VoiceRex—Spontaneous Speech Recognition Technology for Contact-center Conversations

Hirokazu Masataki[†]*, Daisuke Shibata, Yuichi Nakazawa, Satoshi Kobashikawa, Atsunori Ogawa, and Katsutoshi Ohtsuki*

Abstract

For speech recognition technology to be useful in contact centers, it must be capable of recognizing free conversation between humans. However, free conversation is much harder to process than clear speech produced by an announcer reading a prepared text. This article introduces our latest efforts toward implementing speech recognition for free conversation.

1. Introduction

What do you suppose it would be like to talk to a machine and have it understand what you are saying? Scenes in which people start talking to electronic gadgets and robots are a staple feature in science fiction movies, so there are probably lots of people who think this sort of technology is still a long way off. However, speech recognition technology has already started to appear in practical applications. NTT has over 30 years' experience in this field [1] and has already developed various practical applications of this technology. Examples include our V-Portal service [2], which allows users to access a range of information by talking into a telephone, and our real-time subtitle creation system [3] for live broadcasts of baseball matches.

Continuous speech recognition involves recognizing sentences and phrases from natural speech, which tends to contain multiple words strung together in a continuous stream. So far, we have reached the level where it is possible to achieve an accuracy of about 90% in the recognition of continuous speech. However, with conventional technology, this accuracy can be achieved only if the words are spoken clearly and smoothly (like a television announcer, for example).

For speech recognition technology to work in situations such as contact centers, it must be able to recognize free conversation between humans. At NTT Laboratories, we have been researching and developing free conversation speech recognition technology for several years, but our earlier technology was aimed at recognizing conversations between humans and machines [4]. Although it could handle various linguistic expressions such as filler words and hesitations, enabling the system to achieve a good recognition accuracy, it is inadequate for free conversation because when people talk to other humans they use speech that is much freer and rougher than when talking to a machine. To recognize free conversation accurately, the speech recognition technology needed to be improved to handle filler words and hesitations better and extended to handle linguistic problems such as colloquialisms and acoustic problems such as fast speech, indistinct pronunciation, and background noise (**Fig. 1**). These problems made accurate recognition difficult to achieve with earlier technology.

2. Free conversation speech recognition technology

The principle of a standard continuous speech recognition engine is shown in **Fig. 2**. Continuous speech recognition involves the use of three models: an acoustic model that associates phonemes with voice characteristics, a recognition dictionary that defines the words to be recognized, and a language model that expresses the connections between words. These models are used to apply numerical scores to

[†] NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan
Contact: https://www.ntt.co.jp/cclab/contact/index.html

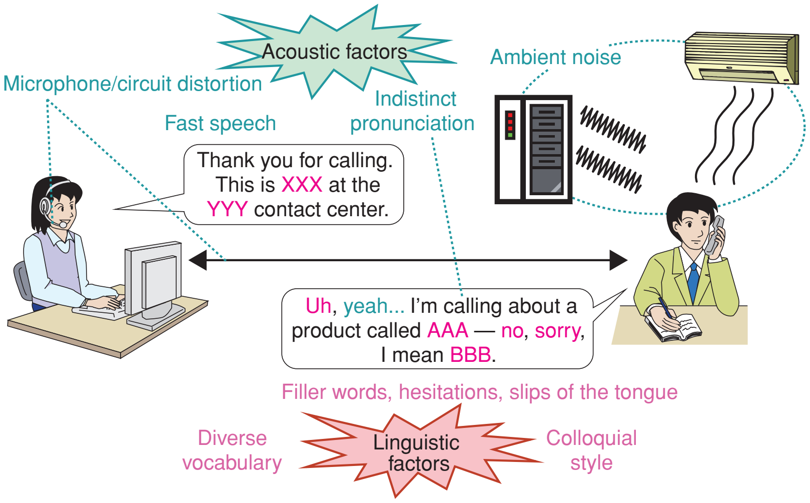
Acoustic factors
Microphone/circuit distortion
Ambient noise
Fast speech
Indistinct pronunciation
Thank you for calling. This is XXX at the YYY contact center.
Uh, yeah... I'm calling about a product called AAA — no, sorry, I mean BBB.
Filler words, hesitations, slips of the tongue
Diverse vocabulary
Linguistic factors
Colloquial style

Fig. 1. Problems of speech recognition for free conversation.

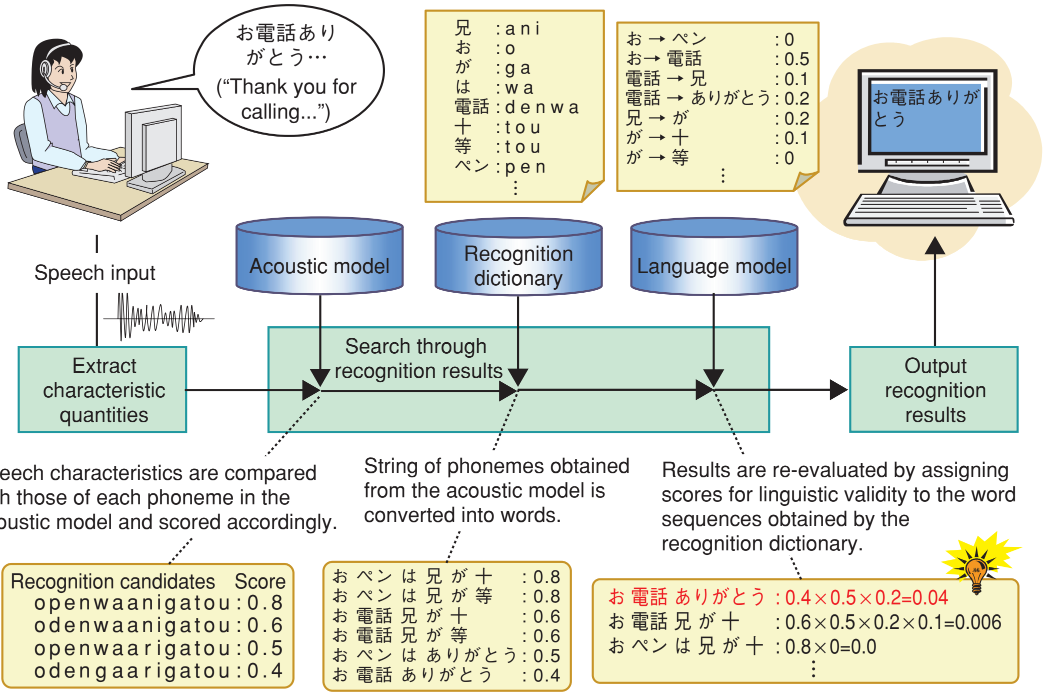
お電話ありがとう…
("Thank you for calling...")
兄 :ani
お :o
が :ga
は :wa
電話 :denwa
十 :tou
等 :tou
ペン :pen
⋮
お → ペン :0
お→ 電話 :0.5
電話 → 兄 :0.1
電話 → ありがとう :0.2
兄 → が :0.2
が → 十 :0.1
が → 等 :0
⋮
お電話ありがとう
Speech input
Acoustic model
Recognition dictionary
Language model
Extract characteristic quantities
Search through recognition results
Output recognition results
Speech characteristics are compared with those of each phoneme in the acoustic model and scored accordingly.
String of phonemes obtained from the acoustic model is converted into words.
Results are re-evaluated by assigning scores for linguistic validity to the word sequences obtained by the recognition dictionary.
Recognition candidates Score
openwaanigatou : 0.8
odenwaanigatou : 0.6
openwaarigatou : 0.5
odengaarigatou : 0.4
お ペン は 兄 が 十 : 0.8
お ペン は 兄 が 等 : 0.8
お 電話 兄 が 十 : 0.6
お 電話 兄 が 等 : 0.6
お ペン は ありがとう : 0.5
お 電話 ありがとう : 0.4
お 電話 ありがとう : 0.4×0.5×0.2=0.04
お 電話 兄 が 十 : 0.6×0.5×0.2×0.1=0.006
お ペン は 兄 が 十 : 0.8×0=0.0
⋮

Fig. 2. Principle of continuous speech recognition.

the input speech with regard to acoustic similarity and linguistic validity, and the string of words with the highest score is output as the resulting recognized text. It is not possible to obtain high accuracy in the recognition of free conversation unless these three models are specifically tuned for free conversation. For each of these models, we describe the efforts that have been made to resolve the characteristic problems of free conversation.

2.1 Adapting an acoustic model to free conversation

The acoustic model consists of a set of phoneme models. A phoneme is roughly equivalent to the sound corresponding to a single letter of a Japanese word written in Roman letters. In a phoneme model, the acoustic features of phonemes are assembled together and expressed as probability values. The acoustic model determines suitable probability values

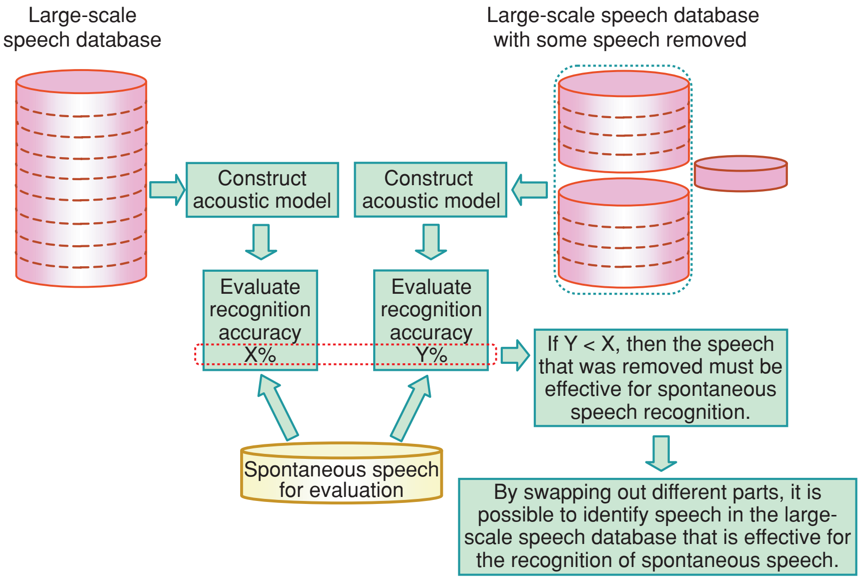
Large-scale speech database
Large-scale speech database with some speech removed
Construct acoustic model
Construct acoustic model
Evaluate recognition accuracy X%
Evaluate recognition accuracy Y%
If Y < X, then the speech that was removed must be effective for spontaneous speech recognition.
Spontaneous speech for evaluation
By swapping out different parts, it is possible to identify speech in the large-scale speech database that is effective for the recognition of spontaneous speech.

Fig. 3. Selection of data for the free conversation acoustic model.

by statistically analyzing large quantities of speech data. The performance of the acoustic model largely depends on the quality and quantity of speech data used in its construction. To improve the performance of the acoustic model, it is essential to collect large amounts of high-quality speech data.

Previously, we have used a large-scale speech database containing over a thousand hours of recordings of speakers reading prepared documents such as news articles in a relatively clear voice. Compared with this sort of careful speech, free conversation is characterized by a faster rate of speech, less distinct pronunciation, and so on. Consequently, with only a speech database of prepared readings, no matter how large, it is not possible to make an acoustic model that is capable of recognizing free conversation with high accuracy. Therefore, we compiled a new free conversation speech database containing recordings of two people conversing freely on particular subjects. However, it was not possible to gather a large volume of high-quality spontaneous speech data, so this database contains only a few tens of hours of speech.

To make up for this lack of speech data, we developed a technique for automatically selecting utterances that are the closest to free conversation from the conventional speech database of prepared readings. This technique is summarized in **Fig. 3**. We conducted tests to compare the accuracy of spontaneous speech recognition using (1) an acoustic model constructed from a speech database of prepared readings and a free conversation speech database and (2) an acoustic model constructed from data excluding part of this database. When excluding some of the data resulted in a model with lower recognition accuracy, we concluded that the excluded data had a positive effect on the accuracy of spontaneous speech recognition. This process was used to evaluate the entire database, allowing us to select speech close to spontaneous speech from the speech database of prepared readings. By using this technique to acquire a total of several hundred hours of speech, we were able to construct a highly accurate acoustic model for recognizing free conversation.

In situations where speech recognition is used in practice, the acoustic characteristics vary from one location to the next due to variations in parameters such as microphone and transmission characteristics and background noise levels. If such variations are reduced when the acoustic model is constructed, it is possible to adapt the model to a wide variety of recording environments. However, when there has been severe distortion such as a high level of background noise or where a different recording microphone has been used, it has not been possible to avoid degraded speech recognition performance. The acoustic model environment adaptation technique shown in **Fig. 4** automatically estimates the background noise and speech distortion contained in the speech data recorded by the end user and derives a numerical matrix representation of the amount of

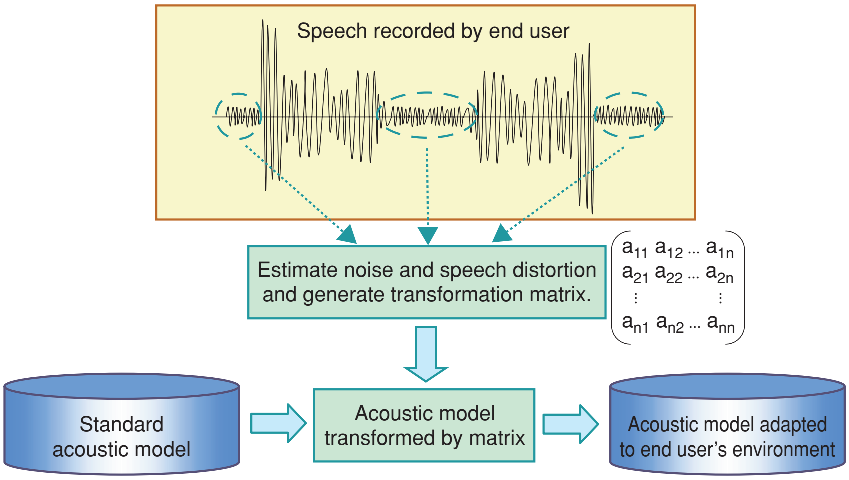
Speech recorded by end user
Estimate noise and speech distortion and generate transformation matrix.
a11 a12 ... a1n
a21 a22 ... a2n
an1 an2 ... ann
Standard acoustic model
Acoustic model transformed by matrix
Acoustic model adapted to end user's environment

Fig. 4. Environment adaptation of the acoustic model.

variation in the sound relative to a suitable standard reference acoustic model. If the standard acoustic model is multiplied by this matrix, it can be transformed into an acoustic model suited to the end user's acoustic environment. This technique makes it easy to adapt the model to a wide variety of recording environments.

2.2 Adapting the recognition dictionary and language model to free conversation

Continuous speech recognition relies on the use of a recognition dictionary (a list of words that can be recognized by the recognition engine) and a language model (which applies constraints to the concatenation of words from the recognition dictionary). In current speech recognition mechanisms, words that are not registered in the recognition dictionary (unregistered words) cannot be output in the recognition result. It is thus necessary for the dictionary to contain as many as possible of the words that are likely to be included in the input speech. However, as the number of words included in the dictionary increases, the processing speed and memory requirements of the speech recognition processing become larger, so the dictionary must be designed to cover as many different expressions and phrases as possible with a limited number of words. We therefore collected texts with a wide variety of contents and entered the words appearing in them into a well-balanced dictionary, thereby designing a general-purpose dictionary having few unregistered words for a wide variety of applications.

To recognize free conversation, the language model must also be adapted to free conversation. Free conversation contains various forms of expression such as repetition, hesitation, and filler words such as "er" and "um", but it would take a great deal of work to make grammatical rules that can handle all of these expressions. Therefore, a statistical language model is used for the recognition of free conversation. In a statistical language model, the ease with which different words can be joined together is represented in the form of statistical concatenation rules by analyzing a large volume of text data.

The language model is trained on large amounts of text gathered from sources such as web pages and transcripts of the speech in the free conversation speech database described above. Although the latter contains many expressions that are found in free conversation, the contents are very limited. Conversely, the former includes a very wide range of content, but the majority of the text is written in a literary style. Therefore, as shown in **Fig. 5**, by training the language model based on a carefully balanced mixture from both sources, it became possible to use words and expressions from a wide variety of fields while dealing with the diverse phraseology of free conversation. Furthermore, by introducing a language model where similar words are handled together as a single class, we further increased the recognition accuracy.

3. Conclusion

The technology described here makes it possible to achieve recognition with fairly high accuracy even for free conversation. In the future, we will continue

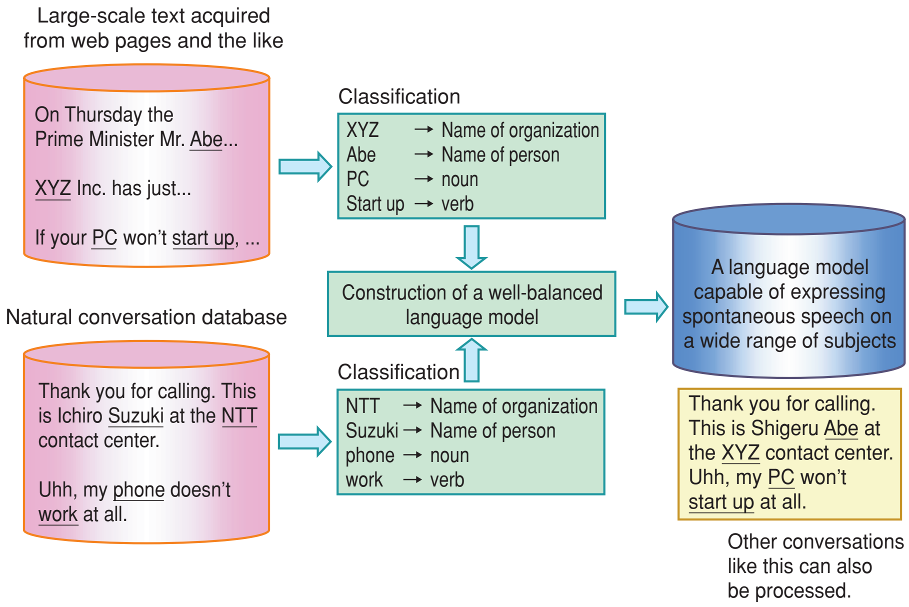
Large-scale text acquired from web pages and the like
On Thursday the Prime Minister Mr. Abe...
XYZ Inc. has just...
If your PC won't start up, ...
Classification
XYZ → Name of organization
Abe → Name of person
PC → noun
Start up → verb
Construction of a well-balanced language model
A language model capable of expressing spontaneous speech on a wide range of subjects
Natural conversation database
Thank you for calling. This is Ichiro Suzuki at the NTT contact center.
Uhh, my phone doesn't work at all.
Classification
NTT → Name of organization
Suzuki → Name of person
phone → noun
work → verb
Thank you for calling. This is Shigeru Abe at the XYZ contact center. Uhh, my PC won't start up at all.
Other conversations like this can also be processed.

Fig. 5. Construction of a language model compatible with free conversation.

to develop this technology to increase its accuracy and allow it to be used in a broader range of situations.

One goal involves increasing the vocabulary of the recognition dictionary. A contact center can be involved in various types of business and may use a large number of words including product names and specialist terminology from its particular field of business. To broaden the applicable scope of speech recognition, it is desirable that as many words as possible can be handled, but as mentioned above due to limits on processing speed and memory, we are currently limited to a vocabulary of roughly 100,000 words. To break through this limit, we are researching a weighted finite-state transducer (WFST) in collaboration with NTT Communication Science Laboratories. This should substantially increase the search efficiency of recognition results by representing the acoustic model, dictionary, and language model as a single compact network. In our experiments, even a commercial PC was able to implement continuous speech recognition with a huge vocabulary of over one million words.

Another area to study is model tuning techniques. To raise the technology to a level where it can be put to practical use, it is sometimes necessary to tune models to suit the environments in which they will be used. In the construction of the acoustic model and language model, we use recorded speech and its transcription, but transcribing this speech is very laborious. We are therefore working to develop technology that trains the acoustic model and language model automatically from the speech without it having to be transcribed.

If these technologies can be achieved, then it should be possible to produce speech recognition systems that can be used in all sorts of situations and that start off with a high recognition accuracy and gradually become even more accurate as they are used.

References

[1] Y. Noda, Y. Yamaguchi, K. Ohtsuki, and A. Imamura, “Development of the VoiceRex speech recognition engine,” NTT Technical Journal, Vol. 11, No. 12, pp. 14-17, 1999 (in Japanese).

[2] V-Portal: http://www.ntt.com/v-portal/

[3] H. Suzuki, H. Kikuchi, and H. Sakaguchi, “Real-time speech recognition subtitle equipment for Japanese television,” Broadcast Engineering, Vol. 58, pp. 711-714, July 2005.

[4] J. Hirasawa, T. Amakasu, S. Yamamoto, Y. Yamaguchi, and A. Imamura, “Cyber Attendant System with Spontaneous Speech Interface,” NTT Technical Review, Vol. 2, No. 3, pp. 64-69, 2004.

Hirokazu Masataki

Senior Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E. and M.E. degrees in electrical engineering, and the Dr.Eng. degree in information engineering from Kyoto University, Kyoto, in 1989, 1991, and 1999, respectively. Since joining NTT in 2004, he has been working on R&D of automatic speech recognition technologies. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the Acoustic Society of Japan (ASJ).

Daisuke Shibata

Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E. and M.E. degrees in electrical engineering from Waseda University, Tokyo, in 2002 and 2004, respectively. Since joining NTT in 2004, he has been working on R&D of automatic speech recognition technologies. He is a member of ASJ.

Yuichi Nakazawa

Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E. degree in environmental information engineering and the M.E. degree in media and governance engineering from Keio University, Kanagawa, in 2001 and 2003, respectively. Since joining NTT in 2003, he has been working on R&D of automatic speech recognition technologies. He is a member of ASJ.

Satoshi Kobashikawa

Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E. degree in electronic engineering and the M.E. degree in information engineering from the University of Tokyo, Tokyo, Japan, in 2000 and 2002, respectively. Since joining NTT in 2002, he has been working on R&D of automatic speech recognition technologies. He is a member of ASJ.

Atsunori Ogawa

Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E. and M.E. degrees in information engineering from Nagoya University, Aichi, in 1996 and 1998, respectively. Since joining NTT Laboratories in 1998, he has been engaged in research on speech recognition. He is a member of IEICE and ASJ and received the ASJ Best Poster Presentation Award in 2003 and 2006.

Katsutoshi Ohtsuki

Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E. and M.E. degrees in electrical engineering from Waseda University, Tokyo, in 1993 and 1996, respectively. Since joining NTT in 1996, he has been working on R&D of automatic speech recognition technologies. From 2004 to 2005, he was a visiting scientist at BBN Technologies, MA, USA, working on statistical language modeling. He is a member of IEEE, IEICE, and ASJ. He received the Awaya Prize from ASJ in 2004.