Evolution of Speech Recognition System—VoiceRex

Takanobu Oba, Tomohiro Tanaka, and Ryo Masumura

Abstract

Speech recognition is a key element of artificial intelligence for contact centers. It is now used in a wide range of scenarios, supporting business in various ways. Research and development of speech recognition has a long history and has been built upon various technologies to reach today's standards. We introduce the VoiceRex speech recognition system developed by NTT Media Intelligence Laboratories, its history, and some technologies employed in the latest VoiceRex system, which are much anticipated for use in contact centers.

Keywords: speech recognition, VoiceRex, contact center

1. History of VoiceRex

Speech recognition is a key technology to understand human communication and is a necessary element of artificial intelligence (AI) for contact centers. Speech recognition is technology to convert speech in an input signal into text. Research and development (R&D) of speech recognition at the NTT laboratories has a long history, spanning half a century. NTT Media Intelligence Laboratories has developed the VoiceRex speech recognition system based on the results of these long years of research and is providing the technology to NTT Group companies, where it will be applied in a wide range of service fields.

The idea of using speech recognition to analyze contact center calls was in mind from the initial stages of this R&D. This capability was almost unthinkable at the time, but it was set as a goal to be reached some decades in the future. VoiceRex (or its predecessor speech recognition library) was first released in the early 1990s, but at that time, it was only able to recognize keywords. The current ability to recognize longer utterances and conversations was achieved in 2000. Even so, it was not at a level where it could correctly recognize conversations between two people. It could only recognize formal language such as newspaper text if it was read out clearly, and the recognizable vocabulary was very limited.

Then VoiceRex went through several technical innovations, and the performance increased dramatically. In 2008, we incorporated a weighted finite state transducer for the first time in Japan. This enabled it to learn approximately 100 times more words than before and to recognize speech from among roughly 10 million words. In 2009, this advancement was used in a system that creates a record of debates in the Japanese House of Representatives. In conditions where one person speaks at a time in question-answer format, the system was able to achieve 90% recognition accuracy. This was the beginning of the use of speech recognition to replace manual shorthand in the main assembly and various committees.

Thereafter, speech databases continued to be extended and consolidated, computing performance improved, and technology was created to utilize large-scale databases efficiently, so the performance of speech recognition continued to increase steadily. Finally, speech recognition for contact center calls started to become practical, and in 2014, NTT Software released its ForeSight Voice Mining* product for contact centers.

For several years before that, another technology had been attracting attention in the speech recognition

* ForeSight Voice Mining is currently distributed by NTT TechnoCross.

research community. Deep neural networks had emerged on the scene. Deep learning caused a great paradigm shift in speech recognition. The performance of conversion from sound signals (air pressure fluctuations) to an acoustic model (sequences of phonemes, which are vowel and consonant sounds) increased sharply, which led to dramatic increases in recognition rates for phone calls.

A commercial version of VoiceRex using deep learning was released in 2014. Then, in 2015, we used a type of neural network called CNN-NIN (convolutional neural network - network in network) to perform speech recognition on sound from a mobile terminal in a noisy public area. This resulted in a first place award among participating research institutions at the CHiME3 (The 3rd CHiME Speech Separation and Recognition Challenge) international technology evaluation event. With the spread of smartphones, people are making telephone calls more frequently while they are out and about, and we are now able to recognize speech accurately in audio signals containing more ambient noise, as when calling from a crowded location.

Through such technical innovations, applications of speech recognition have expanded rapidly. VoiceRex is now used in many products and services. In particular, the number of installed AI-related contact center products have increased rapidly, and these have become a core segment of speech recognition products and services.

However, some new issues have arisen as more and more customers have started using the technology. One such issue concerns the diversity in terms of topics. Current speech recognition technology can work more accurately if the topics in the input call audio are known ahead of time. For example, the names of services being handled will differ for each company operating a call center. Even within one company, the departments for registering or canceling membership, for receiving complaints, and for answering technical questions are all separate. Thus, current technology tunes language models for content separately for each company and also for each contact center.

Another issue is handling the flow of conversation. When two people talk, they often do not speak in clear statements. For example, when a sentence ends in "…-tion," which would be written "tʃən" using a phonetic pronunciation syllable, it might only be pronounced "tʃ," and the remaining "ən," while pronounced, may not be clear and may only be expressed in the rhythm of the speech. This can occur frequently. For such cases of unclear or omitted pronunciation, a mechanism to predict words from the context is needed. Below, we introduce a new technology incorporated in the latest version of VoiceRex to overcome this issue.

2. Conversational context language model

A language model is a model that predicts the sequence of individual words. Intuitively, it has the role of deciding whether a sentence is correct as a language or not. Our speech recognition system uses a probabilistic model called an N-gram language model, which creates models based on the idea that a word is dependent on the N-1 preceding words. The number of word combinations increases exponentially as N increases, so at most, three or four N values are used. As such, models only consider very local context. For short utterances, an N-gram language model is sufficient, but as utterances get longer, it becomes necessary to consider more context.

Thus, language models using neural networks have attracted attention recently, particularly recurrent neural network (RNN) language models, which are able to handle context over longer periods. When these models are used for speech recognition, multiple candidate speech recognition results are first obtained, and each result is scored using an RNN language model to determine the final recognition result. This is called the rescoring method.

One issue with RNN language models is that they are limited to using context in a single utterance. For calls to a contact center, using context that spans utterances is very important. For example, when the operator answers a customer's question, the content of the answer must be related to the content of the customer's question. Therefore, we have developed conversational context language model technology [1] that considers context over longer periods, spanning utterances. When it performs speech recognition, it uses the speech recognition results from successive utterances as context. In VoiceRex, a conversational context language model is applied using the rescoring method for each utterance to select better (closer to correct) sentences, and these results are used as context to perform speech recognition on the next utterance. In this way, the effectiveness of the conversational context language model is gained with the input of each utterance.

3. Neural error corrective language model

As stated before, as fluency increases, cases of

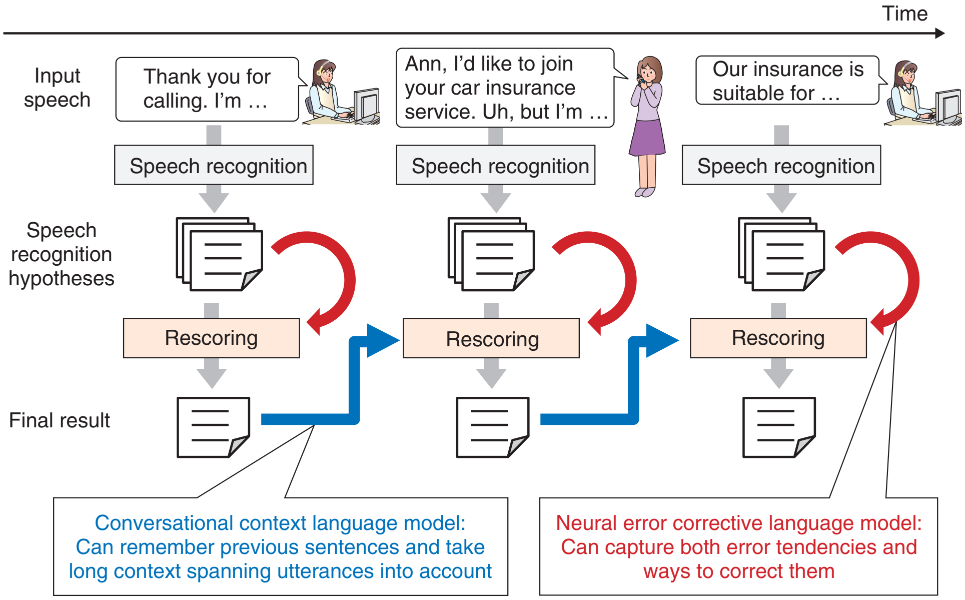

Time
Input speech
Thank you for calling. I'm …
Ann, I'd like to join your car insurance service. Uh, but I'm …
Our insurance is suitable for …
Speech recognition
Speech recognition
Speech recognition
Speech recognition hypotheses
Rescoring
Rescoring
Rescoring
Final result
Conversational context language model: Can remember previous sentences and take long context spanning utterances into account
Neural error corrective language model: Can capture both error tendencies and ways to correct them

Fig. 1. Flow of applying conversional context language model and neural error corrective language model.

unclear pronunciation appear, but certain tendencies also appear in such cases. This frequently occurs for sentences ending in "-tion" (tʃən) as mentioned earlier, and also for prepositions and frequently occurring expressions such as "Thank you very much." Speech recognition tends to make the same mistakes in each of these cases. There are also many other error patterns that occur frequently, beyond those caused by unclear pronunciation. The approach of detecting such error tendencies and correcting them is called error correction.

We were able to improve recognition accuracy by developing neural error corrective language model technology [2] that incorporates a framework for considering speech recognition errors into an RNN language model. Specifically, we introduced a neural network called an encoder-decoder model, which provides a mechanism to select the correct sentence from among results that include speech recognition errors. To train the neural error corrective language model, we used sentences that produced speech recognition errors together with the corresponding correct sentences, and trained for the relationship between them. In this way, we captured both the error tendencies and ways to correct them.

When performing speech recognition, we use the rescoring method, in the same way as for the conversational context language model. This model can be used together with the conversational context language model, producing results that are augmented by both language models for each utterance. This is illustrated in **Fig. 1**. During a call, utterances are extracted from the input signal, multiple recognition result candidates are output for each, and both models are applied to select the final recognition result. For the conversational context language model, this final recognition result is returned as context for recognition of the next utterance.

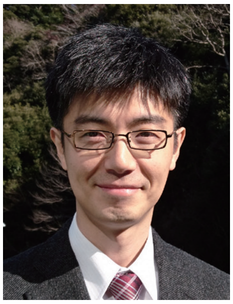
4. Future prospects

For most of the contact centers considered so far, the customers would initiate the call to the contact center. In such cases, the conversation between the operator and customer is one-time-only and relatively formal. This is good for the accuracy of speech recognition. However, as AI products are introduced into contact centers, they are being used more and more by companies to contact their customers. In such cases, an operator is assigned to each customer, and the operator speaks to the customer multiple times to encourage more frank conversation, which complicates speech recognition. As in the past, as the range of applications has broadened, situations are encountered that make recognition more difficult than before. Each time this occurs, R&D is conducted to overcome the difficulty, and the technology continues to advance. We will continue to advance VoiceRex, meeting the new challenges encountered by users as they use speech recognition in real life.

References

[1] R. Masumura, T. Tanaka, A. Ando, H. Masataki, and Y. Aono, “Role Play Dialogue Aware Language Models Based on Conditional Hierarchical Recurrent Encoder-Decoder,” Proc. of the 19th Annual Conference of the International Speech Communication Association (Interspeech 2018), pp. 1259–1263, Hyderabad, India, Sept. 2018.

[2] T. Tanaka, R. Masumura, H. Masataki, and Y. Aono, “Neural Error Corrective Language Models for Automatic Speech Recognition,” Proc. of Interspeech 2018, pp. 401–405, Hyderabad, India, Sept. 2018.



Takanobu Oba

Senior Research Engineer, Audio, Speech and Language Media Project, NTT Media Intelligence Laboratories.

He received a B.E. and M.E. from Tohoku University, Miyagi, in 2002 and 2004, and a Ph.D.(Eng.) from Tohoku University in 2011. He joined NTT Communication Science Laboratories in 2004, where he has been engaged in research on spoken language processing. He received the 25th Awaya Kiyoshi Science Promotion Award from the Acoustical Society of Japan (ASJ) in 2008. He joined NTT DOCOMO in 2015 and worked on service development of spoken dialogue systems such as docomo AI Agent. He is currently working on contact center related solutions at NTT Media Intelligence Laboratories. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and ASJ.

Tomohiro Tanaka

Researcher, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received a B.E. from Tokyo University of Science in 2015 and an M.E. from Tokyo Institute of Technology in 2017. Since joining NTT in 2017, he has been researching automatic speech recognition and spoken language processing. He was the recipient of the 13th Best Student Presentation Award of ASJ. He is a member of ASJ.

Ryo Masumura

Distinguished Research Scientist, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received a B.E., M.E., and Ph.D. in engineering from Tohoku University, Miyagi, in 2009, 2011, and 2016. Since joining NTT in 2011, he has been researching speech recognition, spoken language processing, and natural language processing. He received the Student Award and the Awaya Kiyoshi Science Promotion Award from ASJ in 2011 and 2013, respectively, the Sendai Section Student Awards Best Paper Prize from the Institute of Electrical and Electronics Engineers (IEEE) in 2011, the Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2014, the Young Researcher Award from the Association for Natural Language Processing (NLP) in 2015, and the ISS Young Researcher’s Award in Speech Field from IEICE in 2015. He is a member of ASJ, IPSJ, NLP, IEEE, and the International Speech Communication Association.