Discussion-facilitator: towards enabling students with hearing disabilities to participate in classroom discussions

Mohammad A. Alzubaidi* and Mwaffaq Otoom

Computer Engineering Department,
Yarmouk University,
Irbid 21163, Jordan
Email: maalzubaidi@yu.edu.jo
Email: mof.otoom@yu.edu.jo
*Corresponding author

Abstract: Class participation plays a vital role in the learning process during classroom instruction. Deaf students often have difficulty participating in class discussions. Several studies have shown that deaf people are better able to interpret speech when they can view the lip movements of a speaker. This paper proposes an assistive device, called the Discussion-Facilitator, which aims to enable deaf students to better participate in classroom discussions. This is done by combining the speech-recognised text of the lecture with a live video stream that is zoomed-in on the lecturer’s face. The student is also able to write a text response and play it on loud speakers. Nine deaf students conducted a usability test. The results show that viewing lip movements combined with the speech-recognised text of the lecturer contributed to the understanding of the lecturer’s speech, and that our prototype makes the engagement of deaf students in classroom discussion more effective.

Keywords: assistive technology; class participation and discussion; deaf/hard of hearing; image processing and computer vision; lip reading.

Reference to this paper should be made as follows: Alzubaidi, M.A. and Otoom, M. (2018) ‘Discussion-facilitator: towards enabling students with hearing disabilities to participate in classroom discussions’, Int. J. Technology Enhanced Learning, Vol. 10, Nos. 1/2, pp.73–90.

Biographical notes: Mohammad A. Alzubaidi is an Assistant Professor of Computer Engineering at Yarmouk University, Jordan. He received his PhD degree in Computer Science from the Ira Fulton School of Engineering at Arizona State University in 2012. During his PhD, in joint work with Mayo Clinic Hospital, he developed novel computational methods to solve decision making problems in the medical field, as well as to facilitate medical training. His research areas are medical imaging, computer vision and pattern recognition, anomaly detection, machine learning, bioinformatics, and healthcare applications.

Mwaffaq Otoom received his PhD in Computer Engineering from Virginia Tech in 2012. Since then, he has been an Assistant Professor with the Computer Engineering Department, Yarmouk University, Jordan. His research interests include novel modelling techniques for embedded parallel computing.

Copyright © 2018 Inderscience Enterprises Ltd.
1 Introduction

More than 360 million people worldwide suffer from hearing impairment due to complications at birth, ear infections, exposure to excessive noise, genetic causes and aging (World Health Organization, 2016). One of the major problems that these people often face in their daily life is the inability to participate in discussions with their peers. Examples include discussions that take place during social and work meetings, as well as those that take place among students, or between students and their lecturer during classroom lectures.

Class participation plays a vital role in the learning process during classroom instruction. Students who are actively engaged in class participation and discussion and who are able to ask the lecturer questions (or answer questions put to them by the lecturer) retain information and understand the lectures better. However, students who are hearing impaired might not be able to fully participate in class - possibly discouraging them from even attending the lectures.

Several accommodations have been used to make classroom lectures more accessible to students who are hearing impaired. A student who is deaf might benefit from a sign language interpreter, or a student note-taker. The sign language interpreter might translate spoken English into American sign language (ASL) gestures. The student note-taker is typically another student taking the same class, who takes notes during the lecture and then shares those notes with the student who is deaf. Lecturers can also customise their teaching materials to be more accessible to students who are hearing impaired. For example, video or slide materials presented to the students could include informative captions. Ideally, the lecturer would avoid an oral-intensive lecture style.

Several studies have shown that if people who are hearing impaired are able to closely view lip and tongue movements (as well as the facial expressions of a speaker) they are better able to recognise, interpret and understand what is being spoken (Toba et al., 2015).

Based on that observation, we propose a solution that we call the Discussion-Facilitator, which is a portable device that allows students who are deaf or hard of hearing to better participate in classroom discussions.

This paper is organised as follows: Section 2 surveys the existing assistive technologies, literature and related work. Section 3 describes the methodology in terms of system architecture of the proposed system and its main components, the implementation details of the system and the testing and validation of the proposed system. In Sections 4 and 5, we present and discuss, respectively, the results of validating the different components of the proposed system. The paper is then concluded in Section 6 where future research directions are also discussed.

2 Literature review

This section reviews the existing literature on classroom discussion and participation relevant to students with disabilities. It also presents some popular assistive devices for people who are hearing impaired, along with related research. The section then discusses the process of lip reading. Finally, it presents the research question that is addressed in this paper.
2.1 Classroom discussion and participation

It has been shown that the discussion and participation of students during classroom lectures play a vital role in understanding and comprehending the lecture and helps them retain the information delivered during the lecture (Wassermann, 2010). However, students who are hearing impaired are at a disadvantage, when compared to peer students who are able to hear normally.

There are several accommodations that can be used to help hearing impaired students better benefit from the classroom lectures (Services for Students with Disabilities at the University of Texas at Austin, 2016; Disability Resource Center at Arizona State University, 2016; Knoors, 2016):

1. A sign language interpreter can be hired to provide a real-time translation of the lecture for these students.
2. A human note-taker can be hired, to take notes for these students.
3. The classroom lecturer can be instructed to present the material in textual or visual (as well as oral) form and to make the material less-oral-intensive wherever possible.

While these accommodations are helpful, they have the following disadvantages:

1. They require additional people to attend the lecture (such as an interpreter and/or a note-taker) and they might require a reconfiguration of the classroom set-up.
2. They might not adequately engage the students who are hearing impaired in spontaneous classroom interactions, where questions might be asked and answered.
3. They require classroom lecturers to adapt their presentation and teaching materials.

During the design process of our proposed assistive device, we took into consideration the above-mentioned challenges.

2.2 Assistive technologies and devices for people who are hearing impaired

With the advent of digital, internet and wireless technologies, several types of assistive devices have been developed to help people who are hearing impaired in their daily life activities. The aim of such devices is often to allow these people to better understand what is being said by others and to allow them to better communicate their own thoughts (Hersh, 2006).

The National Institute on Deafness and Other Communication Disorders (2016) categorises popular assistive devices into three main categories:

1. assistive listening devices
2. augmentative and alternative communication devices
3. alerting devices.

2.2.1 Assistive listening devices

Assistive listening devices make sounds more intelligible by amplifying the sounds and by selectively removing background noise. For example, a hearing loop (National Institute on Deafness and Other Communication Disorders, 2016; Myers, 2010) is an
assistive listening device that employs a large magnetic induction loop of cable, which is installed around the area of interest (such as a home, a meeting room or an auditorium). The loop produces a fluctuating magnetic field throughout the area of interest, which can be picked up by a hearing aid (or a Cochlear implant) equipped with a T-coil pickup. The induced audio typically carries a clean audio signal from microphone or a sound system, without any distracting noises from the environment. The logo shown in Figure 1 is used to indicate that a hearing loop system has been installed in an area.

Figure 1  Hearing loop logo (National Institute on Deafness and Other Communication Disorders, 2016) (see online version for colours)

There are two limitations to the use of a hearing loop. It is only useful for people who are not completely deaf, and it can only be used in rooms that have been equipped with a surrounding induction coil.

2.2.2 Augmentative and alternative communication devices

Augmentative and alternative communication devices help people with hearing disabilities to better communicate and express themselves. For example, a speech-generating device (Aetna Inc, 2016; Ashraf et al., 2002; Sutherland et al., 2010) can translate pictures (or text) into speech. This enables people who are not able to speak, to verbally communicate their needs. Figure 2 shows a picture-based speech generation device that allows a deaf person to communicate a limited number of words or phrases by clicking on the corresponding image.

Figure 2  Speech-generating device (see online version for colours)
Discussion-facilitator

This technology might be useful in the classroom, so our proposed system adopts this feature (after extending it to include any text) to engage deaf students in the class discussion, in real time.

2.2.3 Alerting devices

Alerting devices employ non-audio notifications to people with hearing disabilities. For example, Deaf Alert devices are used to notify people with hearing impairment when an important event occurs, using a vibration (Sułkoski et al., 1983) or a flashing light (Sparber, 1983). Figure 3 shows an example - an alarm clock that is augmented with a bed shaker.

Figure 3  Alarm clock with bed shaker (see online version for colours)

While this feature has many useful applications for deaf people in their daily life activities, we did not include this feature in our proposed system, as we did not expect it to play an important role in classroom discussion facilitation.

2.3 Automatic sign translation

Automatic translation systems from speech to sign language have been proposed (Dreuw et al., 2007). The basic idea of these systems includes

1. converting speech to text, using speech recognition algorithms
2. converting that text into sign language text, using machine translation methods
3. converting the sign language text into sign language movements - either by streaming a pre-recorded video, or by animating a pre-programmed 3D avatar.

These systems have been developed for several different languages, such as English (Zhao et al., 2000; Veale et al., 1998), Arabic (Halawani, 2008) and Spanish (San-Segundo et al., 2008).

For example, López-Ludeña et al. (2014a) used sign language machine translation to generate bus schedule information in Spanish sign language. The translation model used is a hybrid of both example-based and statistical machine translation. The system also included a sign animation avatar. The same methodology has been employed in a hotel domain (López-Ludeña et al., 2014b). Zhao et al. (2000) built a machine translation system from English to ASL, which generated all of the visual and spatial information associated with ASL signs. The project employed synchronous tree adjoining grammar
rules to represent an ASL grammar. Othman and Jemni (2011) developed a full system to translate spoken language into ASL, using statistical sign language machine translation.

Often, the sign language for a given spoken language is not standardised, which makes it hard for deaf students to take advantage from such systems during class (Toba et al., 2015). Moreover, the translation, in general, is not fluent, as there are several sources of errors as well as complex components (such as facial expressions) that also need to be translated.

2.4 Lip reading

Lip reading is the process of understanding what a person is saying by watching his/her lip and tongue movements and facial expressions (Dodd and Campbell, 1987).

Speech perception is a subconscious process that involves both hearing and vision. Even people without hearing impairment use visual information from a speaker’s lips and facial expressions to better understand his/her speech. Thus, fluent speakers of any language can lip read, to some degree (Nath and Beauchamp, 2012). Some deaf or hard of hearing people become highly proficient at lip reading, which helps them understand speech from the lip movements of the speaker (Toba et al., 2015).

It has been shown that even young children learn the relationships between lip movements and speech sounds (Dodd, 1979). It has also been shown that the integration of automatic lip reading information into speech processing leads to better performance in speech recognition (Duchnowski et al., 1994).

Conrad (1977) showed that the performance of lip reading for some deaf people is close to those with normal hearing. The study also showed that the performance becomes better when augmenting the viewed speech with text. Kyle and Margaret (Kyle and Harris, 2006) also showed that lip reading is a significant predictor of reading proficiency for deaf children, confirming the conclusions of previous studies (Arnold and Köpsel, 1996; Campbell and Wright, 1988; Geers and Moog, 1989) that there is a strong correlation between lip reading and the ability to read.

2.5 Summary

Participating in classroom discussions is important to a student’s full understanding of the lecture. In this regard, students with hearing impairment are at a disadvantage. Most universities provide a sign language interpreter and/or a note-taker for those students. However, the latter does not engage the student in the learning process during the lecture.

Several assistive devices and technologies and ambient intelligence systems have been developed to help people with hearing impairment to better communicate and express themselves. However, these solutions have not specifically addressed the issue of facilitating classroom discussions.

Lip reading research suggests that viewing lip and tongue movements (as well as facial expressions) facilitates the understanding of speech for people with hearing impairment. It also suggests that presenting the text of the speech enhances performance. With this in mind, this paper poses the following research question:

Would a portable assistive device that facilitates lip reading and that employs both voice recognition and voice synthesis (i.e. ambient intelligence) allow students with hearing disabilities to more effectively participate in classroom discussions?
3 Methodology

3.1 System architecture

This section describes the basic architecture of the proposed system and the overall system functionality. As shown in Figure 4, the system is comprised of two communication paths:

a The first path delivers the lecturer’s speech to the student.

b The second path delivers the student’s response to the lecturer.

Figure 4 Overall system architecture

In the first path, a video camera (with a pan/tilt mechanism and an optical zooming mechanism) captures a close-up video stream of the lecturer’s face (showing lip and tongue movements as well as facial expressions) while a highly directional microphone captures an audio stream of the lecturer’s speech. A speech recognition algorithm is then applied to the audio stream, to produce a real-time text stream of the speech. At the same time, a video enhancement algorithm is applied to the real-time video stream, to enhance its quality. The enhanced video stream (combined with the text stream) is then displayed for the student.

If the student is not able to speak clearly, he/she can use the second path to talk to the lecturer as follows. The student enters a response as text. A text-to-speech convertor then converts the text to synthesised speech, which is delivered to the lecturer through speakers.

The proposed architecture has been implemented as a portable device. The system design (shown in Figure 5) consists of a pan-tilt mechanism, a digital video camera and a Laptop PC. The student places the pan-tilt mechanism on the desk, mounts the video camera onto the pan-tilt mechanism and then connects both the pan-tilt mechanism and the camera to the Laptop PC, through USB cables.
The system user interface (shown in Figure 6) has three main windows:

a. a window that displays live video from the camera
b. a window that is used as a camera control panel, to control the camera’s zoom and pan-tilt movements
c. a window that is used as a text editor and text-to-speech converter.

During a classroom session, as the lecturer discusses the material with the students, a student who is hearing impaired can view a close-up video of the lecturer’s face. Within the same window, the text (from the speech recognition system) is displayed as rolling subtitles. At any time, the student can use the text editor window to enter text and then send this text to the Laptop speakers in the form of synthesised speech.

3.2 System implementation

This section describes the implementation details for the proposed system architecture presented in the previous section.

3.2.1 Video camera with pan/tilt mechanism

The Discussion-Facilitator system uses a Sony digital video camera that allows up to 27× optical zoom and streams video at 30 FPS. The system uses the video zooming feature to provide a magnified view of the lecturer’s face, allowing the user to read the lecturer’s lips.
The pan-tilt mechanism allows the user to move the camera up/down and right/left to follow the lecturer, who might move to a new location from time to time during the classroom lecture.

3.2.2 Speech recognition

Speech recognition is the process of translating spoken language into text (Wikipedia, 2016). Several methods have been proposed in the literature for speech recognition (Besacier et al., 2014).

The first step in this process is to extract the streaming audio from the video stream that is received from the camera. This extracted voice (audio) signal is then used as input to a voice recognition system (VRS) which translates the input voice signal into a text stream. This is done as follows:

1. The input voice signal is tokenised into utterances based on a ‘salience’ delimiter.
2. Voice features are then extracted from each token to generate a feature vector to represent that token.
3. The sequence of feature vectors extracted from the tokens is then compared with a collection of models, each of which is associated with a word or phrase. The best type of learning model for speech is the hidden Markov model, as the process in this model is described as a sequence of states that change, based on probability.
4. The best matching model is then selected and its corresponding words are used to create the English text stream.

In this work, the CMU Sphinx (Sourceforge, 2016) software is used as the VRS to translate the English speech of the classroom lecturer into the corresponding English text. CMU Sphinx software is freely available. First, we downloaded Sphinx-4 and installed it. Next, we downloaded the SphinxTrain tool, which is the acoustic training environment for Sphinx-4. This tool is used to prepare the language dictionary and model, and the acoustic model, which are necessary in the voice recognition process. It is also used to fine tune the acoustic model to handle the variation in speech by training it using several variations of speech samples. In this work, we used the English dictionary provided by CMU, and we used the SRILM (Stolcke, 2002) as a language model with ARPA text format. For the acoustic model, phonetically tied mixtures (PTM) (Digalakis et al., 1996) with 5,000 senones and 128 mixtures was used. This PTM model provides a great balance between decoding speed, accuracy and model size, which fits for mobile devices.

The recognition process is then done as follows. The sound waveform is converted into features. Then we find phonemes that best fit the extracted features and, using the language dictionary and the model, we match the sequence of phonemes to words.

3.2.3 Image processing and enhancement

Some external factors (such as less-than-ideal lighting) might adversely affect visibility within the classroom, causing dim or low contrast video. Video enhancement can be used to address these issues.

Video enhancement is the process of converting low-quality video to higher quality video by enhancing the visual appearance of the video. The selection of an appropriate
algorithm is highly dependent on the application in which we need to apply the video enhancement process, and this can make video enhancement a challenging process.

Generally speaking, video enhancement methods are classified into two main categories: spatial-based and transform-based methods. While spatial-based methods work on the actual pixels of the video frames, the transform-based methods transform the pixels into a new domain before processing them (Rao and Chen, 2012).

Spatial-based methods are simple and have low time complexity. Therefore, spatial methods are more suitable (and preferred) for real-time applications, over transform methods.

Contrast enhancement is a post-processing, spatial-based image enhancement method that increases local contrast within an image, to make visual details more perceptible to the eye of a human viewer. The most commonly used contrast enhancement technique is histogram equalisation.

Histogram equalisation (Russ, 2016) adjusts the distribution of pixel intensities within any given video frame or image to take the fullest advantage of the dynamic range that is available in the image format. Perceptually, this tends to enhance low-contrast features that might otherwise be difficult to perceive.

3.2.4 The live video window

The enhanced real-time video stream of the lecturer’s face is augmented with an English text stream from the voice recognition. The augmented video stream is then presented to the user.

3.2.5 The text editor window and the text-to-speech convertor

Some people who are deaf are not able to speak. Some who can speak prefer not to - possibly because of their inability to control the volume of their voice (If My Hands Could Speak, 2009).

To provide these people with a voice in the classroom, the system provides a feature that converts text that they enter on a keyboard into synthesised speech, using a text-to-speech convertor. Note: This feature can be turned off by students who prefer to speak orally.

In this prototype of the Discussion-Facilitator system, we developed a simple interface (shown on the left side of Figure 6) that accepts text from the student via a simple text editor and then converts that text into speech by calling the Google text-to-speech API (Google, 2016).

3.2.6 Speaker

The resulting synthesised speech is then sent to a loudspeaker. For this prototype, we used the built-in speakers of the laptop PC.

3.3 Experimentation: usability study

This section describes the testing methodology used to evaluate the performance of the prototype Discussion-Facilitator. Since it is always important to employ potential users to evaluate the performance of any assistive technology, the authors contacted a disability
centre that supports students with special needs. We presented the *Discussion-Facilitator* prototype to students who are deaf. We received a positive first impression. We then invited nine of the students to use our prototype during a real classroom lecture and to provide us with feedback about its usefulness for facilitating classroom discussions. Each participant was asked to fill in an evaluation sheet. Evaluation statistics were then summarised to provide an objective measure of the system performance as well as a basis for refining the next version of the *Discussion-Facilitator* system.

Nine students who are deaf participated in the usability study. Three of the participants were females. The ages ranged from 18 to 22. We arranged to have each of them individually use the prototype during a 1-hour classroom lecture of their own choice from their programs of study.

Data was then collected on how helpful the features of the prototype were. This was done by asking each student, after using the prototype, to fill in the evaluation sheet shown in Table 1.

Table 1  
**Evaluation sheet**

<table>
<thead>
<tr>
<th>#</th>
<th>Feature</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Zooming in the lecturer's face to interpret his speech through lip and tongue movements and facial expression.</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>2</td>
<td>Incorporating the recognized text with the video stream of the lecturer's face to understand the lecture.</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>3</td>
<td>The ability to write a response back.</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>4</td>
<td>The ability to convert the written response to speech to be delivered to the lecturer.</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>5</td>
<td>The system user interface – video streaming.</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>6</td>
<td>The system user interface – text editor.</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>7</td>
<td>The system hardware – camera with pan/tilt.</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>8</td>
<td>The system hardware – Laptop PC.</td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>

Comments: ___________________________________________________________________

_____________________________________________________________________________

4 Results

4.1 Speech recognition

To evaluate the performance of the speech recogniser used in our system we sampled classroom video streams, with a total duration of approximately 30 minutes. We extracted the speech into wav files and then applied the speech recognition process. We then manually transcribed the speech and calculated word error rate (WER) using the word_align.pl tool from Sphinxtrain, in terms of the percentage of inserted, deleted and substituted words. We found the WER of the speech recogniser to be 4.31%.
4.2 Image processing and enhancement

We evaluated the real-time video histogram equalisation technique on video captured during classroom lectures. The selected video was of low quality, due to the rather poor lighting in the room, which was optimised for slide projection. Figure 7 shows a before/after example of histogram equalisation. The left side of Figure 7 shows the original low-quality frame, along with its pixel histogram. The right side shows the frame resulting from our histogram equalisation, along with its pixel histogram. Both the enhanced readability of the projected slide and the improved visibility of the instructor’s face are clearly evident.

Figure 7  Histogram equalisation: original frame of low quality and histogram image (left) and the result of histogram equalisation and histogram image (right) (see online version for colours)

4.3 Usability study

Table 2 summarises the results collected with the evaluation sheet. It presents the original scores, average score and standard deviation, across the nine students who used the prototype, for each of the eight features presented in Table 1.
Table 2  Evaluation summary: for the nine participants (P1 to P9)

<table>
<thead>
<tr>
<th>#</th>
<th>Feature</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>Average score</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Zooming in the lecturer’s face to interpret his speech through lip and</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4.33</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>tongue movements and facial expression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Incorporating the recognised text with the video stream of the lecturer</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4.56</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>’s face to understand the lecture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>The ability to write a response back</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4.22</td>
<td>0.44</td>
</tr>
<tr>
<td>4</td>
<td>The ability to convert the written response to speech to be delivered to</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3.89</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>the lecturer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>The system user interface - video streaming</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4.33</td>
<td>0.50</td>
</tr>
<tr>
<td>6</td>
<td>The system user interface - text editor</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3.89</td>
<td>0.33</td>
</tr>
<tr>
<td>7</td>
<td>The system hardware - camera with pan/tilt</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3.44</td>
<td>0.53</td>
</tr>
<tr>
<td>8</td>
<td>The system hardware - laptop PC</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4.44</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Figure 8 summarises Table 2 results visually in the form of a diverging stacked chart - the most commonly used visualisation method for Likert scale type of data (Heibeger and Robbins, 2014).
5 Discussion of results

5.1 Speech recognition

The results presented in Section 4.1 suggest that the system is helpful for facilitating discussions in the classroom environments. However, the results also suggest that there are opportunities to improve the accuracy of the speech recogniser system.

Another important issue is the synchronisation of the speech-recognised text stream with the video stream, which shows the lecturer’s lip movements. We assumed that this synchronisation (which depends upon the speed of the speech recogniser) would be very important to the user experience. However, the users of the system (i.e. the deaf students) were not able to give us precise feedback on the degree of synchronisation, because

1. they cannot compare the lecture’s speech with the recognised text
2. they do not simultaneously view the lecturer’s lip movements and the recognised text stream - focusing, at any point in time, on either the lecturer’s lip movements, or on the text stream.

Typically, the student would focus on the video to lip read and only look at the text stream when unable to interpret the lip movements. This argues that the text stream should actually trail the video somewhat, so that the student can still catch the written text when unable to understand the lip movements. The question of optimum delay will be explored in future research.

5.2 Image processing and enhancement

As shown in Figure 7, histogram equalisation adjusted the distribution of pixel intensities within the video in such a way as to increase the contrast within the low-contrast regions of the frame - thus enhancing the overall quality of the original video frame. Given the adverse lighting present in many classrooms while slides are being projected, we judged this feature to be helpful (if not vital) for facilitating understanding and interaction within the classroom.
5.3 Usability study

The results presented in Table 2 and Figure 8 show that most of the features of the prototype were judged to be helpful. These results also suggest that viewing lip and tongue movements (as well as facial expression) when combined with the speech-recognised text stream of the lecture, contributed to the understanding of the lecturer’s speech.

The results also suggest that the ability of the user to enter text responses to the lecturer, convert the text to synthesised speech and deliver the speech in real time to the lecturer through speakers, did add some value to the system as it allows the deaf students to express their thoughts and makes the engagement of deaf students in classroom discussion more effective.

On the other hand, the average scores for features 4, 6 and 7 were relatively low, compared to the rest of the features. Regarding feature 4, some students commented that, when delivering the synthesised speech to the lecturer, there is no way to know whether that response was delivered or not, and that it would be better to provide some kind of notification - either visual or vibrational. Regarding feature 7, some commented that the zoom and pan/tilt response of the camera was somewhat slow and noisy.

6 Conclusion and future work

This paper has proposed a portable assistive device, called the Discussion-Facilitator, which employs lip reading to allow students who have a hearing disability to better engage in classroom discussions. During a classroom lecture, the device combines a text stream (produced with speech recognition) and a live video stream of the lecturer’s face (showing lip and tongue movements and facial expressions) and then simultaneously presents both streams to the user visually, to facilitate speech understanding. The system also allows the student to communicate his/her thoughts to the lecturer by typing a response and then voicing it out to the lecturer. The results presented in this paper indicate that this prototype was judged to be helpful in the classroom by students who are deaf.

With regards to future work, the results of this research have raised additional research questions. To conduct further research towards finding answers to those questions, the following changes will be made to the system:

1. The zoom and pan/tilt mechanism of the camera will be made faster and less noisy and will be re-evaluated for user satisfaction.
2. A face detection algorithm will be implemented to assist the user, by making automatic pan/tilt adjustments to the camera. The user will be allowed to use manual controls to centre the camera on the lecturer’s face, and the face detection algorithm will then follow the lecturer as necessary.
3. Both a visual and a vibrational notification feature will be added, to notify the user upon successful playback of (text-to-speech) responses to the lecturer.
4. The synchronisation delay between the text stream and the video stream will be made variable, to allow the user to adjust that delay.
Various video enhancement algorithms will be evaluated, to determine their relative benefits with regard to lip reading.

Other alternatives will be explored to provide real-time speech recognition.

References


Discussion-facilitator


Services for Students with Disabilities at the University of Texas at Austin (2016) *Sign Language Interpreter Services*. Available at: http://ddce.utexas.edu/disability/sign-language-interpreter-services/ (access 15 July 2016).


