Text-to-Speech Synthesis Technology Using Corpus-Based Approach

Hideyuki Mizuno†, Hisako Asano, Mitsuaki Isogai, Miki Hasebe, and Masanobu Abe

Abstract

Many people hope that in the future computers, machines, and robots will talk just like humans, letting us communicate with them naturally. Speech synthesis technology is an important component for this. We have developed text-to-speech synthesis technology using a corpus-based approach. This article introduces this technology, which can produce natural synthetic speech that is just like human speech.

1. Introduction

Text-to-speech (TTS) technology generates speech from any text. It is also known as speech synthesis. TTS can be regarded as a form of media conversion, in this case from characters to audio. It is an effective way of handling e-mail, text chat messages, and other information such as events and items on sale, because it is difficult to record such contents in advance.

Our previous TTS system, called Final-Fluet, is already being used in various application systems, e.g., an e-mail reading system and a news reader in a voice portal system. Unfortunately, conventional TTS systems provide synthesized speech that sounds robotic, contains pronunciation errors, and has unnatural intonation, so they are not suitable for incorporation in embedded computers, e.g., in TV set-top boxes and household robots. We need new TTS technology that can synthesize more natural and intelligible speech. This article describes our latest TTS system, which is designed to handle Japanese language.

† NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan
E-mail: mizuno.hideyuki@lab.ntt.co.jp

2. Corpus-based TTS

A flow diagram of the text-to-speech process is shown as **Fig. 1**. First, using the word dictionary and grammar, the text analysis step parses words in the Japanese input text, which consists of kanji (Chinese characters) and kana characters, and extracts word information, e.g., the word's reading, accents, and part of speech.

Next, the positions of pauses and intonation are

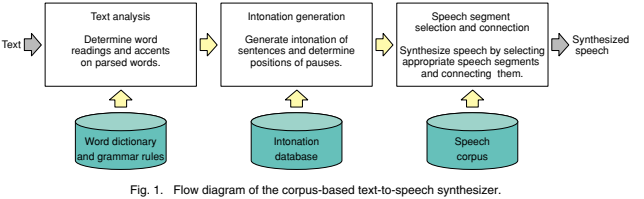

Fig. 1. Flow diagram of the corpus-based text-to-speech synthesizer.

determined based on sentence structure and the word information using an intonation database or rules.

Finally, appropriate speech segments are selected from a speech corpus (a very large database holding a wide variety of speech samples) according to the phoneme[*1] sequence given by the reading of the sentence. The intonation and length of the speech segments are modified as needed and concatenated to form sentences.

The speech synthesis process consists of three major sub-processes—text analysis, intonation generation, and speech segment selection and connection—and a dictionary or database. Unfortunately, none of these sub-processes has been well implemented, and the use of the dictionary/database creates many problems that lead to unnatural synthetic speech. Our advanced TTS technique using the corpus-based approach includes several new methods.

1) Precise transliteration by text analysis based on a statistical approach
2) Semi-automatic formation of a comprehensive speech segment set
3) Modification of speech intonation without degrading the naturalness of speech

These methods overcome several of the problems in speech synthesis and can produce synthesized speech that is as natural as human speech.

3. Text analysis based on a statistical approach

Our previous text analysis method achieves 99% reading accuracy for news texts. However, Japanese web or e-mail texts often include alphabetic words representing shop names, service names, place names, or nicknames, etc. derived from English (e.g., *Grape Bakery*), French (e.g., *Mont Blanc*) or Japanese[*2] (e.g., *Yokosuka*, which is a Japanese place name) etc. Most of them are not registered in the word dictionary (i.e., they are unknown words). As these words change rapidly according to current trends and fashions, not all alphabetic words can be registered in the word dictionary. It is therefore necessary to transliterate unknown alphabetic words using their spelling.

Our previous text analysis method chooses either the spell-out rule or English rule in advance for transliteration and applies the chosen rule to all unknown alphabetic words. For example, the transliteration of *Yokosuka* is "wai (*Y*), oo (*o*), kei (*k*), ..." when the spell-out rule is chosen and "youkosaka" when the English rule is chosen. Both transliterations are wrong because *Yokosuka* is Japanese and inappropriate rules were applied to *Yokosuka*. The most appropriate transliteration rule must be selected for each unknown alphabetic word. The Japanese rule can correctly transliterate *Yokosuka*.

For the reasons mentioned above, we developed a new transliteration method based on a statistical approach to handle unknown alphabetic words. We chose to use SVM (support vector machine)[*3] to categorize unknown alphabetic words. It consists of two steps: Step 1 classifies the unknown word into a "say-as" class, which means the class of the transliteration rule (e.g., spell-out, English, or romaji), based on statistics and Step 2 applies the transliteration rule for the classified say-as class.

Figure 2 shows the flow for transliterating unknown alphabetic words. Step 1 is based on SVM using spelling information and word length as the input. Step 2 applies the transliteration rules using statistical letter bigrams[*4] for English and using manually created letter-to-sound tables for Japanese and spell-out. This new method greatly improves the reading accuracy for unknown alphabetic words. In our previous method, it was 8% for Japanese class words and 75% for spell-out class words. The new method achieves 79% accuracy for Japanese class words and 100% accuracy for spell-out class words.

4. Construction of speech corpus

To construct the speech corpus, we must record a narrator's natural speech. To synthesize high-quality speech, the database must contain a wide variety of speech parts: words, syllables, and phonemes. If the recording script used for recording the parts is random or unbalanced, the recorded data may be full of redundancies and lack critical phonetic elements. We developed a new script generation method that mines a large Japanese text corpus to automatically create a comprehensive and non-redundant set of speech parts.

*1 Phoneme: the set of smallest units of speech in a language that distinguish one pronunciation from another.

*2 *Japanese* here means alphabetic transliteration of Japanese. Any Japanese word can be written in alphabetic transliteration, but this is unusual. Proper nouns are sometimes written in alphabetic transliteration depending on the context.

*3 SVM (support vector machine) is a method for creating functions from a set of labeled training data. The function can be a classification function or a general regression function. For classification, SVMs operate by trying to find a hyper-plane that can split two kinds of data. SVMs deliver state-of-the-art performance in real-world applications such as text categorization, handwritten character recognition, and bio-sequence analysis.

*4 Bigram: a unit consisting of two basic elements.

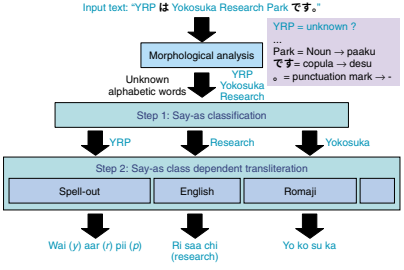

Fig. 2. Transliteration of unknown alphabetic words.

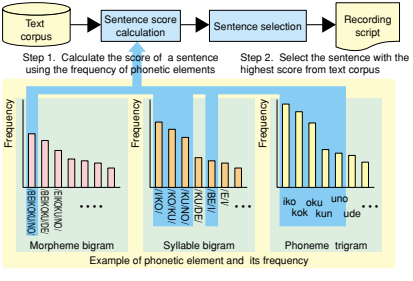

Fig. 3. Generation of a recording script.

We also need phoneme labels that link the script to speech parts, but the conventional labeling method of manually labeling the phonemes is too expensive and time-consuming. We developed a highly accurate automatic phoneme labeling method and a system that semi-automatically constructs the speech corpus. These advances make phoneme labeling cost-effective.

4.1 Recording script generation

Figure 3 overviews the recording script generation method. The frequencies of various elements in the Japanese text corpus are calculated in advance. The text corpus is derived from newspapers, novels, and so on. This method takes into account three types of phonetic elements: morpheme[*5] bigrams, syllable bigrams, and phoneme trigrams[*6].

First, the score of a sentence in the text corpus is calculated. The sentence score is defined as the sum of the frequencies of the phonetic elements in the sentence. In this calculation, we ignore the frequencies of phonetic elements in previously selected sentences as well as the frequencies of phonetic elements that appear often in the sentence. Moreover, the frequen-

*5 Morpheme: the smallest unit of meaning that a Japanese word can be divided into.
*6 Trigram: a unit consisting of three basic elements.

cies of phonetic elements are weighted according to their types. Second, the sentence with the highest score is selected from the text corpus. Finally, this sentence is added to the recording script.

Text selection is iterated until the size of the recording script meets some application-specific requirement. This method yields a recording script that contains a sufficiently wide variety of phonetic elements while simultaneously eliminating as many redundant phonetic elements as possible.

4.2 Semi-automatic construction of speech corpus

Figure 4 overviews the automatic phoneme labeling method. The acoustic feature models of each phoneme are trained using speech data with manually labeled phoneme segments based on a hidden Markov model (HMM)[*7]. Because HMM is the most powerful technique for automatic speech recognition and many techniques based on HMM have been studied, it is known that HMM can give high speech recognition rates. However, our goal is somewhat different; i.e., we want to determine phoneme boundaries that are close to manually assigned ones. Therefore, we optimized the HMM model parameters to develop a high-precision automatic labeling system. Using the acoustic feature models, speech data, and the phoneme label chain derived from recording scripts, the automatic phoneme labeling method determines suitable positions for phoneme boundaries in the speech data and labels phonemes at the boundary positions. An expert operator manually checks the phoneme boundaries and corrects any labeling errors. Finally, the speech corpus is generated using the speech and label data.

*7 A hidden Markov model (HMM) is a Markov chain, where each state generates an observation. You see only the observations; the goal is to infer the hidden state sequence. For example, the hidden states may represent words or phonemes, and the observations represent the acoustic signal. HMMs have recently become the predominant methodology for automatic speech recognition.

5. Intonation modification retaining the quality of natural speech

Our corpus-based TTS technology selects from the speech corpus the parts of speech (speech segments) that are appropriate for the reading and intonation obtained by input text analysis; these parts are concatenated to yield the synthesized speech. This means that the omission of any important speech segment or poor speech segment concatenation degrades the quality of the synthesized speech.

To overcome this problem, we have developed a new intonation modification technique. High-quality speech is obtained by estimating the quality of synthesized speech in selected speech segments and modifying the intonation of only those segments whose intonation is poor. This technique is outlined in **Fig. 5**.

In the phrase “yokosuka tsushin kenkyujo”, the intonation pattern of the directly concatenated speech segments selected from the speech corpus and the target intonation pattern estimated from the given input sentence are first compared. Note that excessive modification of those speech segments that have poor intonation leads to unnatural and mechanical speech. Therefore, we set a limit on the amount of possible

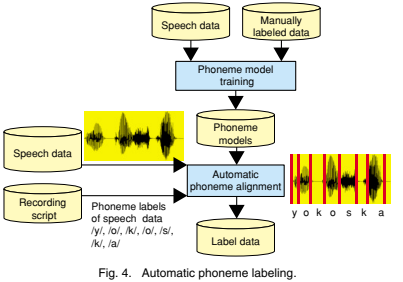

Fig. 4. Automatic phoneme labeling.

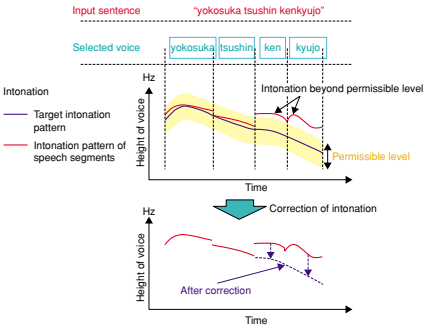

Fig. 5. Outline of intonation modification method.

modification.

In this example, because the speech segments "ken" and "kyujo" exceed the permissible range, the intonation patterns of just these two segments are modified; those of the other segments are not changed. This adaptable intonation modification algorithm yields speech with natural quality and smooth intonation.

6. Conclusion

Our latest text-to-speech technology produces better-quality speech than conventional speech synthesis technologies. Because it yields more audible and clearer speech, we expect various new services that were impossible using conventional TTS to emerge: for example, a virtual announcer or presenter on broadcasts and a personal secretary based on a computer agent. We are aiming to develop speech synthesis technology that achieves natural man-machine communication with emotion, personality, and intention, so that it seems as if a human rather than a machine is talking to us.

References

[1] H. Asano, M. Nagata, and M. Abe, "Say-as classification for alphabetic words in Japanese texts," Proc. Eurospeech 2003, Geneva, Switzerland, Vol. 4, pp. 3181-3184, Sep. 2003.

[2] M. Isogai, H. Mizuno, and M. Abe, "Text script generation based on coverage of morpheme and phoneme chains," ASJ 2003 Spring Meeting,1-6-17, Tokyo, Japan, pp. 255-256, Mar. 2003 (in Japanese).

[3] T. Yonezawa, H. Mizuno, and M. Abe, "Robustness of Automatic Labeling with HMM Phoneme Models," IEICE Technical Report SP2002-74, pp. 17-22, 2002 (in Japanese).

[4] M. Hasebe and M. Abe, "Effect of fundamental frequency differences at phrase concatenation," ASJ 2001 Autumn Meeting, 1-2-19, Oita, Japan, pp. 243-244, Sep. 2001 (in Japanese).

Hideyuki Mizuno

Senior Research Engineer, Media Processing Project, NTT Cyberspace Laboratories.

He received the B.E. degree in electronics engineering and M.E. degree in information engineering from Nagoya University, Nagoya in 1986 and 1988, respectively. In 1988, he joined NTT Human Interface Laboratories. He is currently developing speech synthesis systems. He is a member of the Acoustical Society of Japan (ASJ), and the Institute of Electronics, Information and Communication Engineers (IEICE).

Hisako Asano

Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.

She received the B.E. degree in electrical and computer engineering from Yokohama National University, Yokohama in 1991. Since joining NTT in 1991, she has been working on R&D of text analysis for text-to-speech synthesis and information extraction from e-mails. She is a member of the Information Processing Society of Japan and the Association for Natural Language Processing.

Mitsuaki Isogai

Research Engineer, Media Processing Project, NTT Cyberspace Laboratories.

He received the B.E. and M.E. degrees in electronics information engineering from the University of Yamanashi, Kofu, Yamanashi in 1995 and 1997, respectively. In 1997, he joined NTT Human Interface Laboratories. His current interests include speech synthesis. He is a member of ASJ and IEICE.

Miki Hasebe

Research Engineer, Media Processing Project, NTT Cyber Space Laboratories.

He received the B.E. degree in electronic engineering from Kitami Institute of Technology, Hokkaido in 1998. In 1998, he joined NTT Human Interface Laboratories. Since then, he has been engaged in R&D of speech synthesis algorithms. He is a member of ASJ.

Masanobu Abe

Group leader, Customer Premises Equipment Business Division, NTT East.

He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Waseda University, Tokyo in 1982, 1984, and 1992, respectively. In 1984, he joined NTT Electrical Communications Laboratories. In 1987, he joined ATR Interpreting Telephony Research Laboratories. In 1989, he joined the Laboratory Computer Science, MIT, as a visiting researcher. His research interests include speech synthesis, speaker individuality, and speech enhancement. He is a co-author of "Recent progress in Japanese speech synthesis" (Gordon and Breach Science Publishers, 2000). He is a member of ASJ, the Acoustical Society of America, and IEEE. He received a Paper Award from ASJ in 1996.