TR－H－017

Contextual effects on acceptability for modification of segmental duration in words

Hiroaki KATO　Minoru TSUZAKI
Yoshinori SAGISAKA

1993. 7. 27

ATR人間情報通信研究所
〒619-02 京都府相楽郡精華町光台2-2　☎07749-5-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-7749-5-1011
Facsimile: +81-7749-5-1008

Contextual effects on acceptability for modification of segmental duration in words[1]

Hiroaki Kato, and Minoru Tsuzaki
ATR Human Information Processing Research Laboratories, Hikaridai, Seikacho, Kyoto, 619-02 Japan

Yoshinori Sagisaka
ATR Interpreting Telecommunications Research Laboratories, Hikaridai, Seikacho, Kyoto, 619-02 Japan

---

[1] This paper was submitted to the Journal of the Acoustical Society of America in July 1993.

ABSTRACT

The acceptability of temporal naturalness was measured for various vowel segments in isolated words by modifying original segmental durations. A large-size perceptual experiment employing 1462 stimuli of 70 segments revealed that word acceptability is affected by the segment attributes and context such as the position in a word, phoneme type, and tone accent. An ANOVA test demonstrated that the acceptable range of temporal modification was narrower (1) for the first moraic segment in a word than for the third one, (2) for vowel /a/ than for vowel /i/, and (3) for high-tone segments than for low-tone segments; the effect of (3) was the weakest of the three. An additional experiment showed that the position in a word and phoneme type also affected the temporal discrimination threshold consistently, suggesting that they function at a perceptual stage.

PACS numbers: 43.71.Es, 43.71.Gv, 43.66.Mk

INTRODUCTION

Segmental durations in spoken Japanese will vary according to segment attributes and context. Therefore, extensive studies have been made on some acoustical characteristics of Japanese segmental durations, to achieve high-quality speech synthesis (Kaiki, et al., 1992; Takeda, et al., 1989). In the above studies, the effects of the following control factors were quantitatively confirmed: 1) vowel color, 2) adjacent phonemes, 3) position in a word, 4) mora count in a word, and 5) speaking rate.

However, it would be premature to infer synthesis rules for duration control from acoustical measurements alone. This is because such an approach might extract and imply rules that merely control small deviations, which hardly influence human impressions of naturalness, and conversely, some perceptually important factors might not be picked out because of few physical contributions. We are convinced that control rules need to be evaluated and rearranged, from the viewpoint of perceptual sensitivity, in order to succeed in providing really acceptable synthesized speech.

Many papers have been published on auditory sensitivity to durational modification (Abel, 1972a,b; Creelman, 1962; Ebata, et al., 1974; Small and Campbell, 1962), but studies employing speech segments as stimuli are rare. Through several pioneering studies (Bochner, et al., 1988; Carlson and Granström, 1975; Fujisaki, et al., 1975; Huggins, 1972; Klatt and Cooper, 1975; Sato, 1977), some general characteristics of perceptual sensitivity to the temporal aspect of speech could be roughly estimated; however, these studies used small sets of stimuli to observe perceptual characteristics precisely. Therefore, their arguments on context or attribute dependency of perceptual sensitivity have not always agreed with each other. For instance, Huggins, et al. showed that the just noticeable difference (JND) for segmental durations varies as a function of phoneme type for five types of segments, /ɔ, l, p, m, ʃ/, each employing one or two phonemic contexts e.g., /pɔ**p**, ɔl**ɔ**, ɔp**ɔ**, u**p**i, u**m**u, i**ʃ**i/ (the boldface represents the segments in question). On the contrary, no significant difference due to a variation in phoneme type could be found among the JNDs reported by Fujisaki, et al.; they used four types of segments, /o, s, t, m/, each of which contained only one phonemic context. Since control rules for speech synthesis become more and more precise in later years, the perceptual data should be updated in order to validate such rules. What

is needed is a model that will describe perceptual sensitivity to the temporal aspect of speech in various circumstances.

The aim of this study was to investigate perceptual sensitivity to modification of segmental duration, from the viewpoint of context and attribute dependency using a large set of speech stimuli. We limited the segments in question to vowels in isolated words, because vowels are dominantly varied depending on context variation in spoken Japanese (Sagisaka and Tohkura, 1984). We also limited the number of levels within the factors under study to two; this enabled us to prepare enough stimuli at each level for reliable statistical analyses. Mainly two factors were examined: 1) the position of the segment within a word and 2) the phoneme type of the segment. These two are the major control factors found in acoustical analyses; they have also been touched upon in previous perceptual studies. The segments of interest were chosen from one of two different positions, word-initial or word-medial, to test the effect of the position; they also possessed one of two different vowel colors, /a/ or /i/, to test the effect of the phoneme type. In acoustical measurements (Sagisaka, et al., 1984), inherent vowel durations are known to be shorter at the word-initial position than at the word-medial position and also shorter for /i/ than for /a/. Several psychophysical measurements using steady-state non-speech signals have shown that the shorter the duration is, the shorter the discrimination threshold becomes (e.g. Abel, 1972a; Small and Campbell, 1962). Therefore, if an analogous perceptual characteristic had functioned dominantly in the current experiments, this meant the existence of a higher sensitivity at shorter segmental durations. We also included the factors of tone accent and F0 contour, which possibly influence perceptual sensitivity to the temporal aspect of speech, although they have not been considered as control factors for segmental duration in Japanese.

In the current paper, the first experiment provided a large amount of perceptual data to investigate context and attribute dependency, through an acceptability test on durational modification of the chosen segments. In the second experiment, we measured the temporal discrimination thresholds using a part of the stimulus set employed in the first experiment. This test clarified whether or not the results of the acceptability test really reflected perceptual sensitivity. Then, perceptual sensitivity to durational modification of speech segments was discussed with regard to context and attribute dependency.

I. EXPERIMENT 1: ACCEPTABILITY

A. Method

1. Design

The first experiment was conducted to investigate the effect of segment attributes and context on perceptual sensitivity to durational modification using a large set of samples. According to a statistical analysis on acoustic characteristics, vowels in the first mora are inherently shorter than those in the third mora and vowel /a/ is inherently longer than vowel /i/ (Sagisaka, et al., 1984). Thus, we chose to focus on the first-third contrast for the factor of position in a word and /a/-/i/ contrast for the factor of phoneme type. The contrast for the phoneme type was a difference in vowel colors since we limited the segments to be manipulated to vowel portions. In addition, we included the factor of tone accent; although this factor does not affect Japanese segmental duration control at all (Kaiki et al., 1992), it is still an open question as to whether this factor affects perceptual characteristics such as acceptability. For the factor of tone accent, the contrast between segments with high-tone and low-tone was chosen.

2. Stimuli

The word samples were chosen from the large-scale Japanese speech database constructed at ATR (Sagisaka, et al., 1990). The chosen words were four-mora words, excluding the samples with vowel successions or geminated consonants which may disturb the temporal regularities observed in open syllable successions. We selected 70 segments from 63 sample words. Table 1 shows the distribution of segment attributes concerning the factors considered. Figure 1 shows the distribution of vowel durations, which was defined by manual labeling in the selected segments.

Table 1. Number of selected segments corresponding to each of the conditions considered in the first experiment, the acceptability test.

	1st mora	3rd mora	total
/a/	21 (6)*	22 (15)	43 (21)
/i/	14 (0)	13 (13)	27 (13)
total	35 (6)	35 (28)	70 (34)

*Values in round brackets are numbers of high-tone segments.

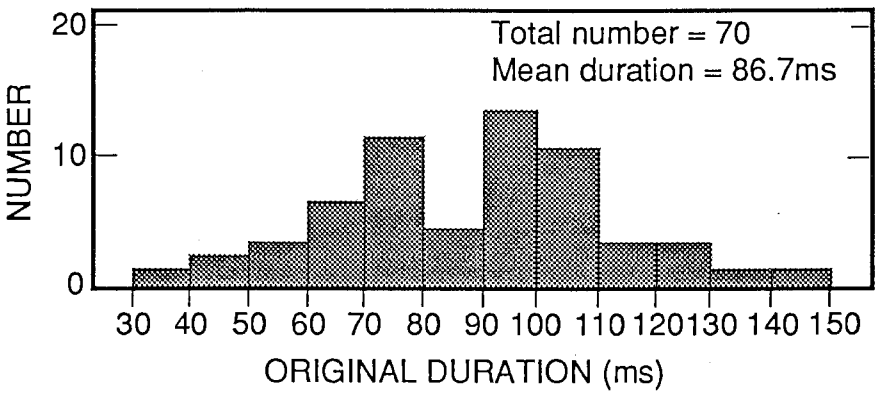


Figure 1. Distribution of original duration of the selected segments in the first experiment, the acceptability test. Original duration was manually defined by trained labelers. (Sagisaka, et al., 1990)

These sample words were uttered by one male professional announcer, sampled at a rate of 12 kHz, and digitized to 16 bits. The word stimuli were prepared using the selected word samples, by means of a cepstral analysis and resynthesis using the Log Magnitude Approximation Filter (Imai and Kitamura, 1978), which were carried out at a 2.5 ms frame interval. The speech quality by this synthesis technique was natural enough for subjects to evaluate acceptability. Then, the duration of the vowel portion in each of the selected segments was varied over a range from -50 ms to +50 ms of the original duration in 5 ms-step, resulting in 21 different steps; a few of the shorter vowels with durations close to or less than 50 ms were not shortened to the full range. The vowel durations were modified by the deletion or repetition of a selected number of frames, which contained synthesis parameters, before resynthesis. The manipulated frames were chosen evenly from the whole range of each vowel portion. In total, 1462 word stimuli were prepared.

3. Procedure

The synthesized stimuli were separated into ten groups; each group contained stimuli derived from seven of the seventy original segments. For each group, stimuli were randomized to four different series and recorded onto a digital audiotape with a DAT recorder (SONY DTC-55ES). One of the randomized stimuli came once after a short marker tone in each trial, with an inter-trial interval of four seconds. A pair of short marker tones was inserted at the beginning of each series and after every ten trials. The recorded stimuli were presented diotically to subjects through headphones (STAX SR-Λ Professional) in a sound-proof room. The presentation level was approximately 81 dB (A-weighted), which was measured (at a nucleus of vowel /a/) by a sound level meter (Brüel & Kjær 2133) mounted on an artificial ear (Brüel & Kjær 4153). The subjects heard four series a day, which were based on one of ten stimulus groups. Hence, each subject participated for ten days in total, and responded four times for each stimulus. Each one-day session lasted about one hour and a quarter, involving four 15-minute sessions and three short breaks.

The task of the subjects was to rate each stimulus as to how acceptable the temporal unnaturalness was, using the following seven subjective categories.

3: very unnatural

2: unnatural
1: rather unnatural
0: undeclared
-1: rather natural
-2: natural
-3: very natural

The obtained scores will be referred to as "acceptability scores" in the following sections.

4. Subject

Seven adult female subjects participated in the experiment. All of them were native speakers of Japanese with normal hearing.

B. Results

1. Parameters representing acceptability

The obtained acceptability scores were pooled over all of the subjects and all of the target segments, then plotted as a function of change in duration of target segments (Figure 2). A parabolic fitting using a least squares criterion was chosen as a good approximation for this plot (superimposed in Figure 2).

To deal with variations in acceptability score due to changes in duration numerically, we adopted a parabola fitting to the acceptability scores obtained for each of the 70 target segments per each of the seven subjects, giving 490 parabolic curves. Each of these curves will be referred to as an "acceptability curve" in this paper. Figure 3 shows a clear example illustrating the difference in acceptabilities between two different segments.

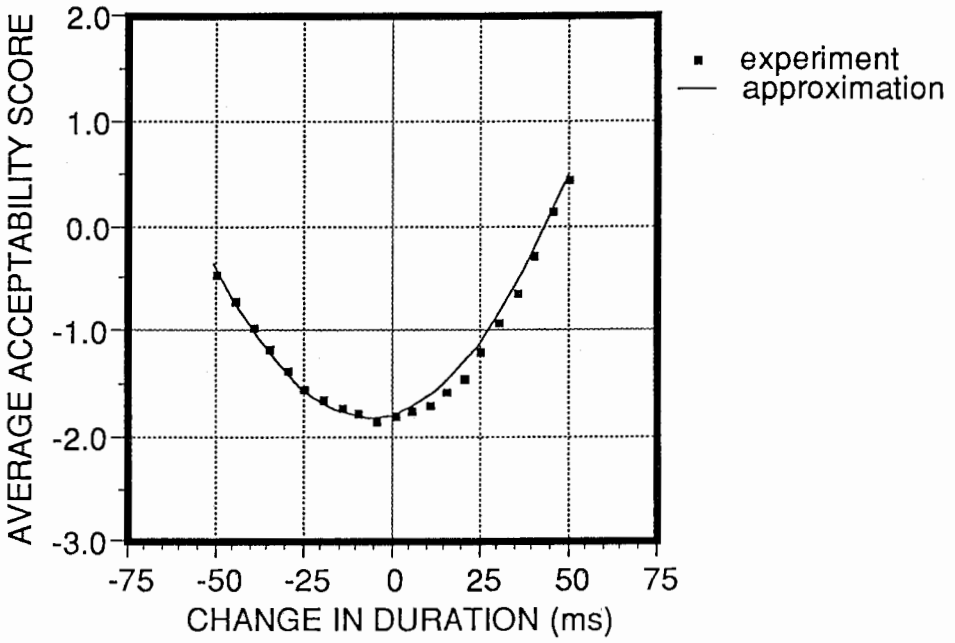


Figure 2. Obtained acceptability scores and their approximation as a function of change in duration. The acceptability scores were pooled over all of the subjects and all of the target segments, then plotted with squares. A parabolic fitting was chosen as a good approximation for this plotting.

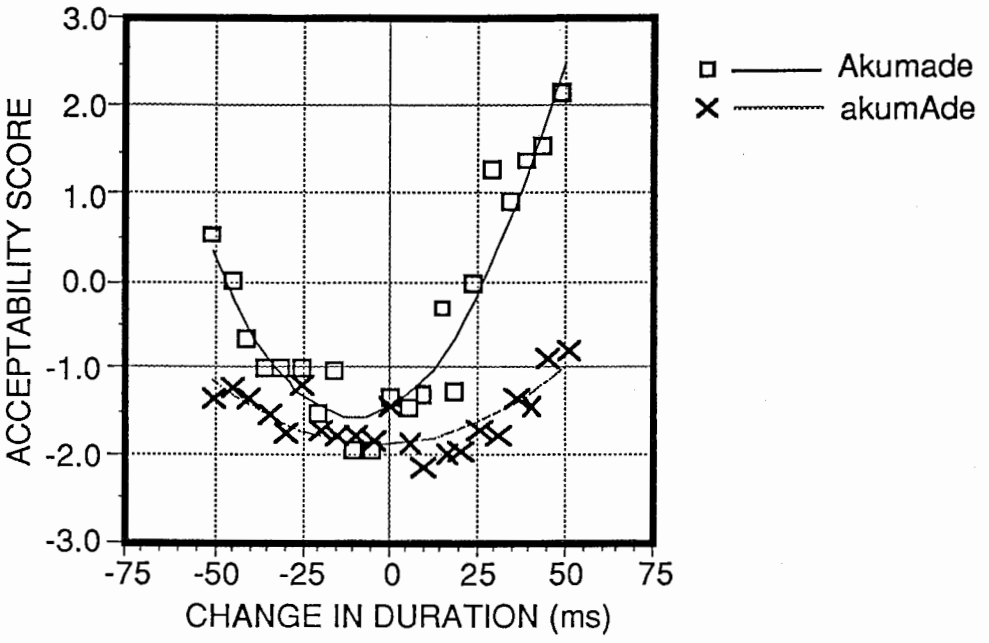


Figure 3. An example illustrating a difference in acceptability curves. The acceptability scores and the acceptability curves for two segments in one word, i.e. the first moraic high-tone segment and the third moraic low-tone segment in the word "akumade" (translated: to the bitter end). The capitals in the legend indicate the segments in question. Scatter plots show the tendency that the acceptability score varies according to the durational change more sensitively to the first moraic segment of "akumade" (plotting with open squares), than to the third moraic segment (plotting with crosses). The two acceptability curves trace this tendency clearly.

Statistical analyses will be shown in the following sections using two indices of the acceptability curve; one is the sharpness of the curve (the second order polynomial coefficient) which reflects the acceptable range, and the other is the axis of the curve which corresponds to the center of the acceptable range. The horizontal value of the axis indicates how the most preferred duration shifts from the original one. Ahead of the analyses, some unreliable data were excluded in accordance with the following two criteria: (1) the acceptability curve whose second polynomial coefficient was not more than zero was not included, because the axis of such a curve never corresponds to the most preferred duration, and (2) the acceptability curve whose axis was extremely apart (more than six times sigma) from the distribution center was not included. Thus, nine curves were excluded by the first criterion and three curves by the second; as two of them were excluded by both criteria, 480 curves were finally obtained for the analyses.

2. Sharpness of the acceptability curve

The influence of the factors considered were examined by an analysis of variance (ANOVA) on the sharpness of the acceptability curve. We included the factor of original as-produced duration in addition to the mentioned three factors. It was because the original duration had a wide range as previously shown in Figure 1. Thus, a four-way ANOVA of repeated-measures was performed with position in a word, phoneme type, tone accent, and original duration as the main factors. As shown in Figure 4, panels (a) to (c), the influence of the factors of position, phoneme type, and tone accent were found.

The factors of position in a word and phoneme type were significant [$F(1,479) = 26.5$, $p < 3.95 \times 10^{-7}$ for position in a word; $F(1,479) = 26.6$, $p < 3.72 \times 10^{-7}$ for phoneme type]. The sharpness was higher for the segment at the first position than for that at the third position, and also higher for vowel /a/ than for vowel /i/. This means the subjects accepted larger temporal modification at the third vowel in a word and at the vowel /i/ than at the above corresponding counterparts. The interaction between these two factors was not significant. Although the factor of tone accent was significant, the effect was much weaker than that of the first two [$F(1,479) = 7.69$, $p < 0.00578$]. This indicated the tendency that the sharpness was higher for the high-tone segments than for the low-tone segments. An interaction between this factor and the factor of

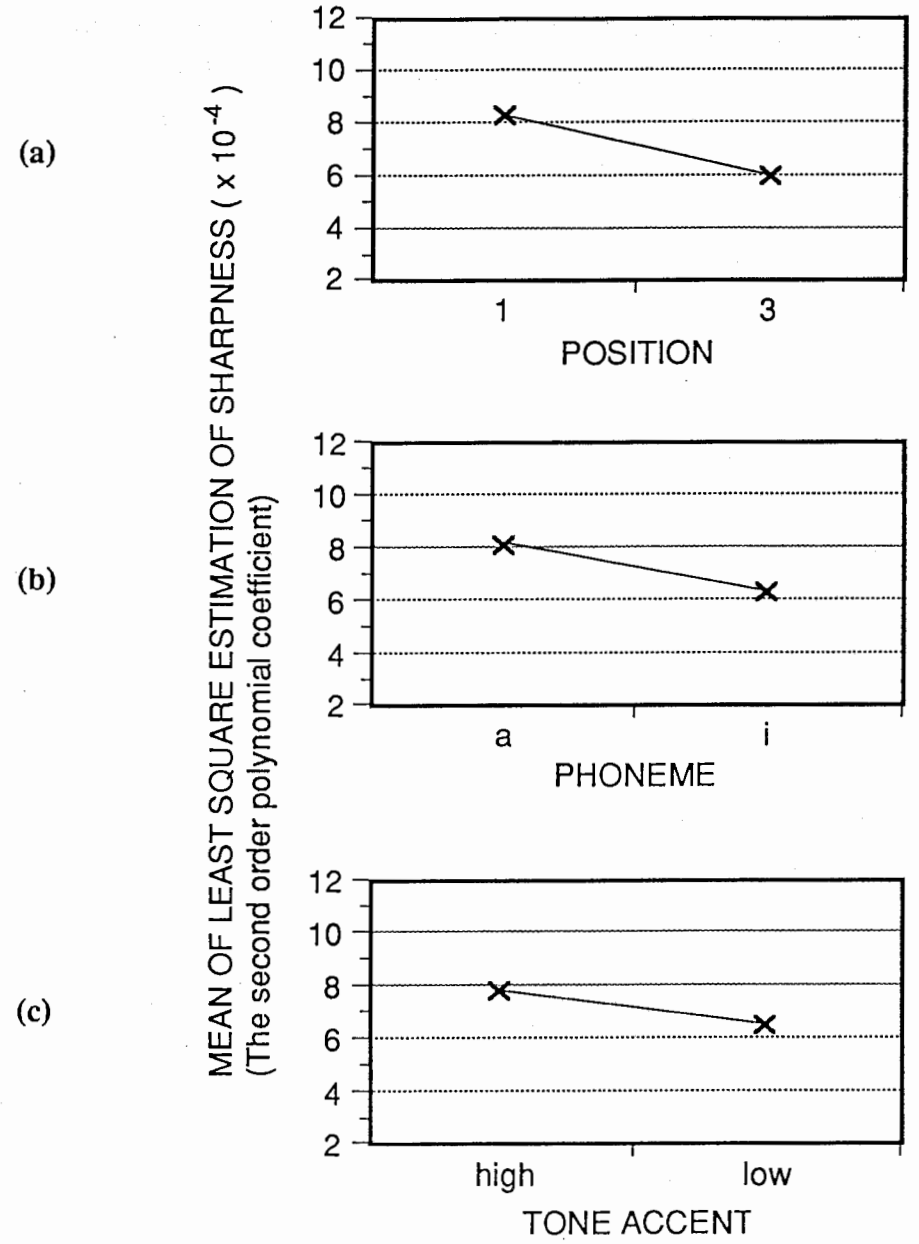


Figure 4. Mean of least square estimations of the sharpness for each level in each factor; they were calculated in the analysis of variance in the first experiment. The following effects are visible. The sharpness is higher (a) for the first mora than for the third mora, (b) for the vowel /a/ than for /i/, and (c) for the high-tone segments than for the low-tone segments. The higher sharpness corresponds with the narrower acceptable range.

position was found [$F(1,479) = 3.97$, $p < 0.0470$], suggesting the effect of tone accent tended to be strong at the third moraic position. Despite the wide variation of the original duration, the effect of the original duration was not significant [$F(1,479) = 0.545$, $p > 0.460$], and the correlation analysis indicated that no significant correlation between the original duration and the sharpness existed [$r = 0.0345$, $p > 0.451$].

3. Axis of the acceptability curve

A four-way ANOVA of repeated-measures was performed again for the horizontal value of the axis of the acceptability curve; the main factors were same as those in the first ANOVA. The results indicated that the factor of original duration was most significant [$F(1,479) = 24.5$, $p < 1.03 \times 10^{-6}$]. The horizontal value of the axis was negatively correlated with the original duration. This means that the longer the original duration was, the shorter the preferred duration shifted, and vice versa. The other two, the factors of position and phoneme type, were relatively less significant [$F(1,479) = 11.9$, $p < 6.19 \times 10^{-4}$ for position in a word; $F(1,479) = 13.4$, $p < 2.83 \times 10^{-4}$ for phoneme type]. The other main factor and the interactions were not significant.

II. EXPERIMENT 2: DISCRIMINATION THRESHOLD

The first experiment showed that two of the main factors found in acoustical analysis, the position in a word and the phoneme type, would also affect the perceptual evaluation. However, this experiment did not conclusively verify that these two factors do in fact affect the perceptual "sensitivity", which should be measured by discrimination thresholds in general. Evaluations on acceptability may be influenced by a cognitive process as well as by a perceptual process. Accordingly, it is possible that these two factors, which were effective in the first experiment, play their roles only in acceptance, but not in detection. A second experiment was therefore needed to measure the discrimination thresholds for the durational modification in order to verify whether the acceptability tests reflected detectability.

A. Method

1. Design

The factors of position in a word and phoneme type were designed in the same manner as in the first experiment; we employed a first-third contrast as the first factor and an /a/-/i/ contrast as the second factor. The factor of F0 contour was also included in the second experiment. The shape of the F0 contour could be considered a cue for discrimination because durational modification causes a change in the F0 slope when the F0 contour is not flat. We chose a contrast between segments with natural F0 contour and segments with flattened F0 contour as the third factor.

2. Stimuli

Two words were chosen from the database in the same manner as in the first experiment. The following set of four segments was employed.

Vowel /i/ in the first and the third mora of the word "shinagire", which means "sold out".

Vowel /a/ in the first and the third mora of the word "nameraka", which means "the state of being smooth".

The synthesis procedure for a half of the stimulus set, the group of segments with natural F0, was the same as in the first experiment, except that the modification range was from -60 ms to +60 ms and the step was 2.5 ms. The other half, with flattened F0 contour, were synthesized in the same manner as that of the first half except that the F0 values were fixed to the mean F0 of each of the target segments in the original utterances.

3. Procedure

Discrimination thresholds were measured by the transformed up-down paradigm[2]

---

[2] We occasionally inserted trials with physically identical pairs as check trials to prevent too short an estimation of the threshold. If a subject responded to a check trial with "same", such a response was regarded as a "false alarm", and the difference between the following pair was set larger than that between the pair previous to the check trial.

(Levitt, 1971) with two response alternatives; "same" or "different". For each trial, subjects were presented with a pair of stimuli which differed only in the duration of one of the four moraic segments. The presentation equipment and levels were the same as those in the first experiment. In each pair, the first stimulus was fixed to the standard, the stimulus with no temporal operation, and the second one was the comparison stimulus, the stimulus with temporal modification. The interval between onsets of paired stimuli was 1.3 seconds; each trial started approximately three seconds after the preceding response. Four series of experimental runs were randomly interleaved to prevent prediction of the target segment's position. Each series was terminated when the tenth up-down turning appeared. Finally, the discrimination threshold was calculated for each series as an average of values at the fifth to the tenth up-down turnings. These procedures in the up-down method were automatically conducted by a personal computer (Macintosh II fx). Separate sessions were prepared for lengthening and shortening directions, and also for the natural and the flattened F0 contours.

4. Subject

The same seven subjects that participated in the first experiment also took part in the second experiment.

B. Results and discussion

Table 2 shows the mean and the standard deviation of the measured discrimination thresholds for each segment pooled over the seven subjects. DT_+, DT_-, and DT range mean the discrimination threshold for the lengthening direction, that for the shortening direction, and the range between the discrimination threshold for the lengthening direction and that for the shortening direction (= $DT_+ + DT_-$) in ms. Since a few subjects could not always succeeded in getting termination for all of the measurements, a certain number of vacant values exist; i.e. three vacancies for DT_+ and two for DT_-; hence five for DT range.

Table 2. Mean and standard deviation of the measured temporal discrimination thresholds for each of the four examined segments. The left part shows the text of the words, phoneme type, position in a word, and F0 contour. The capitals in the texts indicate the segments in question. The right part shows the mean and standard deviation of the discrimination thresholds over the seven subjects.

Segment	Position	Phoneme	F0	DT_+ (ms)	DT_- (ms)	DT range (ms)
shInagire	1	/i/	natural	24.5 (4.0)*	27.5 (10.8)	52.0 (8.5)
			flat	39.2 (13.8)	31.3 (6.9)	67.5 (14.3)
shinagIre	3	/i/	natural	37.0 (8.8)	30.5 (15.6)	67.5 (15.6)
			flat	39.6 (13.5)	39.1 (19.4)	78.7 (22.0)
nAmeraka	1	/a/	natural	23.3 (5.9)	19.8 (9.5)	41.7 (12.1)
			flat	27.8 (10.7)	21.7 (13.2)	50.7 (9.8)
namerAka	3	/a/	natural	28.3 (5.6)	35.9 (11.5)	65.0 (15.6)
			flat	37.2 (6.9)	42.4 (12.8)	78.0 (16.3)

DT_+: Mean discrimination threshold for lengthening direction.
DT_-: Mean discrimination threshold for shortening direction.
DT range: Mean discrimination threshold range (= $DT_+ + DT_-$).
*Values in round brackets are standard deviations.

1. Effects of position, phoneme type, and F0 contour

We performed a three-way ANOVA on the DT range to see whether effects consistent with the effects on acceptability would be found; the main factors were position in a word, phoneme type, and F0 contour. Because of a few vacant values, this was not an ANOVA of repeated-measures. The influences of the three main factors can be observed in Figure 5, panels (a) to (c).

The effect of position in a word was significant [$F(1,50) = 21.3$, $p < 3.55 \times 10^{-5}$]. The inclination of this effect was consistent with the effects observed in the first experiment. The acceptable range was narrower for the segments at the first position than those at the third position in the first experiment. The DT range was also narrower for the segments at the first position than those at the third position in the second experiment. Although the effect of phoneme type was rather weak [$F(1,50) = 3.28$, $p < 0.0771$], the inclination was also consistent with the effects in the first experiment; the narrower DT range corresponded to the narrower acceptable range, and vice versa. The factor of F0 contour was significant [$F(1,50) = 8.45$, $p < 0.00575$]. No significant interaction existed among the three factors.

2. Relationship between discrimination threshold and acceptability

We also measured acceptability for the same stimuli employed in the discrimination threshold measurement. Correlation analyses were performed between the results obtained from the two measurements, discrimination thresholds and acceptability (Figure 6, panels (a) and (b)). A significant positive correlation was found between the center shift of the DT range and the axis shift of the acceptability curve which stands for the center shift of the acceptable range [$r = 0.628$, $p < 5.87 \times 10^{-4}$]. Again, a significant positive correlation was found between the DT range and the acceptable range [$r = 0.667$, $p < 2.00 \times 10^{-4}$]. These values for the acceptability range were defined as the horizontal width of the acceptability curve at a vertical value of 1 up from the bottom of each curve; this definition was adopted as one indicative of the acceptable range.

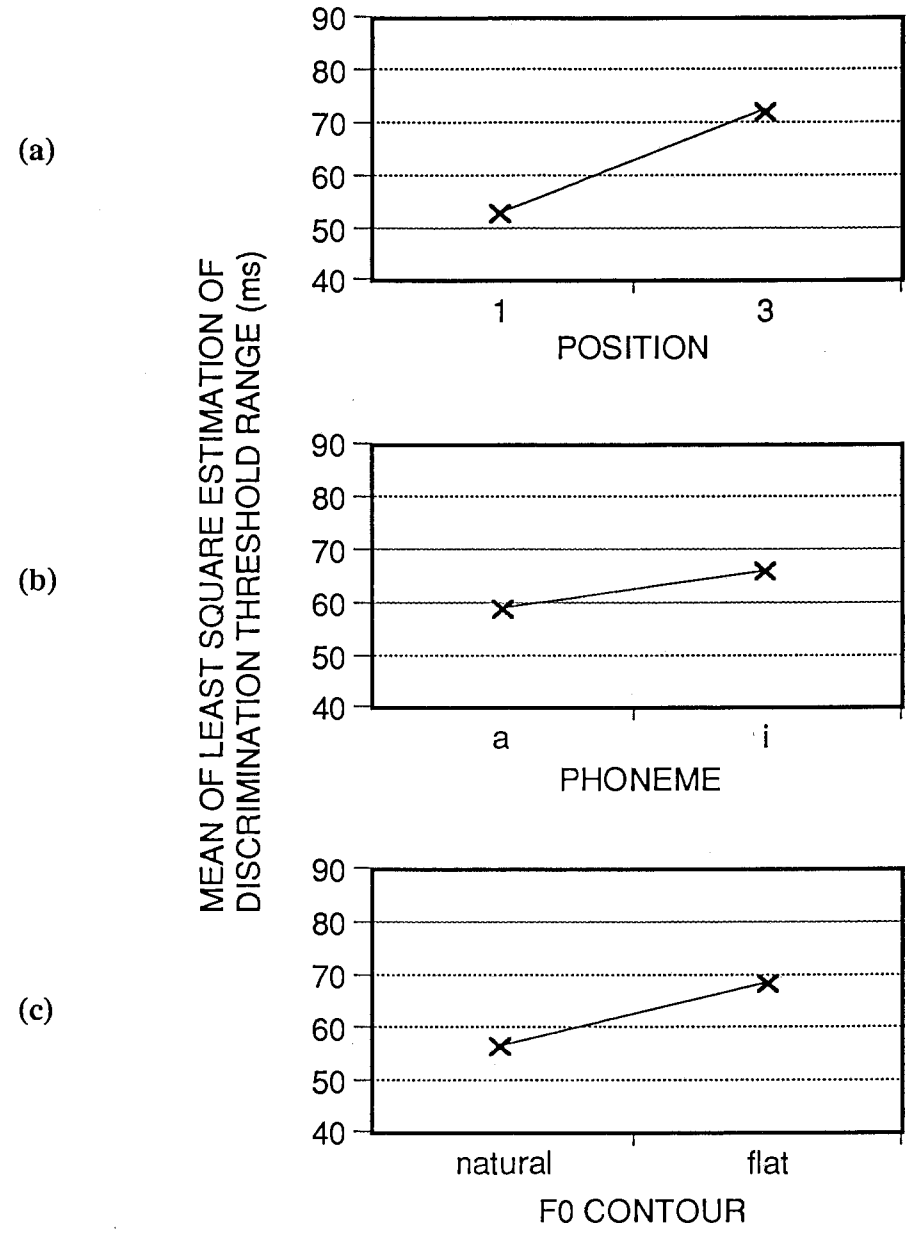
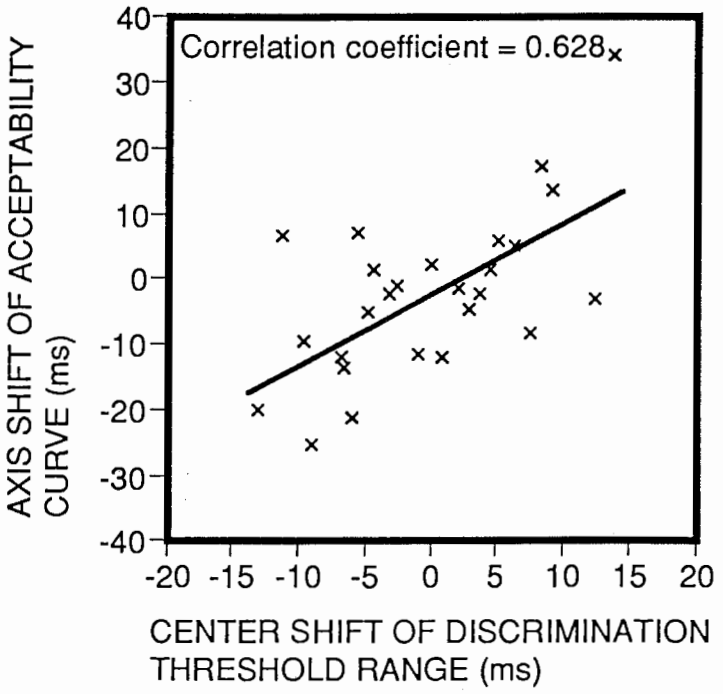


Figure 5. Mean of least square estimations of the discrimination range, the range between two discrimination thresholds, for each level in each factor; they were calculated in the analysis of variance in the second experiment. The following effects are visible. The discrimination range is narrower (a) for the first mora than for the third mora, (b) for the vowel /a/ than for /i/, and (c) for the segments with natural F0 contour than for the segments with flat F0 contour.

(a)



(b)

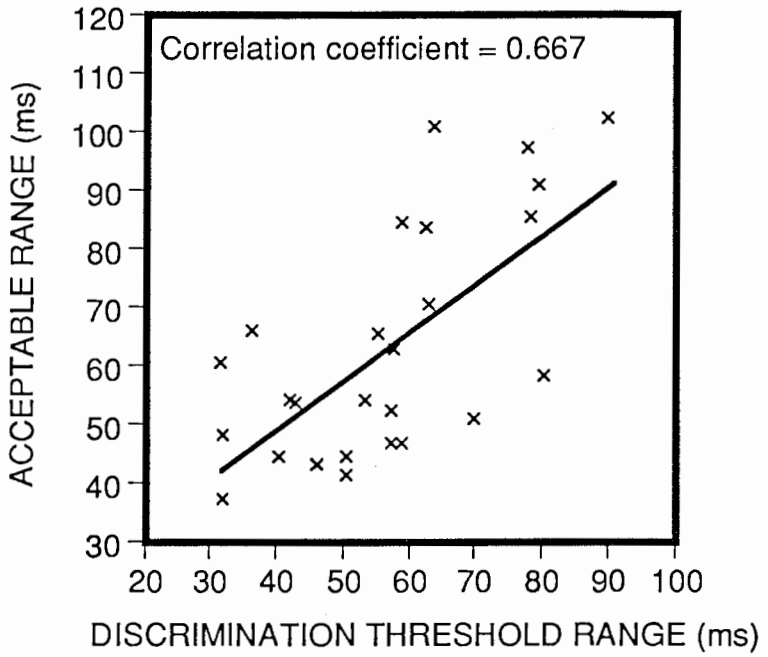


Figure 6. Correlation between acceptability and discrimination threshold. (a) A positive correlation is found between the axis shift of the acceptability curve and the center shift of the discrimination threshold range. (b) A positive correlation is found between the acceptable range and the discrimination threshold range.

These findings suggest that the subjects evaluated acceptability based on perceptual sensitivity. The fact that the position in a word and the phoneme type affected the discrimination thresholds, supports the suggestion that these two factors function at a perceptual stage.

III. GENERAL DISCUSSION

The experimental results showed that the two indices of the acceptability for durational modification, the sharpness and the axis of the acceptability curve, vary according to the segment attributes and context. The consistent influence on the temporal discrimination thresholds was also shown, suggesting that the acceptability tests reflected detectability. The aim of this study was to examine the influence of segment attributes and context on perceptual sensitivity to modification of segmental duration. The index directly linked to perceptual sensitivity might not be the axis of the curve, which corresponds to the preferable duration, but the sharpness of the curve. We therefore would try to mainly explain the influence of segment attributes and context on the sharpness of the acceptability curve, under the headings below. After checking the adequacy of the explanation by original durations, the contents of the two strong effects, the effects of position and phoneme type, would be discussed in reference to previous studies.

A. Original duration

In the current experiments, the original unmodified duration of the segments in question was statistically shorter at the word-initial position than at the word-medial position and also shorter for vowel /a/ than for /i/; these statistical properties agree with the ones found in earlier acoustical measurements for a large amount of Japanese speech data (Sagisaka and Tohkura, 1984). Several studies reported that temporal JND for non-speech sound, such as band noise or tone burst, decreases as base duration is shortened within a certain range (e.g. Abel, 1972a; Small and Campbell, 1962), as predicted from Weber's law. If a similar process conforming to Weber's law functioned dominantly in the current experiments, perceptual sensitivity would have

been higher at the shorter segmental durations. However, this was not always the case in the current study. For vowel color, for instance, higher sensitivity was observed at the longer segments, vowel /a/; for position in a word, higher sensitivity was observed at the shorter segments, the first position. Clearly, this discrepancy cannot be explained by the effect of the original duration. Furthermore, the result of the acceptability test indicated that the sharpness of the acceptability curve had very little correlation to the original duration. Therefore, we would conclude that the explanation of the original duration is not tenable at least under the conditions in the current study.

B. Effect of position

The effect of position on perceptual sensitivity found in this study is consistent with the previous acceptability test reported by Sato (1977). He showed that the acceptable range tended to be narrower at the first position than at other positions using several synthesized words. However, some results of previous studies, in which non-speech stimuli were used, seem to conflict with our findings. Hirsh, et al. (1990) investigated the human's ability to detect temporal deviation occurring at one of six or ten periodic intervals marked by tone bursts; they reported that a listener's detectability for an irregularity was worse in initial intervals than in final intervals when the basic interval was 50 ms. Lehiste (1979) also reported that a listener's ability to detect temporal irregularity was worst in the sequence initial using four successive noise-filled intervals separated by clicks; the basic interval was 300, 400, or 500 ms. By considering this contradiction between the two groups, the "speech stimuli group" and the "non-speech stimuli group", we broke down the characteristics of the positional effect in the following paragraphs.

Although the differences between these two groups apparently came from the type of stimulus, such as speech or non-speech, there was also a practical difference in the type of listener's task. The experimental procedures in the "speech stimuli group" studies commonly required the listeners to compare two temporal patterns as part of each trial. In the current measurements on the discrimination threshold, the listener's task itself was obviously a comparison of two temporal patterns. Although our acceptability test as well as Sato's test presented just one stimulus for each trial, we believe the naturalness evaluation must have asked listeners to compare the presented stimulus with an internal "standard". On the other hand, the "non-speech stimuli

group" studies investigated the listener's ability to detect irregularities within one temporal pattern.

The following statement possibly explains the positional effect that listeners were most sensitive to the modification at the first position in the comparison of two temporal patterns. Since the perceptual duration had to be measured by referring two time markers, the beginning point and the end point, the more precise measurement could be expected when the time markers were represented internally more clearly. Abel's study (1972b) partly supports this hypothesis; he observed the higher temporal discriminability for durations with markers at the higher level. In general, auditory nerves are less activated just before stimulus onset than during the stimulus, because stimulus onset always occurs after a relatively long pause. Therefore, the auditory nerves were more activated by the markers at the sequence initials than at the sequence medials even when they were at a physically identical level (e.g. Delgutte, 1980; Smith, 1979); consequently, internal representations tended to be clearer for the markers at the initial position than those at the other positions. Thus, we presume that because at least one of two markers (in particular, the first one) was always such a clearly represented marker, listeners could achieve the best performance in the measurement of duration at the initial position.

The above statement could be applicable whether the presented temporal patterns were one or two. However, the irregularity detection within one temporal pattern involved a strong restriction that listeners must form an internal regularity in advance of making a judgment; forming regularity is not easy at the sequence initials, but it is easier at the sequence finals, where a sufficient number of markers are available. Since such a restriction, we suppose, overcame the effect of marker clearness in the task of irregularity detection, Lehiste and Hirsh et al. observed the better performance at the medial or final position.

A positional effect similar to one we observed, was reported even for non-speech stimuli when the task was an explicit comparison of two temporal patterns. Tanaka et al. (1992) measured the discrimination threshold for successive intervals separated by clicks which were devised so as to replicate the temporal structures of the word stimuli used in the current discrimination threshold experiment. The methods and the subjects were identical to those in the current measurements. As a result, the discrimination threshold for the first interval was smaller than that for the third one; this indicated the listeners responded more sensitively for the temporal modification

at the first position than for that at the third position.

We would conclude, therefore, that the positional effect we found in this study can be observed regardless of the stimulus type, speech or non-speech, when the task requires the process to compare two temporal patterns.

C. Effect of phoneme type

The listeners responded more sensitively not for vowel /i/, which had a shorter inherent duration, but for vowel /a/, which had a longer inherent duration. Aside from the inherent duration, the loudness might be a candidate that is both a factor that influences the perceptual duration and a contrastive property between /a/ and /i/.

First, loudness probably affects the acuity in perceptual duration measurement; Tyler, et al. (1982) found an increase in the temporal difference limen with stimuli at low levels. The influence of loudness is possibly explained by the clearness of the time markers, which was described in the discussion on the positional effect. Segments with higher loudness usually have a higher level at either end than segments with lower loudness; hence, such an end, which is hardly masked, probably plays a role as a clearly represented time marker. The results of Creelman's study (1962) appear to support this possibility; his subjects achieved better performance in temporal discrimination for tones at higher signal-to-noise ratios.

Next, the power of vowel /a/ is inherently higher than that of /i/ according to previous acoustical measurements (Mimura, et al., 1991). Loudness, which generally correlates with power, is therefore expected to be higher for vowel /a/ than for /i/. The result of the loudness calculation for the stimuli employed in this study showed that the loudness of vowel /a/ was statistically higher than that of /i/; the averages of loudness calculated in accordance with ISO-532B (ISO, 1975; Zwicker, et al., 1991) were 11.6 sone GF for vowel /a/ and 7.3 sone GF for vowel /i/, and the difference between these two averages was confirmed as significant by a t-test [$t(68) = 6.54$, $p < 9.51 \times 10^{-9}$].

Both the positional and phoneme type effects could be commonly understood in terms of the clearness of time markers. However, discussing the perceptual mechanisms governing this possibility is beyond the scope of the current paper. Further research is needed to verify these interpretations.

IV. CONCLUSIONS

Acceptability tests for the durational modification of speech segments were carried out using a large number of stimuli based on 70 different vowel segments, in order to examine the effects of segment attributes and context on perceptual sensitivity. The results of the acceptability tests were consistent with the measured discrimination thresholds which indicate perceptual sensitivity. Hence, it is reasonable to assume that the acceptability data reflected perceptual sensitivity, not completely, but at least in a certain sense. On the basis of this assumption, we made the following conclusions. Perceptual sensitivity to durational modification was not constant regardless of the segment attributes or context but varied as follows: (1) it was higher at the first moraic segments than at the third moraic segments, (2) it was higher at vowel /a/ than at vowel /i/, and (3) it was higher at the high-tone segments than at the low-tone segments; the third difference was less significant than the others. These effects could not be explained in terms of the original as-produced segmental duration. The two strong effects, the effects of position in a word and the phoneme type, could be commonly explained by the following interpretation: the perceptual sensitivity becomes higher at the segments that have more clearly represented time markers.

ACKNOWLEDGMENT

Thanks are due to Dr. W. N. Campbell for his fruitful discussions and suggestions, and to Dr. H. Kawahara for his valuable comments, on earlier versions of this paper.

REFERENCES

Abel, S. M. (**1972a**). "Duration discrimination of noise and tone bursts," J. Acoust. Soc. Am. **51**, 1219-1223.

Abel, S. M. (**1972b**). "Discrimination of temporal gaps," J. Acoust. Soc. Am. **52**, 519-524.

Bochner, J. H., Snell, K. B., and MacKenzie, D. J. (**1988**). "Duration discrimination of speech and tonal complex stimuli by normally hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **84**, 493-500.

Carlson, R. and Granström, B. (**1975**). "Perception of segmental duration," in Cohen, A. and Nooteboom, S. G. (Eds.), *Structure and process in speech perception* (Springer-Verlag, Heidelberg), pp. 90-106.

Creelman, C. D. (**1962**). "Human discrimination of auditory duration," J. Acoust. Soc. Am. **34**, 582-593.

Delgutte, B. (**1980**). "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers," J. Acoust. Soc. Am. **68**, 843-857.

Ebata, M., Sone, T., and Nimura, T. (**1974**). "Temporal summation characteristics in auditory system" (in Japanese with English abstract and English figure captions), J. Acoust. Soc. Jpn. **30**, 662-669.

Fujisaki, H., Nakamura, K., and Imoto, T. (**1975**). "Auditory perception of duration of speech and non-speech stimuli," in Fant, G. and Tatham (Eds.), *Auditory Analysis and Perception of Speech* (Academic Press, London), pp. 197-219.

Hirsh, I., Monahan, C., Grant, K., and Singh, P. (**1990**). "Studies in Auditory Timing: 1. Simple Patterns," Percept. Psychophys. **47**, 215-226.

Huggins, A. W. F. (**1972**). "Just noticeable differences for segment duration in natural speech," J. Acoust. Soc. Am. **51**, 1270-1278.

Imai, S. and Kitamura, T. (**1978**). "Speech analysis synthesis system using the log magnitude approximation filter" (in Japanese with English figure captions), Trans. Inst. Electron. Commun. Eng. Jpn. **J61-A**, 527-534.

International Organization for Standardization. (**1975**). "Acoustics - Method for calculating loudness level," in ISO 532-1975(E), International Organization for Standardization.

Kaiki, N., Takeda, K., and Sagisaka, Y. (**1992**). "Linguistic properties in the control of segmental duration for speech synthesis," in Bailly, G., Benoit, C., and Sawallis, T. R. (Eds.), *Talking Machines: Theories, Models, and Designs* (Elsevier Science Publishers B.V.), pp. 255-263.

Klatt, D. H. (**1976**). "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence," J. Acoust. Soc. Am. **59**, 1208-1221.

Klatt, D. H. and Cooper, W. E. (**1975**). "Perception of segment duration in sentence contexts," in Cohen, A. and Nooteboom, S. G. (Eds.), *Structure and Process in Speech Perception* (Springer-Verlag, Heidelberg), pp. 69-89.

Lehiste, I. (**1979**). "The perception of duration within sequences of four intervals," J. Phonet. **7**, 313-316.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467-477.

Mimura, K., Kaiki, N., and Sagisaka, Y. (**1991**). "Analysis and control of temporal patterns of speech power using statistical methods" (in Japanese with English abstract and English figure captions), Tech. Rep. SP91-4, Acoust. Soc. Jpn.

Sagisaka, Y., and Tohkura, Y. (**1984**). "Phoneme duration control for speech synthesis by rule" (in Japanese with English figure captions), Trans. Inst. Electron. Commun. Eng. Jpn. **J67-A**, 629-636.

Sagisaka, Y., Takeda, K., Abe, M., Katagiri, S., Umeda, T., and Kuwabara, H. (**1990**). "A large-scale Japanese speech database," in Proc. International Conference on Spoken Language Processing, pp. 1089-1092.

Sato, H. (**1977**). "Segmental duration and timing location in speech," (in Japanese with English abstract), Tech. Rep. **S77-31**, Acoust. Soc. Jpn.

Small, A. M. and Campbell, R. A. (**1962**). "Temporal differential sensitivity for auditory stimuli," Am. J. Psychol. **75**, 401-410.

Smith, R. L. (**1979**). "Adaptation, saturation, and physiological masking in single auditory-nerve fibers," J. Acoust. Soc. Am. **65**, 166-178.

Takeda, K., Sagisaka, Y., and Kuwabara, H. (**1989**). "On sentence-level factors governing segmental duration in Japanese," J. Acoust. Soc. Am. **89**, 2081-2087.

Tanaka, M., Tsuzaki, M., and Kato, H. (**1992**). "On the perception of the click sequence simulating moraic structure" (in Japanese with English figure captions), Tech. Rep. H-92-63, Acoust. Soc. Jpn.

Tyler, R. S., Summerfield, Q., Wood, E. J., and Fernandes, M. A. (**1982**). "Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners," J. Acoust. Soc. Am. **72**, 740-752.

Zwicker, E., Fastl, H., Widmann, U., Kurakata, K., Kuwano, S., and Namba, S. (**1991**). "Program for calculating loudness according to DIN 45631 (ISO 532B)," J. Acoust. Soc. Jpn. (E) **12**, 39-42.