MIETA: Multi-modal speech visualization application for deaf and hard of hearing people

Yusuke Toba, Shinsuke Matsumoto, Sachio Saiki,
Masahide Nakamura, Tomohito Uchino, Tomohiro Yokoyama,
Yasuhiro Takebayashi

Supporting deaf and hard-of-hearing (D/HH) people is a crucial social welfare measure. However, currently, communication support for D/HH people is insufficient for a conversation with multiple hearing people. In order to support D/HH people in these situations, this paper proposes a multi-modal speech visualization application, named MIETA, which provides various aspects of information about speech content. For the purpose of evaluating how actually useful the MIETA is, we conduct an evaluation experiment with actual nine D/HH students. As a result, MIETA contributes significantly to understanding, although it increases misunderstanding for D/HH people when there are voice recognition errors. Furthermore, subjects find that the MIETA's strength lies in its main characteristics, which includes multimodal speech visualization, whereas some find weakness in its voice recognition performance speed.

1 Introduction

Supporting deaf and hard-of-hearing (D/HH) people is a crucial social welfare measure. The focus of our study is *assisting D/HH people in understanding the content of real-time, face-to-face conversation with multiple hearing people*. This situation is extremely common and necessary for interacting in society; however, it is particularly difficult for D/HH people [15]. The key to communication support for D/HH people is *sensory substitution* based on visual perception [1]. D/HH people usually retrieve a variety of information from vision to complement their hearing impairment. Sign language and writing are well-known and effective sensory substitution methods. Some D/HH people also supplement their grasp of speech context from lip reading, facial expressions, and gestures.

However, currently, communication support for D/HH people is insufficient for a conversation with multiple hearing people, such as in a professional meeting [16] [21]. Though sign language has a strong advantage in communication speed, it requires significant effort to learn. In addition, not every D/HH person has enough skill to use sign language. A Japanese ministry has reported that the adoption rate of sign language in Japanese D/HH people is only 14.1% [5]. In contrast, writing has a significant advantage for the population who can use it, but has disadvantages in communication speed and in multispeaker situations. Although lip reading does not require hearing people to perform any special actions, mastering lip reading is much more difficult for D/HH people than sign language.

There are already some commercial products [7] [10] and academic studies [11] [13] related to visualization of conversation using some ICT technolo-

MIETA: 聴覚障害者のためのマルチモーダル発話可視化アプリケーション
鳥羽祐輔, 佐伯幸郎, 中村匡秀, 神戸大学大学院システム情報学研究科, Graduate School of System Informatics, Kobe University.
柗本真佑, 大阪大学大学院情報科学研究科, Graduate School of Information Science and Technology, Osaka University.
内野智仁, 横山知弘, 武林靖浩, 筑波大学附属聴覚特別支援学校, Special Needs Education School for the Deaf, University of Tsukuba.
コンピュータソフトウェア, Vol.34, No.4 (2017), pp.116–128.
[ソフトウェア論文] 2016 年 10 月 31 日受付.

gies. Many studies have been carried out in the field of automatic speech recognition (ASR) [8] [9] [23] and its visualization [3] [22] [24]. The primary focus of these studies is improvement of recognition performance. Some open ASR services also are appearing on the Web. However, these ASR systems have a limitation in the term of recognition accuracy and narrowness of vision. At first, current voice recognition technology still cannot achieve 100% accuracy. Furthermore, hearing ability can be omnidirectional. Thus, hearing people can usually grasp a variety of information from both visual and auditory senses simultaneously. In this way, they can understand conversation from multiple modalities. For example, they can search for information about a relevant topic while listening to someone speak. Visualizing only recognized text on a display may prevent a D/HH person from obtaining information like, *"Who is speaking now?"*, *"What kind of expression is s/he making?"*, or *"What does this technical term mean?"*

In this paper, we propose MIETA[†1], which is a *multi-modal speech visualization application* for information support of D/HH people. Here, the term *multi-modal speech visualization* means visualizing a content of speech by using multiple *visualization modes* which display various useful information to grasp conversation. Examples of visualization modes are speech-to-text (STT) mode, face mode and dictionary mode. Visualizing speech from many aspects is expected to help understanding in a mutually complementary manner. For example, recognition error and difficulty of lip reading may help each other by showing both STT mode and face mode.

For the purpose of evaluating how actually useful the MIETA is, we conduct an evaluation experiment to answer the following research questions:

RQ$_1$: *How helpful is* MIETA *for D/HH people in understanding natural conversation with hearing people?*

RQ$_2$: *What strengths and weaknesses do D/HH people find in* MIETA*?*

In order to answer **RQ**$_1$, we conduct a control experiment that uses quizzes to compare the understanding of Japanese conversation under 3 conditions: with MIETA, with sign language interpretation, and with neither tools nor support. Then, to answer **RQ**$_2$, we conduct a survey about how users feel about MIETA after the control experiment.

We find that, for **RQ**$_1$, MIETA contributes significantly to understanding, although it increases misunderstanding for D/HH people when there are voice recognition errors. For **RQ**$_2$, subjects find that the MIETA's strength lies in its main characteristics, which includes multimodal speech visualization, whereas some find weakness in its voice recognition performance speed.

2 Challenge and Scope

To provide equal opportunities for D/HH people to participate in social situations, there are two significant challenges: *helping them understand speech content* and *helping them express their thought*. We tackle the former challenge. In addition, we assume a concrete situation in which to apply our proposed system, because such solutions are highly situation-dependent. Many studies have addressed one-to-many situations, in which a single person speaks and D/HH people try to understand, such as a class in school [13] or an oral presentation at a conference [11]. Our study assumes a **real-time face-to-face situation with multiple speakers**, like a meeting in a company.

Although some communication means can complement each other, it is still difficult to deal with this situation. Sign language has a critical issue in terms of its adoption rate and learning efforts. Making conversation by writing causes reduction of efficiency of communication and requires deeper understanding to cooperate with D/HH people from all participants. Lip reading, method to refer lip movements of a speaker to grasp speech, has difficulty of handling the situation where not all speakers can always show their lips to the lip-reader while they are talking.

3 MIETA: Multi-Modal Speech Visualization Application

3.1 Requirements

The goal of this application is to assist D/HH people in understanding natural conversation with

[†1] "Mieta" means "visualized" in Japanese. Note that this paper is different from our previous work [18] [19] on software architecture and empirical evaluation.

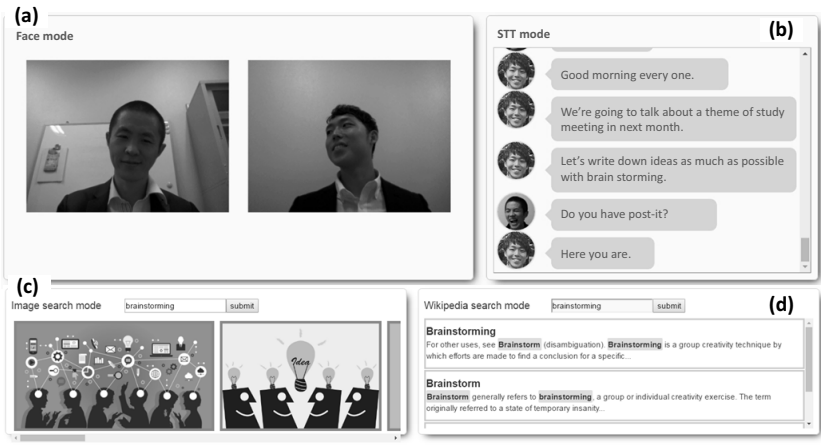


Fig. 1 MIETA graphical user interface (GUI)

multiple hearing people. To achieve the goal, we set the following four requirements, extracted from dozens of years of experience at the Special Needs Education School for the Deaf, University of Tsukuba.

R1: Low initial costs. First, it is important to have a low initial cost for both D/HH and hearing people. Initial cost includes preparing a special device, developing special software, or training a particular skill. Even D/HH people's perception of putting some cost on their group can lead to a mental burden. Therefore, it is ideal to make the initial financial cost not just cheap, but zero.

R2: Variety in information displayed. In general, hearing people deal with various kinds of information in understanding conversation, not only voice information, but also the speakers' facial expressions, gestures, and perhaps some external materials related to the topic. Thus, information support with only text derived from voice recognition is insufficient for fully grasping a conversation. Since D/HH people are accustomed to dealing with visual information, applying visual-perception-based sensory substitution effectively complements the lack of sound information [1]. Additionally, our system should have a rich variety of information and display methods to accommodate a diversity of requirements from D/HH people.

R3: Visible speech history. It is difficult for lip readers to handle situations with multiple speakers since they sometimes speak at the same time. Furthermore, natural conversation information can be too overwhelming to take in at once. Therefore, it is useful to look back on past information, such as the text of remarks or lip images, when users have failed to understand some speech.

R4: Information selection by D/HH people. Although providing varied information is an important requirement, as described in **R2**, there are some cases in which showing too much information can hinder understanding. Effective information for understanding strongly depends on each individual D/HH person. Thus, this application must enable users to select visualization modes.

3. 2 Characteristics

We propose MIETA, which has the following four characteristics, each characteristic corresponds to a requirement described in Section 3. 1.

C1: Available on various devices. This system is deployed as a web application. Thus, MIETA can be used from any device with a Web browser, such as smartphones or laptops. Additionally, MIETA is free of cost and its source code is publicly available. MIETA is designed so that its function can be further developed using its freely available API.

C2: Provides multimodal visualization. This system utilizes multiple visualization modes and provides a variety of information about speech by hearing people for D/HH people to understand. A visualization mode is defined as a unit component of MIETA that visualizes an aspect of speech information. Figure 1 illustrates the MIETA graphical user interface (GUI). The four examples of visual-

ization modes in Figure 1 are:

Speech-to-Text (STT) mode: shows the conversation itself as text using a voice recognition engine, as in Figure 1(b).

Face mode: shows speakers' faces using their devices' cameras and assists in reading lips and face expressions, as in Figure 1(a).

Image search mode: easily shows images about a designated word to abstractly illustrate its meaning, as in Figure 1(c).

Wikipedia search mode: easily quotes a description of difficult words, such as technical terms, extracted from Wikipedia, as in Figure 1(d).

MIETA is designed to accept new visualization modes easily as plugins. For instance, a sign language interpretation mode could be added in the future when an automated sign language interpretation algorithm is designed. We describe more detailed ideas for addressing this evolution in Section 4.

C3: Stores conversation information. Both raw (e.g., voice and face images) and processed (e.g., captions and detected lip movement) data are stored in a database. Users can look back on past conversation information when they forget or are initially unable to understand some speech.

C4: Selectable information. Users can select which information they use to comprehend the conversation from various visualization modes described in **C2** and customize how the modes are displayed.

3.3 Usage and Data Flow

This application is intended for situations such as a meeting in a company attended by multiple hearing and D/HH people with their own laptops or smartphones. Figure 2 shows the dataflow of MIETA and includes the four characteristics specified in Section 3.2. At first, the two hearing people use their devices as a client of the system. Their voice and face image data are stored in a database via these devices through the system's API. Stored data are processed to show visualization modes selected by the D/HH user. This figure includes various examples of visualization modes, such as STT mode and lip-reading mode. Finally, the processed and visualized information are provided to the D/HH user.

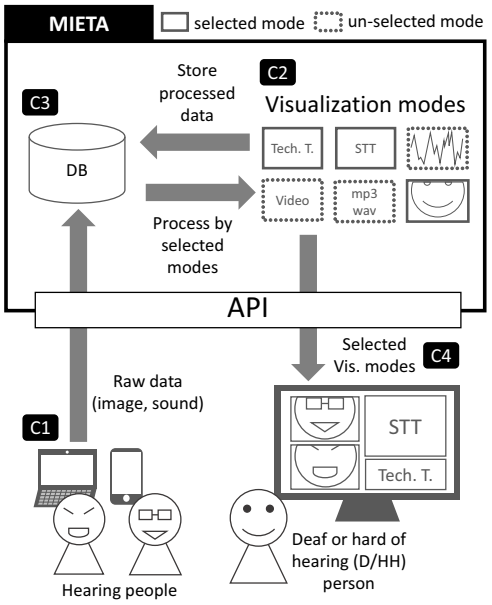


Fig. 2 Data flow of MIETA

4 Extensible Application Design

4.1 Overview

In this section, we design the technical aspects of MIETA, considering the application's prospects for extension over the long term. A rich selection of visualization modes enables MIETA to deal with the diversity of demand based on the preferences and skills of D/HH users. Therefore, it is very important to support the development of visualization modes. Making it easy to develop new visualization modes and encouraging third parties to develop them should lead to more flexible and effective information support in the long term. For example, when an algorithm automatically converting sentences into sign language is invented in the future, a visualization mode that utilizes it can be developed easily.

4.2 Simplification by Inversion of Control

It is essential to define a visualization mode template, to make it easy to develop various visualization modes and encourage third parties to do so. We extract common process for visualization modes by applying inversion of control. In MIETA, the framework controls all visualization modes de-

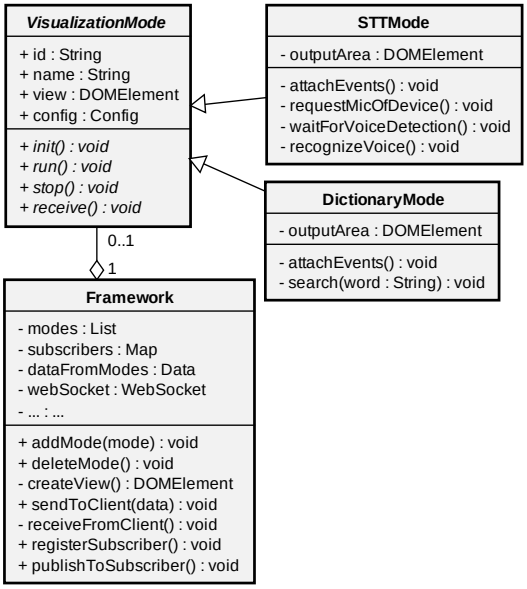


Fig. 3 Abstract and concrete classes of the visualization modes and the framework

veloped by third parties. Common properties and methods in visualization modes are defined as an abstract visualization mode class. All the developers must do is implement the abstract class. They basically do not need to consider how the methods of their visualization mode are called.

Figure 3 illustrates the relationship between abstract and concrete classes of the visualization modes and the framework. For example, when developing STT mode, the developer first implements unique properties and methods to STT mode, such as a request to access the device microphone (`requestMicDevice()`) and processing after voice recognition (`waitForVoiceDetection()`). Then, these functions are called from prepared methods, such as `init()` for initiation processing or `run()` for main processing. Since developers do not have to pay attention to how common methods in the visualization modes are called, they can concentrate on work essential to the visualization. The framework class has common methods related to visualization modes. For example, it contains the methods such as `addMode()` and `deleteMode()`, which add or delete a visualization mode from the display and are called from an operation by users. Furthermore, it also has functions related to communication between clients. Thus, developers can easily implement communication by calling prepared functions, such as `Framework.sendToClient()`, from code for each visualization mode.

4.3 Real-time Communication between Clients with WebSocket

MIETA requires strong real-time communication between clients to successfully support understanding of speech content because conversations with multiple hearing people progress quickly. Furthermore, if a method of realizing communication with clients differs with each visualization mode, the overall system can easily become complicated.

MIETA realizes real-time communication by utilizing **WebSocket**. Developers can easily implement communication with WebSocket by calling `Framework.sendToClients()` to send data. Visualization modes can receive data from other clients in the `receive()` method of each mode, which is called when the framework receives data to the mode.

4.4 Sharing Data between Visualization Modes using Pub/Sub

To combine the functions of visualization modes and realize advanced visualization modes, it is essential for visualization modes to share data, avoid inefficient development, such as a case in which multiple visualization modes each try to obtain the same value in their own ways.

MIETA adopts the **Pub/Sub messaging model** for visualization modes sharing information. The framework unifies the management of all data from modes. When the framework receives data from a visualization mode (publisher), it automatically sends these data to other visualization modes (subscribers). Therefore, to reuse data, developers need only specify which data to provide or obtain.

5 Development

This section describes the development of a MIETA[†2] prototype, which has the partial functions proposed above. The prototype has four visualization modes: STT, Face, Image Search, and Wikipedia Search modes, as described in C2 of Sec-

[†2] https://github.com/usk108/mieta

Table 1 Source code metrics for each component

Component name	Type	LLOC[†6]	CC[†7]
STT mode	mode	144	5
Face mode for observer	mode	88	1
Face mode for speaker	mode	104	2
Wikipedia search mode	mode	82	1
Image search mode	mode	92	1
WebSocket commun	frwk	26	1
Common mode logic	frwk	418	47

tion 3. 2. STT mode has history function to partially realize C3 and the ability to show two speakers' faces. In this prototype, Face mode supports only two speakers owing to screen size constraints on typical laptops. This limitation can be solved by dynamically changing the size of each face, with focus placed on the current speaker. The prototype is designed with IoC simplification (Section 4. 2) and WebSocket communication between clients (Section 4. 3). These are minimum functions to evaluate the concept of multimodal speech visualization in the evaluation experiment described in Section 6. The development environment is as follows.

- **Programming Language**: Java, JavaScript, HTML, CSS
- **Server**: Apache HTTP Server, Apache Tomcat, Node.js
- **STT mode**: Web Speech API[†3]
- **Wikipedia mode**: MediaWikiAPI[†4]
- **Image Search mode**: Custom Search API[†5]

Table 1 shows source code metrics for each component of Mieta. Information includes the type of code (visualization mode or framework), the content of the code, LLOC[†6], and CC[†7].

6 Evaluation

6. 1 Overview

The goal of this evaluation is to investigate Mieta's applicability, and the applicability of the concept of multimodal speech visualization, as information support for D/HH people in actual meetings with hearing people. The research questions we aim to answer are:

RQ$_1$: *How helpful is* Mieta *for D/HH people in understanding natural conversation with hearing people?*

RQ$_2$: *What strengths and weaknesses do D/HH people find in* Mieta*?*

More detailed situation where the two RQs assume is real-time and face-to-face conversation with two speakers. This situation excludes online, one-to-many or unindirectional situation such as classroom or seminar. The term "helpful in understanding" used in **RQ**$_1$ represents the improvement of degree of understanding by the use of Mieta compared with sign language interpreter. This understanding includes not only linguistic context but also non-linguistic context such as face expression and gestures. **RQ**$_2$ is aimed to confirm the qualitative effects of Mieta which cannot be confirmed by **RQ**$_1$. The improvement of understanding can easily measured by conducting some quizzes. However, positive or negative factors affecting on the comprehension process are hard to be measured because of the Mieta's multiple modalities. These factors are qualitatively confirmed by conducting questionnaires. The following three hypotheses are corresponded to C2 to C4 described in Section 3. 2. C1 is excluded because C1 is one of a non-functional feature.

- Multiple visualization helps understanding in a complementary manner
- Visualizing speech history is effective on real-time conversation
- Effective visualization modes vary by individuals

To answer **RQ**$_1$, we conduct a control experiment to compare the understanding of Japanese conversation under three conditions: with Mieta, with sign language interpretation, and with neither tools nor support as a baseline. Then, to answer **RQ**$_2$, we conduct a survey asking users how they feel about using Mieta after the control experiment.

6. 2 Evaluation Design

6. 2. 1 Subjects and Actors

The subjects in this evaluation are nine students (18–20 years old) from the Special Needs Education

[†3] https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API

[†4] https://www.mediawiki.org/wiki/API:Main_page

[†5] https://developers.google.com/custom-search/

[†6] Logical Lines of Code (LLOC) is the number of lines of code excluding blanks and comments.

[†7] McCabe's Cyclomatic Complexity (CC) indicates the complexity of a program.

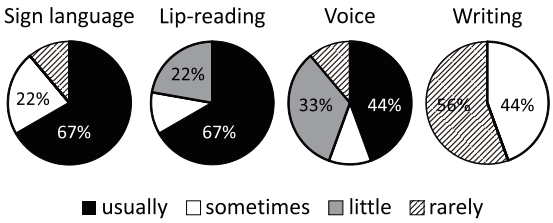


Fig. 4 Means of daily communication for subjects

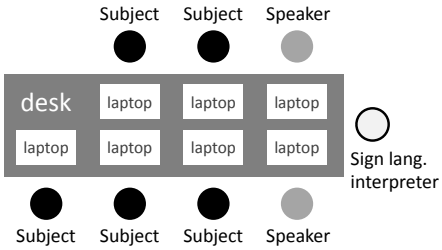


Fig. 5 Seating chart in the experiment

School for the Deaf at the University of Tsukuba. All of them wear hearing instruments or cochlear implants. Their degrees of hearing impairment, which is defined by the Japan Ministry of Health, Labour and Welfare[†8], were two to six.

Figure 4 shows their means of daily communication. Many of the subjects use sign language and lip reading, which shows that they are well-trained. We divided the subjects into two groups (G1 and G2) with four and five members so as to minimize the difference of communication skills between the two groups.

Two hearing people in our research group act as speakers. Speakers consciously speak slowly and smoothly, though not so much as to be unnatural, so that it was easy for the subjects to grasp the conversation. This replicates usual efforts of hearing people in communication with D/HH people.

A teacher from the Special Needs Education School for the Deaf, University of Tsukuba acts as a sign language interpreter. He prepare for translation with given scripts beforehand because he is not a professional sign language interpreter.

Figure 5 shows a seating chart for one group in all conditions. Four or five subjects are seated next to two speakers. In this seating layout, each subject cannot always see the speakers' lips and face, especially the leftmost subject. In addition, the two speakers faced each other because they focused on their conversation during all experiments. This experimental setting is intended to confirm the effects of Mieta in a common situation. When accompanied by a sign language interpreter, he was located between the two speakers, to be visible to all subjects.

6.2.2 Compared conditions

This control experiment aims mainly to compare Mieta and sign language interpretation, because sign language interpretation is common current information support in company meetings. A condition with no support was for practice and to evaluate the basic communication skill of the subjects. The details of the three conditions are as follows:

Using Mieta: Subjects try to understand the conversation with the Mieta prototype described in Section 5. To simplify analysis, Mieta always displays all four visualization modes, which means that subjects do not choose which modes to use. In this experiment, they look at only the display of Mieta but not speakers directly. To control the instability of speech recognition, prerecorded recognition results are used. The prerecording was conducted near a microphone in a silent room, with clear and smooth phonation. Each recognition result is published by the speaker pressing a button in exact timing with his or her speech. Recognition delays between speech and output were also reproduced.

With a sign language interpreter: Subjects try to understand the conversation with a sign language interpreter. Subjects can look at either the interpreter or speakers, or both, as they typically would in daily conversation because we want to compare their daily ways with Mieta.

No support: Subjects try to understand the conversation with neither tools nor support, except for their personal hearing instruments. They can use their preferred methods of understanding speech, such as lip reading.

[†8] http://www.mhlw.go.jp/bunya/shougaihoken/shougaishatechou/dl/toukyu.pdf

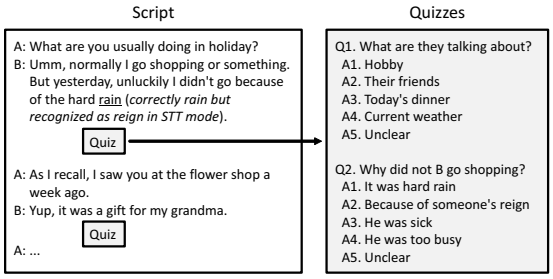


Fig. 6 Example of a script and quizzes

6.2.3 Scripts and Quizzes

Prepared scripts consists of three kind of scripts (S0, S1, and S2). S0 is for the condition with no support. It consists of 16 sentences and has 6 quiz questions. S1 and S2 are for comparing Mieta and sign language interpretation. Each consists of about 25 sentences and has 12 quiz questions. The difficulty of the questions were equalized by ensuring the same number of obstacles, whose variety is described in the next paragraph. The left side of Figure 6 shows an example script.

Figure 6 shows two quiz questions provided relating to some of the script sentences. Each question has five answer choices. Since it is necessary to distinguish having no idea about which answer is correct from misunderstanding, one of the choices is "no idea." The rest of the choices are three wrong answers and one correct answer. Quiz questions are of the following four types: (i) simple question without any obvious obstacles, (ii) question related to sentences when two speakers were talking simultaneously, (iii) question related to technical terms (e.g., "brainstorming") used in the conversation, (iv) questions related to a sentence that includes a voice recognition error. Q1 and Q2 in Figure 6 correspond to quiz types (i) and (iv). All scripts, except for type (iii), only contain common and standard terms. There are no recognition errors in the prerecorded recognition results for types (i) to (iii), to isolate the effects of recognition errors. A single recognition error is consciously injected into the prerecorded results for for each conversation related to a type (iv) quiz question. An example error is illustrated in Figure6. The word "rain" is misrecognized as "reign" and the content of Q2 corresponds to the error.

While answering to quizzes, subjects can search for information with the Image Search or Wikipedia modes. Each script contains quiz question types (i), (ii), (iii), and (iv) in the ratio 3:1:1:1.

Table 2 Combination of groups, scripts and conditions

Script	Group G1	Group G2
S0 (practice)	No support	No support
S1	Sign language	Mieta
S2	Mieta	Sign language

6.2.4 Controlling Experimental Factors

Table 2 shows the combination of two groups, three scripts, and three conditions. The experiment is conducted from S0 to S2 in order. This control experiment aims mainly to compare Mieta and sign language interpretation. We should therefore exclude habituation effects and effects from unintended differences in difficulty between S1 and S2. Therefore, G1 is accompanied by a sign language interpreter before using Mieta and the order is reversed for G2.

6.3 Experiment Result

6.3.1 Quiz Result

Figure 7 shows the degree of understanding for the four question types across the three conditions. We conducted a Wilcoxon signed-rank test about the rate of correct answers and other answers, which consist both of wrong answers and "no idea" answers. The "*" symbol indicates a statistically significant difference at a significance level of 0.05. The "**" symbol indicates a further difference at a significance level of 0.01.In short, this figure illustrates that Mieta contributed to understanding in question types (i), (ii) and (iii), while it increased misunderstanding in questions of type (iv). We describe the result of each of the four quiz types below.

For questions with no obstacles (type (i)), subjects scored high with both Mieta and a sign language interpreter, with no statistical difference between the two.

For type (ii) questions, when two speakers talk simultaneously, subjects got the highest score with Mieta. This situation is hard for a sign language interpreter to handle because he must quickly interpret two sentences while showing who has said

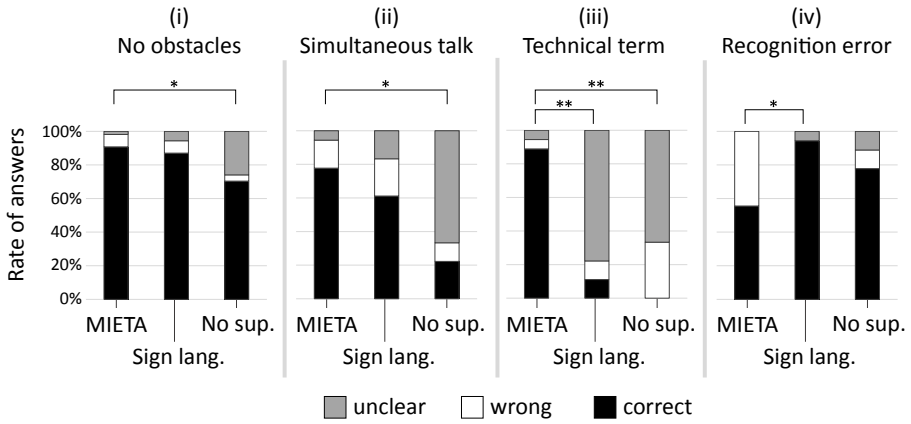


Fig. 7 Degree of understanding for 4 types of quizzes in 3 conditions

each. However, with MIETA, it was easy to understand what was said from STT mode and who was talking from Face mode.

For type (iii) questions, about technical terms in the conversation, subjects got outstanding high scores with MIETA. Technical terms were not previously known by the subjects. This is also difficult for a sign language interpreter because when he does not know a word, he must show each character of the word through sign language and lip movement one-by-one. Even if he knows the word, there is not enough time in natural conversation to describe complex meanings. With MIETA, subjects quickly searched for the meaning and answered quiz questions correctly. Wilcoxon signed-rank test showed that MIETA significantly ($p < 0.01$) contributed to understanding compared with sign language interpretation.

However, there were many wrong and no "no idea" answers to type (iv) questions, which include recognition errors, with MIETA. From statistical tests, MIETA significantly ($p < 0.05$) decreased the correct answer rate compared with sign language interpretation. Subjects were aware that there were recognition errors before starting the experiment. However, subjects misunderstood the conversation easily because they could not know whether a recognition result was correct. Sentences with a recognition error were sometimes still meaningful. For example, when there are two opinions in conversation, even if "I agree with *this* opinion" is recognized with an error to "I agree with *that* opinion", people who see the recognition result should not feel that something is wrong. In addition, current voice recognition algorithms try to make natural sentences. Thus, more than half of subjects misunderstood, while 66% answered that they were "not at all" or "little" disturbed by recognition errors in the questionnaire.

6. 3. 2 Questionnaire Results

We provide some examples of actual answers for each question in the questionnaire after the control experiment. Figure 8 illustrates subjects' feelings about their understanding in the three conditions.

When asked about the **advantages of MIETA**, some wrote that *"We can read face expressions, search for meaning in Wikipedia at that time and refer image search mode for support of grasping.* MIETA *visualized conversation comprehensively"* ($A_{1.1}$), *"*MIETA *visualized not only voice as text, enable us to search for many things when I cannot understand. It's convenient. My parents are also D/HH people, so I want to introduce it"* ($A_{1.2}$), and *"*MIETA *could be referred as conversation history when I failed to hear speech or I did not understand what were said at that time"* ($A_{1.3}$).

As for the **weaknesses of MIETA**, they pointed out that *"I think* MIETA *should add Hiragana mode"* ($A_{2.1}$), *"Speech is not recognized rapidly. So, I think typing is better for some people instead of talking."* ($A_{2.2}$), and *"Recognition result appeared slowly, so it was little difficult to understand"* ($A_{2.3}$).

When asked about the **advantages and weak-**

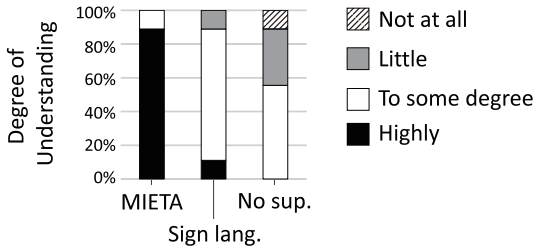


Fig. 8 Degree of overall understanding in 3 conditions as feeling of subjects

nesses of each visualization mode, the comments demonstrate the diversity of demand. For example, regarding Face mode, a subject said *"I hope that Face mode shows who is speaking more apparently, for example, getting the border blinking"* ($A_{3\cdot1}$). Most subjects mentioned Wikipedia mode as a strength: *"When I find unknown word in STT, I can immediately search the meaning of it by just highlighting the word"* ($A_{3\cdot2}$). Conversely, for the weakness of the mode, four of the answers addressed the level of the mode's function, *"existing function is enough"*($A_{3\cdot3}$) while another subject said *"I want to expand the area of Wikipedia mode"* ($A_{3\cdot4}$).

When asked **which modes they want to use next**, while six answered all four visualization modes, two preferred to use only three modes (STT/Face/Wikipedia or STT/Face/Image Search) and one wanted to use only STT mode ($A_{4\cdot1}$).

When asked to **comment about MIETA freely**, some subjects wrote that, *"When* Mieta *become available, I want to use it and it will be a kind of innovation"* ($A_{5\cdot1}$), *"It will help D/HH people. I'd love to want it spread"* ($A_{5\cdot2}$), and *"I was very happy to hear that MIETA can be used for free. I want to use it"* ($A_{5\cdot3}$).

In summary, two of three hypotheses corresponding to C2 (multiple modalities) and C4 (individuality) are confirmed from $A_{1\cdot1}$, $A_{1\cdot2}$ and $A_{4\cdot1}$. We cannot confirmed one hypothesis for C3 (looking back) from the experiment.

7 Discussion and Future Perspective

7.1 Analysis of Result

According to the quiz results in Section 6.3.1, the answer to $\mathbf{RQ}_1$ is that Mieta contributed to D/HH people's understanding of conversation; however, it increased misunderstanding when there are recognition errors.

The result of the questionnaire in Section 6.3.2 suggested that the answer to $\mathbf{RQ}_2$ is that subjects found the effectiveness of Mieta's characteristics to be a strength for information support; however, some found the response speed of STT mode to be a weakness. Figure 8 showed that most subjects felt that Mieta strongly contributed their understanding. Additionally, comments like $A_{1\cdot1}$ and $A_{1\cdot2}$ indicated that they welcome Mieta.

We found comments proved that Mieta's characteristics (C1–C4 in Section 3.2) can satisfy the actual demands of D/HH people. Regarding C1 (initial cost), most subjects, including $A_{5\cdot3}$, said *"I want to use* Mieta*"* because it is free of cost. As for C2 (multimodal speech visualization), many comments (such as $A_{1\cdot1}$, $A_{1\cdot2}$, and $A_{3\cdot2}$) supported the concept and suggested that visualization modes helped understanding in a complementary manner. For C3 (past information), although the history function of STT is only partially developed, comments such as $A_{1\cdot3}$ proved the effectiveness of looking back on conversations. Regarding C4 (free choice of visualization modes), the result suggested that free choice by users can lead to effective personalized information support because there is preferences of visualization modes by users, shown in $A_{4\cdot1}$, and different ways to use each mode, shown in $A_{3\cdot1}$, $A_{3\cdot3}$, and $A_{3\cdot4}$.

7.2 Challenges in STT mode

The quiz result and comments $A_{2\cdot2}$ and $A_{2\cdot3}$ in the questionnaire indicate the quality of information support by Mieta strongly depends on STT mode. We discovered some challenges in STT mode as a result of this experiment.

Mieta's effectiveness, which was confirmed by the experiment, is a potential threat to the validity of the experiment design. While we controlled the occurrence frequency of recognition errors, more errors may possibly occur in natural conversation. Especially, since we isolated and evaluated the effects of recognition error and simultaneous talk, the recognition error may have co-occurence relation with simultaneous talk. Hinton et al. have reported that the accuracy of deep neural network

based recognition methods reaches about 80% at most [8]. We need to confirm the effectiveness of Mieta in more natural situations.

Per $A_{2.2}$ and $A_{2.3}$, some subjects felt that the voice recognition response speed was poor. This is an obvious limitation of the current technology that may require some compromise by hearing people, for example, by setting speech intervals to some degree.

Finally, there is a high possibility of causing misunderstandings with voice recognition errors, as shown in Figure 7. A subject wrote $A_{2.1}$ to request *Hiragana mode*. Hiragana are a type of Japanese character, each of which correspond one-to-one to a phoneme and require no word conversion. Japanese people can understand content from only Hiragana to some extent, though the information can be vague. In other words, his comment means that Mieta should have STT mode without word conversion because there is possibility to make wrong conversions.

Here, we consider an improvement idea of Mieta about the recognition error. Firstly, it can be correcting recognition errors by participants in conversation since voice recognition technology cannot perform at 100% accuracy rate. Clearly, correcting words requires less effort than writing total conversation. Secondly, it can be alerting with `confidence value` of voice recognition. In general, the response from voice recognition API has the value, which shows a conviction degree of the recognition. Therefore, for example, showing the result highlighted as red when the value is less than a particular number is expected to be quick help for D/HH people so as not to misunderstand.

8 Related Works

LiveTalk [7] and UDTalk [17] are commercial products for information support in daily meetings. They display text converted from the speakers' voices using ASR. However, they essentially provide only text and no other situational information. Furthermore, they cost money for organizational usage. The novelty of Mieta is both of information support with various information including captions about speech and basically free of cost.

Some studies have previously been carried out on the topic of information support for D/HH people, including a real-time captioning system [11] [13]. However, these studies are limited to investigating the applicability of the proposed systems, and do not present practical applications. Furthermore, existing studies largely focus on conversations with a single speaker, such as a lecture. The novelty of Mieta is publishing a practical application and targeting conversation with multiple speakers such as a meeting.

Recently, multimodal approaches have been applied to various purposes. Studies have applied multimodal interfaces to implement natural and effective communication systems [12] [14] [20]. In addition, many studies in medical fields utilize multimodal visualization to visualize complex human anatomy [2] [4] [6]. As these studies show, multimodal approaches are well suited for decomposing something complex into a combination of primitive components that are easier for humans to handle. We applied the idea of multi-modal to the problem for D/HH people to grasp conversation and proposed a new type of sensory substitution with combination of visual components about conversation.

9 Conclusion

In this paper, we proposed Mieta, which is a multimodal speech visualization application. It displays information about speech to support D/HH people in understanding conversations with multiple speakers. Additionally, we conducted a control experiment that evaluated Mieta. We found that Mieta contributed to D/HH people's understanding of conversations in real-time and face-to-face situation with two speakers; however, it also increased misunderstandings when recognition errors occurred. Subjects of the experiment found that Mieta is effective for information support; however, some found the response speed of the STT mode to be too slow.

As future work, for effective information support, it is necessary to tackle challenges found from the experiment as in Section 7. 2 and improve Mieta. Since this paper only focuses *assisting understanding*, we need to tackle *assisting statement* to express their opinions in face-to-face conversations.

Acknowledgments This research was partially supported by Grants-in-Aid for Scientific Research (No.16H02908, No.15H02701, No.26280115, No.26730155, 15K12020).

References

[1] Bach-y-Rita, P.: Sensory substitution and qualia, *Vision and mind*, (2002), pp. 497–514.

[2] Baum, K. G., Helguera, M. and Krol, A.: Fusion viewer: a new tool for fusion and visualization of multimodal medical data sets, *J. Digital Imaging*, Vol. 21, No. 1(2008), pp. 59–68.

[3] Brookes, C.: Speech-to-text systems for deaf, deafened and hard-of-hearing people, in *IEE Seminar on Speech and Language Processing for Disabled and Elderly People*, 2000, 5/1–5/4.

[4] Burns, M., Haidacher, M., Wein, W., et al.: Feature Emphasis and Contextual Cutaways for Multimodal Medical Visualization, *EuroVis*, Vol. 7(2007), pp. 275–282.

[5] Department of Health and Welfare for Persons with Disabilities, Minister's Secretariat, Ministry of Health, Labour and Welfare: Overview result of survey on persons with physical disability (in Japanese), http://www1.mhlw.go.jp/toukei/h8sinsyou_9/, 1999. visited on 2016-10-29.

[6] Frank, R., Damasio, H. and Grabowski, T. J.: Brainvox: An Interactive, Multimodal Visualization and Analysis System for Neuroanatomical Imaging, *J. Neuroimage*, Vol. 5, No. 1(1997), pp. 13–30.

[7] Fujitsu: FUJITSU Software LiveTalk (in Japanese), http://www.fujitsu.com/jp/group/ssl/products/software/applications/ud/livetalk/. (visited on 2016-10-29).

[8] Hinton, G., Deng, L., Yu, D., et al.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, *IEEE Signal Processing Magazine*, Vol. 29, No. 6(2012), pp. 82–97.

[9] Junqua, J.-C. and Haton, J.-P.: *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, Vol. 341, Springer Science & Business Media, 2012.

[10] Kurita, S.: Development of IPtalk and PC captioning (in Japanese), *J. Information Processing and Management*, Vol. 59, No. 6(2016), pp. 366–376.

[11] Kuroki, H., Ino, S., Nakano, S., et al.: A Method for Determining the Timing of Displaying the Speaker's Face and Captions for a Real-Time Speech-to-Caption System, *SICE J. Control, Measurement, and System Integration*, Vol. 3, No. 6(2010), pp. 402–408.

[12] Marin, R., Sanz, P. J., Nebot, P. and Wirz, R.: A Multimodal Interface to Control a Robot Arm Via the Web: A Case Study on Remote Programming, *Tran. Industrial Electronics*, Vol. 52, No. 6(2005), pp. 1506–1520.

[13] Miyoshi, S., Ishihara, Y., Nishikawa, S. and Kobayashi, M.: Real-time Captioning System Using Voice Recognition Technique for Hearing Impaired Persons at Lecture, *IEICE Tech. Rep. Education technology*, Vol. 103, No. 600(2004), pp. 15–18.

[14] Perzanowski, D., Schultz, A. C., Adams, W., Marsh, E. and Bugajska, M.: Building a Multimodal Human-Robot Interface, *IEEE Intelligent Systems*, Vol. 16, No. 1(2001), pp. 16–21.

[15] Punch, R., Hyde, M. and Power, D.: Career and Workplace Experiences of Australian University Graduates Who are Deaf or Hard of Hearing, *J. Deaf Studies and Deaf Education*, Vol. 12, No. 4(2007), pp. 504–517.

[16] Sakamoto, N.: Education and Employment of Deaf and Hard of Hearing People (in Japanese), *Report Issued by Research Center for Ars Vivendi*, (2011), pp. 14–30.

[17] studist corporation: UDTalk Online Manual, https://teachme.jp/r/udtalk_en.

[18] Toba, Y., Horiuchi, H., Matsumoto, S., Saiki, S., Nakamura, M., Uchino, T., Yokoyama, T. and Takebayashi, Y.: Considering Multi-Modal Speech Visualization for Deaf and Hard of Hearing People, in *Asia-Pacific Symposium on Information and Telecommunication Technologies*, 2015.

[19] Toba, Y., Matsumoto, S., Saiki, S., Nakamura, M. and Uchino, T.: Evaluating Multi-Modal Speech Visualization Application for Deaf and Hard of Hearing People, in *Applied Computing & Information Technology*, 2016.

[20] Vo, M. T. and Wood, C.: Building an Application Framework for Speech and Pen Input Integration in Multimodal Learning Interfaces, *Int'l Conf. Acoustics, Speech, and Signal Processing*, Vol. 6(1996), pp. 3545–3548.

[21] Weisel, A. and Cinamon, R. G.: Hearing, Deaf, and Hard-of-Hearing Israeli Adolescents' Evaluations of Deaf Men and Deaf Women's Occupational Competence, *J. Deaf Studies and Deaf Education*, Vol. 10, No. 4(2005), pp. 376–389.

[22] Xu, W., Lifang, X., Dan, Y. and Zhiyan, H.: Speech Visualization based on Locally Linear Embedding (LLE) for the Hearing Impaired, *Int'l Conf. BioMedical Engineering and Informatics*, Vol. 2(2008), pp. 502–505.

[23] Yao, K., Yu, D., Seide, F., et al.: Adaptation of Context-Dependent Deep Neural Networks for Automatic Speech Recognition, in *Workshop on Spoken Language Technology* (*SLT*), 2012, pp. 366–369.

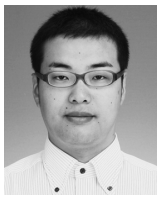
[24] Zshorn, A., Littlefield, J. S., Broughton, M., et al.: Transcription of Multiple Speakers Using Speaker Dependent Speech Recognition, Technical report, Australian Government Department of Defence Technical Report DSTO-TR-1498, 2003.

鳥羽祐輔

2015 年神戸大学工学部情報知能工学科卒業．2017 年神戸大学大学院システム情報学研究科計算科学専攻博士課程前期課程修了．IT を用いた聴覚障害者のコミュニケーション支援などの研究に従事．

松本真佑

2010 年奈良先端科学技術大学院大学博士後期課程修了．同年神戸大学大学院システム情報学研究科特命助教．2016 年より大阪大学大学院情報科学研究科助教．博士 (工学)．エンピリカルソフトウェア工学の研究に従事．

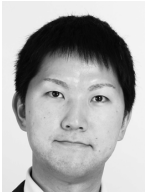


佐伯幸郎

2009 年高知工科大学大学院工学研究科基盤工学専攻博士後期課程修了．同年同大学情報システム工学科助手．2010 年同大学情報学群助教．その後 2013 年神戸大学システム情報学研究科特命助教を経て 2016 年同大学先端融合研究環特命助教．博士 (工学)．ディジタル信号処理，クラウドコンピューティング，ソフトウェア工学教育の研究に従事．IEEE，電子情報通信学会各会員．

中村匡秀

1994 年大阪大学基礎工情報卒．1999 年同大学大学院博士後期課程了．2000 年同大学サイバーメディアセンター助手．2002 年奈良先端科学技術大学院大学情報科学研究科助手．2007 年神戸大学大学院工学研究科准教授．2010 年同大学院システム情報学研究科准教授．2015 年フランス・グルノーブル大学・在外研究員．博士 (工学)．サービス・クラウドコンピューティング，ソフトウェア工学，スマートホーム，スマートシティ，ジェロンテクノロジーの研究に従事．IEEE，ACM，情報処理学会各会員．



内野智仁

2008 年東京工業大学社会理工学研究科人間行動システム専攻修士課程修了．同年より東京福祉大学社会福祉学部専任講師を経て，2013 年より筑波大学附属聴覚特別支援学校教諭 (現職)．2016 年より埼玉大学工学部非常勤講師を兼任．修士 (工学)．聴覚障害の特性を踏まえたデジタル教材開発，情報モラル教育，プログラミング教育などの研究に従事．

横山知弘

1989 年より筑波大学附属聴覚特別支援学校教諭 (現職)．聴覚障害教育に関する研究に従事．

武林靖浩

1990 年東京都立綾瀬ろう学校教諭．その後，2000 年より筑波大学附属聴覚特別支援学校教諭 (現職)．聴覚障害教育に関する研究に従事．