

1. The World Federation of the Deaf, http://www.hearinglossweb.com/res/hlorg/wfd.htm
2. Cornucopia of Disability Information: Disability Statistics, http://codi.buffalo.edu/graph_based/demographics/.statistics.htm
3. Zetterholm, E.: Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success, PhD thesis, Lund University (2003)
4. Zue, V., Cole, R., Ward, W.: Speech Recognition. Cambridge University Press, New York (1997)
5. Hung, J.W.: Optimization of Filter-Bank to Improve Extraction of MFCC Features in Speech Recognition. IEEE Transactions on Intelligent Multimedia, Video and Speech Processing, 675–678 (2004)
6. Rashidul Hasan, M., Mustafa, J., Golam Rabbani, M., Saifur Rahman, M.: Speaker Recognition Using Mel Frequency Cepstral Coefficient. In: 3rd International Conference on Electrical and Computer Engineering, pp. 565–568 (2004)
7. Gray, R.M.: Vector Quantization. Morgan Kaufmann, San Francisco (1990)
8. Wikipedia, Malaysia Sign Language, http://en.wikipedia.org/wiki/
9. Foong, O.I., Low, T.J., Satrio, W.: Hand Gesture Recognition: Signs to Voice System (S2V). In: WASET conference proceedings, vol. 33, article 6, pp. 32–36 (2008) ISSN 2070-3740
10. Sequndo, R.S., Montero, J.M., Guarasa, J.M., Cordoba, R., Ferreiros, J., Pardo, J.M.: Proposing a Speech to Gesture Translation Architecture for Spanish Deaf People. Journal of Visual Languages and Computing 19, 523–538 (2008)