Technology Reports CSR Speech Recognition Telephone Speech-voice to Text Conversion

Development and Testing of the “Mieru Denwa” Service Supporting Users with Hearing Impairment

DOCOMO Technology, Inc. Packet Networks Division Kazue Mikami† Takuya Shinozaki
Service Design Department Junsuke Morita

In an effort to provide products and services that are accessible to everyone from a CSR perspective, NTT DOCOMO has proposed a service that is able to convert the speech-voice of a telephone call to text in real time, and display it on a smartphone screen for users that have hearing disabilities or are hard of hearing. We consulted many users with hearing impairments in development of this system and application, and conducted repeated hypothesis testing on the usability of the application and to tune the speech recognition engine based on a simple prototype.

1. Introduction

The Act for Eliminating Discrimination against Persons with Disabilities (enacted April 1, 2016) requires that in Japan, services in society function with reasonable consideration for people with disabilities, and there are over seven million persons in Japan, including the aged, that have difficulty hearing speech during telephone calls.

We have become an Internet-based society and Web services are common, but there are still many scenarios that can only be handled on the telephone, such as making inquiries or applying for services. This presents obstacles in the lives of those with hearing disabilities. In an actual survey of people with hearing disabilities, the most common response regarding difficulties due to hearing disability was “Situations that require using the telephone” (58.1%).

† Curently Solution Service Division

Particularly in emergencies, when having troubles with life-line services such as loss of a credit card, or plumbing problems, there are issues that cannot be resolved without speech-voice on the telephone, and this can be extremely troubling.

There are existing services for communication through an operator, but they are only available at limited times and they are costly, so they are not well used.

On the other hand, with the maturation of speech recognition technology, it has become possible to convert speech-voice to text in real time, so NTT DOCOMO has begun studying a service for users with hearing disabilities called "Mieru Denwa," which converts speech-voice to text and displays it in real time. In studying this service, we encountered the issue that the accuracy of speech recognition decreased because the people were not aware that speech recognition was being used. As such, we needed to check that the speech recognition was accurate enough to provide as a service that enables telephone communication for people with hearing disabilities. We conducted tests using a prototype Android™*1 application for users with hearing disabilities to verify the service concept, to conduct a user evaluation of the current state of recognition accuracy, and to select the minimum required functionality.

The results indicated that many users would be eager to use such a service if the recognition accuracy was improved. In fact, it was clear that they would use it more in scenarios speaking with someone they did not know, such as making an inquiry to a company, than with friends, family and other people they know.

Taking this into account, we attempted to expand functionality considering such scenarios, and to improve the recognition accuracy. We provided the resulting system as a trial service a second time and it was evaluated highly in terms of quality and performance as a service, attaining a level suitable for a commercial service.

This article describes the speech-voice to text conversion implementation used in the Mieru Denwa trial service, details of measures taken to improve the accuracy of speech recognition, and the system and application developed for providing it as a commercial service.

2. Trial Service

2.1 Overview

To measure the demand for and user satisfaction with speech-voice to text conversion, and to improve the accuracy of speech recognition, we began providing a Mieru Denwa trial service in October 2016.

An overview of the Mieru Denwa trial service is shown in **Figure 1**. The trial service design defined the following functional requirements [1].

(1) Real-time performance requirements

During a call, the system must recognize the speech-voice from the other party and display it as text on the screen of the service-user's smartphone in real time. The call status must also be displayed on the smartphone screen, showing the user when they can start speaking.

*1 Android™: A software platform for smartphones and tablets consisting of an operating system, middleware and major applications. A trademark or registered trademark of Google LLC., in the United States.

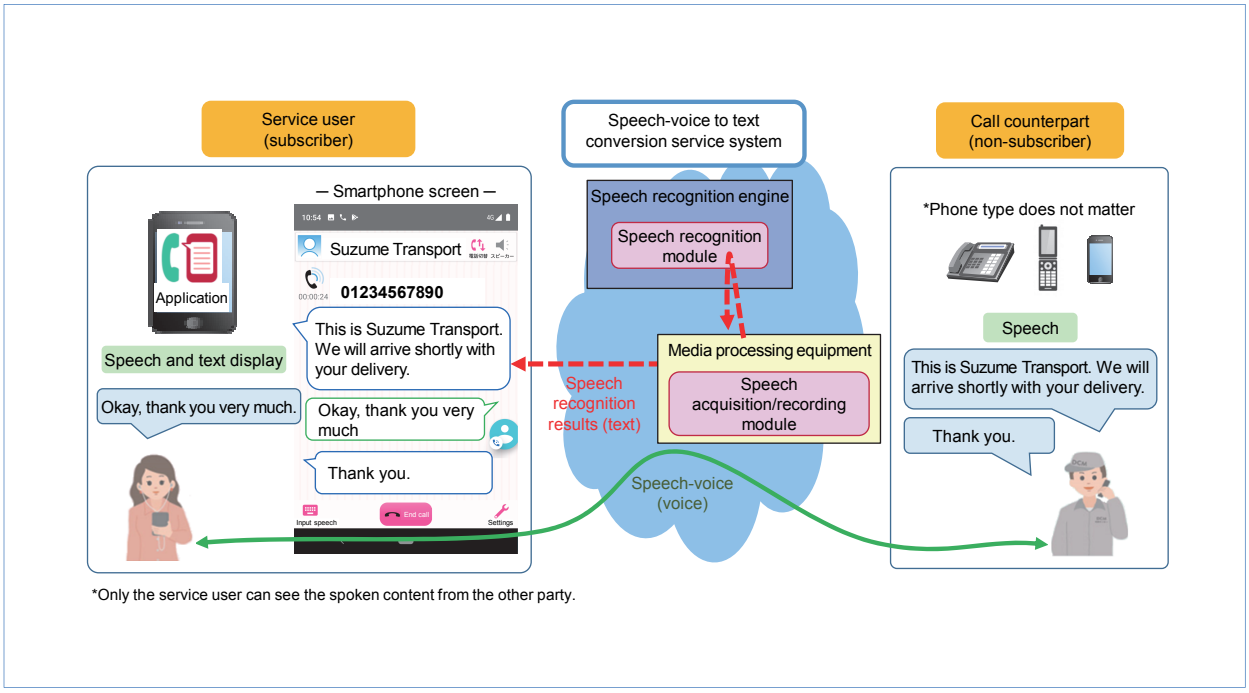

Figure 1 Overview of “Mieru Denwa” trial service

(2) Requirement for terminal independence

The service must be usable on a wide range of smartphone devices, regardless of OS. The other party must be able to use any device, as long as it is capable of speech-voice communication and must not require use of an application or have other requirements.

(3) Requirements for legal considerations

The service must have a mechanism to explain to the service user that speech-voice will be recorded and converted to text and that the recordings could be used to improve service performance, and must obtain agreement from the user. The other parties in calls must also be notified, giving consideration for privacy.

2.2 Implementation of the Speech-voice to Text Conversion Service

We adopted implementation as a network service to satisfy the requirements described above. The network service consists of media processing equipment able to record the speech-voice on the call path, transfer the recorded speech-voice to the speech recognition engine*2, and obtain the resulting text from the speech recognition engine. The trial system architecture is shown in **Figure 2**.

(1) Ensuring real-time performance

The media processing equipment was designed to detect silences in the speech, and at that point, perform speech recognition on the recorded speech-voice and display the result in real time.

*2 Speech recognition engine: Equipment that takes voice data as input, and converts it to text representing what was spoken.

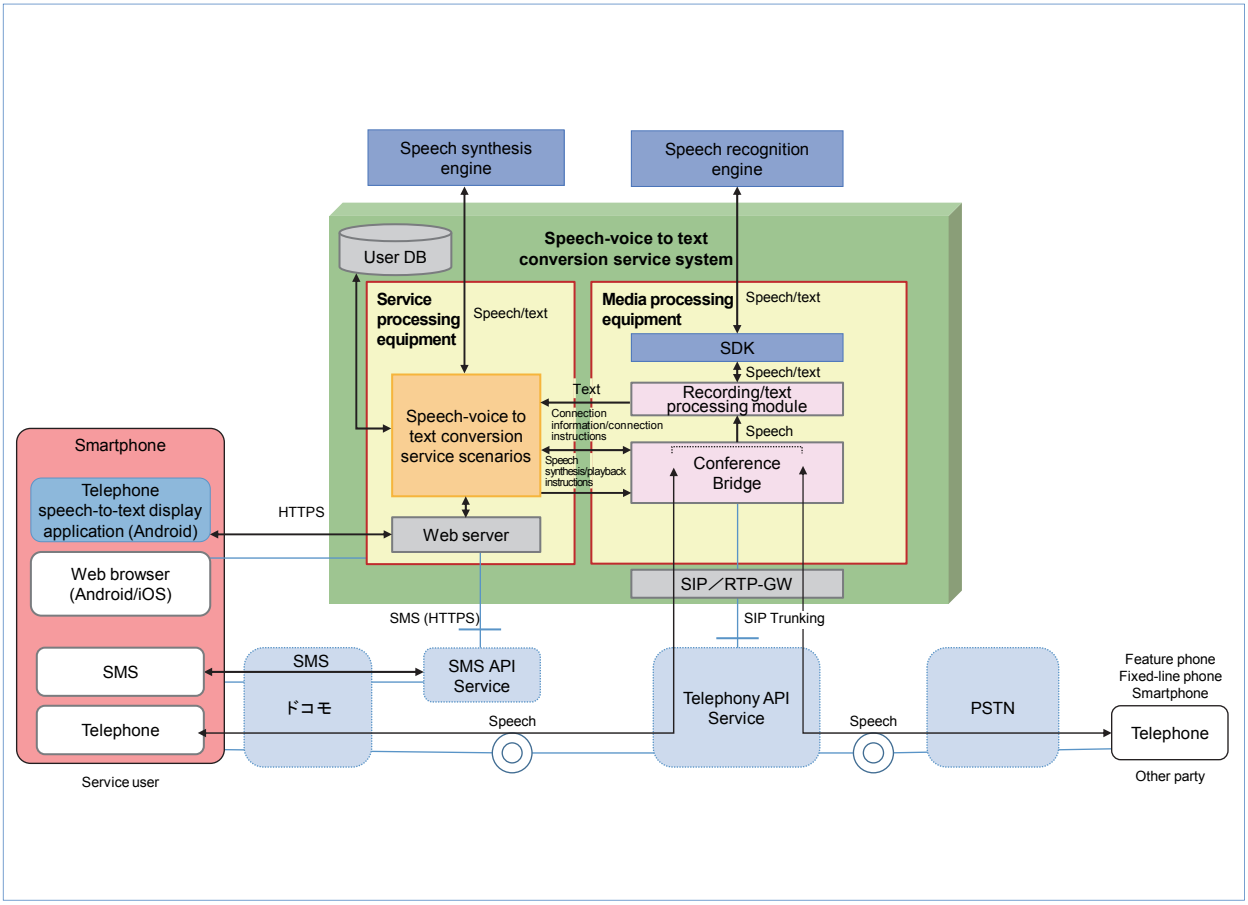

Figure 2　Trial system architecture

Note that the state of call is displayed on the smartphone screen so that the service user knows when playback of guidance has finished and speech can proceed.

(2) Ensuring independence from terminal type

With this architecture, call speech-voice is recorded on the call path, so there is no need for either terminal to have a speech-voice recording function, and terminals need only to support voice calls and the ability to display simple text. Specifically, we used the native telephone application on the terminal and developed a dedicated Android application to display the text of speech-voice.

For devices with a different OS, we built a Web application in the service processing equipment that enables the service function to be used on a standard browser screen and regardless of OS, without using a dedicated application.

This implementation enables the service to be used on any smartphone, regardless

of device type or OS.

Note that by using an architecture able to record speech-voice on the call path, the other party in the call can use any device capable of voice calls, without other requirements such as using an application.

(3) Implementing legal considerations

To obtain consent regarding the confidentiality of communication, the fact that call content will be converted to text is displayed on the service user’s application (or Web application) screen, and they must click an “Agree” button explicitly, each time the service is used.

To consider privacy for the other party, the media processor provides speech guidance indicating that the call will be recorded and converted to text when they accept the call, or when the service user initiates it manually.

2.3 Improving the Accuracy of Speech Recognition

1) Tuning Guidance to Improve the Accuracy of Speech Recognition

The usefulness of Mieru Denwa will vary greatly depending on the correctness of the text, in other words, the accuracy of speech recognition. It is technically difficult to achieve perfect recognition for all types of conversation, so we are currently working to increase the accuracy of speech recognition, defining improvement objectives within a practical scope according to usage scenarios. In particular, since our objective is to support those with hearing disabilities in situations where they have had difficulty, we are aiming for a level of performance that will enable the reader to decide what to do next, based on the results of speech recognition. We are working on two approaches to improve the accuracy of speech recognition. The first is to increase the probability that the speech will be easy to recognize by using guidance and usage scenarios, and the second is to tune the speech recognition engine using actual speech data.

Most of the scenarios in which Mieru Denwa will be used are when the telephone is needed to make inquiries to public institutions, retail businesses or companies. In such situations, speech recognition tends to be accurate because users make clear statements that are easy to recognize. The system also plays a message at the beginning of the call, requesting users to speak clearly because speech recognition is being used, to raise awareness of speaking clearly.

To tune the speech recognition engine, we periodically used the speech logs to improve speech recognition accuracy while we were offering the trial service. Speech collected with permission of users was analyzed, and frequently occurring words such as store names and words associated with the scenarios being used were registered in the dictionary of the speech recognition engine.

By training the speech recognition engine with uttered phrases, we optimized for usage scenarios, and as a result, we saw an improvement of almost 10% in the accuracy of recognized texts, compared with results at the beginning of the trial service. We expect that with the start of a commercial

service, the number of users will increase, allowing us to collect more speech samples, and take measures to improve the accuracy more effectively.

2) Improving the Continuous Recognition Method

Speech-voice requires continuous, real-time speech recognition. Initially, the equipment automatically started recording and speech recognition when the voice call started and then stopped when a silence in speech was detected. It then resumed after the detected silence [2]. In this case, there was a gap of unrecorded sound between detection of the silence and start of the next recording, but it was assumed that this missing section would only be a few tens of milliseconds and fall within the silence in speech-voice. However, when building and testing a real system, the gaps in recorded speech-voice lasted hundreds of milliseconds, causing the beginning of sentences to be clipped (not having been recorded), and resulting in poor speech recognition (speech was not recognized due to the clipping). To resolve this issue, we switched to a continuous process that does not silence recording and speech recognition after the call starts. However, it still detects and uses pauses to finalize conversion of the recorded and recognized speech to text. This approach avoided clipping of the beginning of sentences and improved speech recognition accuracy.

2.4 User Responses and Comments

1) Functional Improvements

During the service trial period, we conducted surveys of the users being monitored and continuously worked to improve the functionality, to develop a service and application suitable for use by people with hearing disabilities. Here we describe the input utterance function that we developed based on comments from many people with hearing or speaking difficulties.

(a) Input utterance function for words to be conveyed as speech

We implemented a text-to-speech function that starts a smartphone application (or Web application) to get text input of what the user wants to say during a telephone call and sends the text to system, which uses a speech synthesis engine to generate and play back the speech to the other party [3]. The synthesized speech is mixed with the spoken voice and sent to both parties, so that it can be transmitted even if the two audio streams overlap. To help conversation flow smoothly, we also incorporated some predefined phrases in the smartphone and Web applications so the service user can say the phrases with a simple tap of a button.

(b) User friendly interface, including for users with difficulty hearing or speech

To facilitate conversation, text from both sources are displayed together on the service user’s smartphone screen, which clarifies the sequence of statements from the other party and entered by the service user using the input utterance function. The applications also show the user exactly when playback of synthesized speech begins and ends, so that they can understand the timing of responses from the other party, and

so they can respond using input utterance at the right time.

In our implementation, we used a WebSocket[*3] for communication between the user's smartphone and the service processing equipment, and we added a mechanism to send signals from the service processing equipment to the smartphone when playback of synthesized speech starts and finishes. The input text is displayed on the screen when the signal indicating that playback of the synthesized speech has started is received, and the color of the bubble displaying the text changes when the signal indicating that playback has completed is received.

2) User Evaluation Survey

The objective of the trial service was to determine receptivity to the service concept and the level of user satisfaction with current speech recognition accuracy, so we conducted a user survey to check these aspects. Although there were still speech-recognition errors, the Mieru Denwa service enabled users to use voice call, which they previously could not or had given up on. Most users encouraged us to continue providing the service, so we decided to proceed with the commercial service.

3. Commercial Development

3.1 Overview

The trial service required use of a dedicated telephone number and had restrictions on services such as emergency and Free-dial®[*4] calling. With the commercial service, ordinary 090/080/070 numbers can be used, and emergency bulletins and other voice call services are supported (work to support emergency bulletins is still in progress). Also for the trial, the application was launched with an SMS[*5] notification, but this was changed to a push notification.

3.2 Service Implementation

1) System Development

The system architecture for the commercial Mieru Denwa service is shown in **Figure 3**. Voice call processing is implemented using a Service Enabler Network (SEN)[*6], which is a platform for executing service scenarios composed, in part, of a virtual Service Composition Node (vSCN)[*7] and a virtual Media Processing Node (vMPN)[*8] [4]. The speech recognition and speech synthesis[*9] engines are within the platform for a speech translation service, and share interfaces with the existing Hanashite Hon'yaku service.

(a) Call processing

When a service user initiates or receives a call, the Mieru Denwa service initiates a voice call connection with the other party through the SEN platform. Ordinarily, the IP Multimedia Subsystem (IMS)[*10] platform makes the voice call connection through U-Plane transport equipment (VGN, SIN) within the IMS platform, but for Mieru Denwa, speech communication must be routed into the vMPN to perform speech recognition. Specifically, when the user enables the Mieru Denwa service on the application, connection

*3 WebSocket: A protocol that realizes real-time full-duplex communication between Web server and a client.
*4 Free-dial®: A registered trademark of NTT Communications Corp.
*5 SMS: A service for sending and receiving short text-based messages. SMS is also used for sending and receiving mobile terminal control signals.
*6 SEN: A platform able to provide added value by combining multiple enablers (See *17). Provides functions such as telecom, Web access, and media control.
*7 vSCN: Equipment that combines enablers (See *17) to provide a service based on a service scenario.
*8 vMPN: Media processing equipment. Provides various audio media services such as Voice Answering, and Melody Call services.
*9 Speech synthesis: Technology for artificially creating speech data from text and verbally reading out text.

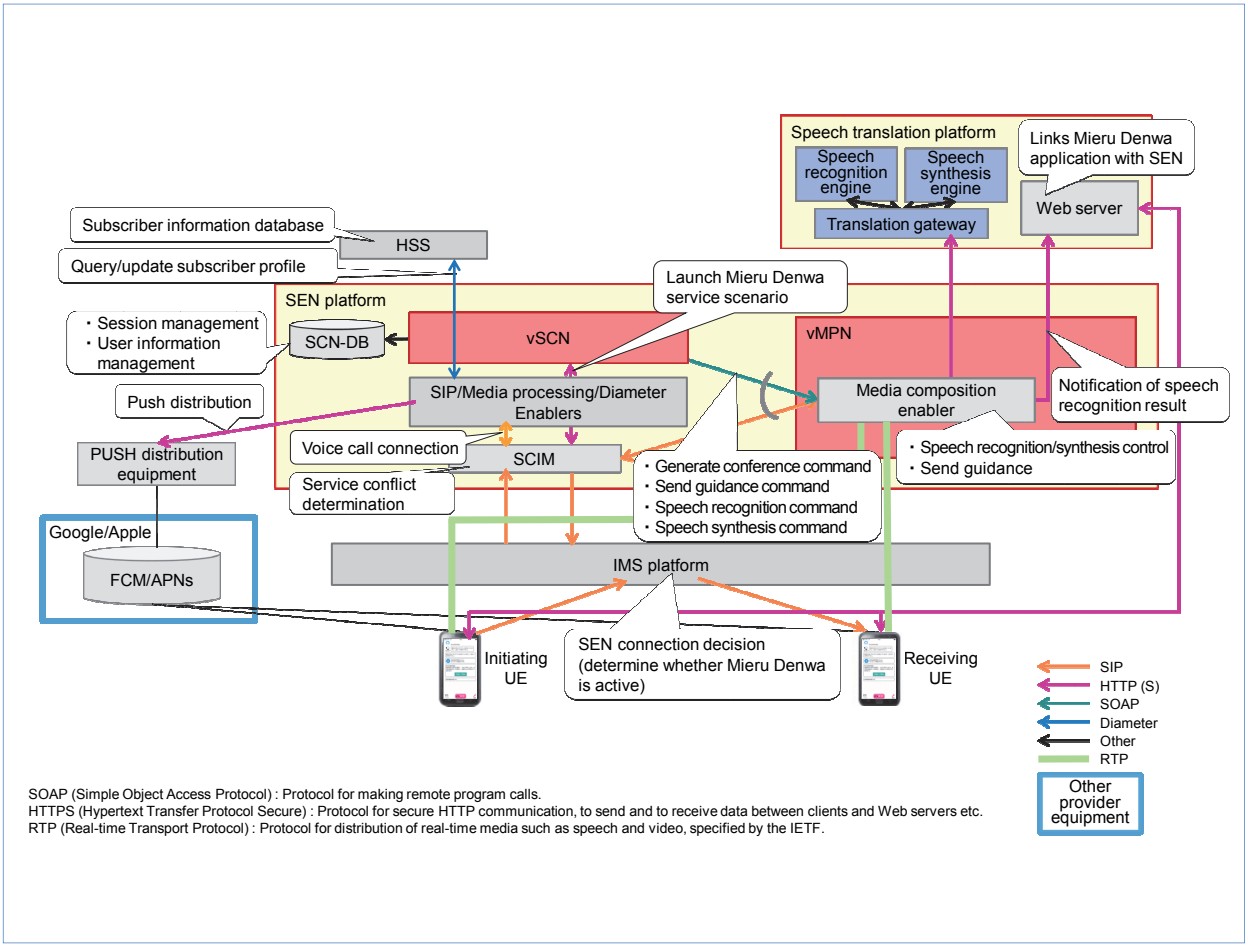

Figure 3 Mieru Denwa system architecture

information for the Service Capability Interaction Manager (SCIM)*11 is registered in the Shared initial Filter Criteria (SiFC)*12 of the IMS platform Service Call Session Control Function (S-CSCF)*13. Then, upon initiation or reception of a call, this causes the Session Initiation Protocol (SIP)*14 INVITE*15 (SIP-INVITE), which requests the voice call connection, to be sent to the SCIM, which is the SIP receiver module in the SEN platform.

In the SCIM, the subscriber's Mieru Denwa contract status and conflict with various other services is determined based on the subscriber profile*16 data, and a Mieru Denwa scenario is launched the vSCN. Based on the SIP-INVITE from the SCIM and with enablers*17 for SIP, media process and Diameter*18 and the service scenario, a conference room is generated in the vMPN, a conference service with the call initiator and

*10 IMS: Standardized by the 3GPP. A call control procedure that realizes multimedia communications by consolidating communication services offered over fixed and mobile networks using SIP (see *14), a protocol used on the Internet and in Internet phones.
*11 SCIM: Function that selects service scenarios according to user requests and controls service conflicts.
*12 SiFC: Criteria for determining which Application Server (AS) to send a request signal to, or the function for doing so.
*13 S-CSCF: A SIP server that performs UE session control and user authentication. A session refers to a continuous period of communication between a client and a server, or between two servers.
*14 SIP: A call control protocol defined by the Internet Engineering Task Force (IETF) and used for IP telephony with VoIP, etc.
*15 INVITE: A SIP signal that requests a connection.
*16 Subscriber profile: Information required for controlling services, including contract, user configuration, and location information.

receiver as participants is launched, and speech-voice is routed to the vMPN.

(b) Guidance control/speech-voice-to-text conversion

At the starting of a voice call, the Mieru Denwa service scenario first instructs the vMPN to playback voice guidance that introduces the service and then to perform continuous speech recognition. The vMPN plays speech guidance according to the service scenario instructions, and begins speech recognition. The speech recognition engine converts input voice data to text representing what was spoken and sends them to the user's smartphone through a Web server, which displays the received text in the smartphone application.

To display text of the voice call continuously in the application, the vMPN sends voice data to the speech recognition engine continuously during the call so it can continue, while performing speech recognition when it detects silent segments[*19] in the voice call. A WebSocket connection is also used between the Web server and the application so that text of the speech-voice can be displayed continuously, in real time in the application.

(c) Input utterance function

The vSCN instructs the vMPN to synthesize and play back text input by the user on the application, based on a speech synthesis requests sent by the Web server. On instructions from the vSCN, the vMPN sends the text to the speech synthesis engine, retrieves the synthesized speech, and starts playback of the speech. The synthesized speech is mixed with the spoken voice and transmitted so that it can be heard by the user and the other party.

(d) Application launch function using standard Android and iOS[*20] push notifications

Application launching using standard OS push notifications was adopted so that the Mieru Denwa application can be displayed in the foreground[*21] during a call. The Mieru Denwa application is launched by a push notification, either Firebase Cloud Messaging (FCM)[*22] or Apple Push Notification service (APNs)[*23]. Device information is stored in the SEN platform beforehand to identify where the notification from FCM or APNs is to be sent. When a call is started, a push notification send containing device information is sent to the PUSH distribution equipment[*24], which is a gateway[*25] within NTT DOCOMO. By launching the application using push notifications, it can be launched without the messaging application having to receive successive SMS messages. Note that on iPhones[*26], the notification must be tapped to launch the application.

2) Application Development

For the trial service, an application was only provided for Android devices, but for the commercial Mieru Denwa service, applications for both Android and iOS were developed. The Mieru Denwa application displays the text of speech from the other party on the screen during a call. Users of

*17 Enabler: A componentized function that can be used by multiple services.

*18 Diameter: An extended protocol based on the Remote Authentication Dial-In User Service (RADIUS), and used for authentication, authorization, and accounting in IMS.

*19 Silent segment: A segment of audio determined to be absent of speech on a telephone connection.

*20 iOS: A trademark or registered trademark of Cisco in the United States and other countries and is used under license.

*21 Foreground: Display of an application in front of other items on a smartphone screen so that the user can operate it immediately, even if other applications are shown on the home screen.

*22 FCM: A PUSH notification service that enables data to be sent from a server to a client, which is an application on an Android device.

*23 APNs: A service that uses PUSH technology to send notifications to an application on an iPhone device from a server through an always-open IP connection.

the service have hearing disabilities, so they converse with the other party by reading the displayed text to understand what was spoken, and then responding. Calls using Mieru Denwa involve the additional effort of converting to text and reading the text, so they proceed at a slower pace than otherwise. To enable calls with Mieru Denwa to proceed at a pace similar to regular voice calls, we also made some adjustments to how text conversion results are displayed and to the UI*27. Rather than waiting for full sentences to be completely converted before displaying them, intermediate results are displayed in real time in units of words (**Figure 4**). Then, when the utterance has completed, revised text for the whole sentence is displayed (**Figure 5**). Recognition results are displayed by words as the other party speaks, so conversation can proceed with as few obstacles as possible. A flashing icon appears on the screen showing clearly that the other party is speaking from the moment they begin. A glance shows whether the other party is speaking or silent, so users can easily know how to time their responses.

4. Conclusion

This article has described details of initiatives to commercialize the Mieru Denwa service and how it has been implemented. The service plays an important role supporting the lives of people with hearing disabilities, but it could potentially also be useful for people without such disabilities in some scenarios, such as environments with noise that makes it difficult to hear what the other

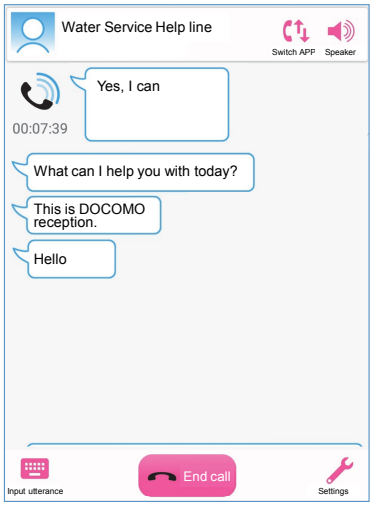

Figure 4 Screen shot with partial results during a call

*24 PUSH distribution equipment: Equipment that sends, receives and responds to SMS from a push client.
*25 Gateway: Equipment having functions such as protocol conversion and data transforming.
*26 iPhone: A registered trademark of Apple, Inc. United States, used within Japan under a license from Aiphone Co., Ltd.
*27 UI: Operation screen and operation method for exchanging information between the user and computer.

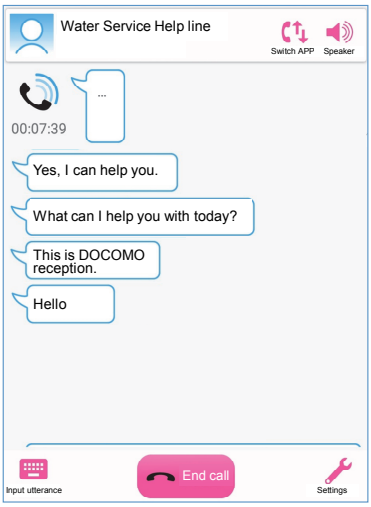

Figure 5 Screen shot with finalized result during a call

person in a phone call is saying. We will continue to improve the accuracy of speech recognition to further enhance the service in the future.

REFERENCES

[1] T. Koiso, K. Mikami, A. Sato and M. Ohta: “Inplementation method of speech-voice to text conversion service for hearing impaired people,” IEICE 2017 General Conference, 2017.

[2] K. Mikami, T. Koiso, A. Sato and M. Ohta: “Improvement of voice recognition method for continuous speech-voice to text conversion service,” IEICE 2017 General Conference, 2017.

[3] T. Koiso, K. Mikami, A. Sato and M. Tada: “A study of input utterance function on speech-voice to text conversion service for hearing impaired people,” IEICE 2017 Society Conference, 2017.

[4] Y. Iimura et al.: “SEN Infrastructure for Configuring a Network Cloud,” NTT DOCOMO Technical Journal, Vol.14, No.2, pp.4–13, Oct. 2012.