Perceptual analysis and evaluation of timing control for speech generation

February, 2005

Makiko Muto

Global Information and Telecommunication Studies, Waseda University
Acknowledgement

This dissertation could be accomplished due to supports received from several people. First, I would like to express my sincere gratitude to Professor Sagisaka of Waseda University, my advisor, for guiding me through the PhD course, for giving me the valuable opportunities to study at ATR and to collaborate with various researchers.

I also would like to express my sincere gratitude to Dr. Hiroaki Kato of the senior researcher of ATR Human Information Science Laboratories (HIS) and Professor Minoru Tsuzaki of Kyoto City University of Arts, my supervisors, for considerately supporting and helping my work, and for teaching me many important things. Without their supports, I could not have completed this work. Also, I would like to thank to Dr. Kato and his family for supporting my stay during trainings at ATR.

I thank to Professor Yoshio Yamasaki, Professor Masaaki Honda, Professor Tetsunori Kobayashi, and Associate Professor Takashi Kawai of Waseda University, my supervisors, for giving valuable advises to write this dissertation.

I thank to Dr. Reiko Akahane-Yamada, the current head of ATR/HIS Multilingual Learning Department for giving opportunity to study and wonderful research environment at ATR. Also, I thank to all members of ATR/HIS Multilingual Learning Department, for their supports and warm encouragements, especially to Takahiro Adachi, Ryo Komaki, Reiko Kubo, Tomoko Takada, Naoko Otsuki, Yoko Ikuma, Yuri Nakamoto, Mami Nishida, Kyoko Fujita and Minako Matsumoto. The experiments in Chapter 2 and 3 could not be accomplished without their supports.

I thank to Professor Hideki Kawahara of Wakayama University for giving valuable advises to use STRAIGHT speech analysis-synthesis method.

I thank to Professor Katsuhiko Shirai for giving opportunity to collaborate with members of Shirai Laboratory. I am indebted to Yasuo Suzuki, Takuro Naito, Daiju Maeki, Aki Kondo and Hiroko Tsuchida, members of CAI group in Shirai Laboratory. The analysis in Chapter 5 could not be accomplished without their technical supports.
I thank to Professor Nobuyoshi Terashima of Waseda University for his support as advisor of graduate course.

I owe a great deal to former and current members of Sagisaka Laboratory. In particular, Takashi Sato, Yoshinori Imai, Yoshihiko Ogawa, and Takumi Yamashita participated and supported the experiments. Also, Yoko Kokenawa, Hiromasa Ogawa, Nagisa Shibuya, and Tatsuhiko Tomita supported my research life at Sagisaka Laboratory.

Finally, I am extremely grateful to my family and friends for their unconditional supports.
Contents

1 Introduction 1

1.1 Need for perceptual studies and objective evaluation on speech timing characteristics 1

1.2 Approach and outline of the dissertation 2

2 Perceptual acceptability analysis of sentence speech I: Effect of intra-phrasal position 5

2.1 Introduction 5

2.2 Experiment 1: Intra-phrase positional effect 7

2.2.1 Experimental conditions 8

2.2.2 Results 9

2.2.3 Discussion 12

2.3 Experiment 2: Intra-phrase positional effect in various contexts 14

2.3.1 Experimental conditions 14

2.3.2 Results and discussion 15

2.4 Experiment 3: Robustness of intra-phrase positional effect in a breath group 18

2.4.1 Experimental conditions 18

2.4.2 Results and discussion 19

2.5 General discussion 20

2.6 Conclusions 21

3 Perceptual acceptability analysis of sentence speech II: Effect of speaking rate 23

3.1 Introduction 23

3.1.1 Effect of original duration and tempo 24

3.1.2 Factors tested in the study 26
### CONTENTS

3.2 Experiment 1: Effects in speech stimuli .............................. 28
   3.2.1 Experimental conditions ........................................... 28
   3.2.2 Results and discussion ............................................. 30
3.3 Experiment 2: Effects in non-speech stimuli ....................... 36
   3.3.1 Experimental conditions ........................................... 36
   3.3.2 Results and discussion ............................................. 37
3.4 General discussion ...................................................... 40
3.5 Conclusion ............................................................. 42

4 Relationship between speech generation and perception in speech timing control 43
   4.1 Introduction .......................................................... 43
   4.2 Measurement of utterance precision ............................... 44
      4.2.1 Speech data ........................................................ 44
      4.2.2 Elimination of control factors’ influence .................... 45
   4.3 Relationship between utterance and perceptual characteristics of segment duration .................................................. 46
   4.4 Conclusion ............................................................. 46

5 Timing naturalness evaluation using objective measures for non-native speech 49
   5.1 Introduction .......................................................... 49
   5.2 Timing control differences and correlation with naturalness evaluation ................................. 51
      5.2.1 Timing control difference between English and Japanese ........ 51
      5.2.2 Beginner’s timing control characteristics ....................... 52
      5.2.3 Duration differences for predicting native’s subjective evaluation .... 52
   5.3 Analysis on correlations between native’s subjective evaluations and duration differences .................. 54
      5.3.1 Speech data ........................................................ 54
      5.3.2 Correlations between duration differences and subjective scores ..................... 56
   5.4 Temporal naturalness prediction of non-native speech using objective parameters .......................... 59
      5.4.1 Modeling using objective parameters ............................ 59
      5.4.2 Modeling using perceptual weighted parameters ................ 60
   5.5 Conclusions ............................................................. 62
6 Conclusions

6.1 Summary of the dissertation .................................................. 67
6.2 Future work ........................................................................ 69

A Properties and correlation coefficients of tested speech portion 71

B Sentence data used in Chapter 5 ............................................. 81

Bibliography .............................................................................. 83

List of Publication ..................................................................... 89
List of Figures

1.1 Outline of the dissertation. ........................................... 3

2.1 Difference in perceptual sensitivities at intra-phrase positions. .... 9
2.2 Vulnerability indices at intra-phrase positions in four-mora phrases. 11
2.3 Vulnerability indices at intra-phrase positions in five-mora phrases. 11
2.4 Time-loudness counter of four-mora phrase /dekirebal/ (dotted line) superimposed with mode value for each segment (solid line). ......... 14
2.5 Vulnerability indices at intra-phrase positions in three-mora phrases. 16
2.6 Vulnerability indices as a function of vowel position in a phrase and phrase position in a breath group. ............................... 20

3.1 An example illustrating differences in acceptability rating due to changes in vowel duration. ........................................... 31
3.2 Difference in acceptability degradation (vulnerability index) due to speaking rate in Experiment 1. ................................. 32
3.3 Schematic example showing the temporal structures of nonspeech stimuli in three different tempi in Experiment 2. ..................... 37
3.4 Difference in detectability of change in duration due to tempo in Experiment 2. ................................................................. 38
3.5 Difference in detectability of change in duration due to stimulus duration and presentation tempo in Experiment 2. ............. 40

4.1 Vowel control precision represented by the standard deviation of the residual errors as a function of the temporal position in the following utterance units: (a) Word, (b) Phrase, (c) Accent phrase, (d) Breath group, and (e) Sentence. .................................................. 47
5.1 Root mean square and correlation coefficient between subjective rating scores and predicted rating scores calculated by linear multiple regression using perceptual weighting parameters by the loudness. .......................... 63

5.2 Root mean square and correlation coefficient between subjective rating scores and predicted rating scores calculated by tree regression using perceptual weighting parameters by the loudness. .......................... 64

5.3 Root mean square and correlation coefficient between subjective rating scores and predicted rating scores calculated by linear multiple regression using perceptual weighting parameters by the loudness and the position. 65

5.4 Root mean square and correlation coefficient between subjective rating scores and predicted rating scores calculated by tree regression using perceptual weighting parameters by the loudness and the position. 66

A.1 Time-loudness counter of four-mora phrases (dotted line) superimposed with mode value for each segment (solid line). .......................... 75

A.2 Time-loudness counter of five-mora phrases (dotted line) superimposed with mode value for each segment (solid line). .......................... 76

A.3 Time-loudness counter of three-mora phrase /harugal/ (dotted line) superimposed with mode value for each segment (solid line). .......................... 77

A.4 Time-loudness counter of three-mora phrase /kuruto/ (dotted line) superimposed with mode value for each segment (solid line). .......................... 78

A.5 Time-loudness counter of three-mora phrase /nodyomel/ (dotted line) superimposed with mode value for each segment (solid line). .......................... 79
List of Tables

2.1 Stimuli of Experiment 2. .................................................. 17
2.2 Target phrases used in Experiment 3. .................................. 19
2.3 Carrier sentences used in Experiment 3. ............................. 19

3.1 Target phrases used in Experiment 1. ................................. 29
3.2 Average durations of speech material parts used in Experiment 1 (ms). . 29
3.3 The numbers of vulnerability indices that are larger than zero (left) and the numbers of all obtained indices (right) in Experiment 1 (7 listeners \times 10 vowels for each carrier condition and position in a phrase). ........ 33
3.4 The number of vulnerability indices (larger than zero/total) and Pearson’s product-moment correlation coefficients between vulnerability index and original duration for each listener in Experiment 1. ............... 35
3.5 Pearson’s product-moment correlation coefficients between vulnerability index (larger than zero) and original duration for each carrier condition and position in a phrase in Experiment 1. ...................... 35
3.6 Pearson’s product-moment correlation coefficients between detectability index and original duration for each amount of duration change in Experiment 2. .......................................................... 39

4.1 Multiple correlation coefficient and decrease of standard deviation (ms). . 45

5.1 Examples of sentences used the in the correlation analyses. ............... 55
5.2 Average and standard deviation of duration difference (ms), and correlation between subjective evaluation score and duration differences reflecting language differences (a-1)–(a-3) and beginner’s control (b-1)–(b–3). . 57
5.3 Root mean square between subjective and predicted rating score. ........ 59
5.4 Correlation coefficient between subjective and predicted rating score. .. 59
A.1 Durations of four-mora phrases and vowels of Experiment 1 (ms) . 72
A.2 Durations of five-mora phrases and vowels of Experiment 1 (ms) . 72
A.3 Durations of three-mora phrases and vowels of Experiment 1 (ms) . 72
A.4 Standard deviations of vowel durations in four-mora phrases of Experiment 1 (ms) . 73
A.5 Standard deviations of vowel durations in five-mora phrases of Experiment 1 (ms) . 73
A.6 Standard deviations of vowel durations in three-mora phrases of Experiment 2 (ms) . 73
A.7 Standard deviations of vowel durations in three-mora phrases of Experiment 3 (ms) . 73
A.8 Loudnesses (left) and loudness jumps (right) of vowels in four-mora phrases of Experiment 1 (sone) . 74
A.9 Loudnesses (left) and loudness jumps (right) of vowels in five-mora phrases of Experiment 1 (sone) . 74
A.10 Loudnesses (left) and loudness jumps (right) of vowels in three-mora phrases of Experiment 2 (sone) . 74
A.11 Correlation coefficients between vulnerability indices and the original durations, loudnesses or loudness jumps for each of the seven listeners, in terms of either Pearson’s product-moment correlation or Spearman’s rank correlation analysis in four- and five-mora phrases . 80
A.12 Correlation coefficients between vulnerability indices and the original durations, loudnesses or loudness jumps for each of the seven listeners, in terms of either Pearson’s product-moment correlation or Spearman’s rank correlation analysis in three-mora phrases . 80
Chapter 1

Introduction

1.1 Need for perceptual studies and objective evaluation on speech timing characteristics

Spoken language information processing systems, such as speech synthesis or speech recognition, have been used in daily various situations. Recently, speech synthesis and speech recognition are achieving high performances using large corpus and statistics method. However, they are still not comparable to human abilities. The more performance goes up, the finer speech characteristics are needed for further improvements. In particular, prosodic characteristics have become more important topics both for speech synthesis and recognition.

The importance of prosodic characteristics has also been increasing in application fields. The performance improvements in spoken language processing technologies have accelerated their use in many applications. For example, in CALL (Compute Assisted Language Learning) systems, speech recognition technologies have been efficiently used to measure speaker’s proficiency in pronunciation. However, most of applications are limited to the technologies related to segmental features. For further improvements in many application systems, in particular, CALL systems for the second language learning, studies on prosody are in urgent need.

To cope with these urgent needs, we need to study prosodic characteristics profoundly relating to acoustically observable features. In the field of speech communication science and technology, considerable amount of studies relating to prosody have been presented in journals such as Speech Communication, Speech language and Speech, in International
conferences such as ICSLP (International Conference on Spoken Language Processing) and Eurospeech (European Conference on Speech Communication and Technology), and workshops. The increase of prosody-related papers shows the needs of studies both in science and in technology.

In prosody studies, most of research efforts have been devoted to the acoustic analysis. The characteristics of fundamental frequency and segmental duration have been studies in speech synthesis field. Their control principles and computational models have been proposed even by reflecting underlying physical control mechanisms. As synthesized speech quality is finally judged by subjective listening tests that usually take a long time, direct relation between perceptual naturalness and prosody control parameters have been longed for. Though the importance on perceptual studies on prosody has been widely acknowledged among researchers in speech synthesis, studies have not been carried out.

In this dissertation, perceptual studies are conducted to investigate fundamental characteristics of timing in sentence. Moreover, direct mapping between subjective naturalness and objective prosodic parameters is tested to confirm the possibility of perceptual characteristics analysis in CALL system for the second language learning.

1.2 Approach and outline of the dissertation

We first investigated the naturalness evaluation of timing control. Previous studies have reported various factors that affect sensitivity to segmental duration distortion. However, their targets were limited to isolated word speech or small size of experiment. We have carried out systematic analyses on naturalness of segmental duration distortion in sentence speech. As we generally observe much larger positional and durational variations in a sentence utterance than those in an isolated word utterance [1][2], two analyses have been carried out. First one is on the effect of intra-phrase position. The other one is on the effect of speaking rate. Their experimental results are described in Chapter 2 and Chapter 3 respectively.

In Chapter 4, for further understanding the nature of the perceptual characteristics found in Chapter 2 and 3, we studied the correlation between speech generation and perception in timing control characteristics. The perceptual characteristics are formed under the strong influence of the exposed environment during the process of growth. Thus, listener’s perceptual sensitivity of segmental duration distortion at a particular context may reflect the duration deviation generally expected for a segment at that context in listener’s
environment. We tried to explain perceptual characteristics by the context-dependent generation precision through the analysis of large-scale speech corpora.

In Chapter 5, as a first step toward modeling of subjective naturalness evaluation using objective measures, we proposed a computational model for temporal naturalness. While we have precisely analyzed perceptual characteristics by focusing on single factors in Chapter 2–4, all these findings are to be harmoniously integrated into a computational model to emulate overall subjective evaluation. In this Chapter, using non-native speech, a computational model was made to correlate objective differences from native’s utterance and the corresponding subjective measurements. Furthermore, to integrate perceptual characteristics found in previous studies and in Chapter 2 and Chapter 3 into the proposed model, an additional computational modeling was tested. For confirmation of the effectiveness of perceptual studies, computational objective scores were compared with native’s subjective evaluation scores.
Chapter 2

Perceptual acceptability analysis of sentence speech I: Effect of intra-phrasal position

2.1 Introduction

In speech synthesis from text, segment duration is determined by using statistical techniques in order to replicate the duration of natural speech. Most of conventional previous duration models are designed to minimize the sum or average of errors between the calculated duration and the observed one as a measure of improvement [3]-[7]. Though this error measure is quite natural engineering viewpoint, it fails to reflect perceptual characteristics. The same duration distortions give different perceptual effect. Previous studies have reported various factors that give different sensitivities to the same segmental duration distortion [1][8]-[12]. Kato et al. systematically analyzed contextual and positional dependency of perceptual sensitivity to segmental duration distortion by using isolated word utterances. These analyses showed that vowel quality, voicing of the following segment, and position in a word had significant relations with acceptability of the segment duration distortion. These results suggest that the error measure should reflect perceptual characteristics.

To integrate these perceptual characteristics found in word speech into segmental duration control for text-to-speech synthesis, further studies are needed in sentence speech as

---

we do not have any knowledge on perceptual characteristics in sentence context. The perceptual characteristics in sentence context are expected to be more serious than those in isolated word, since we generally observe much larger durational and positional variations in a sentence utterance than those in an isolated word utterance [1][2].

Perceptual characteristics for segmental duration distortion in sentence speech have not been studied except in some small works. Sagisaka and Tohkura investigated the acceptability of an overall sentence in which segmental durations of natural speech were modified randomly. They reported that an acceptable range of the distortion was within about 30 ms [13]. Carlson et al. investigated the acceptability of the overall sentence in which vowels or consonant were modified. They reported that the modification of the vowel affected the acceptability more strictly than that of the consonant [14]. Hoshino and Fujisaki reported that the acceptable modification range of a segment in a sentence was smaller than that in an isolated word [15].

Though the above small scale perceptual studies on sentence speech do not give any general perceptual characteristics, perceptual experiments of isolated words by Kato et al. have shown that a few fundamental perceptual principles can be consistently applied to explain superficially diverse differences, i.e. the loudness of the segment in question and the loudness jump between the segments involved can explain many different context-dependent perceptual characteristics [10]–[12]. The loudness of the segment and the loudness jump between consecutive segments were highly correlated to the acceptability to change in segmental duration. That is, a segment with greater loudness or with a larger loudness jump was more susceptible to segmental duration distortion. Since these correlations can be regarded as matters at the psychophysical level, they seem to be applicable not only isolated words but also sentences.

Nevertheless, the effect of temporal position cannot be explained by the loudness properties. In an isolated word, it has been shown that the acceptability decreased more quickly for word-initial than for word-medial vowels as the duration change increased. There was no significant difference in the loudness properties between the word-initial vowels and word-medial ones. We do not know whether a positional effect can be observed in sentence speech as same as in isolated words since there exist further factors in sentence temporal control. For instance, duration variations due to phrasing such as the phrase-final lengthening are remarkable and widely observed in many languages [2][16]-[18].

According to a general law in psychophysics, Weber’s law, a longer segment would need a longer change to yield the same amount of change in perception. However, one
of the few previous studies only showed inconclusive results for sensitivity to phrase-final segment duration [9]. Klatt and Cooper (1975) reported that the temporal jnd (just noticeable difference) became shorter for vowel segments but longer for fricative consonant segments. Thus, the phrase-final effect on the perceptual accuracy of changes in segment duration still remains an open question. In addition to the issue of phrase-final position, there appears to be little consensus on the perceptual effect of temporal positions in sentence-level speech.

As a first step toward understanding the perceptual characteristics of change in segment duration in sentence speech, the present study focused on the temporal position of the segment in question. The major objectives of the study were to examine the effect of positions within a unit that is specific in a sentence rather than a word and to test the robustness of such an effect, if found, in various contexts in sentence speech. The present study chose a phrase as a base unit of interest by taking into account the syntactic nature of Japanese.

Japanese is an agglutinative language, and in a sentence form, any indeclinable content word is generally accompanied by a postposition or case particle to form a minor phrase or, simply, phrase, which is the minimum syntactic unit, and thus a pause can be placed only between phrases [19]. A phrase, therefore, is an appropriate unit in a sentence form for comparison with an isolated word. In this paper, the dependency of the acceptability degradation on the difference in position within a phrase is referred to as the intra-phrase positional effect. In the current study, we conducted three experiments to provide a sufficient range of contextual variation in stimuli. The first experiment investigated the intra-phrase positional effect in a sentence and the difference in the effect due to the phrase length or accent type. The second experiment investigated the intra-phrase positional effect by using multiple carriers and a phrase without a carrier having a fixed phrase accent type. The third experiment investigated the effect of pause position on the intra-phrase positional effect.

2.2 Experiment 1: Intra-phrase positional effect

Since a significant positional difference was only observed between word initial position and word medial position in isolated words [11], the first experiment investigated the difference at the initial, medial and final positions of phrases embedded in a carrier sentence. The isolated word study used fixed length words, i.e., four-mora words, and did
CHAPTER 2. PERCEPTUAL ACCEPTABILITY ANALYSIS OF SENTENCE SPEECH I: EFFECT OF INTRA-PHRASAL POSITION

not systematically alter the accent type. Therefore, the first experiment also investigated the influence of phrase length and accent type by using four- and five-mora phrases with three different accent types.

2.2.1 Experimental conditions

Speech stimuli

The following Japanese phrases were used as speech stimuli: /dekirebal/ (if possible), /kodomogal/ (a child), /tonikakul/ (anyway), /ldokomademol/ (thoroughly), /rokujikaral/ (from six o’clock), and /tomodachigal/ (friend). Three of them were four-mora (four-syllable) phrases and three of them were five-mora (five-syllable) phrases. They had three different accent types (no accent, accented on either first or second mora). These phrases were embedded in the uniform carrier sentence /arewa/ (that) xxx /hanashimasu/ (talk) (xxx was replaced by each of the six phrases). They are natural Japanese sentences and were uttered at a normal speech rate by one female native speaker of Japanese. They were recorded with a DAT device (DTC-2000ES, SONY) using a microphone (SM58, SHURE).

In each stimulus, the duration of one vowel was either lengthened or shortened. Each position of the vowels in the phrases was modified. Accordingly, there were four levels of positional factors for the four-mora phrases and five levels for the five-mora phrases. Original durations of vowels range from 45 ms to 130 ms. There were five levels of change in duration for each direction, i.e., 10, 20, 30, 40, and 50 ms. The modifications of speech stimuli were carried out using the STRAIGHT speech analysis-synthesis method [20] at 12-kHz, 16-bit precision. The shortening manipulation of the shortest vowels did not reach full-step modification. Preliminary listening to all of the manipulated stimuli confirmed that no phonemic shift had occurred in either the target portions or the surrounding phonemes. In total, 274 sentence stimuli were prepared, i.e., 10 modification steps × 27 segments + 6 unmodified - 2 incomplete steps of modification for the shortest vowels.

Listening conditions

Stimuli were presented to the listeners through headphones (STAX SRS-2020 Basic System II ). The average presentation level was 78 dB SPL (A-weighted) measured with a sound level meter (Type 2231, Brüel & Kjær) through a condenser microphone (Type 4134, Brüel & Kjær) mounted on an artificial ear (Type 4153, Brüel & Kjær). The ex-
2.2. EXPERIMENT 1: INTRA-PHRASE POSITIONAL EFFECT

Figure 2.1: Difference in perceptual sensitivities at intra-phrase positions. An example illustrating a difference in acceptability-change between three positions. The vowels subjected to duration modification were initial, medial and final vowel for four- and five-mora phrases.

Experiments were done in a room whose average background noise level was 31 dB SPL (A-weighted), which was measured at the location of the listener with a sound level meter (Type 2231, Brüel & Kjær) and a condenser microphone (Type 4155, Brüel & Kjær). To help the listeners focus on the target of evaluation, textual sentences were also displayed on a computer screen and the phrase containing the modified vowel was underlined. Seven adults listeners with normal hearing participated in the experiments. All of them were native speakers of Japanese. They were asked to evaluate the subjective acceptability of these stimuli on a seven-point rating scale ranging from “1” to “7”, with the larger number corresponding to greater acceptability. All stimuli were randomized and presented to each listener four times.

2.2.2 Results

The obtained rating scores were first pooled over four repetitions, and then the maximum of the pooled scores for each listener was subtracted from each pooled rating score for
each listener to cancel out individual response biases. Hereafter, this normalized value is referred to the *acceptability rating score* which is always negative. Since Kato *et al.* have shown that a parabolic curve can be fitted to the acceptability rating score for isolated word utterances [10], a parabolic curve was fitted to each target vowel and listener. Figure 2.1 shows the acceptability rating scores pooled over seven listeners for four- and five-mora phrases. The three curves in the figure correspond to the initial, medial (4 mora phrase: 2nd - 3rd mora, 5 mora phrase: 2nd - 4th mora), and final vowel for each phrase. As the duration change increases, the acceptability rating scores decrease more quickly for initial vowels than for medial or final vowels.

Although a similar declination was observed in all individual plottings obtained for each combination of listeners and target vowels, the size of the horizontal or vertical shift of the top point varied depending on the listener or target vowel. The horizontal value of the top point reflects the amount of modification required to obtain the most preferred duration of the target vowel from its original, as-produced, duration. The vertical value of the top point reflects the score which the listener gave to his/her best duration. These top point shifts can be regarded as listener’s response biases.

Since the absolute value of the second-order coefficient of the parabolic curve indicated a degree of the acceptability degradation irrespective of these response biases, it was used to find consistent position-dependent acceptability degradation. Hereafter, the absolute value of the second-order coefficient of the parabolic curve is referred to as the *vulnerability index*. The vulnerability indices pooled over seven listeners for four-mora and five-mora phrases are shown in Figure 2.2 and Figure 2.3, respectively. These figures show that vulnerability indices at initial were largest and those at final were smallest. The effect of position in a phrase and target phrase on the vulnerability index was statistically tested by a two-way ANOVA of repeated measures with listeners as a blocking factor. To take an average for comparing the current results with those of the three-mora experiment in the following section, intermediate levels of the positional factor (two for the four-mora phrases and three for the five-mora phrases) were merged into one “medial” level. The main effect of position was significant in both four- and five-mora phrases [F(2, 24) = 23.94, p < 0.05; F(2, 20) = 12.36, p < 0.05]. The main effect of target phrase was significant in the four-mora phrases [F(2, 24) = 10.71, p < 0.05], but not in the five-mora phrases. Multiple comparisons between these averages of the vulnerability indices using Tukey-Kramer’s HSD showed that the difference between the averages of vulnerability indices at the initial and final positions, as well as at the medial and final positions, was
2.2. EXPERIMENT 1: INTRA-PHRASE POSITIONAL EFFECT

Figure 2.2: Vulnerability indices at intra-phrase positions in four-mora phrases.

Figure 2.3: Vulnerability indices at intra-phrase positions in five-mora phrases.
significant in both four- and five-mora phrases \[ p < 0.05 \].

The previous study showed that an initial vowel was more vulnerable to change in segmental duration than a medial vowel [11]. The same tendency was observed for all phrases although the effect was not statistically significant. Furthermore, the least vulnerability at the final position, which had not been tested in the previous studies, was generally observed for all phrases. That is, the acceptability degradation decreased in order of position from initial to final in a phrase. Although the effect of target phrase was significant in the four-mora phrases according to ANOVA, no difference in the acceptability degradation due to the accent was observed. Such positional effect in a phrase (intra-phrase positional effect) was observed in both four- and five-mora phrase regardless of phrase length or accent type.

2.2.3 Discussion

In this section, we first discuss the relation between the vulnerability index and the original as produced duration. According to Weber-Fechner’s law, a general psychophysical law, a longer segment is expected to yield a lower sensitivity to the same absolute change in segmental duration. That is, a vowel with a longer duration would have a smaller vulnerability index. In the present experiment, however, the durations of the phrase-initial vowels were not always shorter than those of the phrase-final vowels, while the vulnerability indices of phrase-initial vowels were generally larger than those of the phrase-final vowels. The results of correlation analyses between the original duration and the vulnerability index did not show any consistent tendency for the correlation. Pearson’s product-moment correlation coefficients for each listener ranged from -0.08 to 0.26.

Bochner et al. reported that the absolute jnd of the vowel duration increased with an increase in the vowel duration [8]. On the other hand, Klatt et al. reported that vowel duration had no significant correlation with the jnd’s of vowel duration [1], and Kato et al. also reported that the original duration had no significant correlation with the vulnerability index [11]. The current results agree with the latter results. The range of vowel durations in Bochner et al.’s, Klatt et al.’s, and Kato et al.’s studies were 75-175ms, 165-340ms, and 35-145ms, respectively. That in the current study was 31-176ms, which is comparable to that of Bochner et al. and Kato et al. but not to that of Klatt et al. Therefore, the range of original duration would not be able to explain the disagreement of these results. Kato et al. explained the disagreement of these results by the difference in context. Bochner et
al. used vowels in only monosyllabic context while Kato et al. and Klatt et al. commonly used those in polysyllabic contexts. The monosyllabic context might make the influence of the original vowel duration effective while the polysyllabic context reduces or hides it. This explanation also seems to be applicable to the current study because the study employed sentence speech, which indeed had polysyllabic context, and found no consistent correlation between the original duration and the vulnerability index. Therefore, it can safely be stated that the intra-phrase positional effect could be observed regardless of the original duration.

Next, we discuss the relation between the vulnerability index and the loudness properties. Previous studies reported that a segment with a greater loudness or with a larger loudness jump, that is, the difference from an adjacent segment, was more susceptible to segmental duration distortion [10]-[12]. To test whether these loudness properties generated the intra-phrase positional effect, we performed a loudness analysis on the stimuli of the current experiment. We first calculated the loudness contour for each stimulus every 2.5 ms with a 20-ms rectangular window. This is done according to the ISO 532 method B [21] while assuming a diffuse field and using Zwicker et al.’s algorithm [22]. Then, we picked out the median value from the entire loudness contour of each segment as the representative loudness of the segment. Figure 2.4 shows one example of the loudness analysis. The results of the loudness analysis showed that the loudness properties of the phrase-initial vowels were not always larger than those of the phrase-final vowels while the vulnerability indices of the phrase-initial vowels were generally larger than those of the phrase-final vowels. Therefore, it is also stated that the intra-phrase positional effect could be observed regardless of the loudness and the loudness jump.

The original duration and loudness properties could not account for the present intra-phrase positional effect. This agrees with a previous non-speech study. Tanaka et al. have reported a positional effect similar to that found in the present study by measuring temporal sensitivity to click sequences with regular intervals [23]. Their listeners were the most sensitive to the initial interval in a sequence and the least sensitive to the final one, with the intermediate one in between. Moreover, no interval was filled with a sound. This was the case even when the test intervals were isochronous and all clicks had the same sound level. That is, the original intervals and loudness-related properties were constant regardless of the temporal position. Therefore, the agreement between the current positional effect and that of the previous non-speech study supports the independence of the positional effect from both original duration and loudness-related properties and, more-
over, suggests that the intra-phrase positional effect is based on a common mechanism with click sequences.

2.3 Experiment 2: Intra-phrase positional effect in various contexts

In sentence speech, the acceptability of the change in the segmental duration possibly depends on the context surrounding the target phrase and the position of the phrase itself in a larger linguistic unit such as a breath group or a sentence. Furthermore, in a fixed context such as that of Experiment 1, listeners may be too sensitive to the change in segmental duration. Klatt et al. randomly presented a phrase embedded in seven sentences, including in isolation, under conditions as close to normal listening as possible [1]. To confirm the intra-phrase positional effect observed in Experiment 1 under various situations, an additional experiment was conducted with larger stimulus variation in a context. In Experiment 2, phrases embedded in three carriers and isolated phrases selected from an existing database were tested using the same experimental procedures as in Experiment 1.

2.3.1 Experimental conditions

Speech stimuli
In Experiment 2, three-mora phrases, rather than four- or five-mora ones, were used as target phrases to minimize the size of the experiment and to prevent the listeners from experiencing unnecessary strain. Three phrases were selected from the ATR speech database for synthesis [24]: /haruga/ (spring), /kuruto/ (come), and /nodemo/ (in the field) that provided the maximal contextual variation in the database. The position of accent was fixed at the first mora because no significant difference in the acceptability degradation due to the accent type was observed in Experiment 1. Three sentences and an isolated phrase were selected per phrase as shown in Table 2.1. In total, nine sentences and three isolated phrases were selected. They are natural Japanese sentences that were uttered at normal speech rate by one male professional narrator. Each vowel at phrase initial, medial and final positions was either lengthened or shortened, as described in the Experiment 1. Original durations of these vowels ranged from 50 ms to 155 ms. The shortening of the shortest vowel did not reach the full-step modification. In total, 371 sentence stimuli were prepared, i.e., 10 modification steps × 36 segments + 12 unmodified - 1 incomplete modification step for the shortest vowel.

**Listening conditions**

Listening evaluation was carried out by seven listeners in the same fashion as in the Experiment 1.

### 2.3.2 Results and discussion

The acceptability rating scores and the vulnerability indices were calculated as in Experiment 1. The vulnerability indices pooled over seven listeners are shown in Figure 2.5. The difference in vulnerability index due to the position was observed in almost every phrase.

The effect of position in a phrase and target phrase on the vulnerability indices was tested by a two-way ANOVA of repeated measures with listeners as a blocking factor. The main effect of position in a phrase was significant \([F(2, 24) = 33.98, p < 0.05]\). The main effect of target phrase was not significant. Multiple comparisons among these averages of vulnerability indices using Turkey-Kramer’s HSD indicated that the difference between the averages of vulnerability indices at the initial and final position, as well as that at the medial and final positions, was significant \([p < 0.05]\).

The vulnerability indices of the phrase-initial segments were the largest, the ones at final were smallest, and the ones at medial were at a level between initial and final.
Figure 2.5: Vulnerability indices at intra-phrase positions in three-mora phrases. (a)-(c), and (d) show sentence stimuli and only phrase stimuli. They correspond to “Phrase ID” of Table 2.1.
Table 2.1: Stimuli of Experiment 2. Target phrases are underlined.

<table>
<thead>
<tr>
<th>phrase ID</th>
<th>sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>haruga (a)</td>
<td>haruga kuruto nodemo yamademo hanaga ichimenni kireini saku</td>
</tr>
<tr>
<td>haruga (b)</td>
<td>haruga kuruto hanaga nodemo yamademo ichimenni kireini saku</td>
</tr>
<tr>
<td>haruga (c)</td>
<td>haruga kuruto</td>
</tr>
<tr>
<td>haruga (d)</td>
<td>haruga</td>
</tr>
<tr>
<td>kuruto (a)</td>
<td>haruga kuruto nodemo yamademo hanaga kireini ichimenni saku</td>
</tr>
<tr>
<td>kuruto (b)</td>
<td>haruga kuruto hanaga nodemo yamademo kireini ichimenni saku</td>
</tr>
<tr>
<td>kuruto (c)</td>
<td>haruga kuruto saku</td>
</tr>
<tr>
<td>kuruto (d)</td>
<td>kuruto</td>
</tr>
<tr>
<td>nodemo (a)</td>
<td>haruga kuruto nodemo yamademo ichimenni kireina hanaga saku</td>
</tr>
<tr>
<td>nodemo (b)</td>
<td>nodemo yamademo kireina hanaga saku</td>
</tr>
<tr>
<td>nodemo (c)</td>
<td>kireina hanaga nodemo yamademo saku</td>
</tr>
<tr>
<td>nodemo (d)</td>
<td>nodemo</td>
</tr>
</tbody>
</table>

though the difference between the initial and medial positions did not reach a significant level, the intra-phrase positional effect was observed in almost all phrases. The effect found in the current three-mora phrases was in good agreement with that of four- and five-mora phrases in Experiment 1, despite the contextual heterogeneity of the current stimuli.

We then moved on to discuss more global contextual effects, i.e., the position of the target phrase in a breath group. In terms of the position in a breath group, the target phrases were divided into two groups. The first group included the phrases at the non-initial position, i.e., kuruto (a, b, c) as shown in Table 2.1, and the second group included those at the initial position (the other phrases). The phrases nodemo (a and c) follow a pause. The sentence-initial phrases and isolated phrases were regarded as members of the initial position group because they follow a pause. Comparing vulnerability indices pooled over the initial or non-initial position in the breath group in each phrase, a clear difference in the vulnerability indices between initial and medial intra-phrase positions was observed for the initial position group. On the other hand, no such clear difference was observed for the non-initial position group. These observations could be accounted for if we assume that the preceding acoustic separation by a pause is the requisite condition for high sen-
sitivity to the phrase-initial segments in the intra-phrase positional effect. However, the current experiment was not systematic enough in terms of the pause position to confirm this hypothesis. Therefore, the next experiment investigated the effect of a separation by a pause in the intra-phrase positional effect.

2.4 Experiment 3: Robustness of intra-phrase positional effect in a breath group

We next investigated the robustness of the intra-phrase positional effect in the breath group. Previous studies have generally focused on perceptual phenomena at sentence final or phrase final positions [1][9], which are not preceded by a pause but followed by a pause. The results of Experiment 2 suggest an effect of a preceding pause, but the experiment was not systematic enough to confirm this suggestion. Experiment 3 focused on the phenomena occurring at the positions after a pause and systematically tested the influence of a preceding pause on the intra-phrase positional effect by comparing the acceptability of changes in duration between the first and second vowels in a phrase.

2.4.1 Experimental conditions

Speech stimuli

Eighteen Japanese phrases were selected as the speech stimuli as shown in Table 2.2. They were three-mora phrases and had one of two different types of accent (no accent or accented on the first mora). The phrases were divided into three groups according to the difference in carrier sentences as shown in Table 2.3. Each carrier sentence had two styles. In one of the two styles, the target phrase was positioned following the pause. In the other style, the target phrase was positioned medially in the breath group not following the pause. These sentences are natural Japanese sentences and were uttered at a normal speech rate by one female native speaker of Japanese. The first and second vowels in the target phrases were tested. The original durations of the tested vowels ranged from 21 ms to 113 ms. The pause durations ranged from 210 to 537. In each stimulus, the duration of one of the two test vowels in the target phrase was either lengthened or shortened as described in the Experiments 1 and 2. In total, 722 sentence stimuli were prepared, i.e., 10 modification steps × 18 phrases × 2 vowel positions in a phrase × 2 phrase positions in
2.4. EXPERIMENT 3: ROBUSTNESS OF INTRA-PHRASE POSITIONAL EFFECT IN A BREATH GROUP

Table 2.2: Target phrases used in Experiment 3.

<table>
<thead>
<tr>
<th>group</th>
<th>phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>betsuni (in addition), hoboga (child minder), kimini (to you), sarani (additionally), sonogo (after that), suguni (in a minute)</td>
</tr>
<tr>
<td>2</td>
<td>chibade (in Chiba), hahaga (mother), karemo (he also), kimimo (you also), moride (in the forest), sakini (in advance)</td>
</tr>
<tr>
<td>3</td>
<td>bujini (in safety), chijimo (governor also), dotede (in the bank), narade (in Nara), soboga (grandmother), t sureto (with companion)</td>
</tr>
</tbody>
</table>

Table 2.3: Carrier sentences used in Experiment 3. Slashees mark the positions of pauses; each target phrase in corresponding group of Table 2.2 was placed at either position 1 or 2.

<table>
<thead>
<tr>
<th>group</th>
<th>carrier sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rikai dekina katta node / position 1 sono mondai dake position 2 denwa de kiita.</td>
</tr>
<tr>
<td>2</td>
<td>Naze shitte iruka to iuto / position 1 tashika position 2 kiita kara desu.</td>
</tr>
<tr>
<td>3</td>
<td>Senshu no nichiyobi ni / position 1 soreo position 2 mita hazu dakedo yoku oboete inai.</td>
</tr>
</tbody>
</table>

a breath group + 36 unmodified - 34 incomplete modification steps for shorter segments.

Listening conditions

Listening evaluation was carried out by eight listeners in the same fashion as in the Experiments 1 and 2.

2.4.2 Results and discussion

The acceptability rating scores and the vulnerability indices were calculated as in Experiments 1 and 2. Figure 2.6 shows the vulnerability indices pooled over eight listeners as a function of the vowel position in phrase and the phrase position in a breath group. The vulnerability indices of the first vowel in a phrase were clearly larger than those of the following vowel whereas the differences in the vulnerability index due to the difference in the phrase position in a breath group was marginal.

The effects of a phrase position in the breath group, vowel position in the phrase, and
target phrase on the vulnerability indices were tested by a two-way ANOVA of repeated measures with listeners as a blocking factor. The main effect of position in the breath group was not significant. On the other hand, the main effects of position in a phrase and target phrase were significant \( F(1,575) = 20.01, p < 0.05 \), \( F(17,575) = 4.50, p < 0.05 \).

The intra-phrase positional effect was observed without the effect of pause. The initial and medial of the acceptability degradation were robust for the pause. These results suggest that acoustic division does not affect the vulnerability. The intra-phrase positional effect seems to be similarly observed regardless of position in a sentence. However, the vulnerability index of a phrase-final position was not measured in the current experiment. Since the phrase-final position can be the final in the breath group or sentence, the intra-phrase positional effect may be affected by a phrase final lengthening or sentence final shortening.

### 2.5 General discussion

The inclination of the intra-phrase positional effect observed in the present three experiments agreed with that observed in the previous non-speech study [23]; in both studies, listeners were stricter with a temporal change in the initial position of a click sequence or phrase than that in the final position. This agreement between perceptual phenomena
2.6. CONCLUSIONS

in speech and non-speech studies suggests that both positional effects share a common perceptual mechanism probably involving the general auditory basis. Therefore, such an effect due to temporal positions in a speech unit would be expected irrespective of difference in language. The agreement also suggests that the unit within which the positional effect functions is not necessarily a linguistic unit such as a phrase but a certain perceptual unit in a broader sense. The results of Experiment 3 further suggests that such a unit does not have to be acoustically separate from its context. In the present experimental situations, all materials were naturally spoken Japanese and all listeners were native speakers of that language who were informed about the position of the target phrase in the context (breath group or sentence) before starting their tasks. Therefore, we can reasonably assume that the listeners could easily pick up the target phrase as a perceptual unit within which the positional effect would be invoked, even when the target was not explicitly signaled by an acoustic separation.

Another important finding of the present study is that the intra-phrase positional effect could be observed independently of the original segment durations as well as the loudness-related properties of the segments. This finding suggests that the effect may not be explained within the scope of a general psychophysical law, i.e., Weber’s law. In particular, the observed broadness in perceptual judgment at the phrase-final position, where a vowel tends to be elongated, may not be solely accounted for by this elongation property. It is reasonable to assume the presence of a general decrement tendency in temporal sensitivity at unit-final positions. This view is consistent with the finding in speech production that the control precision of segment duration would decrease at unit-final positions regardless of the length of the unit in question such as a word, minor phrase, accent phrase, breath group, and sentence [25].

The reason why the effect would emerge in the current direction, i.e., a higher sensitivity at the initial position and a lower sensitivity at the final position, is not clear at this moment. Further extensive and carefully designed studies are needed in both speech and non-speech domains to clarify this issue.

2.6 Conclusions

With the aiming of improving the naturalness criteria for speech synthesis, the effect of position within a phrase on the acceptability of segment duration distortion was investigated in sentence speech. The intra-phrase positional effect was observed in each of three
perceptual experiments, i.e., a duration distortion in the phrase-initial segment was the least acceptable by listeners and that in the phrase-final segment was the most acceptable, with that in the phrase-medial segment in between. This tendency is in good agreement with those observed in works evaluating acceptability evaluation for word speech [11] and the temporal sensitivity to non-speech click sequences [23].

Experiments 1 and 2 showed that the intra-phrase positional effect was observed irrespective of variations in the attribute of a target phrase: (1) length of the phrase (three, four, or five moras), and (2) accent type of the phrase (unaccented, accented on either first or second mora). Experiments 2 and 3 showed that the effect was observed irrespective of variations in the context of a target phrase: (3) with or without a carrier sentence, (4) difference in the carrier sentence, and (5) position in a breath group (initial or non-initial).

The intra-phrase positional effect could be observed independently of the original segment durations as well as the loudness-related properties of the segments. That is, the current effect found in sentence speech could be explained by neither the original duration nor loudness-related properties, since this was also the case for the positional effect in isolated word speech.

These observations and findings suggest the generality of the effect found. A similar positional effect may be observed across languages (or even in non-speech perception) within a certain unit or group providing it was perceptually significant. The extensive variation in the current experimental conditions demonstrated the robustness of the intra-phrase positional effect. This fact implies that a more perceptually valid criterion for duration rules could be expected by incorporating the effect found in this work.
Chapter 3

Perceptual acceptability analysis of sentence speech II: Effect of speaking rate

3.1 Introduction

In speech synthesis from text, segment duration is determined by using statistical techniques to replicate the duration of natural speech. Many previous duration models have adopted the sum or average of errors between the calculated duration and the observed duration as a measure of evaluation [3]–[7]. The rationale of this measure lies in the fact that the physical error in each segment duration has an equivalent weight for perceptual judgment. However, this measure has one problem in that it might not properly reflect a perceptual characteristic for distortions in segment duration. Some experimental findings suggest that this is the case [1][8]–[12][26]. In particular, Kato et al. systematically analyzed contextual and positional dependency of perceptual sensitivity to distortions in segment duration by using isolated word utterances. These analyses showed that vowel quality, voicing of the following segment, and position in a word had significant effects on the acceptability of the distortion in segment duration. Furthermore, they have shown that the loudness of the segment and the loudness jump between consecutive segments were correlated with the acceptability of change in segment duration. That is, a segment with greater loudness or with a larger loudness jump was more susceptible to distortion in

---

segment duration. This suggested that psychophysical characteristics can partly explain context-dependent perceptual characteristics such as vowel quality and voicing of the following segment. These results also suggest that the conventional error measure does not reflect perceptual characteristics correctly.

The works of Kato and his colleagues have been done using isolated words. For assured application of such perceptual characteristics to segmental duration control for text-to-speech synthesis, further studies are needed in sentence speech because speech synthesis systems are in general assumed to produce sentence speech. Moreover, the perceptual characteristics used for sentence speech may be different from those used for word speech, since we generally expect much larger durational and positional variations in a sentence utterance than those in an isolated word utterance [1][2]. Several pioneering studies have investigated the perceptual characteristics of the distortion in segment duration in sentence speech. Sagisaka and Tohkura investigated the acceptability of an overall sentence in which every segment duration of natural speech was modified randomly, and they reported that an acceptable range of modification extended to about 30 ms [13]. Carlson et al. investigated the acceptability of the overall sentence in which every vowel or consonant was modified and reported that modification of vowels affected acceptability more strongly than did that of consonants [14]. Hoshino and Fujisaki reported that the acceptable modification range of a segment in a sentence was smaller than that in an isolated word [15]. Nevertheless, none of these sentence studies addressed the factors reported in the study on isolated words by Kato et al. Therefore, we have studied the perceptual characteristics of sentence speech while addressing the positional factor, which could not be explained by the loudness properties. The results showed that the degradation in acceptability to change in segment duration was greater according to the order of position, from initial to final, in a phrase. As the next step in contributing to segmental duration control, this paper addresses speaking rate, since local speaking rate is also a crucial phrase-level control factor found in natural speech. We investigated the effect of longer-range control of the speaking rate on the acceptability to change in segment duration.

3.1.1 Effect of original duration and tempo

If we define that the speaking rate (or tempo of ongoing sequences) is determined by the rate of occurrence of successive events, then the time intervals between consecutive events would be the critical physical factor involved in judgment of speaking rate. One
could argue that changes in duration themselves may be epiphenomena. However, in
general, changes in segment durations systematically correspond to the speaking rate.
Therefore, it is not anomalous to investigate effects of the speaking rate in the detection
of changes in duration. Assuming that acceptability is evaluated based on the perceived
amount of change in segment duration, which is generally estimated as a relative value to
the original duration, the acceptability degradation caused by the same amount of absolute
change would become smaller when the original duration is longer. According to this
assumption, the acceptability degradation becomes greater as the speaking rate becomes
faster. A previous study by Kato et al. does not appear to support this assumption:
They reported that there was no correlation between the degradation in acceptability of
changes in segment duration and original durations that ranged from 35 ms to 145 ms
[11]. However, Kato et al.’s stimuli were only presented at a normal speaking rate. The
variation in speaking rate was so limited that their results could not conclusively invalidate
the assumption.

Kato et al. reported that the degree of acceptability degradation correlated with the de-
tectability of a duration modification, based on a direct comparison using the same speech
materials and the same listeners [27]. For the absolute just noticeable difference (jnd) of
segment duration, Klatt and Cooper [9] and Bochner et al. [8] have reported a relation
between the original duration and the jnd. Klatt and Cooper showed that the absolute
jnd’s for vowel durations (165-340 ms) had no significant correlations with the vowel du-
rations in “deal(er),” which was embedded in various sentences [9]. This result agreed
with Kato et al.’s acceptability test mentioned above. On the other hand, Bochner et al.
showed that the jnd of vowel duration (75-170 ms) tended to increase with an increase in
the vowel duration in a monosyllabic context, i.e., CVC or isolated vowel segment [8].
This result disagreed with that of Kato et al. Moreover, non-speech studies have reported
that the absolute temporal jnd of a segment-like duration (50–200 ms) roughly correlates
with the corresponding base duration [8][28]–[30]. That is, the results showed that the
jnd for changes in duration or time interval tended to increase with an increase in the
original duration. These results agree with Bochner et al.’s findings. Considering these
results, it cannot be concluded that the original duration does not affect the acceptability
degradation.

Furthermore, since the speaking rate could be defined only by intervals consisting of
signals without the duration and is independent of the segment duration, we have to con-
sider the effect of tempo in addition to duration. There have been many studies dealing
with the jnd of tempo generated by an entire sequence (for example, the works of Drake and Botte [31] and Michon [32]), while few studies have dealt with the effect of tempo on the jnd of an interval within a sequence. Hibi reported that the relative jnd’s of an interval (140–300 ms) within twelve tone bursts was roughly constant [33]. That is, the jnd monotonously becomes shorter as tempo increases (interval lengths decrease). However, a change in speaking rate generally causes change in the segment duration of less than 140 ms. Hibi’s observation lacks data for this smaller range. Therefore, we should be cautious in generalizing his results, partly because of this range mismatch and partly because of the difference in stimuli (non-speech vs. speech).

So far, the effect of speaking rate has not been fully investigated. In the current study, experiments on speech and non-speech imitating speech were conducted while changing the speaking rate or tempo explicitly. Experiment 1 investigated the effect of speaking rate on the degradation in acceptability of change in segment duration. Experiment 2 investigated the same effect in non-speech to find a psychophysical or auditory basis applicable to speech-related phenomena.

### 3.1.2 Factors tested in the study

The current study examined three factors, i.e., speaking rate, position within a phrase, and presence/absence of a carrier sentence. In the non-speech experiment, the corresponding equivalents to these factors were tempo, temporal position, and presence/absence of a carrier.

**Speaking rate**

As mentioned in section 1.1, the previous studies did not deal with speech stimuli in which the speaking rate was changed explicitly. In Experiment 1, the effect of speaking rate on acceptability of change in segment duration was examined by using speech that was changed among three levels of speaking rate. In Experiment 2, the effect of tempo on the detectability of change in an interval filled with a tonal stimulus was examined by using non-speech stimuli at three levels of tempo corresponding to those of speaking rate in Experiment 1.

**Position within a phrase**

Japanese is an agglutinative language, and any indeclinable content word is generally accompanied by a postposition or case particle to form a minor phrase or, simply, a phrase, which is the minimum syntactic unit in a sentence form [19]. Therefore, we have tested
3.1. INTRODUCTION

phrases in various contexts to examine the effect of robustness of the linguistic unit on perception [26]. The results have shown that a duration change in the phrase-initial segment was generally the least acceptable, that one in the phrase-final segment was the most acceptable, and that one in a phrase-medial segment was in between (intra-phrase positional effect). This position-dependent tendency was observed regardless of the variations in phrase length, accent type, carrier sentence, presence of carrier sentence, and position in a breath group (initial or non-initial). The current study examined the intra-phrase positional effect when the speaking rate changed. In non-speech, Tanaka et al. measured temporal jnd’s for one of three successive intervals marked by a four-click sequence, which were designed to replicate the temporal structure of a four-mora word. The results showed that the jnd for the first interval of a sequence was significantly smaller than that for the third interval. The stimuli were sequences comprised of four clicks. The current study investigated the positional effect based on three divisions, i.e., initial, medial and final using sequences of three tones temporally flanked by other tones. These flanking tones can be regarded as providing a similar condition to that by carrier sentences for phrase stimuli.

Carrier

A previous study has investigated the effect of carrier sentence and shown that the intra-phrase positional effect was observed irrespective of the presence or absence of a carrier sentence [26]. However, since that study focused on the robustness of the positional effect within a phrase, the difference due to the presence/absence of a carrier sentence was not systematically controlled. Especially for the phrase-final segment, it may be assumed that the change in segment duration with a carrier sentence was more sensitive than that without it for the following reason. The phrase-final segment with a carrier sentence is not the end of the stimuli, while the phrase-final segment without a carrier sentence is the end of the stimuli, followed by silence. That is, if the carrier part trailing a target segment provides a supplemental cue to change in segment duration, the phrase-final segment could be perceived differently due to the presence or absence of carrier. Therefore, the effect of carrier on acceptability or on detectability was tested in speech or in non-speech, respectively.
3.2 Experiment 1: Effects in speech stimuli

The experiment was designed in a three-way factorial manner. The factors were speaking rate, position within a phrase, and presence/absence of a carrier sentence.

3.2.1 Experimental conditions

Speech stimuli

Ten Japanese phrases were prepared as speech stimuli as shown in Table 3.1. Since the three levels of acceptability degradation due to the position within a phrase are known, three-mora (three-syllable) phrases were selected as target phrases to minimize the size of the experiment. They had two different types of accent (no accent or accented on the first mora). The phrases were embedded in a uniform carrier sentence: sorewo (that) xxx kiita (listened, heard) (which means “xxx heard that”), where xxx was replaced by each of the ten phrases. They are natural Japanese sentences and were spoken at three speaking rates (fast, normal and slow) by a male professional narrator.

In each stimulus, the duration of one vowel was either lengthened or shortened. Original durations of vowels ranged from 21 ms to 300 ms. The average duration of phrases at the slow rate was more than two times as long as that at the fast rate as shown in Table 3.2. There were five steps of change in duration for each direction, i.e., 10, 20, 30, 40, and 50 ms. The modifications of speech stimuli were carried out using the STRAIGHT speech analysis-synthesis method [20] at 22.1-kHz, 16-bit precision. The shortening manipulation of the shortest vowels did not reach the full-step modification (13 vowels). In total, 921 sentence stimuli were prepared, i.e., (10 phrases × 3 positions in a phrase × 3 speaking rates) × 10 modification steps + 30 unmodified (= 10 phrases × 3 speaking rates) - 9 incomplete steps of modification for the vowels whose durations were less than 50 ms. Similarly, 926 phrase stimuli were prepared, i.e., (10 phrases × 3 positions in a phrase × 3 speaking rates) × 10 modification steps + 30 unmodified (= 10 phrases × 3 speaking rates) - 4 incomplete steps of modification for the vowels whose durations were less than 50 ms.

Listening conditions

Stimuli were presented to the listeners through headphones (SR-L Professional, driven by SRM-1 MkII, STAX). The average presentation level was 78 dB SPL (A-weighted)
Table 3.1: Target phrases used in Experiment 1.

<table>
<thead>
<tr>
<th>accent type</th>
<th>phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (no accent)</td>
<td>michide (on the street), nishide (in the west), sakini (on ahead), sonogo (after that), tomoni (together)</td>
</tr>
<tr>
<td>1 (accented on first mora)</td>
<td>doremo (any), hahaga (mother), narade (in Nara), sarani (moreover), sotode (outside)</td>
</tr>
</tbody>
</table>

Table 3.2: Average durations of speech material parts used in Experiment 1 (ms). Vowel durations are pooled by the position in a phrase. The average duration of “slow” speech is in general more than two times as long as its “fast” counterpart.

<table>
<thead>
<tr>
<th>speaking rate</th>
<th>carrier</th>
<th>sentence</th>
<th>phrase</th>
<th>1st vowel</th>
<th>2nd vowel</th>
<th>3rd vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>fast</td>
<td>with</td>
<td>1097</td>
<td>378</td>
<td>68</td>
<td>61</td>
<td>99</td>
</tr>
<tr>
<td>fast</td>
<td>without</td>
<td>-</td>
<td>398</td>
<td>72</td>
<td>69</td>
<td>120</td>
</tr>
<tr>
<td>normal</td>
<td>with</td>
<td>1599</td>
<td>532</td>
<td>93</td>
<td>82</td>
<td>149</td>
</tr>
<tr>
<td>normal</td>
<td>without</td>
<td>-</td>
<td>551</td>
<td>88</td>
<td>100</td>
<td>172</td>
</tr>
<tr>
<td>slow</td>
<td>with</td>
<td>2230</td>
<td>743</td>
<td>121</td>
<td>133</td>
<td>223</td>
</tr>
<tr>
<td>slow</td>
<td>without</td>
<td>-</td>
<td>1051</td>
<td>194</td>
<td>254</td>
<td>316</td>
</tr>
</tbody>
</table>
measured with a sound level meter (Type 2231, Brüel & Kjær) through a condenser microphone (Type 4134, Brüel & Kjær) mounted on an artificial ear (Type 4153, Brüel & Kjær). The experiments were done in a sound-treated room whose average background noise level was 16 dB SPL (A-weighted), which was measured at the location of the listener with a sound level meter (Type 2231, Brüel & Kjær) and a condenser microphone (Type 4155, Brüel & Kjær). Seven adult listeners with normal hearing participated in the experiment. All of them were native speakers of Japanese. They were asked to evaluate the subjective acceptability of each of the prepared stimuli on a seven-point rating scale ranging from “1” to “7”, with the larger number corresponding to greater acceptability. To assist the listeners in focusing on the target of evaluation for a with-carrier stimulus, the corresponding textual sentence was also presented on a computer screen while underlining the phrase containing the modified vowel. Similarly, the corresponding textual phrase was presented for a no-carrier stimulus. All of the stimuli were randomized and presented to each listener four times in total.

3.2.2 Results and discussion

The obtained rating scores were first pooled over four repetitions, and then they were subtracted by the maximum of the pooled scores for each listener to cancel out individual response biases. Hereafter, this normalized value is referred to as the acceptability rating score. In accordance with the procedure of our former works [10]–[12][26], a parabolic curve was fitted to each target vowel and listener. Figure 3.1 shows the acceptability rating scores pooled over seven listeners for the first vowel of ten phrases at a fast rate without a carrier sentence. The acceptability rating scores decrease as the duration change increases. Since the absolute value of the second-order coefficient of the parabolic curve represents a degree of the acceptability degradation for changes in a segment duration, we took this value as the object variable of Experiment 1 and referred to it as the vulnerability index. For cases in which the acceptability rating scores did not decrease with changes in segment duration, a value of zero was assigned to the vulnerability index. In this case, it could be assumed that the listener could not perceive the change in segment duration. Table 3.3 shows the number of cases where the systematic decrease in the acceptability rating score was observed as the degree of modification increased. The vulnerability indices except “0” pooled over listeners and phrases are shown in Figure 3.2. The figure shows that the vulnerability index decreased with the decrement of speaking rate and
Figure 3.1: An example illustrating differences in acceptability rating due to changes in vowel duration. This particular example shows the listeners’ responses to a vowel at initial position in a phrase that was spoken at fast rate without a carrier sentence. Each plot represents acceptability rating score pooled over four repetitions. The line represents a parabolic fitting to these acceptability rating scores. A larger acceptability rating score corresponds to a more acceptable response.

that it also decreased as the target position in a phrase changed from the initial to the final. The average vulnerability indices in the with-carrier condition tended to exceed their counterparts in the without-carrier condition (six cases out of nine).

The effects of speaking rate, position in a phrase, and presence/absence of a carrier sentence on the vulnerability index were statistically tested by a three-way factorial ANOVA of repeated measures with listeners as a blocking factor. The main effects of speaking rate and position within a phrase were significant \([F(2, 12) = 60.76, p < 0.0001; F(2, 12) = 86.78, p < 0.0001]\). On the other hand, the main effect of presence/absence of a carrier sentence was not significant \([F(1, 6) = 2.55, p < 0.162]\). The interactions between speaking rate and presence/absence of a carrier sentence, and those between speaking rate and position within a phrase, were statistically significant \([F(2, 12) = 6.10, p < 0.05; F(4, 24) = 10.38, p < 0.0001, \text{ respectively}]\). Then, multiple comparisons among vulnerability indices using Tukey-Kramer’s HSD (honestly significant difference) indicated
significant differences due to the speaking rate, except for the difference between the fast- and normal-rate \( p < 0.05, q(0.05, 3, 12) = 3.77, \text{HSD} = 9.85 \times 10^{-5} \), as well as significant differences due to the position within a phrase between any two pairs \( p < 0.05, q(0.05, 3, 12) = 3.77, \text{HSD} = 1.02 \times 10^{-4} \).

**Speaking rate**

Through the measurements of the acceptability degradation caused by the vowel lengthening and shortening, it was confirmed that the acceptability declined more rapidly as the speaking rate became faster. The insensitivity of the vulnerability index to the variation in the original duration reported in the previous study [11] might lead to the prediction that the speaking rate does not influence the acceptability degradation. However, the overall tendency of the effect of speaking rate was that the vulnerability index decreased with a decrement of the speaking rate. According to the post-session questionnaires, five of the seven listeners answered that rating the stimuli at the slow rate was most difficult, which implied the effect of speaking rate. This agrees with the larger numbers of vulnerability index “0” in the slow-rate. The difference between fast- and normal-rate was not clear. This is probably due to a sort of ceiling effect. That is, the current experimental measure
3.2. EXPERIMENT 1: EFFECTS IN SPEECH STIMULI

Table 3.3: The numbers of vulnerability indices that are larger than zero (left) and the numbers of all obtained indices (right) in Experiment 1 (7 listeners × 10 vowels for each carrier condition and position in a phrase). A remarkable decrement in the number of positive indices is observed for the slow rate and the third position in a phrase.

<table>
<thead>
<tr>
<th>speaking rate</th>
<th>carrier</th>
<th>1st vowel</th>
<th>2nd vowel</th>
<th>3rd vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>fast with</td>
<td>69/70</td>
<td>70/70</td>
<td>65/70</td>
<td></td>
</tr>
<tr>
<td>fast without</td>
<td>70/70</td>
<td>70/70</td>
<td>55/70</td>
<td></td>
</tr>
<tr>
<td>normal with</td>
<td>69/70</td>
<td>69/70</td>
<td>54/70</td>
<td></td>
</tr>
<tr>
<td>normal without</td>
<td>69/70</td>
<td>68/70</td>
<td>50/70</td>
<td></td>
</tr>
<tr>
<td>slow with</td>
<td>68/70</td>
<td>65/70</td>
<td>37/70</td>
<td></td>
</tr>
<tr>
<td>slow without</td>
<td>56/70</td>
<td>48/70</td>
<td>32/70</td>
<td></td>
</tr>
</tbody>
</table>

may not have a sufficient resolution to clarify the difference in listeners’ performances between these two conditions. In a non-speech click sequence, Hibi reported that a faster tempo yielded a smaller jnd of an interval [33]. This effect of presentation tempo agrees with the effect of speaking rate observed in the current experiment. Both results suggest that the overall tempo affects the sensitivity to a local segmental duration.

**Position within a phrase**

The effect of position within a phrase was shared by all speaking rates and carrier conditions. The vulnerability indices of the phrase-initial segments were the largest, those at the phrase-final were the smallest, and the ones at the phrase-medial were at a level between initial and final for each condition. This tendency agrees with that observed in previous studies using word [11] and sentence stimuli [26]. These results suggest that this intra-phrase positional effect is robust against the variation in speaking rate as well as the difference between with- and without-carrier conditions.

**Carrier sentence**

The effect of the presence/absence of a carrier sentence was not significant. In the questionnaires after the experiment, six of the seven listeners answered that the ratings with a
carrier sentence were easier than those without one. Additionally, the numbers of vulnerability index “0” with a carrier sentence of four listeners were greater than ones without a carrier sentence (Table 3.4). This indicates that the listeners could not perceive changes in segment duration without a carrier sentence as easily as those with a carrier sentence. However, no systematic difference due to presence/absence of carrier sentence was observed. Especially for phrase-final, it might be expected that a change in segment duration with a carrier sentence was more sensitive than one without it since the carrier part after a target segment might provide a supplemental cue of change in segment duration. However, there was no significant difference at phrase-final between with and without a carrier sentence except at the fast-rate \[F(1,132) = 6.10, p < 0.05\]. This ambiguity of the carrier effect might be caused by a floor effect. As the vulnerability indices were at the lowest level in these cases, the difference could not be observed. This might be supported by the numbers of vulnerability index “0” in Table 3.3, where the numbers at normal- and slow-rate were smaller.

**Original duration**

The Pearson’s product-moment correlation coefficients between the vulnerability indices and the original durations for each listener and for each condition are shown in Tables 3.4 and 3.5, respectively. A negative correlation was observed for all of the listeners: The correlation coefficients ranged from -0.313 to -0.760 (Table 3.4). The negative correlation indicates that the vulnerability index is larger when the original duration is shorter, implying a higher sensitivity for the shorter duration. This tendency agrees with the expectation mentioned in the Introduction. In contrast, no consistent correlation was observed when the speaking rate condition was kept constant (Table 3.5), which agrees with the previous study [11]. There are two possible reasons why correlations were observed when the speaking rate changed while no consistent correlation was observed when the speaking rate was constant. One reason is that the correlation was observed as a result of the effect of speaking rate, independent of the original duration. That is, the original duration itself does not affect the acceptability degradation but the speaking-rate does. However, since the speaking rate strongly correlates with the segment duration in natural speech, a correlation between the acceptability degradation and original duration was observed. The other reason is that the correlations can be observed when the original duration has an extremely wide variation. In fact, the range of original durations in the current study
Table 3.4: The number of vulnerability indices (larger than zero/total) and Pearson’s product-moment correlation coefficients between vulnerability index and original duration for each listener in Experiment 1. A negative correlation is clearly observed for all listeners.

<table>
<thead>
<tr>
<th>listener ID</th>
<th>carrier</th>
<th>vulnerability index</th>
<th>correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>with</td>
<td>86/90</td>
<td>-0.442</td>
</tr>
<tr>
<td></td>
<td>without</td>
<td>86/90</td>
<td>-0.651</td>
</tr>
<tr>
<td>02</td>
<td>with</td>
<td>82/90</td>
<td>-0.520</td>
</tr>
<tr>
<td></td>
<td>without</td>
<td>82/90</td>
<td>-0.708</td>
</tr>
<tr>
<td>03</td>
<td>with</td>
<td>85/90</td>
<td>-0.470</td>
</tr>
<tr>
<td></td>
<td>without</td>
<td>72/90</td>
<td>-0.627</td>
</tr>
<tr>
<td>04</td>
<td>with</td>
<td>76/90</td>
<td>-0.313</td>
</tr>
<tr>
<td></td>
<td>without</td>
<td>80/90</td>
<td>-0.760</td>
</tr>
<tr>
<td>05</td>
<td>with</td>
<td>77/90</td>
<td>-0.419</td>
</tr>
<tr>
<td></td>
<td>without</td>
<td>68/90</td>
<td>-0.337</td>
</tr>
<tr>
<td>06</td>
<td>with</td>
<td>80/90</td>
<td>-0.465</td>
</tr>
<tr>
<td></td>
<td>without</td>
<td>63/90</td>
<td>-0.450</td>
</tr>
<tr>
<td>07</td>
<td>with</td>
<td>79/90</td>
<td>-0.397</td>
</tr>
<tr>
<td></td>
<td>without</td>
<td>66/90</td>
<td>-0.663</td>
</tr>
</tbody>
</table>

Table 3.5: Pearson’s product-moment correlation coefficients between vulnerability index (larger than zero) and original duration for each carrier condition and position in a phrase in Experiment 1. No systematic tendency is observed within any single speaking-rate condition.

<table>
<thead>
<tr>
<th>speaking rate</th>
<th>carrier</th>
<th>1st vowel</th>
<th>2nd vowel</th>
<th>3rd vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>fast</td>
<td>with</td>
<td>0.054</td>
<td>0.328</td>
<td>-0.198</td>
</tr>
<tr>
<td>fast</td>
<td>without</td>
<td>-0.003</td>
<td>-0.058</td>
<td>0.004</td>
</tr>
<tr>
<td>normal</td>
<td>with</td>
<td>-0.162</td>
<td>0.048</td>
<td>-0.010</td>
</tr>
<tr>
<td>normal</td>
<td>without</td>
<td>0.033</td>
<td>-0.138</td>
<td>-0.015</td>
</tr>
<tr>
<td>slow</td>
<td>with</td>
<td>-0.092</td>
<td>-0.166</td>
<td>0.014</td>
</tr>
<tr>
<td>slow</td>
<td>without</td>
<td>-0.081</td>
<td>-0.129</td>
<td>-0.312</td>
</tr>
</tbody>
</table>
(20–381 ms) was three times as wide as that in Kato et al.’s study (35–145 ms). However, in speech, since tempo and duration could not be dealt with independently, we can not make conclusions at this time.

3.3 Experiment 2: Effects in non-speech stimuli

Experiment 2 was also designed in a three-way factorial manner. The factors were tempo, temporal position and presence/absence of a carrier.

3.3.1 Experimental conditions

Stimuli

Each stimulus was a complex sound of 500, 1000 and 1500 Hz with amplitude contours comprised of the alternation of slope and steady parts. The steady parts had 97 dB or silence. The target stimuli in this experiment were sequences of three tones with an equal duration arriving at an equal inter-onset interval. Modification in duration was applied to one of the three tones. Since the inter-stimulus interval, i.e., the gap between the two consecutive tones, was fixed, this durational modification caused a change in the inter-onset interval except when the third tone was the target of modification. Using these target stimuli of three tones as a reference can be regarded as a counterpart to the condition without the “carrier” in Experiment 1. To simulate the condition with the carrier, the target part was temporally flanked by two three-tone sequences, each having the same durations, intensities and inter-onset intervals as those of the target part. The spectral tilt of the carrier tones was different from that of the target tones to assist listeners in picking up the target part from the carrier part. The slope of the spectral tilt was rising in the target tones but falling in the carrier tones. There were three levels of inter-onset intervals, i.e., 100, 150, and 225 ms. Each inter-onset interval condition was combined with two duration conditions, i.e., 50, 75, 112.5 and 168.75 ms, as shown in Figure 3.3. Each duration has two 5-ms slopes and fluctuates between 25% of each duration. In total, 12 stimuli were prepared as standard stimuli. For each stimulus, one duration of the target part was either lengthened or shortened. There were three steps of change in duration for each direction, i.e., 12, 26 and 40 ms. Accordingly, 252 stimuli were prepared for comparison stimuli, i.e., (3 inter-onset intervals (tempi) × 3 positions in a target part × 2 tone durations) × (6 modification steps + 1 unmodified) × 2 carrier patterns (with/without carrier).
3.3. EXPERIMENT 2: EFFECTS IN NON-SPEECH STIMULI

Figure 3.3: Schematic example showing the temporal structures of nonspeech stimuli in three different tempi in Experiment 2. The horizontal and vertical axes refer to time and sound pressure level.

**Listening conditions**

The standard and comparison stimuli were presented to the listeners through headphones. The experimental apparatus and room were the same as in Experiment 1. Ten adult listeners with normal hearing participated in the experiments. All of them were native speakers of Japanese. They were asked to rate the difference between standard and comparison stimuli using eight numerical categories: “0” to “7”. The larger number corresponds to a larger subjective difference. To help the listeners focus on the target of evaluation, pattern diagrams of tones using characters were also displayed on the computer screen and the target part containing the modified duration was underlined. All stimuli were randomized and presented to each listener sixteen times in total.

3.3.2 Results and discussion

The detectability index ($d'$) was measured for the difference between each pair of standard and comparison stimuli by the method of constant stimuli. The obtained responses were pooled over all listeners for each category, and then the detectability index $d'$ for each comparison stimulus was estimated in accordance with the theory of signal detection [34][35]. The detectability indices pooled over duration and duration changes are shown
in Figure 3.4. The figure shows that a faster tempo resulted in a greater detectability index and that detectability degraded as the position of the target tone changed backward. The tendency for the tone without a carrier to have a greater detectability index was observed at first and second positions.

The effect of tempo, temporal position, and presence/absence of a carrier on the detectability index was statistically tested by a three-way factorial ANOVA of repeated measures with duration changes as a blocking factor. The main effects of tempo, position and presence/absence of the carrier were significant $F(2, 10) = 11.58, p < 0.001; F(2, 10) = 7.83, p < 0.01; F(1, 5) = 19.46, p < 0.01$. The interactions were not significant.

**Tempo**

Through the measurements of the difference between the standard and the comparison caused by duration lengthening and shortening, it was confirmed at each position with or without a carrier that the detectability further decreased as the tempo became faster. This result also agrees with Hibi’s study using empty intervals. In Experiment 1, the acceptability degradation was greater as the speaking rate became faster. Both results show that listeners are more sensitive to the duration change of a local segment as overall tempo becomes faster.
Table 3.6: Pearson’s product-moment correlation coefficients between detectability index and original duration for each amount of duration change in Experiment 2. A negative correlation is observed for any amount of duration change.

<table>
<thead>
<tr>
<th>change in tonal duration (ms)</th>
<th>-40</th>
<th>-26</th>
<th>-12</th>
<th>12</th>
<th>26</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>correlation coefficient</td>
<td>-0.733</td>
<td>-0.750</td>
<td>-0.098</td>
<td>-0.368</td>
<td>-0.577</td>
<td>-0.519</td>
</tr>
</tbody>
</table>

**Temporal position**

The detectability index at initial position was the largest, that at final was the smallest, and the one at medial was an intermediary level at each tempo. The positional effects were observed for all conditions, as in Experiment 1. This result agrees with the previous non-speech study [23] as well as the result in the speech experiment of the current paper. Taken together, these agreements between perceptual phenomena in speech and in non-speech experiments suggest that both positional effects share a common perceptual mechanism, probably involving a general auditory basis.

**Carrier**

A segment without a carrier has a larger detectability index than one with a carrier at the initial and medial positions. The effect of presence/absence of a carrier turned out to be significant by ANOVA. This disagrees with the result of Experiment 1. On the contrary, as in Experiment 1, no difference at the final position was observed at all tempi, although a carrier might provide a supplemental cue of change in segment duration.

**Original duration**

The Pearson’s product-moment correlation coefficients between the detectability indices and the original durations for each duration change are shown in Table 3.6. The negative correlation shows that the detectability index is greater when the original duration is shorter. Correlations between them for overall stimuli were observed at each duration change except -12 ms.

Since the variation in duration was combined in a factorial manner with the tempo variation in the current experiment, the difference due to the original duration could be investigated independently of the tempo change. The detectability indices pooled over
CHAPTER 3. PERCEPTUAL ACCEPTABILITY ANALYSIS OF SENTENCE SPEECH II:  
EFFECT OF SPEAKING RATE

Figure 3.5: Difference in detectability of change in duration due to stimulus duration and presentation tempo in Experiment 2. Different colors indicate different stimulus durations. Asterisks show pairs of consecutive bars whose differences are statistically significant \( p < 0.05 \).

duration changes for the original duration and tempo are shown in Figure 3.5. The effect of original duration between the same tempo was tested by a one-way ANOVA of repeated measures with duration changes as a blocking factor. The effects of original duration were significant at fast, normal and slow tempi \( F(1, 5) = 17.40, p < 0.01; F(1, 5) = 10.78, p < 0.05; F(1, 5) = 17.33, p < 0.01 \). On the other hand, the effects of tempo between the same durations, i.e., 75 ms and 112.5 ms, were not significant \( F(1, 5) = 0.27, p < 0.62; F(1, 5) = 0.35, p < 0.58 \). These results show that the original duration affected the detectability at a constant tempo. This agrees with Abel’s and Bochner et al.’s studies.

3.4 General discussion

The factors influencing acceptability of changes in segment duration were investigated, and then the equivalent factors influencing detectability using non-speech imitating speech stimuli were investigated to generalize those factors found in speech. The non-speech experiment cannot completely eliminate the influence of language factors because the lis-
teners were native speakers of a single language; however, it did minimize the influence of both speech-related and language-specific factors.

The results of both Experiment 1 and Experiment 2 demonstrate that the listeners were more sensitive to duration change when the speaking rate or tempo of a tonal sequence was faster. Hibi had reported the same effect of overall tempo for more simplified stimuli, i.e., empty intervals. These non-speech experiments showed that the effect of overall tempo on perception of a local segment was not speech-specific. This agreement between perceptual phenomena in speech and non-speech studies suggests that the effects of both speaking rate and the tempo of non-speech sounds share a common perceptual mechanism, probably involving a general auditory basis. Therefore, the effect must be independent of the variation in languages. That is, the effect of speaking rate can be observed in other languages.

The results of the two experiments also demonstrate that the listeners were most sensitive to duration change at the initial segment, least sensitive at the final one, and between the two levels at the medial one, irrespective of tempo or presence/absence of a carrier. As with the effect of tempo, the effect of position within a target part was not speech-specific. The non-speech specific result suggests that the unit within which the positional effect functions was not necessarily a linguistic unit such as a phrase but a certain perceptual unit in a broader sense. That is, the positional effect must also be language-independent. The positional effect on acceptability of change in segment duration can be observed in a minor phrase of other languages which have a temporal division such as initial, medial and final.

The effect of presence/absence of carrier and that of original duration were not common between speech and non-speech. The presence of carrier decreased sensitivity to duration change in non-speech, while it did not appear to function in speech. The reason for this might be that a phrase is a perceptually robust unit irrespective of acoustic separation. The original duration correlated with the index of listeners’ sensitivities in the non-speech experiment but this was not the case in the speech experiment if the speaking rate was kept constant. Kato et al. explained the relation between the original duration and sensitivity as follows. Kato et al.’s and Klatt et al.’s results using polysyllabic contexts showed no correlation between them, while Bochner et al.’s results using a monosyllabic context showed a correlation between them. The monosyllabic context might have facilitated the influence of the original vowel duration, while the polysyllabic context reduces or hides it. This explanation also seems to be applicable to the current study because the
temporal structure of the non-speech stimuli was, in a certain way, simpler than that of the polysyllabic speech stimuli.

Considering together the perceptual studies addressing change in segment duration, three factors have been found to affect the sensitivity to duration change: loudness properties, position within a phrase, and speaking rate [10]–[12][26]. Previous studies have shown the independence of the positional effect from the effects of original duration and loudness-related properties [26]. The current results suggest that the effect of speaking rate is independent of the positional effect and of the effect of loudness-related properties because all of the stimulus sounds in the current non-speech experiment had the same sound level. To adopt these perceptual characteristics of change in segment duration in the field of speech synthesis, the relations among the three effects should be further investigated quantitatively.

### 3.5 Conclusion

Aiming to improve the naturalness criteria for speech synthesis, the effect of speaking rate on the acceptability of distortion in segment duration was investigated. Furthermore, to generalize the effects found in speech, the detectability of duration change was investigated in non-speech imitating speech stimuli. Both the speech and non-speech experiments showed that (1) the acceptability degradation and detectability for the same amount of absolute change were greater when the speaking rate and the tempo were faster, and (2) the listeners were most sensitive to the same amount of absolute change at the initial position within a target portion, least sensitive at the final position and between these levels at the medial position at all speaking rates and tempi. These agreements between the speech and non-speech experiments suggest that the acceptability degradation can be accounted for by psychophysical factors, that is, the effects of speaking rate and position within a phrase were language-independent. These findings will serve as fundamental data for designing an objective measure of speech quality in speech synthesis.
Chapter 4

Relationship between speech generation and perception in speech timing control

4.1 Introduction

For contributing perceptual characteristics to segment duration control of text-to-speech or naturalness of segment duration control, previous studies have investigated the acceptability of change in segment duration [10]–[12][26]. The studies showed that sensitivity or acceptability of change in segment duration was significantly affected by contextual properties of the segment in question, e.g., temporal position within an utterance unit such as word or phrase.

The observed perceptual effects have been accounted for by the psychophysical characteristics of the speech material itself. For instance, a strong correlation has been found between these perceptual effects and psychophysical properties such as the loudness and position of the segment in question. However, perceptual characteristics are formed under the strong influence of the exposed environment during the process of growth. Thus, listener’s perceptual acceptability of a temporal change in a segment at a particular context may reflect the temporal deviation generally expected for a segment at that context in listener’s environment.

On the other hand, the utterance characteristics or strategy of a speaker may reflect the perceptual characteristics as a result of the feedback through speaker’s auditory system.

---

1This chapter is based on Muto, M., Kato, H., Tsuzaki, M., and Sagisaka, Y., “Relationship between duration control characteristics and perceptual characteristics of duration change in sentence speech,” ASJ, 2003.
However, it is still an open question whether such context-dependency of perceptual sensitivity is also consistent with the temporal properties of spoken utterances, i.e., whether the control precision has correlation with the perceptual sensitivity.

Investigating this possibility could yield useful information for understanding the nature of the perceptual factors found, whether it involves an ecological constraint during the development process or solely reflects inherent auditory characteristics. Such analysis could provide useful information not only for exploring the ecological influence on the perceptual factors but also for estimating the requisite precision of durational rules in speech synthesis or naturalness of utterances by a language learner.

For further understanding the nature of the perceptual characteristics, this chapter analyzed the context-dependency of the control precision by using large-scale speech corpora, and tried to explain perceptual characteristics by the context-dependent generation precision. First, Section 2 described the methodology of estimating the control precision. Then, Section 3 showed results of the data analysis and discussed the relationship between utterance and perceptual characteristics.

### 4.2 Measurement of utterance precision

Even the same phoneme (for example “a”), the duration is not constantly uttered at the same length. In general, the duration of the same phoneme vary at each utterance. The variation in segment duration could be caused by two factors: (1) contextual factor such as position of phoneme and preceding or following phoneme, (2) utterance precision factor. Since the current study aims at investigating the relation between perceptual characteristics and variation in segment duration due to the factor (2), we need to eliminate influences of factor (1). Therefore, we estimate a duration in a given context using a linear multiple regression model. Then, the prospective duration is subtracted from observed one to cancel out the factor (1). The residual error was regarded as representing the variation in segment duration due to the utterance precision. The standard deviation of the difference is referred to as control precision of utterance.

#### 4.2.1 Speech data

For speech data, phonetically balanced sentences were selected from the ATR Japanese speech database [36]. They have transcriptions of phonetical boundary, morpheme bound-
4.2. MEASUREMENT OF UTTERANCE PRECISION

Table 4.1: Multiple correlation coefficient and decrease of standard deviation (ms). F\(_x\) (\(x = 1,2,3,4\)) and M\(_y\) (\(y = 1,2,3,4,5,6\)) indicate female and male speaker, respectively.

<table>
<thead>
<tr>
<th>speaker ID</th>
<th>correlation coefficient</th>
<th>decrease of SD</th>
<th>speaker ID</th>
<th>correlation coefficient</th>
<th>decrease of SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F_1</td>
<td>0.734</td>
<td>8.12</td>
<td>M_2</td>
<td>0.866</td>
<td>14.72</td>
</tr>
<tr>
<td>F_2</td>
<td>0.818</td>
<td>11.41</td>
<td>M_3</td>
<td>0.861</td>
<td>15.31</td>
</tr>
<tr>
<td>F_3</td>
<td>0.816</td>
<td>12.98</td>
<td>M_4</td>
<td>0.811</td>
<td>11.95</td>
</tr>
<tr>
<td>F_4</td>
<td>0.772</td>
<td>8.71</td>
<td>M_5</td>
<td>0.856</td>
<td>17.18</td>
</tr>
<tr>
<td>M_1</td>
<td>0.807</td>
<td>11.06</td>
<td>M_6</td>
<td>0.800</td>
<td>9.55</td>
</tr>
</tbody>
</table>

ary, word class boundary, accent phrase boundary and type of accent. Speakers were 10 professional speakers (6 males and 4 females). Each speaker uttered 100 sentences. In total, 5030 sentences were used for the following analysis.

4.2.2 Elimination of control factors’ influence

The current study examined vowel durations because the control of vowel duration is in general more flexible than that of consonant duration. As control factors, the same parameters of previous study [4] and parameters of phrase properties were selected. They were (1) current phoneme category, (2) four neighboring phonemes (preceding, following, pre- and post-adjacent phonemes), (3) gemination of preceding and following consonants, (4) mora count of each utterance group (sentence, breath group, accent phrase, phrase, and word), (5) position (mora) in the utterance group, (6) with or without accent, and (7) word class. A linear multiple regression model was applied to each speaker. Table 4.1 shows the multiple correlation coefficients and decrease of standard deviation for each speaker. The multiple correlation coefficients which show the goodness of linear regression ranged 0.734–0.866. They were comparable with those in the previous study (0.769–0.863) [4]. The standard deviations of residual errors were lower than those of original durations. These show that the influences of contextual factors were successfully eliminated.
4.3 Relationship between utterance and perceptual characteristics of segment duration

Figure 4.1 shows the control precision of each utterance group. The control precision was worse at the final position of an utterance group than at initial or medial position. This tendency was commonly observed for all utterance groups.

Listeners were most sensitive to change in segment duration at phrase-initial position, least sensitive at phrase-final one, and between the two levels at phrase-medial one as shown in Chapter 2 and Chapter 3. On the other hand, the control precision was worse at the phrase-final position than phrase-initial or phrase-medial one. This tendency of the control precision at phrase-final is in good agreement with the perceptual characteristics at phrase-final. The perceptual effect was observed between initial and medial positions, while the effect on control precision was not observed between them.

4.4 Conclusion

For clues to a linkage between control precision and perceptual sensitivity, a comparison was performed between the variability of the segment durations in utterances and the perceptual evaluation of change in segment duration. The results showed that final temporal position in a phrase affected both utterance and perception in a similar way, while the difference in the initial and intermediate intra-unit positions is not common. Clearly, a linkage is not necessarily found in every aspect between perception and production. However, further studies are expected to determine the crucial conditions that are responsible for the existence of agreement between the control precision and perceptual sensitivity.
Figure 4.1: Vowel control precision represented by the standard deviation of the residual errors as a function of the temporal position in the following utterance units: (a) Word, (b) Phrase, (c) Accent phrase, (d) Breath group, and (e) Sentence. Each point represents the standard deviation of each of ten speakers. Triangles and error bars show the averages and standard deviations of the standard deviation, respectively. “initial” and “final” show vowel of first and final position in each utterance group, respectively. “medial” shows vowels of the other positions. Asterisks mark the levels whose averages are significantly different from those of the other levels [$p < 0.05$].
Chapter 5

Timing naturalness evaluation using objective measures for non-native speech

5.1 Introduction

In the study of foreign language learning, subjective evaluations by native speakers have been commonly used. An automatic evaluation of non-native speaker’s speech is expected to make self-study effective and considerably reduce teacher’s efforts. There are quite a few attempts to automatically evaluate English speech uttered by Japanese [37]–[40]. CALL systems have also been proposed for Japanese learning English [41]–[44]. Most of these efforts concentrate on segmental characteristics such as word pronunciation errors using technologies mainly from speech recognition fields. For prosodic-characteristics related evaluations, there exist some studies. Hamada et al. proposed an evaluation model of pronunciation using prosodic parameters [37]. Imoto et al. proposed a detection model of sentence stress [38]. Though these previous studies have used prosodic characteristics, they have not yet evaluated prosodic control directly.

Prosodic control is important to provide not only segmental information accurately but also language specific naturalness. In particular, the timing control is quite different between Japanese and English, and it is one of the biggest problems for learners. As

---

there is no clear objective criteria to specify their differences, it is not easy to learn how to change control strategy. We believe that speech synthesis technology will be able to contribute to specify the differences more directly and show more clear pictures to change timing control for learners. In this paper, as a first step towards objective evaluation of Japanese learner’s English speech, we tried to predict native’s subjective evaluation on naturalness in timing control using objective measures.

Timing control characteristics and statistical properties of segment durations have been studied in speech synthesis for many languages over many years. Recently, corpus-based modeling has successfully been applied to both prosody and segmental characteristics control for synthesizing high quality speech with natural prosody. Large-scale speech data collections and computational modeling have enabled quantitative controls with high precision using statistical optimization techniques through the minimization of objective errors [45]–[47]. Though some of current corpus-based modeling put too much emphasis on optimization and often neglect underlying scientific principles in speech generation, in principle, they can be applied to the quantification of subjective evaluation where so many underlying mechanisms and principles have not yet been known.

For the evaluation of non-native learner’s timing control, it is ideal to directly model their generation characteristics. However, it is quite difficult to create a control model and automatically estimate their control parameters. Non-native’s speech is produced as a result of blending timing controls of a target and learner’s native languages, and this blending is highly dependent on their language proficiency. Furthermore, beginner-specific characteristics such as slow speech rate and frequent pausing are also added to this blended timing control. To avoid these modeling complexities, we adopted a corpus-based approach to predict a native’s subjective score of naturalness in timing using duration differences of speech segments between learner’s and native speakers. In Section 5.2, timing control between English and Japanese and beginner’s characteristics were overviewed and then duration differences were listed to correlate with native’s evaluation scores. In Section 5.3, statistical analyses were carried out for all duration differences correlating with native’s subjective scores. In Section 5.4, both a linear multiple regression model and a regression tree model were generated using subjective scores and duration differences. These models were tested using open data. Finally, we concluded the results of correlation analyses and the experiments of automatic evaluation on naturalness in timing.
5.2 Timing control differences and correlation with naturalness evaluation

To objectively evaluate the naturalness of non-native’s speech, the understanding of underlying control differences is useful. In this section, we overview language-dependent timing control differences between English and Japanese and beginner’s characteristics from our own research experiences. Duration differences between natives and learners are listed from phone-sized units to a whole sentence to examine their correlations with subjective scores.

5.2.1 Timing control difference between English and Japanese

There are big differences in timing control between English and Japanese. In English, stressed syllables give timing regularities by reducing weak syllable length, which is referred to as stress-timed rhythm. While in Japanese, moras are uttered in almost equal length, which is referred to as mora-timed rhythm. Through the analyses of timing characteristics in Japanese speech, Japanese segment durations are found to be controlled by many factors of multiple levels from phoneme to sentence [2][4][13] [48][49].

These studies showed that Japanese segment durations were not only characterized by phoneme intrinsic attributes but also constrained by longer units. In particular, a moraic constraint is dominant in the control of Japanese vowel durations. A strong negative correlation is found between vowel durations and adjacent consonant durations. The timing compensation for the preceding consonant duration is more remarkable than the following one, and this is considered to be an acoustic manifestation of mora-timing. This timing compensation has been observed both in raw duration data and in normalized z-score data for each vowel [50]. Through the statistical analyses, it has been confirmed that the compensation takes place in mora units but not in syllable units. Additionally, it has also been observed that not only segment durations but pause length is under moraic control [51].

On the other hand, in English timing control, a syllable itself does not serve as an isochronous unit but a phrase-sized foot unit gives timing regularities [52]. At the same time, many factors affect a foot duration as same as Japanese. This timing difference is crucial for Japanese learners to master English timing control.

The other big factor in timing control is local timing preset [47]. In Japanese, though each moraic unit shows an almost equal timing interval, this timing is locally preset over a
phrase-sized interval consisting of several words. In Japanese timing control, this locally preset timing becomes shorter as the increase of the mora number in the interval [13]. As the number of syllables is also used in English timing control factor [50], there seems to exist the same kind of local control tendencies. However, the existence of other control factors in English such as syllable complexities make characteristics of segment duration more complex and thus shows differences.

5.2.2 Beginner’s timing control characteristics

Through the observation of Japanese learners’ speech, we have noticed that beginner’s speech has remarkable differences from English native’s one. Beginners speak slower than native speakers. They usually pronounce individual words with unnatural timing. Frequent pause insertions are found even at positions where native speakers do not insert it. We have often heard these beginner’s characteristics from English teachers, though we do not have quantitative precise analysis on beginner’s timing control characteristics. We would like to adopt some of these factors which can supply language-dependent timing control differences shown in the previous section.

5.2.3 Duration differences for predicting native’s subjective evaluation

As shown in the previous sections, language-dependent timing control characteristics and beginner’s timing control characteristics show the differences between learners and native speakers. They can provide good information for naturalness evaluation. For quantitative modeling of subjective evaluation, we have analyzed the correlation between native’s subjective evaluation scores and duration differences between Japanese learners and English natives. They consist of the following factors and corresponding duration differences.

(a) Language dependent factor

(a-1) Difference in function word duration

One of the most remarkable duration differences between Japanese and English can be found in function words. As function words consist of weak syllables, they are considerably reduced to keep timing regularities between stressed syllables in English. It is difficult for Japanese beginners to reduce them due to their custom of moraic timing. The
differences are expected to be clearly observed at vowel positions. To quantify the contribution of duration differences to subjective naturalness evaluation, we analyzed duration differences in vowel durations in the next section. For control experiment, we also analyzed other duration differences such as function words, content words, and consonants both in content words and in function words.

(a-2) Difference in stressed syllable duration

It is known that strong syllable has not only greater loudness but also longer duration [53]–[55]. Since, in Japanese, syllables (moras) are uttered in almost equal length, it is difficult for Japanese learners to control durations of strong syllables. The duration difference in strong syllable is expected to be informative cue of a moraic constraint. To quantify the moraic constraint to subjective naturalness evaluation, duration differences of strong syllables were examined. For analysis in detail, we chose strong vowels, primary strong syllables and vowels, strong syllables excluding primary strong syllables (non-primary strong syllables), strong vowels excluding primary strong vowels (non-primary strong vowels), and weak syllable.

(a-3) Difference in closed syllable duration

Dauer showed that closed syllable accounted for 51% of the syllable structure in English [56]. As most of Japanese syllables are open syllables, Japanese beginners are not accustomed to produce closed syllables. Even an extra-vowel is inserted after the final consonant. The duration difference of closed syllable can also be effective to evaluate naturalness in timing control. To quantify vowel insertions from the aspect of duration, duration differences of closed syllables were examined. Furthermore, duration differences of the latter consonant of closed syllable for detailed analysis and open syllable for comparison were also examined.

(b) Beginner’s control factor

(b-1) Difference in whole sentence duration

Usually beginners cannot speak fast. Non-native beginners tend to utter slower than native speakers. The overall speech rate can be measured by total sentence duration.

(b-2) Difference in word duration

By equalizing phrase break points between speakers, we can compare the change of local speech rate by their constituent units. The duration differences of words can be used to
capture local tempo changes that are buried in overall sentence differences. The sum of word duration differences will be an index of local tempo change characteristics.

(b-3) Difference in pause duration

Since beginners tend to pronounce word by word, a total pause duration of learners can be longer than those of natives. The differences in pause durations are considered to be informative beginner’s timing characteristics.

5.3 Analysis on correlations between native’s subjective evaluations and duration differences

For objective evaluation of Japanese English speech on naturalness in timing, we need to know what duration differences show statistically good correlates with a subjective measure. We measured correlations between native’s subjective evaluations and duration differences listed in the previous section.

5.3.1 Speech data

Speech samples were collected to measure correlations between English native’s subjective evaluations and duration differences between native English speakers and Japanese English learners. To directly measure learner’s prosody control capabilities, we have given phonetic information and prosodic symbols indicating intonation, phrase boundaries, and pause locations to the text. Japanese learners were requested to practice pronunciations of given sentences before recording. In recording, the learners were asked to read repeatedly until they could speak what they thought was the correct pronunciation. During their practices and recordings, speech samples uttered by native had not been presented as reference. Most of Japanese learners were university students. As native reference speech samples, we collected the same ones uttered by English instructors who had long experience in teaching English to Japanese learners.

For English texts, 100 sentences were selected from daily conversation textbooks. These sentences consist of from 3 to 18 words as shown in Table 5.1. These sentences were uttered by every twelve Japanese learners consisting of seven males and five females and three native speakers consisting of two males and one female. In total, 1500 speech samples were recorded.
5.3. ANALYSIS ON CORRELATIONS BETWEEN NATIVE’S SUBJECTIVE EVALUATIONS AND DURATION DIFFERENCES

Table 5.1: Examples of sentences used in the in the correlation analyses.

<table>
<thead>
<tr>
<th></th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>He is a good player, isn’t he?</td>
</tr>
<tr>
<td>2</td>
<td>I don’t want you to say that.</td>
</tr>
<tr>
<td>3</td>
<td>I’m on my way to your hotel, but there was an accident, and I’m stuck in the train now.</td>
</tr>
<tr>
<td>4</td>
<td>Is it possible for me to come to your office to explain our new products?</td>
</tr>
<tr>
<td>5</td>
<td>Take the elevator or escalator to the third floor.</td>
</tr>
<tr>
<td>6</td>
<td>There’re 80, 60, 40 and 25 dollar seats.</td>
</tr>
<tr>
<td>7</td>
<td>This is it.</td>
</tr>
<tr>
<td>8</td>
<td>Why don’t we meet at the coffee shop in the station building at Shinagawa?</td>
</tr>
<tr>
<td>9</td>
<td>Would you like me to invite your friend for dinner this weekend?</td>
</tr>
<tr>
<td>10</td>
<td>Your husband often goes abroad, right?</td>
</tr>
</tbody>
</table>

Naturalness of Japanese learner’s speech was evaluated with respect to timing by five native raters having long experience in teaching English to Japanese. The raters were asked to evaluate naturalness in terms of duration difference, using a five-point rating scale from “1”(very poor) to “5”(excellent). Since the current study aims at measuring correlations between duration differences and subjective naturalness scores, we eliminated speech samples with pronunciation errors from the analysis. Finally we selected 446 speech samples for the following analysis. They consist of 57 sentences having at least six speech samples. One learner uttered at least seventeen sentences, but not more than fifty sentences. The average score and the standard deviation of those sample scores were 4.21 and 0.78, respectively.
5.3.2 Correlations between duration differences and subjective scores

Subjective scores on naturalness in timing were pooled over five natives’ scores for each learner’s utterance. Hereafter, this value is referred to as a *subjective naturalness score*. Each native’s duration was subtracted from each non-native learner’s one. This value is referred to as a *duration difference*. Correlation coefficients between duration differences and corresponding subjective naturalness score were calculated. Averages and standard deviations of duration differences, correlation coefficients between duration differences and the subjective naturalness scores are shown in Table 5.2. A positive average value shows that the corresponding duration of learners is longer than that of natives. A negative correlation value shows that the subjective naturalness score increases when the corresponding duration difference decreases.

From these results, the following correlation characteristics are observed in language dependent duration differences and differences resulting from beginner’s timing control characteristics.

(a-1) Difference in function word duration

Averages of differences showed that durations of function words uttered by the learners tended to be longer than those uttered by the native speakers. Negative correlations were observed between duration differences in function words and subjective naturalness scores. The correlations of function word were comparable to those of content words. By analyzing difference characteristics in detail, correlations of vowels in function words were found to be greater than those of vowels in content words and those of consonant in function and content words. These correlations suggest that natives evaluated the temporal naturalness with particular attention to the vowel duration in function word and that it can express the proficiency of stress-timing control.

(a-2) Difference in stressed syllable duration

Strong syllables have greater negative correlations than weak syllables. By analyzing strong syllables in detail, it is found that strong vowels with a primary stress tend to be uttered slightly longer by the native speakers than by learners. Correlations of strong syllable differences and strong vowel differences with a primary stress were much smaller than those of all strong syllable differences and strong vowels differences, respectively. These tendencies of strong syllable and strong vowel with primary stress suggest that learners can speak them as natives do and that the factors did not affect the subjective
Table 5.2: Average and standard deviation of duration difference (ms), and correlation between subjective evaluation score and duration differences reflecting language differences (a-1)–(a-3) and beginner’s control (b-1)–(b–3). Underlined parameters were used as factor of statistical models in Section 5.4.

<table>
<thead>
<tr>
<th>timing control differences</th>
<th>native_1</th>
<th></th>
<th>native_2</th>
<th></th>
<th>native_3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>average</td>
<td>SD</td>
<td>correlation</td>
<td>average</td>
<td>SD</td>
<td>correlation</td>
</tr>
<tr>
<td>(a-1) Difference in function word</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>function word</td>
<td>37.9</td>
<td>69.2</td>
<td>-0.27</td>
<td>36.3</td>
<td>52.2</td>
<td>-0.39</td>
</tr>
<tr>
<td>content word</td>
<td>47.7</td>
<td>70.1</td>
<td>-0.35</td>
<td>28.3</td>
<td>66.3</td>
<td>-0.37</td>
</tr>
<tr>
<td>vowel in function word</td>
<td>17.0</td>
<td>34.8</td>
<td>-0.38</td>
<td>20.8</td>
<td>33.4</td>
<td>-0.43</td>
</tr>
<tr>
<td>consonant in function word</td>
<td>14.9</td>
<td>43.3</td>
<td>-0.10</td>
<td>10.3</td>
<td>26.6</td>
<td>-0.16</td>
</tr>
<tr>
<td>vowel in content word</td>
<td>6.4</td>
<td>31.8</td>
<td>-0.29</td>
<td>4.5</td>
<td>31.1</td>
<td>-0.29</td>
</tr>
<tr>
<td>consonant in content word</td>
<td>14.8</td>
<td>18.2</td>
<td>-0.22</td>
<td>8.0</td>
<td>16.2</td>
<td>-0.24</td>
</tr>
<tr>
<td>(a-2) Difference in stressed syllable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strong syllable</td>
<td>31.4</td>
<td>54.7</td>
<td>-0.33</td>
<td>19.5</td>
<td>41.2</td>
<td>-0.44</td>
</tr>
<tr>
<td>strong vowel</td>
<td>9.8</td>
<td>28.7</td>
<td>-0.36</td>
<td>7.3</td>
<td>26.1</td>
<td>-0.41</td>
</tr>
<tr>
<td>primary strong syllable</td>
<td>18.8</td>
<td>77.5</td>
<td>-0.10</td>
<td>0.7</td>
<td>67.4</td>
<td>-0.15</td>
</tr>
<tr>
<td>primary strong vowel</td>
<td>-6.4</td>
<td>45.4</td>
<td>-0.05</td>
<td>-8.1</td>
<td>38.5</td>
<td>-0.10</td>
</tr>
<tr>
<td>non-primary strong syllable</td>
<td>35.3</td>
<td>64.0</td>
<td>-0.33</td>
<td>26.8</td>
<td>42.4</td>
<td>-0.47</td>
</tr>
<tr>
<td>non-primary strong vowel</td>
<td>15.2</td>
<td>30.6</td>
<td>-0.41</td>
<td>13.0</td>
<td>28.9</td>
<td>-0.43</td>
</tr>
<tr>
<td>weak syllable</td>
<td>37.4</td>
<td>52.9</td>
<td>-0.25</td>
<td>34.7</td>
<td>49.7</td>
<td>-0.27</td>
</tr>
<tr>
<td>weak vowel</td>
<td>10.8</td>
<td>32.9</td>
<td>-0.30</td>
<td>17.4</td>
<td>32.8</td>
<td>-0.29</td>
</tr>
<tr>
<td>(a-3) Difference in closed syllable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>closed syllable</td>
<td>41.0</td>
<td>61.3</td>
<td>-0.37</td>
<td>24.6</td>
<td>49.5</td>
<td>-0.40</td>
</tr>
<tr>
<td>open syllable</td>
<td>27.5</td>
<td>43.7</td>
<td>-0.33</td>
<td>26.3</td>
<td>37.8</td>
<td>-0.38</td>
</tr>
<tr>
<td>consonant in closed syllable</td>
<td>21.1</td>
<td>44.5</td>
<td>-0.16</td>
<td>10.8</td>
<td>26.5</td>
<td>-0.25</td>
</tr>
<tr>
<td>(b-1) Difference in whole sentence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sentence</td>
<td>593.1</td>
<td>773.3</td>
<td>-0.39</td>
<td>448.0</td>
<td>685.4</td>
<td>-0.43</td>
</tr>
<tr>
<td>(b-2) Difference in word</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>word</td>
<td>44.2</td>
<td>62.2</td>
<td>-0.36</td>
<td>31.7</td>
<td>49.5</td>
<td>-0.46</td>
</tr>
<tr>
<td>(b-3) Difference in pause</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pause</td>
<td>28.0</td>
<td>36.6</td>
<td>-0.29</td>
<td>24.4</td>
<td>37.1</td>
<td>-0.28</td>
</tr>
<tr>
<td>(Reference)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>syllable</td>
<td>33.5</td>
<td>48.7</td>
<td>-0.35</td>
<td>23.3</td>
<td>38.0</td>
<td>-0.46</td>
</tr>
<tr>
<td>vowel</td>
<td>10.7</td>
<td>26.2</td>
<td>-0.41</td>
<td>9.9</td>
<td>25.0</td>
<td>-0.43</td>
</tr>
<tr>
<td>consonant</td>
<td>14.7</td>
<td>21.8</td>
<td>-0.19</td>
<td>8.7</td>
<td>13.5</td>
<td>-0.31</td>
</tr>
</tbody>
</table>
naturalness score. The duration differences in strong syllables excluding primary stressed syllables (non-primary strong syllable) are considered to be effective to predict the proficiency in stress-timing control.

**(a-3) Difference in closed syllable duration**

To measure the vowel insertion, durations of open and closed syllables were examined. It was expected that duration difference of closed syllable between native speaker and non-native learner were greater than that of open syllable because non-native learner inserts a vowel at the end of closed syllable. However, consistent tendencies due to the syllable structure were not observed in the duration difference and correlation. By analyzing in detail, the end consonants of closed syllable were also examined. An influence of vowel insertion was not observed compared with the overall consonant. The contribution of syllable structure to the subjective naturalness score might be smaller than other timing control differences.

**(b-1) Difference in whole sentence duration**

As qualitatively speculated, the averages of duration differences in sentences showed that non-natives’ speech tended to be longer than natives’ one. Negative correlations between duration differences of sentences and the subjective naturalness scores were observed for all native references. This result showed that the subjective naturalness score decreased as learners spoke slower.

**(b-2) Difference in word duration**

The averages of duration differences in words showed that non-natives’ speech tended to be longer than native_1’s and native_2’s one, while native_3’s speech was almost the same as non-natives’ one. Though there is difference of averages of duration differences, negative correlations between duration differences of individual words and subjective naturalness scores were observed for all native references. This result showed that the subjective naturalness score decreased as the word duration difference increased.

**(b-3) Difference in pause duration**

High correlations were not observed between the duration differences of pauses lengths and the subjective rating scores. Pre-assignment of prosodic symbols such as pause locations and phrase boundaries might have reduced this correlation.
5.4 Temporal naturalness prediction of non-native speech using objective parameters

5.4.1 Modeling using objective parameters

Statistical models were generated for automatic naturalness evaluation of English utterance by Japanese in timing control. Linear multiple regression and Regression tree were used to compute naturalness evaluation scores. For control factors of these models, we adopted the duration differences showing reasonable correlations with the subjective naturalness scores. As shown factors underlined in Table 5.2, they consist of vowel duration of function word, non-primary strong syllable, sentence duration, word duration, and pause duration.

Table 5.3: Root mean square between subjective and predicted rating score. “Linear” and “Tree” indicate the case of linear multiple regression model and regression tree model, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>Linear closed</th>
<th>Linear open</th>
<th>Tree closed</th>
<th>Tree open</th>
</tr>
</thead>
<tbody>
<tr>
<td>native_1</td>
<td>0.39</td>
<td>0.52</td>
<td>0.36</td>
<td>0.56</td>
</tr>
<tr>
<td>native_2</td>
<td>0.40</td>
<td>0.52</td>
<td>0.37</td>
<td>0.58</td>
</tr>
<tr>
<td>native_3</td>
<td>0.39</td>
<td>0.51</td>
<td>0.38</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Table 5.4: Correlation coefficient between subjective and predicted rating score. “Linear” and “Tree” indicate the case of linear multiple regression model and regression tree model, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>Linear closed</th>
<th>Linear open</th>
<th>Tree closed</th>
<th>Tree open</th>
</tr>
</thead>
<tbody>
<tr>
<td>native_1</td>
<td>0.48</td>
<td>0.45</td>
<td>0.58</td>
<td>0.42</td>
</tr>
<tr>
<td>native_2</td>
<td>0.44</td>
<td>0.44</td>
<td>0.57</td>
<td>0.30</td>
</tr>
<tr>
<td>native_3</td>
<td>0.50</td>
<td>0.43</td>
<td>0.54</td>
<td>0.23</td>
</tr>
</tbody>
</table>
For open tests, samples of three learners with different skills were selected from twelve learners. Their language skills had been judged by the native speakers. The numbers of open tests set were 17 (low skilled learner), 24 (middle skilled learner), and 43 (high skilled learner). The rest of samples were used for closed test, i.e., the numbers of those were 429, 422, and 403. In the following performance evaluation, the average values of open and closed tests of these three learners were employed. The stop condition of the regression tree was set to 1.0% of samples for closed test.

Table 5.3 shows the root mean square errors for each native’s reference and both test set. The root mean square errors by linear multiple regression models turned out to be 0.39–0.40 for the closed test set and 0.51–0.52 for the open test set. On the other hand, the root mean square errors by regression tree models turned out to be 0.36–0.38 for the closed test set and 0.56–0.60 for the open test set. These errors were turned out to be smaller than the standard deviation of scores among five raters 0.67 for each sample. This result shows the possibility of subjective score prediction by the proposed evaluation models.

To examine the error characteristics in detail, the correlation coefficients were calculated between predicted rating scores and their subjective rating scores for each native’s reference and both test set as shown in Table 5.4. The correlations between them by linear multiple regression models turned out to be 0.44–0.50 for the closed test set and 0.43–0.45 for the open test set. On the other hand, the correlations between them by the regression tree models turned out to be 0.54–0.58 for the closed test set and 0.23–0.42 for the open test set. These values were almost equal to the average of correlations between human ratings 0.22 for each sentence and 0.39 for each learner. Though the correlations of models were not so high, the proposed models show similar scoring characteristics by human. In the above performance evaluation, big differences were not observed between two models.

### 5.4.2 Modeling using perceptual weighted parameters

A series of studies has been done to explore relationships between the physical and perceptual measures of prosodic features in speech, especially on temporal features. Previous studies showed that a segment with greater loudness or with a larger loudness jump was more susceptible to segmental duration distortion [10]–[12]. Chapter 2 and Chapter 3 showed that listeners were most sensitive to duration change at the initial segment, least sensitive at the final one, and between the two levels at the medial one. This section
presents approaches to reflect those human perceptual characteristics onto the objective evaluation of speech prosody.

Linear multiple regression and Regression tree were also used to compute naturalness evaluation scores. For control factors of these models, we adopted the duration differences showing reasonable correlations with the subjective naturalness scores, i.e., vowel duration of function word, non-primary strong syllable, sentence duration, word duration, and pause duration. We did perceptual weighting to these factors by the loudness and position.

We first calculated the loudness contour for each stimulus every 2.5 ms with a 20-ms rectangular window. This is done according to the ISO 532 method B [21] while assuming a diffuse field and using Zwicker et al.’s algorithm [22]. Then, we picked out the median value from the entire loudness contour of each segment as the representative loudness of the segment. In “Loudness model”, we calculated the maximum of loudness in each factor: vowel duration of function word, non-primary strong syllable, and word duration. Then, the maximum of loudness was multiplied by each duration difference. In “Loudness+Position model”, we calculated the sum of loudness in each factor: vowel duration of function word, non-primary strong syllable, sentence, and word duration. Then, loudness of initial vowel in polysyllable was exponentiated.

For closed and open test, same samples of Section 5.4 were selected. In the following performance evaluation, the average values of open and closed tests of three native references were employed. The stop condition of the regression tree was set to 1.0 % of samples for closed test.

Figure 5.1–5.4 show the averages of the root mean square errors and the correlation coefficients between predicted rating scores and subjective rating scores for three native’s reference. At best prediction for training test judging from correlation coefficient, the root mean square errors by linear multiple regression turned out to be 0.51 in “Loudness model” and 0.49 in “Loudness+Position model” for open test set. The root mean square errors by tree regression turned out to be 0.51 in “Loudness model” and 0.51 in “Loudness+Position model”. Those of model in Section 5.4 by linear multiple regression and tree regression were 0.52 and 0.58 respectively. On the other hand, the correlations coefficient by linear multiple regression turned out to be 0.42 in “Loudness model” and 0.37 in “Loudness+Position model” for open test set. The correlations coefficient by tree regression turned out to be 0.42 in “Loudness model” and 0.42 in “Loudness+Position model” for open test set. Those of model in Section 5.4 by linear multiple regression and
Comparing the model in Section 5.4, little improvements were observed in the root mean square errors. However, in the correlation coefficients of linear multiple regression, no improvement were observed. Since the effects of loudness or position on evaluation were observed in local area of speech, those perceptual characteristics might be unable to apply to the current evaluations of overall timing control.

5.5 Conclusions

Aiming at automatic naturalness evaluation of non-native utterance with objective measures, we analyzed the relations between the timing control differences from native speakers’ utterances and the subjective scores given by native speakers. Negative correlations between subjective naturalness scores and the duration differences were observed in vowel duration of function word, strong syllable duration excluding primary stressed syllable, sentence duration, word duration, and pause duration. These parameters correlated with the natives subjective score in temporal naturalness were found, which express the language dependent rhythmic difference and beginner’s timing control tendencies. Then, using those five parameters, we predicted subjective rating scores by the linear multiple regression model and the regression tree model. The prediction accuracy test for open data showed that root mean squares of these models were smaller than errors among native raters. The estimation accuracy test for open data showed the effectiveness of this corpus-based modeling on subjective naturalness evaluation without modeling generation characteristics directly.
Figure 5.1: Root mean square and correlation coefficient between subjective rating scores and predicted rating scores calculated by linear multiple regression using perceptual weighting parameters by the loudness. “org” indicates the model in Section 5.4.
Figure 5.2: Root mean square and correlation coefficient between subjective rating scores and predicted rating scores calculated by tree regression using perceptual weighting parameters by the loudness. “org” indicates the model in Section 5.4.
Figure 5.3: Root mean square and correlation coefficient between subjective rating scores and predicted rating scores calculated by linear multiple regression using perceptual weighting parameters by the loudness and the position. “org” indicates the model in Section 5.4.
Figure 5.4: Root mean square and correlation coefficient between subjective rating scores and predicted rating scores calculated by tree regression using perceptual weighting parameters by the loudness and the position. “org” indicates the model in Section 5.4.
Chapter 6

Conclusions

6.1 Summary of the dissertation

In this dissertation, we have analyzed the perceptual characteristics and naturalness evaluation on timing control in sentence speech for speech synthesis and CALL systems. First, we carried out speech perception experiments to explore how factors influence the human perceptual sensitivity or acceptability of temporal distortions in sentence speech. Chapter 2 and 3 revealed that a duration distortion in the phrase-initial segment was the least acceptable by listeners and that in the phrase-final segment was the most acceptable, with that in the phrase-medial segment in between (intra-phrase positional effect). Chapter 2 showed that the intra-phrase positional effect was irrespective of variations in the attribute of a target phrase and variations in the context of a target phrase. The contextual differences consist of length of the phrase, accent type of the phrase, with or without a carrier sentence, difference in the carrier sentence, and position in a breath group (initial or non-initial). Chapter 3 showed that the acceptability degradation of change in segment duration was greater when the speaking rate was faster and intra-phrase positional effect was irrespective of speaking rates. These findings will serve as fundamental data for designing an objective measure of speech quality in speech synthesis or language learning.

Furthermore, in chapter 3, to generalize the effects found in speech, the detectability of duration change was also investigated using non-speech imitating speech stimuli. Both the speech and non-speech experiments showed that the acceptability degradation and detectability for the same amount of absolute change were greater when the speaking rate and the tempo were faster, and the listeners were most sensitive to the same amount of absolute change at the initial position within a target portion, least sensitive at the
final position and between these levels at the medial position at all speaking rates and tempi. The non-speech experiment cannot completely eliminate the influence of language factors because the listeners were native speakers of a single language; however, it did minimize the influence of both speech-related and language-specific factors. That is, these agreements between the speech and non-speech experiments suggest that the acceptability degradation can be accounted for by psychophysical factor. A similar positional and tempo effect might be observed across languages within a certain unit or group.

In Chapter 4, for further understanding the nature of the perceptual factors found in Chapter 2 and 3, we measured the correlation between segmental control precision of utterance and perceptual sensitivity of segment duration change. We estimated a duration in a given context using a linear multiple regression model, and then regarded the standard deviation of residual error as segmental control precision. The segmental control precision was worse at the phrase-final position than phrase-initial or phrase-medial one. This control tendency was in good agreement with the perceptual characteristics at phrase-final, except at phrase-initial and -medial. The difference of perceptual effect was observed between initial and medial positions, while the difference of control precision was not observed between them. On the other hand, Chapter 3 showed that the difference of perceptual effect within a phrase was good agreement with the psychophysical factor. Although there remains the possibility that the control precision is linked to the perceptual characteristics, the psychophysical factor could account for the intra-phrase positional effect better than the control precision.

In Chapter 5, as a first step toward the objective measurement to evaluate the subjective evaluation by human, we made the evaluation model of temporal naturalness in Japanese English using multiple durations of speech. First, we analyzed the relations between the timing control differences from native speakers’ utterances and the subjective scores given by native speakers. Negative correlations between subjective naturalness scores and the duration differences were observed in vowel duration of function word, strong syllable duration excluding primary stressed syllable, sentence duration, word duration, and pause duration. Those correlations showed that language specific timing control difference could be captured by objective acoustic measures. Using those five objective measures, we predicted subjective rating scores by the linear multiple regression model and the regression tree model. The prediction accuracy test for open data showed that the effectiveness of this corpus-based modeling on subjective naturalness evaluation without modeling generation characteristics directly. Furthermore, to integrate perceptual charac-
teristics found in previous studies and in Chapter 2 and Chapter 3 into the proposed model, we also proposed the experimental model. The comparison between native’s subjective evaluation scores and predicted scores supported the effectiveness of using perceptual findings.

6.2 Future work

The non-speech experiment in Chapter 3 showed that the intra-phrase positional effect was good agreement with the psychophysical factor. This finding suggested that a similar positional effect could be observed within a certain unit in other language. To confirm the generality of positional effect on sensitivity to segmental duration distortion, the same speech experiment using other language and non-speech experiment evaluated by native speaker of other language should be conducted.

Putting together the findings in the perceptual characteristics of segmental duration distortion, three factors have been found to affect the sensitivity to duration change: loudness properties, position within a phrase, and speaking rate. Previous studies have shown the independence of the positional effect from the effects of original duration and loudness-related properties [26]. The current results suggest that the effect of speaking rate is independent of the positional effect and the effect of loudness-related properties as all of the stimulus sounds in the non-speech experiment of Chapter 3 had the same sound level. To make a model of perceptual characteristics of segmental duration distortion integrating findings in this dissertation and the previous studies, quantitatively investigating the relations among the three effects is needed.

Another important direction of future research is to investigate further acoustic features related to naturalness evaluation of prosody. Although proposed model was able to cope with aspect of timing control, that was applicable to limited aspect of prosody. Analyses of intonation or stress by power promote the achievement of a more advanced and realistic model of naturalness in prosody.
Appendix A

Properties and correlation coefficients of tested speech portion

This chapter shows duration properties of tested speech portion in Chapter 2. This chapter also shows correlation coefficients between vulnerability index and the original duration, loudness, and loudness jump.
Table A.1: Durations of four-mora phrases and vowels of Experiment 1 (ms).

<table>
<thead>
<tr>
<th>phrase</th>
<th>1st vowel</th>
<th>2nd vowel</th>
<th>3rd vowel</th>
<th>4th vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>dekireba</td>
<td>510</td>
<td>84</td>
<td>48</td>
<td>88</td>
</tr>
<tr>
<td>kodomoga</td>
<td>552</td>
<td>72</td>
<td>98</td>
<td>85</td>
</tr>
<tr>
<td>tonikaku</td>
<td>553</td>
<td>83</td>
<td>81</td>
<td>99</td>
</tr>
</tbody>
</table>

Table A.2: Durations of five-mora phrases and vowels of Experiment 1 (ms).

<table>
<thead>
<tr>
<th>phrase</th>
<th>1st vowel</th>
<th>2nd vowel</th>
<th>3rd vowel</th>
<th>4th vowel</th>
<th>5th vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>dokomademo</td>
<td>671</td>
<td>103</td>
<td>72</td>
<td>90</td>
<td>71</td>
</tr>
<tr>
<td>rokujikara</td>
<td>641</td>
<td>78</td>
<td>48</td>
<td>61</td>
<td>80</td>
</tr>
<tr>
<td>tomodachiga</td>
<td>642</td>
<td>80</td>
<td>76</td>
<td>103</td>
<td>52</td>
</tr>
</tbody>
</table>

Table A.3: Durations of three-mora phrases and vowels of Experiment 1 (ms).

<table>
<thead>
<tr>
<th>phrase</th>
<th>1st vowel</th>
<th>2nd vowel</th>
<th>3rd vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>haruga (a)</td>
<td>445</td>
<td>83</td>
<td>104</td>
</tr>
<tr>
<td>(b)</td>
<td>436</td>
<td>78</td>
<td>68</td>
</tr>
<tr>
<td>(c)</td>
<td>437</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>(d)</td>
<td>535</td>
<td>88</td>
<td>105</td>
</tr>
<tr>
<td>kuruto (a)</td>
<td>503</td>
<td>49</td>
<td>94</td>
</tr>
<tr>
<td>(b)</td>
<td>531</td>
<td>64</td>
<td>88</td>
</tr>
<tr>
<td>(c)</td>
<td>444</td>
<td>56</td>
<td>96</td>
</tr>
<tr>
<td>(d)</td>
<td>542</td>
<td>72</td>
<td>124</td>
</tr>
<tr>
<td>nodemo (a)</td>
<td>427</td>
<td>81</td>
<td>75</td>
</tr>
<tr>
<td>(b)</td>
<td>407</td>
<td>74</td>
<td>97</td>
</tr>
<tr>
<td>(c)</td>
<td>388</td>
<td>92</td>
<td>85</td>
</tr>
<tr>
<td>(d)</td>
<td>505</td>
<td>99</td>
<td>115</td>
</tr>
</tbody>
</table>
Table A.4: Standard deviations of vowel durations in four-mora phrases of Experiment 1 (ms).

<table>
<thead>
<tr>
<th>phrase</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
</tr>
</thead>
<tbody>
<tr>
<td>dekireba</td>
<td>0.30</td>
<td>0.37</td>
<td>0.27</td>
<td>0.34</td>
</tr>
<tr>
<td>kodomoga</td>
<td>0.36</td>
<td>0.54</td>
<td>0.45</td>
<td>0.17</td>
</tr>
<tr>
<td>tonikaku</td>
<td>0.35</td>
<td>0.48</td>
<td>0.39</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table A.5: Standard deviations of vowel durations in five-mora phrases of Experiment 1 (ms).

<table>
<thead>
<tr>
<th>phrase</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>dokomademo</td>
<td>0.44</td>
<td>0.39</td>
<td>0.40</td>
<td>0.25</td>
<td>0.37</td>
</tr>
<tr>
<td>rokujikara</td>
<td>0.47</td>
<td>0.28</td>
<td>0.39</td>
<td>0.23</td>
<td>0.21</td>
</tr>
<tr>
<td>tomodachiga</td>
<td>0.42</td>
<td>0.34</td>
<td>0.40</td>
<td>0.38</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Table A.6: Standard deviations of vowel durations in three-mora phrases of Experiment 2 (ms).

<table>
<thead>
<tr>
<th></th>
<th>1st vowel</th>
<th>2nd vowel</th>
<th>3rd vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>haruga (a)</td>
<td>0.65</td>
<td>0.49</td>
<td>0.32</td>
</tr>
<tr>
<td>(b)</td>
<td>0.53</td>
<td>0.65</td>
<td>0.33</td>
</tr>
<tr>
<td>(c)</td>
<td>0.61</td>
<td>0.46</td>
<td>0.29</td>
</tr>
<tr>
<td>(d)</td>
<td>0.61</td>
<td>0.22</td>
<td>0.07</td>
</tr>
<tr>
<td>kuruto (a)</td>
<td>0.52</td>
<td>0.42</td>
<td>0.11</td>
</tr>
<tr>
<td>(b)</td>
<td>0.44</td>
<td>0.39</td>
<td>0.19</td>
</tr>
<tr>
<td>(c)</td>
<td>0.66</td>
<td>0.64</td>
<td>0.21</td>
</tr>
<tr>
<td>(d)</td>
<td>0.54</td>
<td>0.39</td>
<td>0.20</td>
</tr>
<tr>
<td>nodemo (a)</td>
<td>0.29</td>
<td>0.32</td>
<td>0.06</td>
</tr>
<tr>
<td>(b)</td>
<td>0.37</td>
<td>0.33</td>
<td>0.15</td>
</tr>
<tr>
<td>(c)</td>
<td>0.29</td>
<td>0.33</td>
<td>0.12</td>
</tr>
<tr>
<td>(d)</td>
<td>0.51</td>
<td>0.26</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table A.7: Standard deviations of vowel durations in three-mora phrases of Experiment 3 (ms).

<table>
<thead>
<tr>
<th></th>
<th>initial</th>
<th>medial</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st mora</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>2nd mora</td>
<td>0.36</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Table A.8: Loudnesses (left) and loudness jumps (right) of vowels in four-mora phrases of Experiment 1 (sone).

<table>
<thead>
<tr>
<th></th>
<th>1st vowel</th>
<th>2nd vowel</th>
<th>3rd vowel</th>
<th>4th vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>dekireba</td>
<td>6.1 5.1</td>
<td>5.2 4.8</td>
<td>6.2 3.8</td>
<td>5.3 4.9</td>
</tr>
<tr>
<td>kodomoga</td>
<td>2.9 2.8</td>
<td>4.4 4.1</td>
<td>4.1 1.4</td>
<td>6.5 5.4</td>
</tr>
<tr>
<td>tonikaku</td>
<td>5.5 5.4</td>
<td>5.3 2.3</td>
<td>4.8 3.9</td>
<td>1.7 1.6</td>
</tr>
</tbody>
</table>

Table A.9: Loudnesses (left) and loudness jumps (right) of vowels in five-mora phrases of Experiment 1 (sone).

<table>
<thead>
<tr>
<th></th>
<th>1st vowel</th>
<th>2nd vowel</th>
<th>3rd vowel</th>
<th>4th vowel</th>
<th>5th vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>dokomademo</td>
<td>5.7 4.9</td>
<td>5.3 5.0</td>
<td>5.6 2.4</td>
<td>4.0 3.3</td>
<td>2.7 1.3</td>
</tr>
<tr>
<td>rokujikara</td>
<td>6.5 2.8</td>
<td>5.5 5.2</td>
<td>4.1 1.9</td>
<td>5.5 5.4</td>
<td>4.4 0.3</td>
</tr>
<tr>
<td>tomodachiga</td>
<td>4.4 4.1</td>
<td>5.4 3.2</td>
<td>6.9 5.6</td>
<td>3.6 2.8</td>
<td>5.9 4.7</td>
</tr>
</tbody>
</table>

Table A.10: Loudnesses (left) and loudness jumps (right) of vowels in three-mora phrases of Experiment 2 (sone).

<table>
<thead>
<tr>
<th></th>
<th>1st vowel</th>
<th>2nd vowel</th>
<th>3rd vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>haruga (a)</td>
<td>14.7 14.7</td>
<td>7.2 3.0</td>
<td>4.8 1.2</td>
</tr>
<tr>
<td>(b)</td>
<td>14.5 14.5</td>
<td>8.2 2.9</td>
<td>6.1 1.9</td>
</tr>
<tr>
<td>(c)</td>
<td>11.4 11.4</td>
<td>6.3 3.2</td>
<td>6.1 2.5</td>
</tr>
<tr>
<td>(d)</td>
<td>12.1 12.1</td>
<td>5.1 1.4</td>
<td>2.1 1.1</td>
</tr>
<tr>
<td>kuruto (a)</td>
<td>2.3 1.8</td>
<td>1.6 0.4</td>
<td>1.0 1.0</td>
</tr>
<tr>
<td>(b)</td>
<td>2.4 2.4</td>
<td>1.3 0.5</td>
<td>1.6 1.6</td>
</tr>
<tr>
<td>(c)</td>
<td>3.3 3.1</td>
<td>3.0 1.4</td>
<td>0.9 0.9</td>
</tr>
<tr>
<td>(d)</td>
<td>6.7 6.7</td>
<td>5.1 1.8</td>
<td>2.2 2.2</td>
</tr>
<tr>
<td>nodemo (a)</td>
<td>5.8 5.4</td>
<td>6.5 5.9</td>
<td>3.7 1.2</td>
</tr>
<tr>
<td>(b)</td>
<td>9.6 5.1</td>
<td>9.6 8.1</td>
<td>6.2 3.0</td>
</tr>
<tr>
<td>(c)</td>
<td>4.8 4.4</td>
<td>5.7 5.3</td>
<td>4.0 2.4</td>
</tr>
<tr>
<td>(d)</td>
<td>9.3 9.3</td>
<td>9.4 8.0</td>
<td>1.6 0.3</td>
</tr>
</tbody>
</table>
Figure A.1: Time-loudness counter of four-mora phrases (dotted line) superimposed with mode value for each segment (solid line).
Figure A.2: Time-loudness counter of five-mora phrases (dotted line) superimposed with mode value for each segment (solid line).
Figure A.3: Time-loudness counter of three-mora phrase /harugal/ (dotted line) superimposed with mode value for each segment (solid line).
Figure A.4: Time-loudness counter of three-mora phrase /kuruto/ (dotted line) superimposed with mode value for each segment (solid line).
Figure A.5: Time-loudness counter of three-mora phrase /nodemo/ (dotted line) superimposed with mode value for each segment (solid line).
Table A.11: Correlation coefficients between vulnerability indices and the original durations, loudnesses or loudness jumps for each of the seven listeners, in terms of either Pearson’s product-moment correlation or Spearman’s rank correlation analysis in four- and five-mora phrases.

<table>
<thead>
<tr>
<th>listener ID</th>
<th>duration Pearson’s</th>
<th>duration Spearman’s</th>
<th>loudness Pearson’s</th>
<th>loudness Spearman’s</th>
<th>loudness jump Pearson’s</th>
<th>loudness jump Spearman’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.26</td>
<td>0.31</td>
<td>-0.16</td>
<td>-0.21</td>
<td>-0.02</td>
<td>0.20</td>
</tr>
<tr>
<td>2</td>
<td>-0.08</td>
<td>-0.05</td>
<td>-0.10</td>
<td>-0.03</td>
<td>0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>3</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.004</td>
<td>-0.07</td>
<td>0.10</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>0.44</td>
<td>0.11</td>
<td>0.17</td>
<td>0.18</td>
<td>0.15</td>
</tr>
<tr>
<td>5</td>
<td>0.03</td>
<td>0.08</td>
<td>-0.14</td>
<td>-0.23</td>
<td>-0.11</td>
<td>0.23</td>
</tr>
<tr>
<td>6</td>
<td>0.23</td>
<td>0.26</td>
<td>0.15</td>
<td>0.14</td>
<td>0.35</td>
<td>0.18</td>
</tr>
<tr>
<td>7</td>
<td>0.15</td>
<td>0.13</td>
<td>0.13</td>
<td>0.09</td>
<td>0.16</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Table A.12: Correlation coefficients between vulnerability indices and the original durations, loudnesses or loudness jumps for each of the seven listeners, in terms of either Pearson’s product-moment correlation or Spearman’s rank correlation analysis in three-mora phrases.

<table>
<thead>
<tr>
<th>listener ID</th>
<th>duration Pearson’s</th>
<th>duration Spearman’s</th>
<th>loudness Pearson’s</th>
<th>loudness Spearman’s</th>
<th>loudness jump Pearson’s</th>
<th>loudness jump Spearman’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.51</td>
<td>-0.43</td>
<td>0.31</td>
<td>0.31</td>
<td>0.43</td>
<td>0.49</td>
</tr>
<tr>
<td>2</td>
<td>-0.51</td>
<td>-0.44</td>
<td>0.31</td>
<td>0.35</td>
<td>0.40</td>
<td>0.51</td>
</tr>
<tr>
<td>3</td>
<td>-0.27</td>
<td>-0.25</td>
<td>0.49</td>
<td>0.53</td>
<td>0.48</td>
<td>0.56</td>
</tr>
<tr>
<td>4</td>
<td>-0.47</td>
<td>-0.34</td>
<td>0.32</td>
<td>0.27</td>
<td>0.43</td>
<td>0.45</td>
</tr>
<tr>
<td>5</td>
<td>-0.40</td>
<td>-0.40</td>
<td>0.63</td>
<td>0.63</td>
<td>0.58</td>
<td>0.63</td>
</tr>
<tr>
<td>6</td>
<td>-0.26</td>
<td>-0.19</td>
<td>0.78</td>
<td>0.74</td>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>7</td>
<td>-0.51</td>
<td>-0.45</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.67</td>
</tr>
</tbody>
</table>
Appendix B

Sentence data used in Chapter 5

1. Half the class has had it.
2. There’s nobody here by that name.
3. There is a restaurant.
4. There is a restaurant on the top floor.
5. Actually, could you make it two?
6. Would you mind explaining it to me once again?
7. He is a good player, isn’t he?
8. Jack can’t understand that.
9. Looks like he likes you.
10. This is it.
11. Let me handle it.
12. Miss Mills thinks it’s big.
13. There’re 80, 60, 40 and 25 dollar seats.
14. I’d like to make an appointment with Dr. Clifton for Friday afternoon.
15. Has Dad had a nap?
16. Your husband often goes abroad, right?
17. Do me a favor, will you please?
18. What time do you expect to check in?
19. Would it be possible to talk to you for an hour or so next week?
20. This tool isn’t useful for many people, is it?
21. Jim studies English everyday, doesn’t he?
22. Open the window, will you?
23. I’m OK for now.
24. I wish I were taller.
25. I’m looking forward to it.
26. The women met last Saturday.
27. Are these seats taken?
28. Will that be cash or credit?
29. What do you like to do in your free time?
30. Is it possible for me to come to your office to explain our new products?
31. The sunshine is dazzling.
32. I can never set it correctly.
33. Is it possible to switch our meeting to sometime in the afternoon.
34. OK, what can I do for you?
35. Can you make some time for further discussion on the price structure?
36. Since we only have a few things to talk about, why don’t we meet over lunch?
37. Let’s go shopping, shall we?
38. It’s got an astringent taste.
39. I don’t want you to say that.
40. I only have 30 minutes available after the meeting on Wednesday at 11:30.
41. Isn’t that a contradiction?
42. Which restaurant do you want to go to?
43. Would you like me to invite your friend for dinner this weekend?
44. Will you be able to come to our office in the near future?
45. I’m on my way to your hotel.
46. I’m on my way to your hotel, but there was an accident.
47. I’m on my way to your hotel, but there was an accident, and I’m stuck in the train now.
48. Why don’t we meet?
49. Why don’t we meet at the coffee shop?
50. Why don’t we meet at the coffee shop in the station building at Shinagawa?
51. Why don’t we meet at the coffee shop in the station building at Shinagawa around 12:30?
52. We’ll knock off 2,000 yen.
53. Get back exactly at midnight Miss Smith.
54. Take the elevator or escalator to the third floor.
55. Is it his?
56. Would you prefer smoking or non-smoking?
57. How long does it take to commute?
Bibliography


Minematsu, N., “Pronunciation Assessment based on upon the phonological distortions observed in language learners’ utterances,” in Proc. 8th International Conference on Spoken Language Processing, 2004.


List of Publication

Journal


International conferences


Technical meetings


