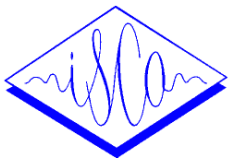

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Improvement of the Synthetic Speech Quality of the Formant-type Speech Synthesizer and Its Subjective Evaluation

Norio Higuchi, Hisashi Kawai, Tohru Shimizu and Seiichi Yamamoto

KDD R&D Laboratories
Kamifukuoka-shi, Saitama, 356 Japan

ABSTRACT

The authors have recently improved the synthetic speech quality of the Japanese speech synthesizer, which was developed for a special-purpose word processor named "Pasokon Talk" three years ago. The peculiarities of this system were using phonemes as synthesis units and generating all acoustic parameters based on production rules.

The major differences between the previous and current systems concern: (1) the method for control of the voice fundamental frequency contour, especially the phrase component of the generation model for the voice fundamental frequency contour proposed by Fujisaki, (2) the method for control of the formant frequencies and formant bandwidths, and (3) the characteristics of the voicing source.

In order to verify the improvement of synthetic speech quality quantitatively, (1) intelligibility tests of Japanese syllables and (2) opinion tests of naturalness have been performed. The results of comparative subjective evaluation tests show that the synthetic speech of the current system has an almost equal intelligibility and a much better grade of naturalness in comparison to that of the previous system.

I. INTRODUCTION

Three years ago, the authors developed a Japanese speech synthesizer which used phonemes as synthesis units and generated all acoustic parameters based on production rules. This speech synthesizer worked as part of a special-purpose word processor named "Pasokon Talk", and made the conversion to synthetic speech from text files which were created with this word processor [1] - [4].

"Pasokon Talk" is the first system developed for practical use in Japan in which the method of "Speech Synthesis by Rule" was used to generate all acoustic parameters. This method was selected because of the possibilities of the speech synthesis system which can synthesize the speech of various languages including Japanese, and also can synthesize synthetic speech with various tones spoken by various speakers such as an young male speaker, an old female speaker and so on. The most suitable tone of the synthetic speech is not always the same. For example, the announcement of an emergency must be stimulating to arouse people's attention while synthetic speech for proof-reading has to be subdued to offset listeners' fatigue.

The intelligibility of the synthetic speech generated by the previous system was good enough to communicate any Japanese sentence, but its naturalness was not so good and it sounded like a robot because of the difficulty of refining all the production rules. The previous system also had an other problem in that it could not convert the text files which were created with other word processors.

Recently the authors have made converting any text files and also improved the synthetic speech quality. In this paper, we will describe the necessary modifications and the results of subjective evaluation tests. The conversion of *kanji-kana*-mixed text files are described in other papers in this conference [5].

II. CONFIGURATION OF THE JAPANESE SPEECH SYNTHESIS SYSTEM

The Japanese speech synthesizer receives an alphabet string which expresses a Japanese sentence in Hepburn style with prosodic symbols from the linguistic processor which determines the readings and accent types of phrases (called *bunsetsu* in Japanese), based on the results of the morphological analysis of *kanji-kana*-mixed text files [5]. Each input character is converted to a standard phonemic symbol with standard segmental and suprasegmental features, and then classified into allophones after the segmental and suprasegmental features which reflect the phonemic environment and the accentuation. Several kinds of

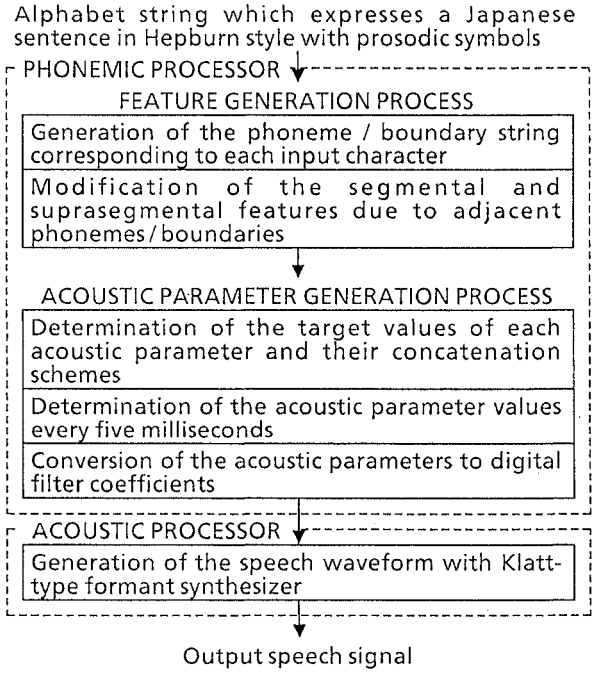

Fig. 1. Block diagram of the Japanese speech synthesizer.

boundaries, whose durations are zero, are also generated and expressed by boundary symbols. For each phoneme acoustic parameters are assigned based on rules using these segmental and suprasegmental features.

The total number of segmental and suprasegmental features including features related to phoneme boundaries is sixty. In addition to the conventional segmental and suprasegmental features, features related to phoneme boundaries, distinction of male / female voice and speech rate are also defined in the broad definition of the term.

The block diagram of the Japanese speech synthesizer is indicated in Fig. 1. The Japanese speech synthesizer is broadly divided into two components, the phonemic processor and the acoustic processor. The phonemic processor executes the feature generation process and the acoustic parameter generation process. The acoustic processor uses the acoustic parameters which are passed from the phonemic processor every five milliseconds to generate the speech waveform using a Klatt-type cascade / parallel speech synthesizer [6].

III. IMPROVEMENT OF SYNTHETIC SPEECH QUALITY

To assess the synthetic speech quality generated by "Pasokon Talk", a systematic diagnostic subjective evaluation consisting of nine kinds of tests was performed in 1988 [7]. The results of the tests indicated that though the system had already reached the required level for intelligibility of speech, there was room for improvement as far as producing a natural-sounding voice.

To improve the synthetic speech quality, the method of controlling each acoustic parameter was examined. The major changes based on the results of the examinations concerns:
(1) the method of controlling the voice fundamental frequency contour,
(2) the method of controlling the formant frequencies and formant bandwidths, and
(3) the characteristics of the voicing source.

3.1 Controlling the Voice fundamental frequency contour

The method of controlling the voice fundamental frequency contour affects naturalness of the synthetic speech significantly. In the previous system there were two major problems in the control of the voice fundamental frequency contour.
(1) The amplitude of the phrase components of the generation model for the voice fundamental frequency contour was smaller than that of natural speech.
(2) The prosodic symbols, which are used as markers to control the phrase components and the length of the pauses, are almost directly related to the punctuation marks of the original text.

The generation model for the voice fundamental frequency contour used here is that proposed by Fujisaki [8], [9]. As a consequence of the second problem, the voice fundamental frequency rises / falls unnaturally in some cases. For example, the voice fundamental frequency rises excessively if commas appear too frequently. On the other hand, the phrase component decays almost to zero and the movement of the voice fundamental frequency becomes monotonous if commas are too distant. Though several elaborations for the position and the amplitude of the phrase component were adopted in the previous system to prevent such phenomena, the result was not satisfactory and the rules were complicated needlessly.

In the current system both the linguistic process and the acoustic process play important roles cooperatively. The former decides an abstract peak level of the phrase component at the symbolic level based on the types of *bunsetsu*, the local syntax and the number of morae, and the latter determines an amplitude of the phrase command based on the distance from the preceding phrase command and its amplitude. Table 1 shows the symbols which are used for the definition of the prosodic symbols. For each boundary between *bunsetsu*, an amplitude of a phrase component and a length of a pause can be controlled independently. Two symbols are therefore assigned to each boundary between *bunsetsu*. Each single symbol used in the previous system to determine both an amplitude of a phrase component and a length of a pause at the same time [1] is converted to a standard combination of two symbols which are used in the current system for an upper compatibility. For examples, symbols " / ", " , " and " ; " with neither preceding nor following boundary symbols are converted into combinations of two symbols "/ @", " , %" and " ; #", respectively.

3.2 Controlling the Formant Frequencies and Formant Bandwidths

The method of controlling the formant frequencies and formant bandwidths has also been changed to improve the synthetic speech quality.

In the previous system, the fourth and fifth formant frequencies and formant bandwidths were constant for each speaker, while the first to the third formant frequencies and formant bandwidths were controlled independently. It is thus difficult to prevent excessive proximity of several formants, and this makes the synthetic speech unpleasant to listeners. To eliminate this problem, all parameter values of formant frequencies and formant bandwidths including the fourth and fifth formants are combined into a set of

Table 1. List of prosodic symbols.

Symbols indicating the amplitude of a phrase component	
/	Reset of the phrase component (mainly at the beginning of a sentence)
,	Addition of large impulse to the phrase component (mainly at the beginning of a clause)
; :	Addition of impulse to the phrase component (mainly at the beginning of a phrase)
+	Addition of impulse to the phrase component (mainly at the middle of a phrase)
<sp>	Without addition to the phrase component
Symbols indicating the length of the pause	
@	Long pause and breath
%	Long pause without breath
#	Short pause
$	Without pause

formants which keep proper distances from each other based on a stochastic analysis of the formant distributions of natural speech. Those values are then substituted for all formant frequencies and formant bandwidths at the same time by applying a formant rule.

Though the number of acoustic parameters including the fourth and fifth formant frequencies and formant bandwidths has increased, the total execution time does not increase so much because of the simultaneous substitution of all formant frequencies and formant bandwidths.

3.3 Characteristics of the Voicing Source

In addition to the above-mentioned two improvements, the characteristics of the voicing source are also changed to make the voice quality softer and to increase the accuracy of the voice fundamental frequency. The voicing source based on the polynomial model proposed by Igarashi *et al.* [10] was adopted here. In the previous system impulses were used for the excitation, but in the case of the female voice the accuracy of the voice fundamental frequency was about 4% which is larger than the just noticeable difference of the voice fundamental frequency. In the current system the minimal step of the time alignment for the waveform of the voicing source is a quarter of the sampling interval using an interpolation of the waveform of the voicing source. The accuracy of the voice fundamental frequency has therefore become about 1%.

IV. SUBJECTIVE EVALUATION TESTS

In order to verify the improvement of synthetic speech quality quantitatively, intelligibility tests of Japanese syllables and opinion tests of naturalness were performed. Four female subjects, who had no experience of listening to synthetic speech, listened to the synthetic speech with a headphone (Sennheiser Type HD-250) in a sound-proof room. The tests continued for ten days, that is five days in a week for two weeks. The total time of the subjective evaluation tests was about forty hours.

4.1 Intelligibility Tests of Japanese Syllables

Intelligibility tests of 100 and 116 Japanese syllables were performed with or without preceding phonemes. The 100 Japanese syllables are exactly the same as those in the conventional list of the articulation score, and the 116 Japanese syllables contain additional 16 syllables which are often used in words of foreign origin. The condition of the preceding phonemes is as follows: nothing (which means utterance-initial), /a/, /i/, /u/, /e/, /o/ and /aN/.

Table 2 gives the results for the 100 Japanese syllables. The subjective evaluation tests show that the synthetic speech of the current system has an almost equal intelligibility, although the method of controlling the formant frequencies and formant bandwidths was changed to make formant trajectories smooth in continuous speech. The correct recognition rate goes down only 1.6 percent. According to the intelligibility tests, correct recognition rates of certain consonants such as /g/ and /d/ are very low. It is therefore possible to make the intelligibility of the synthetic speech generated by the current system higher than that of the previous system if the rules concerning those consonants are improved.

The mean correct recognition rate of the synthetic speech generated by the previous system was lower than that reported previously [7]. The following two reasons are considered.
(1) Since total time of the subjective evaluation tests is 40% of that of the previous tests, the effect of learning by the subjects had not yet reached saturation.
(2) The synthetic speech of words and sentences was used for the subjective evaluation tests in the previous tests, but it was not used in the current tests. There was therefore a difference of learning effects.

4.2 Opinion Tests of Naturalness

To evaluate naturalness of the synthetic speech quantitatively, the following two opinion tests were performed.
(1) Preference test of 116 Japanese syllables [7].
(2) Evaluation test for naturalness of sentences.

Opinion scores used in these tests are listed in Table 3. The results of the preference tests of 116 Japanese syllables are not described in this paper, but they are being used to improve the intelligibility of the synthetic speech. As there is a strong correlation between the preference scores for the syllables of the synthetic speech and their intelligibility [11], the preference scores are indicators of the problems to be eliminated.

Four kinds of synthetic speech were prepared to analyze the effects due to the difference of the voice fundamental frequency contour and the difference of the spectral characteristics. The spectral characteristics mean the formant frequencies, formant bandwidths and the characteristics of the voicing source. Each of them was synthesized using the old or new method to control the voice fundamental frequency contour and the spectral characteristics. The results of the opinion tests for naturalness of the sentences are indicated in Table 4. The results show that the synthetic speech of the current system has a much better grade of naturalness in comparison to that of the previous system, and that the effect due to the difference of the voice fundamental frequency contour which is 0.8 in the opinion score is much larger than that of the spectral characteristics which is 0.3.

The opinion score of the synthetic speech using the combination of the new version of the voice fundamental frequency contour and the old version of the spectral characteristics which had the lowest score shows that proper control of spectral characteristics is a

Table 2. Comparison of percentage of syllables judged correctly by four female subjects in hearing tests for intelligibility of 100 kinds of Japanese syllables.

Preceding phoneme	Previous system	Current system	Difference
-	64.5	63.0	- 1.5
/a/	69.0	67.0	- 2.0
/i/	69.5	67.0	- 2.5
/u/	65.5	70.0	+ 4.5
/e/	74.5	64.5	- 10.0
/o/	70.5	73.0	+ 2.5
/aN/	68.5	66.5	- 2.0
Mean	68.9	67.3	- 1.6

Table 3. Scores used in the opinion tests to evaluate the naturalness of 116 Japanese syllables and sentences.

Opinion scores of 116 Japanese syllables	
4	The synthesized speech is very good as the designated syllable.
3	Small amount of unnaturalness is felt in the synthesized speech, but the synthesized speech has sufficient intelligibility as the designated syllable.
2	Large amount of unnaturalness is felt in the synthesized speech, but the synthesized speech is recognized as the designated syllable.
1	It is possible to recognize the synthesized speech as the designated syllable, but it is also possible to recognize it as other syllables.
0	It is impossible to accept the synthesized speech as the designated syllable.
Opinion scores of sentences.	
4	Very natural.
3	Natural.
2	Neutral.
1	Unnatural.
0	Very unnatural.

Table 4. Result of evaluation test for naturalness of the sentences.

Control of voice fundamental frequency contour	Control of spectral characteristics	
	Old version	New version
Old version	1.8	2.1
New version	1.3	2.9

necessary condition when the voice fundamental frequency is controlled more dynamically over a wider range.

V. CONCLUSIONS

We have described changes in the method of controlling the acoustic parameters in the Japanese speech synthesis system, and the results of subjective evaluation tests.

The major differences between the previous and current systems are (1) the method of controlling the voice fundamental frequency contour, especially the phrase component of the generation model for the voice fundamental frequency contour proposed by Fujisaki, (2) the method of controlling the formant frequencies and formant bandwidths, and (3) the characteristics of the voicing source.

The subjective evaluation tests to verify improvement of the synthetic speech quality quantitatively were (1) intelligibility tests of Japanese syllables and (2) opinion tests of naturalness. The results of the comparative subjective evaluation tests show that the synthetic speech of the current system has an almost equal intelligibility and a much better grade of naturalness in comparison to that of the previous system.

The authors will continue to improve the system by enhancing the naturalness of synthetic speech and diversifying the range of sounds supported by the Japanese speech synthesizer.

ACKNOWLEDGEMENTS

The authors would like to express their appreciation to Director Ono of the KDD R&D Laboratories and Deputy Director Urano. They would also like to thank all members of the Artificial Intelligence Group for fruitful discussions.

REFERENCES

[1] S. Yamamoto, N. Higuchi and T. Shimizu, "A Japanese Speech Synthesis System with Text Editing and Automatic Prosodic Control Facilities," The Second Joint Meeting of Acoust. Soc. Am. and Acoust. Soc. Jpn., J17, Nov. 1988.

[2] N. Higuchi, S. Yamamoto and T. Shimizu, "A Japanese Speech Synthesizer based on the Production Rules," The Second Joint Meeting of Acoust. Soc. Am. and Acoust. Soc. Jpn., J18, Nov. 1988.

[3] N. Higuchi, S. Yamamoto and T. Shimizu, "The Control of the Articulatory Parameters for the Japanese Speech Synthesis System by Rule," J. Acoust. Soc. Jpn., Vol. 45, pp. 426-433, June 1989 (in Japanese).

[4] S. Yamamoto, T. Shimizu and N. Higuchi, "The Control of the Prosodic Features for the Japanese Speech Synthesis System by Rule with Editing Functions," J. Acoust. Soc. Jpn., Vol. 45, pp. 426-433, June 1989 (in Japanese).

[5] T. Shimizu, N. Higuchi, H. Kawai and S. Yamamoto, "The Linguistic Processing Module for Japanese Text-to-Speech System," Proc. of ICSLP 90, Nov. 1990.

[6] D. H. Klatt, "Software for a Cascade / Parallel Formant Synthesizer," J. Acoust. Soc. Am., Vol. 67, pp. 971-995, March 1980.

[7] N. Higuchi, S. Yamamoto and T. Shimizu, "Evaluation of Intelligibility and Naturalness of the Synthetic Speech Generated with a Japanese Speech Synthesizer by Rule," Trans. IEICE Jpn., Vol. J72-D-II, pp. 1133-1140, Aug. 1989 (in Japanese).

[8] H. Fujisaki and H. Sudo, "A Model for the Generation of Fundamental Frequency Contours of Japanese Word Accent," J. Acoust. Soc. Jpn., Vol. 27, pp. 445-453, Sep. 1971 (in Japanese).

[9] H. Fujisaki and K. Hirose, "Analysis of Voice Fundamental Frequency Contour for Declarative Sentences of Japanese," J. Acoust. Soc. Jpn. (E), Vol. 5, pp. 233-242, Oct. 1984.

[10] K. Igarashi *et al.*, "A Study of Voice Quality Control Based on a Polynomial Model of Glottal Source," Trans. Comm. Speech Res., Acoust. Soc. Jpn., S89-142, March 1990 (in Japanese).

[11] S. Yamamoto, N. Higuchi and K. Matsuzaki, "Evaluation Methods for the Synthesized Speech considering the Effect of the Coarticulation," National Conference Record, Communications, IECE Jpn., p.506, Sep, 1986 (in Japanese).