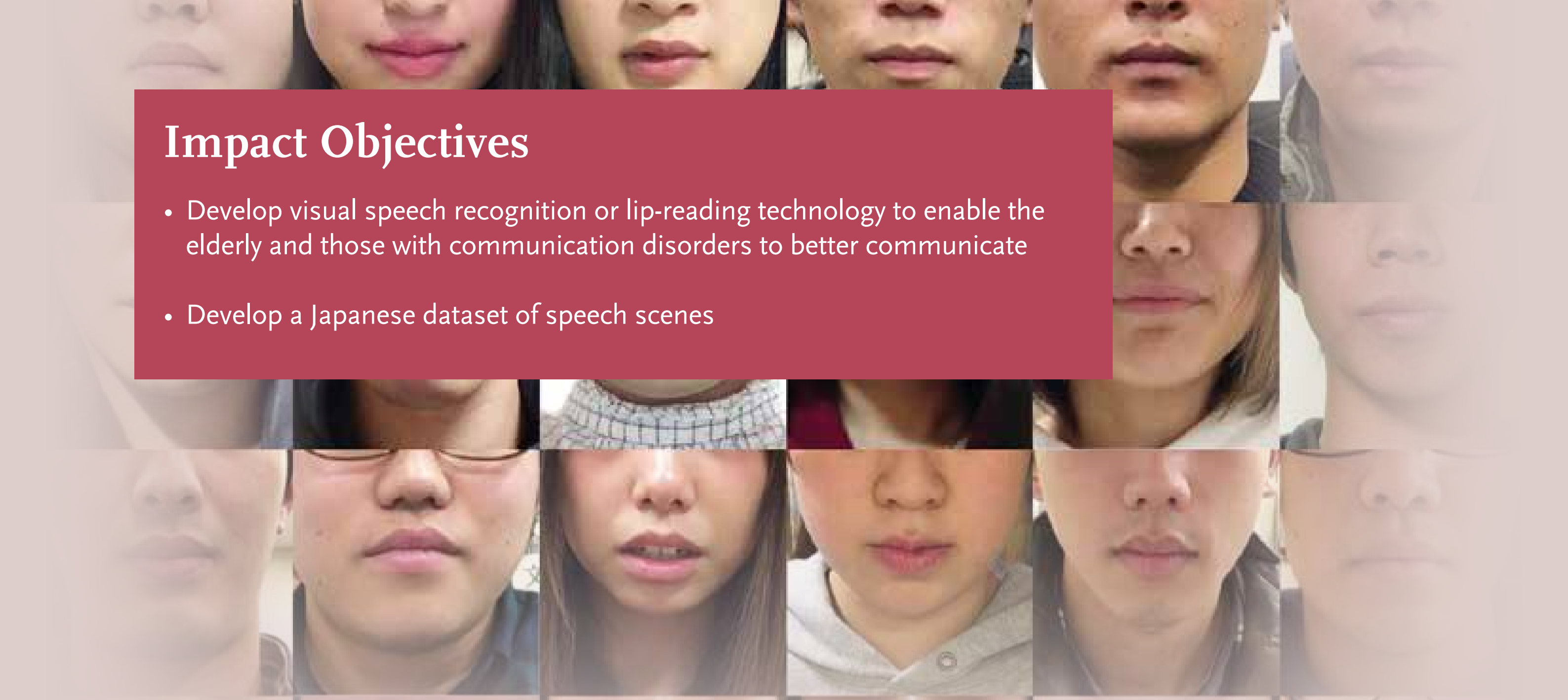

Impact Objectives

- Develop visual speech recognition or lip-reading technology to enable the elderly and those with communication disorders to better communicate
- Develop a Japanese dataset of speech scenes

Lip-reading technology

> *__Associate Professor Takeshi Saitoh__ and his team at the __Kyushu Institute of Technology__ strive to make speech recognition a reality for the speech and hearing-impaired and for human-machine interaction*

What is your background, and how was your interest in visual and silent speech recognition first sparked?

I studied image processing and pattern recognition as a student and when I became Assistant Professor at Tottori University in 2004, I was first introduced to the field of visual speech recognition (VSR). After moving to Kyushu Institute of Technology in 2010, I became interested in silent speech recognition technology using novel approaches, including surface electromyography (EMG) and electrocardiograms (EEC).

VSR is lip-reading using technology, and the purpose of my research is to develop a means by which speech and hearing-impaired people can communicate easily with others. The number of researchers in this field is small, but we are a global community, with the potential to improve many people's lives. I want to do something for these people – to make their lives better and to give them greater employment opportunities. This is my motivation for studying VSR.

What is the objective of your current research project?

Visual and silent speech recognition are techniques that can be used by speech-impaired people, or in noisy environments when normal audio speech recognition cannot work, to communicate with machines, with other humans, or for authentication. With the growth of image recognition technology using deep learning, the recognition accuracy of VSR is very good. In time, the technology will be available for anyone to use, in any situation, whether social, in public or in a clinical setting, and regardless of whether the person is mobile or bedridden. Silent speech recognition technologies include non-audible murmur, colour imaging, depth imaging, ultrasound imaging, surface EMG and electroencephalogram-based brain-computer interfaces. We are working to improve several of these technologies, which are currently quite under-developed.

What have been the main difficulties you have faced during the project?

Audio speech recognition (ASR) is already a mature technology and in general use. VSR faces many more difficulties. For instance, with image processing and pattern recognition techniques, only the lips and mouth shape can be seen by the camera. It is impossible to observe the tongue movements and shapes inside the mouth. The mouth shape depends almost solely on the vowel sounds being made and we can reliably recognise these from a frontal image. Japanese has just five vowels, which makes this task easier. However, consonant recognition is much more difficult. It might also be difficult to recognise vowel sounds expressed in different accents, but we can address this by greatly increasing the amount of data to which the technology has access, in order to find the best match.

Another issue is the difference in sampling frequency between VSR and ASR. The sampling frequency of ASR is 8kHz or greater, whereas the sampling frequency of VSR is approximately 30Hz. Therefore, there is a big difference between ASR and VSR. VSR is much more difficult to achieve than ASR.

How important is collaboration to your work?

When I began to exchange information with other researchers across the world, I was able to solve difficult problems. Moreover, by interacting with people who will be the users of VSR technology, I was able to confirm their needs and correct the direction of my research. My laboratory is fairly small, and many of the assistant researchers are undergraduates and masters course students. By participating in international conferences and particularly by learning and networking through the annual VSR workshops we host in Japan, we have been able to achieve much more than we hoped.

Silent speech recognition

*The **Kyushu Institute of Technology** is working at the cutting edge of visual and silent speech recognition technology with the aim of producing a system to help elderly and disabled people communicate without speech, regardless of their situation or level of technical literacy*

We are all familiar with audio speech recognition technology for interfacing with smartphones and in-car computers. However, technology that can interpret our speech signals without audio is a far greater challenge for scientists. Audio speech recognition (ASR) can only work in situations where there is little or no background noise and where speech is clearly enunciated. Other technologies that use visual signals to lip-read, or that use lip-reading in conjunction with degraded audio input are under development. However, in the situations where a person cannot speak or where the person's face may not be fully visible, silent speech recognition, which uses muscle movements or brain signals to decode speech, is also under development.

Associate Professor Takeshi Saitoh's laboratory at the Kyushu Institute of Technology (Kyutech) is at the forefront of visual speech recognition (VSR) and is collaborating with researchers worldwide to develop a range of silent speech recognition technologies. Saitoh, whose small team of researchers and students are being supported by the Japan Society for the Promotion of Science (JSPS), says: 'The aim of our work is to achieve smooth and free communication in real time, without the need for audible speech.' The laboratory's VSR prototype is already performing at a high level.

VISUAL SPEECH RECOGNITION

There are many reasons why scientists are working on speech technology that does not rely on audio. Saitoh points out that: 'With an ageing population, more people will suffer from speech or hearing disabilities and would benefit from a means to communicate freely. This would vastly improve their quality of life and create employment opportunities.' Also, intelligent machines, controlled by human-machine interfaces, are expected to become increasingly common in our lives. Non-audio speech recognition technology will be useful for interacting with smartphones, driverless cars, surveillance systems and smart appliances.

VSR uses a modified camera, combined with image processing and pattern recognition to convert moving shapes made by the mouth, into meaningful language. Earlier VSR technologies matched the shape of a still mouth with vowel sounds, and others have correlated mouth shapes with a key input. However, these do not provide audio output in real-time, so cannot facilitate a smooth conversation. Also, it is vital that VSR is both easy to use and applicable to a range of situations, such as people bedridden in a supine position, where there is a degree of camera movement or where a face is being viewed in profile rather than full-frontal. Any reliable system should also be user-dependent, such that it will work on any skin colour and any shape of face and in spite of head movement.

Saitoh's team has developed a lip-reading system that combines face detection and active shape matching to achieve a high degree of accuracy and rapid real-time decoding. In any VSR model, the first step is face detection and the well-proven Viola-Jones face detector algorithm is used for this task. Two active appearance models are then applied. These use landmarks or feature points on faces that are compared with images in a database to select the closest match. The first model uses 37 points on the eyes, eyebrows and nose to accurately locate the face, and the second applies 38 points to the lips and nostrils.

TRAINING THE SYSTEM

The system is trained by inputting multiple common phrases many times, so that several samples for each phrase or sound are included in the database. Recognition is achieved by looking for the closest match between the new input and the image representations in the databank. For real-time processing, the system incorporates an utterance section extractor, which uses the shape of the mouth and duration of speech pauses to identify separate phrases for decoding. The camera itself is modified to process the facial images to the same stable light levels, thereby aiding the recognition software.

The prototype system provides a user interface that displays each decoded phrase on a screen for acceptance or deletion. An accurate phrase can be simply accepted by continuing to speak within a short interval or can be manually accepted by pressing a button. A Nintendo Wii controller has been used for ease of interaction. Similarly, an incorrectly rendered phrase can be simply rejected by pushing a button. A user can repeatedly input a phrase while the system tries different interpretations until it is accepted. In a controlled experiment the prototype achieved an average recognition accuracy of 94 per cent and a 1.4 second average time from completion of a statement to its communication as audio.

Saitoh says: 'Our team is now working on further enhancements to make the VSR prototype more accurate and reliable.' Saitoh's team has attempted to apply the system to mobile phones in conjunction with a handheld camera. However, when trialled, the recognition accuracy dropped to 71.5 per cent, possibly owing to camera shake. Saitoh notes: 'More work is needed

> The aim of our work is to achieve smooth and free communication in real time, without the need for audible speech

to differentiate between the mouth shapes for the ‘a’, ‘e’ and ‘i’ vowel sounds, since from the side, these look very similar. We need to apply more feature points around the mouth to pick up the subtle differences.’

We have recently begun collecting a new database named Speech Scene database by Smart Device (SSSD) for lip-reading or VSR. Speech scenes of conventional database available for lip reading or VSR, were record with a video camera fixed on a tripod in a well-maintained environment. On the other hand, VSR is expected to be used as an interface in smart devices such as smartphones and tablets. Therefore, collecting the speech scenes recorded with these devices is an important task for practical use of VSR. We collect word utterance scenes taken with smart device, and build a new database. I expect this database is useful for improving the VSR recognition accuracy.

Experiment scene in bedridden condition

INTERNATIONAL COLLABORATION

Whereas VSR is now very well developed, Saitoh is the first to admit that: ‘Silent speech recognition is currently at a very basic level.’ In 2015, other research teams were able to pick up brain electrical activity using invasive electrocorticography. Non-invasive electroencephalograms, whereby electrodes are placed on the outside of the scalp, can also be used to interpret brain activity, but speech decoding is currently difficult owing to the level of noise. Surface electromyography (EMG) is a promising technology being researched by Saitoh, who explains: ‘EMG uses electrodes on the face to detect the tiny electrical impulses produced by muscle movements, and from these deduce the sounds being formed.’

Saitoh recognises that international collaboration is the way forward and says: ‘We have organised four annual workshops on silent speech technology from 2014 to bring together the people working in this area and specialists from other supporting disciplines. These include sensor technology, signal processing, image processing and machine learning technology.’ These workshops have been very fruitful and have led to advancements such as the technology of deep learning to the Kyutech VSR. Saitoh is hopeful that we will see VSR systems being applied in everyday life in the near future: ‘I aim to overcome the remaining problems and to provide a practical solution for disabled and elderly people, so that they can communicate freely and have a better quality of life.’

Project Insights

FUNDING

This project has been supported by Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research and the Program for Building Regional Innovation Ecosystems.

CONTACT

Associate Professor Takeshi Saitoh
Project Leader

T: + 81 948 29 7713
E: saitoh@ces.kyutech.ac.jp
W: http://www.slab.ces.kyutech.ac.jp/en/index.html

BIO

Professor Takeshi Saitoh is an associate professor of Faculty of Computer Science and Systems Engineering at Kyushu Institute of Technology. His research interests are in image processing and pattern recognition. He is particularly focused in visual speech recognition, and other facial image processing for communication support.