SPEECH SYNTHESIS USING LARGE SPEECH DATA-BASE

Kyu-Keon LEE, Takemi MOCHIDA, Naohiro SAKURAI and Katsuhiko SHIRAI

Department of Electrical Engineering
Waseda University
3-4-1 Okubo, Shinjuku-ku,
Tokyo 169
JAPAN

ABSTRACT In this paper, we introduce a new speech synthesis method for Japanese and Korean arbitrary sentences using the natural speech data-base. Also, application of this method to a CAI system is discussed. In our synthesis method, a basic sentence and basic accent-phrases are selected from the data-base against a target sentence. Factors for those selections are phrase dependency structure (separation degree), number of morae, type of accent and phonemic labels. The target pitch pattern and phonemic parameter series are generated using those selected basic units. As the pitch pattern is generated using patterns which are directly extracted from real speech, it is expected to be more natural than any other pattern which is estimated by any model. Until now, we have examined this method on Japanese sentence speech and affirmed that the synthetic sound preserves human-like features fairly well. Now we extend this method to Korean sentence speech synthesis. Further more, we are trying to apply this synthesis unit to a CAI system.

1. INTRODUCTION

To improve intelligibility and naturalness of synthetic speech sounds, it is essential to realize natural prosodic features as much as possible. In spoken Japanese, it is well known that the global F0 shape or the length of pauses are mainly decided by the depth of contextual gaps at phrase boundaries, the grammatical combination between adjacent words, and so on. From this viewpoint, many schemes have been developed to estimate control parameters for pitch patterns or lengths of pauses from texts. In most of these schemes, pitch patterns are generated by superpositional models and their control regulations[1][2]. However, natural speech has many more complicated pitch patterns and there has so many control factors at different levels. So it is quite difficult to quantify and optimize these factors. From this viewpoint, we have examined a method to realize natural prosody using a large data-base.

Japanese and Korean grammatical structures are similar. So, we are trying to apply Japanese rules of prosodic generation to Korean. Moreover, we examined now, the CAI system for Korean language education as an application can be built into these synthesis systems.

2. GENERATION OF PITCH PATTERN

In this method, we generate the pitch pattern where the accent-phrase is assumed to be a unit. The generation method of appropriate pitch pattern is examined by calculating global F0 shape of the each accent phrase of each sentence in the data-base.

To examine the effect quantitatively, we express global F0 shape as follows (Fig 1). The regression line for the pitch pattern is calculated using a method of least squares phrase by phrase. The slant is assumed to be a, and the altitude of a center position is assumed to be b.

The regression lines of the pitch pattern of the accent phrase preceding and following the target phrase is similarly calculated. A is assumed to be the slant of the line which connects between center points of each accent-phrases and B is assumed to be altitude. (a,b) are normalized to $(a',$ $b')$ by using this A and B (Exp 1).

The difference of the height of a pitch is absolute to the utterance of every sentence. The purpose of this normalization is to reduce the effect on the values of a and b.

$$a' = \frac{a - A}{1 + aA} \qquad b' = b - B \qquad (1)$$

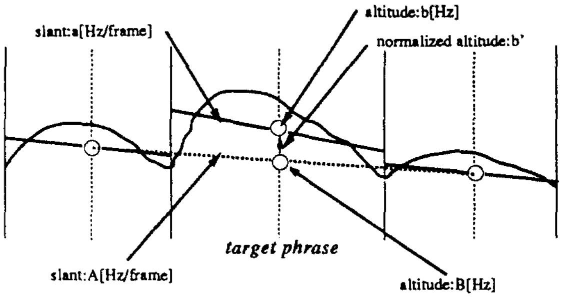

Figure 1. Approximation of Global F0 Shape

3. JAPANESE SPEECH SYNTHESIS

3.1. Sentence Speech Data-base

Isolated sentence speech data which are released by ATR are used as the data-base. This data-base includes 503 sentences spoken by one professional male speaker, and each sentence has an information file about its phrase dependency structure.

3.2. Selection of the Basic Sentence

In Japanese sentence speech, the phrase dependency structure essentially influences the global F0 pattern.

Fig 2 shows the transition of value b' which appears in sentences that consist of 4 accent-phrases . The index represents the phrase dependency structure of these sentences using separation degrees at each phrase boundary. According to this result, it is concluded that the phrase dependency structure contributes to the transition of value b' in a sentence essentially.

Accordingly, we should select a basic sentence from the data-base which is completely identical in the structure of the target sentence.

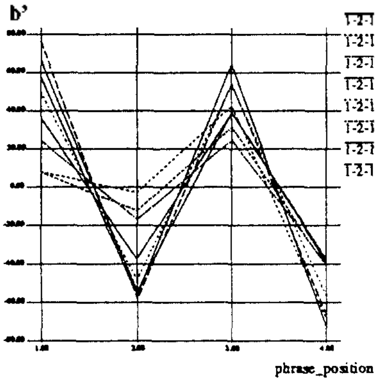

Figure 2. Transition of b'

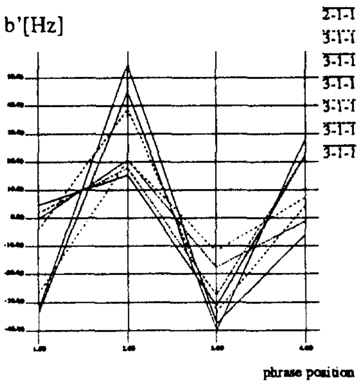

Figure 3. Transition of b'

However, variation of phrase dependency structure increases as the number of accent-phrase increases. Therefore, the basic sentence is not always available in the data-base. Here, we classify the dependency rerations between accent-phrases as D (Direct union) and I (Indirect Union) as shown in Fig 4.

For example, in sentences that consist of 4 accent-phrases, sentences whose structure is 2-1-1 and 3-1-1. These will both become the identical structures of I-D-D, if I and D are used. Fig 3 shows the transition of value b' which appears in sentences whose structures are 2-1-1 and 3-1-1. From this figure, it is found that both transitions of b' are similar.

Therefore, we should express structure of the target sentence using these two rough categories to select the basic sentence.

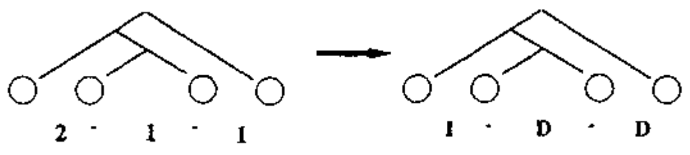

Figure 4. Expression of Structure Using D(Direct) and I(Indirect).

If more than 2 sentences are selected as the target sentence, the most appropriate sentence is selected on the basis of the number of mora, the accent type and the phonemic labels in each accent phrase. Details of selection procedure is described in the next section.

3.3. Selection of the Basic Accent Phrase

Until now, we have examined the method to generate arbitrary word speech using large data-base of word speech [3]. In this method, a basic word is selected from the data-base which has the same number of mora and the same type of accent as those of the target word, and which has similarly matched phonemic labels as much as possible. For synthesis, the pitch pattern extracted from the basic word is used with no modification, and the phonemic parameter series is generated by replacing parameters for mismatching mora between basic and target words.

Here we extend this scheme to the generation of arbitrary accent-phrase. In the case of word, the global F0 slant is approximately flat. However, in the case of accent-phrase, the global F0 slant is strongly affected by the phrase dependency structure. To examine the effect quantitatively, the distribution of global F0 for the same accent-phrase spoken under the various condition of phrase dependency structure is looked over using the above values of a' and b'. Figure 5 shows the distribution of (a', b') for words which have 4 morae and accent type of 3.

The index shows the separation degree at left and right phrase boundaries. From this result, for the target accent-phrase, the accent-phrase which is uttered under completely the same condition of phrase structure should be selected as the basic accent-phrase.

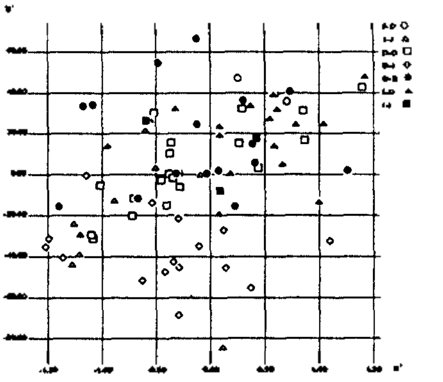

Figure 5. Distribution of (a', b') for 4 Mora and Accent Type 3

For the target phrase, in most case, several accent-phrases which have the same number of mora and the same accent type under the same condition of phrase structure can be found in the data-base. In such a case, the one whose phonemic labels match those of target phrases as much as possible is selected as the basic accent-phrase.

3.4. Replacement and Modification Method

When the target sentence is given in the form of text and phrase dependency structure, the basic sentence is selected under the condition described in Section 3.2. and 3.3.. As for the accent-phrase in the basic sentence which has the same number of mora and the same type of accent as those of the target sentences, there is no necessity for the replacement of modifying the pitch pattern in the basic sentence. However, if those factors of a phrase in the basic sentence is not equal to those of the target, the basic accent-phrase has to be selected under the condition described in Section 3.3.. The pitch patterns of mismatching phrases are replaced by that of the basic accent-phrase with a slight modification. The modification carried out is replaced by that of the basic accent-phrase with a slight modification. The modification is carried out using the value a',b'. The replacement and modification methods are as follows.

It is necessary to preserve the transition of b' in a basic sentence.(from Section 3.2.)Therefore, in any case, the value of b' of the basic accent-phrase normalizes to the value of b' in the basic sentence.

On the other hand, a' and accent type have a mutual relationship. From this viewpoint, the value of a' of the basic accent phrase is not appropriate in the case 2 and case 3. Therefore, the value of a' of basic accent-phrase does not normalize to the value of a' in the basic sentence.

Next, case 1 is examined. Fig 6 shows the distribution of a' in the accent-phrases which have an accent type of 3 and various number of morae. This result shows that the phrase dependency structures and the accent type preceding and following have mutually similar values of a' between the same accent-phrases even if the number of mora is different. Therefore, the value of a' of the basic accent-phrase normalizes to the value of a' in the basic sentence in the case 1.

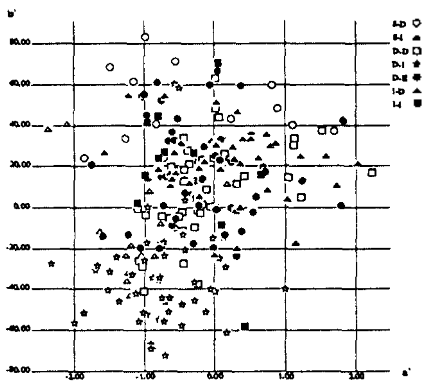

Figure 6. Distribution of (a', b') for 3 Mora

Table 1. Normalization of Pitchpattern of Basic Accent Phrase

	Accent Type	Number of Mora	Normalize to Basic Sentence	
			a'	b'
case1	○	×	Do	Do
case2	×	○	Not	Do
case3	×	×	Not	Do

(○ : match、× : mismatch)

3.5. Experiment

Using the pitch pattern generated by those method, the speech signal is synthesized.

To evaluate the generation method of the pitch pattern, the subjective experiments of the synthetic voice are performed.

The sample data is the sentences which consist of both four accent-phrases. Three of these accent-phrases were extracted from the basic sentence and one accent-phrase was generated using the basic accent-phrase. Two sentences for each of the three cases of replacing method described in Section 3.4., the total of six sentences were used for the evaluation. Five subjects listened the synthetic sounds and forced to answer the naturalness of the prosody with five grade. The average and the standard deviation of the scores of those five subjects are shown in Table 2. The highest score appears in case 1. This result shows that the accent type of each phrase is quite essential for the selection of the basic sentence, and it is important the preservation of the value of a' of the basic sentence for replacing or modifying the pitch pattern. Therefore, it is found that the accent type of the each accent-phrase in a basic sentence is corresponding to the target sentence as much as possible to make the synthetic sound more natural.

Table 2. Subjective Evaluation

	Accent Type	Number of Mora	Average	Standard Deviation
Case1	○	×	3.9	0.89
Case2	×	○	3.1	0.69
Case3	×	×	2.8	1.56

(○ : match、× : mismatch)

4. KOREAN SPEECH SYNTHESIS

4.1. Grammatical Similarities between Japanese and Korean

The grammatical similarities between Japanese and Korean are essential for applying a Japanese speech synthesis method to a Korean speech synthesis. Being based on contrastive analysis of two languages, grammar and phonetic similarities are as follows:

1. Grammatical Similarities
   (a) The word order is almost the same.
   (b) The sentence structure and phrase dependency structure are very similar. Especially, relation of dependency in accent-phrase between two languages does not cross.
2. Phonetic Similarities
   (a) The vowels (a, i, u, e, o) are phonetically very similar.
   (b) The syllable structure has V, VC, CV, CVC.

For example, the phrase dependency structure between two languages are shown in Fig 7

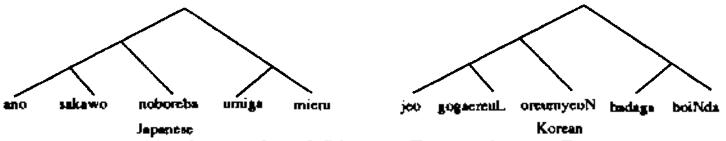

Figure 7. Example of Phrase Dependency Structure

Figure 7 shows the same dependency structure between two languages. It has an accent-phrase structure of 1-1-2-1 and the number of accent-phrase is 5. Based on these similarities, we have investigated the possibility to apply the method of generation of pitch pattern in Japanese to the Korean. In Korean sentence speech not only type of accent and number of mora of each accent-phrase but also the phrase dependency structure between each accent-phrase has essential influences on its prosody. Therefore, we examined the correlation between the dependency structure of each phrase and the pitch pattern on the accent-phrase with identical type of accent and number of mora among accent- phrases in Data-base.

4.2. Application of Japanese Speech Synthesis Method

We construct Korean sentence speech data-base. This includes 100 sentences, which are translated from ATR Japanese data-base into Korean , spoken by one native Korean female speaker, and each sentence has an information file same as the Japanese data-base.

So, we confirmed whether a similar rules of the selection of basic sentences and basic accent-phrases to Japanese can be applied into Korean.

4.2.1. Selection of the Basic Sentence

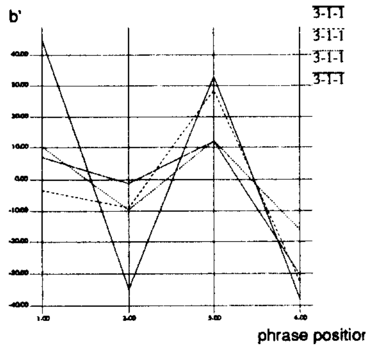

Figure 8. Transition of b' in Korean (1)

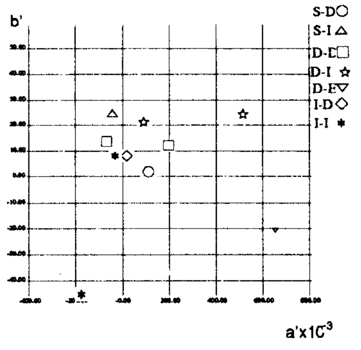

Figure 9. Distribution of (a', b') for Accent Type of 0 and Various Number of Mora

Because the phrase dependency structure is similar to Japanese, it is guessed that a similar rule to Japanese can be applied in Korean in the selection of a basic sentence.

Fig 8 shows the transition of value b' which appears in the sentence. The index shows the phrase dependency structure of sentence using separation degree at each phrase boundary. Thus, the structure also contributes to the transition of value of b' in a sentence essentially in Korean.

Therefore, in Korean we also should select a basic sentence which is completely identical in the structure of the target sentence as the basic sentence.

4.2.2. Selection of the Basic Accent-Phrase

It is similar to select the one to which the number of mora and the type of accent are corresponding in case of Japanese.

The relation between phrase dependency structure and the pitch pattern preceding and following of the accent-phrases is verified. Fig 9 shows the distribution of (a', b') in the accent-phrases which have accent type of 0 and various number of morae. From this result, it is found that the phrase dependency structures and the accent type preceding and following have mutually similar values (a', b') between the same accent-phrases.

Therefore, and accent phrase which has the same phrase dependency structure between the preceding and the following accent-phrases should be a basic accent phrase selected.

5. ICAI SYSTEM WITH SPEECH SYNTHESIS

An Intelligence Computer Assisted Instruction (ICAI) system incorporating speech synthesis may provide an effective new environment for language learning. To transfer information that can not be explained by word characters in the ICAI system, speech synthesis is essential as a part of its interface.

Thus, an ICAI system with speech synthesis capability has the following advantages:

1. Speech generated by the system assists phonetic understanding.
2. Audio information is more effective in maintaining natural concentration than a CRT display.
3. The same verbal information can be simultaneously given to multiple students using a sound speaker.
4. Individual communication is established between a student and the system.

Due to these advantages, we have investigated an ICAI system that performs Japanese and Korean speech synthesis. This system assists a Japanese student to learn compositional writing in Korean based on a principle that Japanese and Korean are grammatically similar. The system provides advice to the student by analyzing errors found in compositional writing. Speech synthesis is applied to give a student compositional exercises in Japanese and to provide advice on compositional errors in Japanese and in Korean.

5.1. Configuration of the Composition Training ICAI System

The composition training ICAI system consists of a student learning environment, a domain knowledge base for target languages, a student diagnosis module, a teaching module, and a speech synthesizer for Japanese and Korean as shown in Figure 10.

5.2. Domain Knowledge Base

The domain knowledge base includes essential background knowledge and compositional exercises. Background knowledge is comprised of a basic Korean text for non-Korean-speaking Japanese students. The basic text in the system is based on two text books, .i.e., one is published in Japan to teach Japanese Korean language and the other is used at the "Korean Linguistic Laboratory" in Korea to teach non-Korean speakers the Korean language. The size of the knowledge base is about 100 Korean grammar rules, 500 basic vocabulary words, and 250 basic example sentences. Our previous research[6][7] has shown the knowledge base contains the necessary grammatical requirements to meet our purposes. It is constructed via bottom-up structures using low-level to high-level sections with the objective being to learn all of its information.

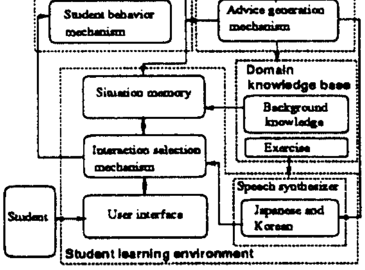

Figure 10. System Configuration

Exercises are simultaneously performed using monitor-displayed sentences and Japanese speech. While learning, a student can freely query the system about the text knowledge. As a student composes a Korean sentence, the system permits the students to access sample Korean texts in the knowledge base, auxiliary texts describing grammatical information and a Japanese-Korean dictionary. The system is sequentially utilized in each lesson in order to give appropriate compositional exercises. The student can learn according to a set of curriculums of teaching strategy, as well as ask grammatical questions and/or refer to grammatical rules.

5.3. Analysis Method of Composition Errors

The student diagnosis module evaluates student comprehension based on the results of his or her responses behavior and analyzes errors contained in the input compositional sentences.

5.3.1. Inferring the Cause of Composition Errors

The basic teaching method involves the student directly substituting Korean words for Japanese ones using grammatical similarities between the two languages. Actually, the student composes a Korean sentence from a problem stated in Japanese doing the following step process; (1) select Korean words, (2) arrange these words, and (3) transform the sentence into a string of Korean characters. This is shown in Fig 11.

Fig 11 shows that compositional errors occur in each step of the composition process. In addition, It is occurred to applied incorrect knowledge of Korean grammar by grammatical differences. The following essential knowledge is required to analyze grammatical differences and compositional errors occurring in the composition process:

1. Transformation of Korean character strings

2. Transformation of associated syntactic structures

3. Semantic relationships between Japanese and Korean

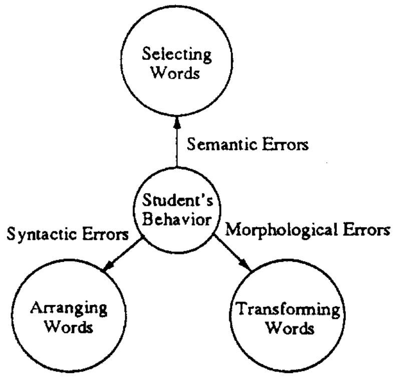

Figure 11. The Composition Process and Error Causes

This knowledge is incorporated into the proposed composition error analysis and advice-based teaching strategy method.

5.3.2. Method of Analysis

The system analyzes the input Korean sentence morphologically. The system uses the following three methods which are able to detect compositional errors and then diagnoses their cause.

1. Analysis using knowledge on transformation of Korean character strings: this method analyzes the cause of composition errors that occurred due to transformation of Korean character strings. Such errors are caused by incorrectly using Korean grammar rules, e.g., conjugation, inflection, and phonemic rules.

2. Analysis using knowledge on transformation of syntactic structures: this method analyzes errors due to differences between the syntactic structures of the two language. Such errors occur by incorrectly arranging the selected Korean words.

3. Analysis using knowledge on semantic relationships between Japanese and Korean: this method analyzes errors due to semantic differences between two language. Such errors occur by incorrectly selecting the Korean words, e.g., select of the postposition, select of auxiliary verbs.

5.4 Teaching Module

The teaching module contains built-in teaching strategy and an advice generation mechanism. Teaching strategy is based on the principle that the Japanese and Korean languages are grammatically similar. Initially, the student directly substitutes Korean words for Japanese ones using grammatical similarities. The advice generation mechanism provides explanations of the cause of composition errors using a text editor and synthesized Japanese and Korean speech. The advice includes information such as error type, error contents, hints, and grammar explanations. This generated advice is based on the results of error analysis. The advice that explains grammatical similarities and differences between the two languages are given through synthesized Japanese and Korean speech.

5.5 Student Learning Environment

The student's learning environment for a student using the system consists of a student interface that provides simulations for presenting domain knowledge and allows the student to take more initiative as learning activities progress. The speech synthesizer helps the student deliver information contained in the system, asks questions related to the composition exercises and gives advice to the student. A situation memory stores the students behavior.

6 CONCLUSIONS

We have proposed two new methods, i.e., one for synthesizing arbitrary Japanese and Korean sentences and another for using the ICAI system with speech synthesis capability. The main feature of our method is to the whole pitch pattern and phonemic environment of real speech as much as possible. Since the system does not require a model to generate the pitch pattern, human-like prosodic features are well preserved in the synthesized speech. To improve the quality of the system, we must account for different grammatical combinations contained in phrases and construct more appropriate rules for selecting basic sentences and/or basic phrases. Towards this end, it is necessary to quantitatively examine the influence of various grammatical combinations on prosodic features.

We believe that the introduction of speech synthesis made the ICAI system more effective for the language training. The method may be applied to other languages by analyzing the grammatical similarities and dissimilarities between the target languages.

REFERENCES

[1] H.Fujisaki and H.Kawai, "Realization of Linguistic Information in the Voice Fundamental Frequency Contour", Proc. ICASSP'88, pp.663-666, 1988.

[2] Y.Sagisaka, "On the Prediction of Global F0 Shape for Japanese Text-to-Speech" Proc. ICASSP'90, pp.325-328, 1990.

[3] H.Inagaki, T.Mochida, T.Kobayashi and K.Shirai, "Prosody Control and Phrase Unit Speech Synthesis", Proc. Autumn Meeting ASJ'92, 1992 (in Japanese).

[4] T.Mochida, T.Kobayashi and K.Shirai, "Speech Synthesis of Japanese Sentences Using Large Waveform Data-Base.", Technical Report of IEICE SP93-91, pp.95-100 1993.

[5] N.Sakurai, T.Mochida, T.Kobayashi and K.Shirai, "Prosodic Pattern Generation Based on Sentence Speech Data-Base", Spring Meeting ASJ'94, 1994 (in Japanese)

[6] K.K.Lee, T.Konishi, A.Takagi, K.Shirai and H.Ohara, "Applying Grammar Similarities to the Japanese-Korean Composition Training ICAI System for the Method of Composition Error Analysis and Advice Strategy", Technical Report of IPSJ SIG Notes, Vol.93, No.9, 93-CE-25-5, pp.41-48, Jan. 1993.

[7] K.K.Lee, T.Konishi and K.Shirai, "Evaluating Composition Error Analysis on the Japanese-Korean Composition Training ICAI System", Technical Report of IPSJ SIG Notes, Vol.94, No.10, 94-CE-31-5, pp.35-43, Jan. 1994.