3-3 Multilingual Speech Synthesis System

SHIGA Yoshinori and KAWAI Hisashi

Adopting the Hidden Markov Model (HMM)-based technique that has become of major interest in the field of speech synthesis technology, we have developed a speech synthesis system with a high degree of flexibility. The technique, which generates speech from the models that were trained statistically on a speech corpus, is capable of acquiring voice characteristics and speaking style from a few hours of speech data, and is also applicable to new languages with relative ease. Exploiting such merits of flexibility, the system currently supports speech synthesis in seven languages and has been used as an output device in several applications such as multilingual speech translation systems and tourist guide spoken dialogue systems.

Keywords
Speech synthesis, TTS, HMM, Multilingual, SSML

1 Introduction

The technology of converting a text document into speech signals is called speech synthesis. The broader definition of the term includes record-playback speech synthesis, where prerecorded voices are simply played back on demand. It also covers speech synthesis by filing and editing, which stores in advance the words or phrases of recorded speech and joins them in a designated order for playback. A typical application of the former is a message response system in answering machines, and that of the latter is an automated announcement system at railway stations, which announces the destinations of arriving trains (by embedding the recorded speech of destinations into that of carrier sentences). While these types of speech synthesis can produce only limited kinds of speech content, the speech synthesis we deal with in this paper can read aloud any arbitrary sentences, and hence requires more advanced and complex processing. This type of speech synthesis is often called text-to-speech synthesis (TTS) to explicitly indicate that the speech synthesis is directly made from text.

Owing to recent improvement in speech quality, speech synthesis is now used in various fields. For example, it is used in the user interfaces of car navigation systems and video game machines. It is incorporated into electronic dictionaries to pronounce words and illustrative sentences. The technology is also applied to creating announcements on public buses and for radio broadcasting on highways. Furthermore, speech translation applications, which have emerged recently with the spread of smart phones, also take advantage of speech synthesis to output translations. Aside from these applications, the multilingualization of speech synthesis is recently in increasing demand due to the growing number of foreign visitors to Japan and the globalization of products.

In this paper we will introduce the multilingual speech synthesis system developed by NICT. This system employs a speech synthesis method based on the Hidden Markov Model (HMM) [1], which has attracted much attention in recent years in the research field, and covers seven languages (as of June, 2012). This paper is organized as follows: In Section **2**, we briefly explain the mechanism of NICT's

speech synthesis system. Section **3** introduces some applications in which the advantageous features of the system, i.e. the diversity and flexibility of synthesizable speech, are made full use of. Section **4** describes current technical problems as well as future issues. In this paper we do not explain any fine details of each particular technique because of space limitations. For further information, see the references listed at the end of this paper.

2 NICT multilingual speech synthesis system

Figure 1 shows a block diagram of the NICT multilingual speech synthesis system. As with most other speech synthesis systems, it consists of two processing units: Text processing and speech signal processing. The role of the text processing unit is to perform language processing on input text and generate a phoneme sequence and some linguistic information to control the phonetic property (timbre) and prosodic property (intonation) of the speech to be synthesized. A phoneme is the smallest unit of speech sound in a language. The role of the speech signal processing unit is to first generate the acoustic features of speech based on the phoneme sequence created by the text processing unit, and then convert them into a speech waveform as the system's final output. In what follows we delve into the details of each unit of the NICT multilingual speech synthesis system. Then we describe the multilingualization of the system and additionally report on the current status of its coverage of the W3C Speech Synthesis Markup Language.

2.1 Text processing unit

The text processing unit is composed of three modules: language processing, pause location estimation and pronunciation generation. The language processing module first

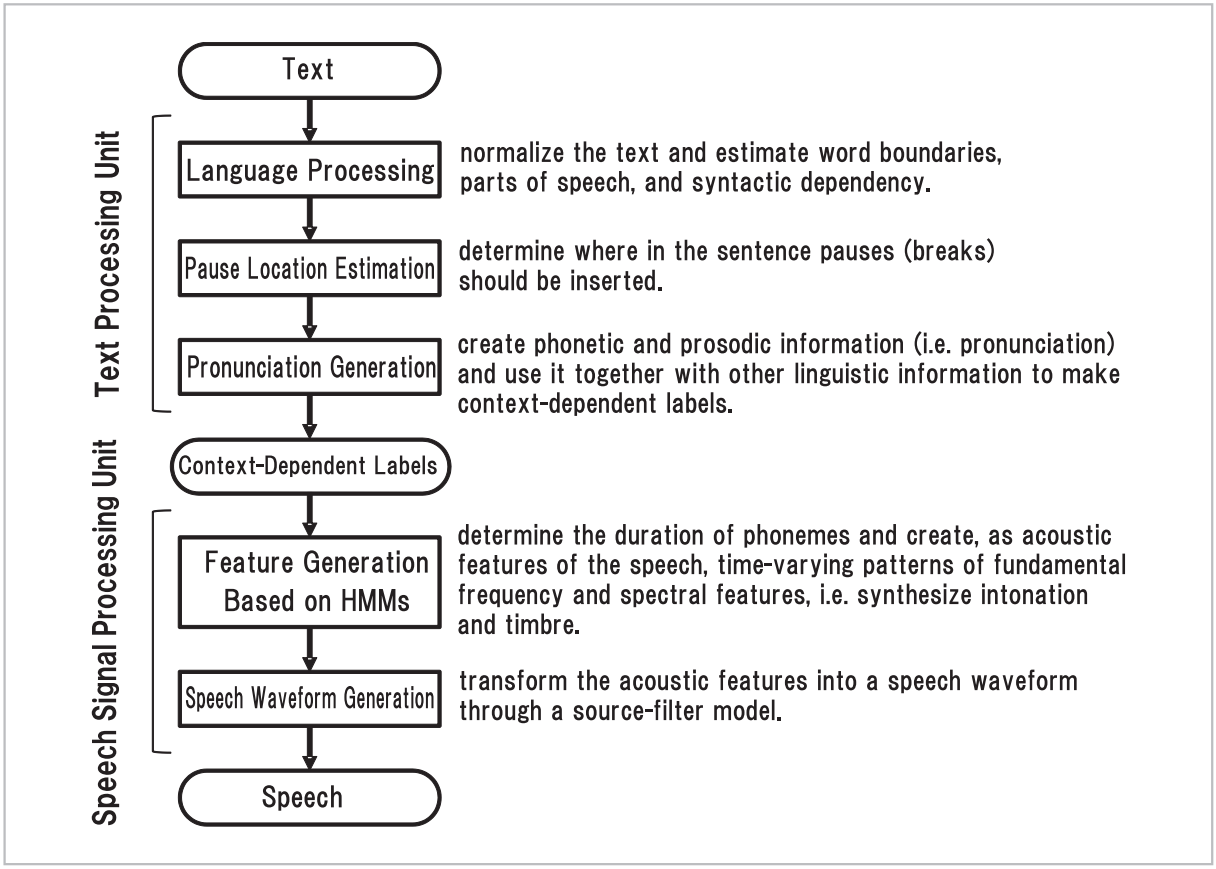

Fig.1 *Block diagram of NICT Multilingual Speech Synthesis System*

normalizes input text. In this normalization, the numeric numbers (e.g. positional representation, telephone number or time), units and symbols (e.g. "cm" and "$") and, in the case of English and the like, contracted forms and abbreviations (e.g. "Mr.", "Ltd." and "Co.") are replaced with appropriate texts. Then, specifically for Japanese and Chinese texts, which have no word segmentation (having no space between words), morphological analysis is performed to identify the ranges of words and/or phrases in the text. Morphological analysis is a process for dividing sentences into morphemes (the smallest meaningful units in a language). For languages with word segmentation (e.g. English), a grouping of some sequential words often expresses a certain concept and therefore a processing similar to morphological analysis is performed on the text.

The pause location estimation module analyzes the syntactic dependency between adjacent words or phrases and, based on the resultant dependencies, determines where to insert pauses in the sentence. The duration of the pauses is not determined at this stage, but is determined statistically together with the duration of phonemes later in the speech signal processing unit. This is because pause duration varies depending on the speaker and his/her speaking style.

The pronunciation generation module creates information about the reading and one of accents (e.g. in Japanese), stresses (e.g. in English), and tones (e.g. in Chinese) by referring to dictionaries within the system. Words that cannot be found in the dictionaries are converted into pronunciation according to rules specifically designed for this conversion. In the case of Japanese for example, readings and accents are provided to such unknown words by first referring to a single-kanji dictionary, where a specific reading is given to every kanji, and then applying rules for assigning a reading and accent to each unknown word. Various other processes are also involved in generating pronunciation in accordance with each language. For example, again in the case of Japanese, the following processes are also applied: the identification of phrase boundaries, the determination of the reading of homographs (words that have more than one reading, such as "市場", which is read as "ichiba" or "shijo"), the processes of euphonic changes and vowel devoicing, and the definition of the accent type of each phrase according to the accent type and accent modification (i.e. sandhi) rules of each morpheme in the phrase. Finally, the module creates a sequence of phoneme labels that include the context, consisting of various kinds of linguistic information (hereafter the labels are referred to as "context-dependent labels"). The details of the context will be given in Subsection ***2.2.1***.

For more information about general text processing for Japanese speech synthesis, see reference [2].

2.2 Speech signal processing unit

The speech signal processing unit determines (1) the duration of each phoneme and generates (2) the time-varying pattern of the fundamental frequency (F_0) and (3) that of the spectral features, based on the phoneme label sequence produced by the text processing unit. To generate these acoustic features we use an HMM-based technique, in which the above features (1) to (3) are simultaneously modeled with the probabilistic model, HMM.

2.2.1 Training HMMs with speech corpus

HMMs are trained statistically with a speech corpus in advance. Figure 2 shows a block diagram of the HMM training. Each HMM corresponds almost to a phoneme of each language, and several HMMs are provided to each phoneme depending on its context in order to model the above three acoustic features. The context of a phoneme is a collection of linguistic information which is likely to affect the acoustic features of the phoneme. For example, because the spectral feature of a phoneme is affected by its adjacent phonemes, the types of the precedent and subsequent phonemes are dealt with as part of the context. We also consider as part of the context the posi-

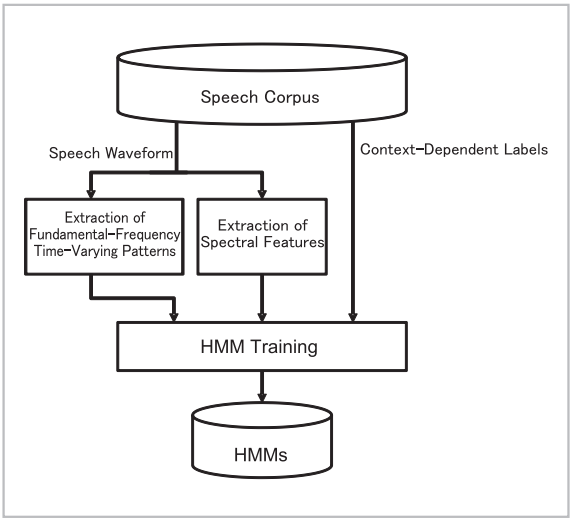

Fig.2 *Training HMMs based on speech corpus*

tion of the phoneme within the sentence, clause and phrase as well as information on the accents/stresses/tones, for the purpose of modeling the duration and F_0 pattern of each phoneme.

However, since a large and diverse amount of linguistic information is actually taken into account (e.g. 50 types of linguistic information compose the context of our Japanese HMMs), their combination produces enormous numbers of contextual varieties. It is therefore difficult to provide sufficient training data for every context-dependent HMM. To cope with this problem, we apply a context clustering technique [3] and train an HMM for each context cluster, thereby ensuring the amount of training data for every HMM.

2.2.2 Speech synthesis from HMMs

For speech synthesis, the trained HMMs are first concatenated according to the context-dependent labels created by the text processing unit in the previous step. From the chain of HMMs, the algorithm in [4] generates the acoustic features (1) to (3) mentioned above. Finally, the features are converted into a speech waveform through a source-filter model which simulates the human speech-production mechanism [5].

2.3 Coverage of multiple languages

The NICT speech synthesis system covers Japanese, English, Chinese, Korean, Indonesian, Malay and Vietnamese (as of June, 2012). We plan to add more languages, in particular Asian languages, in the future.

To cover a new language, we need to mainly develop the following two items. The first item is a text processing unit for the language. To develop this from scratch using a corpus-based approach, it is necessary to (1) build a text corpus that contains morpheme boundaries and part-of-speech information, (2) create a grammar model for morphological analysis based on the text corpus, and (3) create pronunciation generation rules.

The second item is a set of HMMs for the new language. Developing this requires (1) designing the context of the language in accordance with the characteristics of the language and (2) building a speech corpus including the context-dependent labels. It is empirically known that the speech corpus needs to contain at least two-hour speech data for synthesizing speech of acceptable quality. For producing practicable high-quality speech, the corpus is required to have more than five-hour speech data. These data sizes are presented in net time and, to obtain speech data of this amount, three to four times more hours of studio recording is necessary. This is because speakers usually read out sentences at intervals over time and are directed to have ample breaks in order to prevent them from straining their throats. For example, to build a five-hour speech corpus, 15 to 20 hours of studio recording, including intermissions, is necessary.

2.4 Compliance with speech synthesis markup language

The NICT multilingual speech synthesis system accepts the Speech Synthesis Markup Language (SSML) Version 1.1 [6] as input. SSML is an XML-based markup language designed to assist speech synthesis used in Web and other applications. It is recommended by W3C and has been developed continuously for wider coverage of languages. Using SSML as system input allows us to switch the properties of synthesized speech, such as the language

and the speaker, in the middle of reading a document aloud, and moreover, fine-tune the output speech by specifying, for instance, the location and/or duration of pauses in the sentences.

Tables 1 and 2 list SSML tags currently available in the NICT speech synthesis system (as of June, 2012). We intend to increase the availability of the W3C-standard SSML tags to our system in the future. Figure 3 shows an example of SSML that represents a conversation between two people in Japanese.

Table 1 *SSML tags for any language*

Element	Attribute	Description
speak	version	SSML Specification Version (1.1)
	xml:lang	Specifies the Language Used in Documents
voice	languages	Specifies the Language in the Tagged Range
	name	Specifies Speaker Model Name to be Used for Speech Synthesis
	gender	Specifies Gender (M/F)
	f0-mean	Specifies the Mean of Fundamental Frequency (F_0) (NICT's Original Specification)

Table 2 *SSML tags for Japanese only*

Element	Attribute	Description
P	-	Shows Paragraph Boundary
S	-	Shows Sentence Boundary
token (w)	-	Specifies Multiple Morphemes to be Processed as a Single Word
say-as	interpret-as	Specifies How to Interpret a Text (Numeral Phrase)
break	strength	Specifies the Strength of a Break in Six Levels
	time	Specifies the Duration of a Pause

3 Applications

As mentioned above, the HMM-based speech synthesis method has a great advantage in flexibility such that it can acquire the voice quality and speaking style of the speaker from a two-to-three-hour speech corpus. Also, the data size of a model set for a single speaker (hereafter referred to as "speaker model") is relatively small, a few to a dozen megabytes with no compression, while the unit selection speech synthesis method, which is now widely used in commercial products, requires a few hundred megabytes to a few gigabytes. The following introduces some of the applications that utilize these advantages.

3.1 Dialogue speech synthesis

We developed the speaker model

```
<speak version="1.1" xml:lang="ja-JP"> ← SSML Ver. 1.1, in Japanese
  <voice name="JF009">          ← Specifies the Japanese Speaker Model JF009
    <p>                         ← Beginning of a Paragraph
      <s>                       ← Beginning of a Sentence
        よろしければ、<break strength="strong"/>電話番号を教えていただけますか？
      </s>                      ← End of the Sentence   ↑ Break Widely in the Middle of the Sentence
    </p>                        ← End of the Paragraph
  </voice>
  <voice name="JM001">          ← Switch into Japanese Speaker Model JM001
    <p>
      <s>
        良いですよ。
      </s>
      <s>
        <say-as interpret-as="telephone">123-4567</say-as>です。
      </s>                              ↑ Read as Telephone Number
    </p>
  </voice>
</speak>
```

Fig.3 *SSML description example*

"HANNA" in fiscal year 2010 for a spoken dialogue system of Kyoto tourist information. The model was created based on a dialogue speech corpus which was built using simulated dialogues between a tour guide and a visitor. It is an example which utilizes the flexibility of the HMM-based speech synthesis method.

We recorded the dialogue of two professional voice-talents using a special script. The script was created by transcribing a spontaneous conversation between a professional tour guide and a visitor. We then extracted only the sections of the tour guide's speech from the recording, and built a corpus with about four-hour speech for training HMMs. We thereby succeeded in synthesizing speech with a dialogue speaking style that is suitable for tour-guide responses, intonations that are appropriate to the speech content, and the clear pronunciation of the professional voice-talent. We confirmed experimentally that the synthetic speech created in this way evoked natural spontaneous responses from users [7].

3.2 Speech synthesis in different languages with same voice individuality

There has been a demand for a speech translation system to pronounce translation results in the voice of the user (i.e. speaker of the source language) instead of a fixed voice preset by the system. Such a system will, if realized, allow us to easily identify one user from lots of others by the characteristics of the synthesized voice. This demand can be met using the following approach. Taking advantage of the flexibility of HMM-based speech synthesis and its small footprint (i.e. the small data size of a speaker model), we first create a large number of speaker models of the target language and store them in the system in advance. For speech translation, the system chooses one of the models which is acoustically closest to the model of the input speech (i.e. user's voice) in a certain distance measure. The selected speaker model is used to synthesize speech from the translation results.

The system developed experimentally in fiscal year 2011 provides speech translation from Japanese to English. It synthesizes speech by selecting the speaker model of the English voice closest to the input Japanese voice from 50 male and 50 female voice models, as well as adjusting the F_0 reference value (voice pitch) to that of the input speech. For details of the speaker-model selecting method, see reference [8]. The effectiveness of this cross-lingual speaker-adaptation technique was demonstrated in a perceptual experiment conducted to evaluate the speaker similarity between the voices of different languages [8].

4 Conclusions

The HMM-based speech synthesis method used in the NICT system is suitable for the multilingualization and speaking-style diversification as we have seen in this paper. On the other hand, however, it has the disadvantage that its output speech tends to sound muffled or buzzy, i.e. sounding like so-called "vocoded speech." Such speech quality degradation is mainly caused by the process of re-synthesizing a speech waveform from the generated acoustic features, and the process for the statistical training of HMMs where the acoustic features are over-smoothed by the averaging operations.

The former, i.e. the degradation due to the resynthesis process, can be reduced by applying an analysis-synthesis technique using the same resynthesis process to the speech of candidate speakers for studio recording, and selecting speakers who have little degradation in their speech produced through this technique. On the other hand, for the latter, i.e. degradation due to the statistical training, we have not yet found any effective way of reducing it. However, careful selection of speakers with stable voices can alleviate the over-smoothing effect to some extent since such smoothing tends to occur when the acoustic features of the speech in the training corpus have significant variations that cannot be represented by the context. Thus, in the current situation, we audition several tens of professional

narrators and/or voice talents, resynthesize speech through the analysis-synthesis technique from their recorded speech samples, and evaluate the degree of the quality degradation in the resynthesized speech for each speaker. At the same time, our researchers check the stability of the speech samples. Considering both these factors, then, we select speakers judged to be most appropriate.

If viewed from the opposite side, however, this means that current technology is not satisfactory in synthesizing high-quality speech from any person's voice. To resolve this problem, we must cope with the above-mentioned causes of speech degradation. NICT has therefore been conducting research into speech parameterization and resynthesis techniques that are suitable for any type of voice, and research into the modeling of speech that produces little degradation resulting from the over-smoothing [9]-[11]. In addition, we need to establish a method of high-quality speech synthesis from a brief recording. It is difficult for non-professional speakers to make a recording over a long period of time in a stable manner. Speaker adaptation techniques [12] are considered to be effective for this purpose and thus we have introduced some of them [13]. In the techniques, models are trained beforehand with a large-scale corpus built from the speech of different speakers and are adapted to a small amount of speech from a target speaker. However, they have the problem of considerable quality degradation in the synthesized speech since those techniques, in principle, involves deformation of the acoustic features.

To establish a technology that can practically synthesize high-quality natural-sounding speech in any kind of voice of any person from a few to a dozen minutes of their speech data—This will be one of our major future challenges.

References

1 H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," Speech Communication, Vol. 51, No. 11, pp. 1039–1154, Nov. 2009.

2 Y. Sagisaka, "Natural language processing in speech synthesis," IPSJ Magazine, Vol. 34, No. 10, pp. 1281–1286, Oct. 1993.

3 T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in Proc. EUROSPEECH'99, pp. 2347–2350, Sept. 1999.

4 K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in Proc. ICASSP, pp. 1315–1318, June 2000.

5 S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," IECE Trans. A, Vol. J66-A, No. 2, pp. 122–129, Feb. 1983.

6 http://www.w3.org/TR/speech-synthesis11/

7 T. Misu, E. Mizukami, Y. Shiga, S. Kawamoto, H. Kawai, and S. Nakamura, "Toward Construction of Spoken Dialogue System that Evokes Users' Spontaneous Backchannels," IEICE Trans., Vol. J95-A, No. 1, pp. 16–26, 2012.

8 M. Tsuzaki, K. Tokuda, H. Kawai, Y. Shiga, J. Ni, K. Oura, and S. Shiota, "Perceptual evaluation of synthesized speech reflecting "personalities"," IEICE Tech. Rep., Vol. 112, No. 81, SP2012-39, pp. 33–38, June 2012.

9 R. Maia, T. Toda, K. Tokuda, S. Sakai, and S. Nakamura, "A decision tree-based clustering approach to state definition in an excitation modeling framework for HMM-based speech synthesis," in Proc. Interspeech2009, pp. 1783–1786, Sept. 2009.

10 Y. Shiga, T. Toda, S. Sakai, and H. Kawai, "Improved training of excitation for HMM-based parametric speech synthesis," in Proc. Interspeech2010, pp. 809–812, Sept. 2010.

11 Y. Shiga, "Pulse Density Representation of Spectrum for Statistical Speech Processing," in Proc. Interspeech2009, pp. 1771–1774, Sept. 2009.
12 J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," IEICE Trans. Inf. & Syst., E90-D(2), pp. 533–543, Feb. 2007.
13 J. Ni and H. Kawai, "On effects of speaker similarity in average voices on adapted web-based HMM voices," in Proc. Spring Meet. Acoustical Society of Japan, Vol. I, 3-7-3, pp. 303–306, Mar. 2011.

(Accepted on June 14, 2012)

***SHIGA Yoshinori**, Ph.D.*

Senior Researcher, Spoken Language Communication Laboratory, Universal Communication Research Institute

Speech Signal Processing, Speech Synthesis

***KAWAI Hisashi**, Dr. Eng.*

KDDI R&D Labs. Inc./ Former Executive Researcher, Spoken Language Communication Laboratory, Universal Communication Research Institute

Speech Information Processing, Speech to Speech Translation