Cralinet—Text-To-Speech System Providing Natural Voice Responses to Customers

Kazunori Mano†*, Hideyuki Mizuno, Hideharu Nakajima, Noboru Miyazaki, and Akihiro Yoshida*

Abstract

We have developed a text-to-speech synthesis system called "Cralinet" that generates very natural-sounding voices that can be used by various voice response services. This article describes effective instances of voice response services at a contact center and explains a statistical technique for estimating the accents of personal names and an improved technique for generating correct intonations based on statistical intonation evaluation as fundamental techniques of speech synthesis for contact centers. Finally, the future prospects of speech synthesis technology are also discussed.

1. Introduction

Text-to-speech synthesis converts information in written form into artificial speech. NTT has developed a text-to-speech (TTS) system called "Cralinet*" that utilizes a corpus-based approach [1]. This approach uses a very large-scale dictionary of natural speech data to provide the customer with information spoken in natural-sounding synthesized voices. The use of text-to-speech synthesis is highly expected to improve the quality of services at contact centers and other services using interactive voice response (IVR) systems. Some effective instances of voice response services at a contact center are described in section 2. Section 3 overviews the Cralinet system and then explains two novel techniques: a statistical method of estimating the accents of unknown personal names and an improved technique of generating correct intonations based on an intonation evaluation process. Future prospects for speech synthesis in various services are discussed in section 4.

2. Application to contact centers

A typical service area in which speech is important is a consumer service provided by the telephone contact center of a company. In conventional contact centers in Japan, most of the customers' calls are handled by human operators. IVR systems for handling incoming calls, such as by performing automatic call distribution based on the kind of inquiry, are only partially used. Even in this case, a TTS system is not always utilized; instead, pre-recorded human voices are used for automatic responses. One of the main reasons is that the synthesized speech generated by a conventional TTS system is still very artificial-sounding and messages are hard to understand compared with the natural speech of a human operator. The quality of the synthesized speech is not good enough for use in contact centers. As one solution to these problems, we have developed a new text-to-speech system with the development codename Cralinet. Because this system provides natural-sounding synthesized speech that is comparable to pre-recorded human speech, it has been introduced at several contact centers.

One contact center where Cralinet has been introduced provides a ticket sales service. Since its introduction, the operational efficiency of the telephone service has shown a statistically significant increase. The contact center explains about tickets and sells

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

* "Cralinet" stands for "CReate A LIke NEss to a Target speaker".

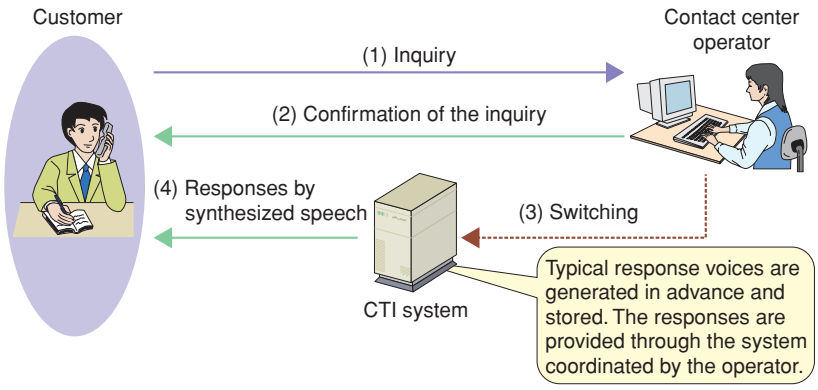


Fig. 1. Voice response system coordinated by the operator.

them, answering many kinds of inquiries about various events. One of the main problems in the service is that operators spend too much working time responding to common inquiries, which leaves less time for more important operations such as ticket sales.

A voice response scheme using a computer-telephony integration (CTI) system coordinated with a human operator is shown in **Fig. 1**. Typical response speeches for expected inquiries are synthesized in advance and stored in the CTI system. When a call arrives, the operator first determines the customer's inquiry. If the inquiry is very typical and matches one of the stored responses, then after obtaining permission from the customer, the operator switches the call over to the CTI system. Detailed information relevant to the inquiry is explained by a synthesized voice generated by the machine and the operator moves on to the next call. Consequently, the time that the operator spends on typical common inquiries can be reduced and total number of calls accepted can be increased.

Systems based on cooperation between operators and speech synthesis systems can be applied to other reception-type services for standard inquiries such as ones about account activity, balances, and customer bonus points in a bank, credit card company, and mail-order store as well as to ticket services. In addition, when an unusually large number of calls suddenly arrive for some reason, such as after an accident, the customer waiting time and the number of calls abandoned because callers gave up waiting both increase and these are seen by contact centers as serious problems that must be solved from the viewpoint of customer services. Customers are happier if they are informed of the expected waiting time until connection or told about the current status of receptionists, and this can be done using synthesized speech. Other advantages of the speech synthesis system for daily contact center operations include always having prepared speech available regardless of the operators' availability and always having stable voice quality, whereas operators' utterances vary with workload, heath conditions, and other factors. These efficient operations and improvements in customer satisfaction can be achieved in future contact centers by fully incorporating speech synthesis technology.

3. Cralinet

3.1 Overview of the synthesis system

When a speech synthesis system is applied to contact center services, besides the quality of the synthesized speech, the correct pronunciation (reading) of customers' names and addresses, numbers, keywords, and brand names is very important. Cralinet estimates correct readings and accents in its text processing part and synthesizes natural intonations. The flow of the speech synthesizer, illustrating how the concatenated speech signal is generated, is shown in **Fig. 2**. Cralinet consists of a text processing part, a speech synthesis part, and a large speech database. The text processing part contains functions for estimating unknown Japanese readings and accents. The speech synthesis part calculates the best combinations of speech waveforms considering both the smooth connection of waveform units and the correctness of the intonation pattern over whole phrases.

3.2 Text processing part

Customers' names often appear at the beginning of synthesized messages: "Mr. (or Ms.) ... , your current usage details are ...". Contact centers usually have customer information databases that include customers' names written in Chinese characters

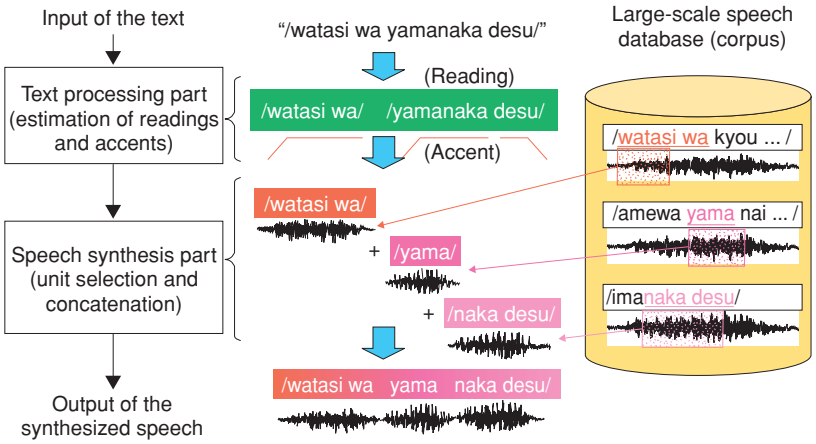
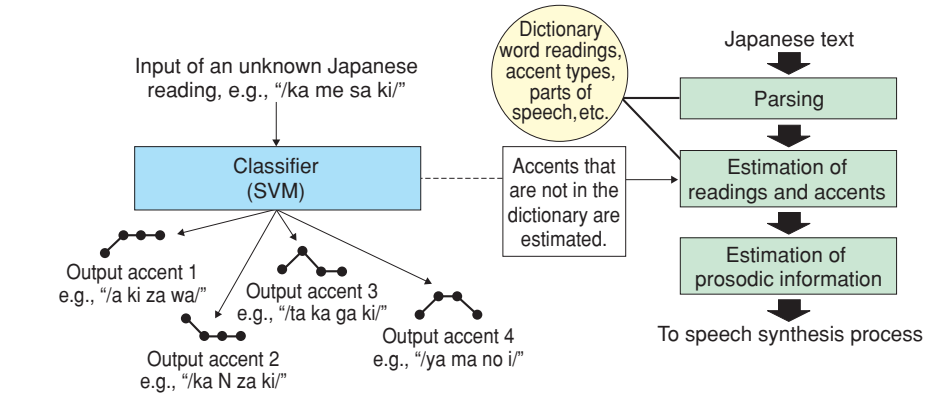


Fig. 2. Flow of Cralinet's text-to-speech synthesis.



(a) Estimation of accents from unknown Japanese readings (b) Text analysis and processing

Fig. 3. Text processing part.

(graphemes) and their pronunciation written in Japanese kana (phonemes) in addition to telephone numbers, addresses, and a list of purchased goods, etc. Although name accents are important for correct speech synthesis, they are usually not recorded in the databases. Thus, the text-to-speech system should estimate the accents from the readings of the names.

An accent is described with binary symbols of voice pitch (high (H) or low (L)). In the case of Japanese Tokyo Standard, for example, when a person's name consists of four morae, the number of pitch pattern candidates (that is, the number of accent type candidates) is the same as the number of morae of the name, i.e., four. The reading /ka me sa ki/ in **Fig. 3(a)** can have four accent types: LHHH, HLLL, LHLL, and LHHL. Conventional accent estimation systems can only give the most frequent accent type or use handcrafted accent rules. However, the most frequent accent type is not always correct. Though handcrafted rules can be revised to stay up to date with new names (and new words), such revisions are expensive because of the complicated modification process, which keeps consistency between old and new rules.

The accent estimation problem is to classify a word whose accent type is unknown into a word group of the same accent type. Cralinet uses a classifier that is automatically constructed by a statistical machine learning method (support vector machine* (SVM)). In Fig. 3(a), the classifier takes the reading of the word whose accent is unknown as an input to calculate similarities between it and representative readings of each accent type group and selects the accent type with the highest similarity. A conventional clas-

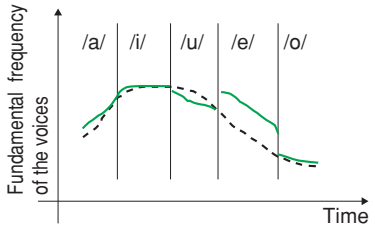
* Support vector machine is a supervised learning method for pattern classification and regression analysis. It finds a hyper-plane that separates two kinds of data with the maximum geometric margin. It provides state-of-the-art performances in real-world applications such as text categorization, handwritten character recognition, and genomic analysis.

sifier, the decision tree, achieves accent estimation accuracies of 82% for family names and 79% for first names. Our system, Cralinet, achieves higher accuracies of 91% and 86%, respectively.

In **Fig. 3(b)**, the text processing part first divides the input sentence into words and gives lexical information such as readings, accents, and parts of speech to the words by using the TTS system dictionary. However, since the number of personal names is usually large, the dictionary cannot store all of them. The accent types of personal names not in the dictionary are estimated here by the above-mentioned statistical method. Next, using this lexical information and estimated accents information of each word, the text processing part determines prosodic information such as phrasal accents, intonations, and pauses in the sentence, and this information is sent to the speech synthesis processing part.

3.3 Speech synthesis processing part

In the speech synthesis processing part, speech units in the speech corpus are selected to match the reading of the input text and its prosodic information, which shows the target fundamental frequency pattern and phoneme duration of the synthesized speech. The synthesized speech consists of concatenated sequences of selected speech units. The larger the speech corpus is, the more variations of the speech units there are. If we can use a huge speech corpus, then the quality of the synthesized speech is expected to be better than when we use a small corpus. However, even if the speech corpus contains tens of hours of speech, the quality of the synthesized speech generated by a conventional unit selection algorithm is occasionally low. This is because even if a chosen speech unit is suitable for the local prosodic target, the concatenated speech does not always output correct and natural intonation over the whole sentence. Careful investigation of the synthesized speech quality has shown that the main cause of the deterioration is unnatural intonation, as shown in **Fig. 4**.



Black broken line: Target intonation pattern.
Green solid line: Intonation of the selected speech units.

In this case, although each speech unit is selected as the most suitable for the local target pattern, the intonation of the concatenated units is unnatural.

Fig. 4. Example of unnatural intonation of concatenated units relative to the target pattern.

To overcome this artifact, we have developed a new unit selection method. It evaluates the naturalness of the overall intonation of speech candidates that were initially generated by a conventional unit selection algorithm and finds the best combination of selected speech candidates that gives the most natural intonation over the whole sentence. The intonation evaluation algorithm was designed using an SVM trained to discriminate the correct intonation from false ones. An example of a flow diagram for evaluating intonation naturalness and selecting speech with correct intonation patterns is shown in **Fig. 5**. First, the system evaluates the naturalness of the intonation of synthesized speech that consists of the combination of speech units selected by a conventional unit selection algorithm per accent phrase. In the example shown in Fig. 5, the intonation of the word "/sizeN na/" included in candidate speech A was evaluated as unnatural. (The red line in the figure indicates the flow of the first process.) Thus, candidate A was not output. Second, the intonation of the next candidate, B, was evaluated (green line in the figure). Candidate B was output as synthesized speech because the intonations of all the accent phrases included in this candidate were evaluated as natural. Experimental results for a preference test showed that 70% of the synthesized speech generated by our method with the reselection mechanism based on the intonation evaluation scheme was preferred over that generated by the conventional unit selection method.

4. Future of speech synthesis technology

The speech synthesis technology described above aims at a simple reading-out speaking style. For instance, this style focuses on situations such as stock price information or a market guide application as well as contact center guidance. Since the typical speaking style required for these applications is rather unemphatic, target applications are restricted to news readings, car navigation services, and so on. However, these represent only the first step toward the ultimate goal of speech synthesis technology, which is to have a machine perform the full range of human utterances. There are more variations in the voices that we usually hear, including the individuality of each speaker and changes in tone to suit the

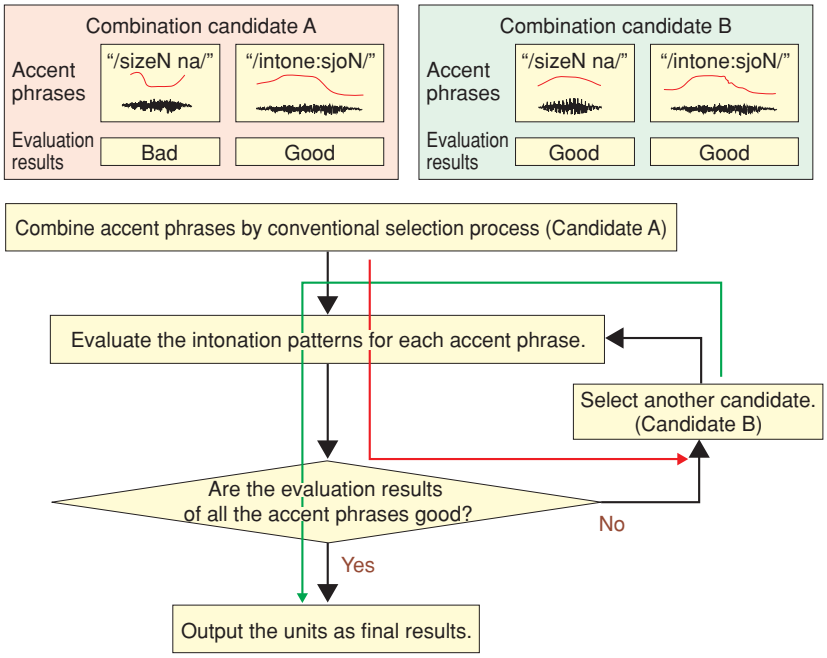


Fig. 5. Flow of speech generation with the intonation evaluation process.

time and place.

Several speech synthesis services that mimic various particular voices and speakers have already appeared, but the content that can be uttered is very limited. Typical services using various intonations and expressions are the call arrival announcement voices of mobile phones and the voices of robots. We believe that we can expand the application areas for speech synthesis technology if it becomes possible to apply the speaking style or tone of a certain speaker to arbitrary text without any limit on the content of the utterances.

For instance, companies could choose to give their IVR systems a speaking style or tone of speech that suits their company image such as a strong fresh voice, a calm and collected voice, or a vigorous voice. Moreover, the individual users of voice guidance systems such as car navigation systems might have very diverse preferences. A new lifestyle of enjoyable synthesized speech might be created in which people can control the tone of the speaker at will or use a tone that is currently in fashion.

Unfortunately, in the present speech synthesis framework, it is very expensive to change the individuality of a synthesized voice that is equal in quality to a human voice. Even if we focus on only one speaking style, for example, a fresh style, there will still remain many unsolved technical problems that prevent synthetic speech with this speaking style being applied to arbitrary text. We will conduct further research to solve these problems. In the future, when every machine can speak with a familiar synthetic voice that suits your wishes or with natural tone appropriate for the time and place, you will find you are in a new world that is completely different from the current IT society enclosed by inorganic interfaces.

5. Conclusion

The Cralinet text-to-speech system has been introduced into a contact center and its efficiency has been confirmed. Two statistical technologies have been developed for the system. The accent estimation technique achieves accuracy of 91% for family names and 86% for first names, and the intonation evaluation technique in the speech synthesis processing part improves the overall naturalness of the speech. We believe that giving people the ability to control the style of generated voices will be a key factor in future services using speech synthesis.

Reference

[1] H. Mizuno, H. Asano, M. Isogai, M. Hasebe, and M. Abe, "Text-to-Speech Synthesis Technology Using Corpus-Based Approach," NTT Technical Review, Vol. 2, No. 3, pp. 70-75, Mar. 2004.

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

He received the B.E., M.E., and Dr.Eng. degrees in electrical engineering from Waseda University, Tokyo, in 1982, 1984, and 1987, respectively. Since joining NTT in 1987, he has been engaged in speech coding standardization activities in Japan and ITU-T. He is currently developing the high-quality speech synthesis system Cralinet. He is a member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of Japan (ASJ), and the Information Processing Society of Japan (IPSJ). He received the NTT President's Award in 1993, the Prize for Outstanding Technological Development in Acoustics from ASJ in 1994, and the Paper Award from IEICE in 1995.

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

He received the B.E. degree in electronics engineering and the M.E. degree in information engineering from Nagoya University, Aichi, in 1986 and 1988, respectively, and the Ph.D. degree in computer science from Tsukuba University, Ibaraki, in 2006. He joined NTT Human Interface Laboratories in 1988. He is currently developing speech synthesis systems. He is a member of ASJ and IEICE.

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

He received the B.E. and M.E. degrees in information science and intelligent systems from the University of Tokushima, Tokushima, in 1990 and 1992, respectively. He joined NTT Information Processing Laboratories in 1992. During 1997-2002, he worked for ATR Interpreting Telecommunications Research Laboratories and ATR Spoken Language Translation Research Laboratories. His research interests include spoken language processing and he is now developing corpus-based speech synthesis. He is a member of ASJ, the Association for Natural Language Processing (ANLP), IEICE, IPSJ, and the Japanese Cognitive Science Society.

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

He received the B.E. degree in information engineering and the M.E. degree in intelligence science from Tokyo Institute of Technology, Tokyo, in 1995 and 1997, respectively. He joined NTT Basic Research Laboratories in 1997. His research interests include speech recognition, speech synthesis, speech understanding, and spoken dialogue processing. He is a member of ASJ, IPSJ, and the Japanese Society for Artificial Intelligence. He was a co-recipient of both the Paper Award and the Inose Award from IEICE in 2001.

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

He received the B.E. degree in information engineering and the M.E. degree in information science from Tohoku University, Miyagi, in 2001 and 2003, respectively. He joined NTT Cyberspace Laboratories in 2003. He is currently focusing on R&D of speech synthesis. He is a member of ASJ and IEICE.