Ann. Bull. RILP (1982) No. 16, 63-71

STUDY ON THE PITCH PERCEPTION OF SENTENCES BASED ON THE F_0-CONTOUR GENERATION MODEL

Hiroshi Imagawa, Shigeru Kiritani and Shuzo Saito

1. Introduction

The pitch pattern of a spoken sentence speech varies with the speaker's identity, sentence structure, dialect, and so on. Fujisaki et al. proposed a model[1,2] which could approximate the pitch pattern of spoken sentences with high accuracy using a rather small number of parameters. In this paper, the effects of the model when the parameters were tentatively changed within a certain range on pitch perception for spoken sentences were studied.

2. Speech Materials and Analysis of Sentence Pitch Patterns

A male adult, speaker of Tokyo dialect, uttered the following two test sentences.

i) aoi umi no e wa ani no ie no mado no mae ni arimasu.
ii) yama no ue no mori ni giri no ane no ie ga arimasu.

Sentence i) has greater pitch inflections, as it is composed of words which have an accent kernel, whereas sentence ii) consists of words of the flat accent type.

These utterances were analized using a PARCOR analysis system. The fundamental frequency was extracted by detecting the maximum peaks in the autocorrelation function of the residual wave from the PARCOR analysis. The PARCOR analysis was made every 6.4 msec frame interval on 19.2 msec of the hamming-windowed speech signal which was sampled in 10 kHz. The extracted pitch patterns for these sentences are shown in Fig. 1.

In the model of sentence F_0-contour generation proposed by Fujisaki,[1,2] sentence F_0-contour is divided into voicing and accent components. By representing the amplitudes of the i-th voicing command and the j-th accent command as A_{vi} and A_{aj}, respectively, the output fundamental frequency $F_0(t)$ as a function of time (t) is given by:

$$\ln[F_0(t)/F_{min}] = \sum_{i=1}^{I} A_{vi}\left\{G_{vi}(t-T_{0i}) - G_{vi}(t-T_{3i})\right\} + \sum_{j=1}^{J} A_{aj}\left\{G_{aj}(t-T_{1j}) - G_{aj}(t-T_{2j})\right\},$$

where

$$G_{vi}(t) = \alpha_i t e^{-\alpha_i t} u(t),$$

$$G_{aj}(t) = 1 - (1+\beta_j(t))e^{-\beta_j t} u(t),$$

u(t): unit step function,
I: number of voicing commands,
J: number of accent commands,
T_{0i}: onset time of the i-th voicing command,
T_{3i}: offset time of the i-th voicing command,
T_{1j}: onset time of the j-th accent command,
T_{1j}: offset time of the j-th accent command.

The parameters of the model are determined by minimizing the mean squared error between the extracted pitch pattern and that of the model on the logarithmic frequency scale. Both the extracted pitch patterns and the best approximations by the model are shown in Fig. 2. The values of the parameters are listed in Table 1.

3. Procedure for the Listening Test

The test stimuli were produced by changing the parameters of the model shown in Table 1. In the present study, the timing parameters (T_{1j}, T_{2j}) and the amplitude parameters (A_{aj}) of the accent command were varied systematically. The pitch patterns were generated by the same model used in the analysis. In each stimulus item, only one of the parameters, T_{1j}, T_{2j} or A_{aj}, was changed and the other parameters were held constant at the values shown in Table 1. The variation range of each parameter for each accent command was selected on the basis of a preliminary listening test. T_{1j} and T_{2j} were changed by an interval of 25.6 msec (for sentence i)), 12.8 msec (for sentence ii)); while A_{aj} was changed by an interval of 0.2 (for sentence i)) and 0.1 (for sentence ii)).

By the use of these parameters from the model, a total of 373 pitch patterns were derived and fed into the PARCOR synthesizer to produce the test stimuli. These test stimuli were presented to the subjects in random order through a loudspeaker in a sound proof room. Four male adults with normal hearing participated in the listening test. The subjects were requested to judge whether the pitch pattern of each test stimulus was 'natural' or 'unnatural'. The choice was forced.

4. Results

The results for the effect of the onset time T_1 variation in each accent command of the stimulus sentence speech on pitch perception for the two sentences are shown in Fig. 3. The value of 0 (msec) in the abscissa represents the reference time for T_1 in each accent command as shown in Table 1. A negative value in the abscissa indicates an early shift in T_1, that is, an early commencement of the accent command, and a positive value represents a retarded shift in T_1.

The ordinate is the averaged percent value of the four listeners' responses as 'natural' for each test stimulus. Similar diagrams were obtained from the experiments on the offset time T_2 and the amplitude A_a variations. It was also observed that the percent responses as 'natural' decreased gradually corresponding to incremental changes in T_2 and A_a.

The permissible threshold for a natural pitch judgment for the sentences was defined as equal to 50% responses as 'natural'. These threshold values for the

three parameters T_1, T_2 and A_a were estimated from the experimental data and are illustrated in Fig. 4. In the Figure for A_a, a negative value in the abscissa represents a decremental variation in the amplitude of the accent command, and a positive one an incremental variation in the command amplitude.

In Fig. 4, the parameter values within the positive and negative permissible thresholds are expressed by solid lines, which are termed the permissible variation range of the model parameters for the pitch perception of spoken sentences, because synthesized speech having the pitch pattern generated by the use of the parameters within these ranges is perceived as 'natural' pitch.

There were several items which had no exact permissible thresholds, such as the negative variations in the amplitudes of the accent commands for the phrases 'umi no', 'e wa' and so on. In these cases, variations in the parameter values hardly resulted in a significant modification of the perceived pitch patterns of the sentences, because the reference values of A_a were relatively small. Similarly, the permissible threshold for the phrase 'arimasu' of the test sentence i) could not determined, as the time interval of the accent command was too short.

From the results shown in Fig. 4, it can be seen that,

(1) As for the direction of the time variation in the onset and offset times of the accent commands T_1, T_2, there was no significant difference between the permissible threshold values for the early and retarded variations in the two spoken sentences tested.

(2) There was no significant difference between the permissible threshold values for the onset and offset times of the accent commands T_1, T_2.

(3) For the two types of test material, i) had greater pitch inflections and ii) was rather flat, resulting in no significant difference in the permissible thresholds for the onset and offset times of the accent commands T_1, T_2.

(4) As for the amplitude of the accent command A_a, the permissible thresholds for sentence ii) were smaller than those for sentence i), whereas the direction of the amplitude variation had no significant effect on the permissible threshold for pitch perception.

Using the permissible threshold values of the model parameters as shown in Fig. 4, the limiting pitch patterns permissible as 'natural' are shown in Figs. 5 (a), (b) and (c) for the parameters T_1, T_2 and A_a, respectively.

5. Discussion

There are several factors causing an 'unnatural' pitch sensation for spoken sentences beyond the limiting pitch patterns shown in Fig. 5. For variations in T_1 and T_2, a perceptible change in the accent type in the phrase was often observed. For variation in A_a, on the contrary, a change in the accent level was perceived for sentence i) and a change in the pitch balance among the phrases was observed for sentence ii).

Using the pitch patterns shown in Fig. 5, the amount of distortion in the pitch patterns of the sentences at the permissible thresholds of the model parameters was calculated. The amount of distortion was defined as the root mean squared error between the reference pitch pattern and that at the permissible threshold. This

was calculated for the three parameters, T_1, T_2 and A_a as shown in Fig. 6.

From these results, it can be seen that,

(1) As for the permissible thresholds for the onset and offset times of the accent commands T_1, T_2, the variance in the threshold values in various phrases was less than that in Fig. 4. It can also be observed that the permissible thresholds for sentence ii) are smaller than those for sentence i). It seems that the permissible threshold of the pitch pattern is larger in sentences having greater pitch inflections than those having relatively flat pitch.

(2) As for the permissible threshold of the amplitude of the accent command A_a, the variance in the threshold values in various phrases was similar to T_1 and T_2, but terminal phrases were affected too by the structure of the clause, as is in 'e wa' or 'ie no' of sentence i).

(3) Comparing the amount of distortion in the time parameters T_1, T_2 and the amplitude parameter A_a at the permissible thresholds for the pitch patterns, the permissible amplitude distortion was larger in general.

Summarizing our present study, it may be concluded that,

(1) The permissible thresholds for the onset and offset times of the accent commands are not affected in pitch perception by differences in spoken sentence or also by the direction of the temporal variation in the onset and offset times. The variance of the threshold values in various phrases decreases with the amount of distortion, which is the root mean squared error of the pitch pattern of the spoken sentence.

(2) The permissible threshold for the amplitude of the accent command A_a is not affected by sentences in terms of the root mean squared error expression of the pitch pattern, whereas several differences can be observed in the simple amplitude expression for different sentences. Comparing the amount of distortion in the permissible thresholds between the onset/offset times and the amplitude of the accent command, the latter is in general larger than the former.

References

1. Fujisaki, H., K. Hirose and K. Ohta (1979); "Acoustic features of the fundamental frequency contours of declarative sentences in Japanese." Ann. Bull. RILP, 13, 163-173.
2. Fujisaki H. and K. Hirose (1981); "Fundamental frequency control in spoken sentences." Trans. of the Committee on Speech Research, Acoust. Soc. Japan. S80-73.

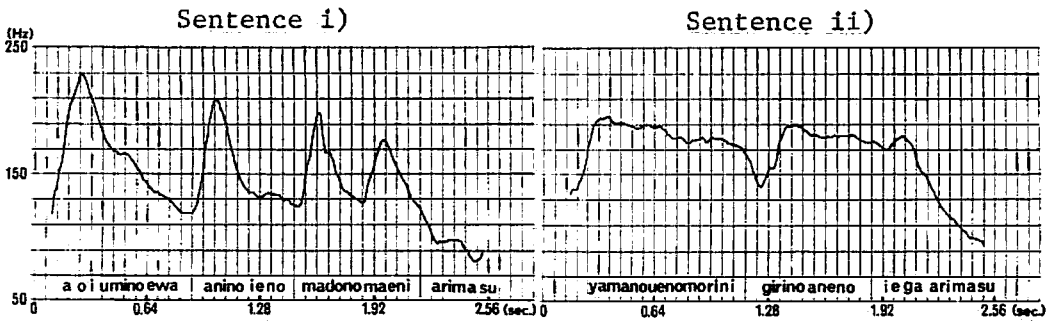

Fig. 1 *Extracted pitch patterns for the two test sentences.*

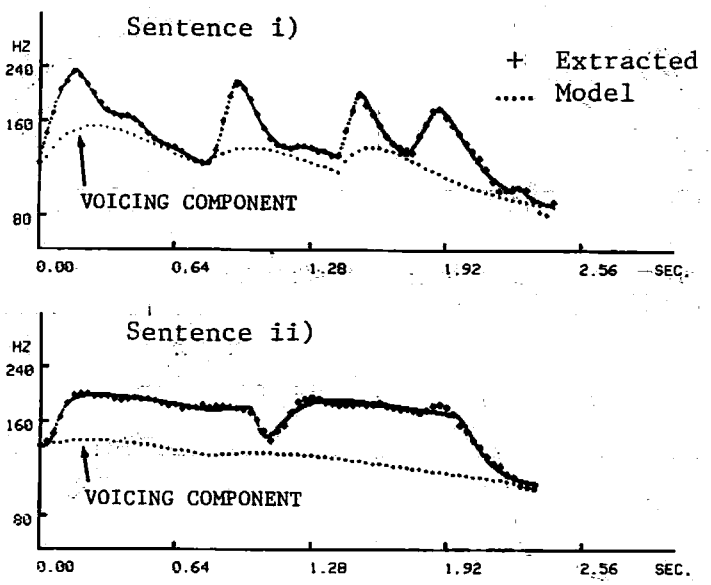

Fig. 2 *Approximated pitch patterns by the model.*

Table 1 *Parameters of pitch patterns for the two test sentences.*

Sentence	Fmin (Hz)	i	T_0 (sec)	T_3 (sec)	Av	α (sec^{-1})	j	Phrase	T_1 (sec)	T_2 (sec)	Aa	β (sec^{-1})
i)	80.0	1	-0.070	2.500	1.77	2.9	1	'aoi'	-0.026	0.166	0.52	17.1
		2	0.794	2.500	0.76	2.9	2	'umi no'	0.314	0.454	0.12	22.4
		3	1.421	2.500	0.72	4.6	3	'e wa'	0.550	0.608	0.10	12.2
							4	'ani no'	0.826	0.922	0.86	18.5
							5	'ie no'	1.152	1.389	0.12	25.1
							6	'mado no'	1.408	1.504	0.61	21.1
							7	'mae ni'	1.728	1.869	0.92	10.9
							8	'arimasu'	2.246	2.266	0.18	45.8
ii)	80.0	1	-0.435	2.500	1.50	1.5	1	'yama....	0.026	0.979	0.33	35.4
		2	0.794	2.500	0.44	1.2	2	'giri....	1.024	1.952	0.44	15.6

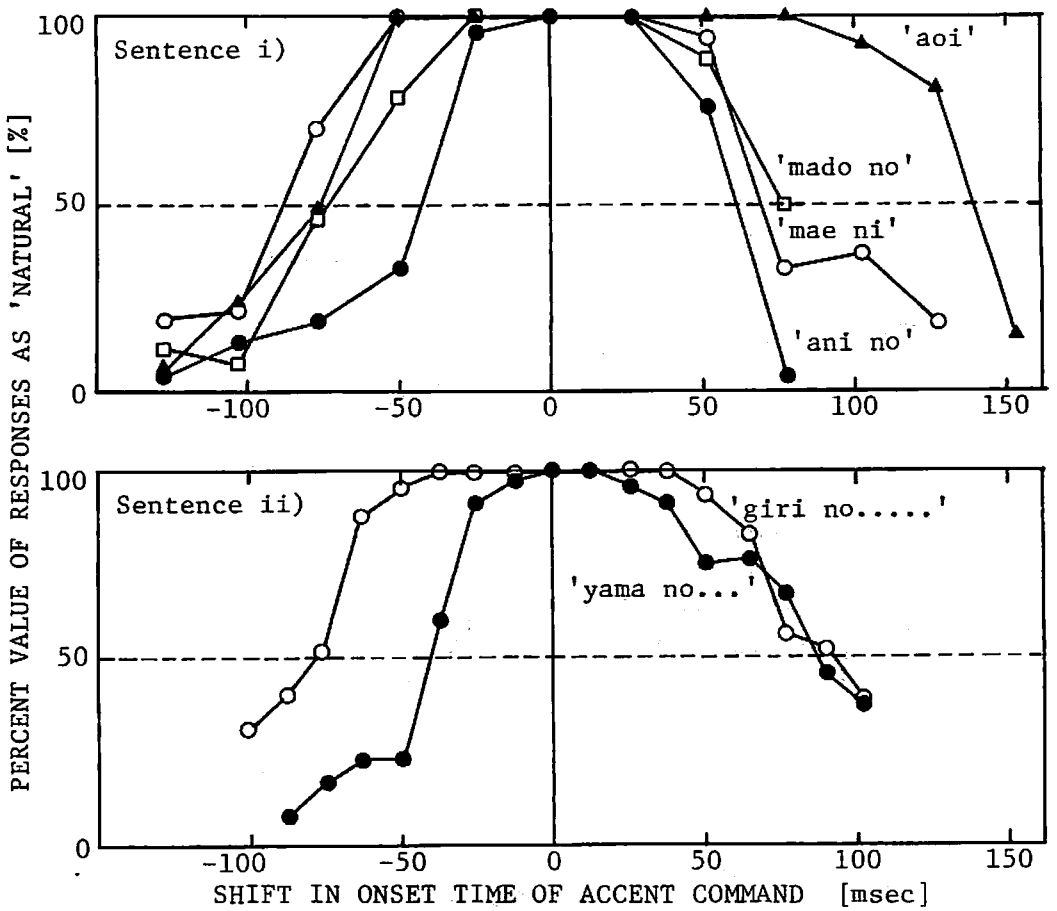

Fig. 3 *Percent value of responses as 'natural' on pitch perception against the onset time (T_1) variation in each accent command.*

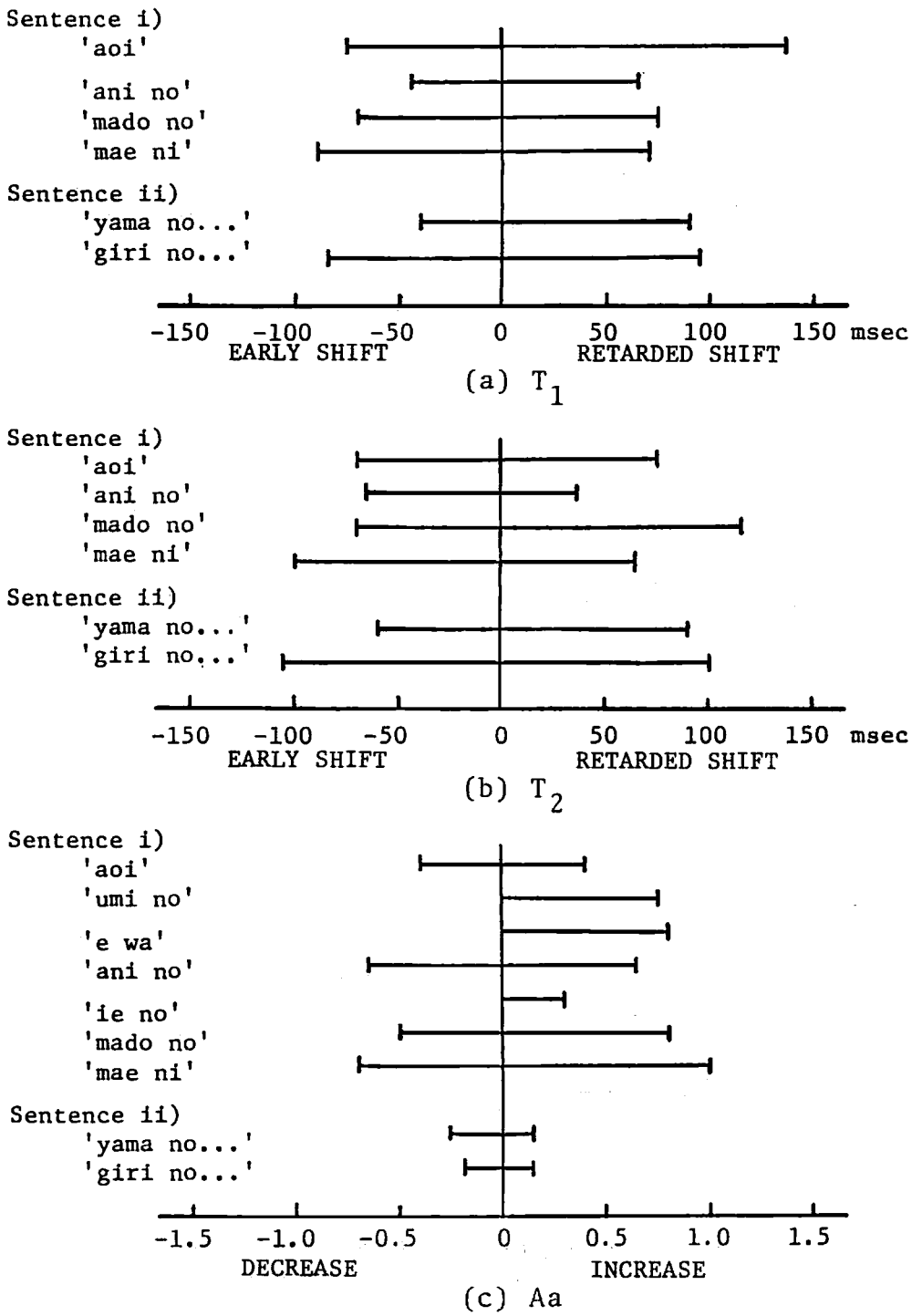

Fig. 4 *Permissible variation range of the model parameter.*

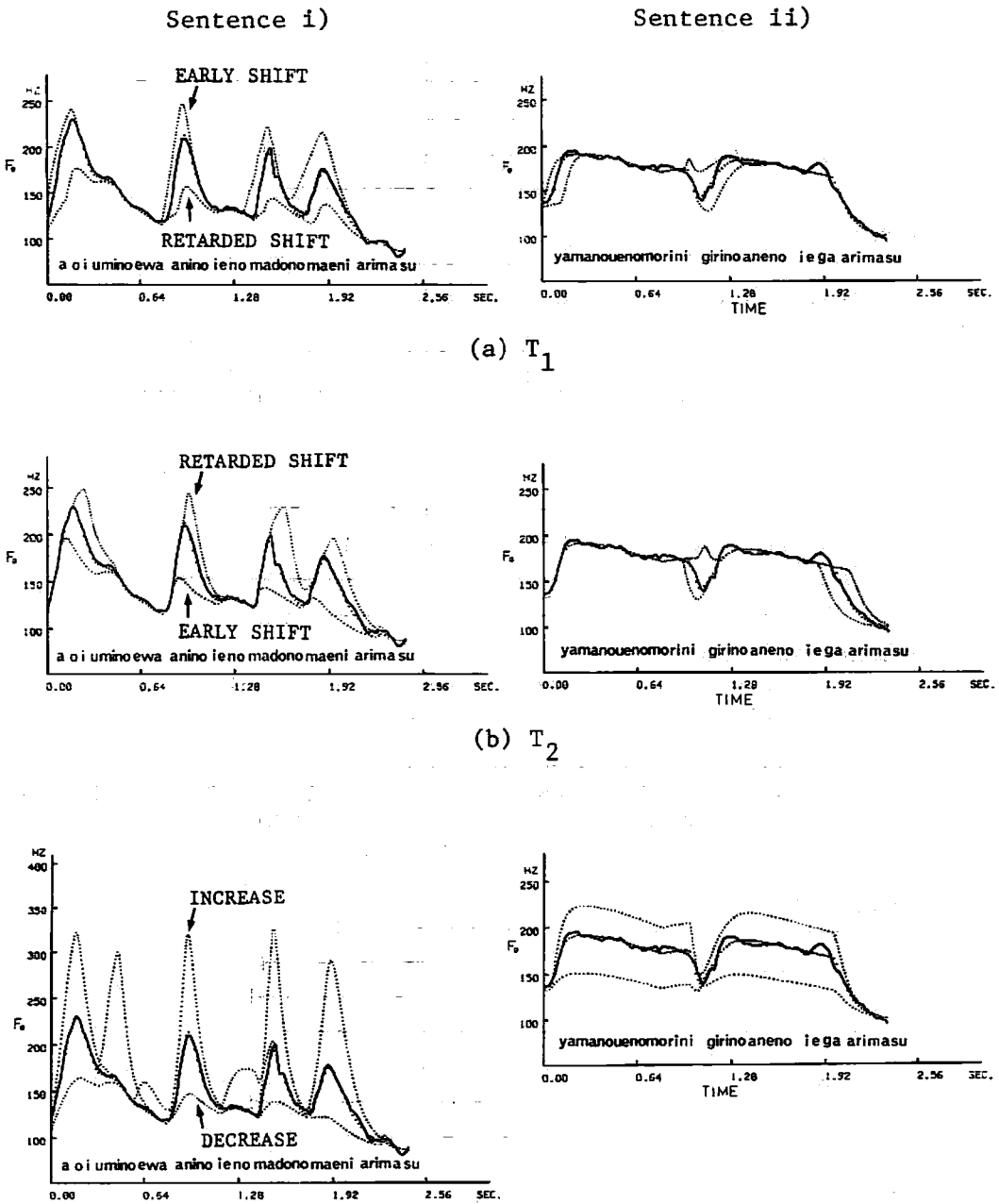

Fig. 5 *Limiting pitch patterns permissible as 'natural'. Solid line represents the extracted pitch pattern.*

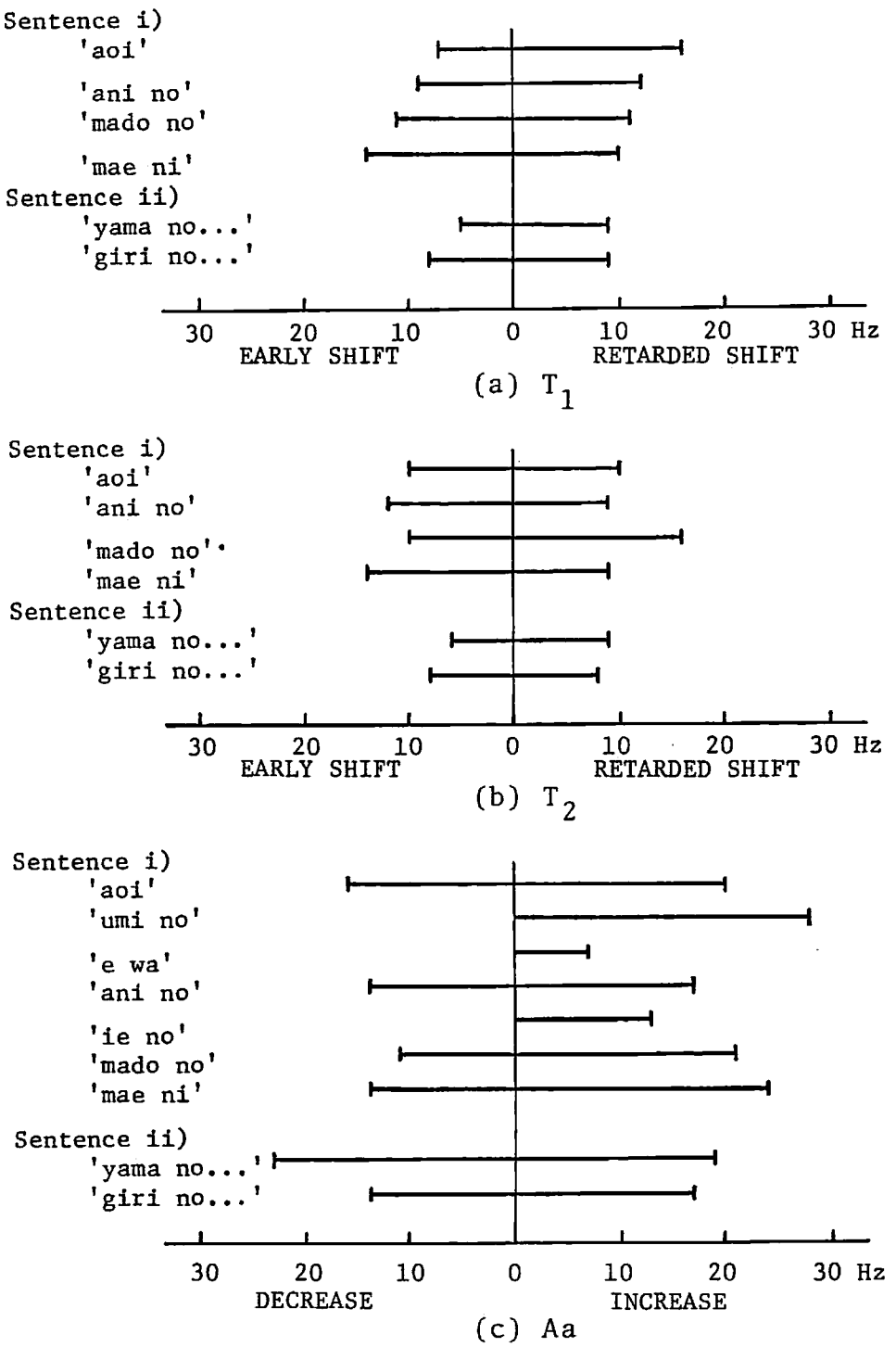

Fig. 6 *Amount of distortion in the pitch pattern of the sentence at the permissible threshold of the model parameter.*