Ann. Bull. RILP, No. 14, 201-207 (1980)

RANDOM PERTURBATIONS ON THE PITCH PATTERN AND THE NATURALNESS OF SYNTHETIC SPEECH

Hiroshi Imagawa, Shigeru Kiritani and Shuzo Saito

1. INTRODUCTION

In the present study, random perturbation was introduced in the pitch pattern of synthetic speech, and the relationship between the amplitude of random perturbation and the naturalness of the pitch pattern of the synthesized speech was investigated.

In the Linear Prediction Speech Analysis-Synthesis System, pitch frequencies are transmitted at every analysis frame. On the other hand, there have been several studies to present the pitch pattern by a small number of parameters[1)2)3)]. Specifically, in the field of speech synthesis by rule, it was reported that synthetic speech with a fairly good quality of pitch pattern can be obtained by using piecemeal linear representation of the pitch curve. In that method, the pitch pattern was obtained by selecting an appropriate time moment within each mora and linearly interpolating the pitch frequencies at these moments. In the following study, as an economized method of expression of the pitch pattern, the points on the pitch curve where the curvature would be the greatest were selected by visual inspection, and the pitch pattern was then approximated by connecting these successive points by straight lines.

In the present study, the effects of random perturbation were examined for the following two cases:

1) Random perturbation was added to the pitch frequency of each frame.
2) Random perturbation was added to the parameter values in the piecemeal linear representation of the pitch pattern.

2. EXPERIMENTAL PROCEDURES

2.1 Speech Analysis-Synthesis Method

In order to synthesize speech having randomly perturbed pitch patterns, the PARCOR analysis-synthesis system was used. A block diagram of the system is shown in Figure 1.

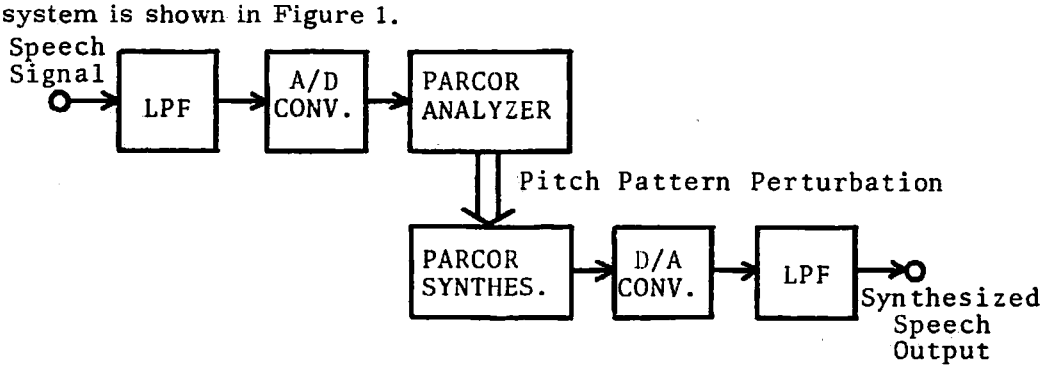


Fig. 1 Schematic diagram of the test system

First, the test sentences were recorded in a sound-proof room. These utterances were then lowpass-filtered to 5kHz, sampled at 10kHz and digitized in 10 bits. PARCOR analysis was made every 6.4 msec. on 19.2 msec. of Hamming-windowed speech. The order of the analysis was 12. The voiced/unvoiced judgment was performed based on the values of the maximum peak in the autocorrelation function of the residual wave.

The random perturbations were introduced by the following method:

Type 1 Random perturbation was added to the pitch frequency of each frame for the original pitch pattern.

Type 2-1 Random perturbation was added to the pitch frequency at selected points on the piecemeal linear approximated pitch pattern.

Type 2-2 The timing of these points was randomly perturbed.

2.2 Speech materials

The test sentences are shown below:

1) Aoi umi no e wa yama no ue no ie ni arimasu.
2) Yama no ue no ie ni wa aoi umi no e ga arimasu.
3) Yama no ue no mori ni giri no ane no ie ga arimasu.
4) Aoi umi no e wa mado no mae no ani no hon ni arimasu.

These sentences were uttered by two adult male speakers of the Tokyo dialect. In these test sentences, unvoiced consonants were avoided in order to obtain a continuous pitch pattern. Sentence 3) consists of words of the flat type accent, while test sentence 4) consists of words having the accent kernel. Thus, sentence 4) shows greater pitch inflections. Sentences 1) and 2) are random accent types. The pitch patterns of these sentences are shown in Figure 2, and the piecemeal linear approximations of the pitch patterns are also indicated in Figure 3. Compared with the original speech, the synthesized speech produced based on the piecemeal linear approximation of the pitch pattern has been recognized as natural.

2.3 Testing Method

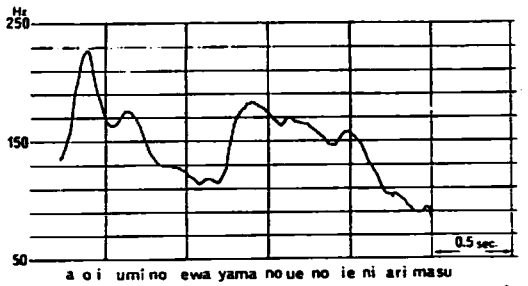
With regard to the aforementioned conditions, three series of test sentences were synthesized, type 1, type 2-1 and type 2-2. In each series, 5 to 6 different degrees of random perturbation amplitude were chosen, and 5 speeches were synthesized for each degree of random perturbation.

Thus 25 to 30 synthesized speech patterns were presented to the subjects, in random order six times each, through a loud-speaker system in a soundproof room.

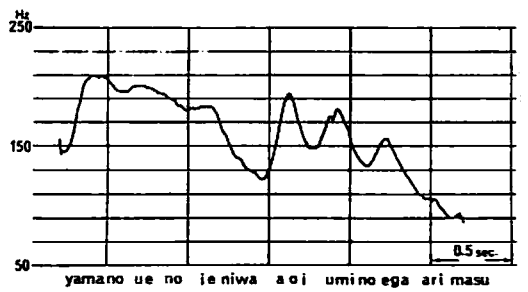
The subjects were requested to decide whether the pitch pattern of the synthesized speech was 'natural' or 'unnatural'. It was a forced choice. Three adult males with normal hearing participated in the listening test.

3. RESULTS

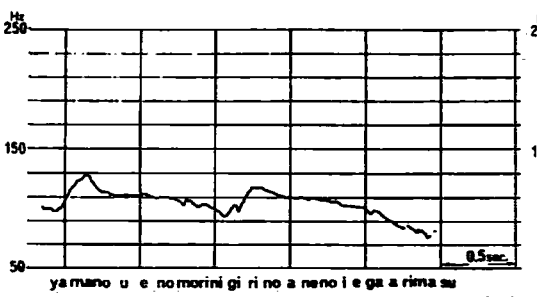
Figure 4, 5 and 6 show the results of the pitch perception test for each type of synthesized speech, for the three different conditions of random perturbation. In each figure the abscissa represents the standard deviation



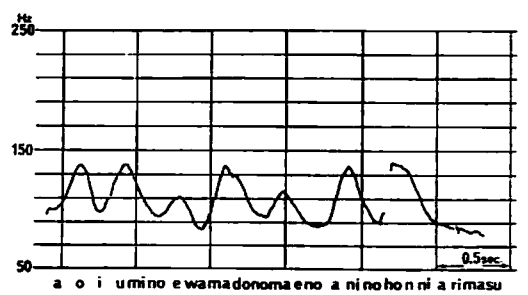
Sentence (1)



Sentence (2)

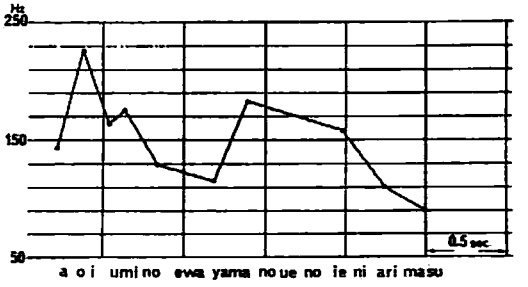


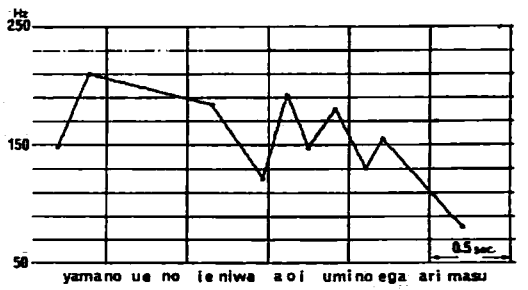
Sentence (3)

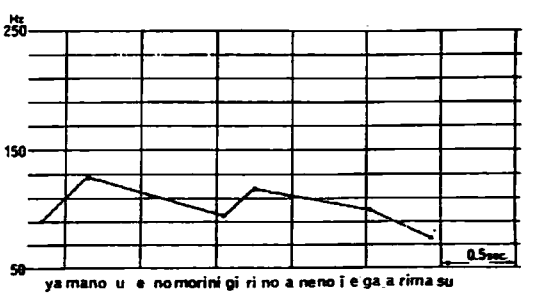


Sentence (4)

Fig. 2 Original pitch pattern







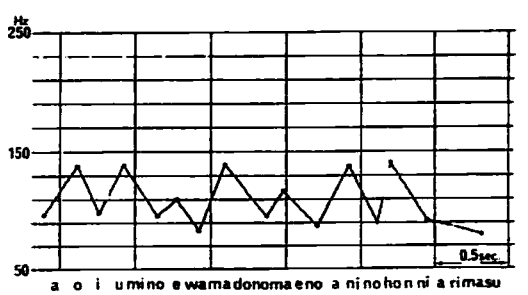


Fig. 3 Piecemeal linear approximated pitch pattern

of random perturbation in Hz or in msec., and the ordinate indicates the rate of allowance for pitch perception. The rate of allowance for pitch perception is defined as the ratio of the number of responses where the pitch pattern of the synthesized speech is identified as 'natural' by the subject to the total number of synthesized speech patterns produced under the same amplitude of random perturbation. The rate of allowance for pitch perception averaged over the three subjects is shown in percentages.

3.1 Pitch perception under random perturbation type 1

Figure 4 shows the relations between the naturalness of each synthesized speech type and the amplitude of random perturbation.

The frequency band width of random perturbation was restricted to 5Hz, for it had been assured that the synthesized speech was natural even if the frequency band width of the original pitch pattern was lowpass-filtered to 5Hz.

For synthesized speech items 1-4, it was observed that the allowance for pitch perception monotonously decreased in accordance with the increase in the amplitude of random perturbation. It was also found that the allowable threshold (i.e., the amplitude of random perturbation at which the allowance for pitch perception is 50%) was around 7Hz for test sentences 1), 2) and 4), and was around 4Hz for test sentence 3) (see Table 1). In other words, it was revealed that the allowable threshold for a test sentence consisting of words having an accent kernel, namely the sentence that had greater pitch frequency inflections, was higher than that of a sentence with flat pitch pattern.

3.2.1 Pitch perception under random perturbation type 2-1.

Figure 5 shows the results of random perturbation type 2-1. This time it was also observed that, in each item, the allowance for pitch perception decreased as the amplitude of the perturbation increased. Table 1 shows the allowable threshold in Hz for each test sentence.

Table 1.

	Type 1	Type 2-1
Sentence 1	7Hz	11Hz
2	7	12
3	4	6
4	7	8

It can generally be said that the allowable thresholds were smaller in type 2-1 than those in type 1.

3.2.2 Pitch perception under random perturbation type 2-2

Figure 6 shows the results of random perturbation type 2-2. Here too it was observed that the allowance for pitch perception decreased as the amplitude of random perturbation increased, for every synthesized speech.

Table 2 shows the allowable threshold in msec. for each test sentence.

Table 2.

	Type 2-2
Sentence 1	50 msec.
2	50
3	55
4	35

The average duration for a mora in the test sentences used in this experiment was about 120 msec. From table 2, it was observed that the allowable threshold was about one half of the average duration of a mora for test sentences 1), 2) and 3), and about one third of the average duration of a mora for test sentence 4).

As a measure of the distortion in the pitch curve introduced by random perturbation, the square-roots of the squared mean values of the differences between the original pitch pattern and the piecemeal linear approximated pitch pattern were calculated. Table 3 shows these values at the allowable threshold for both type 2-1 and 2-2.

Table 3.

	Type 2-1	Type 2-2
Sentence 1	12Hz	10Hz
2	12	10
3	6	5
4	7	6

The values for type 2-2 were smaller than those for type 2-1.

4. CONCLUSION

Generally, the information on pitch frequency that is sampled at 100 – 200Hz is used for speech synthesis. For the sentences used in the present study, the pitch frequency band width restricted to 5Hz appeared to be sufficient. The pitch frequency information sampled at 100 – 200Hz appears to be rather redundant.

In the piecemeal linear representation of pitch pattern, the sampling rate was further reduced to five per second.

Additionally, for the test sentences used in this experiment, it has been shown that the pitch pattern would be more stable against noise when it is expressed by using the characteristic parameters than by using a series of periodically sampled values.

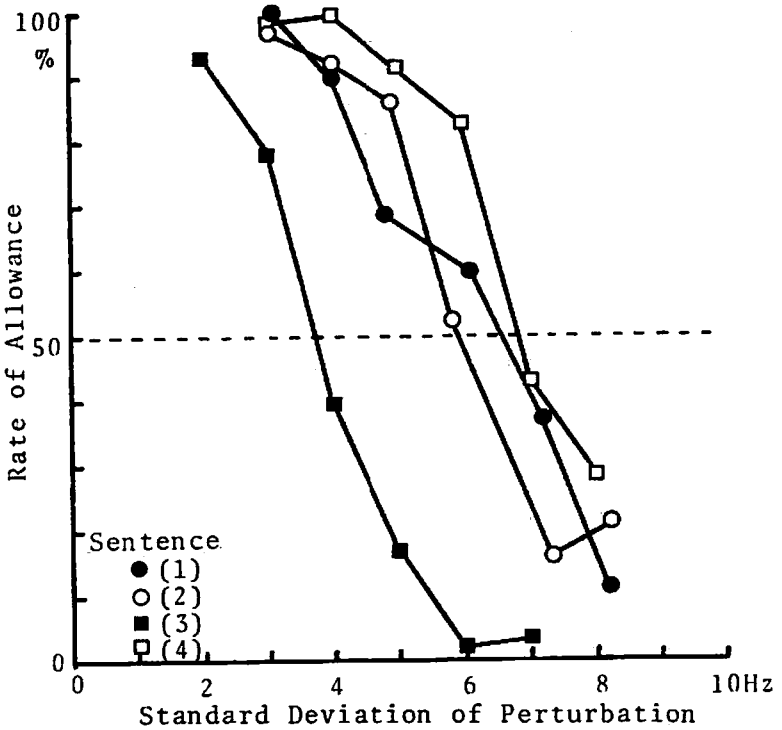


Fig. 4 Results of listening test Type 1

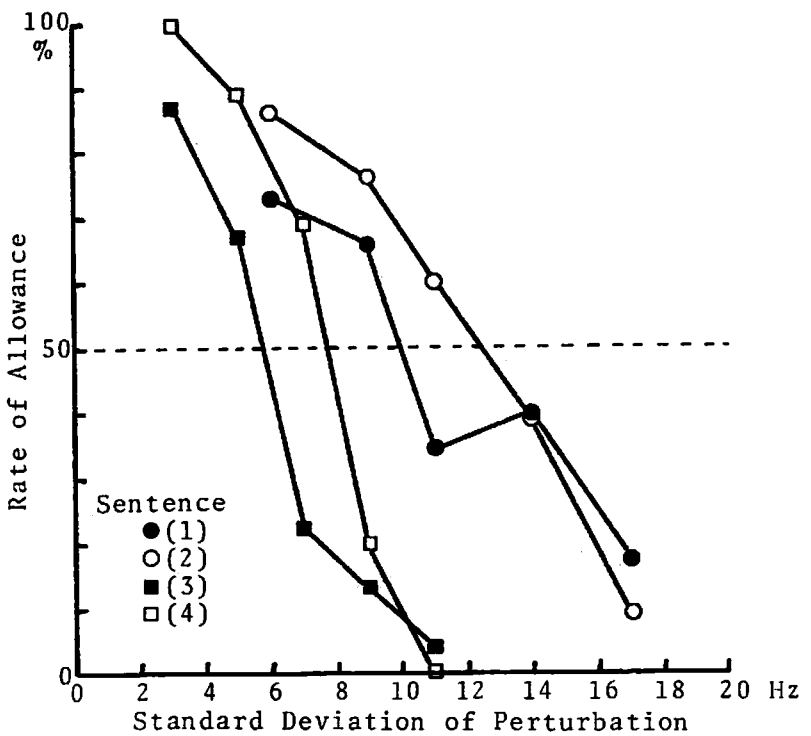


Fig. 5 Results of listening test Type 2-1

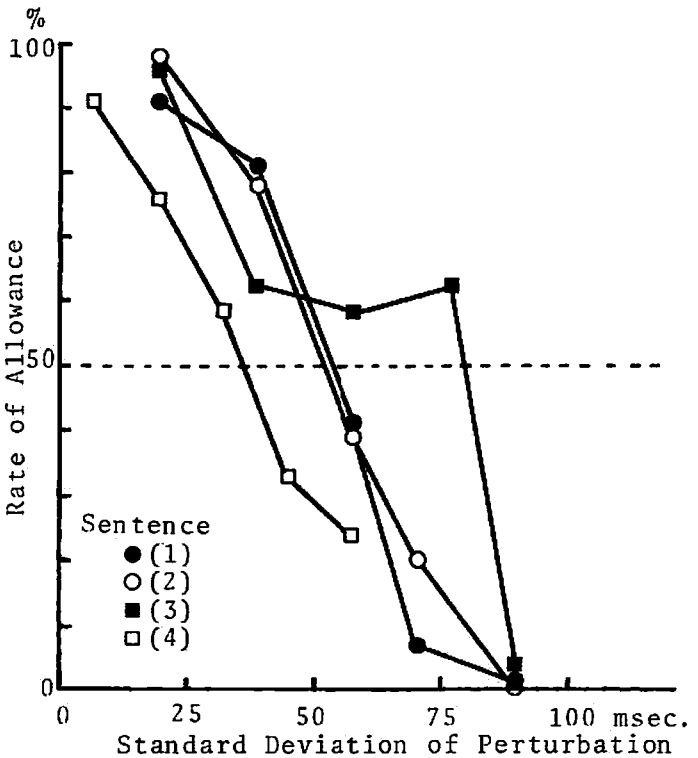


Fig. 6 Results of listening test Type 2-2

References

1) Saito, S., S. Hashimoto and H. Wakita (1967); "On the speech synthesis computer system by interphoneme transition unit," Rec. Fall Meeting, Acoust. Soc. Japan, 1-3-16, 111-112.

2) Fujisaki, H. and H. Sudo (1971); "A Model for the Generation of Fundamental Frequency Contours of Japanese Word Accent," J. Acoust. Soc. Japan 27, 445-453.

3) Hashimoto, S. and S. Saito (1971); "Prosodic rules for speech synthesis," Proceedings of the Seventh International Congress on Acoustics, Budapest, 23C1, 129-132.

Acknowledgment

This study was supported in part by a Grant in Aid for Scientific Research No. 540003 from the Japanese Ministry of Education, Science and Culture.