Corpus-Based Text-to-Speech and Its Application

Satoshi Watanabe　Takeshi Iwaki
Tsutomu Kaneyasu　Kei Miki

At Oki Electric we believe it is important to depict an image of our future society as one of an “e-Society®[*1]” and to be able to provide “desired information in a preferred format” to all people. Furthermore, we intend to provide a rich communication environment that can surpass the mere transmission of information and convey emotions and sympathy.

The voice is one of the most basic modes of communication. We are, therefore, engaged in activities to achieve sophisticated expression through speech synthesis, in order to provide a richer voice communication environment.

Speech synthesis technologies

The development of speech synthesis technologies has been ongoing for a long time. Speech synthesis technologies can be broadly categorized into a slot filling method and rule-based synthesis method (**Table 1**). The slot filling method is a technology that accumulates human speech (natural speech) itself in the database of words and phrases in reusable units and generates synthesized speech by properlyconcatenating the units. It was possible to obtain highly natural synthesized speech, however, it was limited to the available voice message to be synthesized, due to its mechanism. Furthermore, this method also had constraints for adding new words or expressions, as it was necessary to enlist the same speaker using an identical vocalizing style in order to add to the voice recording. The slot filling method has been implemented widely for systems that can specify voice messages to particular strings in advance.

The rule-based synthesis method is a technology that generates synthesized speech by editing the voice waveform segment data and varying it for intonations and such according to synthesis rules established beforehand. Voice recordings are not necessary when expanding with new words or expressions, since it is intended for synthesizing arbitrary texts. In the past, experts analyzed voice waveforms and used such information to determine the formulation of synthesis rules or the preparation of voice waveform segment data. Constraints, such as the limitation of expressive capabilities, can be described by rules or an inability to avoid deterioration of the voice waveforms due to signal processing performed to prepare or edit voice waveform segments. As a result, synthesized speech lacks in tone variations (expressive capabilities) and humanness (natural sounding speech). It is believed that the aforementioned sound quality issues are part of the reasons why the rule-based synthesis has not reached the level of practical implementation initially expected.

Table 1　Categories of speech synthesis technologies

	Unit of synthesis	Subject of synthesis	Sound quality of synthesis	Example of applications
Slot filling method	Words and phrases, etc.	Limited sentences	Equivalent to natural voice	Public announcements at stations Stock price announcements
Rule-based synthesis method (conventional)	Phonemes, etc.	Arbitrary sentences	Smooth but mechanical	Reading emails out loud Screen reader for personal computers
Rule-based synthesis method (corpus-based)	Phonemes, etc.	Arbitrary sentences	Highly natural-sounding speech	Speech portals Personified agent

In the 1990s[1], amidst such developments, research of the Corpus-based Text-to-Speech (hereinafter referred to as “CTTS”) technology started and is currently the mainstream[2] synthesis method, of all the rule-based synthesis methods, with which the algorithms and database are determined by a statistical approach based on a large-scale speech database (speech corpus) . There are high expectations for CTTS as a technology able to synthesize arbitrary text strings while maintaining expression capabilities and a natural-sounding speech that can rival the slot filling method. Although the problems relating to calculation and memory capacities were initially considered issues, these have been resolved as a result of advancements in information devices and currently this synthesis method is in the practical implementation stage.

Furthermore, although a definitive definition of the CTTS technology is unclear[3], for the purpose of this paper it is defined as a “method that obtains synthesized speech by directly connecting, in phonemic element units, a large quantity of voice waveforms that have been accumulated beforehand”.

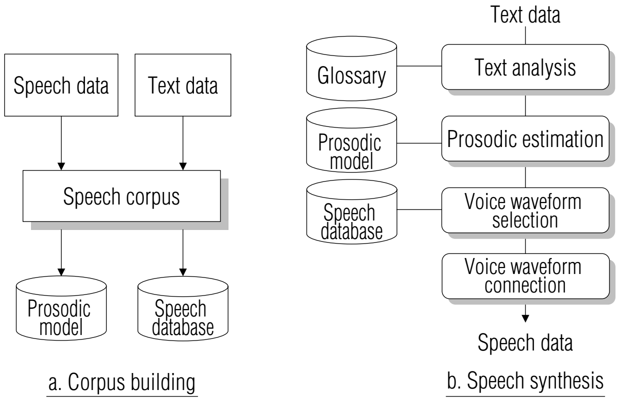
Features of Corpus-based Text-to-Speech Technology

(1) Outline of CTTS technologies

Technologies relating to CTTS can be broadly categorized into a speech corpus building technology (**Figure 1a**) and speech synthesis technology (**Figure 1b**).

The speech corpus building technology is used to prepare prosodic models and a speech database necessary for speech synthesis by building a speech corpus from vast amounts of speech data. During the

[*1] e-Society is a registered trademark of Oki Electric Industry Co., Ltd.



a. Corpus building

b. Speech synthesis

Fig. 1 Outline of CTTS processes

building of the speech corpus the timing of text data is aligned with speech data collected previously and tens of minutes to several tens of hours, on various levels, such as phonemes, prosodics and morphemes, are required before individual parameters are extracted and statistical data is calculated. Prosodic models and speech databases are subsequently created by implementing procedures objectively stipulated, such as the learning of statistical models, to the built speech corpus. Prosodic models are used for determining the voice tone segment ofsynthesized speech, such as the intonation forsynthesized speech, length of time sustained by each sound, as well as the length of a pause. The speech database is an accumulation of speech data, which is actually directly connected during synthesis, in a searchable format.

The speech synthesis technology actually converts text data into speech data. Based on the results of an analysis on entered text the prosodic information is estimated using prosodic models prepared using the corpus building technology. Then phoneme and voice waveform pairs that best match the estimated prosodic information are selected from the speech database and synthesized speech is obtained by connecting these pairs.

(2) Sound quality features of CTTS

A highly natural-sounding speech, in comparison with conventional rule-based synthesis methods, is obtained with the CTTS technology, since the voice waveform of natural speech is directly connected in phonemic units to obtain synthesized speech. Even though synthesized speech obtained using conventional rule-based synthesis methods did not trail far behind naturalspeech, in terms of comprehensiveness (can whatever is being said be understood?), it is not possible to state that it was widely accepted. It is believed that this is heavily influenced by the mechanical sound quality unique to synthesized speech. If we were able to assure an highly natural-sounding speech, then synthesized speech would be accepted for applications on a much wider scale.

The ability to efficiently realize speech synthesis with personality or tone expressions is also a sound quality feature of the CTTS technology. Since it was necessary to determine synthesis rules for specific individuals, particular tones or expert analysis for audio technology specialists to produce individual voice waveform segment data, it was not realistic to consider developing speech synthesis for expressing various personalities or tones with the conventional rule-based synthesis method. However, it is possible to systematically develop speech synthesis that can express the personalities of specific persons or particular tones with CTTS by building a speech corpus using strategically collected speech data. Personalities and tones are major aspects of speech media, which enable services using voices that suit a particular overall image for services or to convey emotions and intentions through the manipulation of subtle differences in nuance.

Trends of CTTS technology development

(1) Improving naturalness

The naturalness of synthesized speech has demonstrated marked improvements since it has become possible to handle adequate amounts of speech corpus. Some achieve a sound quality equivalent to that of a natural speech by limiting the scope of applications to the reading of particular subjects (such as weather forecasts or ticket bookings). Furthermore, in some circumstances natural speech is realized by specializing in voices for reading news only[4)]. Since no discomfort arises from patching together fixed-pattern voice messages, accumulated as a recorded voice beforehand, some composite systems are emerging that use CTTS only for the synthesis of personal names and place names.

The sound quality is still not adequate in comparison with naturalspeech, however, when the scope of reading is expanded to include arbitrary sentences. Also, when the smoothing process is performed in order to eliminate discontinuity, which arises when voice waveforms are connected, deterioration to the naturalness occurs. In order to solve such issues activities leading to the efficient building of a large-scale corpus[5)] for obtaining very high naturalness, even for arbitrary sentences, the development of waveform connecting methods[6)], with minimal deterioration in naturalness as well as methods for applying the estimated prosodic to elements in a speech database, are gradually yielding results.

(2) Personality

Systems that synthesize speech with tones of particular individuals are being developed in order to reproduce the personality of individuals by building a corpus with the voice of a specific television personality or announcer. The technology has reached a level of ability to adequately express with a voice that can be identified as the voice of a particular person, even if the subject is an ordinary person who is not familiar with voice recording, as long as the speech data is made with favorable recording conditions without noise, etc.

In order to reproduce a personality with a high quality synthesized speech, however, it is necessary to secure an adequate amount of speech data with a high sound quality. The problem is that it is not always easy to do so. To deal with this problem research and development is being conducted with a speaker exchange technology[7)], which converts the sound quality of speech synthesized with an existing speech corpus by using parameters for individuals that have been obtained from small amounts of speech data of particular individuals.

(3) Emotional expressions

The majority of conventional CTTS was built using the corpus of speech in narrative mode and read by announcers, which is known as “narrative voice”. However, systems developed recently use an individual corpus built for each emotion with a voice that is vocalized with specific emotions (“joy”, “sadness”, etc.), making it possible to synthesize speech expressing emotions to a certain degree by switching the corpus.

When considering even more detailed emotional expressions, however, it becomes necessary to design the corpus based on detailed analysis, which includes various types of emotions and the extent of necessary emotions, interrelationships between emotions, etc. A lot of research is being done in order to deal with such issues.

Anticipated applications

With the realization of CTTS, which has more expressive capabilities and higher naturalness through the resolution of the aforementioned issues, new applications become available.

- **Voice guidance**

Progressively more voice guidance functions are being incorporated into public instruments to improve usability. In order to make such systems easier to use voice messages are required that can provide detailed responses to users and conditions of use at a low cost. Most of the current voice guidance systems are either synthesized by the slot filling method or through actual voice recordings. For this reason, in order to add or change voice messages, the issue of cost arises relating to re-recording by the same announcer. CTTS can dramatically reduce production costs for speech data and make it possible to add, replenish or change a large number of voice messages, thereby contributing to the dissemination of voice guidance that is easy to use. By using a networked CTTS batch updates of voice messages can be performed in real-time (**Figure 2**).

- **Interface for portable information devices**

Due to the popularization of mobile phones and devices a growing need to enjoy digital content outside the home or office has emerged. Devices used in such environments typically have inadequate space to facilitate display functions, thus audio output becomes a dominant means for the presentation of information. Voice content that is relatively long in terms of time, such as vocalized digital books or digital newspapers, previously were considered unsuitable for conventional rule-based synthesis as they lacked in naturalness and the tone of voice was monotonous. However, using the CTTS technology, with its naturalness and high level of expressive capabilities, such content can be enjoyed while commuting on trains or driving a car.

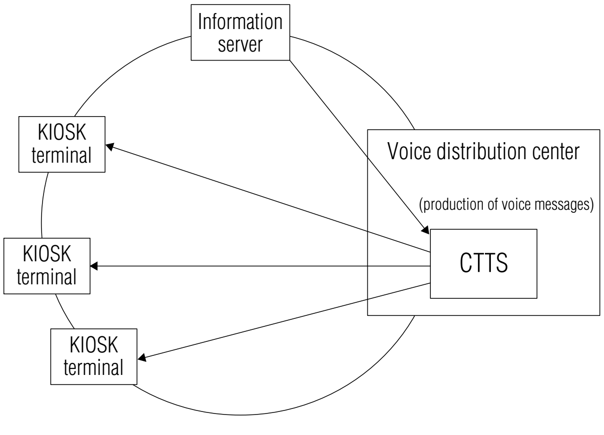


Fig. 2 An example of batch updating of voice messages

- **Application for welfare instruments**

Due to the advancement of information technologies a new product is about to be created in the field of welfare instruments. Since a diverse range of voices are available synthesized with CTTS, it is possible to provide the reading of e-mails or web pages with voices that offer more of an affinity to persons with a visual impairment. Also, more voice books and vocalized educational materials could be made available at libraries. Furthermore, voice boxes could be provided that would be useful for expressing the personality of vocally impaired users. Speech synthesis systems[8] for persons who have had their pharynx extracted have already been developed with expectations for the development of systems in numerous applications.

Issues surrounding practical implementation

Oki Electric has been researching and developing rule-based synthesis methods for many years, providing various products around the world, while engaging in activities to popularize the application of such products[9],[10]. Other issues, besides the technical items mentioned in the section “Trends of CTTS technology development”, need to be addressed in order to implement speech synthesis in a practical manner. Issues surrounding the practical implementation of CTTS are indicated below.

Furthermore, since the sound quality required varies greatly with the CTTS technology, depending on the application, it is essential for the technology to be developed with a clear understanding of the intended application. At Oki Electric, we are promoting validation tests for services that use CTTS, in collaboration with service and content providers, while clarifying the voice attributes and standards required for individual services as we proceed with technology developments for CTTS.

(1) Synthesized sound quality designing technologies

With the current CTTS it is difficult to make an estimate, such as “a speech corpus built with this many recorded speeches is needed to synthesize speeches with a quality on this particular level”. Even though the general guidelines based on the coverage[11] in units of synthesis and amount of speech data are about to be clarified, they are still inadequate for practical implementation with the assumption that the applications are for business purposes. In order to determine whether or not a desired speech corpus can be provided at a reasonable cost for business purposes, it is necessary to formulate a practical framework that can quantify relationships, such as the amount of speech data and content to be recorded, variations to texts that are subject to speech synthesis, sound quality required for synthesized speech, etc.

As for the evaluation of synthesized speech quality in particular, evaluation methods based on subjective evaluations relating to clarity and comprehensiveness are being developed at the present time[12)], but there is a need for the development of a new sound quality evaluation index that can respond to the high level of expressive capabilities of the CTTS technology.

(2) Interaction design technology

Among the systems that use the current speech synthesis quite a few have no consideration for the characteristics of speech media, such as redundancy, volatility or promptness, which are aspects inhibiting the popularization of speech synthesis as a result. We can hope to improve the practicality of speech synthesis for numerous speech synthesis applications by carefully designing interaction into the overall system while making considerations for the attributes of media and functions to take advantage of such attributes. Such a design for interaction is particularly important in order to bring out the highly expressive capabilities of the CTTS technology.

For example, speech instruments with highly expressive capabilities can only be realized by incorporating text entry, speech start and stop, as well as emotional speech control and similar functions as proper GUIs for speech instruments for which the CTTS technology has been applied. Similarly, in order to realize multimedia information terminals that offer a pleasant feeling when operated, it is also necessary to design a proper distribution method for the display information, as well as a synchronization method for the timing. In order to create attractive digital novel reading services it is also necessary to provide functions for assigning the optimum emotional information or pace of speech to individual segments of text.

(3) Rights issues

Once the popularization of the CTTS technology gets under way problems regarding various rights relating to speech corpus can be anticipated. They include the rights to the speech corpus, including recorded speeches, as well as the rights to prosodic models and speech databases, the rights to speech synthesis systems that incorporate such items, along with the rights to synthesized speech data (speech files) generated as a result of the speech synthesis. In order to proceed aggressively with practical implementation, guidelines for the contracts of these rights are called for.

Furthermore, the CTTS technology may be used for malicious purposes. For example, if CTTS is used for a particular individual to produce and distribute speech data containing details that are immoral or for the purpose of fraudulent impersonation (such as telebanking fraud, etc.). It is necessary to reach a social consensus on the use of CTTS, along with considerations for technical solutions, such as electronic watermarks or personal authentication.

Conclusion

An introductory outline of the CTTS technology was provided, along with descriptions on issues relating to anticipated applications and practical implementation. Naturalness, personality reproduction and emotional expressions, which are features of CTTS, have been realized to a certain degree and CTTS is entering a phase of practical implementation in search of applicable domains as well as for the purpose of business use.

At Oki Electric, we intend to keep on promoting practical implementation of CTTS at business levels by extracting and resolving new service issues early on through joint validation tests performed in collaboration with various interested parties.

References

1) N. Campbell and A. Black: “CHATR: Natural Speech Waveform Connecting Type Arbitrary Voice Synthesis System”, Shingaku Giho, SP96-7, May 1996.

2) Hirose: “Flexible Speech Synthesis”, Partner Robot Material Corpus (NTS), pp. 58-85, December 2005.

3) Kawai et al: “[Tutorial Lectures] Building Large-scale Speech Corpus for Speech Synthesis”, Shingaku Giho, SP2005-9, May 2005.

4) Yogi et al: “A Study on Voice Waveform Connecting Type Speech Synthesis that Variable Length Phonemic Environment Dependent Handle Phonemic Strings”, Onkoronshu (Sound Lecture Collection), 1-8-9, September 2003.

5) Hirai et al: “Corpus Based Speech Synthesis System XIMERA”, Shingaku Giho, SP2005-18, May 2005.

6) Mizuno et al: “Speech Synthesis from Text by Corpus Based Approach”, NTT Journal, pp. 23-26, January 2004.

7) Isogai et al: “Study on Model Adaptations and Adaptive Learning Algorithms for Diversified Speech Synthesis”, Shingaku Giho, SP2005-50, August 2005.

8) Kimura: “Fusion of Medicine and Engineering in Otolaryngological Domain: Speech Synthesis”, Jibika Sosetsu, 2003.

9) Yazu et al: “Text to Speech Conversion Technology and Applications”, Oki Technical Review, Issue 181, Vol. 66, No. 2, pp. 59-62, October 1999.

10) Watanabe et al: “Designing and Prototyping Speech Synthesis for 'Use with TV or Radio'”, Collection of Papers, Human Interface Symposium 2003, pp. 467-470, September 2003.

11) Kawai et al: “Trends of Corpus Based Speech Synthesis Technology [III] - Design and Evaluation Standards of Corpus -“, Shingakushi, Vol. 87, No. 3, March 2004.

12) “Performance Evaluation Method Guidelines for Speech Synthesis Systems”, JEITA-IT-4001, March 2004.

Authors

Satoshi Watanabe: Corporate Research & Development Center, Human Interface Lab., Team Leader.

Takeshi Iwaki: Corporate Research & Development Center, Human Interface Lab.

Tsutomu Kaneyasu: Corporate Research & Development Center, Human Interface Lab.

Kei Miki: Corporate Research & Development Center, Human Interface Lab., Manager.