Saxe: Text-to-Speech Synthesis Engine Applicable to Diverse Use Cases

Yusuke Ijima, Nozomi Kobayashi, Hiroko Yabushita, and Takashi Nakamura

Abstract

Technological advances such as deep learning and changes in the social structure where artificial intelligence supports and substitutes human activities have recently been accompanied by changes in use cases requiring speech-synthesis technology and in the functions and performance they require. Key issues in implementing these new use cases are understanding according to context, reproduction of speaker characteristics, and support for diverse operation environments. NTT Media Intelligence Laboratories is developing a speech-synthesis engine (development codename “Saxe”) based on deep neural networks to address these issues. This article provides a technological overview of Saxe along with several application examples and touches upon future developments.

Keywords: speech synthesis, deep learning, diversity

1. Introduction

Speech synthesis is technology for generating speech from input text, therefore it is sometimes called text-to-speech synthesis. NTT has a long history of research and development in the field of speech synthesis and has developed a variety of speech-synthesis technologies that have found widespread use in services with the purpose of conveying information accurately. These include the telephone services web171 (disaster emergency message board) and 177 (weather forecast telephone service) and IVR (interactive voice response) systems.

There have also been a variety of technological advances typified by deep learning and changes in the social structure where artificial intelligence (AI) is increasingly being used to support and substitute human activities. These developments are being accompanied by changes in use cases requiring speech-synthesis technology and in the functions and performance they require. Uses cases that aimed to convey information accurately have required the generation of speech for standard text in a specific speaker’s voice. However, use cases that support and substitute human activities require the generation of speech for all kinds of text in a user-desired speaker’s voice in diverse operation environments. NTT Media Intelligence Laboratories is developing a speech-synthesis engine (development codename “Saxe”) based on deep neural networks (DNNs) in response to these issues and is promoting its practical application in diverse use cases. In this article, we outline the technologies behind this engine, introduce several application examples, and touch upon future developments.

2. Technology overview

2.1 High-accuracy reading of heteronyms according to context

Speech synthesis can be broadly divided into a text-analysis section that infers the reading and accents of characters from input text and a speech-synthesis section that generates synthesized speech from the

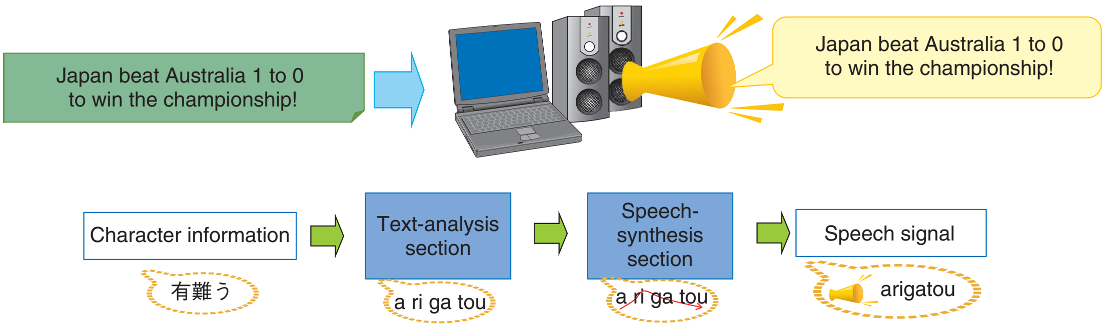
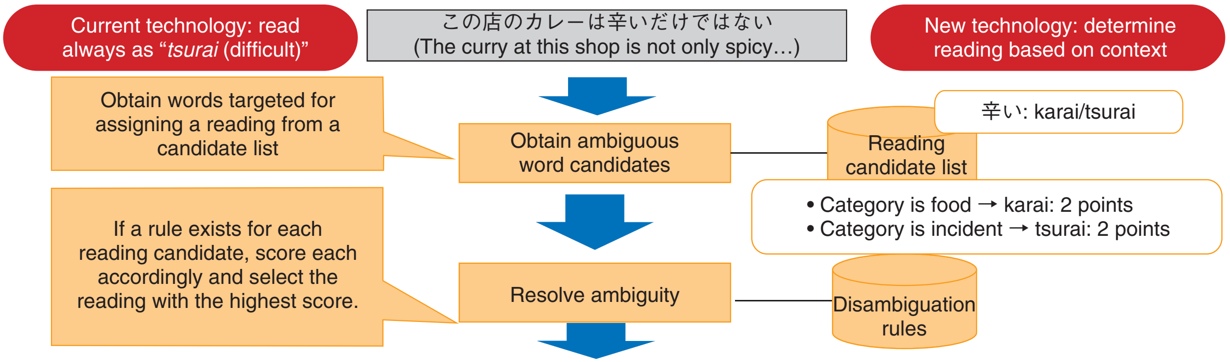


Fig. 1. Overview of speech-synthesis technology.


Current technology: read always as “*tsurai* (difficult)”
この店のカレーは辛いだけではない
(The curry at this shop is not only spicy…)
New technology: determine reading based on context
Obtain words targeted for assigning a reading from a candidate list
Obtain ambiguous word candidates
辛い: karai/tsurai
Reading candidate list
• Category is food → karai: 2 points
• Category is incident → tsurai: 2 points
If a rule exists for each reading candidate, score each accordingly and select the reading with the highest score.
Resolve ambiguity
Disambiguation rules

この	店	の	カレー	は	辛い	だけ	では	ない
kono	mise	no	karee	wa	karai	dake	dewa	nai

Fig. 2. Overview of word-reading disambiguation technology.

inferred reading and accents (**Fig. 1**). In the text-analysis section, an erroneous inference of the reading or accents of characters can result in the transmission of incorrect information to the listener of that synthesized speech, so there is a need for high-accuracy inference of reading and accents with respect to input text. The Japanese language, however, includes heteronyms, which are words for which reading and accents differ according to context despite being written the same. For example, the word “辛い” can be pronounced “*karai* (spicy)” or “*tsurai* (difficult)” and the word “寒気” can be pronounced “*samuke* (chills)” or “*kanki* (cold air).” For this reason, achieving high-accuracy inference of reading and accents has become a major objective.

To this end, we developed word-reading-disambiguation technology that infers the correct reading against an obvious misreading. This technology can infer the reading of a word having ambiguity through the use of dictionaries and rules that make use of linguistic knowledge. For example, a rule can be prepared beforehand that says if the word “カレー,” which is pronounced “*karee* (curry),” should appear near the word “辛い,” points are given to reading the latter as “*karai* (spicy).” Consequently, the word “辛い” appearing in the sentence “この店のカレーは辛いだけではない” (The curry at this shop is not only spicy…) would then be correctly read as “*karai*” (**Fig. 2**). Since it would be difficult to exhaustively write out all words considered to be associated with “*karai* (spicy),” an alternative means would be a framework that uses word categories (e.g., “food”) as rules. Therefore, we can reduce the number of rules while improving comprehensiveness. This technology can infer readings with high accuracy while saving on memory.

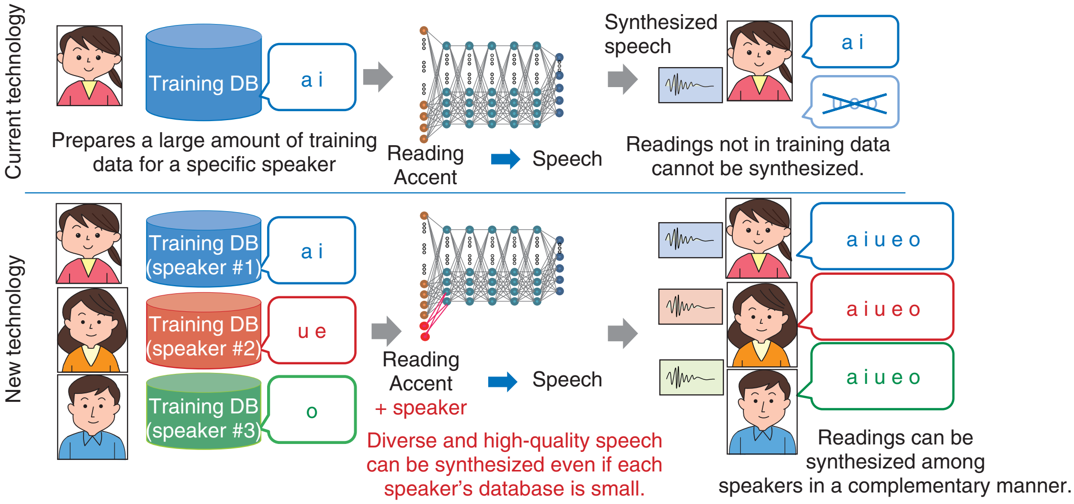

Current technology
Training DB
a i
Prepares a large amount of training data for a specific speaker
Reading Accent
Speech
Synthesized speech
a i
a i u e o
Readings not in training data cannot be synthesized.
New technology
Training DB (speaker #1)
a i
Training DB (speaker #2)
u e
Training DB (speaker #3)
o
Reading Accent + speaker
Speech
Diverse and high-quality speech can be synthesized even if each speaker's database is small.
a i u e o
a i u e o
a i u e o
Readings can be synthesized among speakers in a complementary manner.

Fig. 3. DNN speech-synthesis technology for reproducing diverse speaker characteristics.

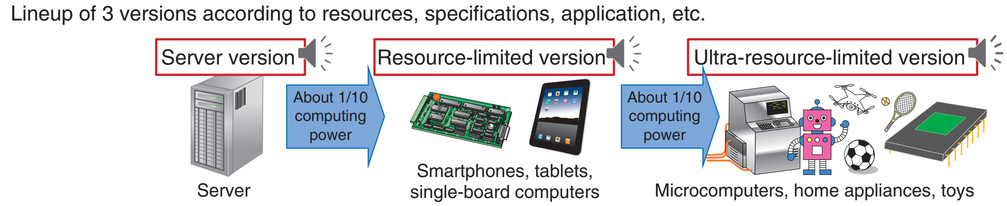
2.2 DNN speech-synthesis technology for reproducing diverse speaker characteristics at low cost

In contrast to the text-analysis section that is required to infer the reading and accents of input text with high accuracy, the speech-synthesis section is required to reproduce with high accuracy the voice of a desired speaker according to the wishes of the customer or other factors. However, achieving high-quality speech synthesis for a desired speaker requires a large amount of speech data uttered by that speaker. For example, in Cralinet [1], a unit-selection speech-synthesis system, up to 20 hours of speech data are needed to generate high-quality synthesized speech. As a result, achieving speech synthesis of a variety of speaker characteristics in an interactive speech system requires speech recordings and database (DB) construction, which can be a major issue due to the costs involved.

In response to this issue, we have achieved high-quality speech synthesis of a desired speaker from 20–30 minutes of speech data (about 2 hours of speech recordings) by using a previously constructed multi-speaker speech DB together with a DNN. A key feature of this system is that the speech data of multiple speakers can be modeled with a single DNN (**Fig. 3**). The information needed to generate speech (readings and accents) is learned from previously prepared multi-speaker speech data, and features of the desired speaker, such as voice quality and speaking style, are learned from the speech data of that desired speaker. Therefore, we have achieved high-quality speech synthesis even with a relatively small amount of speech data of the desired speaker [2]. Additionally, by combining this system with generative adversarial networks that have been found to be effective in image generation and other tasks, we have achieved improvements in the quality of synthesized speech and in the reproducibility of a speaker's voice [3].

2.3 DNN speech-synthesis technology for diverse environments

Depending on the environment in which speech synthesis is actually being used, a variety of constraints (e.g., no network access, need for high-speed response) require that the system be operated not on a computer server having abundant computing resources (central processing unit (CPU), read only memory (ROM), random access memory (RAM), etc.) but rather on devices such as smartphones or robots that are very limited in such resources. To address this issue, we developed an embedded type of DNN speech-synthesis library that can operate at practical speeds on devices with limited computing resources while maintaining the quality of synthesized speech to the extent possible. Specifically, we created a lineup of three types of libraries. In addition to the server-specific library, we developed a library for resource-limited terminals for operation on devices such as smartphones and tablets and a library for ultra-resource-limited terminals for devices such as microcomputers, home appliances, and high-grade toys that are greatly limited in computing resources


Lineup of 3 versions according to resources, specifications, application, etc.
Server version
Resource-limited version
Ultra-resource-limited version
About 1/10 computing power
About 1/10 computing power
Server
Smartphones, tablets, single-board computers
Microcomputers, home appliances, toys

Features	High-quality speech synthesis, multilingual support (Japanese, English, Chinese, Korean)	Lightweight speech-synthesis library enabling high-quality speech synthesis for smartphones, etc.	Ultra-lightweight speech-synthesis library enabling operation on low-end CPUs having only integer arithmetic
CPU	x86_64	ARM Cortex-A series	ARM9, MIPS1, etc.
Memory	2GB (ROM)/4GB (RAM)	About 10 MB (ROM/RAM)	Up to 10 MB
Applications	Content creation, robots, interactive speech, entertainment, information processing, and more	Robots requiring high-speed response, use in environments with no network access	Use on inexpensive, low-end devices

Fig. 4. Speech-synthesis engine operable on diverse devices.

(**Fig. 4**).

Many microcomputers and similar devices have no built-in floating-point unit (FPU), so the issue was how to speed up DNN-inference processing. In the library for ultra-resource-limited terminals, we used fixed-point arithmetic to achieve high-speed DNN-inference processing without using floating-point operations. Additionally, by incorporating speed-enhancement measures into the text-analysis section, the library for ultra-resource-limited terminals can operate at high speeds with small memory (ROM: 7 MB or greater) even on devices equipped with no FPU.

3. Application examples

3.1 Application to news-reading speech spoken by computer-generated (CG) announcer

FutureVoice Crayon [4] from NTT TechnoCross provides the text-to-speech-synthesis technology we developed as a service. It has been used since February 2020 for synthesizing the speech of TV Asahi's "AI×CG Announcer Hanasato Yuina" [5]. This is a news program, so there is a need for rich expressive power close to that of a human announcer as well as the ability to correctly read news copy in various categories. The word-reading-disambiguation technology described above automatically assigns readings that fit the category of that news, which helps to reduce the work involved in correcting readings and accents that up to now has been performed by humans.

Note that the voice of the CG announcer in this example took form by mixing the voices of several announcers at TV Asahi. This project therefore created an original voice independent of the rights of a specific person, which demonstrates new possibilities of speech-synthesis technology.

3.2 docomo AI Agent API

NTT DOCOMO's "docomo AI Agent API" [6] is an interactive-type AI-based application service provider service that packages a speech/text user interface (UI). This service incorporates our speech-synthesis technology as its speech-synthesis engine. The application programming interface (API) includes more than 50 types of voices as preset speakers, which enables a user to implement voice UIs that fit a variety of characters and environments without having to go through the trouble of recording speech or managing rights for each UI. The DNN speech-synthesis technology described above enables the reproduction of speaker characteristics with much variation ranging from the voices of elementary school students to those of the elderly.

3.3 Disaster Mitigation Communication System

NTT DATA's Disaster Mitigation Communication System [7] is an announcement broadcasting system for transmitting administrative and disaster-prevention information from local governments to residents. It delivers this information from a transmission system inside a government building or from remote operation terminals to outdoor speaker equipment installed within the community and to tablets, smartphones, mobile phones, and other devices. The information delivered in this manner is then used as a basis for conducting speech synthesis from the outdoor speaker equipment and each device, such as a tablet or house-specific receiver, then conveying that information to local residents with synthesized speech.

4. Future developments

This article described the recent technological developments in speech-synthesis technology at NTT Media Intelligence Laboratories and presented examples of its application. Thanks to these efforts, speech-synthesis technology has reached a certain level from the viewpoint of generating synthesized speech in the voice of a desired speaker.

Nevertheless, if we were to compare the output of current speech-synthesis technology with human speech, we would find that there are still significant differences. For example, when reading the news or a script in the case of a television announcer or voice actor, the goal would be to understand, for example, the intent included in the text and to then express that text in a manner of speaking that includes emotion. However, current speech-synthesis technology is incapable of interpreting intent and can only generate synthesized speech with the same tone. As expectations are growing for means of supporting and substituting human activities, the further spread of speech-synthesis technology in society will require speech synthesis that can achieve a level of expression equivalent to or better than that of humans. Going forward, we aim to expand the range of speech-synthesis applications by pursuing technologies that can enable expression based on context, intent, and emotion and expression that take into account the attributes and receptiveness of the listener.

References

[1] K. Mano, H. Mizuno, H. Nakajima, N. Miyazaki, and A. Yoshida, “Cralinet—Text-to-Speech System Providing Natural Voice Response to Customers,” NTT Technical Review, Vol. 5, No. 1, pp. 28–33, 2007. https://ntt-review.jp/archive/ntttechnical.php?contents=ntr200701028.pdf

[2] N. Hojo, Y. Ijima, and H. Mizuno, “DNN-based Speech Synthesis Using Speaker Codes,” IEICE Trans. Inf. and Syst., Vol. E101-D, No. 2, pp. 462–472, 2018.

[3] H. Kanagawa and Y. Ijima, “Multi-speaker Modeling for DNN-based Speech Synthesis Incorporating Generative Adversarial Networks,” Proc. of the 10th ISCA Speech Synthesis Workshop, pp. 40–44, Vienna, Austria, Sept. 2019.

[4] FutureVoice Crayon (in Japanese), https://www.futurevoice.jp/

[5] TV Asahi’s “AI×CG Announcer Hanasato Yuina, https://news.tv-asahi.co.jp/news_international/articles/000175834.html

[6] docomo AI Agent API (in Japanese), https://docs.sebastien.ai/

[7] Disaster Mitigation Communication System (in Japanese), https://www.nttdata.com/jp/ja/lineup/disaster_mitigation_c/

Yusuke Ijima

Distinguished Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

He received a B.E. in electric and electronics engineering from National Institution for Academic Degrees and University Evaluation from Yatsushiro National College of Technology, Kumamoto, in 2007, and an M.E. and Ph.D. in information processing from Tokyo Institute of Technology in 2009 and 2015. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2009, where he engaged in the research and development of speech synthesis. His research interests include speech synthesis, speech recognition, and speech analysis. He received the Awaya Prize Young Researcher Award of the Acoustical Society of Japan (ASJ) in 2018. He is a member of ASJ and the International Speech Communication Association (ISCA).

Nozomi Kobayashi

Senior Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

She received an M.E. and Dr.Eng. in information science from Nara Institute of Science and Technology in 2004 and 2007. She joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2007. Her current research interests include natural-language processing, especially information extraction. She is a member of the Information Processing Society of Japan (IPSJ) and the Association for Natural Language Processing (NLP).

Hiroko Yabushita

Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

She received a B.S. and M.S. in information science from Ochanomizu University, Tokyo, in 2008 and 2010. She joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2010. She engaged in the research of image recognition and received the IEEE Japan Council Women in Engineering Best Paper Award in 2013. Her current interest is in the digitization of human capabilities. She is a member of the Institute of Electronics, Information and Communication Engineers (IEICE).

Takashi Nakamura

Senior Research Engineer, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.

He received an M.E. in information and communication engineering from the University of Electro-Communications, Tokyo, in 2005 and joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) the same year. His interests include the practical development and business development of speech synthesis. He is a member ASJ.