Speech Synthesis Technology to Produce Diverse and Expressive Speech

Hideyuki Mizuno, Hideharu Nakajima, Yusuke Ijima, Hosana Kamiyama, and Hiroko Muto

Abstract

We have been developing a new text-to-speech synthesis system based on *user-design* speech synthesis technology that can be extensively applied to various fields. The technology yields speech with rich expression and various characteristics and thus replaces existing synthesized speech systems that have a limited range of voices or speaking styles. This article introduces this new system that represents the future of speech synthesis technology.

1. Introduction

The use of mobile phone telecommunication services is continuing to increase, and this is driving demand for various speech synthesis services, for example, speech guidance and speech dialogue services. For such services, speech synthesis must offer not only high quality but also variety. For example, synthesized speech that remains audible even in noisy environments and that has a characteristic voice quality and speaking style is required. The Cralinet (CReate A LIkeNEss to a Target speaker) system originally developed by the NTT Media Intelligence Laboratories as a telephone speech guidance service can generate high quality speech [1]. Cralinet has been broadly used in a safety confirmation system and in an automatic speech guidance system for business contact centers. The main feature of Cralinet is the production of synthesized speech that is as natural as that of humans. This was achieved by using a lot of speech waveforms uttered by a narrator and properly connecting them. Unfortunately, only one voice, that of a female speaker, is output, and the speaking style is limited to reading. Hereafter, the main aim of our research and development activities will be to introduce various new speech services that satisfy a far wider range of demands. The immediate goals are to generate any kind of voice or style while retaining the usability of speech even in noisy environments. In this article, we introduce the *user design* speech synthesis technology, which can generate expressive synthesized speech.

2. Outline of user-design speech synthesis technology

The framework of our technology is shown in **Fig. 1**. First, when the target speaker's voice is input, a source model is selected according to the voice quality of the speaker. The source model is then trained using the acoustic features of the voice. Next, the texts given as the speech synthesis target are analyzed to determine the most appropriate speaking style, e.g., a reading, storytelling, or sales pitch style. The resulting synthesized speech thus has the voice quality of the target speaker and also the appropriate speaking style. If the end-use environment is discovered to be noisy, the speech is enhanced to permit clear discernment.

The *user-design* speech synthesis technology comprises three novel techniques: automatic model training, speaking style assignment, and speech clarity enhancement. Some conventional text-to-speech technologies that provide similar functions have been

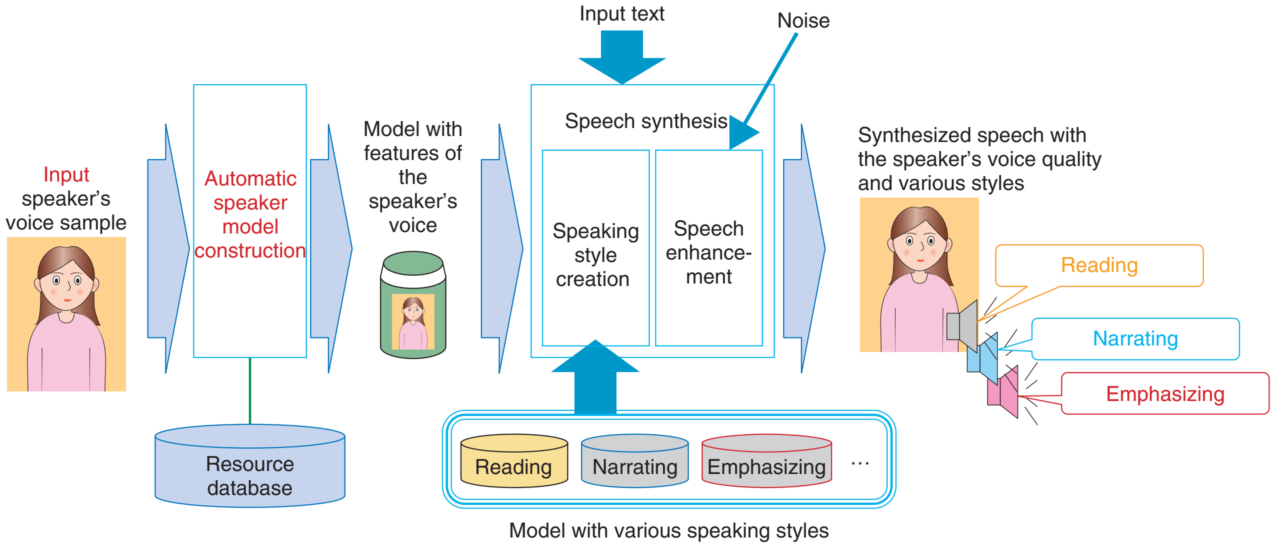
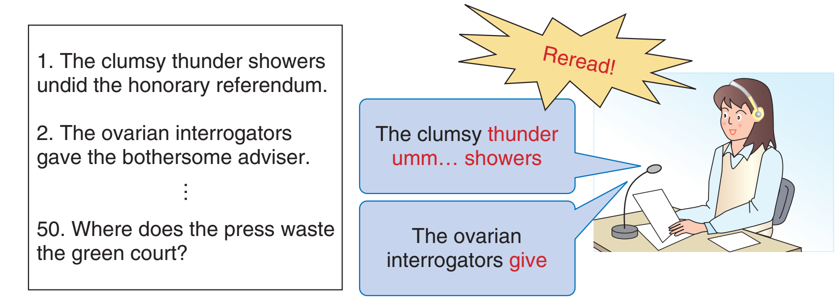

Input text
Noise
Input speaker's voice sample
Automatic speaker model construction
Resource database
Model with features of the speaker's voice
Speech synthesis
Speaking style creation
Speech enhance-ment
Synthesized speech with the speaker's voice quality and various styles
Reading
Narrating
Emphasizing
Reading
Narrating
Emphasizing
…
Model with various speaking styles

Fig. 1. Framework of user-design speech synthesis technology.


1. The clumsy thunder showers undid the honorary referendum.
2. The ovarian interrogators gave the bothersome adviser.
⋮
50. Where does the press waste the green court?
Reread!
The clumsy thunder umm… showers
The ovarian interrogators give

Example of text for speech recording [2]

Fig. 2. Example of the speech recording process for speech synthesis.

developed, but they have several problems. For instance, the time required to train the speaker model is excessive, speaking styles are limited, and speech clarity enhancement is effective only for a specific kind of noise. We apply our three new techniques to produce synthetic speech with rich expression and various characteristics and to realize speech synthesis with a voice similar to a user-specified speaker's voice based on very little speech data from the target speaker. Moreover, we can generate prosody, which refers to the rhythm, stress, and intonation of speech, in order to produce speech with various speaking styles, and we can enhance speech by using noise characteristics to set speech features and thus maintain high voice quality.

3. Synthesis of various speakers

Recent advances in speech synthesis technology mean that it is now possible to achieve reading out of various texts in a specified speaker's synthesized speech, but only if about one hour of speech data is uttered by that speaker. Moreover, as shown in **Fig. 2**, the desired speaker must utter the set text precisely word-for-word. Reading errors are common, so the same text must be reread until the samples are error-free. This is not a problem for professional narrators, but it is impractical for the general public (family members or friends). Clearly, the amount of recorded speech data required must be reduced. Our solution is called arbitrary speakers' speech synthesis. This solution can synthesize speech that sounds like any

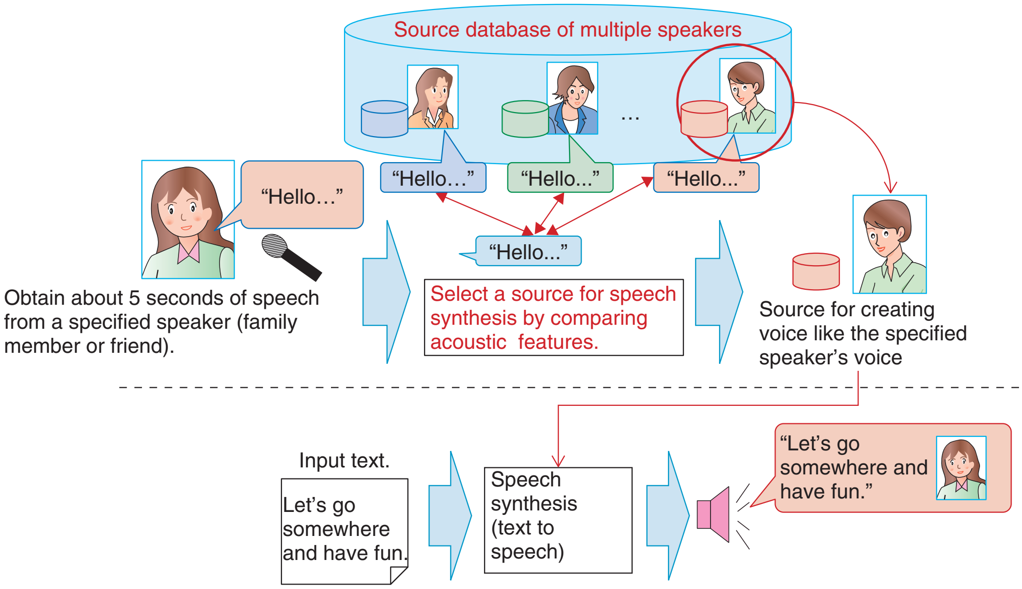

Source database of multiple speakers
…
"Hello…"
"Hello…"
"Hello…"
"Hello…"
"Hello…"
Obtain about 5 seconds of speech from a specified speaker (family member or friend).
Select a source for speech synthesis by comparing acoustic features.
Source for creating voice like the specified speaker's voice
Input text.
Let's go somewhere and have fun.
Speech synthesis (text to speech)
"Let's go somewhere and have fun."

Fig. 3. Overview of arbitrary speakers' speech synthesis technique.

particular speaker from just 5 sec of the speaker's speech data.

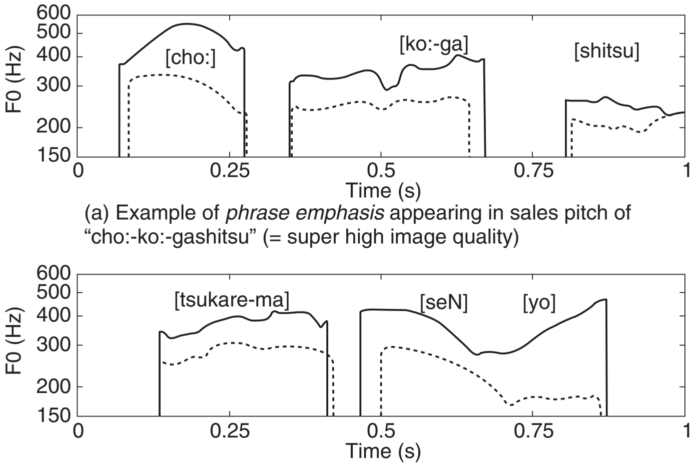
An overview of the technique is shown in **Fig. 3**. Samples of speech from multiple speakers are obtained in advance. These become the sound sources for speech synthesis that form the source database. By obtaining just a very small amount of a specified speaker's speech data, i.e., a speaker not in the database, and comparing the acoustic features of that data with the previously obtained samples, a sound source can be selected from the source database to create a voice very similar to the specified speaker's voice. This technique therefore makes it possible to synthesize speech that sounds like any particular speaker based on only 5 sec of the speaker's speech data. Experiments confirmed that 70% of the speakers selected from the database using this technique had a similar voice to the target speaker's voice.

4. Expressive speaking style

The style of speech depends on where the speech is uttered and what the intended purpose is, so adding a natural speaking style for various domains yields expressive synthesized speech. This kind of synthesis research is known as *expressive speech synthesis* and is being researched worldwide. Style can be expressed by three factors: i) intonation, ii) speed, and iii) loudness of speech. The style is determined by combining these three factors. Of these factors, intonation is known to be the most perceptible factor.

We recorded both conventional reading style speech and expressive style speech, and compared their intonations. We observed that 1) expressive speech had higher intonation than reading speech in many phrases when a phrase-by-phrase comparison of fundamental frequency (F0) was done, and 2) there are various F0 movements at phrase-end positions, for example, rise, fall, rise-fall, and rise-fall-rise. The first observation is described as *phrase emphasis*, and an example of higher F0 is shown in **Fig. 4(a)**. The second is called the *phrase boundary tone*. As shown in **Fig. 4(b)**, although the phrase boundary tone in reading style speech falls towards the phrase end, e.g., "Tsukare-masen-yo↘" (in English, "You may not be tired".), the tone in expressive style speech rises around the end of the phrase, e.g., "Tsukare-masen-yo↗ " to strongly emphasize the message.

These two phenomena of *phrase emphasis* and *phrase boundary tone* are found to be useful as F0 generation control factors when synthesizing speech with the hidden Markov model (HMM), which is commonly used in many studies [3]. To achieve expressive text-to-speech synthesis (which takes text as input and generates synthesized speech as output), we investigated a method for predicting whether or not the phrase boundary tone rises at each phrase end [4]. For this prediction, it is not sufficient to know the



(a) Example of *phrase emphasis* appearing in sales pitch of "cho:-ko:-gashitsu" (= super high image quality)

(b) Example of *phrase boundary* tone appearing in sales pitch of "tsukare-ma-seN-yo" (= You may not be tired.)

Fig. 4. Fundamental frequency (F0) difference between reading and expressive style speech. Solid line: expressive, dashed line: reading.

identities of the phrase-end particles. Phrase boundary tones change with the context surrounding the phrase boundary and the situation in which the speech is uttered e.g., "A-ka↗ (asking question)" vs. "A -ka↘ (with disappointment)". Though some phrase boundary rise/fall prediction rules can be written by human experts, the variation is too large due to speaker individuality and the diversity of domains and situations. Thus, we use speech/linguistic data and a machine learning method to construct models to predict phrase boundary rise/fall. Through experiments targeting expressive speech such as sales pitches and telephone call center operation conversations, we confirmed that the proposed method can accurately predict phrase boundary rise/fall labels.

5. Enhanced synthesized speech

To apply synthesized speech to a wide variety of speech services, the synthesized speech must not only be expressive but also intelligible. In noisy environments, speech can be hard to follow unless some form of noise cancellation is used. Hence, we have been developing a technique to enhance synthesized speech. It yields synthesized speech that remains discernible while retaining as much of the distinctive characteristics of a speaker's voice and speaking style as possible. As the first step, we analyzed the attributes of easily discerned speech in noisy environments. It is well known that some speakers have voices that carry exceptionally well; they cut through noise and are easily heard.

We investigated the attributes of such *carrying voices* using many speakers and several types of noise. The experiments revealed that the carrying voices had a higher power spectrum in specific frequency bands occupied only by vowel sounds (which are produced by vibrating the vocal cords) than the noise [5]. As the next step, we developed a technique that uses the results of analysis to reproduce the carrying voice without changing the unique characteristics of the speaker's voice. This is done by accentuating the power spectra of the specific frequency bands so that they dominate the identical frequency bands of the noise.

The direct enhancement of power causes unexpected and unwanted changes in voice quality. Therefore, our enhancement algorithm first identifies the vowel parts of the synthesized speech from pronunciation information generated in the speech synthesis process. Next, the power spectra of the frequency bands are increased. It is difficult to precisely determine the frequency bands from actual speech in real time because the bands vary with the pronunciation. However, with speech synthesis they can be accurately determined prior to their use by analyzing the speech source. Therefore, synthesized speech that is both intelligible and high in quality can be achieved by using both pronunciation information and the frequency characteristics at specific bands, as shown in

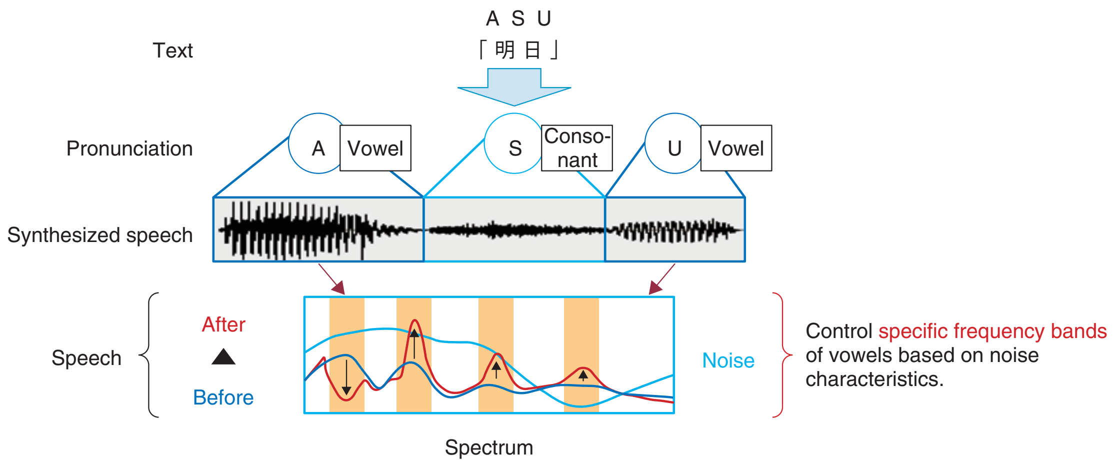


Fig. 5. Process of clarifying speech synthesis result.

Fig. 5. The result is conductive to an increase from 50–60% to 80% in word discernment. This result indicates that the technique significantly increases the appeal and utility of synthesized speech.

6. Conclusion

The techniques introduced in this article are able to yield expressive synthetic speech with high voice quality and various speaking styles and that offers excellent clarity even in noisy environments. With these techniques, the range of applications of speech synthesis will expand greatly from the conventional applications, which have been restricted by the limited variety of speech and use environments. When the expressive speech synthesis technique described here is refined, the usage of speech synthesis will expand to encompass speech dialogue systems that can talk with various voice qualities and speaking styles in accordance with user requests. Optimizing the techniques introduced in this article is our immediate goal.

References

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

[2] A. W. Black and K. Tokuda, “The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets,” Proc. of Interspeech 2005, pp. 77–80, Lisbon, Portugal, 2005.

[3] Y. Maeno, T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, “A study on prosodic contextual factors for HMM-based speech synthesis with diverse speaking styles,” Proc. of the Acoustic Society of Japan 2011 Spring Meeting, pp. 385–386, 2011 (in Japanese).

[4] H. Nakajima and H. Mizuno, “Predicting phrase boundary tone labels for expressive text-to-speech synthesis,” Proc. of the Acoustic Society of Japan Autumn Meeting, pp. 361–362, 2011 (in Japanese).

[5] H. Kamiyama, Y. Ijima, M. Isogai, and H. Mizuno, “Analysis of the correlation between various acoustic features and the audibility of speech with noise,” IEICE Technical Report, Vol. 122, No. 81, pp. 69–74, 2012 (in Japanese).

Hideyuki Mizuno

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

He received the B.E. and M.E. degrees from Nagoya University, Aichi, in 1986 and 1988 and the Dr.Eng. degree in systems and information engineering from the University of Tsukuba, Ibaraki, in 2006. In 1988, he joined NTT Human Interface Laboratories, where he engaged in R&D of speech synthesis and voice quality conversion. During 1994–1997, he worked for NTT Intelligent Technology Co. Ltd. developing speech application systems. He is currently researching text-to-speech synthesis and its applications. Since 2009, he has been an Associate Editor of the Editorial Board of the Acoustic Society of Japan (ASJ). Since 2010, he has been a chair of the Speech Synthesis Group of the Technical Standardization Committee on Speech Input/Output Systems in the Japan Electronics and Information Technology Industries Association, Japan. He received the Technical Development Award from ASJ in 1998. He is a member of ASJ and the Institute of Electronics, Information and Communication Engineers (IEICE).

Hideharu Nakajima

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

He received the B.E. and M.E. degrees in computer engineering and information science from the University of Tokushima in 1990 and 1992 and the Ph.D. degree in global information and telecommunication studies from Waseda University, Tokyo, in 2010. He joined NTT Information Processing Laboratories in 1992. During 1997–2002, he worked for Advanced Telecommunications Research Institute International (ATR). His research interests include spoken/natural language processing based on clear principles and he is now investigating corpus-based text-processing for speech synthesis. He is a member of ASJ, the Phonetic Society of Japan, the Association for Natural Language Processing (NLP), IEICE, the Information Processing Society of Japan, and the Japanese Cognitive Science Society.

Yusuke Ijima

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

He received the B.E. degree in electrical and electronics engineering from the National Institution for Academic Degrees and University Evaluation after graduating from Yatsushiro National College of Technology, Kumamoto, in 2007, and the M.E. degree in information processing from Tokyo Institute of Technology in 2009. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2009 and began researching speech synthesis. His research interests include speech synthesis, speech recognition, and speech analysis. He is a member of ASJ and the International Speech Communication Association.

Hosana Kamiyama

NTT EAST.

He received the B.E. and M.E. degrees in computer science from Tokyo Institute of Technology in 2008 and 2010, respectively. Since joining NTT in 2010, he had been engaged in researching synthesized speech enhancement. He is a member of ASJ and IEICE. He moved to NTT EAST in July 2013. At the time of this research was conducted, he was Researcher, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

Hiroko Muto

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

She received the B.E. and M.E. degrees in computer science from Tokyo Institute of Technology in 2009 and 2011, respectively. Since joining NTT in 2011, she has been engaged in researching text processing for speech synthesis. She is a member of ASJ.