Technology Reports

AI | Voice Recognition | Dialogue

"AI Phone Service" to Automate Telephone Reception and Monitoring

Service Innovation Department Tomoko Kawase
5G & IoT Business Department Shin Oguri
Service Design Department Yuki Saito

The number of businesses implementing contact centers is increasing with the spread of cloud-based contact center systems, but the shortage of human resources for operators has become a problem. To automate routine telephone answering and post-answering office work, NTT DOCOMO has developed an automated telephone answering service called "AI Phone Service," which features identity verification through voice recognition. This service makes it possible to automate tasks such as accepting reservations and applications, and monitor the elderly.

1. Introduction

In improving customer satisfaction, telephone contact points such as contact centers are significant as they are customer contact points that do not require IT literacy. In recent years, cloud-based contact centers[*1] have become popular, and the number of businesses implementing contact centers is increasing. However, to respond to an increasingly diverse range of customers in an easy-to-understand and prompt manner, and to improve customer satisfaction, in addition to business knowledge, operators need communication skills as well as IT skills to carry out post-call operations, but there are not enough human resources to fill these positions. Also, staffing must be flexibly adjusted during busy and off-season periods. As well as that, the number of people in contact center offices must be reduced to prevent the spread of the novel coronavirus. Against this backdrop, there are growing demands for automated telephone answering services using Artificial Intelligence (AI). Overseas, services using AI to support telephone operations have been available since

around 2018, and in Japan, many cases implementing AI telephone application acceptance were announced in 2020. Using AI to take over routine answering duties conventionally performed by operators means operators can focus on non-routine answering duties. In addition, AI can limit the IT skills required of operators by taking over post-call tasks, which will help reduce the shortage of IT-literate operators.

NTT DOCOMO has already implemented a voice recognition Interactive Voice Response (IVR)*2 [1] function for IVR mechanisms for general inquiry counters (general centers). These systems automatically connect users to the appropriate specialist center to handle their inquiries, and utilize the spoken dialogue service know-how accumulated through the AI agent services “Shabette Concier*3” and “my daiz*4” [2]. Voice recognition IVR implementation has had the effect of reducing waiting time before being connected to an operator, reducing the time spent answering the phone at the general centers, and reducing the amount of work transferring to specialist centers.

NTT DOCOMO has developed a new cloud service called “AI Phone Service” to provide customers with a solution for improving the efficiency of their telephone answering services with voice interaction technology. This article provides an overview of the AI Phone Service and its mechanism, and describes approaches to speech recognition technology to support various use cases.

2. Service Overview

The AI Phone Service is designed to be used by businesses such as local governments, retailers, restaurants and companies with call centers. As shown in **Figure 1**, use cases with received phone

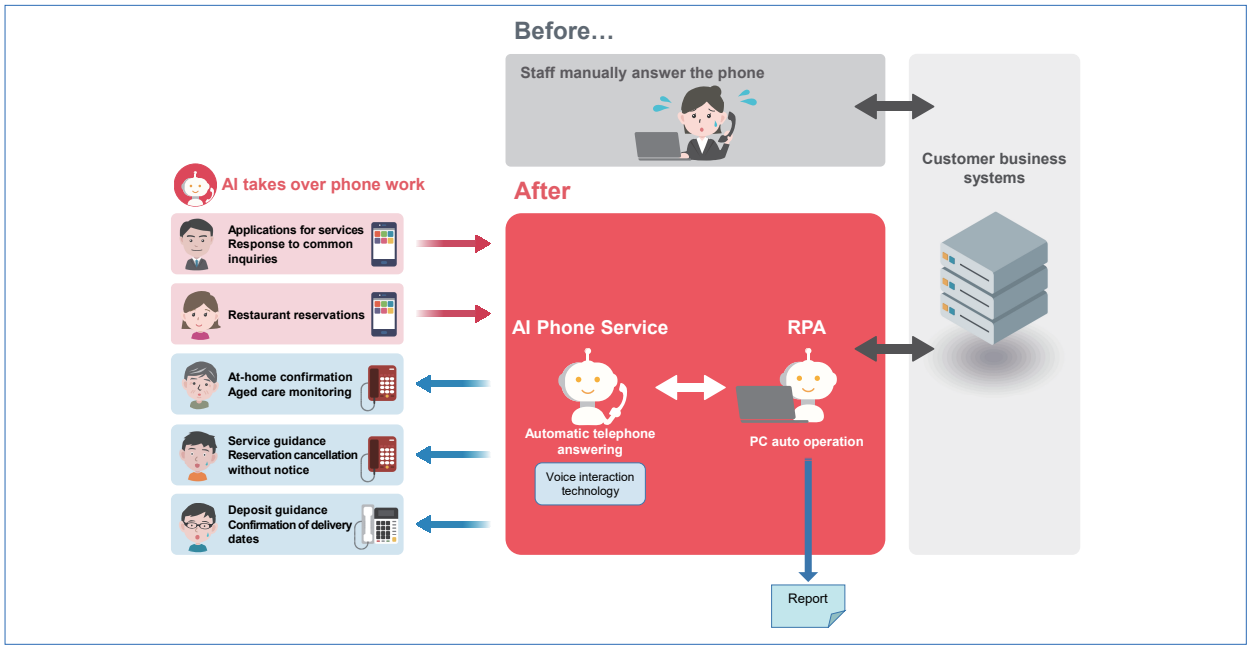

Figure 1 Overview and use cases for AI Phone Service

*1 Cloud-based contact center: A system for responding to customers over the phone that operates using servers on a network, rather than servers owned by a company.

*2 IVR: A system that provides voice guidance over the phone, such as “For XX, please press Y.”

*3 Shabette Concier: A speech dialogue agent that runs on smartphones or tablets, providing conversation with characters, making phone calls through conversation, setting alarms, searching for transfers, and fortune telling.

*4 my daiz: A speech dialogue agent that runs on smartphones and tablets, providing a wide range of information suited to the user.

calls include service application/modification, responding to common inquiries, and accepting reservations for restaurants and vehicle dispatch. This system can be used not only for receiving calls but also for making calls and can be applied to a wide range of applications beyond contact center operations, such as checking up on the elderly in the homes, monitoring their health, providing customers with service information, and confirming and reminding customers of reservations and payment information. In addition, the system not only automates the routine tasks that operators formerly performed, but also automates post-call tasks by linking spoken dialogue technology with Robotic Process Automation (RPA)*5. For example, it is possible to automatically create reports based on dialogue content logs and link with the customer business systems. Implementing the AI Phone Service will not only help solve the problem of securing human resources but will also enable 24/7 support.

3. System Configuration

The AI Phone Service system configuration is shown in **Figure 2**. The automatic telephone answering service is realized by linking the "AI phone core application" to the "docomo AI Agent Application Programming Interface (API)*6" [2], which provides the NTT DOCOMO dialogue technology, and the "Amazon Connect" cloud-based call center service. As telephones are used, voice interface functions are provided by a voice recognition engine.

The docomo AI Agent API provides functions that enable AI to respond to the user according to

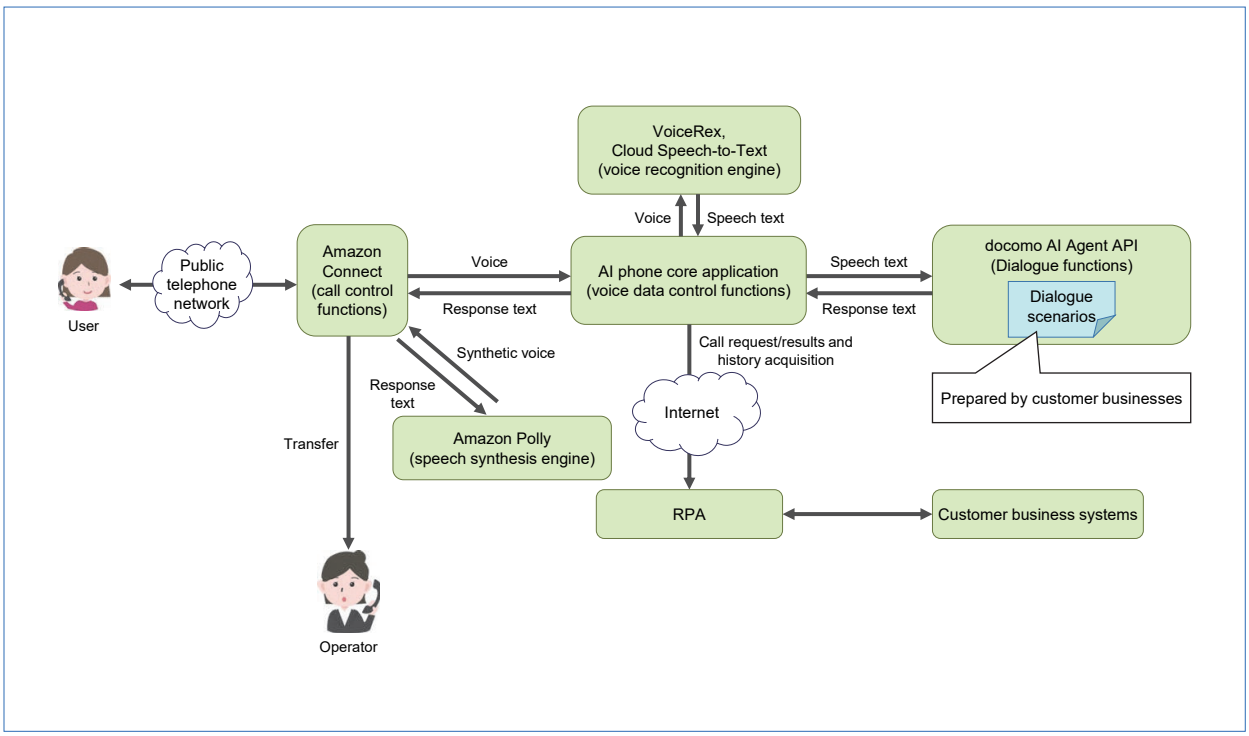

Figure 2 AI Phone Service system configuration

*5 RPA: A mechanism to automate routine tasks.

*6 API: An interface that enables the functions of software to be used by other programs.

predetermined dialogue scenarios, and notably, customer businesses can flexibly create their own dialogue scenarios.

Amazon Connect provides call control functions that give customer businesses using the AI Phone Service the advantages of easy expansion and not having to manage call control servers. There is also a function to transfer calls to operators when cases arise that are not easily handled automatically by AI. However, one of the limitations of using Amazon Connect is that the only text-to-speech*7 engine that can be used is "Amazon Polly." Amazon Polly cannot reproduce the synthesized voice of a specific person to use as the AI voice, and in the case of Japanese language, there is only a choice between one male and one female speaker. Nevertheless, by tuning the speaking speed, pauses, and volume within the range of the selected speaker's voice, it is possible to make the AI speak slowly and loudly for important words, for example.

In sequence, the AI phone core application feeds the user's voice obtained from Amazon Connect to the voice recognition engine*8 and then receives the speech text as the recognition result. Once the voice recognition engine detects the end of the user's speech, the AI phone core application works with the docomo AI Agent API and RPA to perform the subsequent processing.

The system uses NTT "VoiceRex*9" and Google Cloud Speech-to-Text as voice recognition engines. It mainly uses the former, but it can also select the latter each time it is a user's turn to speak. Since each engine has different areas of strength, it is possible to specify an engine to utilize its advantages in the scenario in advance, depending on what is to be heard. For example, when a person needs to be identified over the phone, it is necessary to listen to his or her name. VoiceRex can output voice recognition results in both kanji and kana and can also output information that the word is classified as "first name" or "last name," which is useful for accurately recognizing Japanese names.

4. Approaches to Speech Recognition Technology for Various Use Cases

In the AI Phone Service, VoiceRex is applied with the NTT DOCOMO original language model*10, which has been proven in Shabette Concier, my daiz, and voice recognition IVR. To apply VoiceRex to automate telephone answering in a variety of use cases, we worked on the following three things.

4.1 Speech Recognition for Names

In use cases such as service application/modification and taking reservations, it is important to complete the identity verification scenario, which requires accurate voice recognition of the caller's name. Therefore, we tuned the language model by adding names as training data to the aforementioned language model. Real names cannot be used due to restrictions on personal information, so fictitious Japanese names were generated and used as training data. We evaluated name recognition performance before and after tuning and found that the error rate fell to less than 70% of the error rate before tuning.

4.2 Speech Recognition for Scenario-specific Words

Dialogue scenarios differ with each customer business, and user speech assumed in each scenario

*7 Text to speach: Technology for artificially creating speech data from text and verbally reading out text.
*8 Voice recognition engine: Equipment that takes voice data as input and converts it to text of what was spoken.
*9 VoiceRex: A voice recognition engine developed by NTT Media Intelligence Laboratories.
*10 Language model: A model that represents the frequency of word order.

also differs for each dialogue scenario. If words, which are frequently used in a particular scenario, are rare in general dialogue, it is not uncommon for AI agents to misrecognize them. For example, in the use case of restaurant reservations, users might frequently utter the word "private room." However, since the frequency of occurrence of "koshitsu (private room)" is not so high inside the aforementioned language model, it may be misrecognized as words such as "hoshitsu (moisturizer)" or "koushitsu (imperial family)" (similar sounding words in Japanese). Also, in principle, service names that are not included in the language model cannot be recognized. Therefore, it is generally necessary to tune the language model in advance, as mentioned above. However, tuning the language model and implementing it in the voice recognition engine every time a dialogue scenario is added is not practical from either the computational or operational perspectives. Therefore, VoiceRex has a function that enables specification of a list of expected words for each speech recognition request, which makes it easier to output specified words without tuning the language model. This function enabled improved speech recognition performance.

4.3 Appropriate Timing of Responses Based on Speech Content

In spoken dialogue with AI agents, the speed of the response, i.e., the replay of AI speech a short time after user speech finishes, is a factor that leads to better user experience. However, when a user pauses (to breath, etc.) while saying such things as an address, a sequence of numbers or an open-ended response, if AI regards the pause as the end of the speech, it might not hear the speech after the pause, or it may interrupt the user's speech and start responding. In other words, there is a trade-off between the success rate of listening to pause-containing speech and response speed.

Therefore, in the AI Phone Service, with each user's turn to speak in a scenario, an allowable pause length is set according to the expected speech content, which is dynamically specified by the AI phone core application to the voice recognition engine. For example, after the AI asks, "Is your name Taro Tanaka?", short user answers such as "Yes" or "No" are expected, so the allowable pause length should be set to a few hundred milliseconds. In contrast, after the AI asks, "Is there anything you are taking care of for your health?", the user is expected to think and speak for a long time, so the allowable pause length should be set to longer than one second. This eliminates the aforementioned trade-off and allows the system to respond to short user speech at a good tempo while still being able to hear user speech that contains pauses until the end.

5. Verification Experiments

Before providing the AI Phone Service commercially, NTT DOCOMO established a test environment and conducted verification experiments for two use cases.

5.1 Accepting Applications with Identity Verification

To confirm the effectiveness of reducing the telephone answering workload of businesses that provide monthly services, we conducted a verification experiment with the use case of accepting

an application over the phone. Since the application process here involves identity verification linked to a user database in a customer service system, we designed and applied the scenario shown in **Figure 3**. If the user can be uniquely identified by simply searching the name in the user database, acceptance is complete. Even if the user cannot be uniquely identified by name confirmation, customer number confirmation or address confirmation, acceptance is complete if the user can be uniquely identified by combining them with the confirmation of a fee payment. In this verification experiment, we obtained an acceptance dialogue completion rate of 77%. In the verification experiment, dialogue was carried out using voice input only. However, commercial systems also support dial key input which holds promise for dialogue completion rate of 88% when used in combination with voice input.

5.2 Monitoring the Elderly

Elderly people who live alone tend to have less communication with others and need support such as daily calls, but local support organizations do not have enough staff to take care of each elderly person. Thus, to verify whether AI telephone support can solve the problem of monitoring elderly people living alone and the burden on support organizations, we conducted a verification experiment involving automatically calling elderly people living alone at regular intervals to check on their health and safety. Over the phone, the AI asks the questions shown in **Table 1** and converses with the elderly person. Since it was difficult to quantitatively

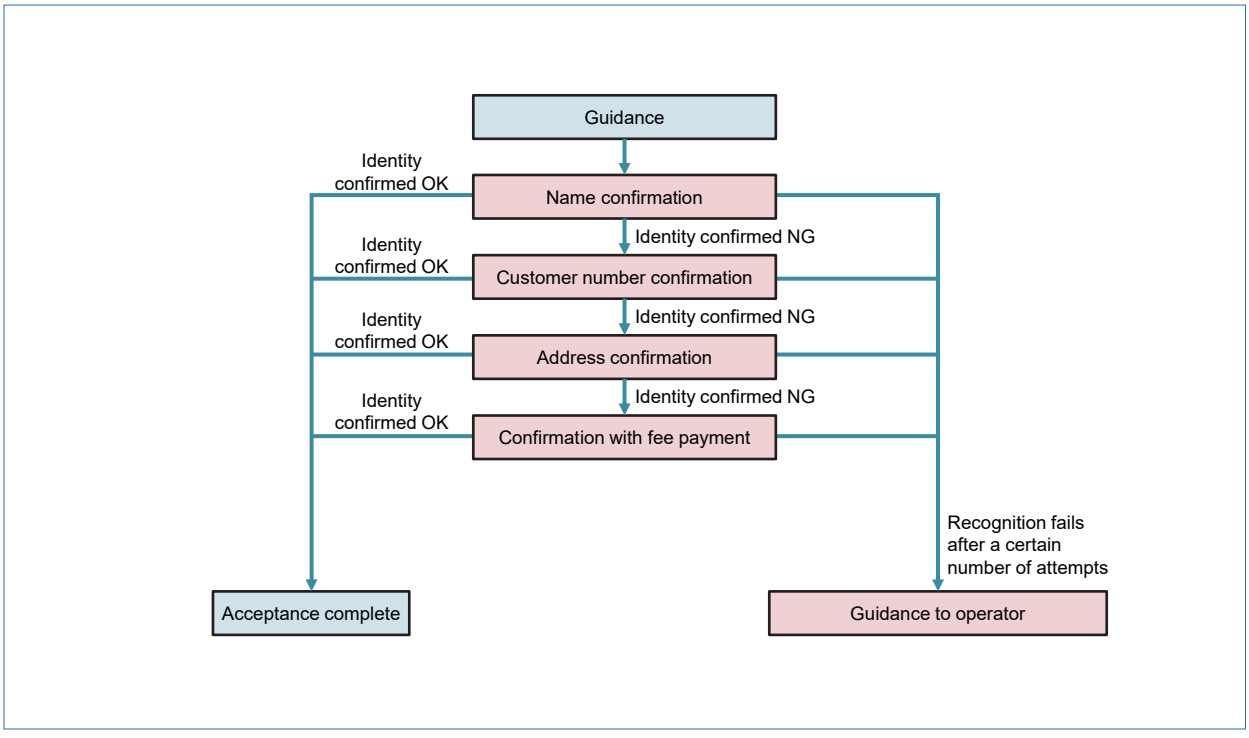

Figure 3 Accepting an application with an identity verification scenario

Table 1 Elderly monitoring questions

Item	Question
Sleep	Did you sleep well last night?
	What time did you go to bed last night?
	Did you wake up during the night?
	Is there anything else that is bothering you about your sleep?
Meals	Did you eat three meals yesterday?
	Have you eaten any protein this week, such as meat, fish, or eggs?
	Do you have an appetite?
	Is there anything else about your diet that concerns you?
Activities	Did you go out yesterday?
	Do you have any plans to go out today?
	Have you talked to your family and friends?
	Is there anything you would like to try next week?
	What would you like to do?
Physical condition	How are you feeling today?
	Have you had a bowel movement?
	Have you been to hospital recently?
	Is there anything particular that you do for your health?
	What kind of things do you pay attention to?
Body care and grooming	Did you take a bath yesterday?
	Do you soak in the bathtub?
	Do you take care of your teeth every day?
	Do you have any other concerns about body care and grooming?

measure the effects of monitoring the elderly with AI, we conducted interviews with the people subjected to the experiment. Based on the results of these interviews, we have been conducting a second verification experiment since February 2021.

6. Conclusion

In this article, we discussed the AI Phone Service to automate phone answering. These services have the potential to alleviate the problem of the shortage of human resources engaged in telephone

answering services. NTT DOCOMO began providing commercial AI Phone Service in December 2020 and plans to continue verification experiments and officially launch services for the use cases of accepting applications and reservations, and monitoring the elderly. Going forward, we will continue to work on the technological challenge of further improving speech recognition performance.

REFERENCES

[1] T. Hashimoto, et al.: “Improving Customer Satisfaction and Operator Efficiency in Call Centers Using AI - Speech Recognition IVR -,” NTT DOCOMO Technical Journal, Vol.19, No.4, pp.4-10, Apr. 2018.

[2] T. Oba, et al.: “docomo AI Agent Open Partner Initiative,” NTT DOCOMO Technical Journal, Vol.20, No.3, pp.4-9, Jan. 2019.