EFFECTS OF PHONETIC QUALITY AND DURATION ON PERCEPTUAL ACCEPTABILITY OF TEMPORAL CHANGES IN SPEECH

Hiroaki Kato[1,3] *Minoru Tsuzaki*[1] *Yoshinori Sagisaka*[2,3]

[1]ATR Human Information Processing Research Laboratories
[2]ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seikacho, Kyoto 619-0288, Japan
[3]Kobe University, 1-1 Rokkodaicho, Nada, Kobe 657-8501, Japan
E-mail: kato@hip.atr.co.jp, tsuzaki@hip.atr.co.jp, sagisaka@itl.atr.co.jp

ABSTRACT

To establish a perceptually valid rule for the durational control of synthetic speech, it is necessary to know the degree to which a given temporal error or distortion is acceptable to human listeners. Two perceptual experiments were conducted to estimate the acceptability of modifications in either vocalic or consonantal durations as a function of two attributes of the modified portions, i.e., the phonetic quality and the original (unmodified) duration. The results showed that the listeners' acceptable modification ranges were narrowest for vowels, and widest for voiceless fricatives and silent closures, with nasals in between. They were also narrower for those portions with shorter base durations. The effect of the original duration was larger for the vowel stimuli than for the voiceless fricative stimuli. The perceptual mechanism mediating these results is discussed with regard to the dependency of the listeners' temporal sensitivity on the stimulus loudness and base duration. [Re: http://www.hip.atr.co.jp/~kato/single_duration/]

1. INTRODUCTION

Rules to assign segmental durations have been proposed for speech synthesis [2, 6, 10]. The effectiveness of a durational rule should be evaluated in terms of human perception. With almost all previous rules, however, the average of the absolute difference, or error, of each segmental duration from its reference has been adopted as the measure of any objective evaluation. A potential problem with this measure is that it gives every error the same weighting regardless of the attributes of the segment in question which possibly affect human sensitivity to durational changes. Toward more perceptually valid evaluation measures, we have examined the influence of various segment attributes and contexts on durational sensitivity. Among them, the current paper concentrates on the following two attributes: (1) phonetic quality and (2) original duration. [1]

Typical examples of the dependency of temporal sensitivity on phonetic quality seem to be found when comparing vowel and consonant segments. Both Huggins [4] and Carlson and Granström [3] reported that the just noticeable differences (jnd's) for segmental durations are smaller for vowels than for consonants. However, there seems to be no report of this sort for the acceptability, which can be considered as a more direct measure than a jnd in the evaluation of durational rules. The current study, therefore, focused on the acceptability measure and addressed the difference among phonetic quality types within consonants as well as that between vowels and consonants.

Although our intuition may predict that the acceptable modification range would be wider for a longer original or base duration, a previous study [8] did not find any influence of the original duration. This previous study, however, only used word stimuli having a homogeneous temporal structure (CVCVCVCV), and therefore, the temporal variations of the tested portions were limited. Considering this, the current study employed a stimulus set having more diverse temporal variations and reexamined the relation between the original duration and the acceptable modification range.

[1]An unmodified duration of a portion whose duration is subject to temporal modification is referred to as an "original duration."

2. EXPERIMENT 1: EFFECT OF PHONETIC QUALITY

Experiment 1 aimed to test the dependency of acceptability evaluations for durational modifications on the difference in phonetic quality.

2.1. Methods

Stimuli. Forty-nine four-mora Japanese word utterances of a male speaker were selected from the ATR speech database [9] as the original material. One of the acoustically continuous portions in each utterance was shortened or lengthened over a range from -75 ms to $+75$ ms from the original duration in 7.5-ms steps using a log magnitude approximation (LMA) analysis-synthesis technique [5], resulting in 20 different modification steps. The modified portions were chosen from the second moraic position in the words and had one of the following four phonetic quality types: (1) vowel, (2) nasal, (3) voiceless fricative, or (4) silence. To prevent the phonemic quality of the test portions from suffering for the temporal modifications, the consonantal durations were chosen from phonemically longer segments, i.e., moraic nasals, devoiced vowels (having a voiceless fricative quality), and geminate obstruents. Table 1 summarizes the profiles of the test portions. In total, 1029 word stimuli were synthesized: i.e., (20 modifications + 1 unmodified) $\times$ 49 portions.

Subjects. Six adults with normal hearing participated in experiment 1. All of them were native speakers of Japanese.

Procedures. The subjects listened to each of the word stimuli and were asked to rate the acceptability of modification using seven subjective categories; i.e., "quite acceptable" to "unacceptable." Each subject rated each stimulus four times in total.

2.2. Results and Discussion

Measure of acceptability. The measure of acceptability was the same as that used in a previous study [8] to maintain consistency among the studies. First, the subjects' evaluation scores were plotted against the change in duration of the test portion, and then

Table 1: The number of test portions for each of the stimulus groups in experiment 1, and the averages and standard deviations of their acoustic durations.

	Stimulus group					
	Short vowel	Moraic nasal	Devoiced vowel portion	Geminate stop	Geminate fricative	Total
Number of samples	10	14	11	7	7	49
Phonetic quality type	vowel	nasal	voiceless fricative	silence	voiceless fricative	
Average duration (ms)	115.5	121.6	113.0	192.9	224.3	
S.D. of durations (ms)	11.6	30.8	15.8	18.6	29.2	

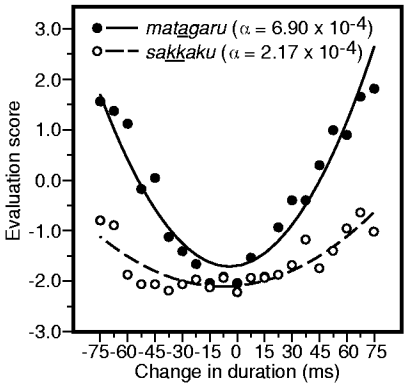

Figure 1: An example illustrating a difference in acceptability-change between two different speech portions. The portions subjected to durational modification are marked with underlines in the legend. The scatter plots show that the evaluation score varies according to the durational change more drastically for the second vowel of the word "*matagaru* (to ride)" (filled circles), than for the silent closure of the geminate stop consonant in the word "*sakkaku* (illusion)" (open circles). The two parabolic regression curves trace this tendency.

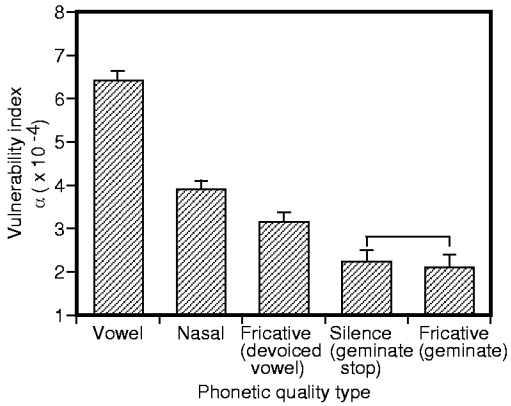

Figure 2: The least squares means of the vulnerability index (α), i.e., the second-order polynomial coefficient of the fitting curve, for each phonetic quality type; they were calculated in the ANOVA procedure. The error bars show the standard errors. A larger α implies a narrower acceptable range. The difference between the two bridged bars is not statistically significant.

a parabolic regression as generally formulated below was applied for each token and for each subject,

$$Evaluation\ score = \alpha(\Delta T - \beta)^2 + \gamma, \qquad (1)$$

where ΔT denotes the change in duration; the unit of ΔT is not the relative duration but milliseconds. The coefficient of the second-order term (α) was, then, taken as the "vulnerability index," the objective variable of the current study. As derived from Eq. 1, the vulnerability index serves the width between the longer and shorter limits of the temporal modification that yields a certain level of acceptability, i.e., an acceptable range. Figure 1 shows examples of individual fittings. In total, 287 vulnerability indices (or α scores) were obtained, i.e., (49 tokens $\times$ 6 subjects) $-$ 7 unreliable data.

Effect tests. A one-way ANOVA of repeated measures with *subject* as the blocking factor showed the effect of *phonetic quality type* on the vulnerability index (α) as significant [$F(4, 20) = 53.1, p < 0.001$]. As shown in Fig. 2, α was greatest for the vowels, next for the nasals, third for the fricatives, and smallest for the silent portions and fricatives in geminate consonants. Multiple comparisons using Tukey–Kramer's HSD (the honestly significant difference) indicated the difference between any two average αs to be significant [$p < 0.01$], except for the difference between those of the geminate fricative and silence groups.

Discussion. The listeners, in general, evaluated the temporal modifications of vowel portions as less acceptable than those of consonant portions. This tendency is in agreement with that predicted from literature reporting that vowel durations are more accurately discriminated than consonant durations are [1, 3, 4].

Some results, however, could not be accounted for by the factor of *phonetic quality type*. There was a significant difference between the αs of the devoiced vowel portions and the geminate fricatives, although they have the same phonetic quality type, i.e., the voiceless fricative type. This difference in the αs is likely due to their difference in the original duration because their durational difference is, as seen in Table 1, notable. Experiment 2 was, therefore, designed to test the effect of the original duration separately from that of the phonetic quality type.

3. EXPERIMENT 2: EFFECT OF ORIGINAL DURATION

3.1. Methods

Design. A two-way factorial design was applied. The first factor, *phonetic quality type* of the test portion, had two levels: vowel and voiceless fricative. The second factor, *original duration* of the test portion, also had two levels: short and long. The test portions for the short and long levels in the vowel type were chosen from phonemically short and long vowels, and those for the short and long levels in the voiceless fricative type were chosen

Table 2: The number of test portions for each of the stimulus groups in experiment 2, and the averages and standard deviations of their acoustic durations.

	Stimulus group				Total
	Short vowel	Long vowel	Short fricative (Devoiced vowel portion)	Long fricative (Geminate fricative)	
Number of samples	10	10	5	5	30
Phonetic quality type	vowel	vowel	voiceless fricative	voiceless fricative	
Duration category	short	long	short	long	
Average duration (ms)	115.5	251.8	125.0	219.0	
S.D. of durations (ms)	11.0	25.7	17.7	17.1	

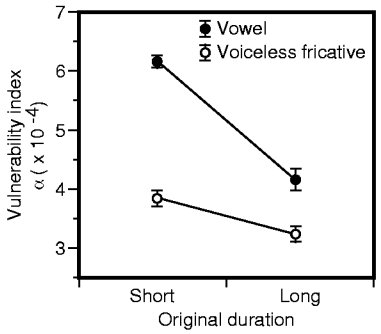

Figure 3: The least squares means of the vulnerability index (α), i.e., the second-order polynomial coefficient of the fitting curve, for each stimulus group; they were calculated in the ANOVA procedure. The error bars show the standard errors. A larger α implies a narrower acceptable range.

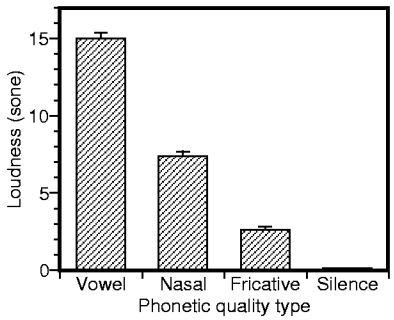

Figure 4: The average loudness of the speech portions whose durations were subjected to temporal modification in experiment 1, as a function of *phonetic quality type*. For the *silence* type, the background noise level was adopted. The error bars show the standard errors.

from devoiced vowel portions and geminate obstruents.

Stimuli and procedures. Thirty four-mora Japanese words were selected as the original materials from the same database as in experiment 1. Table 2 summarizes the profiles of the test portions. The speaker of the original materials, the manipulation method, and the procedure for the experimental run were the same as those in experiment 1.

Subjects. Nine adults with normal hearing participated in experiment 2. All of them were native speakers of Japanese. None of them participated in experiment 1.

3.2. Results

Effect tests. In accordance with the same procedures as in experiment 1, the vulnerability index (α score) was computed for each of the 30 test portions and each of the nine subjects, resulting in 270 αs. A two-way factorial ANOVA of repeated measures was performed with *phonetic quality type* and *original duration* as the main factors, and with *subject* as the blocking factor. The main effects of *phonetic quality type* and *original duration* were significant [$F(1,8) = 51.9, p < 0.0001$; $F(1,8) = 67.0, p < 0.0001$, respectively]. As shown in Fig. 3, α was greater for the vowels than for the voiceless fricatives, and similarly greater for the short portions than for the long portions. There was a significant interaction between both main factors [$F(1,8) = 14.7, p < 0.005$]; the effect of *original duration* was larger for the vowels than for the voiceless fricatives. Multiple comparisons among the average αs of four stimulus groups using Tukey–Kramer's HSD indicated the difference between any two average αs to be significant [$p < 0.05$], except for the difference between the αs of the long vowel and short voiceless fricative (devoiced vowel) portions.

To summarize, a similar effect of *phonetic quality type* to that in experiment 1 was replicated. The effect of *original duration* for the voiceless fricatives was also replicated. Additionally, *original duration* was found to affect the temporal vulnerability of vowel portions.

4. GENERAL DISCUSSION

This section tries to relate the acceptability measure to measures of human sensitivity against changes in non-speech durations. Such an auditory-based approach has a potential advantage in providing perceptually valid notions that can be generalized across languages.

4.1. Effect of Phonetic Quality

We chose loudness as the candidate variable representing differences in the phonetic quality type from among many psychoacoustical features of the tested speech portions.[2] A previous study had shown that the acceptability of modifications in a vowel duration correlates with the loudness inherent in each vowel quality (/i/ or /a/) [8]. Figure 4 shows the average loudness of the test portions pooled for each phonetic quality type. Interestingly, the order of the phonetic quality types by these loudness values is identical with that by the vulnerability indices except for the relation between the voiceless fricative and silence types (c.f., Fig. 2). These similar characteristics suggest that these loudness values are likely to also correlate with the vulnerability index (α). The Pearson product-moment correlation coefficient (r) between the average loudness and the α score based on the 49 test portions

[2]Any usage of the word "loudness" in the current study means the loudness calculated by ISO-532 method B, unless otherwise stated.

was 0.889. This accountability of the loudness for the vulnerability index was comparable with that of the phonetic quality type where r was 0.888. The psychoacoustical validity of this correlation can be found in a previous study [7] which reported a clear correlation between temporal sensitivity to non-speech auditory durations and their intensity.

4.2. Effect of Original Duration and Its Interaction with Phonetic Quality

A larger vulnerability index, i.e., a narrower acceptable modification range, was observed for the shorter test portions. This tendency seems to be reasonable in the light of a general psychophysical law, i.e., Weber's Law. Conforming to this law, a longer physical change is necessary for a longer base duration to yield the same amount of perceived change. Note, however, that an acceptable range is not exactly proportional to the corresponding original duration. The ratio of two acceptable ranges was considerably smaller than that of the corresponding original durations.

To account for the observed interaction, on the other hand, the properties within the tested portion itself, including its phonetic quality and original duration, appear to be insufficient. The temporal structure that exceeds the tested duration should additionally be taken into account. From among such global temporal properties, we chose the vowel-onset asynchrony (VOA) immediately surrounding the test portion; literature has suggested that VOA is especially useful for measuring some temporal structures, e.g., inter-syllable timing or speaking-rate. As seen in Fig. 5, whereas a clear contrast in the VOA between the 'short' and 'long' groups is observed for the vowel type, no such contrast seems to be observed for the voiceless fricative type. Acoustic measurements of the actual stimulus words confirmed that the same tendency was found in the material of experiment 2. Therefore, the observed interaction can be accounted for if we consider the difference in the VOA contrast as the source enlarging the effect of the original duration for the vowel type compared to the voiceless fricative type.

An alternative global source that potentially accounts for the observed interaction is the degree of deviation from a regular rhythm formed by the short C and V alternation. We, however, cannot provide a fuller discussion on this matter in this brief paper.

5. CONCLUSIONS

The modification range for which a certain decrement of acceptability would be expected, i.e., the acceptable range, expanded as the phonetic quality type changed from vowel, nasal, then voiceless fricative or silence. The observed acceptability variations with the phonetic quality type correlated with the variation in loudness of the portion in question; the acceptable range narrowed as the portion became louder. The acceptable range also expanded as the original, as produced, duration increased. The effect of the original duration was larger for the vowel type than for the voiceless fricative type. This dependency could be accounted for by another source of the temporal structure, i.e., the vowel-onset asynchrony (VOA).

An important implication of the current research is that an expanding acceptable range observed with changes in the phonetic quality or original duration can be mostly accounted for by psychoacoustical terms, i.e., a reduced capability to discriminate temporal modifications as the loudness decreases or the base duration increases. The current results demonstrate that we can expect a more valid (closer to human evaluation) measure than the traditional simple average of acoustic errors in evaluating durational rules by accounting for the loudness and original duration as weighting factors.

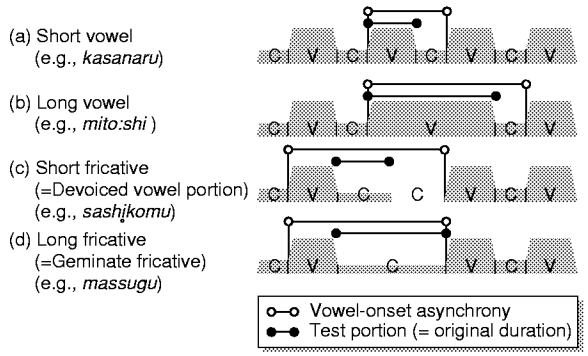

Figure 5: Schematic examples showing the temporal structures of four-mora Japanese words for each stimulus group in experiment 2. The horizontal and vertical axes roughly refer to the time and loudness, respectively. 'C' and 'V' represent consonant and vowel portions, respectively. Note that the temporal alignment of each segment is highly idealized in these examples and that such rigid isochronous relations are rarely observed in actual Japanese speech.

6. REFERENCES

1. J. H. Bochner, K. B. Snell, and D. J. MacKenzie. Duration discrimination of speech and tonal complex stimuli by normally hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 84:493–500, 1988.
2. W. N. Campbell. *Multi-level timing in speech.* doctoral dissertation, University of Sussex, Brighton, UK, 1992.
3. R. Carlson and B. Granström. Perception of segmental duration. In A. Cohen and S. Nooteboom, editors, *Structure and Process in Speech Perception*, pages 90–106. Springer-Verlag, Berlin, 1975.
4. A. W. F. Huggins. Just noticeable differences for segment duration in natural speech. *J. Acoust. Soc. Am.*, 51(4):1270–1278, 1972.
5. S. Imai and T. Kitamura. Speech analysis synthesis system using the log magnitude approximation filter. *Trans. Inst. Electron. Commun. Eng. Jpn.*, J61-A(6):527–534, 1978. (in Japanese with English figure captions)
6. N. Kaiki and Y. Sagisaka. The control of segmental duration in speech synthesis using statistical methods. In Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, editors, *Speech Perception, Production and Linguistic Structure*, pages 391–402. IOS, Amsterdam, 1992.
7. H. Kato and M. Tsuzaki. Intensity effect on discrimination of auditory duration flanked by preceding and succeeding tones. *J. Acoust. Soc. Jpn. (E)*, 15(5):349–351, 1994.
8. H. Kato, M. Tsuzaki, and Y. Sagisaka. Acceptability for temporal modification of single vowel segments in isolated words. *J. Acoust. Soc. Am.*, 104(1):540–549, 1998.
9. A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano. ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech Commun.*, 9(4):357–363, 1990.
10. J. P. H. van Santen. Assignment of segmental duration in text-to-speech synthesis. *Comput. Speech Lang.*, 8:95–128, 1994.