Technology Reports
Call Centers Speech Recognition AI

Improving Customer Satisfaction and Operator Efficiency in Call Centers Using AI —Speech Recognition IVR—

Service Innovation Department Takanori Hashimoto Yuuki Saitou Yuriko Ozaki

With recent advancements in AI technology, customer interaction systems using speech have started to be implemented. NTT DOCOMO has developed a Speech Recognition IVR system that can be introduced to receive telephone inquiries, and uses speech recognition to forward calls to the appropriate specialized call center. This improves customer satisfaction by reducing wait times and eliminating the need to enter numbers, and improves operator efficiency. This article describes the structure of Speech Recognition IVR and how it came to be introduced.

1. Introduction

The speech guidance for NTT DOCOMO's general inquiries line, called Interactive Voice Response (IVR)*1, has begun providing a Speech Recognition IVR functionality that is able to forward calls to the most suitable number by simply having the caller say what they need. This service can be used by dialing 151 from a DOCOMO mobile phone, or by calling the Free-Dial®*2 number (0120-800-000) and selecting the Speech Recognition IVR number.

Previously, customer inquiries were handled through the main, general inquiries call center together with various other specialized call centers such as the 113 Center (for repairs) and the DOCOMO Hikari call center. The basic process was for customers to call the general inquiries number and be forwarded to the appropriate center by selecting a number based on the voice guidance, but because there were many numbers to choose from and it was often difficult to know which one was the best, many customers just asked the operator to forward them to the appropriate center. In fact, of all the calls handled by the general call center,

*1 IVR: Automatic voice response equipment. A system that provides guidance on the telephone, with phrases such as, "For help with ⋯, please press number⋯."

*2 Free-dial®: A registered trademark of NTT Communications Corp.

approximately 20% of the customers were not able to select the correct option after voice guidance for reasons like those stated above. These callers needed to be forwarded from the selected specialized center to another one. As a result, operators needed to handle call forwarding and other inquiries for these callers. This created more congestion, increased customer wait times, and occupied specialist operators understanding customer inquiries a second time after the general reception center had forwarded the calls to them.

Speech interaction services in various forms have been implemented and offered in recent years. NTT DOCOMO has provided the "Shabette-Concier" service, which infers the intention of user utterances and performs searches for weather and restaurant information, and can launch the telephone applet.

NTT DOCOMO also provides the "DOCOMO DriveNet" [1] application, which provides useful information to drivers by simply talking to it, and has developed OHaNAS®*3 in collaboration with TOMY Co. Ltd., as part of NTT DOCOMO's "+d"*4 initiative to work with partners to create new value. OHaNAS is a toy that operates through a smartphone or tablet and is able to have natural conversations. NTT DOCOMO is developing a natural-language dialogue platform, as described above [2], and is accumulating knowhow in speech interaction services. As such, we have introduced Speech Recognition IVR utilizing AI technology, to improve issues with speech guidance systems as described earlier.

The system enables customers to simply state their issue, without needing to do complex button operations, to be connected to the appropriate call center. This reduces customer wait times before reaching an operator that can handle their issue, compared with forwarding from the general inquiries center. It also reduces work transferring calls by operators at the general inquiries center, and the customer's inquiry is converted to text and can be sent to the operator it is forwarded to, which can help reduce time in initial handling of the call.

This article describes the structure of the Speech Recognition IVR system.

2. Overview of Speech Recognition IVR

Speech Recognition IVR is a system combining IVR, which can receive inquiries by telephone, and speech recognition/intention interpretation*5, and is able to transfer calls to the appropriate specialized call center. The basic process of using the system is shown in **Figure 1**. After the IVR guidance is played, the user selects the number for "Speech Recognition IVR," and speech recognition begins. The speech recognition function converts the customer's statement of their issue to text. Then, the intention interpretation function determines which operator to forward the call to among all of the specialized call centers, who has skills appropriate for the issue. The telephone reception server receives the above result from speech recognition and intention interpretation, and forwards the call to the appropriate center. The text version of the customer's issue is also sent to the operator.

The above system connects the customer to the appropriate center, when they simply select the

*3 OHaNAS®: A registered trademark of TOMY Company, Ltd.
*4 +d: Name of NTT DOCOMO initiative for creating new value together with partner companies.
*5 Intention interpretation: A function that interprets the intention behind the text and assigns it to an appropriate task. An example of text and a task is, "Will it rain today?" and "Weather report search."

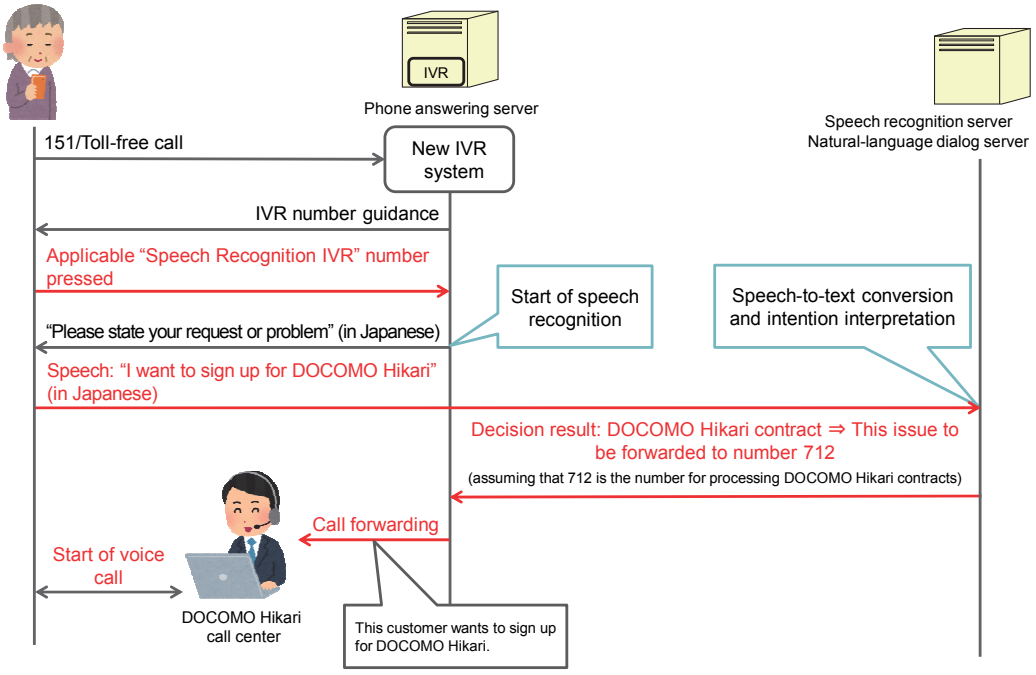

Figure 1 Overview of Speech Recognition IVR service

Speech Recognition IVR number and stating the issue. The operator is also able to check a summary of the issue before speaking with the customer.

3. Speech Recognition Function

The speech recognition function converts the recorded speech of the calling customer to text. This process is shown in **Figure 2**. Conversion to text uses an acoustic model*6, a pronunciation dictionary, and a language model*7. The speech recognition technology first uses the acoustic model to match phonemes*8 to the utterance waveform, and then converts the phonemes to words using the pronunciation dictionary. The language model is then used to statistically determine the word order, which is finally converted to text in sentences. Features of each of these steps are described below.

(1) Customers calling the call center vary in aspects such as gender, age, and location, so the sound of the utterance also varies. Calls also come from different phones, whether fixed-line or smartphone, and each use different types of audio compression, so the sound quality varies. The acoustic model

*6 Acoustic model: Statistical model comprising frequency characteristics of phonemes targeted for recognition possesses.
*7 Language model: Statistical model comprising morpheme arrangements and frequency of those arrangements.
*8 Phoneme: The smallest units of speech, such as vowels and consonants.

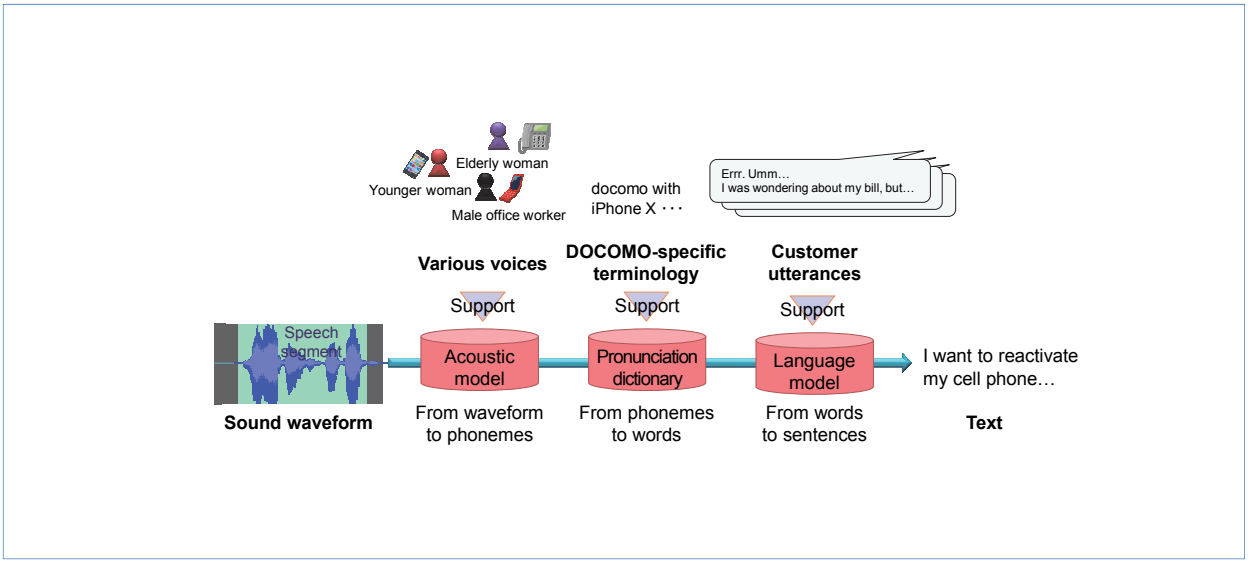

Figure 2 Speech recognition system processing

handles such differences in sound quality.

(2) Words that can be expected to occur in a telephone inquiry and related to NTT DOCOMO (handset models, service names, etc.) together with variations in pronunciation are registered in the pronunciation dictionary and used to improve word recognition accuracy.

(3) The customer utterance is not a well-formed, carefully recited sentence, but is filled with fillers[*9] and other modified phraseology, much like free conversation. To analyze such sentences, as many as possible anticipated phrases of this type are gathered and used to train the language model.

Accurate and appropriate ending points in the customer's utterance are also found, and it is forwarded without unnecessary sections to further improve usability. In doing so, there is also a function to segment utterances naturally, taking noise (ambient, other speakers, music, etc.) in the customer's environment into consideration.

The calling customer's statements are converted to text using these speech recognition functions and passed to the intention interpretation function.

4. Intention interpretation Function

1) Overview

The intention interpretation function determines what sort of inquiry is intended in the utterance, from the customer text information, and decides where the call should be forwarded. Destinations for forwarding calls are decided based on the numbers provided by NTT DOCOMO telephone inquiry reception, as shown in **Table 1** (as of September 2017). Our customer inquiries span an extremely wide range, so to increase coverage, we have added items for inquiries not included in Table 1.

Interpreting the intention of the customer's utterance can be solved as a sentence classification problem, classifying the input text. Specifically, a

*9 Fillers: Words used to connect utterances, like "umm…" and "you know…."

Table 1 DOCOMO inquiry categorization

Major category	Minor category	Sub-category
Inquiries regarding DOCOMO Hikari	Moving, new applications and other procedures and inquiries	Moving procedure (transfer)
		New subscriber procedure
		Cancellation procedure
		Inquiry
	Construction date inquiries	Construction scheduling
		Construction rescheduling
		Other construction inquiries
	Connecting and settings	
	Problems	
Suspending/ Resuming service		
Operation of telephones and data communications devices	iPhone and iPad® operation	
	Smartphone and tablet operation	
	docomo Feature Phone operation, settings	
	DOCOMO Hikari Internet and other settings	
	Mobile Wi-Fi® router and other data communication device operation and settings	
Problems/ Service area	Smartphone and tablet problems	
	docomo Feature Phone and other terminal problems	
	Service areas	
	iPhone and iPad® problems	
	DOCOMO Hikari problems	
Guidance on various services	Rate plans, discount services	
	Mobile number portability	
	Various network services	
	International services	
	Unlocking a SIM lock	
Other order procedures/ inquiries	Order procedures	
	Inquiries	

iPad®: A trademark of Apple Inc. registered in the USA and other countries.

Wi-Fi®: A registered trademark of the Wi-Fi Alliance.

model is trained using the language dictionary and a large number of sample utterances matched to forwarding numbers, as shown in **Figure 3**. It is able to determine the intention and a forwarding number from the input text at high speed. As with tuning of speech recognition, the sample utterances and language dictionary are created using NTT DOCOMO specific terms and covers the diversity of utterances occurring in inquiries.

2) Discriminating Ambiguous Statements

There are cases when it is difficult to determine a suitable forwarding number from the limited information in a customer's statements. For example, for an inquiry like, "How do I use my device?" although it is clear that it is about using a device, there is a wide range of devices handled

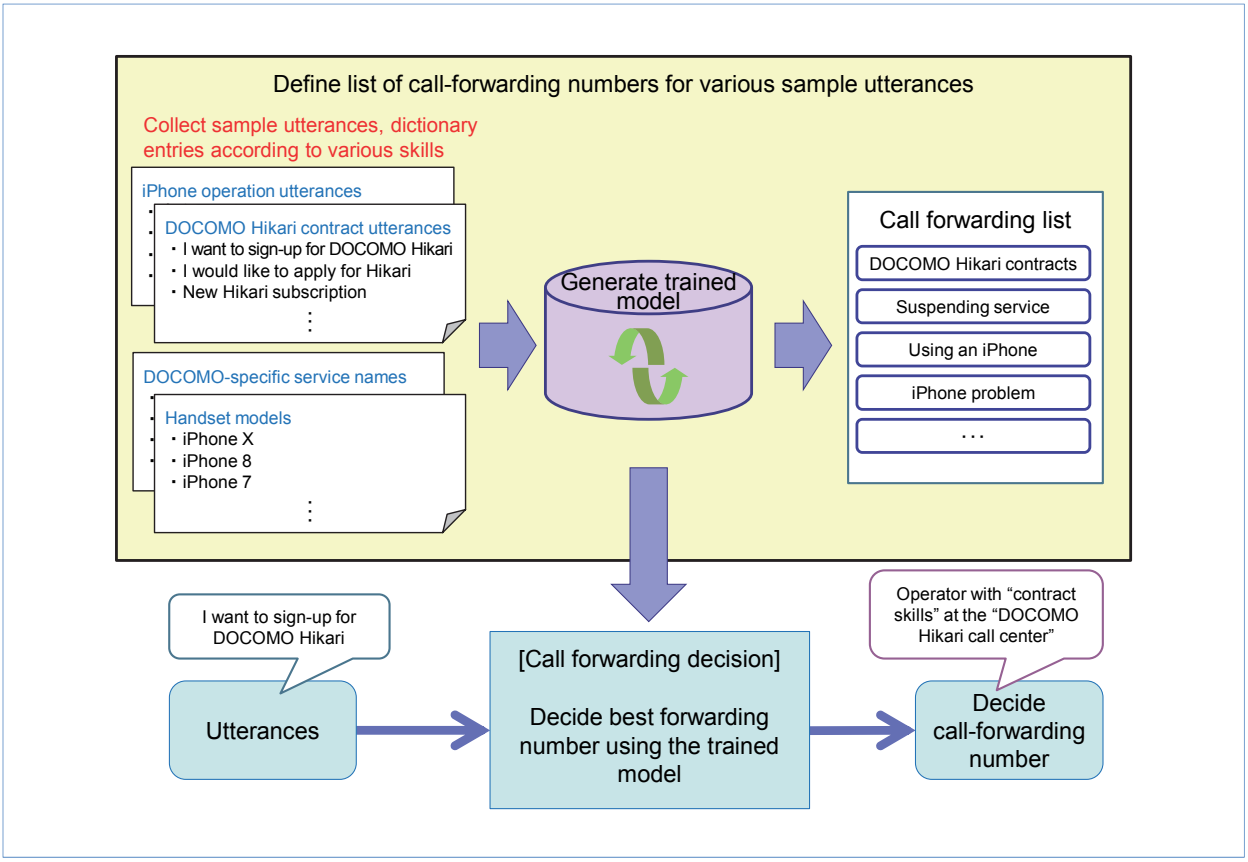

Figure 3 Deciding call forwarding number

by NTT DOCOMO and it can be difficult to narrow this down. As such, we use two methods to discriminate ambiguous utterances.

- The first is to prompt the user. For example, if it is about how to use a device, we can narrow down to specialized centers by asking about major categories (iPhone®*10, Android™*11, docomo Feature Phone, etc.).
- The second is to narrow-down and forward the call according to operator-specific skills. Within each specialized center, each operator is able to handle a different range of inquiries. Inquiries that are more ambiguous can be forwarded to operators with broader skills, and if the utterance contains enough information, it is forwarded to an operator with suitable skills.

These two approaches are used to forward inquiries to an appropriate specialized center according to the content of the customer's utterance.

The utterance and forwarding number obtained using the speech recognition and intention interpretation processes described above are used by the telephone reception server to decide where the customer's call will be forwarded and are displayed on the operator's screen.

*10 iPhone: A registered trademark of Apple, Inc. United States, used within Japan under a license from Aiphone Co., Ltd.
*11 Android™: A trademark or registered trademark of Google Inc., United States.

5. Conclusion

This article has described a Speech Recognition IVR system that introduces AI into call center work and is able to transfer a call to an appropriate specialized call center and send details to the operator by just having the caller speak on the telephone. After the system was introduced, 10 to 20% of customers were transferred to the specialist center, and this was the goal we initially intended, without using the general reception center, reducing the work of operators transferring calls. We will continue efforts to improve performance, increasing use of Speech Recognition IVR and the rate of successful call forwarding.

This article has only dealt with forwarding of calls, but we are studying ways to automate parts of call center work, by expanding the interaction of customers with the AI to be bi-directional. We are also studying the potential to provide this functionality to other companies as an NTT DOCOMO corporate solution.

REFERENCES

[1] R. Kurita et al.: "DOCOMO DriveNet Info," NTT DOCOMO Technical Journal, Vol.16, No.1, pp.4–10, Jul. 2014.

[2] K. Onishi et al.: "Natural-language Dialogue Platform for Development of Voice-interactive Service," NTT DOCOMO Technical Journal, Vol.17, No.3, pp. 4–12, Jan. 2016.