Perceptual Characteristics of Temporal Structures in Speech : Towards Objective Assessment of Synthesis Rules

加藤, 宏明

(Degree)
博士（工学）

(Date of Degree)
1999-03-31

(Date of Publication)
2009-10-14

(Resource Type)
doctoral thesis

(Report Number)
甲1927

(JaLCDOI)
https://doi.org/10.11501/3156328

(URL)
https://hdl.handle.net/20.500.14094/D1001927

PERCEPTUAL CHARACTERISTICS OF TEMPORAL STRUCTURES IN SPEECH: TOWARDS OBJECTIVE ASSESSMENT OF SYNTHESIS RULES

（合成規則評価のための音声の時間構造知覚の研究）

A DISSERTATION

SUBMITTED TO THE DIVISION OF INFORMATION AND MEDIA SCIENCE

GRADUATE SCHOOL OF SCIENCE AND TECHNOLOGY

KOBE UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Hiroaki Kato

January, 1999

要旨

合成規則評価のための音声の時間構造知覚の研究

本論文は，自然な音声の合成への指針を得るために行なわれた音声の時間構造知覚の研究をまとめたものである。任意のテキストから音声を生成する規則音声合成では，自然性の高い音声を実現するために，時間構造を自然発話の場合に近いものに制御する方法がとられている。この制御に要求される精度は，合成された音声の最終的な受益者である人間が許容する誤差の量によって決まる。したがって，最適な制御を行うためには，種々の誤差に対する人間の許容特性の知識が不可欠であるが，これまでは極めて限られた例の実験報告が得られるのみであった。

本研究では，人間が音声の時間構造の乱れを感じ取る際に影響される諸要因を心理実験により系統的かつ多面的に調査することで，合成規則における時間構造制御に必要な知識の集積を図り，さらに，それらの知識を合成規則の評価に反映させるための枠組みを提案し，実際に評価モデルを試作することで得られた知識の有用性を実証した。また，実験で得られた知識が今後分野を越えて広く使われることを想定し，心理音響学的な解釈の可能性を検討することによって当該知識の適用可能領域についての指針を与えた。本論文は，以下に述べる 8章からなる。論文の構成を Fig. 0.1に示す。

第 1章では，まず音声コミュニケーションにおける時間構造の役割とそれに対する知覚の特性を概説し，次に従来の音声合成規則の評価方法を概観し，それらが内包する知覚的側面からの不十分さを明らかにすることによって，本研究で提起する問題の所在を示す。

第 2章から第 5章では、第 1章で明らかにされた音声聴取における時間知覚に関する問題に取り組むため，“ 変形に対する感じやすさ ”という切り口により種々の条件の音声を実験刺激として知覚特性を調査する。章が進むにしたがい，音声刺激条件は基本的なものから複雑なものへと移行する。

第 6章では，第 2章から第 5章の音声知覚実験で得られた知見の心理音響学的妥当性を音声の時間構造を模した非音声音を用いた実験により検証する。この章での調査は，音声実験において得られた知識の適用可能領域を見積もるために重要と考えられる。

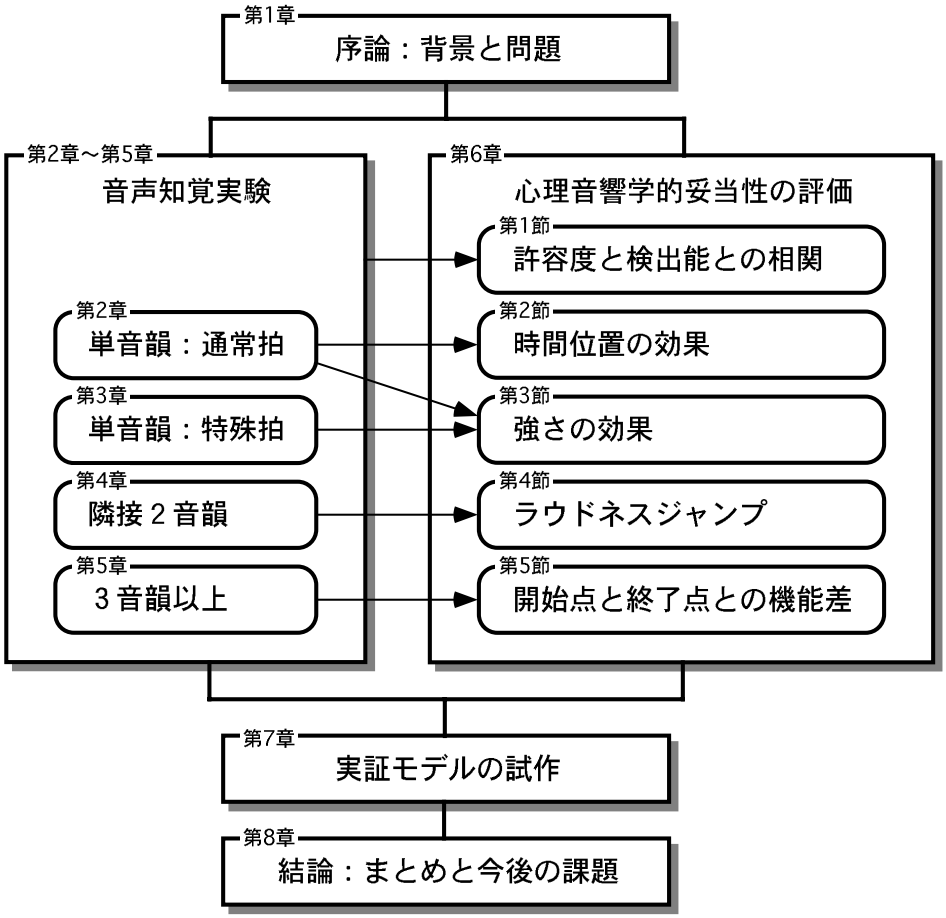

Figure 0.1: 論文の構成

第 7章では，ここまでの実験結果をまとめ，人間の知覚特性を考慮した音韻長歪みの評価モデルを試作することによって，得られた知見の合成規則の評価における有用性を実証する。

第 8章で以上の結果を総括し，今後の課題と将来の方向を展望する。

本研究の目的は，言い換えれば，人間が聴覚を通じて外界の変化を感じ取る能力の鋭敏さと遅鈍さとが音声の知覚にいかに巧みに組み込まれているかを知ることでもある。今後ますます需要が高まっていくであろう人と機械との音声を媒介としたコミュニケーションの場面において，このような人間の知覚メカニズムの知識が使用者に負担を強いない技術を確立するための基盤となることが期待される。以下に章毎の題を示し内容を概説する。

第 1章　序論

第 1章では，まず研究の動機と目的を述べた。音声コミュニケーションにおいて，時間構造は情報交換を円滑に行うために送り手と受け手とが共有すべきプロトコルの役割を果たす。しかし，機械が送り手となる音声合成では，時間構造は受け手の人間にとって常に最適に制御されるわけではない。すなわち，物理的時間と人が感じる時間すなわち心理的時間とは必ずしも線形に対応するわけではないため，現在広く行われている物理的時間のみに頼る客観評価法では真に最適な制御へのガイドラインを与えることはできない。本研究は音声の時間的歪みに対する人の知覚特性を系統的に調査することで，人の感じ方を取り入れた客観評価法の確立を図るものである。

本章では，加えて，この論文で用いた心理実験の枠組みを概説し，最後に，論文の構成を述べた。

第 2章　単母音長変形に対する許容特性
— 単一音韻長変形の知覚：通常拍 —

第 2章から第 5章では音声の時間構造に対する知覚感度の測定を行なった。第 2章では，最も基本的な実験条件として，日本語の通常拍，すなわち子音と母音とが規則的に交替する音声に含まれる単一の音韻に歪みが加わった場合を対象とした。その結果，母音長の変化に対する知覚の感度に影響を与える要因として，変化させた母音の種類（例：‘a’ か ‘i’ か），その単語内での時間的位置（例：１文字目か３文字目か），後に続く音韻の種類（例：有声音か無声音か），が挙げられることが分かった。変化させた母音の元々の長さの違いは知覚感度に影響を与えなかった。

第 3章　特殊拍を含む区間長の変形に対する許容特性
— 単一音韻長変形の知覚：特殊拍 —

日本語には通常拍に加えていわゆる特殊拍が存在し，その出現頻度も高い。第 3章では，第 2章と同様の方法で，特殊拍，すなわち撥音，促音，長音，無声化母音を含む音韻の変形に対する知覚感度を測定した。その結果，特殊拍では許容される時間歪みの範囲が通常拍の場合よりも一般に広く，それは特殊拍の定常部分の心理音響的性質，すなわち小さなラウドネス値，と特殊拍に固有の時間構造に起因するものとして説明できることが分かった。

第 4章　隣接 2 音韻の時間変形に対する許容特性
— 隣接 2 音韻長変形の知覚 —

合成音における時間歪みは一般には単一音韻のみではなく複数音韻に同時に生じるものである。第 4章では，複数音韻に生じた歪みの知覚に対する相互作用として，隣接 2 音韻間での知覚的音韻長補償効果を調査した。その結果，補償効果の生じやすさが，関係する２音韻間の心理音響的特徴の差（ラウドネスの落差）に依存しており，仮名一文字に相当するような音韻論上の単位とは無関係であることが分かった。

第 5章　音声の時間構造知覚における母音開始点と母音終了点との機能差
— 多数音韻長変形の知覚 —

音声合成における音韻長の制御は，各音韻の開始点と終了点の時間位置を決定することに他ならない。日本語では，子音の終了／開始点は大概母音の開始／終了点に一致する。第 5章では，さらに多くの音韻に同時に変形が加わった場合を想定して，母音開始点と終了点との時間構造知覚の手がかりとしての機能について調査した。その結果，音韻長の知覚的測定には母音終了点は開始点と同様に重要な役割を果たすが，発話テンポの把握などより広い範囲の処理には母音開始点のみが貢献することが分かった。

第 6章　音声知覚実験の心理音響学的解釈

第 2章から第 5章で調査した知覚特性に基づき，次章では所与の音韻長歪みに対する主観値を予測するモデルの構築を目指す。モデルの構築に際しては，より広い適用可能性を確保するため，人間の聴知覚の一般的特性に準拠したものを狙う。このため，本章では，第 2章から第 5章の各々で観察された音声の知覚特性が，人間の聴覚一般に備わった特性が反映されたものとして解釈可能かどうかを非音声音を用いた心理音響実験により検討した。ここで行なった実証的検討は，音声知覚実験で得られた結果の普遍性の吟味に役立つだけでなく，広く一般的な人間の時間知覚特性の解明にも貢献することが期待される。

第 7章　音韻長設定規則の客観評価モデル

第 7章では，ここまでの実験結果をまとめ，得られた知見が合成規則の評価に有用であることを音韻長歪みの主観評価モデルを試作することにより実証した。このモデルは，客観的指標のみによりながら，人間の主観に近い音韻長設定誤差の評価を可能にするものである。まず，音声の時間構造表現の一つの枠組みとして，時間 – ラウドネスマーカ表現を提案する。これは，ラウドネスの時間変化を音韻毎に量子化したものである。さらに，

この枠組みに則って音韻長設定誤差に対する客観評価モデルを実現し，このモデルにより心理実験から得られた主観評価データの予測を行った結果，従来型の音韻長誤差のみに基づく予測よりも一貫して精度の良い，すなわち人間の評価に近い，成績が達成された。

第 8章　結論

まず論文を手短かに総括し，最後に将来への課題を示して結びとした。

Acknowledgments

This dissertation is the outcome of a jointly operated graduate program between the Graduate School of Science and Technology, Kobe University and ATR Laboratories.

The work reported here could not be accomplished without the support received from several people. First, I would like to thank:

Shinzo Kitamura, one of my PhD advisors, for giving me the opportunity to write this dissertation and for helping me with a number of bureaucratic issues despite his busy schedule. He was also an advisor of my ME thesis.

Yoiti Ando, another of my PhD advisors, for taking the time to give me valuable feedback throughout the submission process.

Yoshinori Sagisaka, for guiding me through the PhD course as my principal advisor, for teaching me many important things as a supervisor at ATR Laboratories, and for helping me improve my thinking and writing as a colleague of this work. No part of this dissertation could have been possible without him.

I am equally indebted to Minoru Tsuzaki, my supervisor and colleague in my work over almost ten years at ATR Laboratories. None of the published or unpublished works incorporated in this dissertation could be presented without his “everyday” advice and encouragement.

I am also indebted to Masako Tanaka, my colleague at ATR. The experimental studies incorporated in section 6.2 were primarily conducted by her. She also assisted and encouraged me in performing most of the other experimental studies.

I also thank Yoh’ichi Tohkura, the former president of ATR Human Information Processing Research Laboratories (HIP), and the former head of the Hearing & Speech Perception Department, ATR Auditory and Visual Perception Research Laboratories (AVP), for giving me the opportunity to come and work at ATR and to join the PhD program of Kobe University, for teaching me what research life should be as well as all other technical affairs, and

for encouraging me to start this dissertation.

My first step in the field of speech research was taken while I was in the undergraduate program at Kobe University. I would like to express my gratitude to Haruya Matsumoto for his support as advisor for both the undergraduate and graduate (ME) programs, and to Kazuyoshi Tsutsumi for actually guiding me towards the first step of speech research. I would especially like to thank Hisao Kakehi, professor emeritus of the faculty of letters, Kobe University, for making me aware of the mysteries of spoken language.

This dissertation is based entirely on work done at ATR/AVP and ATR/HIP. I would like to express my sincere gratitude to the former and current directors of ATR Laboratories for providing such an excellent research environment and for their support and encouragement in the work, especially to Kohei Habara, the former chairman of the board of ATR/AVP and ATR/HIP, Yasuyoshi Sakai, the chairman of the board of ATR/HIP, Eiji Yodogawa, the former president of ATR/AVP, and Yutaka Ichinose, the president of ATR/HIP.

I would also like to thank former and current members of ATR Laboratories. First of all, Tatsuya Hirahara guided my first steps in speech perception research with patience and sincere passion during my starting years at ATR. Hideki Kawahara and Shigeru Katagiri, the former and current heads of the Auditory and Speech Communications Department, ATR/HIP, always gave warm support on my current work. Norio Higuchi and Nick Campbell of ATR Interpreting Telecommunications Research Laboratories taught me many important things in speech synthesis research. In particular, Dr. Campbell kindly read almost all of my English manuscripts and provided many helpful comments and suggestions. In addition, Alain de Cheveigné, a guest researcher at ATR, occasionally provided many helpful comments on my manuscripts. I am, indeed, indebted to many other members of ATR Labs., including Reiko Akahane-Yamada and Erik McDermott who gave me innumerable valuable things for more than ten years.

Some of the ideas and motivations propelling my work were picked up from the monthly meetings of a book-reading and discussion group of psychoacousticians in the Kansai area. I thank former and current members of this group including Kenji Kurakata, Michinao F. Matsui, and Masashi Yamada.

Finally, I am extremely grateful to all members of my family for their unconditional support and encouragement. In special, I thank my grandmother Chie Suzuki, parents-in-law Taiji and Sei Tazoe, and parents Akira and Chiyoko Kato.

This dissertation is dedicated to my family, Satoko and Yutaka , and my late sister Reiko Kurita and her family.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Abstract

Aiming to contribute to the assessment of rules for assigning segmental durations for speech synthesis, the present study examines the cues underlying the perception of temporal structures of speech, both from a speech perceptual and from a psychoacoustical point of view. In speech perception experiments, auditory sensitivity to different kinds of distortions in temporal structures of real speech is determined. In psychoacoustical experiments, the internal mechanism mediating perceptual effects on the speech stimuli is estimated.

In this chapter, we briefly describe the problems motivating this study and the approach taken in our experiments along with an outline of this dissertation.

1.1 Why temporal structures of speech?

Kaki-kue-ba	Tasting a persimmon
Kane-ga Naru-nari	I heard the bell
Horyuji	of Horyuji Temple toll

Masaoka, Shiki (1867–1902)

Stylized poetry such as a traditional verse form (either *haiku* or *tanka*) may remind some of us of the presence of speech rhythms, although we are rarely aware of them in our everyday oral communications.

Speech is, needless to say, one of the most important ways of human-to-human communications. Rhythms or, more generally, temporal structures of speech sounds serve an essential role as a form of protocol that ensures smooth and effortless information exchanges between humans. A transmitter (speaker) and a receiver (listener) sharing the same protocol (rhythms) can communicate successfully. The sharing of speech rhythms is naturally and easily established in human-to-human communications.

Recently, however, "high-tech" environments have been providing us a way, or even requiring us, to communicate with machines orally. In human-machine communications, machines are unable to intuitively understand our rhythms, and because of this, humans have to follow the rhythms fixed by the machines. Unfortunately, this situation can increase the burden for some, if not all, humans. For instance, a person might become irritated or even sort of angry toward an automatic answering system that continues to respond with monotone and somewhat unnatural rhythms irrespective of whether the person is in a rush or not. To lighten such loads on humans, machines should learn the way that humans perceive or produce temporal structures of speech.

In the current study, we aim at collecting empirical and systematic knowledge about human speech perception, and try to develop a framework for teaching this knowledge to machines. More specifically, we try to guide machines to produce naturally sounding speech by evaluating their performance in terms of temporal naturalness from the standpoint of human listeners. In other words, the general mission of the current study is to bridge the gap between the physical world, which machines deal with, and the internal world of humans, in the temporal aspect of speech. The collected knowledge in the study will comprise resources of a unified perceptual model of temporal structures in speech. This study, therefore, is expected to contribute not only to the assessment of synthetic rules but also to the design of general speech-oriented human-machine systems.

1.2 Potential problems in the assessment of synthesized speech

Although state-of-the-art synthesis technologies can provide fairly intelligible synthetic speech, the speech is, at present, still far from real speech in terms of naturalness or acceptability. Historically, the assessment of rule-based synthetic speech has generally taken two types of strategies. One is referred to as "subjective evaluation" which uses the judgment of human listeners and the other is "objective evaluation" which uses only the physical properties of the speech sounds to be tested. The former method has been widely and deeply investigated (Bailly *et al.*, 1992; Kasuya, 1993; Kasuya, 1992; Nusbaum *et al.*, 1995; Pols, 1991) because this type of evaluation is inevitable for any commercial synthesis system to ensure that its final performance is certainly tolerable for human listeners. In these years, especially, discussions on methods for this type have become extremely active in academic meetings of this field (van Santen *et al.*, 1997; COCOSDA, 1998). However, these methods can only occasionally be introduced into development stages because they usually require a large amount of human and time resources.

The latter method, in contrast, can provide a concrete criterion of development without consuming human resources because this method usually evaluates the differences between natural (target) and synthetic (tested) speech sounds in acoustic domains. Therefore, this method has been widely introduced in the development cycles of speech synthesis techniques, typically in corpus-based ones (Sagisaka, 1998). On the other hand, this objective evaluation method potentially holds a fatal, we would say, risk in that its outcome is not assured to be perceptually valid because its criterion, a physical measure, does not necessarily have a linear relationship with the corresponding perceptual measure or the perceived amount of a given temporal distortion. Although this potential inconsistency between physical and perceptual measures has been certainly considered as serious, few systematic studies have addressed this issue besides a small number of exceptions (e.g., van Wieringen, 1995, and Bakkum *et al.*, 1993, 1995).

In the current study, we conduct a series of experiments to investigate relationships between physical and perceptual measures of speech, especially in terms of the temporal structures, and try to provide a modeling of the temporal error evaluation for synthetic rules that can predict the acceptability to humans (a subjective measure) from only objective measures (physical properties) of speech signals. Such a model, if achieved, will enable a means of evaluation with the advantage of both objective evaluation and subjective evaluation, i.e., simplicity and perceptual validity.

1.3 General approach

We started an empirical and systematic approach in an attempt to reveal clues that govern temporal structures of speech during oral communications. By way of breaking down the problem, the first step was to narrow down the candidate features embedded in speech sounds and used by humans as temporal cues.

Speech sounds, in general, have neither obvious rhythmic structures nor clear beat points like musical sounds do. In a most strict sense, temporal cues can exist at every acoustic change in speech sounds. It is highly probable, however, that major temporal cues do not locate at steady-state (or gradually changing) parts of speech but instead locate at rapidly changing parts. Human perception, in nature, tends to ignore gradual or small changes, but will pick up rapid and large changes in auditory stimuli (Bregman, 1990). In addition, such cues in spoken sounds should be able to be found repeatedly and generally, i.e., with a reasonably high frequency.

As fair candidates conforming to these two conditions, we chose boundaries between consonant (C) and vowel (V) segments which usually have large acoustic changes and do not consider possible cues at the central portions of segments, which are relatively steady-state. What we should know about the perceptual effects of these cues can be, then, summarized into the following two issues: (1) Attributes of a duration between two consecutive cues or temporal markers, i.e., a single speech segment, and (2) attributes of a single cue. The latter issue can further be divided into two aspects of cue attributes: (2-1) the perceptual salience of each cue, and (2-2) the functional differences among cues. The following chapters are devised to explore each of these issues.

The current study, as a general methodology, utilized two measures of human temporal sensitivity from both psychoacoustical and speech-perceptual measures, i.e., detectability and acceptability. The detectability measure is precisely defined in psychophysical terms and can be estimated by many established methods. The acceptability measure, on the other hand, is not generally defined but is practically useful for the purpose of assessing of synthetic speech. This speech-perceptual measure is used to explore factors affecting the subjective evaluation of distortions given in temporal structures of speech, typically from the phonetic attributes of the stimuli. The psychoacoustical measure is applied to non-speech stimuli that replicate some aspects of the temporal structures in the speech stimuli as well as to the speech stimuli themselves, to investigate the internal mechanisms mediating the factors found in speech-perceptual experiments.

The current study also provides direct comparisons between these psychoacoustical and speech-perceptual measures to estimate the extent to which they reflect each other.

1.4 Outline of the dissertation

The subsequent four chapters, i.e., Chapters 2 through 5, measure the perceptual sensitivity to diverse temporal distortions given in speech sounds and try to explore the factors yielding inconsistencies between physical distortions and their perceptual consequences. As the chapters move forward, the complexities of the speech potions whose durations are subjected to modification increase as follows:

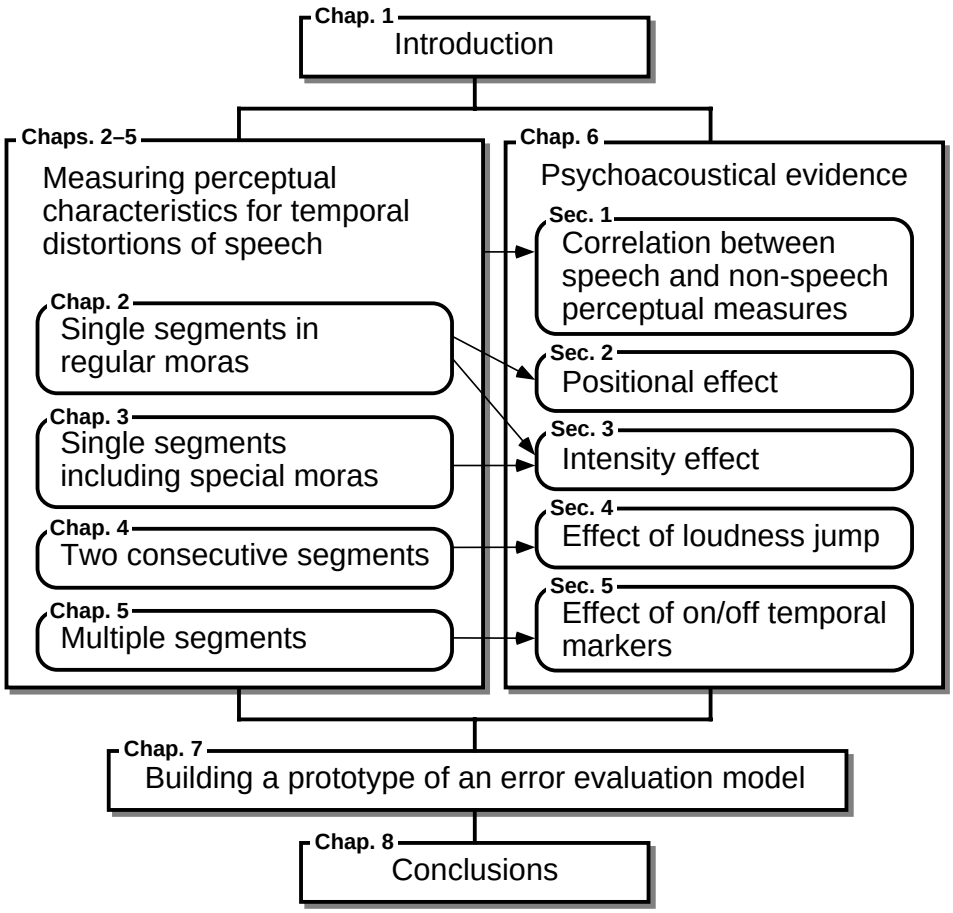

Figure 1.1: Outline of the dissertation.

- Single vowel segments.
- Single steady-state portions of speech. They are either vocal or consonantal parts and may include “special moras” as well as regular moras.
- Two consecutive segments.
- Multiple (more than two) segments.

These topics are addressed one by one in the following chapters as shown in Fig. 1.1. The results obtained are then evaluated and interpreted in terms of the psychoacoustical validity provided in Chapter 6. Such psychoacoustical interpretation of the results from speech experiments is extremely important to estimate the generality of the perceptual phenomena observed in speech cases. Then, on the basis of the obtained speech and non-speech experimental knowledge, a modeling of an evaluation for durational rules is proposed in Chapter 7. Finally, Chapter 8 summarizes the major findings of the dissertation. In this final chapter, some ideas are presented for future research.

Chapter 2

Acceptability of temporal modification of single vowel segments in isolated words[1]

Abstract

We measured the perceptual acceptability of changes in the segmental durations of vowels in regular (CV) moras as a function of the segment attributes or contexts, such as the base duration, temporal position in a word, vowel quality, and voicing of the following segment. Seven listeners estimated the acceptability of word stimuli in which one of the vowels was subjected to temporal modification from -50 *ms (for shortening) to +50 ms (for lengthening) in 5 ms steps. The temporal modification was applied to vowel segments in 70 word contexts; their durations ranged from 35 ms to 145 ms, the mora position in the word was the first or third position, the vowel quality was /a/ or /i/, and the following segment was a voiced or an unvoiced consonant. The experimental results showed that the listeners' acceptable range for durational modification was narrower for vowels in the first moraic position in the word than for vowels in the third moraic position. The acceptable range was also narrower for the vowel /a/ than for the vowel /i/, and similarly narrower for vowels followed by unvoiced consonants than for those followed by voiced consonants. The vowel that fell into the least vulnerable class (the third /i/, followed by a voiced consonant) required 140 % of the modification of that which fell into the most vulnerable class (the first /a/, followed by an unvoiced consonant), to yield the same acceptability decrement. In contrast, the effect of the original vowel duration on the acceptability of temporal modification was not significant despite its wide variation (35–145 ms).*

[1]This chapter is published as Kato, Tsuzaki, and Sagisaka (1998a).

2.1 Introduction

Rules to assign segmental durations have been proposed for speech synthesis to replicate the segmental durations found in naturally spoken utterances (Allen *et al.*, 1987; Carlson and Granström, 1986; Campbell, 1992; Fant and Kruckenberg, 1989; Higuchi *et al.*, 1993; Kaiki and Sagisaka, 1992; Klatt, 1979; Riley, 1992; van Santen, 1994; Takeda *et al.*, 1989). Each of the segmental durations achieved by such durational rules generally has some error compared with the corresponding naturally spoken duration.

The effectiveness of a durational rule should be evaluated by how well such error is accepted by human listeners, who are the final recipients of synthesized speech in general. In almost all previous cases, however, the average absolute error of each segmental duration from its standard has been adopted as the measure for objective evaluation.

One possible problem with this measure is that it gives every segment the same weight in the error evaluation. In other words, it neglects factors that may affect the perception of segmental durations, such as (1) interactions among errors in different segments (Kato *et al.*, 1997) and (2) variations in segment attributes (Bochner *et al.*, 1988; Carlson and Granström, 1975; Huggins, 1972a; Klatt and Cooper, 1975). If the perceptual sensitivity to durational modification largely depends upon these factors, the reliability of the traditional measure, which gives every segment the same weight, should be seriously reconsidered. At the same time, if such a segmental effect on the perceptual evaluation could be quantitatively specified, we could obtain a more valid (closer to human evaluation) measure than the traditional one to evaluate durational rules.

In our previous study (Kato *et al.*, 1997), we focused on the first factor, i.e., interactions among errors, and investigated to what degree the temporal modification in one segment is compensated by the modification in an adjacent segment. As a result, the amount of perceptual compensation was found to be inversely correlated with the difference in loudness between the two segments of interest.

The current study looks at the second factor and examines whether the attributes and contexts of individual segments affect the perceptual sensitivity to durational modification, and, if so, how large these effects are.

2.1.1 Factors affecting perceptual evaluation of changes in segmental durations

First, we will consider factors affecting auditory sensitivity to durational changes. It is widely acknowledged that the absolute just noticeable differences (jnd's) of auditory durations will increase with an increase in the base duration for stimuli ranging in the duration of speech segments (e.g., from 50 to 300 ms) (Abel, 1972a; Abel, 1972b; Creelman, 1962; Fujisaki *et al.*, 1975; Small and Campbell, 1962). According to these studies, although the relative jnd's or Weber fractions, i.e., the proportions of jnd's to their corresponding base durations, may decrease slightly with an increase in the base duration of this range, the absolute jnd's still keep increasing. Abel (1972a), for example, suggested that the absolute jnd was roughly proportional to the square root of the base duration.

An analogous tendency was also observed for durational jnd's of vowel segments (Bochner *et al.*, 1988). Bochner *et al.* showed that jnd's tend to increase with an increase in the base duration for six vowels (75–170 ms) presented in isolation or in a consonant-vowel-consonant (CVC) sequence. Klatt and Cooper (1975), on the other hand, reported absolute jnd data for vowel or syllable nucleus durations not appearing to depend on the base duration. Their target was the stressed vowel /i/ in the word "deal(er)" embedded in various sentences; the durations of the syllable nuclei which included the target vowels ranged from 165 to 340 ms. The jnd's varied almost independently from the durations of the syllable nuclei, i.e., the base durations.[2] These jnd studies on vowel or syllable nucleus durations cannot be directly compared to each other, however, owing to differences in the range of stimulus durations and in the stimulus context that embedded the target duration. Therefore, the influence of the base duration on temporal sensitivity, particularly for a speech segment, warrants testing by further systematic studies.

In addition to the factor of the base duration, three factors have been reported which can affect auditory sensitivity to durational modification in speech segments. The first factor is the temporal position in a word; Klatt (1976) reported that a durational jnd is smaller for the segments in the first syllable of a two-syllable word than for the segments in the word-final syllable. The second factor is the effect of the following word; the jnd for a vowel duration is smaller at the sentence end than elsewhere in the sentence (Klatt, 1976). Klatt suggested that this effect probably involves a backward recognition masking (Massaro and Cohen, 1975);

[2] Although the authors of the original reference did not mention this irrelevancy explicitly, our analysis of their data showed no significant correlation between the base duration and the absolute jnd. The Pearson-product moment was very small $[r = 0.011]$. The linear regression also turned out to be not effective $[F(1,5) = 0.0007, p = 0.98]$.

i.e., if other words follow the crucial segment, the jnd will increase. The third factor, which has been mentioned by both Huggins (1972) and Carlson and Granström (1975), is the type of segment. They reported that their subjects were more sensitive to durational changes in vowel segments than to those in consonant segments. This third factor, however, did not appear to be significant in Fujisaki *et al.*'s experiments (Fujisaki *et al.*, 1975); the jnd obtained was almost equal for all types of segments, regardless of whether the crucial segment was a vowel or a consonant. Fujisaki suggested that such a disagreement was probably due to whether durational differences served as cues for the phonemic distinctions; the segmental durations investigated by Fujisaki *et al.* were phonemic in Japanese while those tested by the other studies were not phonemic (Carlson and Granström, 1975, commentary section).

Although jnd's can be precisely defined in psychoacoustical terms, acceptability is another useful measure for evaluating errors in durational rules. This measure can be considered more practical than jnd's, because there are many cases in which we can accept two tokens as natural utterances even though they are clearly discriminable. Acceptability, however, has seldom been investigated, except for several pioneering studies (Sato, 1977; Hoshino and Fujisaki, 1983; Sagisaka and Tohkura, 1984). Sato (1977) investigated the acceptability of temporal modification in vowel segments within isolated words. The acceptable modifications reported ranged from 15 to 30 ms and were smaller for word-initial segments than for word-medial or -final segments. This kind of positional effect is consistent with that observed for durational jnd's by Klatt and Cooper (1975). Hoshino and Fujisaki (1983) used vowel or consonant segments and reported that shortening modifications were more acceptable than lengthening modifications. Sagisaka and Tohkura (1984) used sentence stimuli in which every segmental duration was subjected to random modification and reported that the acceptable modification size was about 30 ms.

Since these studies on evaluation by acceptability were rather elementary and were primarily designed to measure a "rough" range of acceptable modification able to be instantly applied to their own durational rules, however, the number of speech samples employed in each of the studies was not very large. Sato's study used four words, all starting with the same phoneme sequence (*saka, sakana, sakanaya,* and *sakanayasan*), Hoshino *et al.* used three nonsense words (*hatapaka, hatabaka,* and *hapakka*), and Sagisaka *et al.* used two sentences. Consequently, one should be prudent in generalizing the tendencies observed in these studies. Furthermore, they did not provide any information for evaluating whether or not factors affecting temporal jnd's would also affect acceptability, except for the factor of the temporal position in a word.

2.1.2 Factors to be tested in the study

Based upon the above background discussions, the current study was designed to provide a direct and reliable test as to whether segment attributes or contexts affect the acceptability of durational modification in speech segments. For this purpose, we limited the number of factors to be tested and their constituent levels in order to obtain sufficient data for each level of the factors. The tested factors were chosen from the factors themselves or from relevant factors that were reported as being effective in previous jnd or acceptability studies. Some of the chosen factors are also important in view of the control of segmental durations for speech synthesis (Crystal and House, 1988; Klatt, 1979; van Santen, 1992; Kaiki and Sagisaka, 1992), so experiments using these factors are of practical benefit for establishing durational rules based on perceptual characteristics.

Original duration We felt it was necessary to first examine the factor of the base or original duration because doing so is crucial for evaluating segmental errors, in that it determines the unit in which the errors should be considered, i.e., the absolute duration (e.g., milli-seconds) or relative duration (e.g., percentage). To test this factor, generally speaking, a wide variety of durations are necessary for the test segments of the original materials. Therefore, the segments to be tested were chosen from vowel segments, which generally have the widest durational dispersions in spoken Japanese (Sagisaka and Tohkura, 1984). We then examined the following three factors.

Temporal position within a word The tendency for a word-initial segment to be more susceptible to temporal modification than a word-medial or -final segment was previously observed both in a jnd study (Klatt and Cooper, 1975) and in an acceptability study (Sato, 1977). Both studies found this positional effect only in a single stimulus context. Accordingly, we included this factor to test whether it is statistically robust.

Vowel quality The factor of phoneme difference was suggested in previous jnd studies (Carlson and Granström, 1975; Huggins, 1972a). Although the contrast observed in these studies was between vowels and consonants, we employed the contrast between two different vowels because the current study treated only vowel segments owing to the requirement of a wide durational variation. The vowel quality is, indeed, a major control factor in terms of durational rules (Peterson and Lehiste, 1960; Umeda, 1975; Sagisaka and Tohkura, 1984).

Voicing of the following consonant Klatt *et al.* (1975) suggested the backward influence of words immediately following the segment in question. In accordance with this suggestion, the factor of the following segment was examined in the current study. To obtain enough samples to enable a reliable statistical analysis for each of the constituent levels, we focused on the contrast between voiced and unvoiced consonants. The voicing of postvocalic consonants has been acknowledged to be a control factor for vowel durations in many languages (Delattre, 1962; Luce and Charles-Luce, 1985).

2.2 Method

2.2.1 Subjects

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

2.2.2 Stimuli

The original materials were taken from the same speech database as reported in our previous paper (Kato *et al.*, 1997). The current study modified only one segment within a word while the previous study additionally modified either the preceding or following segment.

Seventy words were selected from the ATR speech database (Kurematsu *et al.*, 1990). All of them were commonly used four-mora Japanese words[3], excluding words with doubled vowels, geminated consonants, or moraic nasals[4] which have heterogeneous syllable structures and, consequently, may disturb the temporal regularities observed in the open syllable sequences. The selected words were spoken naturally in isolation by one male speaker and were digitized at a 12 kHz sampling frequency and with 16-bit precision.

One segment out of four vowel-segments in each stimulus word was subjected to durational modification. Each segmental duration of these target vowels was manually measured from spectrographic images by well-trained labelers. The measured durations of the target vowels ranged from 35 ms to 145 ms as shown in Fig. 2.1.

The three factors other than the original duration were represented in the selected materials by the following three contrastive aspects of the target vowels: (1) the temporal position in a word was either the first or third moraic position, (2) the vowel quality was either /a/ or

[3]To maximize the freedom in the word selection, we chose the materials from the four-mora words which are lexically the most frequent in contemporary Japanese (Yokoyama, 1981).

[4]In the orthography, each of them has a separate character with the same status as the CV units.

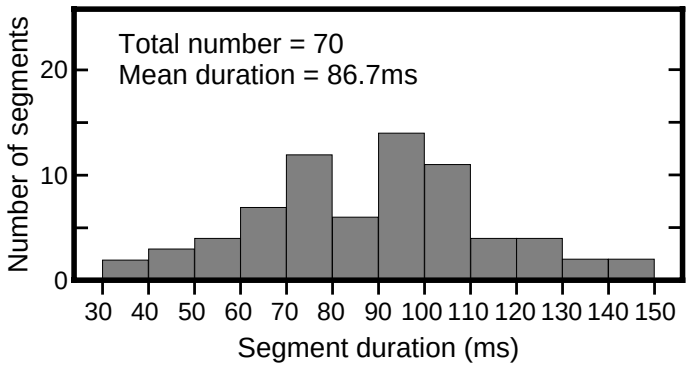

Figure 2.1: Histogram in 10 ms bins showing the distribution of the segmental durations of the original vowel materials. Each duration was manually measured from spectrograms by trained labelers.

/i/ (i.e., the lowest or highest vowel in Japanese), and (3) the following consonant was either voiced or unvoiced. Table 2.1 shows the number of target vowels with these contrasts. The 70 selected tokens are listed in Appendix A (Table A.1) with the attributes of each of the target vowels.

The temporal modifications were made by a cepstral analysis and resynthesis technique with the log magnitude approximation (LMA) filter (Imai and Kitamura, 1978), and were carried out at 2.5 ms frame intervals. The duration change was achieved by deleting or doubling every n-th frame in the synthesis parameters throughout the whole vowel.

Each target vowel duration was shortened or lengthened over a range that extended from -50 ms to $+50$ ms from the original duration in 5 ms steps, resulting in 21 different modification steps. Since five of the 70 target durations were less than 50 ms, i.e., their modification did not reach -50 ms, their modification steps were less than 21. In total, 1452 word stimuli were prepared.[5] As a result of a preliminary listening session, it was confirmed that there was no phonemic shift in either the target vowels or the surrounding phonemes caused by these manipulations.

2.2.3 Procedure

The prepared stimuli were first recorded onto a digital audiotape (DAT) through a D/A converter (MD-8000 mkII, PAVEC) and a low-pass filter (FV-665, NF Electronic Instruments,

[5]They were (20 modification steps + 1 unmodified) $\times$ 70 vowels - 18 (incomplete steps for the target vowels whose durations were less than 50 ms).

Table 2.1: The number of selected vowel segments falling into each of the cells defined by the factors to be tested. The number of segments followed by voiced consonants is in parentheses.

	First mora	Third mora	Total
/a/	21 (15)	22 (12)	43 (27)
/i/	14 (10)	13 (10)	27 (20)
Total	35 (25)	35 (22)	70 (47)

$f_c = 5700$ Hz, -96 dB/octave) with a DAT recorder (DTC-55ES, SONY), and then presented diotically to the subjects through headphones (SR-Λ Professional, driven by SRM-1 MkII, STAX) in a sound-treated room. A four-second interval was inserted after each presentation for the subjects' response. The average presentation level was 73 dB (A-weighted) which was measured with a sound level meter (type 2231, Brüel & Kjær) mounted on an artificial ear (type 4153, Brüel & Kjær).

The subjects were told that each stimulus word was possibly subjected to temporal modification. Their task was to evaluate how acceptable each stimulus was, as an exemplar of the token of that stimulus on a seven-point rating scale ranging from -3 to 3, where -3 corresponded to "quite acceptable" and 3 corresponded to "unacceptable."[6] The subjects were asked to limit their responses to the temporal aspects of the stimuli, as much as possible.

A total experimental run for each subject comprised of ten sessions. Seven of the 70 tokens were chosen for each session, and four repetitions of their 21 modified versions were randomly presented in the session. Accordingly, each subject evaluated each stimulus four times in total. Each of the seven tokens within a session was carefully picked from each constituent level of the factors to be tested, assuring that the seven tokens were as uniformly distributed as possible throughout the levels; the primary criterion was the uniformity in the variation of the original duration.

[6] If the listeners were asked to evaluate the "naturalness," they might have tended to use a strict criterion making it difficult for an informative evaluation to be maintained for the whole range of temporal modifications to be tested. To obtain information for a reasonably wide range of modifications, therefore, we chose the "rating of acceptability" over the "rating of naturalness."

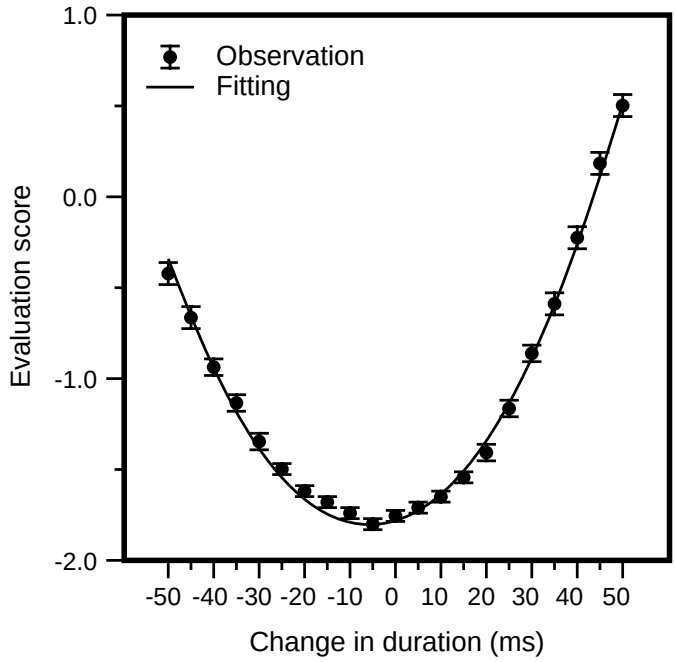

Figure 2.2: Evaluation scores of acceptability pooled over stimuli and subjects as a function of the change in the duration of vowel segments. Each point consists of 1960 observations (70 stimuli $\times$ 7 subjects $\times$ 4 observations). The error bars show the standard errors. The fitting curve by second-order polynomial regression is superimposed. The curve is formulated as: $y = 0.000743(x + 5.68)^2 - 1.77$.

2.3 Results

2.3.1 Measure of acceptability

Figure 2.2 shows the obtained evaluation scores pooled over all subjects and all target vowels, plotted as a function of change in the duration of the target vowels. As shown in the figure, the evaluation scores had a general inclination; i.e., the bottom of the scatter plot, the most acceptable point, is located around the center of the horizontal axis, and the scores increase in an accelerated manner from that bottom point as the absolute change increases. The subjects' sensitivity to durational modification can be represented by the sharpness of the rise from the bottom point.

Although a similar inclination was observed in all of the individual plottings obtained for each combination of subjects and target vowels, the size of the horizontal or vertical shift of the bottom point varied depending on the subject or target. These bottom point shifts

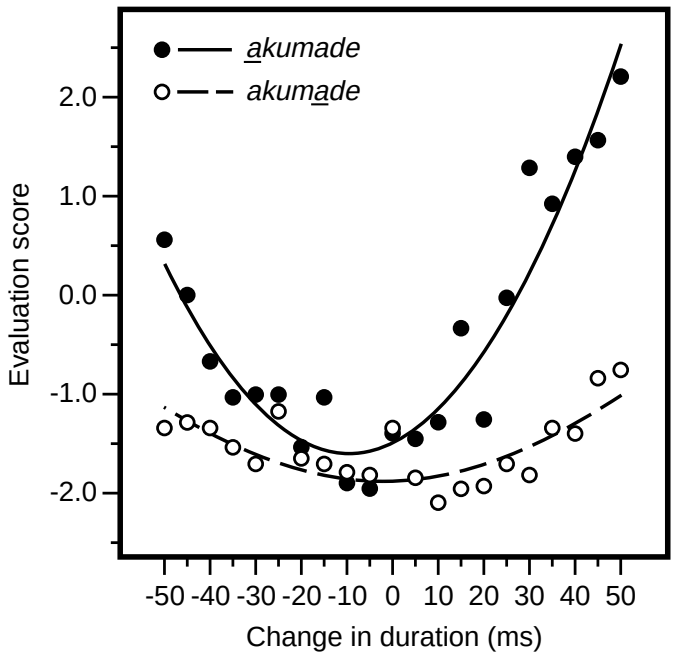

Figure 2.3: An example illustrating a difference in acceptability-change between two different vowels. Those vowels subjected to durational modification are underlined in the legend. The scatter plots show that the evaluation score varies according to the durational change more drastically for the first vowel of the word "*akumade* (to the bitter end)" (filled circles), than for the third vowel of the same word (open circles). The two regression curves trace this tendency. These are formulated as: $y = 0.00116(x + 9.48)^2 - 1.44$ (solid line), and $y = 0.000322(x + 2.02)^2 - 1.84$ (dashed line).

can be regarded as subject response biases.[7] To parameterize the subjects' sensitivity or the vulnerability of the target vowels irrespective of these response biases, a parabolic regression as generally formulated below was applied to the plot (superimposed) in Fig. 2.2,[8]

$$Evaluation\ score = \alpha(\Delta T - \beta)^2 + \gamma, \tag{2.1}$$

where ΔT denotes the change in duration; the unit of ΔT is not the relative duration but milliseconds. This regression was the best fitted polynomial function to this plot on the basis

[7] The horizontal value of the bottom 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 bottom point reflects the score which the subject gave to his/her best duration. These two aspects might be influenced by factors that are difficult to control in experiments, i.e., subjects' response biases.

[8] Although we could choose fitting functions other than polynomial fittings and/or could rescale the vertical axis to obtain an interval scale, we adopted the parabolic fitting on the raw evaluation scores because of the advantage of its directly reflecting the subjects' responses and its goodness of fitting.

of the *F*-ratio criterion, i.e., the second-order model achieved the smallest probability of the null hypothesis $[F(2, 10231) = 1699.4, p < 0.0001]$. The coefficient of the second-order term or α of this parabolic curve shows the rate of change in the evaluation score with a change in the durational modification; both the horizontal and vertical response biases can be separated out as β and γ. As derived from Equation 2.1, α represents the average decrement[9] of the acceptability for a certain size of temporal modification, and also the width between the longer and shorter limits of the temporal modification which yields a certain amount of acceptability decrement, i.e., an acceptable range. Therefore, α indicates the vulnerability of the vowel durations from the temporal modification.

From this, we adopted the second-order polynomial coefficient of the fitting curve as the object variable of the current study and refer to it as the vulnerability index or simply α, hereafter. We then applied a parabolic fitting to the evaluation scores for each of the 70 target vowels and each of the seven subjects, obtaining 490 fitting curves. Figure 2.3 shows typical examples of individual fittings illustrating the difference in the acceptability change between two different targets. Prior to the statistical analyses, we dropped unreliable data (fitting curves) on the basis of two criteria: (1) a fitting curve in which α was not positive was dropped because it suggests that the subject probably sensed no durational change for that particular token, and (2) a fitting curve in which the axis was extremely remote (more than six times sigma) from the distribution center of the entire data was dropped. In all, ten fitting curves were excluded, i.e., seven eliminations due to the first criterion, one elimination due to the second, and two eliminations due to both criteria, resulting in 480 fitting curves. Consequently, the object variable of this study consists of 480 α scores.

2.3.2 Effect tests

First, the effect of the original segmental duration on the vulnerability index was tested. Its Pearson's product-moment correlation coefficient was extremely small $[r = 0.0189]$ (e.g., McCall, 1980) and the general linear model (SAS Institute Inc., 1990) with subject as the random factor showed no significant effect $[F(1, 6) = 0.127, p > 0.73]$. (A more tolerant criterion, Spearman's rank correlation test, also showed no significant effect $[\rho = 0.077]$.)

To examine these analyses in more detail, we further computed correlations between the original duration and the vulnerability index within each of the seven subjects. As shown in Table 2.2, the absolute values of the correlation coefficients were fairly small and their signs

[9] As we assigned a larger evaluation score to a more unacceptable or less acceptable impression, an increment of the evaluation score implies a decrement of the acceptability.

Table 2.2: Correlation coefficients between the original duration and the vulnerability index (α) for each of the seven subjects, in terms of either Pearson's product-moment correlation or Spearman's rank correlation analysis.

Subject ID	Pearson's r	Spearman's ρ
A	0.25	0.31
B	0.081	0.17
C	-0.17	-0.12
D	0.31	0.36
E	-0.27	-0.21
F	0.17	0.23
G	-0.060	-0.041

did not agree with each other, i.e, four positive correlations and three negative ones.

Next, the effects of the other three factors on the vulnerability index were tested. A three-way factorial ANOVA of repeated measures was performed with position in a word, vowel quality, and voicing of the following consonant as the main factors, and with subject as the blocking factor. The main effects of position in a word, vowel quality, and voicing of the following consonant were significant [$F(1, 42) = 22.5, p < 0.001$; $F(1, 42) = 16.3, p < 0.001$; $F(1, 42) = 30.9, p < 0.001$, respectively]. As shown in Fig. 2.4, panels (a) to (c), the vulnerability index was greater for vowels at the first moraic position in a word than vowels at the third moraic position. It was also greater for the vowel /a/ than for the vowel /i/, and similarly greater for vowels followed by unvoiced consonants than for those followed by voiced consonants. There was a significant interaction between the factors of vowel quality and voicing of the following consonant[$F(1, 42) = 10.7, p < 0.003$]; the effect of voicing of the following consonant was larger for the vowel /a/ than for the vowel /i/. No other interaction was significant.

The target vowel falling into the most vulnerable (i.e., susceptible to durational modification) combination of these three factors was /a/ followed by an unvoiced consonant at the first moraic position in a word. That falling into the least vulnerable combination was /i/ followed by a voiced consonant at the third moraic position in a word. The ratio of the averaged α score of the most vulnerable targets to that of the least vulnerable targets was 1.96. This means that the least vulnerable targets required 140 % of the temporal modifications of

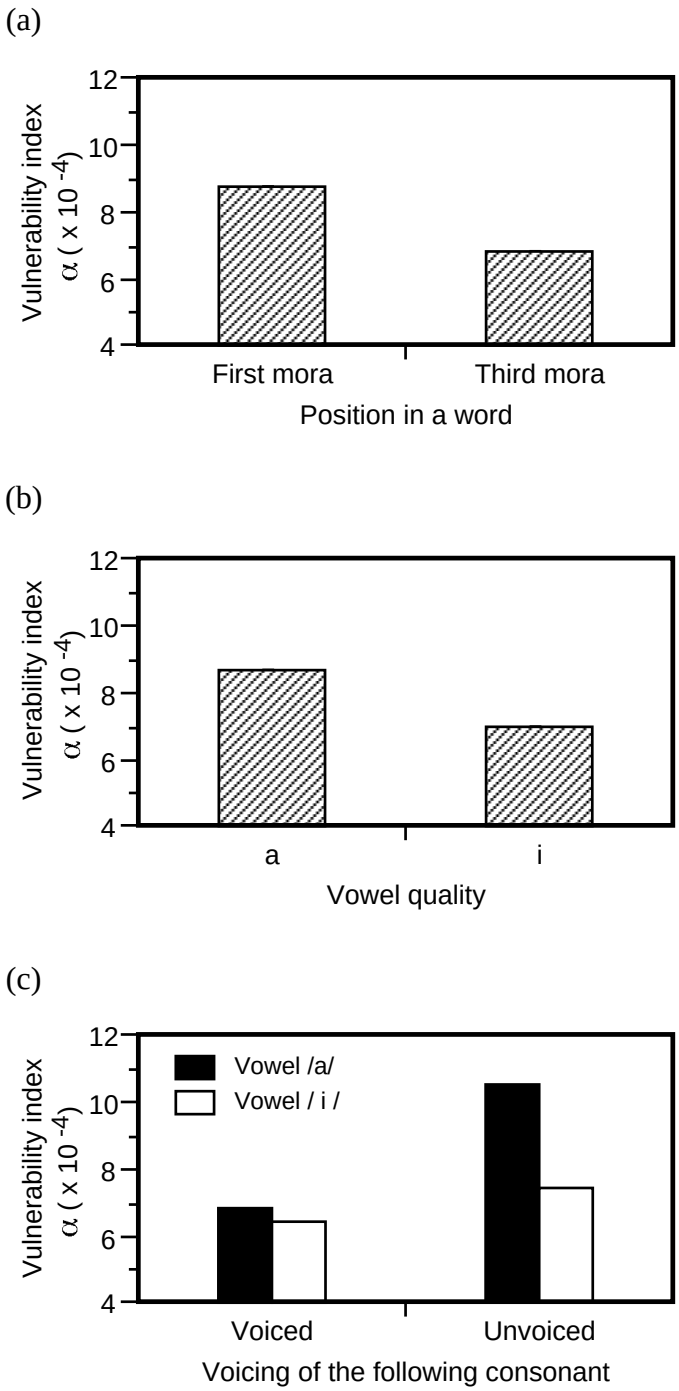

Figure 2.4: The least squares mean of the vulnerability index (α), i.e., the second-order polynomial coefficient of the fitting curve, for each level in the factors of (a) temporal position in a word, (b) vowel quality, and (c) voicing of the following consonant; they were calculated in the ANOVA procedure. A larger α implies a narrower acceptable range.

the most vulnerable targets to yield the same acceptability decrement. From another aspect, a certain modification size evaluated as acceptable for one vowel, say 50 ms, can become unacceptable for another vowel depending on the vowel's attributes and context.

2.4 Discussion

The primary purpose of the current study was to test whether or not the segment attributes or the context affects the perceptual sensitivity to modification in vowel durations, and if so, how large these effects are. The results showed that the vulnerability index (α score) was affected by the three experimental factors: (1) position, (2) quality, and (3) voicing. They indicated that perceptual acceptability changes as a function of temporal modification at a different rate depending on the attributes and context of the segment, such as its temporal position in the word, its vowel quality, and the voicing of the following consonant. In contrast, the original vowel duration had no significant effect on the vulnerability index despite its wide variation, i.e., 35 to 145 ms.

The following two subsections will try to provide perceptual implications for each of the test factors that did or did not yield a significant effect, in relation to previous studies on auditory temporal perception.[10]

2.4.1 Original duration

The present results revealed that the original vowel duration had no significant linear relation with the vulnerability index. This finding implies that the absolute acceptable range depends little on the original vowel duration. This appears to disagree with the result of Bochner *et al.*'s (1988) study and agrees with that of the study of Klatt and Cooper (1975). Bochner *et al.* reported that the absolute jnd of the vowel duration increases with an increase in the original vowel duration. Klatt *et al.*, on the other hand, found no significant correlation between the jnd's of vowel durations and their original durations. The stimulus conditions of these two studies differed from each other principally in two aspects: the range of vowel durations tested, and the context in which the tested vowels were embedded.

[10]Most of the studies included in the current discussion were based on the measurement of duration discriminability. Although discriminability studies may not be directly compared with acceptability studies in general, the vulnerability index used in the current study has been reported to reflect, to some extent, variations in duration discrimination thresholds or jnd's (Kato et al., 1992). We, therefore, recognize that it is valuable to discuss the agreements and discrepancies between the current results of acceptability tests and the previous results of jnd studies.

The range of vowel durations tested in the current study, 35 – 145 ms, was closer to that of Bochner *et al.*'s study, 75 – 175 ms, than that of Klatt *et al.*'s study, 165 – 340 ms, although the current results did not agree with Bochner *et al.*'s but with Klatt *et al.*'s. Therefore, the difference in the range of vowel durations is not likely to explain the disagreement between the results of Bochner *et al.*'s study and those of both Klatt *et al.*'s study and the current study.

For the presentation context of the tested vowels, Bochner *et al.* used a monosyllabic context, i.e., CVC or isolated vowel segment. Such a single-nucleus or isolated context is comparable to the stimulus presentations of previous jnd studies using non-speech filled durations (Abel, 1972b; Ruhm *et al.*, 1966; Small and Campbell, 1962); they presented single durations in isolation. The results of these non-speech studies commonly showed a tendency similar to that of Bochner *et al.*'s study, i.e., the jnd's were roughly proportional to the base durations. The current study, on the other hand, presented a vowel segment in a polysyllabic context, i.e., a four-vowel word. Klatt *et al.* also employed a polysyllabic context, i.e., a sentence. This contrast in the presentation context might suggest a general tendency underlying the perception of durational changes; i.e., an isolated or monosyllabic context makes the influence of the base duration effective while a polysyllabic context reduces or hides it. A polysyllabic presentation generally provides widely distributed information that spans a two-segment range or more, which listeners can utilize for their judgment in addition to the target duration itself. Although it would be difficult to specify the valid cues in these wider processes owing to the limitation of the stimulus manipulation in the current experiment, it can be assumed that the listeners would tend to depend on temporal cues distributed in a range wider than a single segment. The presence of such wider-ranging processes can be suggested by the temporal compensation phenomenon between two consecutive segments as Huggins (1968) and Kato *et al.* (1997) pointed out.

2.4.2 Psychophysical implications of the three factors affecting acceptability

The experimental results showed that at least three aspects of vowel segments affect the acceptability evaluation for temporal modification in the segments: (1) the temporal position within a word, (2) the vowel quality, and (3) the voicing of the following consonant. Implications of these effects are discussed below. The discussions focus on whether the effects can be interpreted within psychophysical or auditory-based knowledge instead of speech-specific features. Psychophysical interpretations do tend to provide a wider generalization of the effects than speech-specific ones do; they enable wide applicability in evaluating new

materials that have not actually been examined.

Positional effect

The effect of the segment's position in the word found in the current study is consistent with that observed in previous studies on both acceptability estimation (Sato, 1977) and jnd measurement (Klatt and Cooper, 1975). The listeners responded more critically to modification in word-initial vowels than to modification in the following vowels. A positional effect of this sort has also been reported for non-speech stimuli (Tanaka *et al.*, 1994). Tanaka *et al.* measured temporal jnd's for one of three successive intervals marked by four click sequences, which were devised so as 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, i.e., the listeners responded more sensitively to the temporal modification at the initial position than to that at the following position. Consequently, the positional effect can be explained by a speech-specific mechanism.

Effect of vowel quality

The modifications in the vowel /a/ had a stronger influence than those in the vowel /i/. In Japanese, the high vowel /i/ has a smaller intrinsic power, which correlates highly with loudness, than the low vowel /a/ has (Mimura *et al.*, 1991) as observed in the contrast between high and low vowels in English (Lehiste and Peterson, 1959). This contrastive property of the power or loudness between these two vowels is likely to explain the effect found. Several studies have reported that the power or intensity of the stimuli will affect the discrimination performance for different filled durations (Creelman, 1962; Kato and Tsuzaki, 1994; Tyler *et al.*, 1982).[11] Although the stimuli in these studies were not taken from speech, these studies commonly pointed out that a higher intensity level in the stimuli yielded a higher discrimination performance for the stimulus duration. This tendency can qualitatively explain the current effect of the vowel quality; i.e., the vowel /a/ is more susceptible to temporal changes than the vowel /i/.

In the current experiment, the loudness of the target vowel /a/ was found to be, certainly,

[11] For auditory intervals filled with noise or pure tones, the discrimination performance for stimulus durations does not appear to be affected by a change in the intensity, as long as the stimulus whose duration is to be judged is presented in isolation and is clearly audible (Allan, 1979; Abel, 1972a; Creelman, 1962). However, it has been reported to be affected if the duration to be judged is not clearly marked (Creelman, 1962; Tyler *et al.*, 1982) or if it is presented between preceding and following sounds, just like a speech segment in a word is usually surrounded by other segments (Kato and Tsuzaki, 1994).

greater than that of the target vowel /i/. The average loudness of the vowel /a/ and the vowel /i/ was 11.6 sone and 7.3 sone, respectively, which were calculated in accordance with ISO-532B (ISO, 1975) using Zwicker *et al.*'s (1991) algorithm. The difference between these two averages was confirmed to be significant by Student's *t*-test [$t(68) = 6.54, p < 0.001$].

Effect of voicing of the following consonant

The listeners responded more sensitively to a duration change when the vowel whose duration was to be judged was followed by a voiceless consonant than when it was followed by a voiced consonant. This tendency appears to be consistent with the observation of Klatt and Cooper (1975); their listeners tended to be less sensitive to temporal changes in segments with following words than to those without following words. Klatt *et al.* suggested that the effect found might be due to some sort of backward recognition masking (Massaro and Cohen, 1975). Accordingly, the subsequent event, the presence of the following word, might have disturbed the listeners' judgment of the target durations by overwriting new information onto the listeners' auditory storage before he/she had completed the processing.

The current effect of the following consonant may be partially interpreted by an explanation of this sort, however, a more basic or lower-level influence should also be taken into account because the effect in the current study spans a relatively short range, i.e., a segment period, while that in Klatt *et al.*'s study probably spans a word period. Looking at the changes in acoustic features on the vowel-to-consonant boundaries, greater changes are observed for the unvoiced-consonant cases than for the voiced-consonant cases, in terms of either F0, power, or spectrum. Therefore, we can reasonably assume that the offsets of vowels are perceptually more salient when they are followed by unvoiced consonants than when they are followed by voiced consonants. A greater perceptual effect can apparently be assumed for the temporal displacement of a boundary having a greater perceptual salience. This view is consistent with the previous finding that the perceptual salience for a temporal displacement of a segment boundary is high when there is a large loudness difference or jump at the boundary (Kato *et al.*, 1997).

2.4.3 Interactions between the original duration and the three factors

Table 2.3 shows the means and standard deviations of the original durations of the vowels tested, for each of the constituent levels of the three factors: position in a word, vowel quality, and voicing of the following consonant. These averaged values suggest that neither

Table 2.3: The averages (and standard deviations in parentheses) of the original durations of the tested vowels in milli-seconds, for each of the constituent levels of the three factors: position in a word, vowel quality, and voicing of the following consonant.

Position in a word	
First mora	Third mora
77.1 (23.4)	96.3 (22.8)
Vowel quality	
/a/	/i/
93.4 (22.0)	75.9 (25.7)
Voicing of the following consonant	
Voiced	Unvoiced
82.1 (28.0)	96.1 (12.9)

of the three factors is independent of the original duration. The difference in the average durations between two levels of each factor turned out to be significant as follows:

- temporal position within a word: first mora $<$ third mora
- vowel quality: /a/ $>$ /i/
- voicing of the following consonant: voiced $<$ unvoiced[12]

Therefore, one might argue that the observed effects of these three factors are not substantial, and that the current result could be interpreted solely by the factor of the original duration without considering the other three factors. This subsection will discuss this possibility.

Assuming the influence of the original duration, a shorter original duration was expected to yield a narrower absolute acceptable range. As a matter of fact, the reverse tendency was observed for the factor of vowel quality; the acceptable range was generally narrower for the vowel /a/, which has a longer inherent duration, than for the vowel /i/, which has a shorter inherent duration. A similar discrepancy was found between the expected effect

[12] Although the duration of a vowel or syllable nucleus in English is generally shorter when it is followed by an unvoiced segment than when it is followed by a voiced segment (House, 1961; Peterson and Lehiste, 1960), a vowel duration in Japanese has no clear tendency of this sort. Even the reverse tendency has been reported, by Campbell and Sagisaka (1991); they showed that vowels followed by unvoiced consonants are longer than those followed by voiced consonants, as a result of a statistical analysis on a spoken Japanese database comprising 503 sentences.

of the original duration and the observed effect in the case of the voicing of the following consonant. Therefore, these two factors may not be interchangeable with the factor of original duration.

The positional effect, on the other hand, appeared to be in agreement with the expected effect of the original duration. To examine this possibility specifically, the obtained vulnerability indices (αs) were separated into four subsets corresponding to four cells constituting the combination of the factors of vowel quality and voicing of the following consonant, i.e., (/a/, /i/) $\times$ (voiced, unvoiced). As a result, no factors other than the position and original duration had any influence within each of these subsets. We then applied a linear model onto the α score with either the position in a word or the original duration as the explanatory variable, for each subset. The results showed that the position in a word correlated more highly with the α score than the original duration in all subsets. These results support the view that the positional factor is a more plausible explanatory variable than the original duration.

However, another problem then arises. As the effects of the three factors other than the original duration are large, they may obscure a genuine effect of the original duration. To test this possibility, we computed correlations between the original duration and the α score within each of the eight conditions defined by the factors of position in a word, vowel quality, and voicing of the following consonant. The results showed that there is no condition where the original duration has a significant effect in terms of both linear correlation and rank correlation as shown in Table 2.4.

Although these discussions may not be enough to reject the possibility of an influence of the original duration, they are enough to show that the difference in the original duration can not provide a consistent explanation of the observed tendencies in the vulnerability indices.

2.5 Summary

The acceptability of temporal unnaturalness was measured for word stimuli in which one of the vowel durations was systematically changed. The changes in acceptability depended on at least three factors: (1) the temporal position of the modified vowel within the word, (2) the quality of the modified vowel, and (3) the voicing of the consonant preceded by the modified vowel. For a given amount of modification, the listeners evaluated the modification of vowels at the third mora position in a word as more acceptable than the modification of vowels at the initial mora position in the word. They also evaluated the modification of /i/

Table 2.4: Correlation coefficients between the original duration and the vulnerability index (α) for each of the eight conditions defined by the three factors other than the factor of the original duration, i.e., position in a word (1 or 3), vowel quality (a or i), and voicing of the following consonant (v or uv), in terms of either Pearson's product-moment correlation or Spearman's rank correlation analysis.

Condition	Pearson's r	Spearman's ρ
1-a-v	0.043	0.055
1-a-uv	-0.092	-0.052
1-i-v	-0.10	-0.075
1-i-uv	0.082	0.0066
3-a-v	0.067	0.038
3-a-uv	-0.19	-0.10
3-i-v	0.016	0.023
3-i-uv	0.031	0.029

segments as more acceptable than that of /a/ segments, and they evaluated the modification of vowels followed by a voiced consonant as more acceptable than those followed by an unvoiced consonant. The original duration of the target vowel itself had no systematic effect on the rate of the acceptability change.

The primary impact of the current study is that the least vulnerable vowel, the third /i/ followed by a voiced consonant, required 140 % of the temporal modifications or durational errors required by the most vulnerable vowel, the first /a/ followed by an unvoiced consonant, to yield the same acceptability decrement. These results suggest that the traditional measure for durational rules, the average acoustic error, is insufficient in terms of perceptual evaluation, i.e., acceptability rating. We expect a more valid (closer to human perception) measure for durational rules to be achieved by taking into account the perceptual factors suggested in the present study as weighting values in the error evaluation.

Chapter 3

Acceptability of temporal modification in special mora segments[1]

Abstract

We examined effects of the phonetic quality type (e.g., vowel or fricative) and duration (single or double length) on human's perceptual acceptability of temporal modification given to steady-state speech portions including special moras[2] *such as a moraic nasal, a devoiced vowel, a geminate obstruent, and a long vowel. In experiment 1, the effect of phonetic quality of four types, i.e., vowel, nasal, voiceless fricative, and silence, on the acceptable modification range was tested. Six listeners evaluated the temporal acceptability of each of 49 words where one of the steady-state portions was subjected to durational modification from* -75 *ms (for shortening) to +75 ms (for lengthening) in 7.5 ms steps. The results showed that the listeners' acceptable modification ranges were narrowest for vowels, and widest for voiceless fricatives and silent closures, with nasals in between. The mean acceptable ranges for the least vulnerable phonetic quality types, i.e., voiceless fricative and silence, reached 143 % or more of that for the most vulnerable type, i.e., vowel. The observed variation in the acceptable modification range due to the different phonetic quality types was highly correlated with the inherent loudness in each phonetic quality type. A larger inherent loudness yielded a narrower acceptable range. Experiment 2 tested the effect of the original, as produced, duration of steady-state speech portions using 30 words where the factors of phonetic quality and original duration were designed in a factorial way. The results showed that the original durations affected the listeners'*

[1] This chapter is based on Kato, Tsuzaki, and Sagisaka (1998b) and Kato, Tsuzaki, and Sagisaka (a).

[2] Included are any of the non-initial moras of a (super-) heavy syllable (Crystal, 1997; Kubozono, 1999). Although a devoiced vowel seems to not be, in general, regarded as a special mora, we included it in the current experiments because its temporal structure is different from that of a regular (CV or V) mora and common with that of other "proper" special moras in being without a vowel onset.

absolute acceptable ranges; the ranges were narrower for shorter original durations. There was a significant interaction between the factors of phonetic quality and original duration. The effect of the original duration was larger for vowel portions than for fricative portions. This interaction could be accounted for by the difference in the temporal structure spanning beyond the modified portion itself.

3.1 Introduction

Rules to assign segmental durations have been proposed for speech synthesis to replicate the segmental durations of naturally spoken utterances (Allen *et al.*, 1987; Bartkova and Sorin, 1987; Carlson and Granström, 1986; Campbell, 1992; Fant and Kruckenberg, 1989; Higuchi *et al.*, 1993; Kaiki and Sagisaka, 1992; Klatt, 1979; van Santen, 1994; Takeda *et al.*, 1989). The segmental durations provided by these rules generally have a certain amount of difference from the corresponding, naturally spoken durations. The effectiveness of a durational rule should be evaluated by how much such a difference is perceptually salient. With almost all previous rules, however, the average of the absolute difference, or error, of each segmental duration from its reference had been adopted as the measure of any objective evaluation.

In previous studies (Kato *et al.*, 1997; Kato *et al.*, 1998a), we pointed out a possible problem with this measure, i.e., it gives every segment the same weighting in evaluating errors. That is, it neglects the following factors which possibly affect the perceptual sensitivity to segmental durations: (1) interactions among errors in different segments, and (2) variations in segment attributes and phonemic contexts. Kato *et al.*, (1997) examined the first factor and demonstrated that two durational errors occurring in consecutive vowel (V) and consonant (C) segments can be perceptually compensated by each other. Following that, Kato *et al.* (1998) focused on the second factor and revealed that perceptual sensitivity to durational errors is affected by a segment's temporal position in a word, its vowel quality, and the voicing of the following segments. This second study, however, tested only phonemically short vowels in words consisting of an open and light syllable succession, i.e., CVCVCVCV. The current study, therefore, continues the investigation of the second factor and expands the variety of segments or speech portions to be tested in the following two aspects: (1) phonetic quality—consonantal portions are included in addition to vowel portions, and (2) duration—phonemically long portions are included in addition to short vowels.

3.1.1 Influence of phonetic quality on temporal sensitivity

Typical examples of the dependency of temporal sensitivity on the phonetic quality can be found when comparing the durational discriminability between vowel and consonant segments. Both Huggins (1972) and Carlson and Granström (1975) reported that just noticeable differences (jnd's) for segmental durations are smaller for vowels than for consonants. Huggins manipulated the duration of a vowel /ɔ/ or a consonant /m, l, p, or ʃ/ embedded in a

naturally spoken sentence, and asked his listeners to judge whether it was “normal”, “too long”, or “too short”. The listeners were more sensitive to changes in the vowel duration than to changes in the consonant duration. Carlson and Granström employed a discrimination task for the differences in the duration of a vowel /a/ or a consonant /m, s, or t/ in isolated words, and found that their listeners’ discriminability was remarkably higher for the vowel duration than for the consonant duration. Similarly, Bochner *et al.* (1988) reported that their listeners demonstrated greater acuity for changes in the durations of vowels (/i, ɪ, u, ʊ, a, ʌ/) than consonants (/p, t, k/).

The results of a study by Fujisaki, Nakamura, and Imoto (1975), however, appear to contradict these three studies. Fujisaki *et al.* reported that durational jnd’s are almost equal for all phonetic quality types, regardless of whether they are vowels, nasals, voiceless fricatives, or silent closures. However, the task of their listeners in experiments was to judge the phonemic contrast depending on the segmental duration. This categorical judgment can obscure any potential effect of a difference in the phonetic quality (Carlson and Granström, 1975, commentary section). To summarize, therefore, previous studies have suggested a general tendency of durational jnd’s being shorter for vowels than for consonants, except in particular cases like when durational changes supply phonemic contrasts.

Although jnd’s are precisely defined in the psychophysical sense, “acceptability” can be considered as another practical (and more direct) measure in the evaluation of durational rules. This “acceptability” measure, however, has rarely been investigated, except for several pioneering studies (Carlson and Granström, 1975; Sato, 1977; Sagisaka and Tohkura, 1984; Hoshino and Fujisaki, 1983). Carlson and Granström made compensatory temporal modifications on a vowel-consonant pair /as/ and on two consonants /st/ in a word *plasta* (to cover something with plastic). The listeners evaluated the acceptability of each modification on a scale from zero (“acceptable”) to ten (“not acceptable”). The results showed that the acceptable ranges were narrower when the modifications included a vowel segment. The results therefore suggest that the durational change in the vowel influenced the acceptability more than that in the consonants. However, no direct comparisons were provided between the vowel and consonant segments. Neither did Sato’s nor Sagisaka *et al.*’s study compare the effects between vowels and consonants. Hoshino *et al.*, in contrast, did measure the acceptability of the modification in vowels and consonants separately, but they did not address the differences between vowels and consonants.

In summary, no direct experimental data exists showing the difference between vowels and consonants in evaluating the acceptability of durational changes. In particular, there

seems to be no comparative study on the effect of different consonants, whereas the influence of different vowel qualities on the evaluation of acceptability has been investigated (Kato *et al.*, 1998a). The first objective of the current study is, therefore, to provide a direct comparison among phonetic quality types in terms of the acceptability of durational modifications. Four phonetic quality types were chosen, i.e., vowel, nasal, voiceless fricative, and silence.

3.1.2 Influence of the original duration on temporal sensitivity

The current study also provides a comparison among different temporal properties of speech portions. The independent temporal variable to be tested is the duration of a speech portion that is acoustically continuous, and therefore, phonetically inseparable. In what follows, we refer to this portion as the "continuous portion" and to its duration as the "continuous duration."[3] This continuous portion starts with a segmental boundary and ends with either a segmental boundary or the end of the closure term in a stop articulation. It mostly coincides with a segment itself (e.g., a vowel or fricative) or a part of a segment (e.g., the closure of a stop), but it may span over linguistic (phonological and/or phonemic) boundaries. For instance, a geminate fricative as a whole is a continuous portion because it is phonetically inseparable; nevertheless, it can be phonologically separated into two parts by a syllabic boundary.

In a previous study, Kato *et al.* (1998) did not find any influence of the original duration or continuous duration on the perceptual sensitivity to durational modifications. However, all of their stimuli were taken from words having a homogeneous temporal structure (CVCVCVCV) and, therefore, the temporal variations of the tested portions were limited. The current study, in addition to using CV-syllable words, also uses words including much longer continuous portions, such as geminate obstruents and phonemically long vowels, and, naturally, there is a wide diversity in the stimulus duration. If a listener's acceptability judgment is based on the perceived distortion or temporal modification, and the temporal sensitivity roughly conforms to Weber's Law, i.e., proportional to the base duration, then a wider variation of the base duration should show the perceptual effect as more salient. That is, the second objective of the current study is to reexamine the influence of the original duration on the acceptability of durational modifications using a sufficient variation

[3] What "continuous portion" refers to is very similar to the "continuant sound" that was defined by Jacobson *et al.* (1954) in their distinctive feature theory, but differs from it in that the continuous portion may refer to the silent hold portion of a stop consonant; also, the continuant sound was implicitly defined as ranging within a single segment whereas the continuous portion may not.

of stimulus durations.

One should note that the duration of a continuous portion can also be described by discrete or linguistic measures, i.e., lengths of sounds quantized in terms of linguistic contrasts (e.g., the mora counting), instead of the continuous measure, i.e., the acoustic duration in milliseconds. Linguistic durations, however, are not always proportional to the corresponding acoustic durations, and even linguistic measures themselves are sometimes subject to controversy. Therefore, it is difficult to control both continuous (acoustic) and discrete (linguistic) duration measures simultaneously in designing experimental conditions.

In the current two experiments, we controlled the stimulus duration so as to have coherency in terms of the acoustic measure. This allowed us to estimate the extent of the accountability of psychoacoustical (i.e., non language-specific) factors. Such an auditory-based approach has a potential advantage in providing perceptually valid notions that can be generalized across languages. Note, however, that we do not discard the notion of discrete or linguistic duration measures but supply an extensive discussion (including them) later in the general discussion section.

3.2 Experiment 1: Effect of phonetic quality

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

3.2.1 Method

Material

The following four phonetic quality types were chosen from among types tested in previous jnd studies: 1) vowel, 2) nasal, 3) voiceless fricative, and 4) silence. To make our listeners focus on the temporal aspect of the material as much as possible, we chose test portions from among candidates whose durations were long enough, i.e., the phonemic quality of the test portions would not suffer from temporal manipulations over a reasonably wide range. If any phonemic quality changed, the listeners' judgments could no longer rely on a single criterion, i.e., the temporal cue. In these respects, the following groups of speech portions were employed for each of the quality types tested.

(1) Vowel type: Phonemically short /a/ vowels were used. The Japanese language has five vowels /a, i, u, e, o/ each of which has short and long phonemically contrastive

variations. The short /a/ vowel has a relatively long inherent duration along with /e/ and /o/ in comparison with the short /i/ or /u/ (Sagisaka and Tohkura, 1984) and, therefore, can provide a sufficient durational margin for manipulation.

(2) Nasal type and (3) voiceless fricative type: Since consonants in Japanese CV moras are usually shorter than those in the languages used in previous studies (e.g., English), it is difficult to manipulate their durations over a sufficient range to examine the temporal acceptability without any shift in the phonemic quality. Therefore, syllabic or moraic nasals and continuous portions including devoiced vowels were chosen for the nasal and voiceless fricative types, respectively, to obtain sufficient consonantal durations comparable to those tested in previous studies.

A moraic nasal (as in *Honda*) has the same phonological status as the CV-mora and has a comparable acoustic duration with a short vowel /a/. A devoiced vowel (as in *Hi̥tachi*) usually has a voiceless fricative quality like [ɸ], [s], [ʃ], [ç], [x], and [h]; in Japanese, short high vowels /i, u/ are mostly devoiced when they are placed between voiceless consonants (Tsujimura, 1996). When the preceding consonant is a voiceless fricative, a devoiced vowel continues from it, keeping the same phonetic quality. Note that the tested continuous portions were chosen from such concatenated pairs of a devoiced vowel and an adjacent voiceless fricative. In what follows, the term "devoiced vowel portion" refers to this concatenated, but continuous, voiceless fricative portion. A devoiced vowel portion, therefore, also has the same phonological status as the CV-mora and has a comparable acoustic duration with a short vowel /a/.

(4) Silence type: The silent targets were chosen from the closures of geminate voiceless stops /pp, tt, kk/ (as in *Sapporo*) which have considerably longer silent closures than their single counterparts.

In addition to these four groups of speech portions, we included a fifth group of test portions, i.e., geminate fricatives. While the acoustic durations of vowels /a/, moraic nasals, and devoiced vowel portions are comparable with each other, those of silent closures in geminate stops are generally longer than the other three (150 % or more). This means that there might be another explanatory variable, i.e., the base duration, in addition to the primary explanatory variable, i.e., the phonetic quality. The geminate fricative /ss/ (as in *Nissan*) was, therefore, chosen for the fifth group to probe the influence of the base duration. The duration of the geminate fricative /ss/ is, in general, comparable with those of the longer test

portions, i.e., silent closures in geminate stops, and its phonetic quality type matches one of those of the shorter three groups, i.e., voiceless fricative type.

Stimulus manipulation

The speech database from which the original materials were taken and the method of stimulus manipulation were the same as those in our earlier papers (Kato *et al.*, 1997; Kato *et al.*, 1998a). Forty-nine words were selected from the ATR speech database (Kurematsu *et al.*, 1990). All of them were commonly used four-mora Japanese words.[4] The selected words were spoken naturally in isolation by one male speaker and were digitized at a 12 kHz sampling frequency with 16-bit precision.

The continuous portions whose durations were subjected to modification ("target portions") were chosen from the second moraic position in the words. This positional condition was introduced to prevent the influence of the temporal position in the word on the acceptability evaluation, which was reported in a previous study (Kato *et al.*, 1998a), as much as possible. Each duration of these target portions was manually measured from the spectrographic images of the original materials by professional phoneticians. The 49 selected tokens are listed in Appendix A (Table A.2) with the phonetic quality and measured continuous duration of each of the target portions. Table 3.1 summarizes the number of target portions and the average and standard deviations of the continuous durations for each of the stimulus groups.

The temporal modifications were made by a cepstral analysis and resynthesis technique using a log magnitude approximation (LMA) filter (Imai and Kitamura, 1978), carried out at 2.5 ms frame intervals. The durational changes were achieved by deleting or doubling the synthesis parameters frame by frame. Each of the target portions was shortened or lengthened over a range from -75 ms to $+75$ ms from the original duration in 7.5 ms steps, resulting in 20 different modification steps. Preliminary listening to all of the manipulated stimuli assured us that no phonemic shift had occurred in either the target portions or the surrounding phonemes. All of the stimuli were produced by a computer (SPARC Station 10, Sun Microsystems) at a 12 kHz sampling frequency with 16-bit precision. In total, 1029 word stimuli were prepared; i.e., (20 modification steps + 1 unmodified) $\times$ 49 portions.

[4] To maximize the freedom in word selection, we chose the materials from four-mora words which are lexically the most frequent in contemporary Japanese (Hashimoto, 1973; Yokoyama, 1981).

Table 3.1: The number of selected continuous portions for each of the stimulus groups in experiment 1, and the averages and standard deviations of their acoustic durations.

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

Procedure

The experimental procedure was the same as that in a previous study (Kato *et al.*, 1998a). The stimuli were randomized and recorded onto a digital audiotape (DAT) through a D/A converter (MD-8000 mkII, PAVEC) and a low-pass filter (FV-665, NF Electronic Instruments, $f_c = 5700$ Hz, -96 dB/octave) with a DAT recorder (DTC-55ES, SONY), and then presented diotically to the subjects through headphones (SR-Λ Professional, driven by SRM-1 MkII, STAX). A four-second interval was inserted after each presentation for the subjects' response. The average presentation level was 73 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 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 subject with a sound level meter (Type 2231, Brüel & Kjær) and a condenser microphone (Type 4155, Brüel & Kjær).

The subjects were told that each stimulus word was possibly subjected to temporal modification. Their task was to evaluate how acceptable each stimulus was as an exemplar of the token of that stimulus, using a seven-point rating scale ranging from -3 to 3, where -3 corresponded to "quite acceptable" and 3 corresponded to "unacceptable."[5] The subjects

[5]If the listeners were asked to rate the "naturalness," they might have tended to use a strict criterion making it difficult for an informative evaluation to be maintained for the whole range of temporal modifications to be tested. To obtain information for a reasonably wide range of modifications, therefore, we chose the "rating of acceptability" over the "rating of naturalness."

were asked to respond regarding only the temporal aspects of the stimuli, as much as possible.

A total experimental run for each subject comprised of seven sessions. Seven of the 49 tokens were chosen for each session, and four repetitions of their 21 modified versions were randomly presented in the session. Therefore, each subject evaluated each stimulus four times in total. The seven tokens within a session were chosen from all of the five stimulus groups.

Subjects

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

3.2.2 Results

Measure of acceptability

The acceptability measure, referred to as the "vulnerability index," was the same as that used in a previous study (Kato *et al.*, 1998a) to maintain consistency among the studies. To compute the vulnerability index, we first plotted the listeners' evaluation scores against the change in duration of the portion in question, and then a parabolic regression method as generally formulated below was applied to the plot ,[6]

$$Evaluation\ score = \alpha(\Delta T - \beta)^2 + \gamma, \tag{3.1}$$

where ΔT denotes the change in duration; the unit of ΔT is not the relative duration but milliseconds. This regression was the best fitted polynomial function to the plot on the basis of the *F*-ratio criterion $[F(2, 6171) = 1408.0, p < 0.0001]$ (e.g., McCall, 1980). The coefficient of the second-order term or α of this parabolic curve was taken as the "vulnerability index," the objective variable of this study. It shows the rate of change in the evaluation score with a change in the durational modification; both the horizontal and vertical response biases can be separated out as β and γ. As derived from Equation 3.1, α serves the average decrement[7] of the acceptability for a certain temporal modification size, and also the width between the longer and shorter limits of the temporal modification, which

[6] Although we could choose fitting functions other than polynomial fittings and/or could rescale the vertical axis to obtain an interval scale, we adopted the parabolic fitting on the raw evaluation scores owing to the advantage of its directly reflecting the subjects responses and its goodness of fitting.

[7] As we assigned a larger evaluation score to a more unacceptable or less acceptable impression, an increment of the evaluation score implies a decrement of the acceptability.

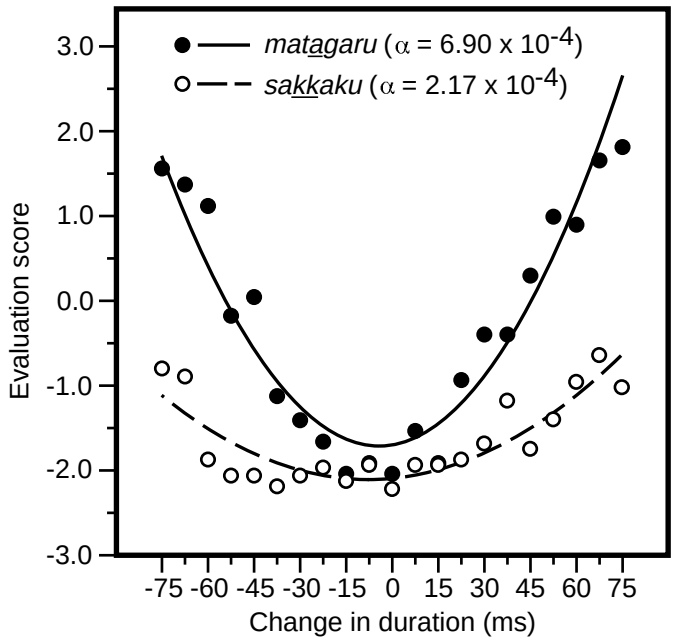

Figure 3.1: An example illustrating a difference in acceptability-change between two different speech portions. Those portions subjected to durational modification are underlined in the legend. The scatter plots show that the evaluation score varies according to the durational change more drastically for the second vowel of the word "*matagaru* (to ride)" (filled circles), than for the silent closure of the geminate stop consonant in the word "*sakkaku* (illusion)" (open circles). The two regression curves trace this tendency. These are formulated as: $y = 0.000690(x + 4.61)^2 - 1.71$ (solid line), and $y = 0.000217(x + 7.63)^2 - 2.10$ (dashed line).

yields a certain amount of acceptability decrement, i.e., an acceptable range. Therefore, α represents the vulnerability of a given continuous portion from the temporal modification. Figure 3.1 shows typical examples of individual fittings illustrating the difference in the acceptability change between two different test portions.

We applied a parabolic fitting to the evaluation scores for each of the 49 target portions per each of the six subjects, obtaining 294 fitting curves. Prior to the statistical analyses, we dropped unreliable data (fitting curves) on the basis of two criteria: (1) when α was not positive the fitting curve was dropped; because this suggests that the subject probably sensed no durational change for that particular token, and (2) when the axis was extremely remote (more than six times sigma) from the distribution center of the entire data the fitting curve was dropped. In all, seven fitting curves were excluded, i.e., five eliminations by the

first criterion, one elimination by the second, and one elimination by both criteria, resulting in 287 fitting curves. Therefore, the dependent variable of experiment 1 consisted of 287 α scores.

Effect tests

The effect of *phonetic quality type* on the vulnerability index (α) was tested by a one-way ANOVA of repeated measures with *subject* as the blocking factor. The effect of *phonetic quality type* was found to be significant $[F(4, 20) = 53.1, p < 0.001]$. As shown in Fig. 3.2, α was greatest for the vowels, next for the nasals, third for the fricatives, and smallest for the silent portions and fricatives in geminate consonants. Multiple comparisons among these average αs using Tukey–Kramer's HSD (the honestly significant difference) (SAS Institute Inc., 1990) indicated the difference between any two average αs to be significant $[p < 0.01]$, except for the difference between the average αs of the geminate fricative and silence groups.

The averaged α score of the most vulnerable (i.e., susceptible to durational modification) quality, i.e., the vowel, became more than twice that of the least vulnerable quality, i.e., the voiceless fricative or silence. This means that the least vulnerable quality type required more than 143 % of the temporal modification of the most vulnerable quality type to yield the same acceptability decrement.

3.2.3 Discussion

The primary objective of the current study was to examine whether the acceptability for the temporal modification of continuous portions is affected by the phonetic quality type of modified portions. The results of experiment 1 showed significant effects due to the difference in the phonetic quality type. The listeners evaluated the temporal modifications of vowel portions as less acceptable than those of consonant portions regardless of whether they were nasals, voiceless fricatives, or silent closures. This tendency is in agreement with that predicted from literature on jnd's for vowel and consonant durations in English or Swedish (Bochner *et al.*, 1988; Carlson and Granström, 1975; Huggins, 1972a). In addition, the current experiment revealed nasals to be different from voiceless fricatives or silent closures.

On the other hand, some results could not be accounted for by the factor of *phonetic quality type*. A significant difference was observed between the average αs of the devoiced

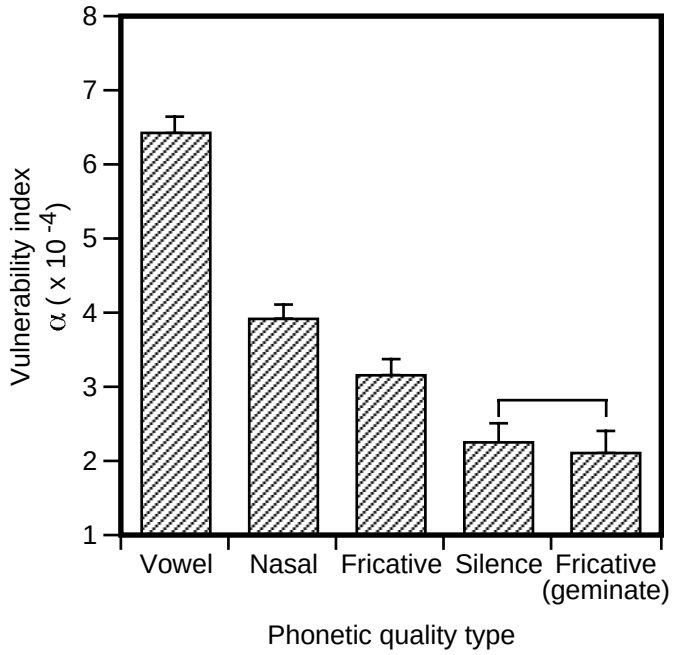

Figure 3.2: The least squares mean of the vulnerability index (α), i.e., the second-order polynomial coefficient of the fitting curve, for each phonetic quality type; they were calculated in the ANOVA procedure. The error bars show the standard errors. A larger α implies a narrower acceptable range. The difference between the two bridged bars is not statistically significant.

vowel portions and geminate stops; these two stimulus groups differed from each other in both the phonetic quality type and the original duration. The difference between the average αs of the geminate stops and the geminate fricatives was, on the other hand, not significant; these two stimulus groups only differed in the phonetic quality type. Therefore, the difference of the average αs between the devoiced vowel portions and the geminate stops is more likely due to their difference in the original duration than to their difference in the phonetic quality type.

However, a problem now arises. If the difference between the αs of devoiced vowel portions and geminate stops is due to their durational differences, a similar durational effect should be observed in any type of phonetic quality including the vowel type. However, such an effect of the original duration was not observed in a previous study (Kato *et al.*, 1998a) which measured α scores for vowel segments using the same procedures as those of the current study. Note, however, that the stimulus condition of the previous study differed

from those of the current experiments. The previous study used only short vowels in a homogeneous syllable context, i.e., CVCVCVCV, and therefore the durational variation of the target segments was limited to a shorter range (less than 150 ms) than that used in the current study (87.5 – 220 ms). Experiment 2 was, therefore, designed to examine whether the effect of the original duration would be generally observed for similarly ranging portions as those of experiment 1, irrespective of differences in the phonetic quality type.

3.3 Experiment 2: Effect of the original duration

Experiment 2 systematically examined the effect of the original continuous duration on the acceptability of temporal modification in a given portion using two phonetic quality types, i.e., vowel and voiceless fricative.

3.3.1 Method

Design

A two-way factorial design was applied. The first factor was the *phonetic quality type* of a continuous portion subjected to temporal modification. There were two levels in this factor, i.e., vowel and voiceless fricative. The second factor was the *original duration* of the portion in question. There were also two levels in this factor, i.e., short and long. The target portions for the short and long levels in the vowel type were chosen from phonemically short and long vowels, respectively. Those for the short and long levels in the voiceless fricative type were chosen from devoiced vowel portions and geminate obstruents, respectively. Other phonetic quality types, e.g., nasal or silence, were ignored as they are unlikely to provide a comparable extent of durational variations in Japanese speech.

Material and procedure

Thirty four-mora Japanese words were selected as the original materials from the same database as in experiment 1. The target portions included ten short and ten long vowels, and five short and five long voiceless fricatives. The vowel quality of the vowel target portions was either /a/ or /o/. Five phonemically short /a/ materials were taken from among those used in experiment 1 in accordance with a criterion, i.e., that their α scores had been the five nearest to the median within that stimulus group. Five phonemically short /o/ materials were also chosen. For the long vowel materials, three phonemically long /a/ and seven

Table 3.2: The number of selected continuous portions for each of the stimulus groups in experiment 2, and the averages and standard deviations of their acoustic durations.

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

phonemically long /o/ materials were chosen, i.e., ten /a:/ and /o:/ portions in total. The reason why the stimuli were taken from vowels other than /a/ vowels was that the source speech database did not include a sufficient number of long /a/ materials. A phonemically long /a/ is not lexically so frequent in Japanese. The vowel quality /o/ was chosen as a substitute because its inherent loudness, which has been suggested to affect the temporal acceptability (Kato *et al.*, 1998a), is the closest to the inherent loudness of vowel quality /a/ among the four other vowel qualities in Japanese.

The original materials for the voiceless fricative type were a subset of those used in experiment 1. Five devoiced vowel portions and five geminate fricatives were taken according to the same criterion as the short /a/ case mentioned above. To reduce the size of the experiment and prevent the subjects from unnecessary strain, a smaller number of tokens were taken for the fricative groups (five per group) than for the vowel groups. Relatively stable responses had been expected for the fricative groups because their tests were replications of those in experiment 1. The 30 selected tokens are listed in Appendix A (Table A.3). Table 3.2 summarizes the number of target portions and the average and standard deviation of the continuous durations for each of the stimulus groups.

The speaker of the original materials, the recording procedure, the manipulation method, and the procedure for the experimental run were the same as those in experiment 1.

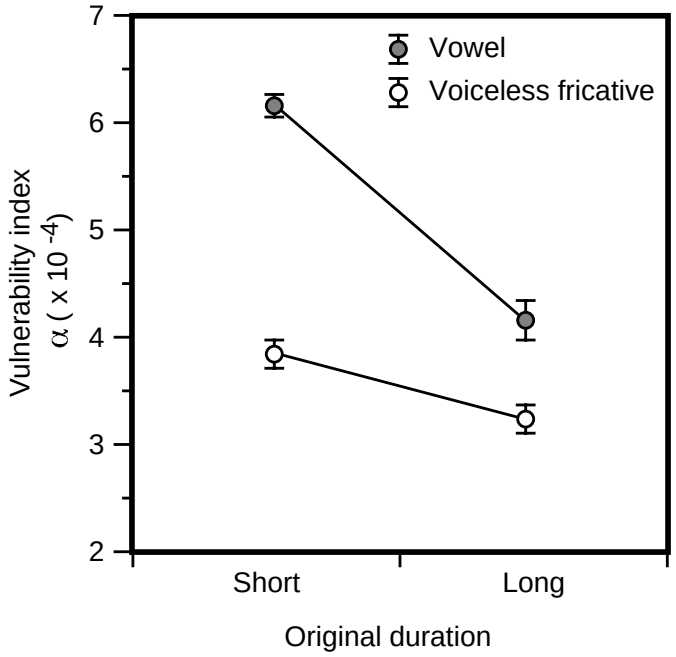

Figure 3.3: The least squares mean of the vulnerability index (α), i.e., the second-order polynomial coefficient of the fitting curve, for each stimulus group; they were calculated in the ANOVA procedure. The error bars show the standard errors. A larger α implies a narrower acceptable range.

Subjects

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

3.3.2 Results

In accordance with the same procedures as in experiment 1, the vulnerability index (α score) was computed for each of the 30 target portions and each of the nine subjects, resulting in 270 αs. A two-way factorial ANOVA of repeated measures was performed with *phonetic quality type* and *original duration* as the main factors, and with *subject* as the blocking factor. The main effects of *phonetic quality type* and *original duration* were significant [$F(1,8) = 51.9, p < 0.0001$; $F(1,8) = 67.0, p < 0.0001$, respectively]. As shown in Fig. 3.3, α was greater for the vowels than for the voiceless fricatives, and similarly greater for the short targets than for the long targets. There was a significant interaction between both

main factors (i.e., *phonetic quality type* and *original duration*) [$F(1,8) = 14.7, p < 0.005$]; the effect of *original duration* was larger for the vowels than for the voiceless fricatives. Multiple comparisons among the average αs of four stimulus groups using Tukey–Kramer's HSD indicated the difference between any two average αs to be significant [$p < 0.05$], except for the difference between the αs of the long vowel and short voiceless fricative (devoiced vowel) portions.

To summarize the results, an effect of the phonetic quality type similar to that in experiment 1 was replicated in experiment 2. The effect of the original duration for the voiceless fricatives was also replicated. It was additionally found that the original duration also affects the temporal vulnerability of vowel portions.

3.4 General discussion

This section tried to relate the acceptability measure, a perceptual measure of changes in phonetic durations, to measures of human sensitivity against changes in non-speech durations. We thought it necessary to include this psychophysical discussion because it is important to estimate the extent to which conclusions and predictions based on the current study may be generalized. If they are psychophysically accountable, then we can infer what should happen in unknown languages or, e.g., other phonetic qualities that have not been actually tested.

3.4.1 Effect of phonetic quality

Differences in the phonetic quality type affected the acceptability of the durational modification in both experiments 1 and 2. We chose loudness [8] as the candidate variable to represent differences in the phonetic quality type from among many psychoacoustical features of the target speech portions. A previous study had shown that the acceptability of modification in a vowel duration correlates with the loudness inherent in each vowel quality (Kato *et al.*, 1998a).

To estimate the inherent loudness for each phonetic quality type, we first calculated the loudness contour for each of the 49 target portions in experiment 1 every 2.5 ms with a 30 ms window in accordance with ISO 532B (ISO, 1975) using Zwicker *et al.*'s (1991) algorithm.

[8] Any usage of the word "loudness" in the current study means the loudness calculated by ISO-532 method B, unless otherwise stated. Although ISO-532B does not always provide excellent approximations for non-steady-state signals like speech, we adopted this method due to the advantage of its psychophysical basis instead of adopting power or intensity which incorporates no psychophysical considerations.

Then, we picked up the median value from the entire range of each of the target portions as the representative loudness of that portion. Figure 3.4 shows these representative loudness values pooled for each phonetic quality type. Multiple comparisons among the phonetic quality types using Tukey–Kramer's HSD clearly indicated the difference between any pair of pooled loudness values as significant $[p < 0.001]$. In fact, these loudness values, i.e., the estimated inherent loudness values, highly correlated with the phonetic quality difference ($r = 0.979$).

Furthermore, interestingly, the order of the phonetic quality types by these loudness values was identical to that by the vulnerability indices (see Fig. 3.2) except for the relation between the voiceless fricative and silence types. Therefore, these loudness values also have to correlate with the vulnerability index (α). The Pearson product-moment correlation coefficient (r) between the representative loudness and the α score based on the 49 target portions was 0.889. This accountability of the loudness for the vulnerability index was comparable with that of the phonetic quality type where r was 0.888.

These facts suggest that the loudness measure can be a good index to predict differences in the temporal vulnerability of speech portions due to differences in the phonetic quality. However, one should not generalize this accountability to any phonetic quality other than those tested in the current study before confirming the psychophysical validity of the correlation between the stimulus loudness and the acceptability evaluation. We, therefore, tried to validate it according to the following two steps: the first step examined the correlation between acceptability and temporal sensitivity, and the second step examined the correlation between temporal sensitivity and loudness.

The first correlation seems to be plausible because the evaluation of the acceptability has to be based on the distortion detected by listeners. That is, the durational jnd determines the baseline of the acceptable range. To support this notion, there are at least two examples showing the correlation in question. First, the influence of the phonetic quality found in previous jnd studies (Bochner *et al.*, 1988; Carlson and Granström, 1975; Huggins, 1972a) is generally in agreement with that of the current acceptability study; i.e., the listeners respond more sensitively to modifications in vowels than to those in consonants. Secondly, although the number of word tokens was not very large, a correlation has been reported between the vulnerability index (α) and durational jnd as a result of a direct comparison using the same speech materials and the same listeners (Kato *et al.*, 1992).

As for the second correlation, little evidence seems to be given by literature. Quantitative models dealing with the auditory acuity of filled durations, as long as the range of speech

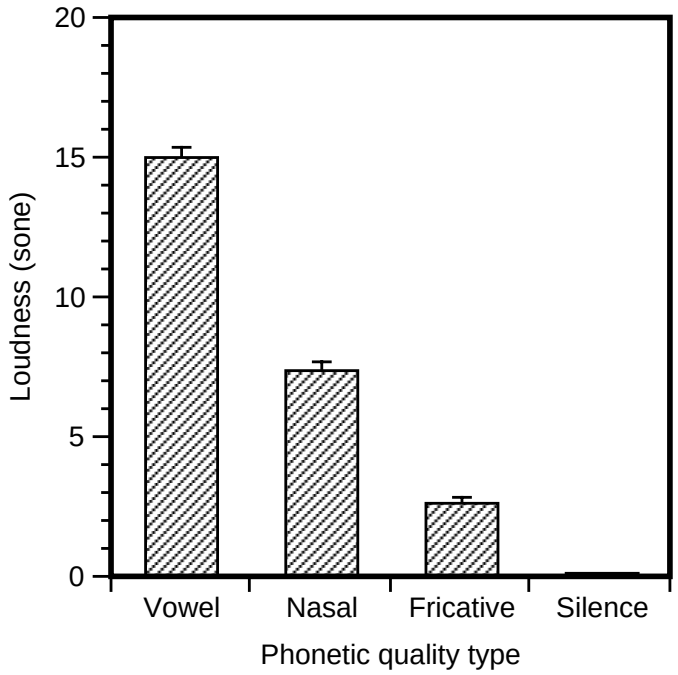

Figure 3.4: The average loudness of those speech portions whose durations were subjected to temporal modification in experiment 1, as a function of the phonetic quality type. For the *silence* type, the background noise level was adopted. The error bars show the standard errors.

segments (about 50–300 ms) is involved, have not taken into account stimulus loudness or intensity (Allan, 1979; Allan and Kristofferson, 1974). Additionally, experimental data has not appeared to support the relation between duration discrimination and stimulus intensity (Abel, 1972b; Henry, 1948; Rammsayer, 1994). In these cases, although some intensity dependency has been observed during target stimulus presentation at an extremely low level (Henry, 1948) or under a low S/N condition (Creelman, 1962), no intensity effect has been found for a clearly audible stimulus.

However, it is important to point out that all of these studies presented the target signals only in isolation, while a segment in speech generally has preceding and/or succeeding sounds, i.e., adjacent segments. The target portions in experiment 1 were also of the same case, because they were placed at the second moraic position within the four-mora words. Therefore, it appears necessary to examine the intensity effect under the condition the target signals are temporally flanked by other signals, in addition to the traditional isolated

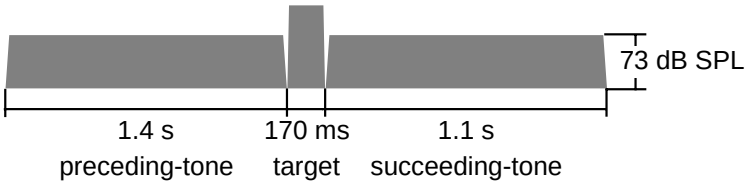

Figure 3.5: Schematic examples showing the amplitudes of stimulus sequences used in the temporal discrimination test in Kato and Tsuzaki (1994). The horizontal and vertical axes refer to the time and level, respectively. All stimuli were 1 kHz tones. The level of the target portion was either 79, 76, 70, 67, 55 dB SPL, or silence. (Reproduced from Kato and Tsuzaki (1994).)

presentation.

There is one example that deals with auditory temporal sensitivity under the above mentioned stimulus conditions. Kato and Tsuzaki (1994) demonstrated the intensity effect on temporal discriminability for auditory durations surrounded by other sounds. Their stimuli were 1 kHz pure tones having an amplitude contour as shown in Fig. 3.5. The level of the target part whose duration was subjected to temporal modification was either 79, 76, 70, 67, or 55 dB SPL, corresponding to about 15 to 2.8 sone in loudness, or silence. The level and duration of the preceding and succeeding tones were fixed. The measured temporal jnd was correlated highly with the level of the target part as shown in Fig. 3.6. This figure was reproduced from Kato and Tsuzaki (1994) to clarify the difference among loudness categories, with each corresponding to the inherent loudness of each phonetic quality type tested in experiment 1. That is, 12–15, 6.5–8, and 2.8 sone referred to the loudness of the vowel, nasal, and voiceless fricative portions, respectively [see also Fig. 3.4]. These results can be considered as evidence for the second correlation. Therefore, the accountability of loudness in the observed effect of the phonetic quality seems to be valid from the psychophysical viewpoint.

3.4.2 Effect of the original duration

The effect of the original duration on the acceptability of durational modification was observed for voiceless fricatives in experiments 1. A similar effect was observed for both vowels and voiceless fricatives in experiment 2. A larger vulnerability index, i.e., a narrower acceptable modification range, was yielded for portions having a shorter original duration. This tendency seems to be reasonable in the light of a general psychophysical law, i.e.,

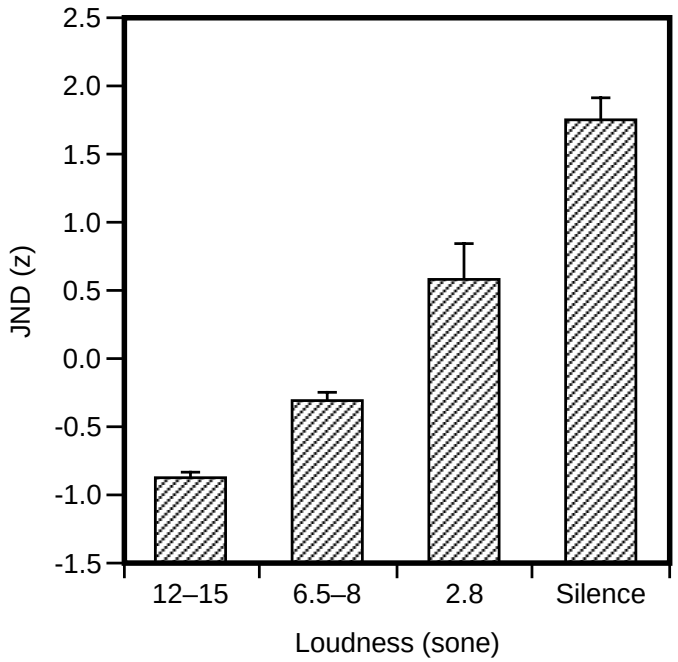

Figure 3.6: Results of the temporal discrimination test in Kato and Tsuzaki (1994). Normalized just noticeable differences are shown. They are pooled over six subjects as a function of the level (translated to the loudness measure) of the target portion. Each category of the loudness measure roughly corresponds to the average loudness value of the vowel, nasal, voiceless fricative, or silence portions tested in the current study. The error bars show the standard errors. (Reproduced from Kato and Tsuzaki (1994).)

Weber's Law. Conforming to this law, a larger physical change is necessary for a longer base duration, i.e., the original continuous duration, to yield the same amount of perceived change. Note, however, that the acceptable range of a target portion, which is derived from α, is not exactly proportional to the corresponding original continuous duration, although they positively correlate with each other. The ratio of two acceptable ranges is considerably smaller than that of the corresponding original durations.

3.4.3 Interaction between phonetic quality and the original duration

In experiment 2, a significant interaction was found between the factors of *phonetic quality type* and *original duration*. The effect of *original duration* was larger for the vowel type than for the voiceless fricative type. As seen in Table 3.2, the difference in the average acoustic duration between the 'short' and 'long' groups is slightly larger for the vowel type.

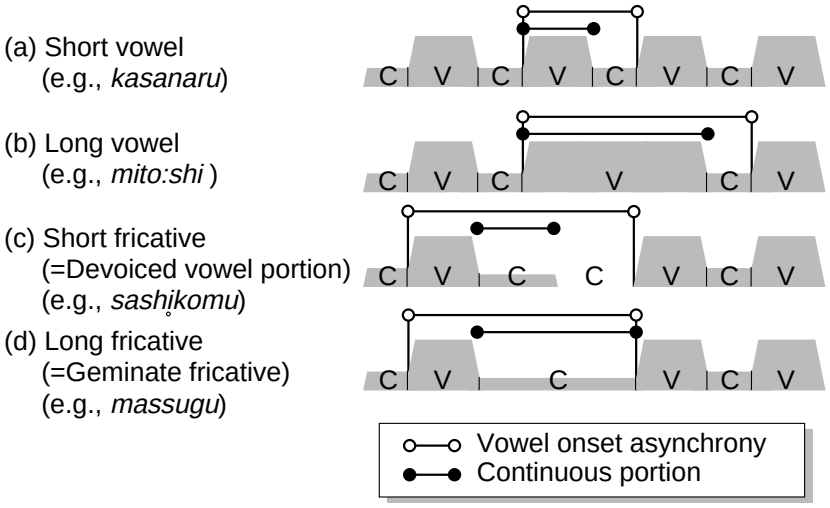

Figure 3.7: Schematic examples showing the temporal structures of four-mora Japanese words for each stimulus group in experiment 2. The horizontal and vertical axes roughly refer to the time and loudness, respectively. "C" and "V" represent consonant and vowel portions, respectively. Note that the temporal alignment of each segment is highly idealized in these examples and that such rigid isochronous relations are rarely observed in actual Japanese speech.

Nevertheless, the durational contrast in the vowel groups is not sufficiently larger than that in the voiceless fricative groups to account for the observed interaction on the basis of Weber's Law.

An alternative source, therefore, should be taken into account for this interaction. We consider, as a possible candidate, temporal cues that span beyond the target portion itself. In evaluating the temporal acceptability, the listeners might use the relative timings among the multiple portions or syllables surrounding the target portion. A major cue forming the perceptual timing of speech has been suggested to be the interval between vowel-onsets or vowel-onset asynchrony (VOA) by Sato (1977) through an observation of the production process; this was, then, empirically confirmed by Kato *et al.* (1996). The contribution of vowel onsets to the perceived timing has also been reported for English speech (Allen, 1972b; Morton *et al.*, 1976).

To examine the role of VOA cues, we schematically illustrated the temporal structures of the speech materials used in experiment 2 and marked, thereon, the target portions and their VOAs (Fig. 3.7). As clearly seen in this figure, the VOA spanning over the short vowel

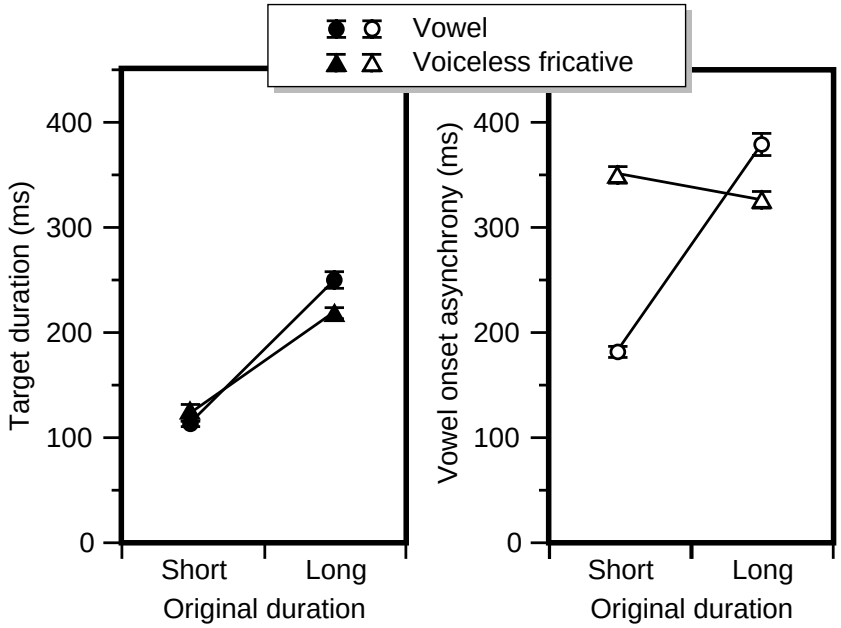

Figure 3.8: The average and standard errors of the measured durations of target portions in experiment 2, and the measured intervals between the two vowel onsets (VOA) surrounding the target portions, as a function of the categorized target duration (short or long), i.e., the original duration. The short–long contrast of the original duration yields a much larger difference in VOA for the vowel type stimuli (open circles) than that for the voiceless fricative type stimuli (open triangles), while the difference in the target duration by the same contrast of the vowel type stimuli (closed circles) is similar to that of the voiceless fricative type stimuli (closed triangles).

(the vowel in the CV-mora) seems to be considerably shorter than that over the long vowel, whereas the VOA over the short fricative (devoiced vowel portion) is comparable with that over the long fricative (geminate fricative). Acoustic measurements of the actual stimulus words confirmed that the same tendency was found in the materials of experiment 2 as summarized in Fig. 3.8. Whereas a clear contrast in the VOA between the ‘short’ and ‘long’ groups was observed for the vowel type, no such tendency, or even the inverse tendency, was observed for the voiceless fricative type. Therefore, the observed interaction can be accounted for if we consider the difference in the VOA contrast as the source enlarging the effect of the original duration for the vowel type compared to the voiceless fricative type. These results suggest that the perceptual consequences of a given local modification or distortion may not solely be accountable by the change in the local duration itself but

Table 3.3: The mora counting and the number of dropped temporal markers (phoneme boundaries) compared with the intact alternation of short consonants and vowels, for each of the stimulus groups used in experiment 2. The short and long fricative groups refer to the devoiced vowel portion and geminate fricative groups, respectively. The sequential structures of the words embedding the tested portions are also shown along with example words. C and V stand for consonant and vowel segments. The subscript numerals show temporal moraic positions in a word. The mora boundaries are marked by hyphens. A devoiced vowel is marked with an under-ring. Those target continuous portions whose durations were subjected to temporal modification are underlined.

Stimulus group	CV sequence in the whole word	Example token	Mora counting	Number of marker droppings
Short vowel	C_1V_1-$C_2\underline{V_2}$-C_3V_3-C_4V_4	*ka-s<u>a</u>-na-ru*	1	0
Long vowel	C_1V_1-$C_2\underline{V_2}$-$\underline{V_3}$-C_4V_4	*mi-t<u>o-o</u>-shi*	2	2
Short fricative	C_1V_1-$\underline{C_2\mathring{V}_2}$-$C_3V_3C_4V_4$	*sa-<u>shi̥</u>-ko-mu*	1	1
Long fricative	C_1V_1-$\underline{C_2}$-$\underline{C_3}V_3$-C_4V_4	*ma-<u>s-s</u>u-gu*	1	2

also by the changes in the intervals among the widely distributed multiple cues whereby the timing or rhythm is supplied.

3.4.4 Coherency of durational effects with discrete measures

To introduce an alternative implication about the effects of the original duration and the interaction observed in experiment 2, this subsection attempts to apply discrete or linguistic measures to the durations of target speech portions, rather than a continuous one, i.e., the acoustic duration.

Mora counting

First, mora counting is examined. The notion of mora counting is used in phonology to handle the syllable weight, the relative durations of syllables when they are linguistically contrastive. The analysis of segments into moras is usually applied only to the syllabic nucleus (core vowel) and coda (final consonants), and not to the syllable onset (initial consonants), so that the presence or absence of an initial consonant does not change the mora counting or weight of a given syllable (Hyman, 1975; Kubozono, 1998). For instance,

both of the two types of light syllables, V and CV, are counted as monomoraic and the heavy syllables ((C)VV or (C)VC) are counted as bimoraic.

The mora, in several languages including Japanese, is frequently regarded as a basic unit of temporal regulation, as well as a unit by which the phonological distance is defined. If a quantity based on this unit were to be referenced as a base duration during the perceptual evaluation of temporal modifications, the observed durational effects would correlate with the mora counting of the target continuous durations. To test this possibility, we measured target continuous durations by mora counting and summarized them by stimulus groups in Table 3.3.

The duration of a short vowel, a vowel in a CV-mora, is counted as monomoraic; although this duration shares one CV-mora with its consonantal partner, the initial consonant does not contribute to the mora counting according to the definition. The duration of a long vowel is counted as bimoraic in a similar way. The duration of a short fricative, i.e., a devoiced vowel portion, is counted as monomoraic because this portion has the same phonological status as a CV-mora. The duration of a long fricative, i.e., a geminate fricative, is counted as monomoraic; although it consists of a moraic obstruent (Kubozono, 1999; Vance, 1987), which is the final consonant of the first syllable of the word ($C_1V_1C_2$; subscript numerals show temporal moraic positions in a word), and the following short obstruent, which is the initial consonant of the CV-mora (C_3V_3), the initial consonant (C_3) again does not contribute to the mora counting.

These measurements reveal a sort of incoherence between the mora counting and the observed effects. That is, the short–long contrast in the voiceless fricative type does not show any difference in terms of the mora counting while it shows a significant difference in the acceptability evaluation. This fact, accordingly, does not support the notion that mora counting plays a role as a base duration in the current perceptual evaluation of temporal modifications.

Deviation from intact C-V alternation

The orthodox mora counting given above did not succeed in accounting for the observed effects probably because this counting does not consider any contribution of syllable-initial consonants. The reason why initial consonants, in general, have not been taken into account that much by phonology is that they are irrelevant in determining the phonological properties of a syllable (Hyman, 1975). However, their relevance to acoustical properties is obvious, and therefore, some psychoacoustical influence from their presence or absence seems to

be inevitable. More specifically, the transitional part or boundary between consonant and vowel portions generally has a rapid and large acoustic change, e.g., a jump in intensity. Such an acoustic discontinuity can be used as one of the major markers that indicates the temporal structure or rhythm of a whole utterance. In short, the presence or absence of syllable-initial consonants is always accompanied by the appearance or disappearance of such major temporal markers.

For a coherent implication of the observations in experiment 2, we compared the appearance or disappearance of these markers among different stimulus groups. The words that included short vowel targets consisted of only CV syllables as seen in Fig. 3.7-(a). Since a stable and regular temporal alignment of the temporal markers was achieved in this alternation of short C and V, a clear and constant rhythm could be easily perceived from it. On the other hand, the temporal structures of those words that included the targets in the other three stimulus groups, more or less, deviated from the regular C-V alternation (Figs. 3.7-(b–d)). The degree of such structural deviations can be defined, as given below, using the number of temporal markers, i.e., C-to-V or V-to-C boundaries, that are dropped from the short C and V alternation.

In the case of words including short fricative targets (devoiced vowel portions), their temporal structures appeared to be the same as those including short vowel targets, C_1V_1-$C_2\mathring{V}_2$-C_3V_3-C_4V_4. However, as the devoiced vowel $\mathring{V}_2$ was actually a voiceless fricative and fused with the preceding consonant, there was, in fact, no acoustic discontinuity between the second consonant (C_2) and the following vowel ($\mathring{V}_2$). Accordingly, one temporal marker (C_2-to-$\mathring{V}_2$ boundary) could be regarded as having been dropped in comparison with the intact C-V alternation. The boundary between the devoiced vowel portion ($C_2\mathring{V}_2$) and the following consonant (C_3) had remained as a marker because C_2 and C_3 were different consonants in this stimulus group.

In the case of words including either long vowel or long fricative (geminate fricative) targets, two temporal markers could be regarded as having been dropped because their structures were either C_1V_1-C_2V_2-V_3-C_4V_4 or C_1V_1-C_2-C_3V_3-C_4V_4, where the boundary between the second and third mora (V_2-V_3 or C_2-C_3) did not have acoustic discontinuity. More specifically, both the V_2-to-C_3 and C_3-to-V_3 boundaries, and both the C_2-to-V_2 and V_2-to-C_3 boundaries could be regarded as having been dropped from the intact C-V alternation in the long vowel and long fricative cases, respectively. These marker droppings from the intact C-V alternation are summarized in Table 3.3.

Droppings of temporal markers from an intact C-V alternation imply degradation of the

temporal regularity. A temporal modification given in an irregular temporal sequence is, in general, perceived less sensitively than that in a regular one (Tanaka *et al.*, 1994, for example). Assuming a similar degrading tendency of the temporal sensitivity with marker droppings also in experiment 2, the predicted effects on the α scores due to the short–long contrast in each phonetic quality type agree with the observed ones.

Finally, it should be emphasized that the above introduced two ways of explanations using continuous and discrete durational measures are unlike those that exclude each other; they are two views explaining a general source that affects the acceptability judgment, i.e., the temporal structure that spans beyond the "target" portions. Therefore, the latter explanation, i.e., using a discrete or linguistic measure, is not necessarily regarded as language specific but as one of the universal ways of figuring out the physical variations of temporal structures in speech.

3.5 Conclusions

The extent of acceptability decrement with temporal modification in speech portions depended on the phonetic quality type of those portions whose durations were subjected to modification. The modification range for which a certain decrement of acceptability would be expected, i.e., the acceptable range, expanded as the phonetic quality type changed from vowel, nasal, then voiceless fricative or silence. The observed acceptability variations with the phonetic quality type correlated with the variation in loudness of the portion in question; the acceptable range narrowed as the portion became louder.

The extent of acceptability decrement with temporal modification in speech portions also depended on the original, as produced, durations of the portions in question. The acceptable range expanded as the original duration increased. Interestingly, the degree of the effect by the original duration depended on the phonetic quality type. The effect was larger for the vowel type than for the voiceless fricative type. This dependency could be accounted for by another source of the temporal structure, i.e., the vowel onset asynchrony (VOA). This dependency could be alternatively accounted for by a discrete measure representing the structural deviations of stimulus utterances from the regular C-V alternation.

An important implication of the current research is that an expanding acceptable range observed with changes in the phonetic quality or original duration can be mostly accounted

for by psychoacoustical terms, i.e., a reduced capability to discriminate temporal modifications as the loudness decreases or the original duration increases. It is probable that the acceptability of a speech portion coming in a different phonetic quality and duration from those tested in the current study is, to a considerable extent, predictable from the psychoacoustical properties. There may, indeed, be other factors capable of affecting the acceptability evaluation. However, the results presented here demonstrate that we can expect a more valid (closer to human evaluation) measure than the traditional simple average of acoustic errors in evaluating durational rules by accounting for the loudness and original duration as weighting factors.

Chapter 4

Acceptability of temporal modification of two consecutive segments[1]

Abstract

Perceptual sensitivity to temporal modification in two consecutive speech segments was measured in word contexts to explore the following two questions: (1) is there interaction between multiple segmental durations, and (2) what aspect of the stimulus context determines the perceptually salient temporal markers? Experiment 1 obtained acceptability ratings for words with temporal modifications. The results showed that the compensatory change in duration of a vowel (V) and its adjacent consonant (C) is not perceptually so salient as expected for the simultaneous modifications in the two segments. This finding suggests the presence of a time perception range wider than a single segment (V or C). The results of experiment 1 also showed that rating scores for compensatory modification between V and C do not depend on the temporal order of modified pairs (VC or CV), but rather on the loudness difference between V and C; the acceptability decreased when the loudness difference between V and C became high. This suggests that perceptually salient markers locate around major jumps in loudness. Experiment 2 further investigated the influence of the temporal order of V and C by utilizing the detection task for the speech stimuli instead of the acceptability ratings.

[1]This chapter is published as Kato, Tsuzaki, and Sagisaka (1997).

4.1 Introduction

Rules to assign segmental durations have been proposed for speech synthesis to replicate the segmental durations found in natural speech (Allen *et al.*, 1987; Campbell, 1992; Fant and Kruckenberg, 1989; Higuchi *et al.*, 1993; Kaiki and Sagisaka, 1992; Klatt, 1979; Sagisaka and Tohkura, 1984; van Santen, 1994). Each of the segmental durations produced by such durational rules will have a certain amount of error compared with the corresponding naturally spoken duration. The effectiveness of a durational rule should be evaluated by how much such an error is acceptable to human listeners, who are the final recipients of synthesized speech in general. While some durational rules have been perceptually evaluated (Carlson *et al.*, 1979; Sagisaka and Tohkura, 1984), in almost all of the previous cases, the average of the durational errors has been adopted as the measure for evaluation. Although we will not deny the effectiveness of this traditional approach, i.e., the effort to minimize the average acoustic error, we also find it crucial to investigate the "perceptual" basis to the evaluation of durational modification and to test the validity of the implicit premise in the traditional approach.

The implicit premise of this approach is that the sum of the perceived distortions corresponding to each segmental error becomes equal to the perceptual distortion for the entire speech. There are two possible problems with this premise. The first problem is in its giving every segment the same weighting in the error summation. In other words, it neglects segment attributes capable of affecting perceptual sensitivity to durational modification. For example, both Huggins (1972a) and Carlson and Granström (1975) reported that their subjects were more sensitive to durational changes of vowel (V) segments than to those of consonant (C) segments. The results of perceptual studies by Kato, Tsuzaki, and Sagisaka (1992) and Klatt and Cooper (1975) also suggested that durational modifications in word initial syllables are more critical than those in word medial or word final syllables. The second problem with the traditional premise is that it neglects dependencies among multiple errors. Relation factors between errors in adjacent segments, e.g., differences in the relative directions of deviations (the same or opposite), may affect the total impression of the perceived distortions, even when the average amounts of errors remain the same. If such a contextual effect on perceptual evaluation could be specified quantitatively, we could obtain a more valid (closer to human evaluation) measure than the traditional simple average of acoustic errors in evaluating durational rules.

While the first problem can be addressed by perceptual studies on temporal modification

in single segments, the second problem needs to be addressed by studies on simultaneous modification in multiple segments. However, only few studies have so far been made on the latter topic. In the current study, therefore, we focused on this second problem, and conducted experiments to examine perceptual sensitivity to modification made onto two consecutive segmental durations in a word.

The current study utilized two measures for perceptual sensitivity, i.e., detectability and acceptability. Detectability is fairly intuitive and psychophysically well-defined. Acceptability, on the other hand, is not generally defined and may be evaluated based on different norms. A word can have many quite different emotional or emphatic realizations that are judged to be equally acceptable according to the speaker's intention. We, therefore, limited our materials to normally pronounced word samples without any special emotional or emphatic intention from the speaker and measured the rate of acceptability for each manipulated stimulus as a normally stated exemplar of that token, focusing on the temporal aspect of the stimulus. In all of this, we tried to explore two questions concerning contextual effects on perceptual evaluation for the temporal modification of speech segments: (1) is there interaction between multiple segmental durations, and (2) what aspect of the stimulus context determines the perceptual salience of temporal markers?

4.1.1 Processing range in perceptual evaluation of temporal modifications

The first purpose of the current study was to explore whether there is a processing range wider than a single segment, i.e., a phoneme, in the perceptual evaluation of temporal modification in speech. A number of acoustic studies have pointed out that the duration of a given segment may depend on the surrounding contexts at various levels (Campbell, 1992; Fant and Kruckenberg, 1989; Hiki, 1967; Kaiki *et al.*, 1992; Takeda *et al.*, 1989; van Santen, 1992). Hiki (1967), for example, measured segmental durations in running speech comprising 424-mora text data and found a compensatory inclination between C durations and V durations within a mora. In addition, the effect of the number of moras in an utterance group on segmental durations has been reported at the word level (Takeda *et al.*, 1989) and at the sentence level (Kaiki *et al.*, 1992), as the tendency of each segmental duration to be inversely proportional to the number of moras. These results are consistent with the assumption that there are processing ranges wider than a single segment, i.e., a mora, a word, or a sentence, in the domain of speech production. However, these studies did not provide direct evidence for such a wide processing range in the domain of speech perception, because they were limited to the description of naturally spoken speech.

Perceptual studies, on the other hand, have looked at the perceptual consequence of temporal modification in speech segments (Carlson and Granström, 1975; Fujisaki, Nakamura, and Imoto, 1975; Huggins, 1972a, b; Klatt, 1976), but only a few have addressed perceptual phenomena caused by interaction among multiple modifications. Several studies have shown results suggesting the presence of perceptual compensation between V durations and their adjacent C durations. Huggins (1972b), Hoshino and Fujisaki (1983), and Sagisaka and Tohkura (1984) each reported that speech stimuli with multiple durational modifications in opposite directions between V and C tend to be heard as more natural than those with multiple durational modifications in the same direction. Sato (1977) moreover found that a lengthening of a consonant duration may cancel out the unnaturalness brought by the same amount of shortening of the adjacent vowel. A preliminary study by Carlson and Granström (1975), however, contrasts these studies. Carlson and Granström employed a discrimination task and found that their listeners' sensitivity to a change in the duration of a vowel was not affected by the presence of a compensatory change in the duration of the succeeding fricative. The discrepancy among these studies may be attributed to the differences in the speech utterances employed. However, one has yet to obtain sufficient information for deciding whether such temporal compensation between V and C is common because each of the previous studies employed a fairly small number of speech samples; i.e., two sentences in Sagisaka *et al.*'s study, three nonsense words in Hoshino *et al.*'s, one sentence in Huggins', the first to third syllables of one word *sakanayasan* (a fishmonger) in Sato's, and one word *plasta* in Carlson *et al.*'s.

In the current study, therefore, we tried to provide a direct test of the hypothesis that there is a wider processing range than a single segment in the perceptual evaluation of temporal modification in speech, by collecting a sufficient number of subjective responses using a sufficient number of stimulus samples. For this purpose, we measured perceptual compensation effects in accordance with the following procedure. First, we chose 30 V and C pairs from 15 four-mora Japanese words. Each of the chosen pairs was then temporally modified in four ways: (1) single V, (2) single C, (3) V and C in opposite directions, and (4) V and C in the same direction, as shown in Fig. 4.1. Secondly, temporal acceptability was rated for each of the modified words by human listeners. The obtained rating scores were processed and mapped on an interval scale using a psychological scaling method, then pooled for each of the four modification conditions. If the traditional premise, i.e., adopting the average acoustic error as the evaluation measure of durational rules, were valid in terms of perception, then the estimation score for multiple modifications as a whole would be the

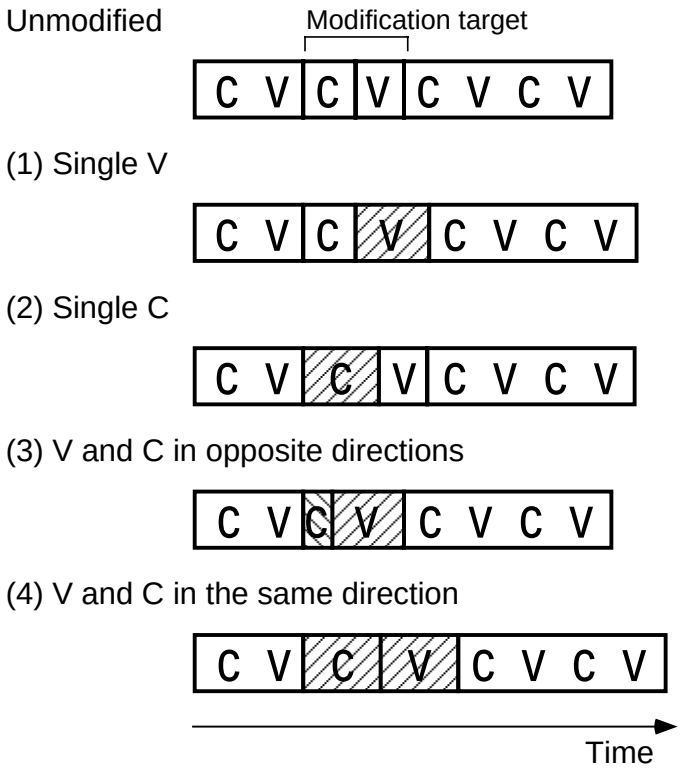

Figure 4.1: Schematic examples of different modification types performed on each of the selected word samples: (1) single V, (2) single C, (3) V and C in opposite directions (compensatory modification), and (4) V and C in the same direction. Each C or V stands for a consonant segment or a vowel segment, respectively, comprising a four-mora word. In the examples, the second consonant and the second vowel were chosen as the modification targets. The hatched segments were temporally modified.

same as the sum of the estimation scores for each of the single modifications. Therefore, the estimation scores for both "double modified" conditions (see Fig. 4.1, (3) and (4)), would each be expected to become equal to the sum of the scores for the "single modified" conditions (see Fig. 4.1, (1) and (2)). Otherwise, it would be suggested that the contextual factor among the multiple modifications had affected the perceptual evaluation; this supports the presence of a wider processing range than a single segment in the time perception of speech. In particular, if the average estimation score for condition (3) were significantly lower than that for condition (4), this would imply a general tendency of the perceptual compensation effect (Hoshino and Fujisaki, 1983; Sagisaka and Tohkura, 1984). Prior to this, although the stimuli were click sequences and the results may have been unable to be directly generalized

for speech studies, Schulze (1978) did an experiment analogous to the current design. He reported results supporting the existence of such a perceptual compensation effect, or the existence of a processing level that can cope with patterns distributed in a wide time span, involving multiple click intervals, rather than just a local dispersion, within a single interval.

4.1.2 Contextual effect on the perceptual salience of temporal markers in speech

As temporal structures such as rhythm or tempo can be perceived in speech, there should be temporal markers giving us reference points for such temporal structures in speech. Therefore, one can deal with issues on the time perception of speech by considering these temporal markers. Problems then arising from this standpoint are that one cannot explicitly specify the locations of such temporal markers and that the markers do not necessarily have the same perceptual salience.

Numerous attempts have been made to specify temporal markers in speech perception. The following items serve as examples: (1) earlier studies have assumed perceptual beats to be located at the vowel onsets or release points of a consonant into the succeeding vowel (Allen, 1972a, b; Rapp, 1971; Sato, 1977), (2) studies on "perceptual centers (P-centers)" have tried to calculate the precise locations of stress beats (Barbosa and Bailly, 1994; Fowler, 1979; Morton *et al.*, 1976; Scott, 1993); an analogous idea has been examined for the beat locations of mora timings in Japanese (Kato and Hashimoto, 1992), and (3) another group of studies has attempted to find connections between perceptual beats and production information such as the beginning of vowel articulations or the timings of motor commands (Fowler, 1983; Fujisaki and Higuchi, 1980; Tuller and Fowler, 1980). These studies have commonly assumed the presence of perceptual beats each of which has a one-to-one correspondence with some linguistic unit such as a syllable, stress, or mora and/or assumed the presence of perceptual isochronism of the beat sequence. Naturally, the major interests of these studies have been centered on the acoustic/articulatory features or contexts affecting beat locations and on the degree of isochronism of the predicted beat locations; little attention has been given to the perceptual salience of each temporal marker, i.e., the susceptibility of each marker to temporal displacement.

However, the perceptual salience of each marker seems to be of practical importance in evaluating durational rules because it can serve as an inherent perceptual weighting for individual temporal markers. In the current study, therefore, we first defined temporal markers as rapidly changing parts of speech in between steady-state parts and assumed, as

a first-order approximation, that their locations were to be at both the V-to-C and C-to-V boundaries. The markers were not necessarily a single point in time but could be a part having a certain duration, e.g., the burst part of a stop or affricate consonant in a CV sequence was included into the temporal marker at the C-to-V boundary. Then, we tried to explore what aspect of the stimulus context determines the perceptual salience of temporal markers without any assumption about the isochronous beat.

As shown in Fig. 4.1(3), the compensatory modification of two consecutive segments does not destroy the temporal structure outside the modified pair. Therefore, if each modification were made on stable parts, such as vowel plateaus, fricatives, nasals, or pre-burst closures, then generally speaking, only the boundary between the two modified segments would be temporally displaced. In such a case, if the boundary part were to contain a perceptually salient temporal marker, this modification would have a strong effect on the perceptual evaluation such as the rating magnitude of the perceptual distortion.

In this way, any change in perceptual evaluation caused by the compensatory modification can be a measure for the perceptual salience of the temporal markers located between the two modified segments. Utilizing this measure, the current study tested the following two possible models for predicting the perceptual salience of temporal markers in speech.

The first model is called the CV model; “CV” stands for a pair of a consonant and its succeeding vowel. This model assumes that the consonant onset is perceptually the most salient and tries to explain differences in perceptual effects such as the acceptability of temporal modification mainly by this factor. Therefore, the model assumes that such differences arise after the processes of segmentation and categorization which utilize speech-specific knowledge. The model is likely to be supported by linguistic considerations because a CV unit usually forms a mora, a phonological segmentation unit in Japanese. Several studies have repeatedly mentioned the importance of CV units in the domain of speech production (e.g., Campbell and Sagisaka, 1991; Sagisaka and Tohkura, 1984); the compensatory relation between a C duration and its succeeding V duration was observed in both Sagisaka *et al.*’s study and also Campbell *et al.*’s study which performed acoustical analyses on large databases of spoken Japanese. These previous studies apparently support the CV model. These studies, however, can not directly support the dominance of the CV unit in speech perception, because they are based on the observation of physical characteristics of “naturally spoken” speech and do not make an empirical assessment in the subjective evaluation for these stimuli.

On the other hand, in the domain of speech perception, several pioneering studies

by Hoshino and Fujisaki (1983) and Sato (1977) have so far looked into the temporal compensation between V and its adjacent C. Although both studies suggest the presence of a compensation effect, they seem to disagree in terms of the results for the compensation unit. Hoshino and Fujisaki investigated the tolerability of changes in the durations of V and C in nonsense words. Their results showed that the tolerability is higher for complementary duration changes of a C and its succeeding V than for those of a C and its preceding V. Sato, on the other hand, found that the decrease in naturalness caused by the lengthening of the first or the third vowel in the word *sakanayasan* (a fishmonger) could be recovered more by the shortening of its succeeding consonant than by the shortening of its preceding consonant. The former study supports the advantage of CV-unit compensation over VC-unit compensation, while the latter supports the advantage of VC-unit compensation. Therefore, it is still an open question whether CV is a more significant unit for perceptual compensation than VC; the CV model therefore needs to be tested. Using a relatively large number of speech samples, the current study compared perceptual evaluations involving compensatory modifications on VC and on CV. If the compensatory modification on CV made a smaller perceptual effect than that on VC, then the CV model would be supported.

The second model is called the loudness model. This model assumes that the perceptual salience of a temporal marker correlates with the amount of change in the perceived intensity, i.e., loudness[2], before and after the marker in question. Therefore, the second model focuses on the psychophysical property of speech sounds, while the first model is based on the linguistic property of speech, i.e., the mora. In the current study, we chose the magnitude of the loudness difference or jump between two modified segments from among various psychophysical variables, because previous studies have shown that perceptual sensitivity to durational modification on a single segment is correlated with the loudness of the modified segments relative to their surrounding segments (Kato *et al.*, 1998a; Kato and Tsuzaki, 1994). If the temporal modification on the segment boundary having a larger loudness jump made a larger effect on the perceptual evaluation, then the loudness model would be supported.

[2] Any usage of the word "loudness" in the current study means the loudness calculated by ISO-532 method B, unless otherwise stated. ISO-532B provides a loudness level or its equivalence in loudness as an instantaneous one. Although ISO-532B does not always provide excellent approximations for non-steady-state signals like speech, we adopted this method due to the advantage of its psychophysical basis instead of adopting power or intensity which incorporates no psychophysical considerations.

4.2 Experiment 1

In experiment 1, the acceptability of the temporal modification of V, C, or both V and C within a four-mora word was measured to test whether two consecutive temporal modifications interact with each other and to test the two models (CV and loudness) each describing a stimulus context that possibly correlates with the perceptual salience of temporal markers in speech.

4.2.1 Method

Subjects

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

Stimuli

Fifteen words were selected from the ATR speech database (Kurematsu *et al.*, 1990) as the original materials (Table 4.1). All of them were commonly used four-mora Japanese words, excluding words with doubled vowels, geminated consonants, or moraic nasals[3] which may disturb the temporal regularities observed in open syllable sequences. The selected words were spoken naturally by one male speaker and were digitized at a 12 kHz sampling frequency and with 16-bit precision. One of the V segments in each of the selected words and either the preceding or succeeding C segment were chosen as the targets of durational modification. Therefore, a set of target segments in experiment 1 comprised 15 CV pairs and 15 VC pairs. All of the target vowels were /a/ and their temporal positions in a word were chosen from the first three out of four moras.

Each of the paired target segments was temporally modified in four ways: (1) single V, (2) single C, (3) V and C in opposite directions, and (4) V and C in the same direction, as shown in Fig. 4.1. Each modification was either to lengthen or to shorten the segment(s). The size of a modification was either 15 ms or 30 ms. When two segments were modified, i.e., (3) or (4), the absolute modification size of one segment was equal to the other.

The modifications were made by a cepstral analysis and resynthesis technique with the log magnitude approximation (LMA) filter (Imai and Kitamura, 1978), and were carried out at 2.5 ms frame intervals. Durational changes were achieved by deleting or doubling

[3] In the orthography, they each have a separate character with the same status as the CV units.

Table 4.1: Speech tokens selected in experiment 1. The underlined CVC sequences are the target parts. The left column shows the temporal positions of targets in a word, where C_i or V_i stands for the ith consonant or ith vowel in the word.

Target position	Roman transcription				
$C_1V_1C_2$	bakugeki	gakureki	hanareru	nagedasu	sakasama
$C_2V_2C_3$	hanahada	imasara	kasanaru	katameru	mikakeru
$C_3V_3C_4$	hanahada	korogasu	rokugatsu	tachimachi	tamatama

the synthesis parameters frame by frame; the frames of interests were evenly chosen from the entire target part of each segment. The target parts were carefully trimmed out so as to exclude the transient parts at both ends of vowels and the on-and-after burst parts of plosives or affricates. That is, since the temporal markers were assumed to be at the VC or CV boundaries, the editing procedures modified durations at locations remote from these boundaries. In addition to the above modified stimuli, we prepared unmodified stimuli for reference. In total, 435 word stimuli were prepared.[4]

Procedure

The stimuli were randomized and fed diotically to the subjects through a D/A converter (MD-8000 mkII, PAVEC), a low-pass filter (FV-665, NF Electronic Instruments, $f_c = 5700$ Hz, -96 dB/octave), and headphones (SR-Λ Professional, driven by SRM-1/MkII ATR Version, STAX) in a sound-treated room. The average presentation level was 73 dB (A-weighted) which was measured with a sound level meter (type 2231, Brüel & Kjær) mounted on an artificial ear (type 4153, Brüel & Kjær). A four-second interval was inserted after each presentation for the subjects' response.

The subjects were told that each stimulus word was possibly subjected to temporal modification. Their task was to rate how acceptable each stimulus was, as an exemplar of

[4]They were 15 CVC's × 29 variations of modifications; i.e., 2 modification sizes (= 15 ms, 30 ms) × 2 modification directions (= lengthening, shortening) × 7 modification types (= single V, single C succeeded by the target V, single C preceded by the target V, V and the preceding C in the same direction, V and the succeeding C in the same direction, V and the preceding C in opposite directions, and V and the succeeding C in opposite directions) + 1 (= unmodified for reference).

that token on a scale of seven subjective categories ranging from -3 to 3, where -3 corresponded to "quite acceptable" and 3 corresponded to "unacceptable."[5] The subjects were asked to respond only about the temporal aspect of the stimuli, as much as possible. Each subject rated each stimulus eight times in total. The obtained raw scores were pooled over all subjects for each category, and then each stimulus was mapped on a unidimensional psychometric scale to assure an interval scale in accordance with Torgerson's law of categorical judgment[6] (Torgerson, 1958). The scaled estimation score of each "modified" stimulus was then adjusted by subtracting the scaled estimation score of its corresponding "unmodified" stimulus. Consequently, the score finally obtained for each stimulus corresponded to the difference in acceptability from the unmodified reference stimulus.

4.2.2 Results and discussion

Figure 4.2 shows the estimated acceptability scores pooled over the 15 stimulus words for each of the four types of temporal modifications, i.e., single V, single C, V and C in opposite directions (V and C opposite), and V and C in the same direction (V and C same). Multiple comparisons among these four modification conditions using Tukey–Kramer's HSD indicated the differences between "V and C same" and the other three conditions, and between the conditions of "single C" and "V and C opposite" to be significant $[p < 0.05]$. If the two single modifications of each "double modified" condition were to undergo the acceptability evaluation independently of the other, then no difference would be observed between the decreases in the acceptability scores for the two "double modified" conditions, i.e., "V and C opposite" and "V and C same." This, however, was not the case. The acceptability score for "V and C same" decreased more drastically than that for "V and C opposite" as clearly shown in the figure.

These results mean that, simultaneous modifications in a V duration and its adjacent C duration are not independent of the other in terms of the acceptability evaluation. They seem to either perceptually compensate each other when in opposite directions or perceptually enhance each other when in the same direction. This suggests that a process having a time span wider than a single segment (V or C) is involved in the time perception of speech.

[5] If the listeners were asked to rate the "naturalness," they might have tended to use a strict criterion making it difficult for an informative evaluation to be maintained for the whole range of temporal modifications to be tested. To obtain information for a reasonably wide range of modifications, therefore, we chose the "rating of acceptability" over the "rating of naturalness."

[6] This is a method of psychological scaling using the outputs of a rating scale method. Each of the categorical boundaries and the stimuli used in the rating are mapped on a unidimensional interval scale by the method.

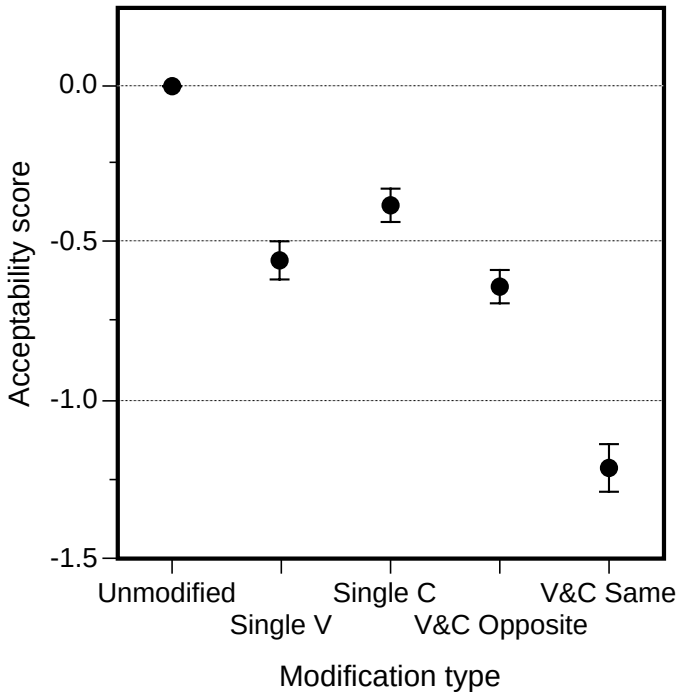

Figure 4.2: Estimated acceptability scores pooled over 15 word stimuli for each type of temporal modification. The dots and error bars show the group averages and the standard errors, respectively.

To test the two models (CV and loudness) each predicting a stimulus context that correlates with the perceptual salience of temporal markers, the following analyses focused on the modification type of "V and C opposite", the compensatory modification. If a large amount of decrease in acceptability were observed for this type of modification, this would mean that a perceptually salient temporal marker was located at the boundary of modified V and C. This is because the compensatory modification does not change the temporal structures outside of the target V and C, but mainly displaces the part between the two target segments. According to the CV model, a lower acceptability would be predicted for VC pair modifications than for CV pair modifications because the latter case preserves the duration of CV which is assumed to be a perceptual unit in the model while the former case changes the unit duration. According to the loudness model, on the other hand, a lower acceptability would be predicted for a stimulus with a large loudness jump between modified V and C than for a stimulus with a small loudness jump between modified V and C. The temporal order between V and C does not have any constraints in this model.

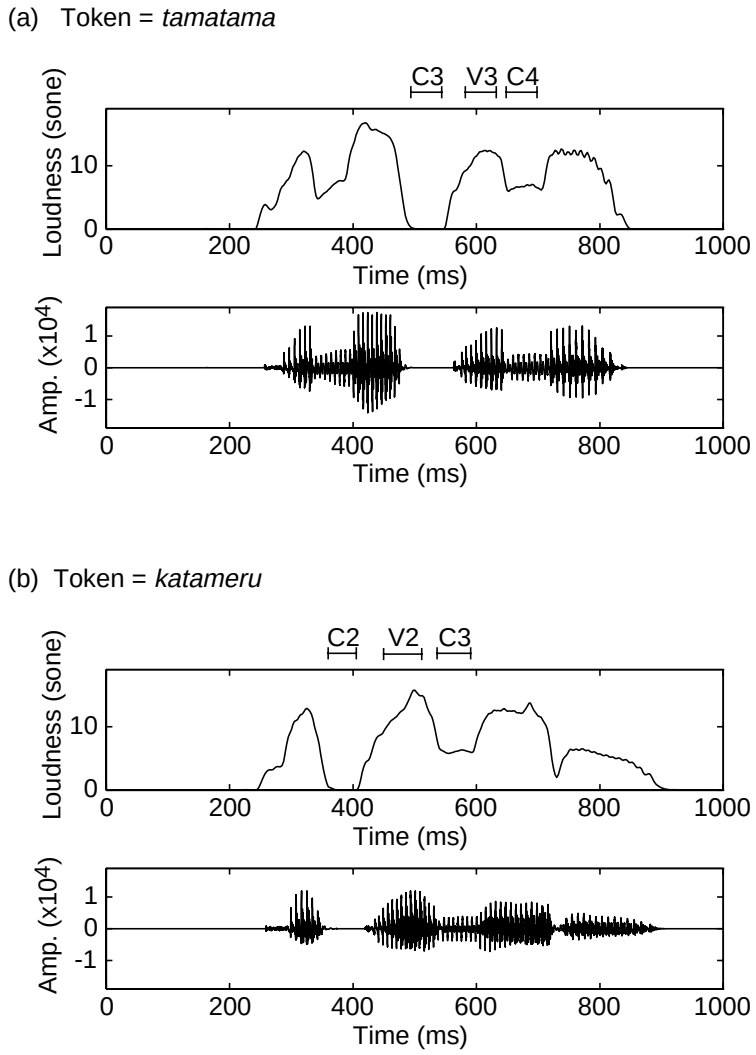

Figure 4.3: Time waveforms and loudness contours of the word stimuli used in experiment 1. The horizontal bars at the top of each figure indicate the target parts to be modified. The sampled tokens are "*tamatama* (by accident)" and "*katameru* (to make hard)."

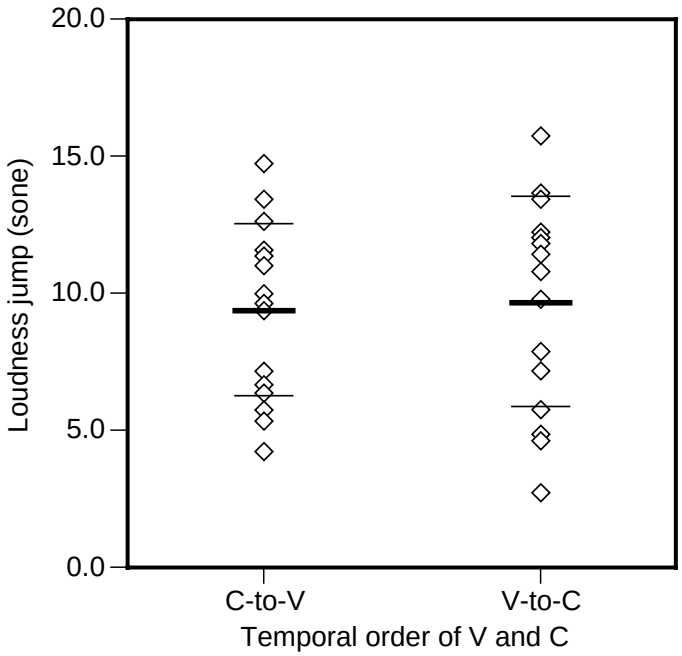

Figure 4.4: Loudness jump as a function of the temporal order of V and C. The thick and thin lines show the group averages and standard deviations of the loudness jump.

To quantify the explanatory variable of the loudness model, the loudness contour was calculated for each of the 15 original word samples in accordance with ISO 532B (ISO, 1975) using Zwicker *et al.*'s (1991) algorithm. Figure 4.3 shows examples of the obtained loudness contours and their corresponding waveforms. We then obtained the "loudness jump" by subtracting the median loudness of the C part from that of the V part. As every V target in this experiment was louder than its adjacent C parts, the employed loudness jumps were always positive. These obtained loudness jumps and the temporal order of modified V and C were adopted as the explanatory variables for the two models considered. Although we could not evaluate whether the populations of these two explanatory variables are independent of each other, the original word samples of the current experiment were selected so that these two variables were as little correlated with each other as possible, in order to make the subsequent statistical analyses reliable. Figure 4.4 shows the calculated loudness jumps as a function of the temporal order of modified V and C. A *t*-test did not indicate the difference between the average loudness jump at the VC boundaries and that at the CV boundaries to be significant $[t(28) = 0.28, p = 0.80]$.

In addition to the factors of loudness jump and temporal order of V and C, we included the following two factors which may affect the acceptability evaluation into a statistical test: the mora position of the modified vowel (1, 2, or 3), and the size of each single modification (15 or 30 ms). The effects of the above four factors and their interactions on the decrease in acceptability were tested by a four-way factorial General Linear Model (GLM)[7](SAS Institute Inc., 1990). The main effect of the loudness jump was significant $[F(1, 96) = 10.5, p < 0.005]$. The acceptability decreased with increasing loudness jump as shown in Fig. 4.5(a). The interaction between the loudness jump and the amount of modification was significant $[F(1, 96) = 4.15, p < 0.05]$. The effect of the loudness jump was stronger for the longer (30 ms) modification condition. The temporal order of V and C was not significant $[F(1, 96) = 0.087, p = 0.77]$ as shown in Fig. 4.5(b). No other main effect or interaction was significant.

There was no evidence of the CV model within the scope of experiment 1. On the contrary, the results of the GLM analysis supported the loudness model; a large loudness jump between modified segments generally caused a considerable decrease in acceptability. This suggests that perceptually salient temporal markers tend to locate around major loudness jumps. Note, however, that several factors capable of affecting the perception of temporal aspects of speech were not included in the current analysis; e.g., temporal discriminability is higher around phoneme boundaries when the phonemic distinction depends on the durational cue (Fujisaki *et al.*, 1975). Although we selected the stimulus tokens of experiment 1 so as to balance such factors, they were not completely factored out. Furthermore, we should not overlook the possibility that the effect observed might have depended on the particular language of the materials or the subjects because prosodic patterns, in general, carry different loads in different languages.

Experiment 2 was therefore designed to test whether the factor of loudness jump really affects time perception, using non-speech stimuli replicating the time-loudness features found in the speech stimuli of experiment 1. Although language factors cannot be completely eliminated provided the subjects are the native speakers of a single language, such non-speech studies minimize the influence of both speech-related and language-specific factors. If the effectiveness of the loudness jump were to be confirmed in this non-speech experiment, it would suggest that the effect found was based on general perceptual processes instead of speech-related or language-specific ones.

[7]This is an extended version of the analysis of variance or ANOVA. GLM copes with continuous values as explanatory variables as well as nominal values.

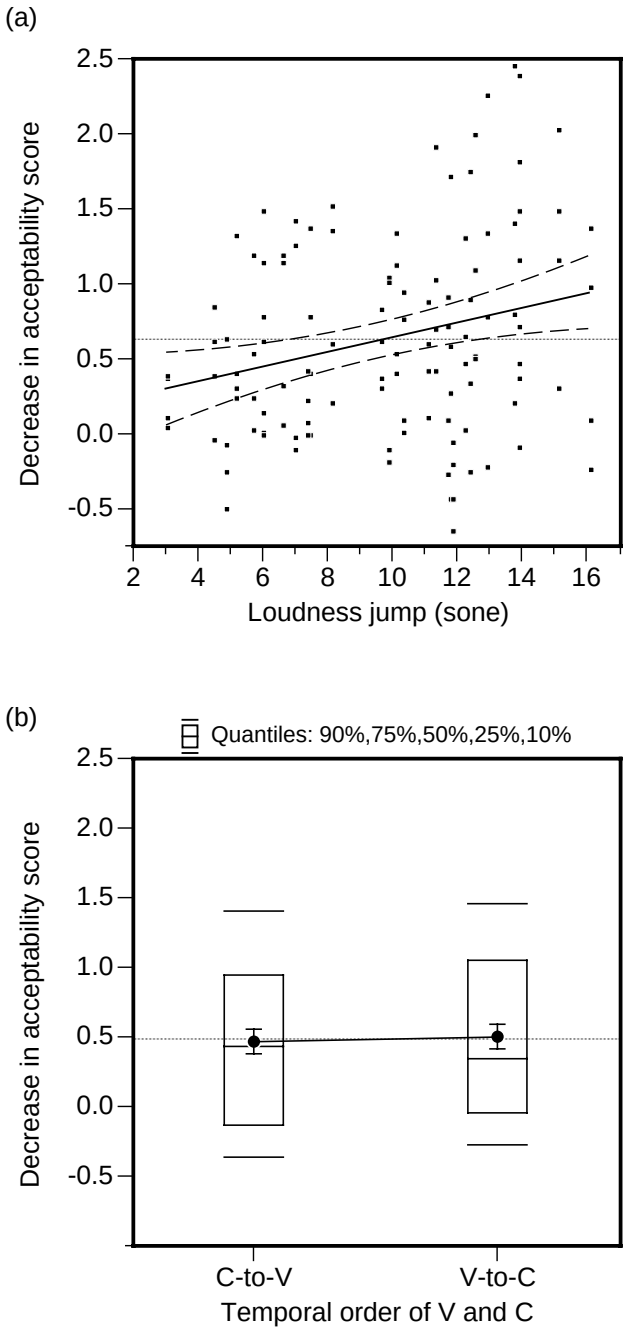

Figure 4.5: Decrease in acceptability caused by compensatory modification as a function of the loudness jump between V and C (panel a) or the temporal order of V and C (panel b). In panel a, the thick solid line and dashed lines show the regression line and its 95 % confidence curves. In panel b, the dots and error bars show the group averages and the standard errors, respectively. Quantile boxes are also shown in the figure. The horizontal thin dotted line in both panels marks the average of all samples.

4.3 Experiment 2 [8]

The purpose of experiment 2 was to test the effect of loudness jump on perceptual sensitivity to the displacement of temporal markers under controlled experimental conditions.

4.3.1 Method

Subjects

Six adults with normal hearing participated in experiment 2. All of the subjects also participated in experiment 1, and, therefore, there was a period of one month between experiment 1 and experiment 2 to prevent the carry-over of judgment strategies as much as possible.

Design

The experiment was designed as a four-way factorial one. The first factor was the loudness jump between two modified segments (large jump or small jump). Three other factors were included mainly to test their interactions with the first factor; they were, the direction of the marker slope (rising or falling), the steepness of the marker slope (steep or broad), and the temporal position of the marker in a sequence (first half or second half).

Stimuli

Each stimulus was a 1 kHz tone with one of two types of overall amplitude contours as shown in Fig. 4.6. These two types (type I and type II) were modeled on typical loudness contours of four-mora word stimuli (see Fig. 4.3), and enabled us to complete the factorial design described above. Each stimulus comprised the alternation of slope and steady parts. As in experiment 1, temporal markers were defined in experiment 2 as rapidly changing parts of a signal in between steady-state (either silence or sounding) parts; i.e., only the slope parts could become temporal markers. The steady parts each had one of the following three levels: 73 dB SPL (9.85 sone), 64 dB SPL (5.28 sone), or silence, where each was employed as an approximation for the average loudness of vowels, nasals, and pre-burst closures found in the speech stimuli used in experiment 1. The duration of the slope was 10 ms (steep) or 20 ms (broad). The duration of the loud part (the V part in Fig. 4.6) including rise-fall slopes and that of the soft or silent part (the C part in Fig. 4.6) were 100 and 50

[8] This section also appears as Section 4 of Chapter 6.

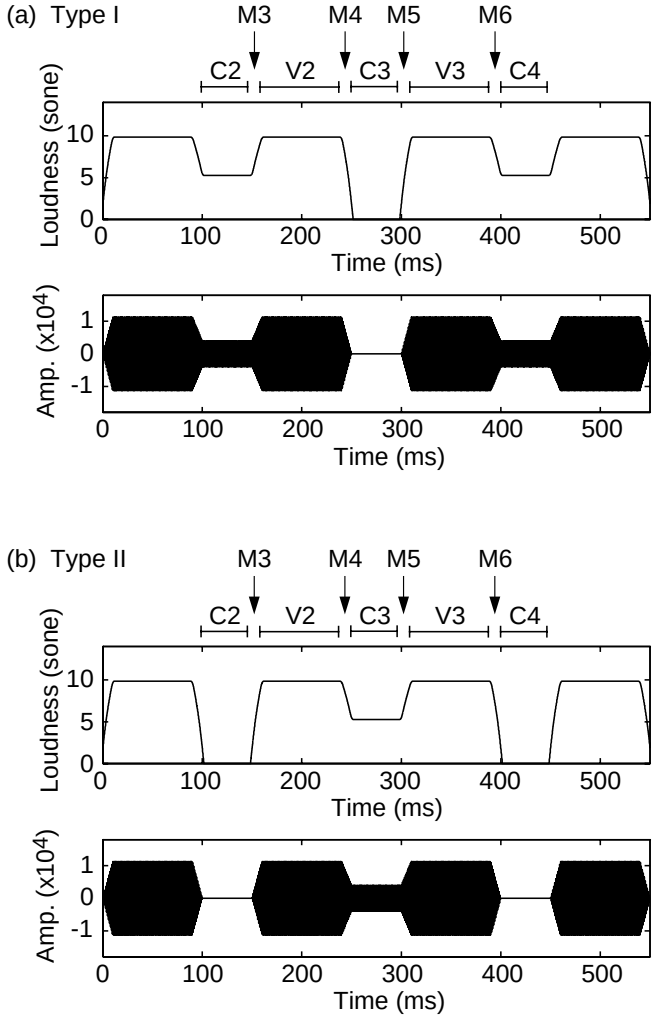

Figure 4.6: Time waveforms and loudness contours of the two types of stimuli used in experiment 2. Each "V" or "C" indicates a target part to be modified. Each "M" indicates the location of the temporal marker considered. The level of V is 73 dB SPL (= 9.85 sone), the level of louder C is 64 dB SPL (= 5.28 sone), and that of softer C is "silence." All signals are 1 kHz pure tones. The eight markers in the figures (type I and type II) comprise an orthogonal set for three (loudness jump, slope direction, temporal position) of the four factors considered. The fourth factor (slope steepness) is included by considering another set of type I and II stimuli of which the slope duration is 20 ms (broad slope); that of the above stimuli is 10 ms (steep slope).

ms, for the standard stimuli. One of the V durations in each of the comparison stimuli and either its preceding or succeeding C duration were modified in opposite directions with 30 ms for each. The modification target was limited to the steady part in the second V (V2) or the third V (V3) and either of its adjacent C's (Fig. 4.6). As a result, the marker (transient part) between the modified segments was solely displaced forward or backward by 30 ms from the standard. In total, 32 stimuli were prepared.[9]

Procedure

The detectability index (*d'*) was measured for the difference between each pair of standard and comparison stimuli by the method of constant stimuli. The experimental apparatus was the same as in experiment 1. The subjects listened to the standard and comparison stimuli and were asked to rate the difference between them using eight numerical categories: "0" to "7"; the larger number corresponding to a larger subjective difference. Since the stimuli were complex and unfamiliar to the subjects, the experimental trials were preceded by a 1-hour practice session to familiarize the subjects with the stimuli. In each experimental trial, the subjects listened to the presentation of four successive stimuli, the first three each being the standard and the last one being a comparison. This repetition of standard stimuli served to effectively familiarize the subjects with the stimuli. The inter-onset interval of four stimulus sequences was 1400 ms each which was chosen so as to prevent temporal markers in the standard sequences from coinciding to a perfect isochronous rhythm. Twenty percent of the trials were control trials in which each comparison stimulus was the same as the standard stimulus. Twelve judgments were collected from each subject for each stimulus. The obtained responses were pooled over all subjects for each category, and then the detectability index, *d'*, for each comparison stimulus was estimated in accordance with the Theory of Signal Detection (Green and Swets, 1966).

4.3.2 Results and discussion

A four-way completely randomized factorial analysis of variance (ANOVA) was performed for the obtained detectability indices *d'*. The factor of loudness jump was significant $[F(1,16) = 99.5, p < 0.0001]$. The other three factors also turned out to be significant; they were, the direction of the slope $[F(1,16) = 52.2, p < 0.0001]$, the steepness of the slope

[9] They were 2 types of amplitude contours (= type I, type II) × 2 steepness conditions (= steep, broad) × 2 slope directions (= rising, falling) × 2 target positions (= first half, second half) × 2 displacement directions (= forward, backward).

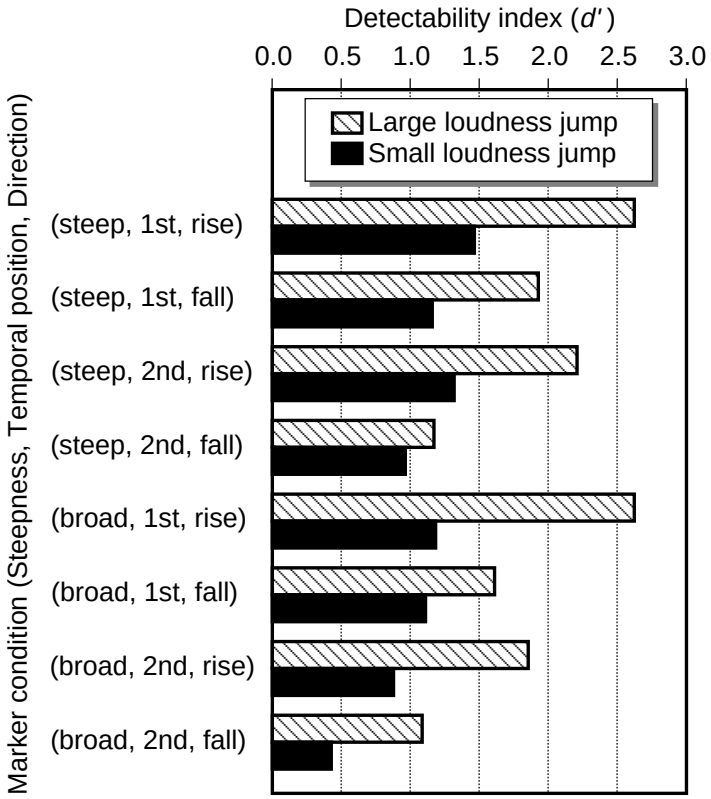

Figure 4.7: Detectability index *d'* for a 30 ms displacement of a temporal marker, for each combination of the marker conditions, as a function of the loudness jump between both sides of the marker. A larger *d'* implies easier detection.

$[F(1,16) = 6.30, p < 0.05]$, and the temporal position $[F(1,16) = 36.7, p < 0.0001]$. Besides these main effects, a significant interaction was observed between the factors of loudness jump and slope direction $[F(1,16) = 7.88, p < 0.05]$. No other interaction was significant.

Figure 4.7 shows *d'* for each combination of the marker conditions pooled over the marker displacement directions as a function of the loudness jump. As clearly shown in the figure, the effect of loudness jump agrees with the observed one in experiment 1; i.e., a larger loudness jump causes a higher sensitivity. The effect of temporal position in a sequence was significant; i.e., displacements of the markers in the first half were detected more easily than those of the markers in the second half. This effect is consistent with the finding reported by Tanaka, Tsuzaki, and Kato (1994) that the temporal discrimination for the initial interval is easier than that for the succeeding intervals in a click sequence.

The detectability of temporal displacement on a rising marker was significantly higher than that on a falling marker. Therefore, we thought that by applying this effect directly to experiment 1, the displacements of the C-to-V transition (always a rising slope) would have a greater effect on the perception than the displacements of the V-to-C transition (always a falling slope). However, this was not the case.

So what brought about such an inconsistency between the factors of marker direction and the temporal order of V and C? Two major differences existed between experiment 1 and experiment 2. The first one involved a physical difference between the speech stimuli and the pure tone stimuli. While the rising and falling slopes compared in experiment 2 were the exact mirror images of each other in the time axis, experiment 1 used 30 different slopes (V-to-C or C-to-V transitions). Such a wide stimulus variation in experiment 1 possibly obscured the effect of the marker direction.

The second difference was in the experimental procedure; experiment 1 employed the acceptability ratings of single stimuli while experiment 2 used a detection task. The task in experiment 1 could be broken down, from an analytical viewpoint, into the following two stages: 1) a detection stage — each subject had to detect the difference between the temporal structure of the presented stimulus and that of his/her internal exemplar of that token even though a single stimulus was presented in each trial, and 2) a rating stage — the degree of acceptability was rated. That is, experiment 1 required the subjects to do a rather central or higher level process in addition to a simple detection task similar to the one used in experiment 2. Therefore, even though the displacements of the C-to-V transition were detected more easily than those of the V-to-C transition, the rated scores possibly showed no difference with regard to the temporal order of V and C if the subjects were more tolerant of the former displacements than the latter ones owing to any factor related to central processes.

The influence of the mora unit is likely to be a candidate for such factors. Because a mora unit is usually comprised of a consonant and its succeeding vowel, a displacement of the C-to-V transition (rising marker) preserves the unit duration while that of the V-to-C transition (falling marker) changes it. Consequently, if the mora were a significant unit for the acceptability rating, a displacement of the V-to-C transition should be regarded as more critical than a displacement of the C-to-V transition.

Experiment 3 was therefore designed to test the second possibility: whether the difference in task between experiment 1 and experiment 2 yielded the inconsistency between the factors of temporal order of V and C and marker direction. This experiment adopted the same task as experiment 2 and employed stimuli similar to experiment 1's, i.e., we tried to separate

out the influence of the central processes possibly functioning at the rating stage. If the displacements of the C-to-V transition were detected more easily than those of the V-to-C transition, the hypothesis that the inconsistency between the results of experiment 1 and those of experiment 2 was due to the difference in task between these experiments would be supported. This would suggest the possibility that the CV unit (mora) functioned at the stage of the acceptability ratings in experiment 1.

4.4 Experiment 3

The purpose of experiment 3 was to test whether the difference between the results of experiment 1 and those of experiment 2 had been produced by the difference in task between these experiments.

4.4.1 Method

Subjects

Six adults with normal hearing participated in experiment 3. All of the subjects participated in all three experiments, and, therefore, there was a period of one month between two experiments to prevent the carry-over of judgment strategies as much as possible.

Stimuli

The stimuli were a reduced set of those used in experiment 1; the modification type was compensatory and the amount of each modification was 30 ms. In total, 60 word stimuli were employed (15 tokens $\times$ 2 temporal orders $\times$ 2 modification directions of vowel).

Procedure

The experimental apparatus was the same as in experiments 1 and 2. The experimental procedure and detectability calculation were the same as in experiment 2 except that a standard stimulus was presented once in each trial.

4.4.2 Results and discussion

The effects of the following three factors: the loudness jump, the temporal order of V and C, and the temporal position of the modified vowel, and their interactions on the obtained

detectability indices, *d'*, were tested by a three-way factorial General Linear Model (GLM). The main effect of the loudness jump was significant $[F(1, 48) = 33.1, p < 0.0001]$. On the other hand, the main effect of the temporal order of V and C was not significant $[F(1, 48) = 0.65, p = 0.42]$. No other main effect or interaction was significant. Figure 4.8 shows the obtained *d'* as a function of the loudness jump (panel a) and the temporal order of V and C (panel b).

These results are in good agreement with those obtained in experiment 1. Even though the detection task of experiment 2 was adopted in experiment 3, there was no significant effect for the temporal order of V and C. Therefore, we can safely state that the inconsistency between the results of experiment 1 and those of experiment 2 was not due to the difference in experimental task but to the difference in stimuli. This finding, therefore, does not support the hypothesis that the mora unit functioned as a factor cancelling the effect of the slope direction at the acceptability rating stage. Yet, we cannot exclude the possibility that the mora unit functioned in experiment 3 even though the task was of the detection type. We are, however, willing to say in a practical sense that such influence of the mora unit should be taken as a secondary effect preceded by more general processes based on the loudness jump.

Note that we adopted the loudness jump as a representative of the psychophysical auditory basis in contrast with more central or speech-specific ones. Further investigations are warranted to explore whether the loudness jump has an advantage or not over other psychoacoustical indices, e.g., the change in an auditory spectrum. In addition, the acoustic microstructures at the segmental boundaries such as the presence or magnitude of explosions or aspirations, possibly affect the perceptual salience of temporal markers. The effect of such detailed differences should also be explored in further investigations, probably by means of psychoacoustical indices that can deal with fine temporal and spectral changes.

4.5 General discussion

4.5.1 Possible evidence for the universality of the effect of loudness jumps

The results of both experiment 1 and experiment 3 demonstrated that the listeners were generally more sensitive to compensatory modifications of paired segments having a large loudness jump than to those having a small loudness jump. A similar effect of loudness jump was also observed for non-speech stimuli in experiment 2. These observations suggest that the effect found is not a speech-specific one. Naturally, the effect must be independent

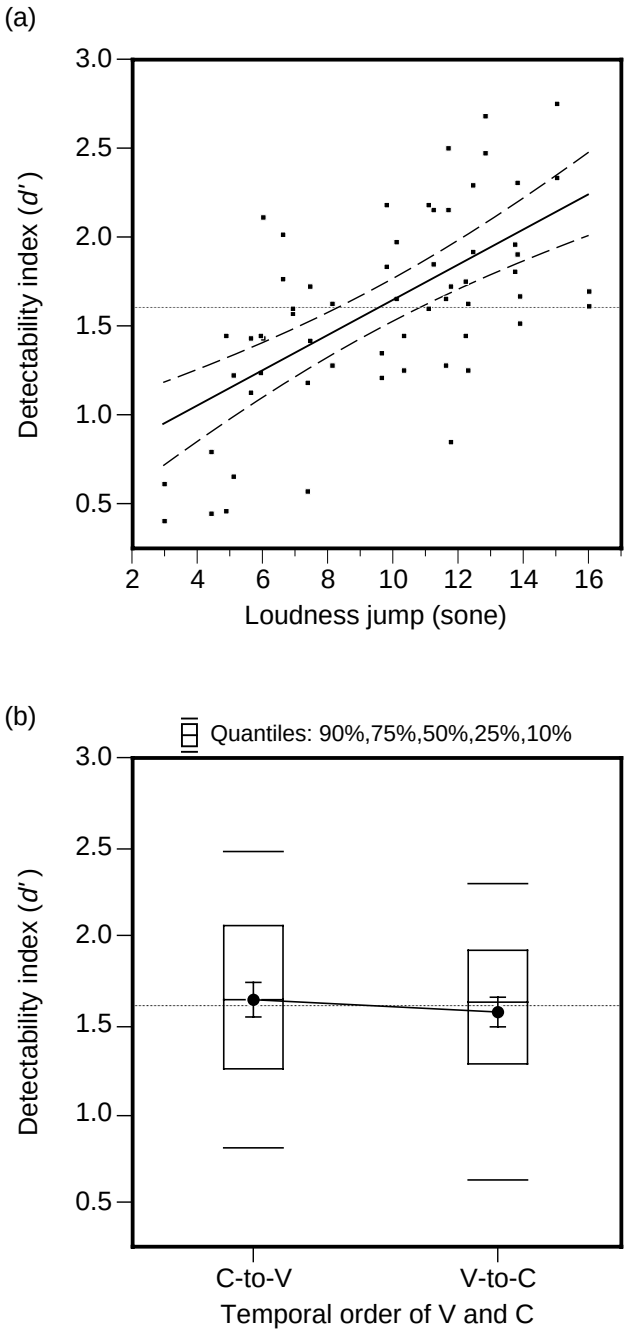

Figure 4.8: Detectability index *d'* for a 30 ms compensatory modification as a function of the loudness jump between V and C (panel a) or the temporal order of V and C (panel b). In panel a, the thick solid line and dashed lines show the regression line and its 95 % confidence curves. In panel b, the dots and error bars show the group averages and the standard errors, respectively. Quantile boxes are also shown in the figure. The horizontal thin dotted line in both panels marks the average of all samples.

of the variation of languages. Although such language independency cannot be proven by the current monolingual study, a phenomenon possibly reflecting the effect of loudness jumps can be found in the previous study where English materials and English listeners were employed.

Huggins (1972b) investigated perceptual compensation between a vowel duration and its adjacent consonant duration using the sentence "the hostel for paupers strives for perfection" which was spoken by a single male. His results showed that the vowel duration in the unstressed syllable "per" in the phrase "for perfection" was successfully compensated by the pre-burst closure duration of preceding consonant "p" while this was not the case for the vowel duration in the stressed syllable "pau" in the phrase "for pauper." Huggins gave two ways to account for the observed inconsistency; i.e., perceptual compensation was not observed for "paupers" but it was for "perfection." One was to reject the data from the case of "perfection" because the compensation found was highly dependent on a single pair of data points. The second was to argue that the compensation had the effect of not restoring the duration of the unstressed first vowel but of restoring the onset of the stressed second vowel in "perfection."

Both ways ignored the presence of perceptual compensation in the first syllable of "perfection." However, we can give an alternative explanation. Although a high intensity does not seem to be the primary requirement for the presence of syllable stress in English (Lehiste, 1970), a nucleus vowel in a stressed syllable generally has a greater intensity than that in an unstressed syllable (Fry, 1955; Lieberman, 1960). In addition, the intrinsic intensity of vowel 1/ɔ/ has been reported to be greater than that of vowel /ə/ (Lehiste and Peterson, 1959). Furthermore, the initial /p/ in a stressed syllable tends to be solidly unvoiced while the initial /p/ in an unstressed syllable can be pronounced weakly, sometimes reducing to a fricative. In the case of Huggins' study, therefore, it can plausibly be assumed that the difference in loudness, which is highly correlated with the intensity, between C (pre-burst p-closure) and V was larger in the stressed /pɔ/ of "paupers" than in the unstressed /pə/ of "perfection." This assumption is reinforced by the fact that both syllable materials were obtained from a single token uttered without any special emphasis, i.e., the conditions capable of affecting the syllable intensity, other than the presence of stress and the intrinsic intensity, were almost the same for both materials.

Our loudness model predicts a high sensitivity to the temporal displacement of a C-to-V transition having a large loudness jump even though the C duration and the V duration had been complementarily modified. Taking this model into account, it can be expected that

Huggins' listeners perceived the temporal displacement of the p-to-V transition in the first syllable of "paupers" to be larger than that of "perfection." This difference in sensitivity to the temporal displacement possibly yielded the difference in the significance level of the compensation effect. More specifically, in the case of "paupers," the compensation effect did not turn out to be significant because the perceptual salience of the boundary part between the two modified segments was relatively high. In the case of "perfection", on the other hand, the compensation effect was observed because the boundary part was not so perceptually salient.

This is merely an example possibly demonstrating the language-independency of the effect of loudness jump. However, the effect observed in the current paper would lend support for the presence of language-independent universal rules in evaluating the temporal modification of speech segments.

4.5.2 Validity of judgments based on energy differences

The stimulus modification in the current study changed not only the segmental duration but also the overall energy of the stimulus. Therefore, one could claim that the listeners' strategy to distinguish stimuli was not based on differences in the temporal structure but on differences in the overall energy. This claim appears to be consistent with the current experimental results, i.e., the listeners are more sensitive to the temporal modifications of markers having large loudness jumps than to those having small loudness jumps, because a compensatory temporal modification brings a larger energy change for a temporal marker having a large loudness jump than for a marker having a small loudness jump.

First, let us reexamine the results of experiment 2 under the assumption that the listeners did not process the differences between the standard and comparison stimuli in terms of temporal differences but differences in the overall energy. The ratio of the energy difference between the standard and comparison stimuli to the overall energy of the standard stimulus, i.e., the Weber fraction, was 8.42 % or 0.367 dB when modified markers had the larger jumps. This difference is equivalent to or smaller than the just noticeable differences (jnd's) for energy differences[10] found in former psychophysical studies, i.e., approximately 10–20 % (Florentine, 1983; Green, 1993; Jesteadt *et al.*, 1977; Rabinowitz *et al.*, 1976; Riesz, 1928; Schroder *et al.*, 1994). Nevertheless, our listeners could achieve an average *d'* of 1.89.

[10] We actually referred to studies on intensity discrimination because energy comparisons can be interchangeable with intensity comparisons when the stimuli involved are the same in terms of duration, and a lot of reliable jnd data can be obtained from studies on intensity comparisons.

Assuming the standard score of the listeners' response was linear to the energy difference, the Weber fraction of the energy difference required for our listeners to produce a *d'* of unity would be 4.46 %. This is, indeed, fairly smaller than formerly published jnd's of auditory energy. Such a high discriminability was also obtained for many of the speech stimuli in experiments 1 and 3.

In addition, the explanation based on comparisons of overall energy values does not predict many of the results in experiment 2. The experiment showed that the listeners' performance was significantly affected by at least three attributes of the displaced temporal markers, i.e., the slope direction (rising or falling), the steepness of the slope (steep or broad), and the temporal position in a sequence (first half or second half). None of these three attributes, however, are dependent on the differences in the overall energy between the standard and comparison stimuli. These results could hardly be obtained if the listeners did not pay attention to the temporal markers and did not consider them as cues of distinction.

Therefore, we should state that the claim that listeners could distinguish the stimuli by differences in the energy without processing temporal information is not plausible. It is reasonable to assume that the listeners made judgments by taking advantage of the differences in the temporal structures of the stimuli.

4.6 Conclusions

The experimental results showed that temporal modifications of two consecutive segments (V and C) are more acceptable when they are made in opposite directions than when they are made in the same direction. This suggests that a range having a wider time span, corresponding to a moraic range or wider, than a single segment (V or C) functions in the perceptual evaluation of temporal modifications in speech. The results additionally showed that the listeners' perceptual sensitivity to the compensatory modifications between V and C does not depend on the temporal order of V and C but rather on the loudness difference or jump between V and C; the sensitivity increased when the loudness jump between V and C became high. This suggests that the perceptual salience of temporal markers in speech is more closely related to an acoustic-based psychophysical feature (the loudness jump) than to a phonological or phonetical feature (the level of CV or VC).

Large jumps in loudness generally coincide with V-to-C and C-to-V transitions. This is probably one reason why previous studies which assumed a unit comprising CV or VC, had, to some extent, succeeded in explaining perceptual phenomena. However, the current

study implies that general auditory perception principles should be reexamined, as well as linguistic information, as efficient variables to explain the temporal aspects of speech perception. The important and practical suggestion of the current research is that when evaluating a durational rule objectively, the traditional measure, the average of acoustic errors, is not sufficient from a perceptual viewpoint. A measure more valid (closer to human evaluation) than the traditional average acoustic error could be expected by taking into account the perceptual effect described above.

Chapter 5

Functional difference between vowel onsets and offsets in perceiving temporal structure of speech[1]

Abstract

Controlling multiple segmental durations is interchangeable by aligning the start and end points of the segments. In Japanese, most consonants start with vowel (V) offsets except for post-pausal cases and end with V-onsets. The temporal alignment of V-onsets and V-offsets, therefore, mostly covers issues on the segmental duration control.

We examined the functional differences between V-onsets and V-offsets in the perception of temporal structures in speech stimuli. Listeners were required to estimate the perceived difference between a four-mora Japanese word and its temporally modified counterpart using two perceptual clues: (1) the simple difference and (2) the speaking rate. In the V-onset condition, the inter-onset intervals of vowels were uniformly changed (either lengthened or shortened) while preserving their inter-offset intervals, and vice versa in the V-offset condition. These manipulations did not change the duration of the entire word. Each of the modified words was paired with its unmodified counterpart and was given to the listeners. In the simple discrimination task, the listeners' ability was not affected significantly by the difference in the modification condition (onset or offset). In the speaking rate task, on the other hand, the influence of the marker condition on the listeners' performance was obvious. Changing the V-onset intervals correlated with a change in the perceived speaking rate despite the fact that the duration of the entire word was unchanged. However, the modifications to the V-offset intervals had no significant effect on the perceived speaking rate. These results suggest that V-onsets

[1]This chapter is based on Kato, Tsuzaki, and Sagisaka (1998c) and Kato, Tsuzaki, and Sagisaka (b).

and V-offsets equally contribute to the detection of a temporal distortion while V-onsets are the dominant cue in determining the speaking rate, i.e., the tempo of events. The results were interpreted by assuming two types of temporal measures in auditory perception, i.e., the between-event timing and the within-event duration.

5.1 Introduction

As temporal structures such as rhythm or tempo can be perceived in speech, there should be temporal markers giving us reference points for such temporal structures of speech. Therefore, one can deal with issues on the time perception of speech by considering these temporal markers. Problems then arising from this standpoint are that one cannot explicitly specify the locations of such temporal markers and that the markers do not necessarily have the same perceptual function.

In the previous study (Kato *et al.*, 1997), we assumed that the boundaries between vowels (Vs) and consonants (Cs) are markers that give us temporal information about speech, and estimated their perceptual saliency, i.e., the susceptibility to temporal displacement. The results showed that there was no significant functional difference between C-to-V boundaries, or V-onsets, and V-to-C boundaries, or V-offsets, in the detection of local changes in the temporal structure. The listeners were equally sensitive to both V-offset displacements and V-onset displacements, where each displacement was introduced by a compensatory change in a pair of V and its preceding/following C.

However, a number of studies so far have provided evidence for functional differences between V-onsets and V-offsets or the dominance of V-onsets. Allen (1977) and Sato (1977) each found locations of perceptual beats existing at V-onsets (or release of the consonants) by several methods such as the synchronization of finger tapping to speech. In addition, many of the perceptual centers (P-centers) studies (Marcus, 1981; Scott, 1993) have commonly regarded the contribution of V-onsets as greater than that of V-offsets; the subject of controversy for this group of studies has been the accurate locations of perceptual beats in relation to acoustical features of speech.

Discrepancies between the results of Kato *et al.* (1997) and those of previous perceptual studies can be attributed to differences in the stimulus context and in the listeners' task. In the former study, the listeners could achieve their task by detecting just a local (within a two-segment range) temporal distortion in an isolated word while the latter studies generally used longer stimulus sequences such as sentences or repetitions of a single syllable; i.e., the tasks of the latter studies required the listeners to process the target sequences as a whole. Such global processing by the listeners possibly yielded the dominance of V-onsets.

In the current study, therefore, we tried to test this possibility by utilizing a stimulus manipulation and an experimental task making it easier for listeners to process global

Table 5.1: The total durations and total inter v-onset/-offset intervals of speech tokens chosen in experiments 1 and 2 in ms.

Token	bakugeki	hachigatsu	hanahada	minogasu	monosashi
Total duration	720	695	675	730	710
Total inter-v-onset interval	520	515	445	530	550
Total inter-v-offset interval	535	550	545	605	595
Token	nagedasu	nakanaka	sakasama	samazama	sashidasu
Total duration	740	670	735	722.5	735
Total inter-v-onset interval	550	470	495	480	535
Total inter-v-offset interval	575	530	565	535	565

information. Experiment 1 used a global stimulus manipulation, i.e., all V-onsets or V-offsets in a word were subjected to temporal modification while Kato *et al.* (1997) only manipulated one of the V-onsets and V-offsets in a word. The listeners' task was the same as in Kato *et al.* (1997), i.e., the detection of the manipulation. Then, experiment 2 additionally employed the task of speaking-rate estimation, which requires listeners to use global information while the detection task can be achieved by observing only a local difference of the stimuli. The stimulus manipulation was the same as in experiment 1.

5.2 Experiment 1: Detection

The purpose of experiment 1 was to test whether there would be any difference between the V-onset and V-offset conditions in the detectability of a global temporal modification.

5.2.1 Method

Subjects

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

Stimuli

Ten words were selected from the ATR speech database (Kurematsu *et al.*, 1990) as the original materials (Table 5.1). All of them were commonly used four-mora Japanese words, each comprised of four C and V alternations. The selected words were spoken naturally by one male speaker and were digitized at a 12 kHz sampling frequency and with 16-bit precision.

Each of the original tokens was temporally modified in three ways: (1) entire modification, (2) inter V-onset interval (V_{on}-V_{on}) modification, and (3) inter V-offset interval (V_{off}-V_{off}) modification, as shown in Fig. 5.1. In the entire word condition, the entire word duration was either lengthened or shortened evenly, by 20, 40, 60, or 80 ms. This condition was included to confirm whether or not the subjects judged the speaking rates consistently with the physical tempi, and was designed as the reference condition.

In the V-onset or V-offset condition, each of the three V_{on}-V_{on}s or V_{off}-V_{off}s was either lengthened or shortened by 5, 10, or 15 ms, i.e. the change in total V_{on}-V_{on} or V_{off}-V_{off} was 15, 30, or 45 ms. To preserve the entire word durations and the intervals among the temporal markers of no interest, each of the V duration modifications was compensated with the corresponding modification of either the preceding C duration (V-onset condition) or the following C duration (V-offset condition). To lengthen every V_{on}-V_{on} by 15 ms, for example, the first to fourth Vs in a word were modified by $+22.5, +7.5, -7.5, -22.5$ ms while the first to fourth Cs were modified by $-22.5, -7.5, +7.5, +22.5$ ms (positive and negative values mean lengthening and shortening).

The ranges of all of these temporal modifications were chosen on the basis of preliminary experiments, as ranges within which no phonetical transitions could occur. In addition to the above modified stimuli, we prepared unmodified stimuli for reference. In total, 210 word stimuli were prepared (10 tokens $\times$ (20 variations of modification + 1 unmodified)). The modifications were made by a cepstral analysis and resynthesis technique with the Log Magnitude Approximation (LMA) filter (Imai and Kitamura, 1978), and were carried out with a 2.5 ms frame interval. The durational changes were achieved by deleting or doubling the synthesis parameters frame by frame.

Stimulus presentation

Each of the prepared stimuli was paired with its counter unmodified stimulus (in the order of unmodified first) and was presented to the subjects diotically through a D/A converter

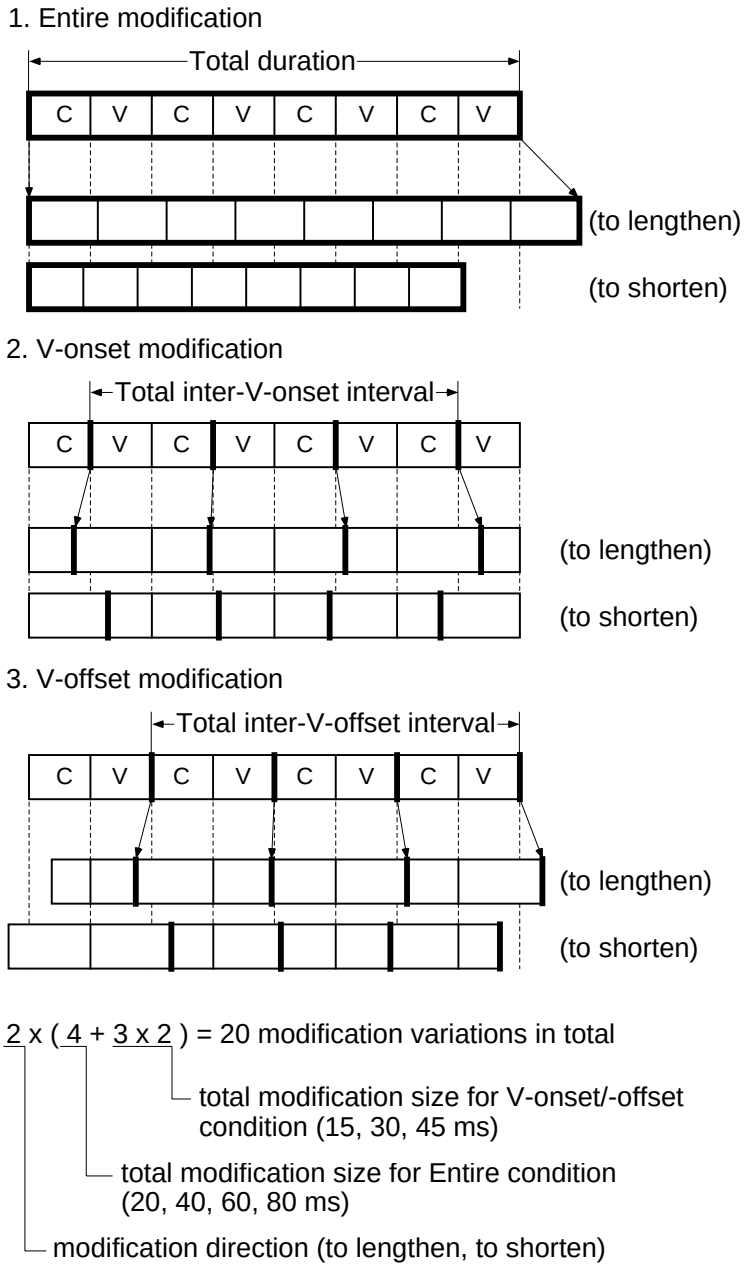

Figure 5.1: Schematic diagrams of the three modification types in a four-mora word: (1) entire modification of the total word duration; (2) inter V-onset interval (V_{on}-V_{on}) modification (V-offsets were preserved); (3) inter V-offset interval (V_{off}-V_{off}) modification (V-onsets were preserved).

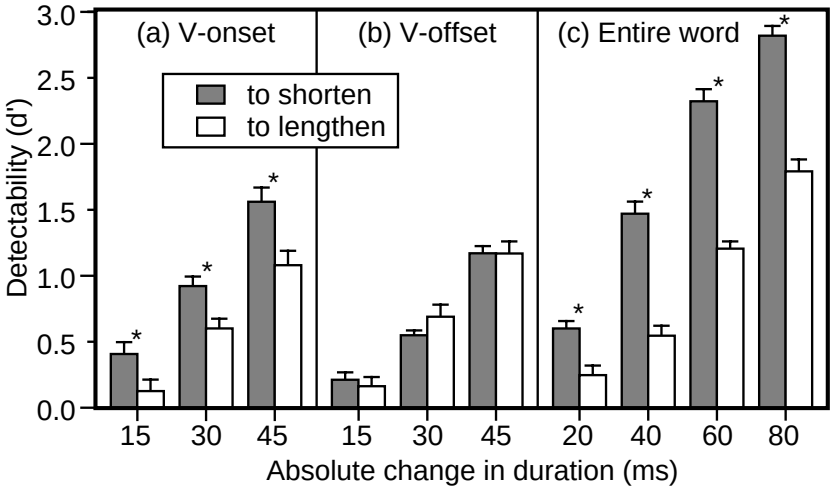

Figure 5.2: Mean detectability index (and standard error) for each modification type as a function of change in the (a) total inter-V-onset interval (V_{on}-V_{on}), (b) total inter-V-offset interval (V_{off}-V_{off}), or (c) entire word duration. The asterisks show pairs of bars whose differences are statistically significant ($p < 0.05$).

(MD-8000 mkII, PAVEC), a low-pass filter (FV-665, NF Electronic Instruments, $f_c = 5700$ Hz, -96 dB/octave), and headphones (SR-Λ Professional, driven by SRM-1 MkII, STAX) in a sound-treated room. The average presentation level was 73 dB (A-weighted) which was measured with a sound level meter (Type 2231, Brüel & Kjær) mounted on an artificial ear (Type 4153, Brüel & Kjær). All three stimulus conditions were tested in the same experimental sessions. Fifteen percent of the trials were control trials in which the presented two stimuli were identical.

Task

In each trial, the subjects were asked to rate the difference between the paired two word stimuli using eight numerical categories: "0" to "7"; a larger number corresponded to a larger subjective difference. Twelve judgments were collected from each subject for each stimulus. The obtained responses were pooled over all subjects for each response category, and then the detectability index, *d'*, for each comparison stimulus was estimated in accordance with the Theory of Signal Detection (Green and Swets, 1966).

5.2.2 Results

Figure 5.2 shows *d'* for each of the modification types, modification sizes, and modification directions. An ANOVA for the *d'* data of both Figs. 5.2 (a) and (b) with the modification type (V-onset or V-offset) and the absolute modification size (15, 30, or 45 ms) as main factors showed no significant effect of the modification type while the effect of the modification size was significant. The *d'* increased with increasing size of the modification.

Studying the results more specifically, the modification direction affected the detectability of the V-onset modifications; they were more easily detected when the direction was to shorten than when it was to lengthen. A similar tendency could be clearly observed in the entire word condition (Fig. 5.2(c)), although there was no such tendency in the V-offset condition.

5.2.3 Discussion

Experiment 1 showed that the listeners' detectability, on average, did not differ between the V-onset and V-offset modifications, even though they were globally manipulated modifications. However, this fact does not necessarily mean that the listeners used the same strategy in both conditions.

A specific examination of the results revealed a systematic difference in the detectability due to the modification direction in the V-onset condition. A similar systematic difference in *d'* was also observed in the entire word condition but not in the V-offset condition. This discrepancy suggests that the listeners' strategy used in the V-onset condition was similar to that in the entire word condition but differed from that in the V-offset condition. The most reliable cue for changes in the entire word duration seems to be changes in the speaking-rate. Therefore, if the speaking-rate criterion was truly used in the entire word condition, this criterion was probably used also in the V-onset condition and not used in the V-offset condition. To test this assumption, experiment 2 restricted the listeners' task to speaking-rate estimation.

5.3 Experiment 2: Speaking rate estimation

The purpose of experiment 2 was to test whether there is any difference in listeners' speaking-rate estimation performance between the V-onset and V-offset conditions.

5.3.1 Method

Subjects

Eight adults with normal hearing participated in experiment 2. All of the subjects participated in experiment 1, and, therefore, there was a period of three month between the two experiments to prevent the carry-over of judgment strategies as much as possible.

Stimuli and their presentation

The stimuli and their presentation procedures were the same as in experiment 1 except that there was no additional control trial in which the presented two stimuli were identical.

Task

The subjects were asked to estimate the speaking rate of the second word of each of the paired stimuli compared to that of the first word using eleven numerical categories from -5 to +5; a higher number corresponded to a faster rate. Each subject estimated each stimulus pair ten times in total. The obtained responses were pooled over all subjects for each category, and then each stimulus was mapped on a unidimensional psychometric scale in accordance with Torgerson's Law of Categorical Judgment (Torgerson, 1958).[2]

5.3.2 Results

The scaled estimation scores for the speaking rate of each stimulus of the entire word condition, i.e. the reference condition, are plotted as a function of change in the total word duration in Fig. 5.3(b). The estimated speaking rate was highly correlated with the change in the total duration ($r = -0.97$); the speaking rate decreased in proportion to the change in the total duration. This relation demonstrates that the subjects' judgments on the speaking rates were based on the physical rate or total duration.

The scaled estimation scores for the speaking rate of each stimulus of the V-onset condition are plotted as a function of change in V_{on}-V_{on} in Fig. 5.3(a). Although the word durations remained unchanged, the estimated speaking rate varied in inverse proportion to the change in V_{on}-V_{on} ($r = -0.91$); this showed the same tendency as the reference condition.

[2]This is a method of psychological scaling that uses the outputs of a rating scale method. Each of the categorical boundaries and stimuli used in the ratings is mapped on a unidimensional interval scale.

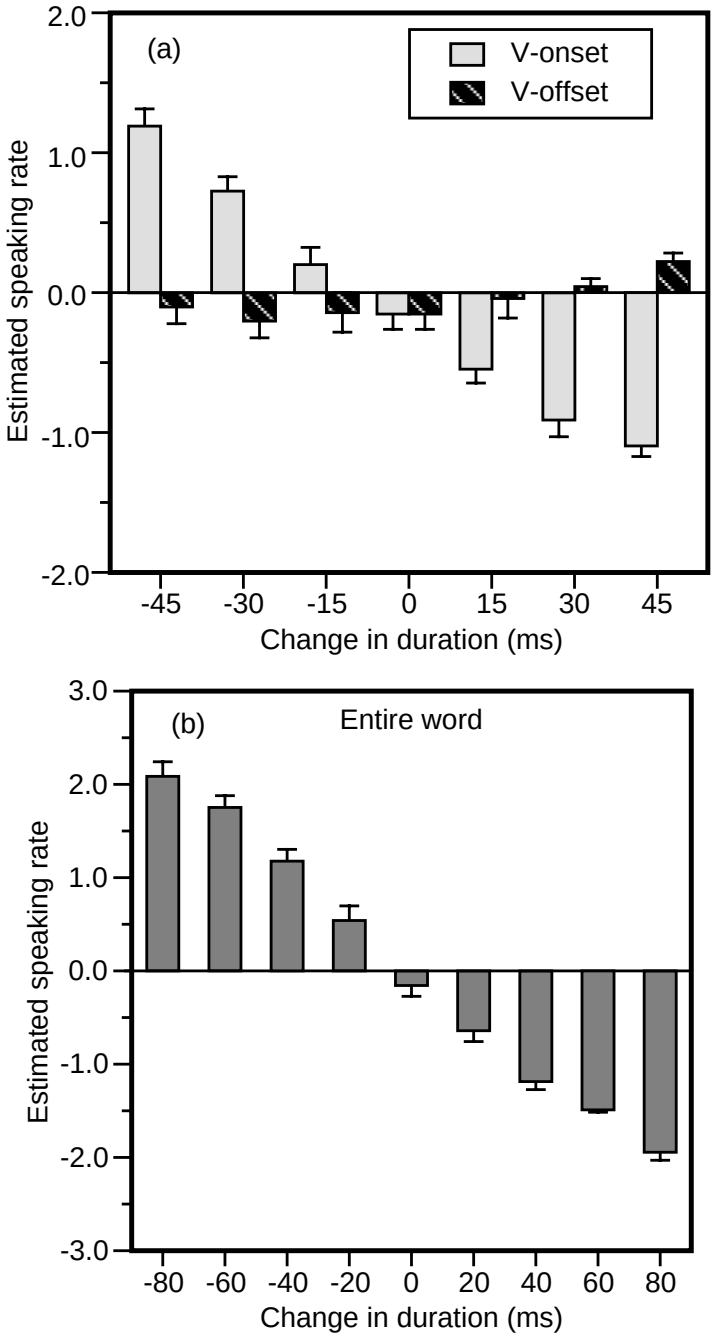

Figure 5.3: Means of estimated relative speaking rates (and standard errors) of modified stimuli to their corresponding unmodified stimuli for each modification type, as a function of change in the (a) total inter-V-onset/offset interval or (b) entire word duration.

In the V-offset condition, on the other hand, the correlation between the estimated speaking rate and the change in V_{off}-V_{off} was low ($r = 0.30$) as shown in Fig. 5.3(a). An ANOVA with the token being a blocking factor showed the effect of temporal modification, V_{off}-V_{off}, on the estimated speaking rate to be significant [$F(5, 45) = 5.12, p < 0.001$]. However, multiple comparisons using Tukey–Kramer's HSD indicated that the significant differences in the speaking rate existed only between the $+45$ ms level and either the -15 ms or -30 ms level ($p < 0.05$). We, therefore, are prudent in stating that the estimated speaking rate shows a consistent relation with V_{off}-V_{off}.

5.3.3 Discussion

In the speaking-rate estimation, a clear difference could be observed between the V-onset and V-offset conditions. Perceived speaking rates slowed as V_{on}-V_{on} increased, even though only slight linear relations were observed between perceived speaking rates and V_{off}-V_{off}. This finding along with the results of Kato *et al.* (1997) suggest that listeners tend to depend on V-onset locations if they are required to process temporal patterns distributed in a global range rather than just a local (two segments or a mora) range. This notion agrees with those in previous studies that the V-onsets are crucial in perceiving speech timing.

5.4 General discussion

Experiment 1 used a detection task and showed that the listeners equally performed for both V-onset and V-offset displacements. Experiment 2 used speaking-rate estimation and showed that the contributions of V-onset displacements were much larger than those of V-offset displacements.

As mentioned before, these two tasks differed in the extent of time that the listeners had to process. That is, the detection was a locally processed task, which could be achieved if any local difference were found between the standard and comparison stimuli, while the speaking-rate estimation was a global task, which required the listeners to compare the global relationships of temporal markers distributed over whole stimuli.

Therefore, roughly speaking, the current experimental results imply that V-onsets and V-offsets equally contribute to locally processed tasks while the importance of V-onset locations increases when a global processing is required. This notion agrees with those in previous studies that the V-onsets are crucial in perceiving speech timing.

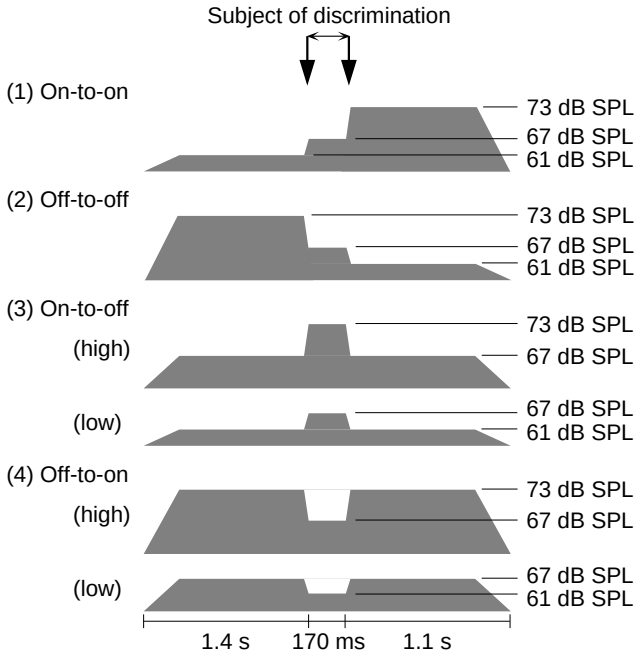

Figure 5.4: Schematic examples showing amplitude envelopes of non-speech stimuli used in Kato and Tsuzaki (1998).

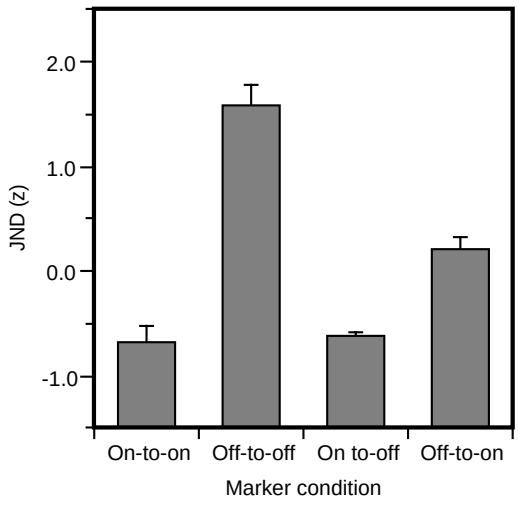

Figure 5.5: Normalized just noticeable differences showing their dependencies on combinations of temporal markers (from Kato and Tsuzaki (1998)).

Table 5.2: Cue intervals for the listeners to achieve each experimental task under each modification condition. The reliable cue intervals are also listed (by applying the criterion of how accurate each cue interval is measured, according to Kato and Tsuzaki's (1998) study).

Modification type	Task	Cue interval(s)	Reliable cue(s)
V-onset intervals	Detection	on-to-on, on-to-off, off-to-on	on-to-on, on-to-off
	Speaking rate	on-to-on	on-to-on
V-offset intervals	Detection	off-to-off, on-to-off, off-to-on	on-to-off
	Speaking rate	off-to-off	none

To obtain a deeper understanding about the mechanisms underlying the functional differences between V-onsets and V-offsets, we introduce a non-speech study on the perceptual measurement of time intervals marked by rising amplitude changes (onsets) and/or falling amplitude changes (offsets) (Kato and Tsuzaki, 1998). This study measured temporal discriminability using 1 kHz tone stimuli whose amplitude contours are described in Fig. 5.4 and obtained jnd data as in Fig. 5.5. To summarize the results, the listeners were very accurate in both on-to-on and on-to-off measurements, much less accurate in off-to-on measurements, and extremely inaccurate in off-to-off measurements (or it was almost impossible to measure this interval).

These results suggest that offset markers are utilized only in combination with onset markers as their counterparts. In other words, offset markers can be effectively used in the measurement of within-event durations; otherwise, especially in the measurement of between-event timings, they are not effective, because the on-to-off interval is usually regarded as the duration of an auditory event (Tsuzaki and Kato, 1998). Onset markers, on the other hand, can be effectively used in both measurements of within-event durations and between-event timings.

Table 5.2 lists the effective cues to achieve each task under each stimulus condition in the current two experiments. Referring to the discriminability data of the above non-speech study (Kato and Tsuzaki, 1998), both the V-onset and V-offset conditions allowed the listeners to use at least one reliable cue in the detection task. In the speaking-rate estimation task, however, the only cue to show the changes in the (physical) global rate in the V-offset condition, i.e., the inter V-offset interval, was not reliable with regard to the accuracy of measurement while it was reliable in the V-onset condition. This notion can

totally account for the observed phenomena in the current two experiments.

The motivation taking the correspondence between V-onsets/offsets and non-speech on/off-markers can be found in our previous studies. Kato *et al.* (1997, 1998a) suggested that quite a few aspects of perceptual sensitivity to temporal modification of a given speech signal can be accounted for by the loudness contour of that signal. In general, a V-onset is a rising marker and a V-offset is a falling marker in terms of the loudness contour expression [see also Appendix B].

Chapter 6

Psychoacoustical evidence for the factors affecting perceived temporal distortions of speech

Abstract

This chapter intends to provide psychoacoustical evidence for the factors (found in Chapters 2 through 5) affecting the subjective evaluation of temporal modifications in speech segments. Section 6.1, at first, presents evidence for the correlation between the acceptability measure, which was used in the speech cases, and the discriminability measure, which is precisely defined in psychophysical terms and is commonly used in the studies of the following four sections. Section 6.2 provides psychophysical evidence for the effect of the temporal position in a word observed in Chapter 2. In Section 6.3, we show an effect the stimulus intensity has on duration discrimination that can account for the effect of the different segment types observed in Chapters 2 and 3. Section 6.4 replicates the effect of loudness jumps on the perceptual salience of temporal markers (observed in Chapter 4) using non-speech stimuli. Section 6.5 presents possible evidence accounting for the functional difference between vowel onsets and vowel offsets observed in Chapter 5. Each of the studies introduced in the current chapter can be regarded as a new finding or a novel idea in psychoacoustics, and assures the universality of the perceptual effects found in the speech experiments.

6.1 Correlation between discriminability and acceptability [1]

6.1.1 Introduction

The purpose of this section was to show direct evidence for the correlation between the discriminability of segmental durations and the acceptability of their modifications.

The results of acceptability evaluation experiments have shown that there are quite a few factors largely affecting the subjective evaluation of temporal modifications given in speech segments, e.g., the temporal position of the segment in the word and the vowel quality of the segment (Kato *et al.*, 1997,1998a,1998b). These experiments, however, have not been sufficient to conclude that these factors actually affect the perceptual "sensitivity," which would be measured by discrimination thresholds. Evaluations on acceptability can be influenced more by a rather higher or "cognitive" process than a perceptual process. Accordingly, one can say that the factors found play their roles only at the "cognitive" stage rather than at the "perceptual" stage.

In the current study, therefore, we measured discrimination thresholds for durational modification using part of the stimuli employed in the acceptability evaluation. First, we examined whether the factors affecting the perceptual evaluation of temporal distortions also affect temporal discriminability in the same way. From among the factors found in the previous experiments, we focused on the factors of temporal position in a word and vowel quality. Then, we directly compared the discriminability data and the acceptability data assuming the same speech materials.

6.1.2 Experiment 1: Discrimination threshold

Design

Adding to the factors of position in a word and vowel quality, the factor of F0 contour was included. When a segmental duration is modified, the F0 contour of the segment is also modified. This difference can be a cue for discrimination. We chose the contrast between a segment with a natural F0 and one with a flat F0 as the third factor.

Stimuli

Two words were chosen from the ATR speech database (Sagisaka *et al.*, 1990), i.e., *shinagire* (sold out) and *nameraka* (state of being smooth). They are commonly used four-mora

[1] Part of this section is published in Kato, Tsuzaki,and Sagisaka (1992).

Japanese words without doubled vowels, geminated consonants, or moraic nasals which make heterogeneous syllable structures and, as a result, may disturb the temporal regularities observed in open syllable sequences. The selected words were spoken naturally in isolation by one male speaker and were digitized at a 12 kHz sampling frequency and with 16-bit precision. The target segments whose durations were subjected to be modified were /i/ in the first and third moras of *shinagire* and /a/ in the first and third moras of *nameraka*.

The temporal modifications were made by a cepstral analysis and resynthesis technique with the log magnitude approximation (LMA) filter (Imai and Kitamura, 1978), and were carried out at 2.5 ms frame intervals. The durational changes were achieved by deleting or doubling every n-th frame in the synthesis parameters throughout the whole vowel. Each target vowel duration was shortened or lengthened over a range that extended from -60 ms to $+60$ ms from the original duration in 2.5 ms steps, resulting in 49 different modification steps. These stimuli were synthesized with either a natural or flat F0 contour; the F0 value of the latter case was fixed to the mean of the original F0 of the target segment. In total, 392 word stimuli were prepared, i.e., (48 modification steps + 1 unmodified) $\times$ 2 temporal positions (1 and 3) $\times$ 2 vowels (/a/ and /i/) $\times$ 2 F0 conditions (natural and flat).

Procedure

Discrimination thresholds were measured by the up-down paradigm with two response alternatives; “same” or “different.” The subjects were presented two stimuli which differed only in the duration of one of the four moraic segments. Four series of stimulus pairs were randomly presented to prevent prediction of the position of the target segment. As a check, trials with physically identical pairs were inserted occasionally, to prevent too short an estimation of the threshold. Each series was tested with both a natural and a flat F0 contour.

Subjects

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

Results and discussion

Table 6.1 shows the mean and the standard deviation of the measured discrimination thresholds for each segment. A three-way factorial ANOVA of repeated measures was performed

Table 6.1: Means and standard deviations of the measured discrimination thresholds for each target segment. The target segments are underlined. DT^+, DT^-, and DTR denote the average discrimination thresholds for the lengthening direction, shortening direction, and the average discrimination threshold range (= $DT^+ + DT^-$), respectively. The values in parentheses are the standard deviations.

Segment	Position	Vowel	F0	DT^+ (ms)	DT^- (ms)	DTR (ms)
sh<u>i</u>nagire	1	/i/	natural	25.4 (4.4)	26.8 (10.2)	52.2 (7.9)
			flat	40.8 (13.5)	30.4 (6.7)	69.0 (13.6)
shinag<u>i</u>re	3	/i/	natural	36.9 (8.2)	30.7 (14.5)	67.6 (14.4)
			flat	40.5 (12.8)	37.3 (18.6)	77.8 (20.5)
n<u>a</u>meraka	1	/a/	natural	22.9 (5.5)	18.7 (9.1)	40.4 (11.6)
			flat	27.9 (9.7)	22.1 (12.3)	51.1 (9.0)
namer<u>a</u>ka	3	/a/	natural	28.9 (5.3)	37.9 (12.0)	67.7 (15.9)
			flat	36.0 (7.1)	40.6 (13.0)	75.0 (16.9)

with position in a word, vowel quality, and F0 contour as the main factors, and with subject as the blocking factor. The main effects of position in a word, vowel quality, and F0 contour on the range between discrimination thresholds (DTR) were significant [$[F(1,51) = 25.5, p < 0.0001$; $F(1,51) = 4.72, p < 0.04$; $[F(1,51) = 9.08, p < 0.004$, respectively]. Each tendency was obtained as follows:

- temporal position: DTR(1st) $< DTR$(3rd)
- vowel quality: DTR(/a/) $< DTR$(/i/)
- F0 contour: DTR(natural F0) $< DTR$(flat F0)

No interaction among the three factors was significant.

As shown above, the tendencies of the first two factors agreed with those observed in the acceptability evaluation (Kato *et al.*, 1998a). Such agreements imply the correlation between the temporal discrimination and the acceptability of temporal distortions. One, however, should be careful in generalizing this implication because the number of word samples employed in the discrimination experiment was not sufficiently large to claim the tendencies observed. We, therefore, performed direct comparisons between the discriminability and the acceptability to confirm their correlation.

6.1.3 Experiment 2: Acceptability evaluation

To investigate the relationship between discrimination threshold and acceptability more precisely, we also measured acceptability using the same stimuli and subjects employed in the discrimination threshold measurement of experiment 1.

Stimuli

The original speech materials and the synthesis procedure were the same as in experiment 1, except that the modification range and step were -50 to $+50$ ms and 5 ms, respectively, which were identical with those in Kato *et al.* (1998a). The stimuli with flat F0's in experiment 1 were not used.

Procedure

The experimental procedures were the same as those in our previous studies (Kato *et al.*, 1997, 1998a, 1998b).

Subjects

Eight adults with normal hearing participated in experiment 2. All of them also participated in experiment 1.

Results and discussion

The vulnerability index was calculated in accordance with Equation 2.1 (or 3.1) for each of four target segments and for each of eight subjects, resulting in 32 α values. A smaller vulnerability index implies a narrower acceptable range. Therefore, a negative correlation was expected between the discrimination threshold range (DTR) and vulnerability index if there was any positive relationship between the acceptability evaluation and detection of a given distortion.

Correlation analyses were performed on the results obtained from experiments 1 and 2. A negative correlation was found between the range between the discrimination thresholds (DTR) and the vulnerability index (α), where Pearson's product-moment correlation coefficient r was -0.556, as shown in Fig. 6.1.

To examine the correlation between discriminability and acceptability from another aspect, we compared the center of the range between the discrimination thresholds and

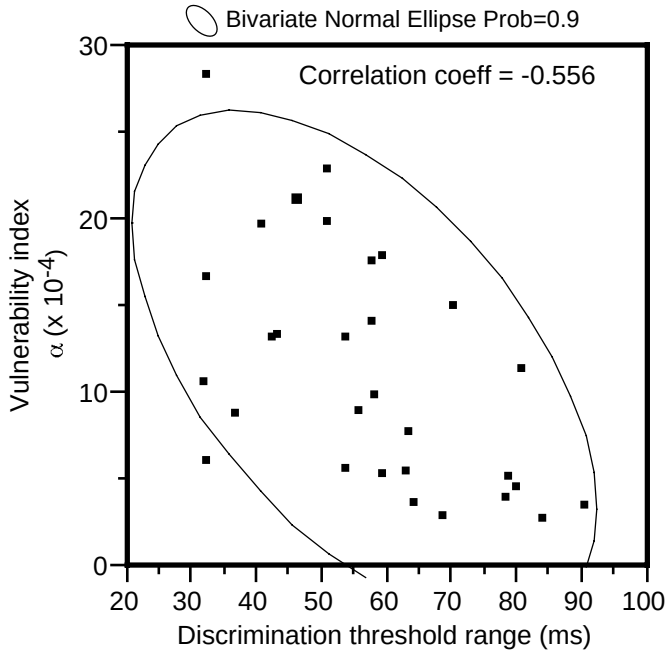

Figure 6.1: Correlation between discrimination and acceptability threshold (a): A negative correlation between the threshold range and the vulnerability index.

the axis of the acceptability curve, the parabolic fitting to the experimentally obtained acceptability evaluation scores, of which the second-order coefficient was taken as the vulnerability index. Since a parabolic curve is symmetrical over its axis, the horizontal position of the axis can be regarded as the center of the acceptable range (see Fig. 6.2). We, then, found a positive correlation between the center shift of the range between the discrimination thresholds and the axis shift of the acceptability curve, where Pearson's product-moment correlation coefficient r was 0.563, as shown in Fig. 6.3.

6.1.4 Summary

The current study successfully demonstrated the correlation between the measures of temporal discrimination and acceptability of temporal modification. Experiment 1 measured the discrimination thresholds of vowel durations in the word context and showed that the listeners' sensitivity was higher (1) to the vowel /a/ than to the vowel /i/, (2) to the first moraic segments than to the third moraic segments, and (3) to the segments with natural F0's than to those with flat F0's. The first two findings were in good agreement with those

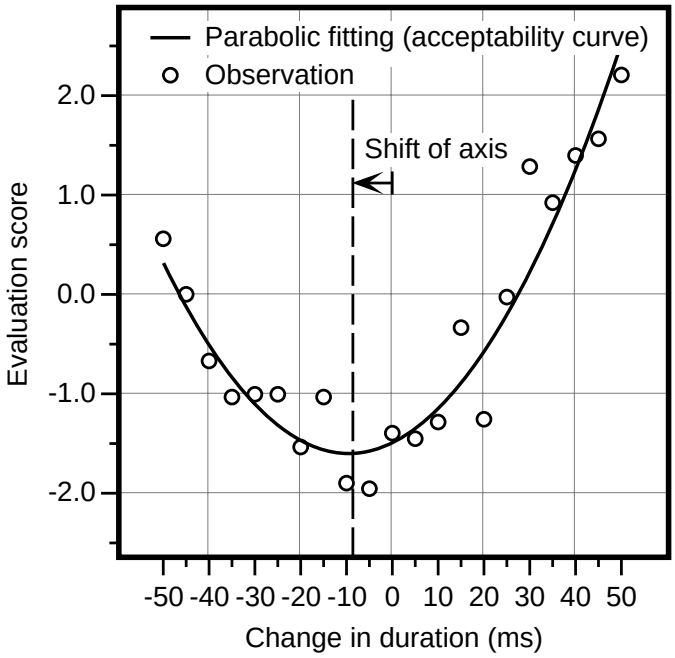

Figure 6.2: An example of the acceptability curve. The curvature represents the acceptable range of temporal modifications. The shift of the axis represents the deviation of the most "acceptable" segment duration from the original.

observed in the acceptability evaluations of previous studies. Experiment 2 measured the acceptability (vulnerability index) of temporal modifications using the same segments and listeners as in the first experiment. Direct comparisons of the results of experiments 1 and 2 revealed that there was a negative correlation between the discrimination threshold range (the range between the upper and lower discrimination thresholds) and the vulnerability index which inversely reflects the acceptable range. It was also revealed that there was a positive correlation between the center shift of the discrimination threshold range and the axis shift of the acceptability curve which reflects the center shift of the acceptable range.

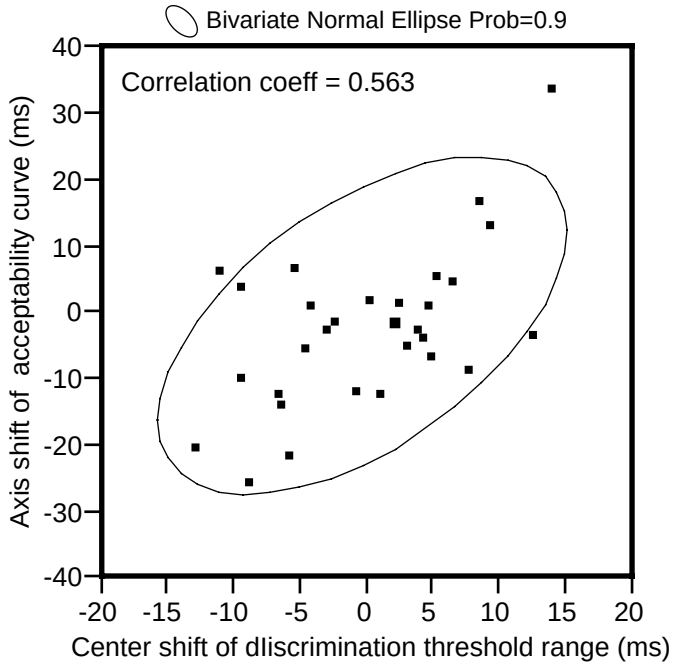

Figure 6.3: Correlation between discrimination and acceptability threshold (b): A positive correlation between the center shift of the threshold range and the axis shift of the acceptability curve.

6.2 Positional effect — Chapter 2 [2]

6.2.1 Introduction

The purpose of this section was to examine the effect of temporal position in a word, which was observed in Chapter 2, on perceptual sensitivity to temporal modification in single speech segments. To test the effect found under controlled experimental conditions, we achieved temporal structures observed in word stimuli using non-speech click sequences.

Speech and music, for example, consist of successive intervals. Although speech and music have different physical characteristics, we can relate singing and hand clapping with musical rhythm. This suggests the possibility of extracting the temporal structure from each different stimulus and comparing all of them in terms of a common mental representation. It is possible to hypothesize the existence of such a common mental representation by finding common rules for the general perception of time intervals regardless of whether speech or non-speech stimuli are involved.

Hirsh *et al.* (1990) studied the human's ability to detect the timing deviation of a single element in a periodic series of tones. Using click sequences, they reported that the detectability of an irregularity in the final interval was higher than that in the initial interval for a standard interval of 50 ms. Lehiste (1979) also studied the detectability of temporal deviations in a single interval within sequences of four isochronous intervals using noise sequences separated by clicks. She observed the highest detectability for the third intervals and the lowest detectability for the initial intervals for all base durations, i.e., 300, 400, and 500 ms. Kato *et al.* (1992, 1998a), on the other hand, studied perceptual sensitivity to the segmental duration in words; the length of a single mora in Japanese words was modified. They reported a higher sensitivity to an initial mora position than to an intermediate mora position in words in both a discrimination test (Kato *et al.*, 1992) and an acceptability evaluation (Kato *et al.*, 1998a).

Although these two groups of studies apparently contradicted one another, there were differences in the types of employed stimuli such as speech or click sequences, and differences in the procedures to detect irregularities within a sequence and to detect (Kato *et al.*, 1992) or evaluate (Kato *et al.*, 1998a) the difference of a test sequence from the standard.[3]

[2] This section is an extended version of Tanaka, Tsuzaki, and Kato (1994) with data from Tanaka, Tsuzaki, and Kato (1992).

[3] Although Kato *et al.* (1998a) used a single-stimulus paradigm in the acceptability evaluation experiment, the subjects were not able to achieve their tasks without referring to their own "internal" standards for each presented speech token [see also Subsection 2.2.3 for details of the experimental procedure].

Table 6.2: The durations of three intervals for each stimulus condition in ms. A (Het) and B (Het) denote the sequences simulating the temporal structures of Japanese words *shinagire* and *nameraka*, respectively. A (Hom) and B (Hom) denote the homogeneous, i.e., physically isochronous, sequences comprised of the average intervals of A (Het) and B (Het), respectively.

Stimulus condition	Temporal position		
	$t1$	$t2$	$t3$
A (Het)	185	160	175
A (Hom)	173	173	173
B (Het)	155	182	163
B (Hom)	166	166	166

In this study, we measured the discrimination thresholds for click sequences by a method and procedure similar to those of Kato *et al.* (1992) to investigate whether the difference in stimulus type causes the different results. We also investigated effects caused by the factors of homogeneity and base words, in addition to the position of the element modified in the sequence.

6.2.2 Method

Subjects

Seven adults with normal hearing participated in the experiment.

Stimuli

A standard stimulus consisted of three consecutive empty intervals ($t1, t2, t3$) divided by four clicks. Table 6.2 shows the physical durations of these intervals for four types of standard stimuli. A(Het) and B(Het) each consisted of a sequence simulating the temporal structure of a Japanese word, *shinagire* or *nameraka*. These two words were previously used in experiments by Kato *et al.* (1992). The clicks (duration markers) for the starting points of individual moras were located at power-dips close to the starting points of individual consonants. A(Hom) and B(Hom) were sequences comprised of the average durations for the three intervals of A(Het) and B(Het), respectively.

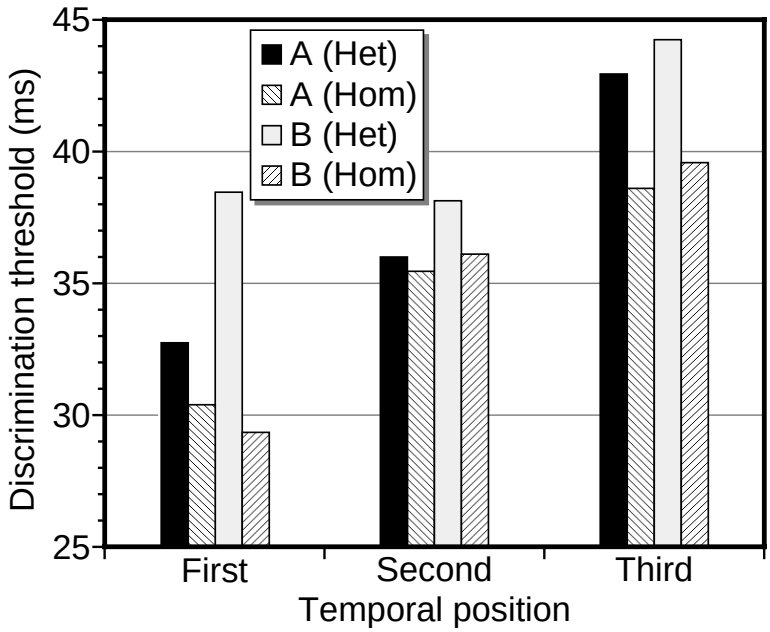

Figure 6.4: Discrimination threshold for the temporal interval of a click sequence pooled over seven subjects as a function of the temporal position, homogeneity, and base words.

The comparison stimulus had the same sequence as the standard stimulus, and the physical duration of either the first, second, or third empty interval ($t1, t2$, or $t3$) was modified from -60 ms to $+60$ ms in 2.5 ms steps for each standard sequence. The duration markers had a single rectangular waveform of 1 kHz, and had a duration of 1 ms.

Procedure

These markers were presented diotically through headphones (SR-Λ Professional, driven by SRM-1 MkII, STAX). Each of them had a sound level measured by an artificial ear (Type 4153, Brüel & Kjær), of about 78 dB (A-weighted). The subjects were required to respond whether the two stimuli were the "same" or "different." The up-and-down method was used to measure the discrimination threshold.

6.2.3 Results and discussion

The discrimination thresholds were calculated by averaging the upward threshold and its downward counterpart. A repeated measures analysis of variance (ANOVA) (3 modified

positions $\times$ 2 base words $\times$ 2 homogeneity/heterogeneity factors), with the subjects as the blocking factor for these discrimination thresholds were performed. The significant main effects were the modified position $[F(2, 66) = 15.05, p < 0.0001]$ and the homogeneity $[F(1, 66) = 8.89, p < 0.0040]$. No significant interaction was observed among the factors. Figure 6.4 shows the discrimination thresholds obtained; they are averaged for the seven subjects.

The result concerning the effect of the modified position, i.e., the discrimination threshold increased with increasing temporal position, agreed with the experimental results obtained by Kato *et al.* (1992). This suggests the possibility that a common mechanism works to perceive the timing structures of speech and click sequences.

The results of Hirsh *et al.* (1990), however, indicated no significant effect by the position for a standard interval of 200 ms, which approximately corresponds to the temporal interval used in our experiment. Lehiste (1979) even reported the reverse tendency. This difference might have been caused by the different procedures used in the two groups of experiments; both Hirsh *et al.* and Lehiste investigated the human's ability to detect irregularities in temporal patterns and we investigated the human's ability to discriminate two temporal patterns.[4]

Adding to the positional effect, the discrimination thresholds for sequences simulating the temporal structures of words were larger than those for homogeneous sequences. The physical variance in the heterogeneous conditions was able to increase the variance of a mental representation and increase the threshold. The effect of homogeneity suggests the discrimination process was affected by the other intervals surrounding the target interval in the sequence, although physical differences in the two temporal patterns could only be found in the target interval. Assuming that the degree of influence by the other intervals was larger for the later intervals than the earlier intervals, we could also explain the results that the discrimination threshold correlated with the temporal position.

[4] ten Hoopen *et al.* (1996) recently reported experimental results that apparently disagree with this explanation. They replicated part of the current study in both "without standard" (single stimulus) and "with standard" (paired comparison) conditions and observed no difference between these two conditions. The difference between ten Hoopen *et al.*'s results and ours can be attributed to the homogeneity of stimulus conditions. As ten Hoopen *et al.* only used homogeneous, i.e., physically isochronous, conditions, their subjects could always rely on a single cue, i.e., the detection of irregularities, throughout an experimental session even in "with standard" conditions. In the current study, on the other hand, as we randomly changed homogeneous and heterogeneous conditions in every trial, the subjects could hardly use the irregularity cue even in "homogeneous" trials. This implication, however, still remains to be confirmed by additional experiments.

6.2.4 Summary

On the discrimination of temporal structures in two sequences, we found that (1) the discrimination for an earlier interval is more sensitive than that for a later interval in a sequence, and that (2) the discrimination thresholds for heterogeneous conditions are larger than those for homogeneous conditions.

6.3 Intensity effect — Chapters 2 and 3[5]

6.3.1 Introduction

This section reports on an intensity effect found in the duration discrimination of auditory stimuli temporally flanked by other sounds. Such intensity effect has not been observed in previous studies which have employed isolated durations.

Kato *et al.* (1992,1998a) found that perceptual sensitivity to durational change is dependent on the intensity, using speech segments as stimuli. Up to that time, no clear evidence showing intensity-dependency on duration discrimination had ever been reported (see, e.g., Allan and Kristofferson, 1974, for a detailed review). The difference between the previous studies and Kato et al.'s studies is that the former usually dealt with a non-speech signal presented only in isolation while the latter used a segment in spoken words; i.e. the target is preceded and succeeded by adjacent segments.

This difference has two major aspects; one is concerned with whether the stimulus is speech or non-speech and the other is concerned with whether the target is isolated or flanked by other sounds. In the current study, we address the latter aspect. An experiment using non-speech stimuli was designed to test whether an intensity effect on duration discrimination can be clearly observed in the presence of flanking sounds. For this purpose, the performance on temporal discrimination was evaluated for tones at several levels (including silence) preceded and succeeded by long tones.

6.3.2 Method

Subjects

Six adults with normal hearing participated in the experiment.

Stimuli

— Flanker Condition — The target level was 79, 76, 70, 67, 55 dB SPL, or silence. The standard duration of each target was 170 ms including rise and fall slopes. The comparison duration was 7, 22, 37, or 52 ms longer or shorter than the standard, which was chosen on the basis of preliminary experimental results. As shown in Fig. 6.5, each target stimulus (T) was temporally flanked by two tones (F1 and F2). There was no interval either between F1 and

[5]This section is published as Kato and Tsuzaki (1994).

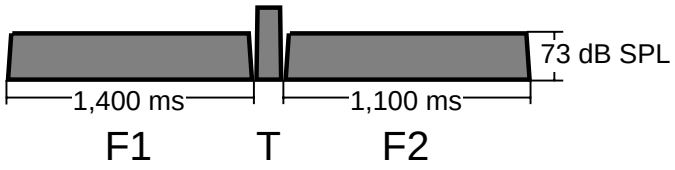

Figure 6.5: Schematic diagram showing the amplitude contour of a stimulus sequence used in the flanker condition. All signals were 1 kHz sinusoids with 10 ms linear rise and fall slopes.

T, or between T and F2. The durations of F1 and F2 were 1400 and 1100 ms, respectively. Both had a level of 73 dB SPL.

— No-flanker Condition (Control Condition) — In this condition, only the target was presented. The target stimuli were the same as in the flanker condition excluding those at 76 and 70 dB SPL which were not employed. For the silent target level, an empty 170 ms interval marked by two identical tone bursts was used; the marker tones each had a duration of 21 ms and a peak level of 79 dB SPL.

All signals were 1 kHz sinusoids with 10 ms linear rise and fall slopes; they were digitally generated by a workstation (SPARC Station 2 or 10, SUN Microsystems) at a 12 kHz sampling rate and with 16-bit precision. The presented levels of the stimuli were adjusted using a sound level meter (Type 2231, Brüel & Kjær) mounted on an artificial ear (Type 4153, Brüel & Kjær).

Procedure

Each subject listened to a pair of standard and comparison sequences, and was required to judge which target was longer. The interstimulus interval between the standard sequence and comparison sequence (ISI) was 1500 ms in the flanker condition and 4000 ms in the no-flanker condition to make the interval between the two target durations constant in both conditions. The durations of flanking tones and ISIs were chosen so as to be sufficiently longer than the target durations and also to not be simple integer proportions of each other. The paired stimuli were randomized and fed diotically to the subjects through a D/A converter (MD-8000 mkII, PAVEC), a low-pass filter (FV-665, NF Electronic Instruments, $f_c = 3$ kHz, -96 dB/octave), and headphones (SR-Λ Professional, driven by SRM-1 MkII, STAX) in a sound-treated room. The order of presentation of the standard and comparison sequences was random but equiprobable within each session.

Table 6.3: Individual and pooled discrimination thresholds in ms and Weber fractions for pooled data.

Condition	Level	Subject ID						Mean	Weber
		A	B	C	D	E	F		fraction
With flankers	79 dB	21.6	11.1	31.4	24.3	20.5	14.9	20.6	0.121
	76 dB	22.4	10.8	64.9	45.1	17.7	15.9	22.6	0.133
	70 dB	27.8	12.2	72.1	25.8	26.8	20.6	30.9	0.182
	67 dB	32.7	14.3	57.4	30.7	30.9	20.0	31.0	0.182
	55 dB	50.4	15.7	95.6	88.8	27.5	26.5	50.8	0.299
	silent	103.7	22.5	103.0	74.3	55.0	39.8	66.4	0.391
No flanker	79 dB	21.5	12.9	22.5	16.3	14.7	15.8	17.3	0.102
	67 dB	17.1	14.3	23.6	18.2	13.3	20.2	17.8	0.105
	55 dB	14.5	12.9	27.4	22.6	14.1	21.9	18.9	0.111
	silent	19.7	12.3	21.7	17.6	19.2	23.4	19.0	0.112

In this experiment, there were fourteen sessions for each of the flanker and no-flanker conditions, the first one being for training. In each session, each pair of standard and comparison sequences was repeated four times. Accordingly, each point on the psychometric function comprised 52 judgments for each subject. The discrimination threshold was estimated as the point of 75 % correct on the approximation line, fitted according to the least-squares criterion with Müler-Urban weighting, for the obtained probability of a correct response plotted on normal coordinates. Then, the upper and lower 75 % discrimination thresholds were pooled.

6.3.3 Results and discussion

The obtained 75 % discrimination thresholds for each subject and for each experimental condition (flanking target level) are shown in Table 6.3. The data indicates a general tendency in the flanker condition, i.e., for the thresholds to increase as the target levels decrease. However, the range of thresholds varies widely among the subjects. Therefore, each of the discrimination thresholds was normalized by the mean and the standard deviation calculated for each subject. The normalized thresholds are pooled over six subjects and

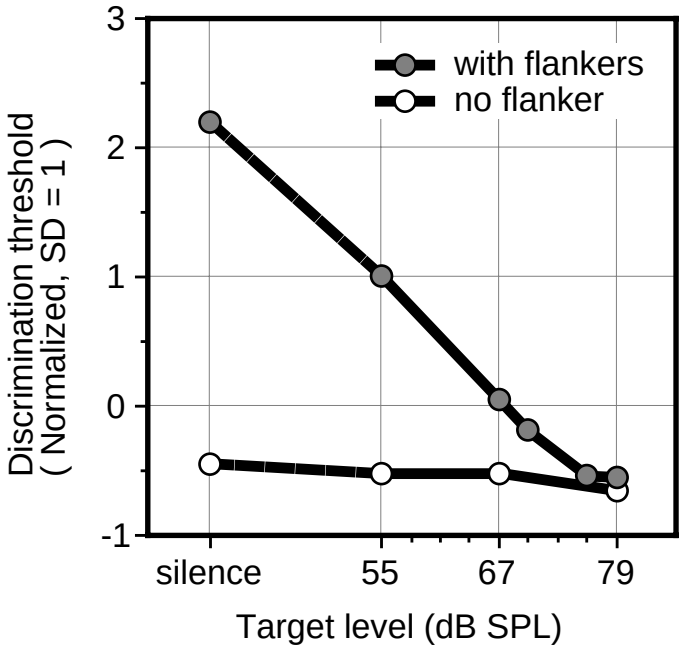

Figure 6.6: Normalized discrimination thresholds pooled over six subjects as a function of the target level.

plotted as a function of the target level in Fig. 6.6 for both conditions. A one-way ANOVA with the factor of target level was performed for the normalized data separately for each of the two flanking conditions.

In the flanker condition, a significant effect of the target level was found $[F(5, 30) = 35.28, p < 0.001]$. Multiple comparisons for all pairs using Tukey–Kramer's HSD indicated the discriminability for silent targets to be significantly worse than that for other conditions and also the discriminability for the 55 dB target to be significantly worse than that for targets at higher levels (p < 0.01 for both). In the no-flanker condition, the target level had no significant effect $[F(3, 20) = 0.42, p = 0.74]$.

The results in the no-flanker condition showing no intensity effect and Weber fractions around 10 %, are in agreement with other published results (e.g. Abel, 1972b). In the flanker condition, on the other hand, a clear intensity effect could be observed even for a 12 dB difference in levels; no effect could be found in previous studies with no flankers for 20 dB differences (Abel, 1972b). Clearly, the observed effect is due to the existence of preceding and succeeding tones and might be unable to be predicted by conventional types of duration discrimination models (e.g., Creelman, 1962) which take no account of influences

from flanking sounds. Further studies should be undertaken to investigate how the flankers function in the perceptual measurement of target durations.

6.3.4 Summary

An intensity effect was found on auditory duration discrimination using a short target tone between long preceding and succeeding flanking tones. The measured discrimination thresholds significantly increased with decreasing target levels; the largest discrimination threshold was obtained in the silence condition where the target portion was an empty interval. Such a systematic intensity effect could not be observed in the control condition where the target was presented with no flanking sounds.

6.4 Effect of loudness jump — Chapter 4[6]

6.4.1 Introduction

The purpose here was to test the effect of a loudness jump, which was observed in Chapter 4, on perceptual sensitivity to the displacement of temporal markers under controlled experimental conditions.

6.4.2 Method

Subjects

Six adults with normal hearing participated in the experiment. All of the subjects also participated in experiment 1 of Chapter 4, and, therefore, there was a period of one month between these two experiments to prevent the carry-over of judgment strategies as much as possible.

Design

The experiment was designed as a four-way factorial one. The first factor was the loudness jump between two modified segments (large jump or small jump). The other three factors were included mainly to test their interactions with the first factor; they were, the direction of the marker slope (rising or falling), the steepness of the marker slope (steep or broad), and the temporal position of the marker in a sequence (first half or second half).

Stimuli

Each stimulus was a 1 kHz tone with one of two types of overall amplitude contours as shown in Fig. 6.7. These two types (type I and type II) were modeled on typical loudness contours of four-mora word stimuli (see Fig. 4.3), and enabled us to complete the factorial design described above. Each stimulus comprised the alternation of slope and steady parts. As in experiment 1 of Chapter 4, temporal markers were defined as rapidly changing parts of a signal in between steady-state (either silence or sounding) parts; i.e., only the slope parts could become temporal markers. The steady parts each had one of the following three levels: 73 dB SPL (9.85 sone), 64 dB SPL (5.28 sone), or silence, where each was employed as an approximation for the average loudness of the vowels, nasals, and pre-burst closures

[6]This section is published as part of Kato, Tsuzaki, and Sagisaka (1997).

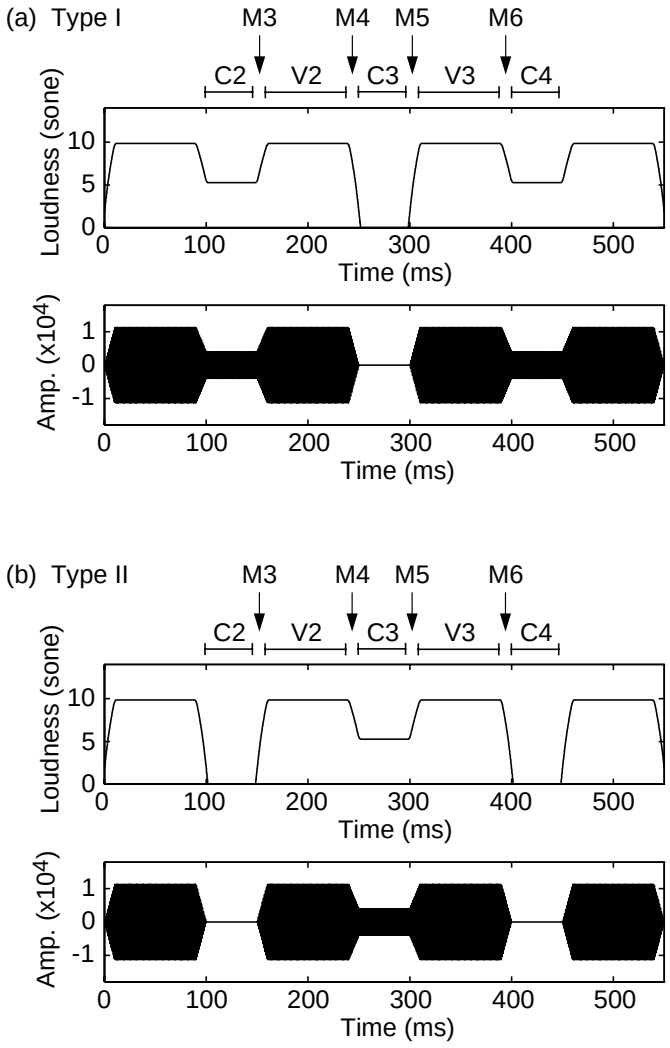

Figure 6.7: Time waveforms and loudness contours of the two types of stimuli used in the experiment. Each "V" or "C" indicates the target part to be modified. Each "M" indicates the location of the temporal marker considered. The level of V is 73 dB SPL (= 9.85 sone), the level of louder C is 64 dB SPL (= 5.28 sone), and that of softer C is "silence." All signals are 1 kHz pure tones. The eight markers in the figures (type I and type II) comprise an orthogonal set for three (loudness jump, slope direction, temporal position) of the four factors considered. The fourth factor (slope steepness) is included by considering another set of type I and II stimuli of which the slope duration is 20 ms (broad slope); that of the above stimuli is 10 ms (steep slope).

found in the speech stimuli used in experiment 1 of Chapter 4. The duration of the slope was 10 ms (steep) or 20 ms (broad). The duration of the loud part (V part in Fig. 6.7) including rise-fall slopes and that of the soft or silent part (C part in Fig. 6.7) were 100 and 50 ms, for the standard stimuli. One of the V durations in each of the comparison stimuli and either its preceding or succeeding C duration were modified in opposite directions with 30 ms for each. The modification target was limited to the steady part in the second V (V2) or the third V (V3) and either of its adjacent C's (Fig. 6.7). Therefore, the marker (transient part) between the modified segments was solely displaced forward or backward by 30 ms from the standard. In total, 32 stimuli were prepared.[7]

Procedure

The detectability index (*d'*) was measured for the difference between each pair of standard and comparison stimuli by the method of constant stimuli. The experimental apparatus was the same as in experiment 1 of Chapter 4. The subjects listened to the standard and comparison stimuli and were asked to rate the difference between them using eight numerical categories: "0" to "7"; a larger number corresponded to a larger subjective difference. Since the stimuli were complex and unfamiliar to the subjects, the experimental trials were preceded by a 1-hour practice session to familiarize the subjects with the stimuli. In each experimental trial, the subjects listened to the presentation of four successive stimuli, the first three each being the standard and the last one being a comparison. This repetition of standard stimuli served to effectively familiarize the subjects with the stimuli. The inter-onset interval of the four stimulus sequences was 1400 ms each which was chosen so as to prevent temporal markers in the standard sequences from coinciding to a perfect isochronous rhythm. Twenty percent of the trials were control trials in which each comparison stimulus was the same as the standard stimulus. Twelve judgments were collected from each subject for each stimulus. The obtained responses were pooled over all subjects for each category, and then the detectability index, *d'*, for each comparison stimulus was estimated in accordance with the Theory of Signal Detection (Green and Swets, 1966).

[7]They were 2 types of amplitude contours (= type I, type II) $\times$ 2 steepness conditions (= steep, broad) $\times$ 2 slope directions (= rising, falling) $\times$ 2 target positions (= first half, second half) $\times$ 2 displacement directions (= forward, backward).

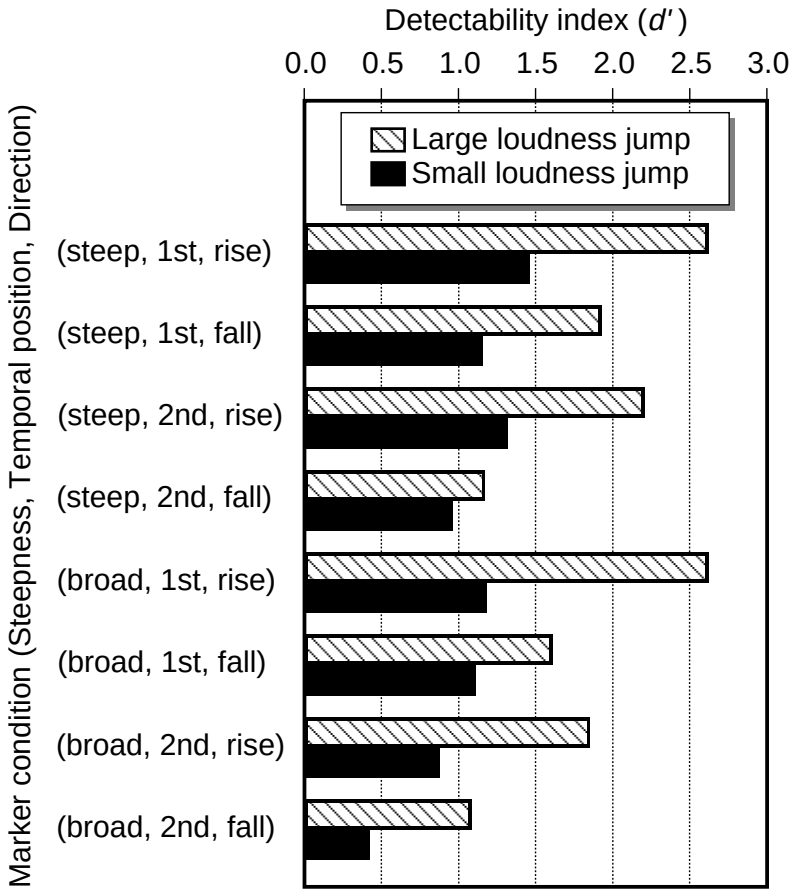

Figure 6.8: Detectability index *d'* for a 30 ms displacement of a temporal marker, for each combination of the marker conditions, as a function of the loudness jump between both sides of the marker. A larger *d'* implies easier detection.

6.4.3 Results and discussion

A four-way completely randomized factorial analysis of variance (ANOVA) was performed for the obtained detectability indices *d'*. The factor of loudness jump was significant $[F(1, 16) = 99.5, p < 0.0001]$. The other three factors also turned out to be significant; they were, the direction of the slope $[F(1, 16) = 52.2, p < 0.0001]$, the steepness of the slope $[F(1, 16) = 6.30, p < 0.05]$, and the temporal position $[F(1, 16) = 36.7, p < 0.0001]$. Besides these main effects, a significant interaction was observed between the factors of loudness jump and slope direction $[F(1, 16) = 7.88, p < 0.05]$. No other interaction was significant.

Figure 6.8 shows *d'* for each combination of the marker conditions pooled over the marker displacement directions as a function of the loudness jump. As clearly shown in the figure, the effect of loudness jump agrees with the observed one in experiment 1 of Chapter 4; i.e., a larger loudness jump causes a higher sensitivity. The effect of temporal position in a sequence was significant; i.e., displacements of the markers in the first half were detected more easily than those of the markers in the second half. This effect is consistent with the finding reported by Tanaka, Tsuzaki, and Kato (1994) that the temporal discrimination for the initial interval is easier than that for the succeeding intervals in a click sequence.

6.4.4 Summary

The current experiment successfully demonstrated the effect of loudness jump on the sensitivity to the temporal displacement of temporal markers or rapid amlitude changes. A larger loudness jump corresponded to a higher detectability of temporal displacements. This tendency agrees with that observed for the acceptability evaluation of durational changes in speech (Chapter 4). The current experiment additionally found three other factors affecting the temporal displacement of temporal markers, i.e., the direction of the slope, the steepness of the slope, and the temporal position in the sequence.

6.5 Effect of on/off temporal markers — Chapter 5[8]

6.5.1 Introduction

This section investigated functional differences between a sudden energy rise and a sudden energy fall (each was assumed to be a temporal marker) in the detection of change in auditory filled durations. If one assumes that the perceptual duration is mentally calculated by measuring the interval between the two temporal markers, i.e., one at the rising point and the other at the falling point, there will be no difference whichever marker precedes the other. In auditory information processing, however, temporal asymmetry is preserved at quite a few stages.

This study aimed to reveal functional differences between the two types of markers. It was hoped the findings would provide information on important factors that affect the perceptual acuity of auditory durations, and would also be useful in predicting the detectability of temporal distortion in complex temporal patterns such as those found in speech.

The two markers considered in the study, i.e., rising and falling markers, are the exact mirror image of the other along the time axis. Therefore, they have the same transition duration and the same amount of level change. However, they do not necessarily have the same perceptual markability; for example, Kato *et al.* (1997) found that listeners were more sensitive to the temporal displacement of a rising marker than to that of a falling marker, in detecting a shift in the temporal position in a sequence of alternating rising and falling markers. This finding suggests that the rising marker is perceptually more salient than its falling counterpart. If such saliency were to affect the perceptual measurement of a given duration marked by rising/falling markers, the duration bounded by two rising markers, i.e., a rise-rise type, would be measured more accurately than that bounded by two falling markers, i.e., a fall-fall type. The first objective of the current study was to test this possible advantage of the rise-rise type over the fall-fall type.

On the other hand, effects of the temporal order of markers have been reported in studies investigating the discrimination of durations bounded by two different intermodal markers, i.e., auditory and visual markers (Grondin *et al.*, 1996). The current study, therefore, also examined whether the temporal order of the two different markers affected the listeners' acuity in measuring durations. This time, the difference was between two intramodal markers, i.e., rising and falling auditory markers.

No temporal order effects of different intramodal markers on duration discrimination

[8]This section is published as Kato and Tsuzaki (1998).

had been reported, and therefore the current design was particularly interesting in that two different viewpoints predicted different results which contradicted each other.

The first viewpoint predicted that the fall-rise type is advantageous. Even when a rise-to-fall duration is equal in acoustical duration to a fall-to-rise duration, the former is likely to have a longer sensory or perceptual duration than the latter as predicted by filled-duration illusion (Goldfarb and Goldstone, 1963). Given Weber's law for time perception, the discrimination threshold in terms of an absolute value can be expected to be smaller for the fall-to-rise duration than the rise-to-fall one. This assumption is in line with the prediction of the *internal-marker hypothesis* (Grondin, 1993) in duration discrimination.

The second viewpoint, in contrast, predicted that the rise-fall type is advantageous. A rising change in the amplitude envelope, in general, coincides with the start of an auditory event while a falling one means the end of an event. Therefore, a temporal interval marked by rising and falling markers (rise-to-fall interval) can be regarded as an attribute of a single event, while the fall-to-rise interval is a relation between two different events. Following the argument by Divannyi and Danner (1977) that temporal discrimination is easier for a duration bounded by markers likely to be perceived as one event than for that bounded by markers unlikely to be perceived as one event, a higher performance can be expected for rise-to-fall durations than for fall-to-rise durations.

This second view was referred to as the global viewpoint, i.e., it took the consequence of the perceptual integration of the two temporal markers into account. On the other hand, the first view, in comparison with the second view, was referred to as the local viewpoint, i.e., it only looked at the individual characteristics of the markers.

In summary, the current study provided a direct comparison of four types of marker combinations, i.e., rise-rise, fall-fall, rise-fall, and fall-rise, in terms of duration discrimination.

6.5.2 Method

Subjects

Five adults with normal hearing participated in the experiment.

Stimuli

All of the stimuli were 1 kHz sinusoids and started with a 100 ms linear rising transient and ended with a 100 ms linear falling transient. They were digitally generated by a workstation (SPARC Station 2 or 10, SUN Microsystems) at a 12 kHz sampling rate and with 16-bit

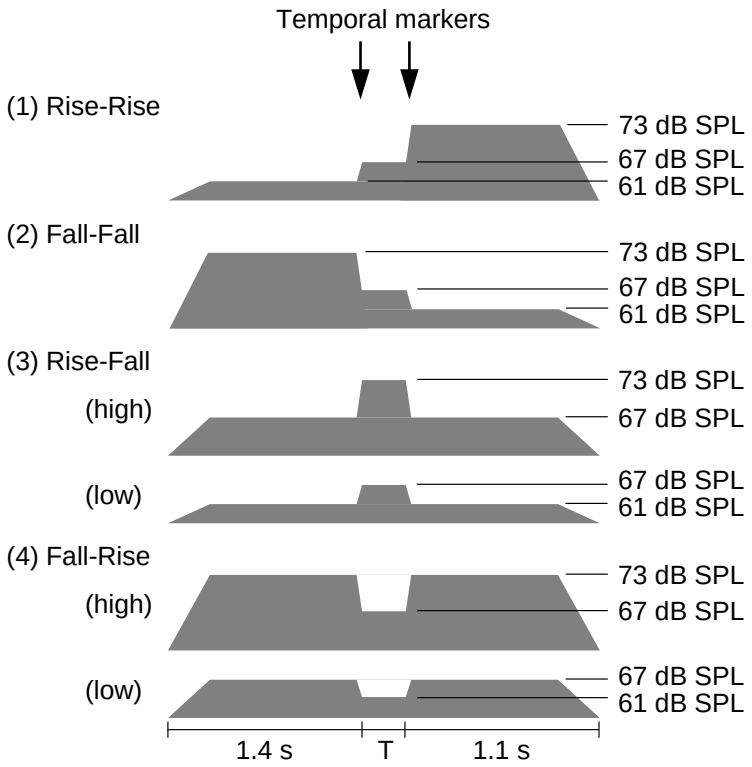

Figure 6.9: Schematic examples showing amplitude envelopes of stimuli in each experimental condition. All signals are 1 kHz sinusoids. T shows the target part of discrimination. Its duration is 170 ms in the standard stimuli, and is either longer or shorter by 8, 24, 50, or 100 ms in the comparison stimuli.

precision.

Each stimulus had two rapid transitions in the amplitude envelope in the middle. The target segment whose duration was subjected to temporal modification was the steady-state part bounded by these two transitions. Each transition was achieved by a linear change in the amplitude envelope in 10 ms. Each change either doubled the amplitude in rising transitions or halved it in falling transitions, i.e., there was a +6 dB or -6 dB change in the sound pressure level. The phase of carrier tones at the starting or ending of each transition was 0 rad. Each of these transitions was referred to as a temporal maker.

There were four combinations (marker conditions) of these rising and falling markers: (1) rise-rise, (2) fall-fall, (3) rise-fall, and (4) fall-rise, as schematically shown in Fig. 6.9. The last two conditions came in two different overall levels because the temporal markers

employed had two different absolute levels, which were necessary to achieve the rise-rise and fall-fall conditions under the restriction that the inter-marker region, i.e., target part, be steady-state.

The durations of the target steady-state parts were 170 ms for the standard stimuli, and were either longer or shorter by 8, 24, 50, or 100 ms for the comparison stimuli. This modification range was chosen on the basis of preliminary experimental runs.

Procedure

Just noticeable differences (jnd's) were estimated for each subject and for each stimulus condition by the method of constant stimuli with a two-interval forced-choice (2IFC) task. The paired stimuli were randomized and presented diotically to the subjects through a D/A converter (MD-8000 mkII, PAVEC), a low-pass filter (FV-665, NF Electronic Instruments, $f_c = 3$ kHz, -96 dB/octave), and headphones (SR-Λ Professional, driven by SRM-1 MkII, STAX) in a sound-treated room. The presentation level was calibrated with a precision sound level meter (Type 2231, Brüel & Kjær) mounted on an artificial ear (Type 4153, Brüel & Kjær). In each trial, the subjects listened to a pair of standard and comparison stimuli which were sequentially presented with a 1.5 s inter-stimulus interval, and were then asked to answer which of the paired stimuli included the longer duration. The correct answer was visually fed back to the subject after each trial. The temporal order of presentation of the standard and comparison stimuli changed randomly trial by trial but was counter-balanced throughout the experimental sessions.

The total experimental run took 13 days for each subject, the first day being for training. Each one-day session comprised of eight subsessions and it took, in total, about one hour including instructions and breaks. Each subject made, in total, 64 judgments for each of the prepared stimulus pairs. The jnd was estimated as the threshold of 75 % correct on the approximation line, and fitted according to the least-squares criterion with Müller-Urban weighting, for the obtained probability of a correct response plotted on normal coordinates. Then, the upper and lower 75 % thresholds were pooled as a single jnd.

6.5.3 Results and discussion

The obtained jnd's for each subject and for each stimulus condition are shown in Table 6.4. As clearly shown in the table, the data indicates a general tendency for the jnd's to be greatest in the fall-fall condition and least in the rise-rise or rise-fall condition. However, the range

Table 6.4: Just noticeable differences in ms for each condition and for each subject.

Marker condition	Overall level	Subject ID					Average	Weber fraction
		A	B	C	D	E		
Rise-Rise		40.5	43.0	26.4	92.7	42.7	50.1	0.26
Fall-Fall		262.1	423.7	135.5	449.4	61.9	255.5	1.34
Rise-Fall	high	36.4	35.4	21.8	76.5	35.9	43.4	0.23
	low	43.7	48.0	30.5	129.7	32.9	53.9	0.28
Fall-Rise	high	114.4	84.0	47.6	157.3	45.9	108.4	0.57
	low	112.6	156.2	39.6	151.6	49.6	128.1	0.68

of jnd's varies widely among the subjects. Therefore, each of the jnd's was normalized by the mean and the standard deviation calculated for each subject. Figure 6.10 shows the normalized jnd's pooled over the five subjects for each marker condition.

A one-way ANOVA of repeated measures showed the effect of the marker conditions on the normalized jnd's to be significant $[F(4, 19) = 114.2, p < 0.0001]$. Multiple comparisons for all pairs of the marker conditions using Tukey–Kramer's HSD indicated the difference between any pair of average jnd's to be significant ($p < 0.01$) except for those between the rise-rise and rise-fall conditions.

The current experimental results showed that the listeners measured tone durations bounded by two rising markers more accurately than those bounded by two falling markers. This advantage of the rising markers over the falling markers in duration measurement probably resulted from the higher perceptual salience of the former markers than that of latter markers as reported by Kato *et al.* (1997).

This tendency of duration discrimination is also supported by physiological data. Amplitudes of evoked potentials have been reported as smaller upon being elicited by auditory stimulations of rising amplitude changes than when elicited by those of falling ones that are mirror images of the rising ones along the time axis, when taking brain stem responses into account (Brinkmann and Scherg, 1979; Kodera *et al.*, 1977). If one assumes that perceptual saliency is related to these evoked potentials, these physiological data are consistent with the observed advantage of rising markers over falling markers in the markability of auditory durations.

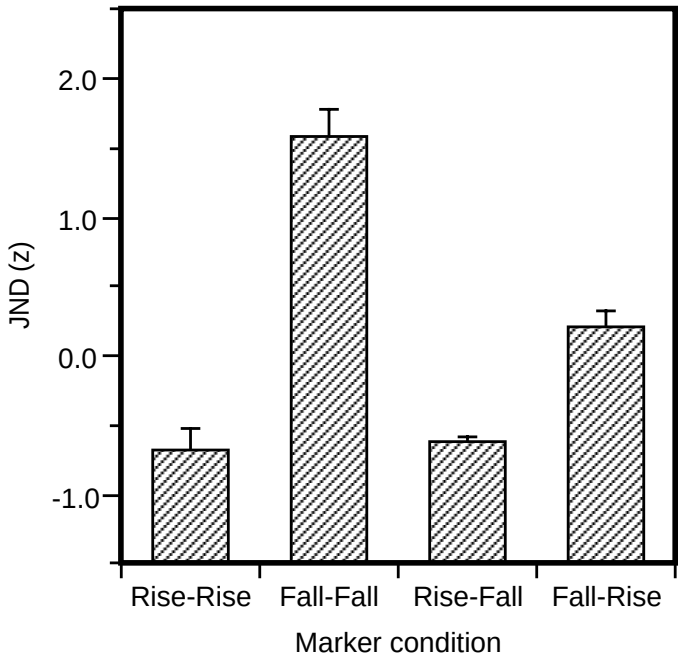

Figure 6.10: Normalized just noticeable differences pooled over five subjects as a function of combinations of rising and falling markers. The error bars show the standard errors.

The results also showed that a greater acuity was obtained for the measurements of the rise-to-fall durations than for the fall-to-rise durations. The local viewpoint, which takes only sensory or perceptual inter-marker intervals into account, predicts the advantage of the fall-rise type. The results, however, did not agree with this viewpoint. All of this suggests that the factors that are considered in the global viewpoint function more effectively than those in the local viewpoint under the current experimental conditions. Introducing such a global viewpoint may be important if one wants to establish a model of time perception that can cope with human behaviors in a more realistic environment than just a laboratory environment.

6.5.4 Summary

Marker combination effects were found in the discrimination of auditory durations using short target parts of tones marked by rising and/or falling level changes. The obtained jnd's were the smallest for durations marked by two rising changes (rise-rise type) or for those of the rise-fall type; moderate for those of the fall-rise type; and largest for those of the fall-fall

type.

Although the variations in the experimental conditions were limited, they certainly provided direct evidence for (1) the predominance of rising markers to falling markers, and for (2) the predominance of rise-to-fall durations to the reverse case, fall-to-rise durations, in the discrimination of auditory durations.

Chapter 7

A Modeling of Subjective Evaluation for Temporal Distortions of Speech[1]

Abstract

Integrating the results of both the speech perceptual and psychoacoustical experiments in the preceding five chapters, we proposed a modeling of the temporal error evaluation for synthetic rules that can predict, to some extent, the acceptability to humans (a subjective measure) from only objective measures (physical properties) of speech signals. To take into account the perceptual factors in an error measure for evaluation, a description of a speech signal was proposed by simplifying the loudness contour, i.e., the time-loudness marker model. Using this model, the acceptability could be predicted from the changes in the inter-marker intervals in conjunction with the defined perceptual salience of the markers involved. An objective evaluation model achieved in accordance with the proposed time-loudness marker description is presented in this chapter, and an evaluation experiment conducted to test the effectiveness of the model is discussed. The results showed that the proposed model consistently achieved a better prediction (i.e., closer to human evaluation) than the reference model which only used the average acoustic errors without any perceptual consideration.

[1]This chapter is based on Kato, Tsuzaki, and Sagisaka (1999) and Kato, Tsuzaki, and Sagisaka (c).

7.1 Introduction

The quality of any synthesized speech should be evaluated by human listeners who are the final recipients of that speech. Nevertheless, objective evaluations, which are widely utilized in the development stages of rule-based speech synthesis systems, must rely on the examination of merely acoustic errors. This chapter presents a framework to reflect human perceptual characteristics onto the objective evaluation of synthetic speech.

The evaluation methods used in the assessment of rule-based synthetic speech, in general, can be separated into two types, i.e., subjective and objective evaluations. Methods of the former type have, so far, been widely and deeply investigated (Bailly *et al.*, 1992; Kasuya, 1993; Nusbaum *et al.*, 1995; van Santen *et al.*, 1997; COCOSDA, 1998) because this type of evaluation is indispensable to confirm the final performance of any synthesis system. However, the requirements of human and time resources in these method have prevented taking the methods at the development stage.

Methods of the latter type, in contrast, have been widely introduced in the development cycles of speech synthesis techniques, typically in corpus-based ones (Sagisaka, 1998), because they can show "some" criteria of development without human resources. These objective evaluation methods, however, potentially hold a serious problem in that their outcome is not assured to be perceptually valid because their criterion, a physical measure, does not necessarily have a linear relationship with the corresponding perceptual measure, such as naturalness or acceptability. Although this potential inconsistency between physical and perceptual measures has certainly been considered as serious, few systematic studies have been done about the issue besides a few exceptions (e.g. van Wieringen, 1995).

Recently, however, 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, as we have seen in Chapters 2–6. To extend this series of research, the current study tried to build a prototype of an error evaluation model for synthetic rules able to predict acceptability to humans (a subjective measure) from only objective measures (physical properties) of speech signals. Such a model, if it could be achieved, would enable an evaluation with both the advantages of objective and subjective evaluations, i.e., simplicity and perceptual validity.

Firstly, we summarized the major factors yielding the inconsistencies between physical and perceptual measures of temporal structures in speech and illustrated strategies to

compensate those inconsistencies (in the following section). Next, we proposed a framework to incorporate the mentioned factors into the objective error evaluation model, i.e., **the time–loudness marker description**, and built a prototypical model in accordance with the proposed framework. Finally, we performed a simulation of error evaluation to confirm the effectiveness of the model.

7.2 Auditory perceptual characteristics for durational errors

Rules to assign segmental durations have been developed to replicate the durations of natural speech (Sagisaka and Tohkura, 1984; Kaiki and Sagisaka, 1992; Higuchi *et al.*, 1993; Kato and Hashimoto, 1992). They have commonly adopted the sum or average of acoustic errors as the measure of an objective error evaluation. This strategy which minimizes average errors is certainly valid if the final goal is to make a synthetic duration identical to the corresponding original. If some amount of error is allowed, however, the implicit promises of this evaluation criterion need to be examined from the auditory perceptual point of view. The implicit premises of the average error criterion can be summarized into the following two points: (1) a single durational error linearly correlates with the perceived distortion regardless of the attributes of the segment in question, and (2) multiple durational errors affect the perceived distortion independently of each other. We examined these two premises in the light of the perceptual characteristics obtained in Chapters 2 through 6, and found ways to achieve a perceptually valid evaluation measure.

7.2.1 Perceptual weighting of each error

Previous studies reported that perceptual sensitivity to the durational change of a speech segment is affected by the segment quality; e.g., the temporal discriminability of vowel segments is higher than that of consonant segments (Huggins, 1972a; Carlson and Granström, 1975; Bochner *et al.*, 1988), excluding a special case where the durational change also affects the phonemic category (Fujisaki *et al.*, 1975). On the other hand, for the influence of a segment quality, little has been known about the acceptability of a change in the segmental duration, which can be regarded as a more practical measure for the evaluation of a synthetic error, although some other aspects of this measure have been explored by several pioneering studies (Sagisaka and Tohkura, 1984; Carlson and Granström, 1975; Sato, 1977; Hoshino and Fujisaki, 1983).

Kato *et al.* (1998a, 1998b) recently showed, as can be seen in Chapters 2 and 3,

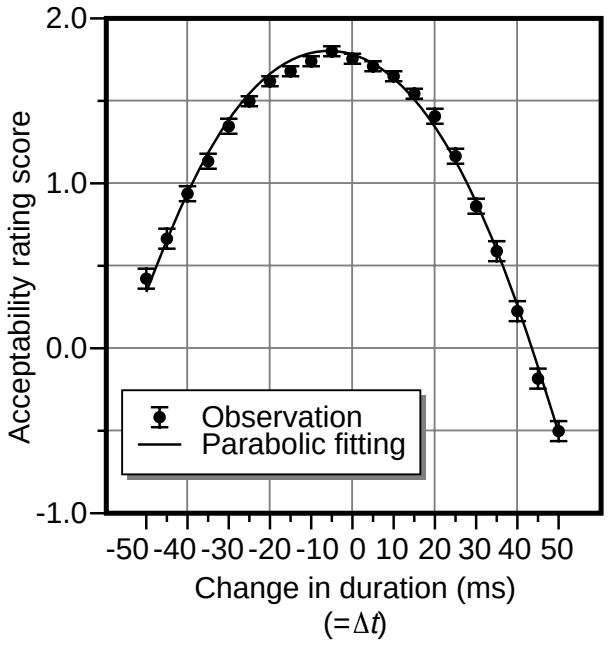

Figure 7.1: An example of an acceptability rating profile as a function of change in the segmental duration. The dots and error bars show the means and standard errors of rating scores by six listeners using 70 segments in words. The parabolic fitting line is superimposed. (Reproduced from Kato *et al.* (1998a).)

that listeners' rating scores of acceptability against changes in segmental durations can be accurately traced by a parabolic curve Fig. 7.1. They also showed that the absolute value of the second-order coefficient of this approximation curve, namely, the vulnerability index is generally larger for vowel segments than that for consonant segments (Fig. 7.2, the left-hand scale). This tendency agrees with previous discrimination studies that vowel durations are more accurately discriminated than consonant durations.

Furthermore, this variation in the vulnerability index has been found to be highly correlated with the loudness that is intrinsic to the segment quality, as shown in Fig. 7.2. A non-speech study on temporal discriminability, on the other hand, showed that an auditory duration with large loudness is more accurately discriminated than a softer duration (Kato and Tsuzaki, 1994). This tendency in the temporal discriminability agrees with that of the acceptability measure found in Fig. 7.2. All of these results suggest that the correlation observed between the vulnerability index (acceptability measure) and the segment loudness can be accounted for as a reflection of the general characteristics of the auditory perception.

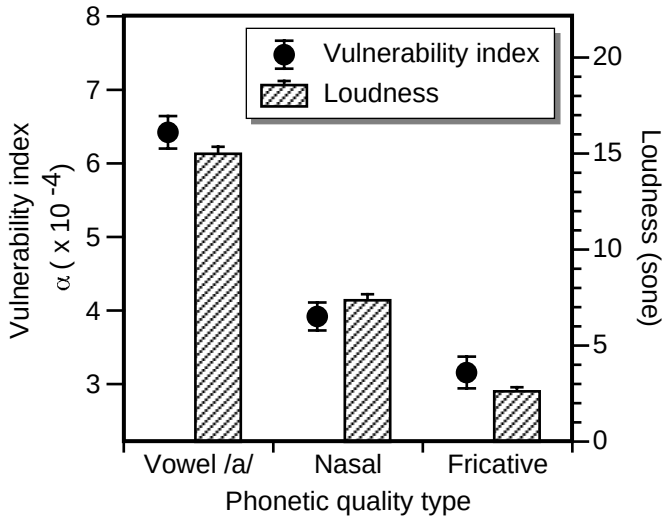

Figure 7.2: The temporal vulnerability (the second-order coefficient of a parabolic fitting to acceptability rating scores with change in the segmental duration; dots, left-hand scale) and the loudness (bars, right-hand scale) of a speech segment as a function of the phonetic quality type. The error bars show the standard errors. A larger vulnerability index implies a lower perceptual acceptability for a given change in the segmental duration. (Reproduced from Kato *et al.* (1998b).)

To take into account these perceptual characteristics, i.e., the dependency of durational sensitivity on the segment quality, for the evaluation model, we adopted the loudness as a weighting factor for each segmental error.

7.2.2 Perceptual interactions among multiple errors

A typical example of the interaction among multiple segmental errors may be the perceptual compensation effect in two consecutive segmental durations. This effect is like that when two segmental durations are modified in a compensatory manner, i.e., to lengthen one segment and to shorten the other by the same size; the total perceived distortion does not become very large in comparison with that expected from the sum of two independent modifications (see Fig. 7.4(b) for an example of compensatory modification).

The perceptual compensation effect between consecutive vowel and consonant durations has been reported for both detectability of the modification (Huggins, 1972b; Carlson and Granström, 1975) and acceptability rating (Sato, 1977; Sagisaka and Tohkura, 1984; Hoshino and Fujisaki, 1983). The compensation effect of this sort indicates that the influence

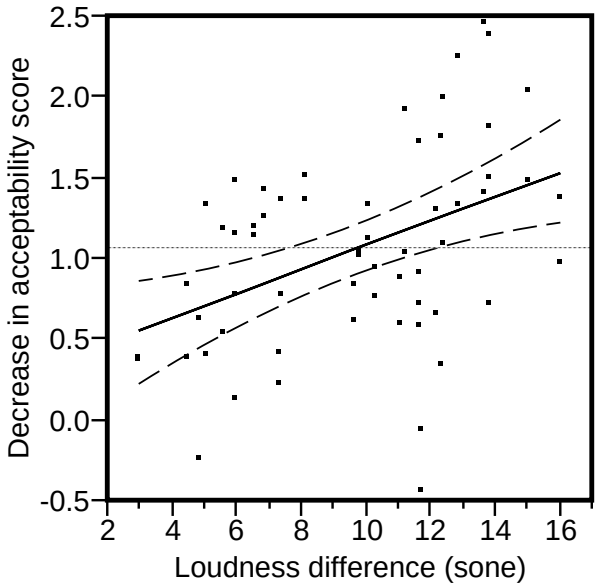

Figure 7.3: Decrease in the acceptability score yielded by a compensatory durational modification (30 ms lengthening and shortening) as a function of the loudness difference between two modified segments. The thick solid and dashed lines show the regression line and its 95 % confidence intervals. The horizontal dotted line marks the average of all samples. (Reproduced from Kato *et al.* (1997).)

of a durational error is not trapped within a segment but may interact beyond segmental boundaries, and also suggests that an evaluation criterion regarding each segmental error as independent is not perceptually valid.

Furthermore, as can be seen in Chapter 4, Kato *et al.* (1997) found that the amount of the perceptual compensation effect between two consecutive segments inversely correlates with the loudness difference or jump at the segmental boundary, in both detectability and acceptability tasks. The amount of compensation decreased with increasing loudness jump as shown in Fig. 7.3.

A non-speech study also showed that the detectability of a compensatory temporal modification correlates with the loudness jump at the displaced boundary (Kato *et al.*, 1997). This suggests that the correlation observed between the perceptual compensation effect of speech and the loudness jump can be accounted for as a reflection of the general

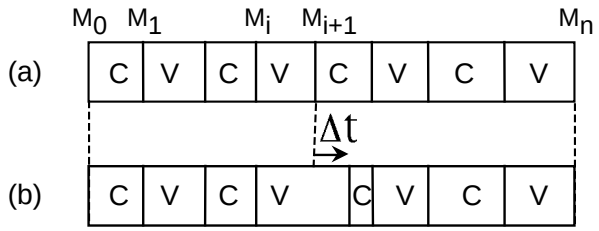

Figure 7.4: Schematic examples showing compensatory durational modification given to two consecutive segments in a four-mora word. C and V stand for consonant and vowel segments. The compensatory modification solely displaces the marker M_{i+1} to the right by Δt.

characteristics of the auditory perception.

Conventionally, while segmental errors have been regarded as "the changes of a segmental duration" (Fig. 7.5-(a)), all of the above notions suggest that they can also be regarded as "the displacements of segmental boundaries" (Fig. 7.5-(b)). Especially for describing the relationship among multiple errors, the former view may not be sufficient but the latter view appears to be useful. In the subsequent section, we will propose a framework to describe the temporal structure of speech, i.e., **time–loudness marker model**, by consulting this novel view in dealing with temporal errors.

7.3 Building an evaluation model based on perceptual characteristics

7.3.1 Framing the temporal marker model

The proposed model describes the temporal structure of a given speech token as a sequence of the perceptual cues embedded in that token, i.e., the temporal markers. Using this description, any modification in the temporal structure can be uniformly expressed by a single variable, that is, the mutual relationship between the temporal markers (Fig. 7.5-(b)). However, a problem arising for such marker-based expressions is that one cannot explicitly specify the location of each marker. Temporal markers, in a most strict sense, can exist at every acoustic change in the speech sounds. Nevertheless, human perception, in nature, tends to ignore gradual or small changes, but will pick up rapid and large changes in an auditory

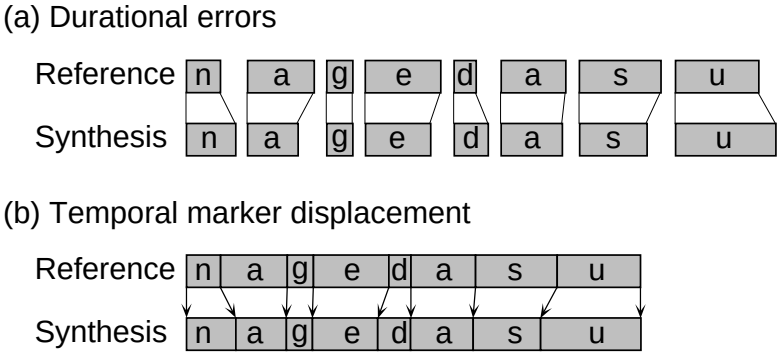

Figure 7.5: Examples schematically showing two expressions of differences in the temporal structure of a speech token (the word *nagedasu* in the examples). (a) In the conventional error evaluation procedure, the durational difference in each segment is measured independently, and then, all of them are summed or averaged throughout the token. (b) In the current model, the differences are totally expressed by the relative displacements among all of the temporal markers in the token, i.e., the segment boundaries in these diagrams.

stimulus (Bregman, 1990). Therefore, the present modeling, as a first-order approximation, assumes the markers to locate at segmental boundaries (which usually have large acoustic changes) and does not consider any possible marker at the central portions of a segment (which is relatively steady-state).

The remaining part of this section assumes a scaled loss of acceptability of a given temporal distortion as the subjective evaluation value. Any formulation for the modeling assumes it as the dependent variable.

The temporal displacement of a single marker in general makes multiple inter-marker intervals change. Assuming an independent element loss of acceptability for the change in each marker interval, the overall loss of acceptability L for given temporal changes in a word can be defined by the summation of all elements throughout that word as

$$L = \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} l_{ij}, \tag{7.1}$$

where l_{ij} denotes the element loss of acceptability corresponding to the temporal change in the interval between the ith and jth markers.

Notice that the summation range of Eq. (7.1) needs to be limited because the perceptual measurement between remote markers becomes difficult. Therefore, the implementation procedure in the next section will omit the terms l_{ij} where $|i - j| > 3$, for which no empirical data has currently been obtained.

7.3.2 Modeling the perceived error for each marker interval

This subsection formulates the element loss of acceptability l_{ij} in Eq. (7.1). As previously shown in Fig. 7.1, the loss of acceptability of a change in a single segmental duration increases not linearly but acceleratedly with the linear increasing of the modification size. Therefore, the whole-word acceptability L can be approximated as a second-order polynomial function of Δt, the size of the modification. Since l_{ij} has a linear relationship with L as derived from Eq. (7.1), l_{ij} can also be approximated as a second-order polynomial function of Δt. The vertex of this parabolic function should be placed at the origin of the coordinate axes because the loss of acceptability can be assumed as minimum when no temporal distortion is given, that is, Δt is zero.

In addition, the size of the target inter-marker interval itself, i.e., the original duration denoted by t_{ij}, may affect the loss of acceptability l_{ij}. Although the original duration does not affect the decrement speed of acceptability for regular-length (within the regular mora) segments (Kato *et al.*, 1998a), it can affect the decrement tendency for an extra-long speech portion as found in special moras (Kato *et al.*, 1998b).

Therefore, in the current modeling, we adopted the number of regular-mora segments between the markers in question to measure any inter-marker interval as a first order approximation. For the time range dealt with in the current study, the absolute discrimination threshold of a given auditory interval has been reported as not proportional to the base duration, but approximately to the square-root of the base duration (Abel, 1972a; Abel, 1972b). The effect of the original duration is, therefore, taken into account by normalizing l_{ij} with the reciprocal of $\sqrt{t_{ij}}$.

All of the above notions formulate the element loss of acceptability $l_{ij}(\Delta t)$ for a modification Δt as

$$l_{ij}(\Delta t) \cong \frac{a \cdot w_{ij} \cdot \Delta t^2}{\sqrt{t_{ij}}}, \tag{7.2}$$

where w_{ij} denotes the weighting factor due to the variation of the markers involved and a is a constant to adjust the difference of scales between both sides of the equation.

7.3.3 Weighting function of each marker and marker interval

This subsection formulates the weighting function w_{ij} in Eq. (7.2) that reflects the two kinds of perceptual factors introduced in section 7.2. Firstly, subsection 7.2.1 demonstrated that the vulnerability index, an index of the loss of acceptability, correlates with the loudness

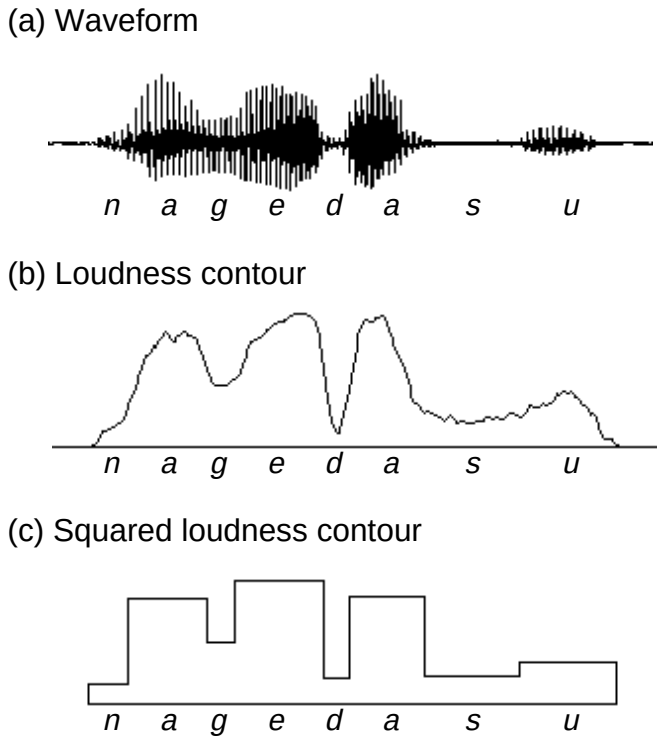

Figure 7.6: An example showing the process to extract the time-loudness marker expression from a given speech waveform. (a) The given waveform. (b) The loudness contour calculated every 2.5 ms with a 30 ms window in accordance with the ISO-532(B) method. (c) Simplified result by taking a representative loudness (i.e., the median loudness) for each segment.

of the segment in question as shown in Fig. 7.2. This correlation can be denoted as a linear weighting to l_{ij} by the representative loudness between markers M_i and M_j, e.g., median loudness, C_{ij}. Second, subsection 7.2.2 demonstrated that the loss of acceptability correlates with the loudness difference or jump at the displaced segmental boundary, namely, the temporal marker as shown in Fig. 7.3. This correlation can also be denoted as a linear weighting to l_{ij} by the loudness jumps at the markers M_i and M_j; they are, I_i and I_j, respectively. These notions can formulate the weighting function due to the marker variations w_{ij} as

$$w_{ij} = b(I_i + I_j) + C_{ij}, \tag{7.3}$$

where b is a constant to adjust the difference of scales between I and C. Notice that I and C are assumed as independent.

Table 7.1: Speech tokens used in the performance test. All of them are real Japanese words. The underlined CVC sequences are the targets of modification. The left column shows the temporal positions of targets in a word, where C_i or V_i stands for the ith consonant or ith vowel in the word.

Target position	Roman transcription				
$C_1V_1C_2$	bakugeki	gakureki	hanareru	nagedasu	sakasama
$C_2V_2C_3$	hanahada	imasara	kasanaru	katameru	mikakeru
$C_3V_3C_4$	hanahada	korogasu	rokugatsu	tachimachi	tamatama

To summarize, the above formulations simply require a reduced description of the loudness contour of a speech signal as shown in Fig. 7.6(c). To obtain this description, first the loudness contour (Fig. 7.6(b)) is calculated from the time waveform of a speech signal (Fig. 7.6(a)), and then, the representative loudness is sampled and held by every segment. In what follows, we refer to this simplified description of a speech sound as **the time–loudness marker description**[2] and any model constructed based on this description as **the time–loudness marker model**.

7.4 Effectiveness test of the time–loudness marker model

In accordance with the frame proposed in section 7.3, the current section achieves an evaluation model of duration setting errors and tests the effectiveness of the model using real data measured in perceptual experiments.

7.4.1 Procedure

The test data are indices of the loss of acceptability adopted from previous experiments (Kato *et al.*, 1997) that have assured an interval scale. Table 7.1, Fig. 7.7, and Table 7.2 show the speech materials, their manipulations, and the procedures of the listening experiments. In brief, the loss of acceptability was obtained for each of 28 temporally modified versions of 15 word materials from six listeners.

The evaluation model was then achieved according to Eqs. (7.1, 7.2, 7.3). First, the

[2]The time–loudness marker descriptions of all of the speech tokens used in the experiments of this thesis are shown in Appendix B.

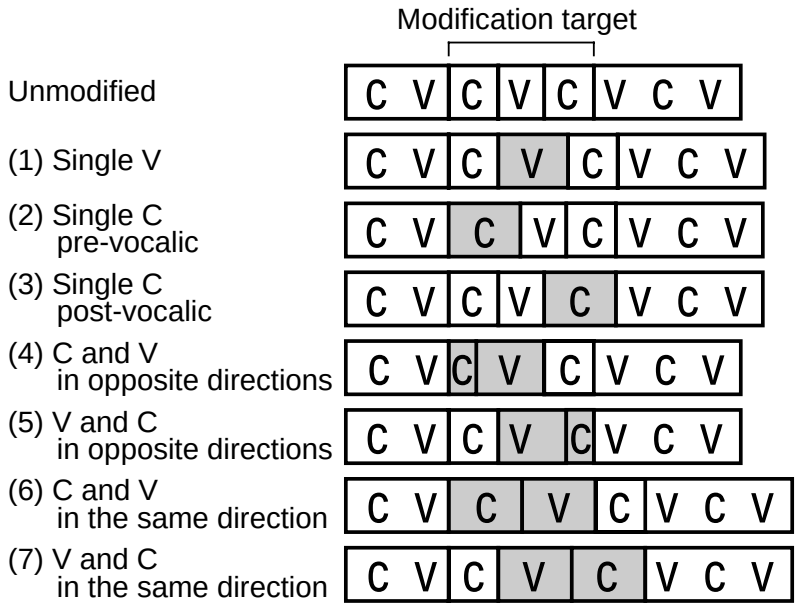

Figure 7.7: Schematic diagrams showing the seven types of temporal modifications made on each of the word samples. The hatched parts represent the segments whose durations were modified.

loudness contour was calculated for every material in accordance with ISO-532 B method (ISO, 1975) using Zwicker *et al.*'s (1991) algorithm. Then, each time-loudness marker description was obtained by sampling representative loudness values that were the median values in each segment. The constant values, i.e., a of Eq. (7.2) and b of Eq. (7.3), were appropriately chosen such that the root mean square prediction errors would be minimized for the given data.

The predictability of the proposed model was tested through several kinds of simulations. The prediction simulations were separated into two parts: a closed condition group and an open condition group. In the closed condition, the model was optimized by choosing the constant values as optimum using all of the prepared data. The open condition comprised two parts, i.e., open for listeners and open for word materials. In the open condition, the model was optimized using a half of the prepared data separated regarding listeners or words, and then used to predict the other half of the data. Similar procedures were repeated five times in each part for different combinations of listeners or words.

Table 7.2: Experimental conditions used in the performance test of the proposed model.

TEST DATA	
Material	
Tokens [Re: Table 7.2]	15 four-mora Japanese words
Speaker	1 male announcer
Temporal manipulation	
Modification types [Re: Fig. 7.7]	7 types (single V, single pre-vocalic C, single post-vocalic C, C-V in the same direction, V-C in the same direction, C-V in opposite directions, V-C in opposite directions)
Modification directions (sign of Δt)	to lengthen or to shorten
Modification sizes ($\|\Delta t\|$)	30 or 15 ms per segment
Total N of modification variations	28 (7 types$\times$2 directions$\times$2 sizes)
Total N of stimuli	420 (15 tokens$\times$28 modifications)
Evaluation by human	
Subjects	6 native speakers of Japanese
Scoring and scaling (loss index)	rating each stimulus on a seven-point scale, and then converting the scores to an interval scale using the law of categorical judgment.
Total N of loss index data	2520 (420 stimuli$\times$6 subjects)
Evaluation by model	
Loudness calculation	ISO-532, B method
Loudness contour	2.5 ms step, 30 ms Blackman window
Squared loudness	picking up the median of the loudness contour for each segment
Data usage in model optimization	open for subject, open for token, or closed

On the other hand, similar prediction simulations were performed as a reference test using a model that simply uses the average acoustic errors. This reference model, which is referred to as **the simple average model**, represents the conventional evaluation measure, namely, the average acoustic error and is given by

$$L = \frac{1}{n} \sum_{i=0}^{n-1} l_{i,i+1}. \tag{7.4}$$

Let

$$l_{i,i+1}(\Delta t) \cong w \cdot \Delta t, \tag{7.5}$$

where w is a constant to adjust the scale difference between both sides of Eq. (7.4) and is chosen so as to minimize the root mean square errors.

7.4.2 Results and discussion

The root mean square prediction errors of the proposed model are shown in Fig. 7.8 under each testing condition. Those of the reference model are also shown (only the closed condition). As clearly seen in the figure, the prediction errors of the proposed model are smaller than those of the reference model in any of the experimental conditions.

As generally expected, better predictions were found in the open conditions than the closed condition. Within the open conditions, the prediction error in the open-for-listeners condition was larger than that in the open-for-words condition. This is because the data deviation by the listeners was larger than that by the words.

To examine the effectiveness of the proposed model more specifically, the one-to-one correspondences between the observed and predicted loss indices are shown in Fig. 7.9 for each of both the simple average model and the proposed psychoacoustical model. The horizontal distance from the diagonal indicates the amount of prediction error; the left and right directed ones correspond to underestimation and overestimation, respectively.

It can be seen that the proposed model predicted the largest group of the observed loss indices (marked with crosses) more accurately than the simple average model did. The simple average model significantly underestimated these “dangerous” loss values. This difference in predictability between the two models, in fact, mostly came from the advantage of the proposed psychoacoustical model to properly deal with the relationship among multiple errors. As such, the proposed model is advantageous in picking up errors that are acoustically not so large but perceptually serious.

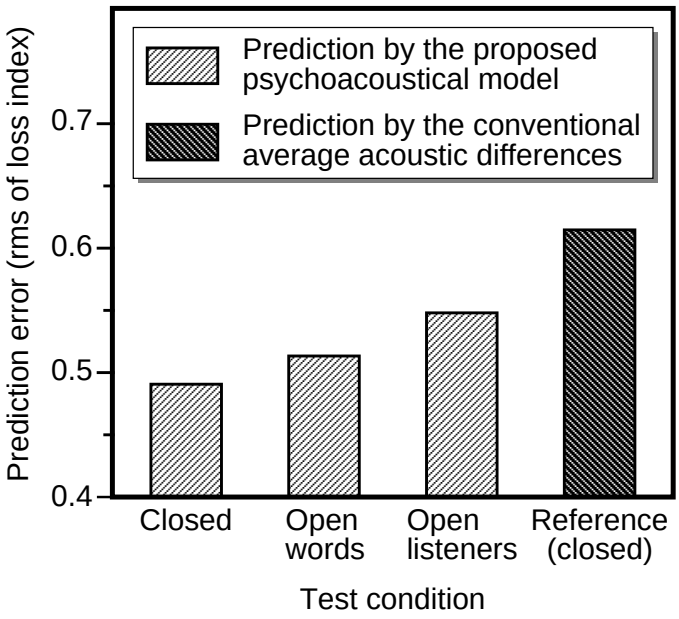

Figure 7.8: Mean absolute prediction errors to observed loss indices. The reference condition shows the errors produced by prediction using the conventional simple average model.

7.5 Conclusions

To assess a given durational error in segmental duration controls, we proposed a frame that can be utilized to predict the perceptual amount of degradation caused by **the time–loudness marker model**.

First, we examined the perceptual invalidity of the conventional evaluation measure, the simple average of acoustic errors; the problems were summarized into the following two points: (1) to give every segment the same importance, and (2) to treat each segment independently. Second, we proposed one description of a speech signal, i.e., **the time-loudness marker model**, to take into account the perceptual factors in a measure of evaluation; this description was obtained by simplifying the loudness contour of speech. Finally, we achieved an example of the evaluation model using the proposed description and tested its effectiveness by the prediction simulations of real perceptual data. The results showed that the proposed model consistently achieved a better prediction (i.e., closer to human evaluation) than the reference model which only used the average acoustic error without any perceptual consideration.

The proposed framework can be expected to contribute to an improvement in the naturalness of synthesized speech because it provides more perceptually valid (closer to human) evaluation criteria. An important direction of future work is the generalization of the current evaluation model by evaluation tests with newly collected data that includes more realistic errors, e.g., ones occurring randomly at more than two segments. In addition, there remain a couple of factors to address, which affect the perceptual sensitivity to temporal structures of speech, but which are not included in the current modeling, e.g., functional differences between vowel onsets and offsets (Kato *et al.*, 1998c, also in Chapter 5) and the temporal position of a marker in the word (Kato *et al.*, 1998a, also in Chapter 2). An implementation of these factors awaits additional experiments designed to quantitatively confirm the effects of these factors.

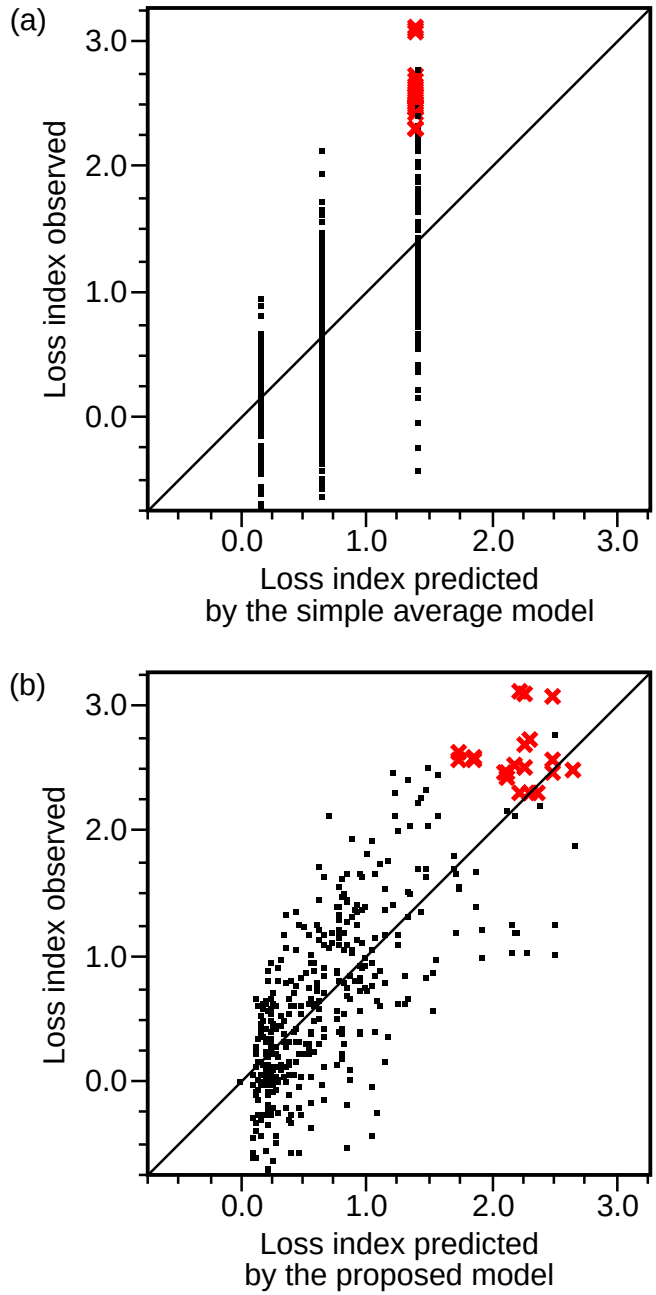

Figure 7.9: Observed versus predicted loss indices using two models: (a) A simple average model, and (b) the proposed psychoacoustical model. The diagonal lines show predictions free of errors. A point over the diagonal means underestimation, i.e., the model predicted the loss index as too small, and a point under the diagonal means overestimation. The psychoacoustical model (b) predicts the largest, i.e., most dangerous, loss values (marked with crosses) more accurately than the average model (a) does. The simple average model significantly underestimates these “dangerous” loss values.

Chapter 8

Conclusions

Abstract

We have seen how perceptual sensitivity to temporal distortions is affected by differences in stimulus attributes and contexts, and we concluded that psychoacoustical bases of auditory perception play an important role in the subjective evaluation of temporal distortions in speech segments. In our study, perceptual sensitivity varied with different phonetical or phonological properties, but most of the influence allowed a psychoacoustical accountability even when the listeners used a speech-specific evaluation such as the acceptability rating. Nevertheless, there remain quite a few phenomena that are not accountable in psychoacoustical terms. Further investigations are therefore necessary to determine whether they require speech-specific implications.

In this final chapter, we summarize each experimental result with a brief overview of the dissertation, and try to direct future investigations towards addressing the remaining problems.

8.1 Summary of the dissertation

The problem addressed in this dissertation is that of the gap between human perception and the machine production of speech from temporal aspects. The aim of the current study was, then, to guide machines to produce more naturally sounding speech by bridging the gap between physical and perceptual spaces.

We first performed speech perception experiments to explore how phonetic factors influence the human perceptual sensitivity or acceptability of temporal distortions in speech. Chapter 2 revealed three factors affecting the acceptability of distortions in a single vowel duration, i.e., the vowel quality, the temporal position in a word, and the voicing of the following consonant. Chapter 3 investigated the influence of phonetic quality and the original duration of a modified speech portion which may include "a special mora," and demonstrated both factors as perceptually significant.

Chapter 4 investigated the perception of temporal modification in two consecutive segments. The results, first, demonstrated that two consecutive segments can compensate their durations with each other and, then, showed that the amount of the compensation effect can be accounted for by the loudness difference between the two modified segments. The most important implication of this chapter's results is that changes in the segmental duration can also be regarded as displacements of the segmental boundaries or temporal markers. This marker-oriented viewpoint for temporal error description could successfully be applied to the error evaluation modeling in Chapter 7. Chapter 5, then, revealed the functional differences between two kinds of temporal markers in speech, namely, vowel onsets and vowel offsets or consonant onsets.

We also performed psychoacoustical non-speech experiments in order to examine the extent to which general auditory limitations influence the speech perceptual factors found. This consideration is important to assure psychological validity in designing a perceptual model in the following chapter. The first two experiments (Section 6.1) related the acceptability measure, which was generally used in the preceding speech perception experiments, to the detectability measure, which was to be used in the following non-speech experiments. They, in turn, provided psychoacoustical evidence for each of the effects observed in the speech cases, i.e., the effects of temporal position in a word, phonetic quality, loudness jump, and direction (on or off) of the temporal markers.

We finally applied the obtained speech perceptual and psychoacoustical knowledge to an evaluation model for temporal rules of speech synthesis. To take into account the

perceptual factors, we first proposed a description of a speech signal, i.e., the time-loudness marker model. We then achieved an objective evaluation model in accordance with the proposed time-loudness marker description. The final evaluation test demonstrated that the proposed model consistently achieved a better prediction (i.e., closer to human evaluation) than the reference model which only used the average acoustic error without any perceptual consideration.

8.2 Future directions

One of the most important future research activities is to reexamine the current perceptual results from a linguistic or cross-linguistic point of view. Although much of the results could be accounted for from the psychoacoustical point of view, as seen in Chapter 6, and some phonological factors have already been examined, we still feel it necessary to include such a consideration for two reasons. First, much debate surrounds the question of whether phenomena in speech perception are accountable in psychoacoustic terms (e.g., Schouten, 1980) or whether different, speech-specific explanations are called for (e.g., Liberman and Mattingly, 1989). Second, it is of practical importance to estimate the extent to which the conclusions and predictions drawn from the current speech study can be generalized and also to estimate the extent to which linguistic or speech-specific factors may affect them. Language-specific phonotactic constraints, for example, may vary perceptual accessibility to different kinds of segment chunks, and, therefore, possibly affect the perceptual sensitivity to changes in the temporal structure of speech.

Another important direction of future research is to get into a more sophisticated internal model of general time perception. Early models of time perception have only been able to cope with relatively simplified stimulus patterns, such as a single interval marked by two tones or noise bursts and a single filled duration (Allan and Kristofferson, 1974; Creelman, 1962; Divenyi and Sachs, 1978; Getty, 1975; Kristofferson, 1977). These models have only been applicable to limited aspects of speech cases. Nevertheless, in recent years, theoretical or conceptual frameworks have been proposed to account for the perception of more complicated, closer to daily-life, stimulus patterns (McAuley and Kidd, 1998; Nakajima, 1987; Nakajima and Sasaki, 1996; Tsuzaki and Kato, 1998). There, however, still remains a considerable gap between their stimulus patterns and natural speech. Both the speech and non-speech experimental data obtained from the current study may play a role of filling up this gap and promote the achievement of a more advanced and realistic model

of time perception.

Appendix A

Properties and Vulnerability Index of Each Tested Speech Portion

Table A.1 shows the word texts used in the experiment of Chapter 2 in alphabetical order, and the position in a word, vowel quality, voicing of the following consonant, duration, and mean of vulnerability indices for each of the tested vowel segments.

Table A.2 shows the words used in experiment 1 of Chapter 3 in alphabetical order, and the phonetic quality, phonetic quality type, acoustic duration, and mean of vulnerability indices for each test portion.

Table A.3 shows the words used in experiment 2 of Chapter 3 in alphabetical order, and the phonetic quality, phonetic quality type, categorized duration (short or long), acoustic duration, and mean of vulnerability indices for each test portion.

Table A.1: The words used in the experiment in Chapter 2 in alphabetical order, and the position in a word, vowel quality, voicing of the following consonant, duration, and mean of vulnerability indices (αs) for each of the tested vowel segments. The transcription of the Japanese text is based on the Hepburn system. Those vowels whose durations were subjected to modification are marked in bold face.

Text	Position	Vowel	Following consonant	Duration (ms)	α ($\times 10^{-4}$)
akumade	1	a	uv	102.5	11.56
akum**a**de	3	a	v	100	3.22
atsumaru	1	a	uv	102.5	7.85
b**a**kugeki	1	a	uv	105	12.93
b**a**rabara	1	a	v	95	4.96
barab**a**ra	3	a	v	80	3.42
ch**i**jimeru	1	i	v	70	5.61
fumik**i**ri	3	i	v	75	5.67
g**a**kureki	1	a	uv	100	12.83
hanah**a**da	3	a	v	70	4.87
h**a**nareru	1	a	v	55	6.01
h**a**rahara	1	a	v	75	7.78
harah**a**ra	3	a	v	75	3.40
h**a**rigane	1	a	v	65	5.20
hatar**a**ki	3	a	uv	130	6.26
hatsur**a**tsu	3	a	uv	115	8.65
h**i**rogaru	1	i	v	65	6.64
h**i**romeru	1	i	v	55	4.97
horob**i**ru	3	i	v	102.5	8.87
imas**a**ra	3	a	v	90	6.72
iriguchi	1	i	v	95	4.88
kakur**i**tsu	3	i	uv	95	6.99
kakuj**i**tsu	3	i	uv	90	6.48
k**a**nashimu	1	a	v	55	9.76
kasan**a**ru	3	a	v	120	8.83
k**a**tameru	1	a	uv	77.5	10.55
ketob**a**su	3	a	uv	92.5	9.65

Table A.1: (continued)

Text	Position	Vowel	Following consonant	Duration (ms)	α ($\times 10^{-4}$)
k**i**nodoku	1	i	v	57.5	5.92
kogat**a**na	3	a	v	75	6.91
korog**a**su	3	a	uv	95	10.51
kotog**a**ra	3	a	v	135	5.19
kurush**i**mu	3	i	v	60	5.35
m**a**monaku	1	a	v	70	6.10
m**a**rumeru	1	a	v	120	8.89
matag**a**ru	3	a	v	125	5.87
m**i**kakeru	1	i	uv	80	7.64
m**i**nogasu	1	i	v	60	10.37
m**i**tomeru	1	i	uv	85	10.12
m**i**tsumeru	1	i	uv	90	6.45
muras**a**ki	3	a	uv	95	8.32
n**a**gedasu	1	a	v	110	10.25
n**a**nishiro	1	a	v	100	13.56
n**a**raberu	1	a	v	90	7.45
n**a**ruhodo	1	a	v	95	8.69
n**i**semono	1	i	uv	75	7.54
nokog**i**ri	3	i	v	145	4.90
osam**a**ru	3	a	v	140	3.60
rokug**a**tsu	3	a	uv	105	12.41
s**a**bireru	1	a	v	85	7.40
s**a**kasama	1	a	uv	80	14.23
sashid**a**su	3	a	uv	100	10.32
setsur**i**tsu	3	i	uv	105	6.80
sh**i**baraku	1	i	v	37.5	8.36
sh**i**mekiri	1	i	v	40	5.21
shimek**i**ri	3	i	v	70	3.00
sh**i**mekiru	1	i	v	35	6.73
shimij**i**mi	3	i	v	65	7.19
sh**i**nabiru	1	i	v	47.5	8.56

Table A.1: (continued)

Text	Position	Vowel	Following consonant	Duration (ms)	α ($\times 10^{-4}$)
shinab**i**ru	3	i	v	90	7.29
tachim**a**chi	3	a	uv	110	8.39
t**a**dachini	1	a	v	45	9.99
t**a**matama	1	a	v	60	8.95
tamat**a**ma	3	a	v	75	6.88
tanosh**i**mi	3	i	v	70	4.99
tanosh**i**mu	3	i	v	65	7.02
tomok**a**ku	3	a	uv	90	9.07
tonik**a**ku	3	a	uv	90	9.17
urag**i**ru	3	i	v	125	5.72
z**a**razara	1	a	v	117.5	7.67
zaraz**a**ra	3	a	v	105	2.42

Table A.2: The words used in experiment 1 of Chapter 3 in alphabetical order; those portions whose durations were subjected to modification are marked in bold face. The attributes of each test portion, i.e., phonetic quality, phonetic quality type, and acoustic duration, and mean vulnerability indices (αs) are also given. The transcription of the Japanese text is based on the Hepburn system (except that a moraic nasal is transcribed by an upper-case N). A devoiced vowel is marked with an under-ring. A top tie-bar marks a phoneme pair or triad that is inseparable in terms of phonetic segmentation. The phonetic symbols basically follow IPA usage.

Text	Phonetic quality	Phonetic quality type	Duration (ms)	α ($\times 10^{-4}$)
ba**N͡**gumi	[ŋ]	nasal	145.0	3.94
ba**ku̥͡**hatsu	[x]	fricative	125.0	1.86
bu**tsu̥͡**karu	[s]	fricative	107.5	3.47
cho**k͡k**aku	silence	silence	182.5	2.29
da**N**ketsu	[ŋ]	nasal	95.0	3.17
ga**N**jitsu	[n]	nasal	125.0	4.35
ga**k͡k**ari	silence	silence	220.0	1.90
ha**N**doru	[n]	nasal	145.0	3.64
han**a**reru	[ɑ]	vowel	127.5	6.04
i**N**sotsu	[ɴ]	nasal	85.0	3.71
im**a**sara	[ɑ]	vowel	105.0	7.06
ji**s͡s**eki	[s]	fricative	210.0	2.11
ka**N͡**geki	[ŋ]	nasal	157.5	2.43
ka**N**kaku	[ŋ]	nasal	102.5	3.45
ka**N**toku	[n]	nasal	90.0	4.34
kan**a**shimu	[ɑ]	vowel	120.0	5.18
kas**a**naru	[ɑ]	vowel	97.5	7.09
ka**shi̥͡**kiri	[ʃ]	fricative	117.5	2.38
kat**a**meru	[ɑ]	vowel	102.5	7.30
ke**s͡s**aku	[s]	fricative	195.0	2.20
ki**N͡**motsu	[m]	nasal	157.5	3.38
ko**k͡k**aku	silence	silence	175.0	2.22
ko**shi̥͡**kake	[ʃ]	fricative	115.0	3.62

Table A.2: (continued)

Text	Phonetic quality	Phonetic quality type	Duration (ms)	α ($\times 10^{-4}$)
ma**N**naka	[n]	nasal	155.0	4.75
ma**ss**ugu	[s]	fricative	270.0	0.74
mat**a**garu	[ɑ]	vowel	105.0	6.90
mik**a**keru	[ɑ]	vowel	125.0	5.19
mi**ss**etsu	[s]	fricative	190.0	2.44
mi**tsu**keru	[s]	fricative	87.5	2.92
mo**chi**komu	[ʃ]	fricative	95.0	3.67
nar**a**beru	[ɑ]	vowel	122.5	6.54
na**tt**oku	silence	silence	190.0	2.93
o**shi**komu	[ʃ]	fricative	135.0	3.01
ri**N**kaku	[ŋ]	nasal	85.0	5.06
sa**N**butsu	[m]	nasal	162.5	3.16
sa**kk**aku	silence	silence	172.5	2.17
sa**pp**ari	silence	silence	215.0	2.58
sa**shi**komu	[ʃ]	fricative	117.5	3.34
sa**ss**oku	[s]	fricative	220.0	1.64
sa**ss**uru	[s]	fricative	235.0	2.22
se**kk**aku	silence	silence	195.0	1.59
shi**N**jiru	[n]	nasal	102.5	4.66
shin**a**biru	[ɑ]	vowel	127.5	6.28
ta**chi**kiru	[ʃ]	fricative	102.5	3.69
ta**ss**uru	[s]	fricative	250.0	1.53
ta**su**keru	[s]	fricative	137.5	2.62
u**chi**komu	[ʃ]	fricative	102.5	4.14
ur**a**giru	[ɑ]	vowel	122.5	6.81
za**N**koku	[ŋ]	nasal	95.0	3.78

Table A.3: The words used in experiment 2 of Chapter 3 in alphabetical order; those portions whose durations were subjected to modification are marked in bold face. The attributes of each test portion, i.e., phonetic quality, phonetic quality type, duration category, and acoustic duration, and mean vulnerability indices (αs) are also given. The transcription of the Japanese text is based on the Hepburn system (except that a long vowel is marked with a subsequent length mark ":"). A devoiced vowel is marked with an under-ring. A top tie-bar marks a phoneme pair or triad that is inseparable in terms of phonetic segmentation. The phonetic symbols basically follow IPA usage.

Text	Phonetic quality	Phonetic quality type	Duration category	Duration (ms)	α ($\times 10^{-4}$)
ap**a**:to	[ɑ]	vowel	long	205.0	4.13
dep**a**:to	[ɑ]	vowel	long	200.0	4.35
fug**o**:ri	[o]	vowel	long	260.0	3.67
hir**o**garu	[o]	vowel	short	132.5	5.45
im**a**sara	[ɑ]	vowel	short	105.0	6.26
im**o**:to	[o]	vowel	long	227.5	3.75
ji**s͡s**eki	[s]	fricative	long	210.0	3.66
ka**s͡hi̥**kiri	[ʃ]	fricative	short	117.5	3.65
ke**s͡s**aku	[s]	fricative	long	195.0	3.86
kor**o**gasu	[o]	vowel	short	100.0	5.78
mat**a**garu	[ɑ]	vowel	short	105.0	6.53
min**o**gasu	[o]	vowel	short	120.0	6.60
mit**o**:shi	[o]	vowel	long	232.5	3.80
mit**o**reru	[o]	vowel	short	112.5	6.43
mi**ts͡u̥**keru	[s]	fricative	short	87.5	4.98
mon**o**:ki	[o]	vowel	long	255.0	2.45
moy**o**:shi	[o]	vowel	long	220.0	3.47
nar**a**beru	[ɑ]	vowel	short	122.5	6.26
o**s͡hi̥**komu	[ʃ]	fricative	short	135.0	3.56
ot**o**:to	[o]	vowel	long	227.5	3.93
rek**o**:do	[o]	vowel	long	210.0	4.25
sa**s͡hi̥**komu	[ʃ]	fricative	short	117.5	4.42
sa**s͡s**oku	[s]	fricative	long	220.0	2.63

Table A.3: (continued)

Text	Phonetic quality	Phonetic quality type	Duration category	Duration (ms)	α ($\times 10^{-4}$)
sa**s͡s**uru	[s]	fricative	long	235.0	3.22
shin**a**biru	[ɑ]	vowel	short	127.5	6.68
sum**a**ːto	[ɑ]	vowel	long	202.5	4.77
ta**s͡s**uru	[s]	fricative	long	250.0	2.87
ta**s͡u̥**keru	[s]	fricative	short	137.5	4.30
tod**o**keru	[o]	vowel	short	107.5	6.14
ur**a**giru	[ɑ]	vowel	short	122.5	5.55

Appendix B

Time-Loudness Profiles of Speech Materials

The time–loudness marker descriptions introduced in Chapter 7 were calculated for the speech materials used in the experiments and are shown in the following pages, in alphabetical order of the tokens. The corresponding loudness contours are superimposed. The horizontal and vertical axes represent the time in ms and the loudness in sone, respectively. In each panel, the thin line shows the loudness contour calculated every 2.5 ms using a 30 ms rectangular window in accordance with ISO 532(B). The thick line shows the time-loudness marker description by taking a representative loudness, i.e., the median loudness, from each segment. The text of the token is supplied within each panel at the top-left corner.

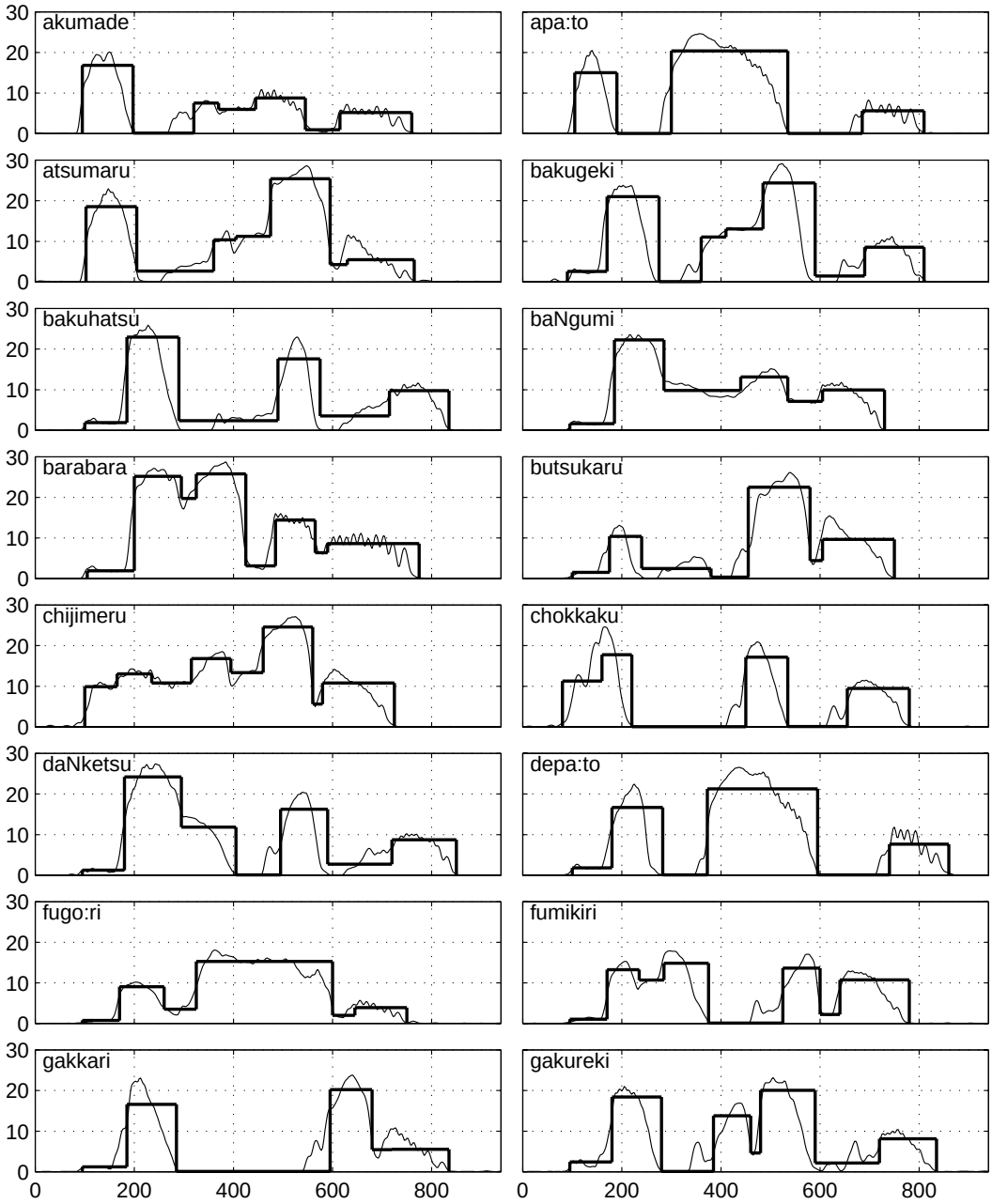

Figure B.1: The time–loudness marker descriptions (thick lines) of the speech materials used in the experiments superimposed with the corresponding loudness contours (thin lines). The horizontal and vertical axes represent the time in ms and the loudness in sone, respectively.

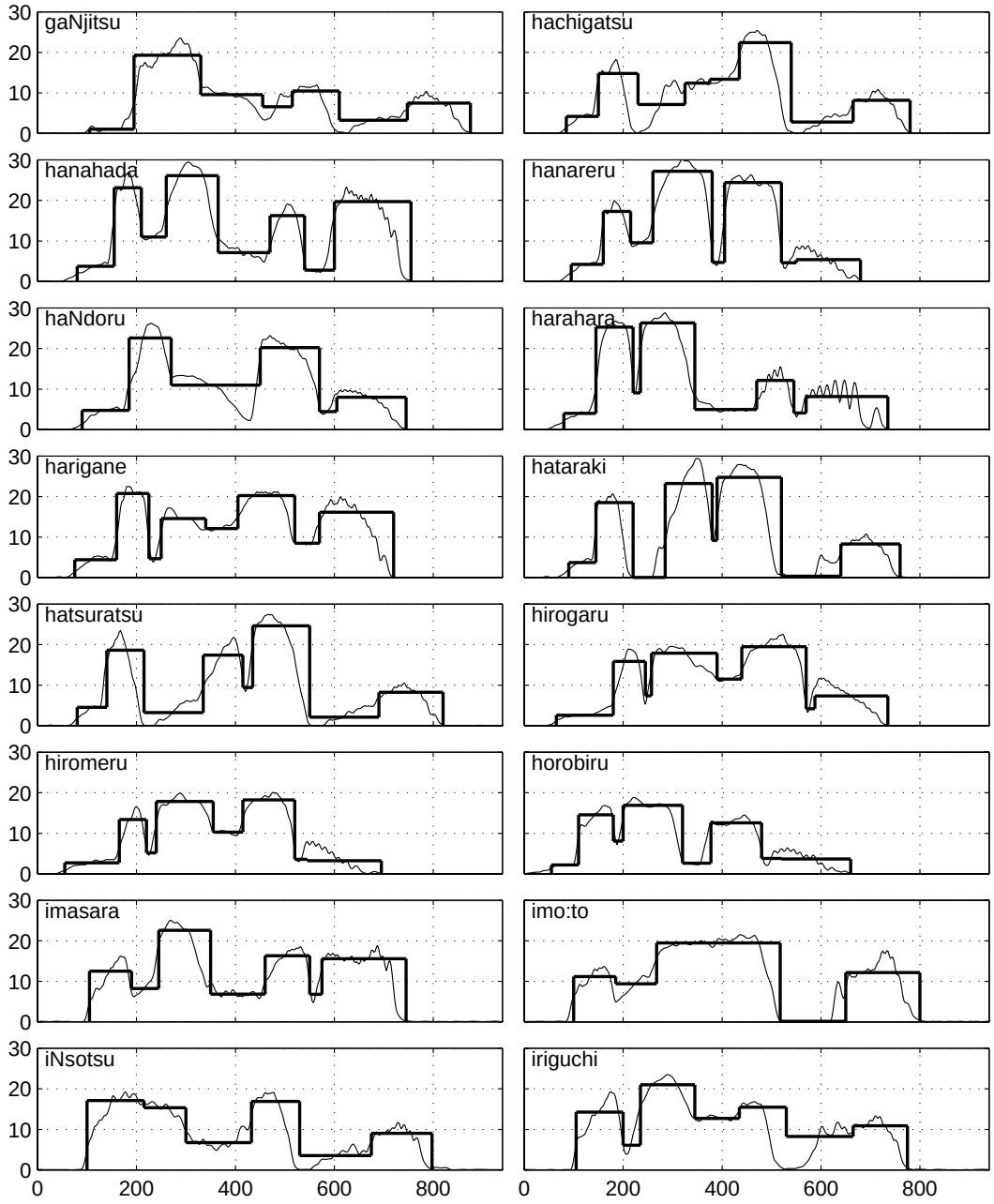

Figure B.2: The time–loudness marker descriptions (thick lines) of the speech materials used in the experiments superimposed with the corresponding loudness contours (thin lines). The horizontal and vertical axes represent the time in ms and the loudness in sone, respectively.

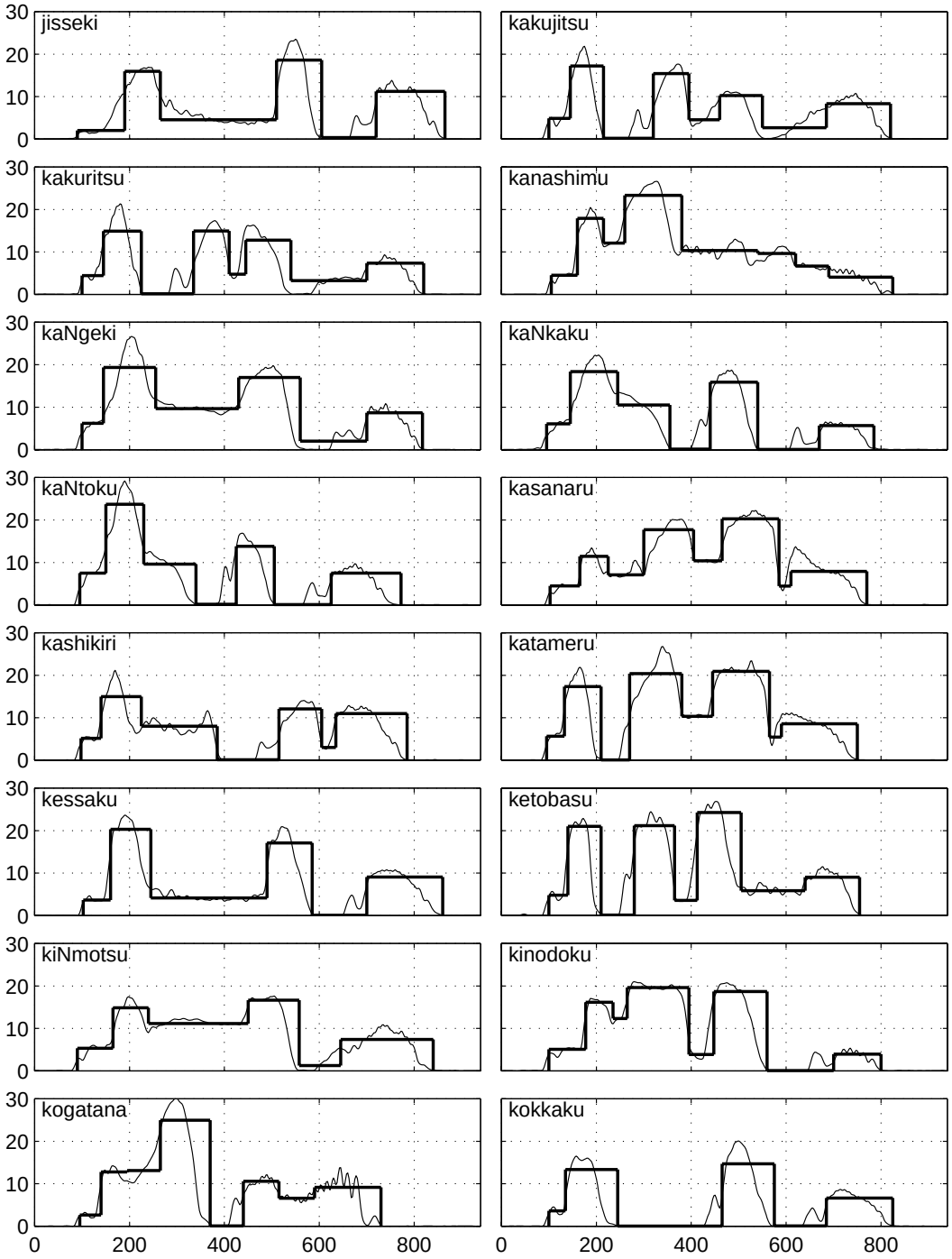

Figure B.3: The time–loudness marker descriptions (thick lines) of the speech materials used in the experiments superimposed with the corresponding loudness contours (thin lines). The horizontal and vertical axes represent the time in ms and the loudness in sone, respectively.

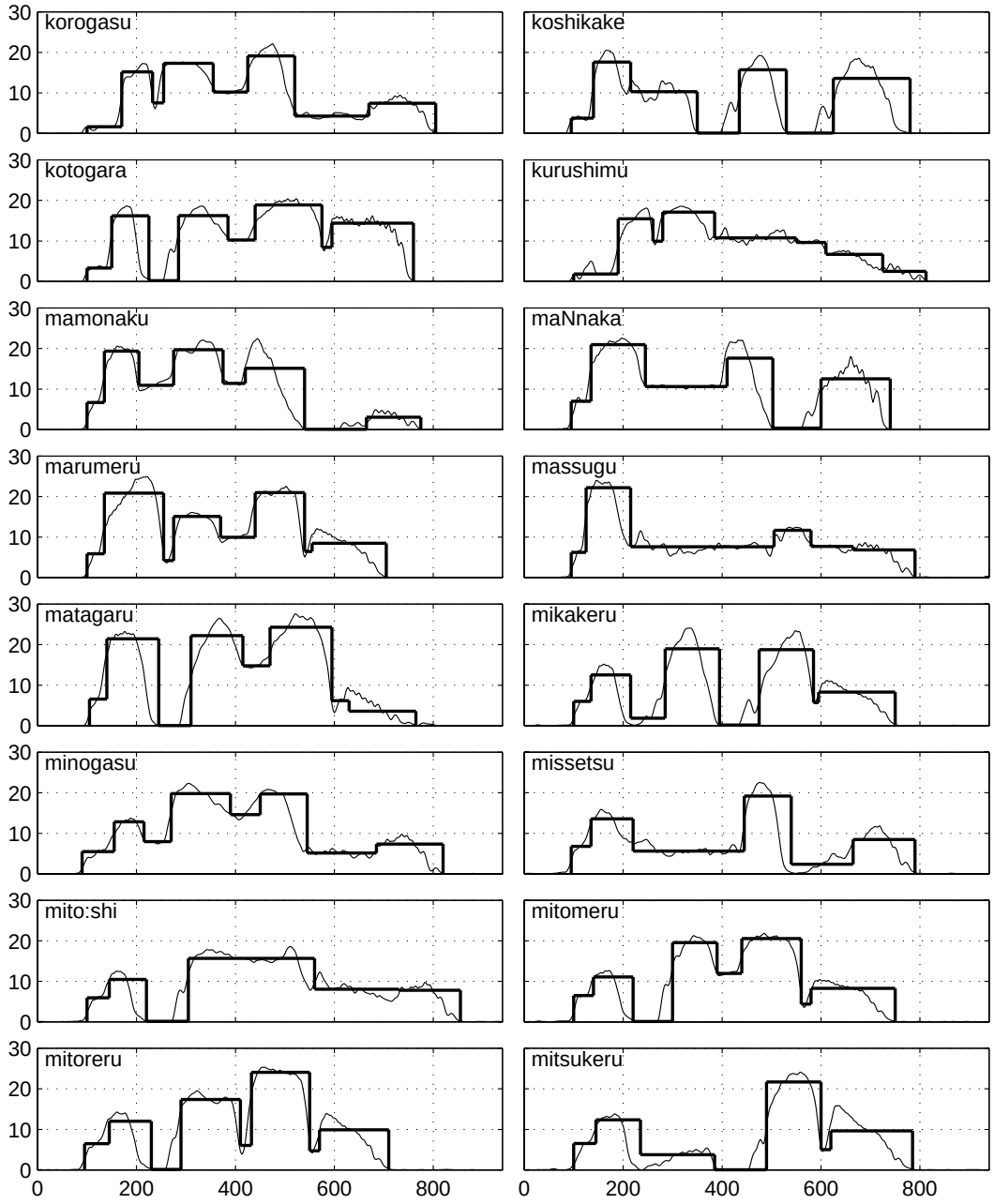

Figure B.4: The time–loudness marker descriptions (thick lines) of the speech materials used in the experiments superimposed with the corresponding loudness contours (thin lines). The horizontal and vertical axes represent the time in ms and the loudness in sone, respectively.

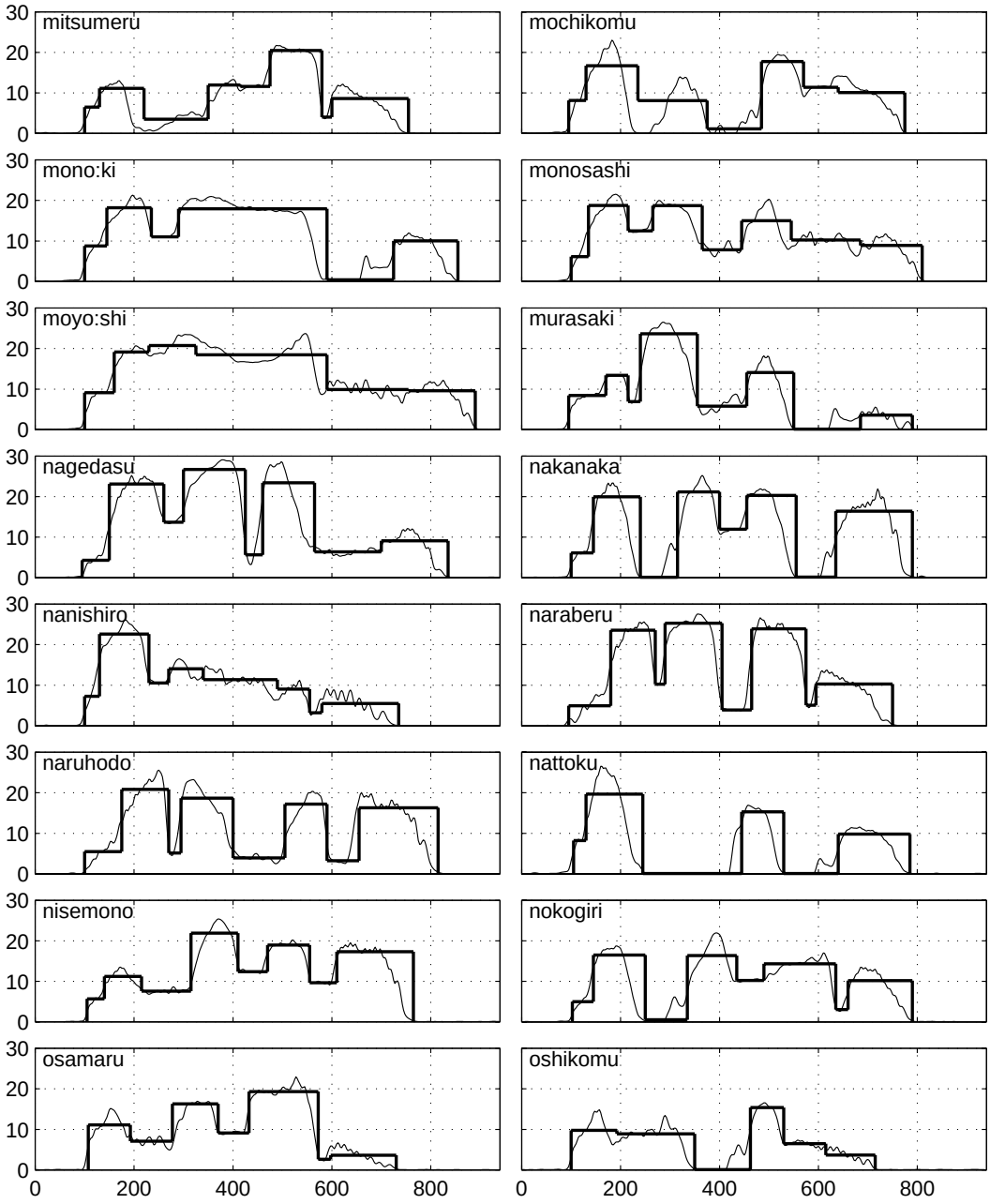

Figure B.5: The time–loudness marker descriptions (thick lines) of the speech materials used in the experiments superimposed with the corresponding loudness contours (thin lines). The horizontal and vertical axes represent the time in ms and the loudness in sone, respectively.

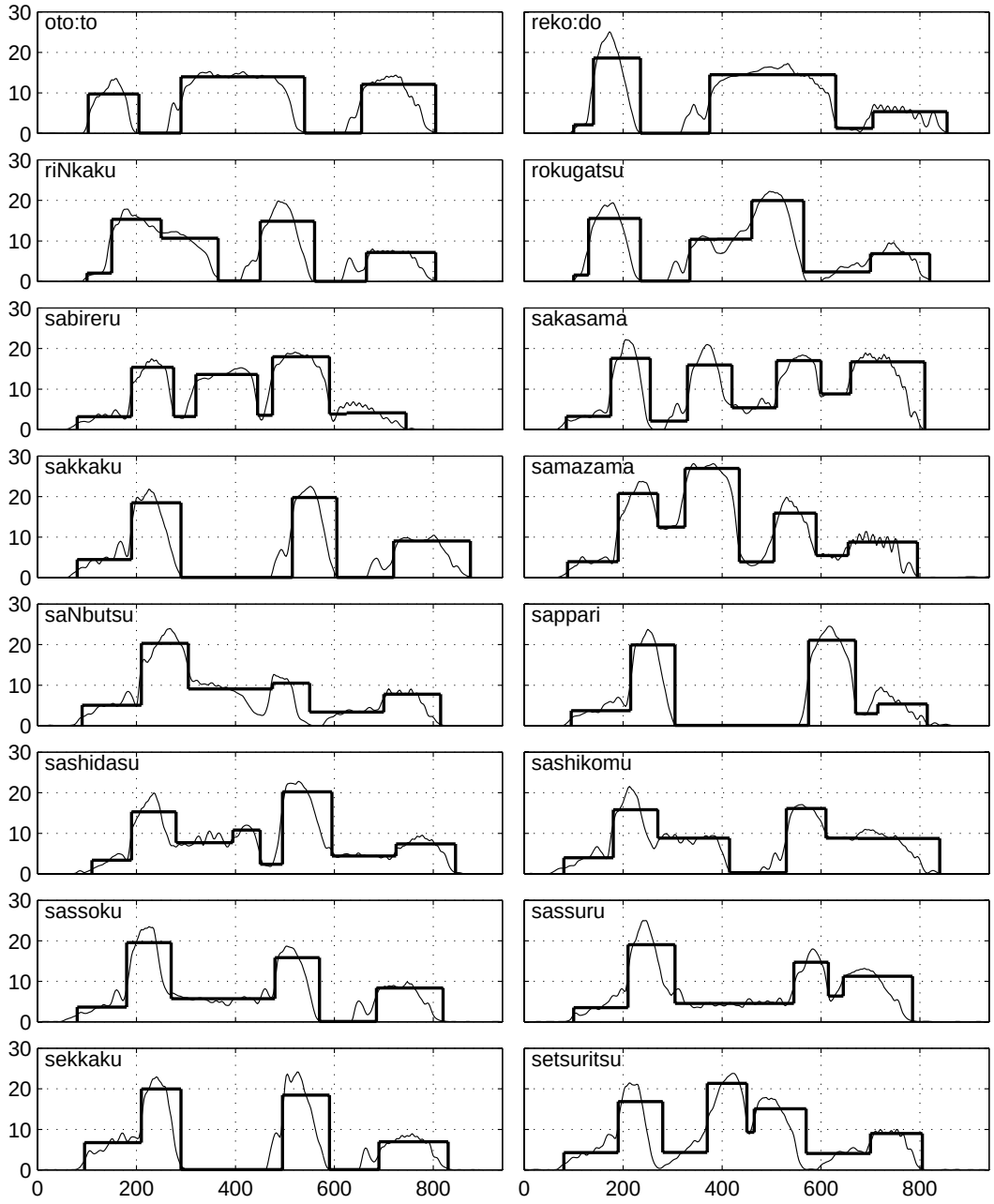

Figure B.6: The time–loudness marker descriptions (thick lines) of the speech materials used in the experiments superimposed with the corresponding loudness contours (thin lines). The horizontal and vertical axes represent the time in ms and the loudness in sone, respectively.

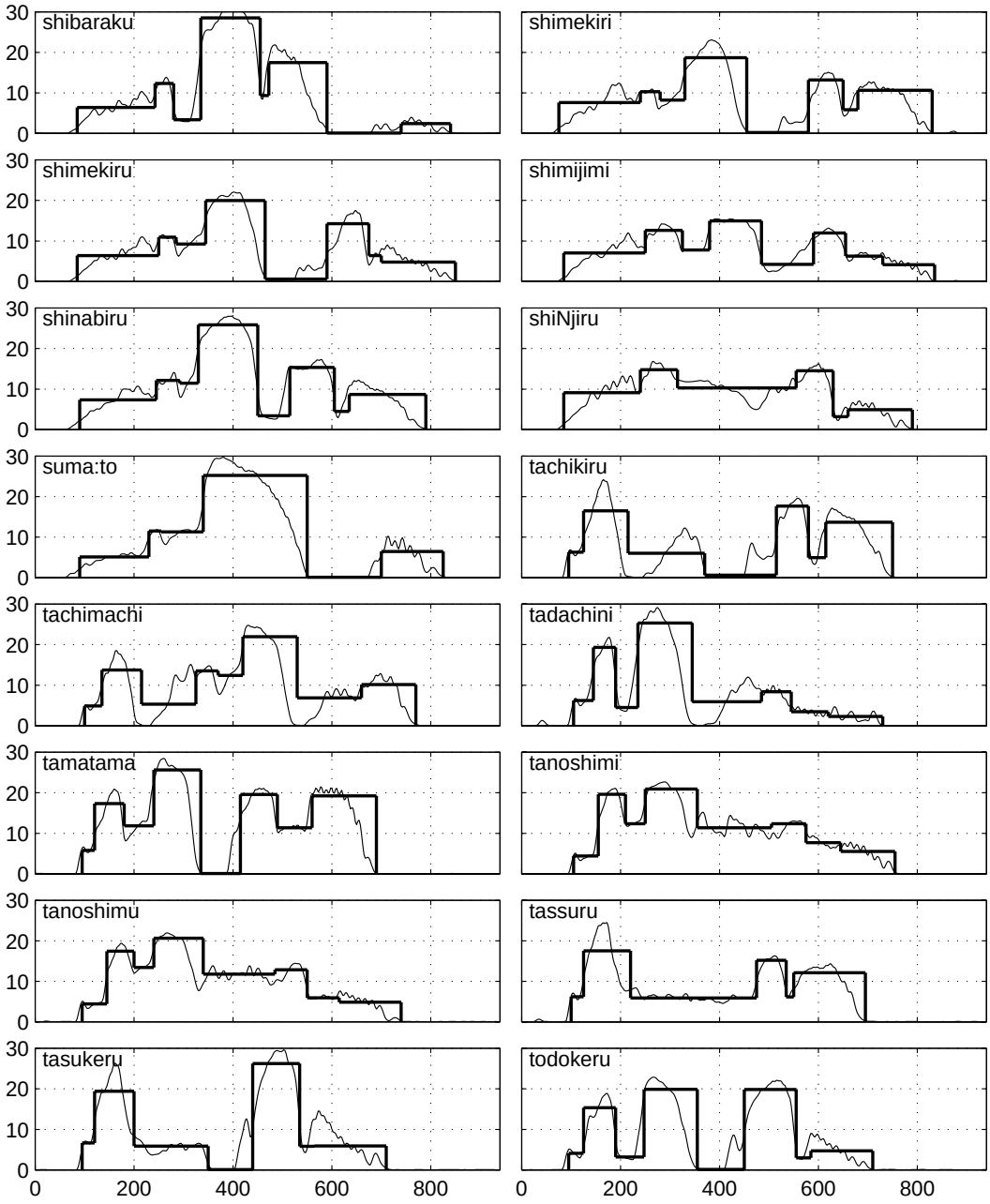

Figure B.7: The time–loudness marker descriptions (thick lines) of the speech materials used in the experiments superimposed with the corresponding loudness contours (thin lines). The horizontal and vertical axes represent the time in ms and the loudness in sone, respectively.

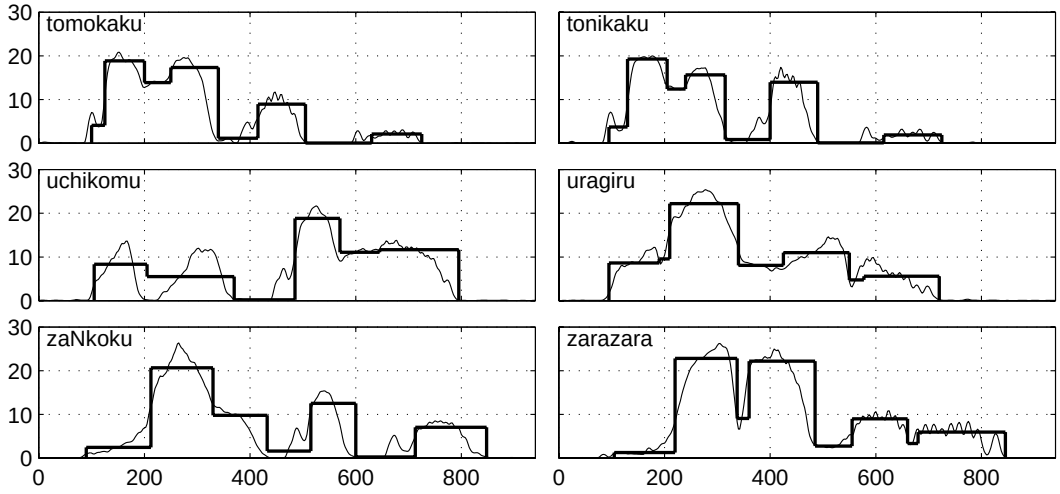

Figure B.8: The time–loudness marker descriptions (thick lines) of the speech materials used in the experiments superimposed with the corresponding loudness contours (thin lines). The horizontal and vertical axes represent the time in ms and the loudness in sone, respectively.

Bibliography

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

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

Allan, L. G. (**1979**). "The perception of time," Percept. Psychophys. **26**, 340–354.

Allan, L. G., and Kristofferson, A. B. (**1974**). "Psychophysical theories of duration discrimination," Percept. Psychophys. **16**, 26–34.

Allen, G. D. (**1972a**). "The location of rhythmic stress beats in English: An experimental study I," Lang. Speech **15**, 72–100.

Allen, G. D. (**1972b**). "The location of rhythmic stress beats in English: An experimental study I and II," Lang. Speech **15**, 72–100, 179–195.

Allen, G. D. (**1972c**). "The location of rhythmic stress beats in English: An experimental study II," Lang. Speech **15**, 179–195.

Allen, J., Hunnicutt, M. S., and Klatt, D. H. (**1987**). *From Text to Speech: The MITalk System* (Cambridge U. P., Cambridge, UK).

Bailly, G., Benoît, C., and Sawallis, T. R. (**1992**). *Talking Machines: Theories, Models, and Designs* (Elsevier, Amsterdam).

Bakkum, M. J., Plomp, R., and Pols, L. C. W. (**1993**). "Objective analysis versus subjective assessment of vowels pronounced by native, non-native, and deaf male speakers of Dutch," J. Acoust. Soc. Am. **94**, 1989–2004.

Bakkum, M. J., Plomp, R., and Pols, L. C. W. (**1995**). "Objective analysis versus subjective assessment of vowels pronounced by deaf and normal-hearing children," J. Acoust. Soc. Am. **98**, 745–762.

Barbosa, P., and Bailly, G. (**1994**). "Characterisation of rhythmic patterns for text-to-speech synthesis," Speech Commun. **15**, 127–137.

Bartkova, K., and Sorin, C. (**1987**). "A model of segmental duration for speech synthesis in French," Speech Commun. **6**, 245–260.

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

Bregman, A. S. (**1990**). *Auditory Scene Analysis — The Perceptual Organization of Sound* (MIT, Cambridge).

Brinkmann, R. D., and Scherg, M. (**1979**). "Human auditory on- and off-potentials of the brainstem," Scand. Audiol. **8**, 27–32.

Campbell, W. N. (**1992**). "Multi-level timing in speech," doctoral dissertation, University of Sussex, Brighton, UK.

Campbell, W. N., and Sagisaka, Y. (**1991**). "Moraic and syllable-level effects on speech timing," Acoustical Society of Japan, Trans. Tech. Committee Speech SP90-107, 35–40.

Carlson, R., and Granström, B. (**1975**). "Perception of segmental duration," in *Structure and Process in Speech Perception*, edited by A. Cohen and S. Nooteboom (Springer-Verlag, Berlin), pp. 90–106.

Carlson, R., and Granström, B. (**1986**). "A search for durational rules in a real-speech database," Phonetica **43**, 140–154.

Carlson, R., Granström, B., and Klatt, D. H. (**1979**). "Some notes on the perception of temporal patterns in speech," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhman (Academic, London), pp. 233–243.

COCOSDA (**1998**). "COCOSDA's home page," The International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques for Speech Input/Output, WWW document http://www.itl.atr.co.jp/cocosda/ (edited and maintained by N. Campbell).

Cohen, A., and Nooteboom, S. G. (**1975**). *Structure and process in speech perception* (Springer-Verlag, Berlin).

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

Crystal, D. (**1997**). *A Dictionary of Linguistics and Phonetics* (Blackwell, Oxford), 4th ed.

Crystal, T. H., and House, A. S. (**1988**). "Segmental durations in connected-speech signals: Current results," J. Acoust. Soc. Am. **83**, 1553–1573.

Delattre, P. C. (**1962**). "Some factors of vowel duration and their cross-linguistic validity," J. Acoust. Soc. Am. **34**, 1141–1143.

Divenyi, P. L., and Danner, W. F. (**1977**). "Discrimination of time intervals marked by brief acoustic pulses of various intensities and spectra," Percept. Psychophys. **21**, 125–142.

Divenyi, P. L., and Sachs, R. M. (**1978**). "Discrimination of time intervals bounded by tone bursts," Percept. Psychophys. **24**, 429–436.

Fant, G., and Kruckenberg, A. (**1989**). "Preliminaries to the study of Swedish prose reading and reading style," Royal Institute of Technology, Speech Transmission Lab. Q. Prog. Status Report 2/1989, 1–83.

Fant, G., and Tatham, M. A. A. (**1975**). *Auditory Analysis and Perception of Speech* (Academic, London).

Florentine, M. (**1983**). "Intensity discrimination as a function of level and frequency and its relation to high-frequency hearing," J. Acoust. Soc. Am. **74**, 1375–1379.

Fowler, C. A. (**1979**). "'Perceptual centers' in speech production and perception," Percept. Psychophys. **25**, 375–388.

Fowler, C. A. (**1983**). "Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet," J. Exp. Psychol.: General **112**, 386–412.

Fry, D. B. (**1955**). "Duration and intensity as physical correlates of linguistic stress," J. Acoust. Soc. Am. **27**, 765–768.

Fujisaki, H., and Higuchi, N. (**1980**). "Temporal organization of segmental features in Japanese disyllable," J. Acoust. Soc. Jpn. (E) **1**, 25–30.

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

Getty, D. J. (**1975**). "Discrimination of short temporal intervals: A comparison of two models," Percept. Psychophys. **18**, 1–8.

Goldfarb, J. L., and Goldstone, S. (**1963**). "Time judgment: A comparison of filled and unfilled durations," Percept. Mot. Skills **16**, 376.

Green, D. M. (**1993**). "Auditory intensity discrimination," in *Human Psychophysics*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer-Verlag, New York), pp. 13–55.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Grondin, S. (**1993**). "Duration discrimination of empty and filled intervals marked by auditory and visual signals," Percept. Psychophys. **54**, 383–394.

Grondin, S., Ivry, R. B., Franz, E., Perreault, L., and Metthé, L. (**1996**). "Markers' influence on the duration discrimination of intermodal intervals," Percept. Psychophys. **58**, 424–433.

Hashimoto, S. (**1973**). "Several features of Japanese word accent," Trans. Inst. Electron. Commun. Eng. **56-D**, 654–661 (in Japanese with English figure captions).

Henry, F. M. (**1948**). "Discrimination of the duration of a sound," J. Exp. Psychol. **38**, 734–743.

Higuchi, N., Shimizu, T., Kawai, H., and Yamamoto, S. (**1993**). "Control of phoneme duration based on the movement of speech organs," J. Acoust. Soc. Jpn. (E) **14**, 281–283.

Hiki, S. (**1967**). "Effects of the context on the duration of phonemic segment," J. Acoust. Soc. Jpn. **23**, 317–318.

Hirsh, I., Monahan, C., Grant, K., and Singh, P. (**1990**). "Studies in auditory timing: 1. Simple patterns," Percept. Psychophys. **47**, 215–226.

Hoshino, M., and Fujisaki, H. (**1983**). "A study on perception of changes in segmental durations," Acoustical Society of Japan, Trans. Tech. Committee Speech S82-75, 593–599 (in Japanese with English abstract and English figure captions).

House, A. S. (**1961**). "On vowel duration in English," J. Acoust. Soc. Am. **33**, 1174–1178.

Huggins, A. W. F. (**1968**). "The perception of timing in natural speech: Compensation within the syllable," Lang. Speech **11**, 1–11.

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

Huggins, A. W. F. (**1972b**). "On the perception of temporal phenomena in speech," J. Acoust. Soc. Am. **51**, 1279–1290.

Hyman, L. M. (**1975**). *Phonology: Theory and Analysis* (Holt, Rinehart and Winston, New York).

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

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

Jacobson, R., Fant, G., and Halle, M. (**1954**). *Preliminaries to Speech Analysis: the Distinctive Features and Their Correlates* (MIT Press, Cambridge, MA).

Jesteadt, W., Wier, C. C., and Green, D. M. (**1977**). "Intensity discrimination as a function of frequency and sensation level," J. Acoust. Soc. Am. **61**, 169–177.

Kaiki, N., and Sagisaka, Y. (**1992**). "The control of segmental duration in speech synthesis using statistical methods," in *Speech Perception, Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (IOS, Amsterdam), pp. 391–402.

Kaiki, N., Takeda, K., and Sagisaka, Y. (**1992**). "Linguistic properties in the control of segmental duration for speech synthesis," in *Talking Machines: Theories, Models, and Designs*, edited by G. Bailly, C. Benoît, and T. R. Sawallis (Elsevier, Amsterdam), pp. 255–263.

Kasuya, H. (**1992**). "Assessment of speech synthesis technology," J. Acoust. Soc. Jpn. **48**, 46–51 (in Japanese).

Kasuya, H. (**1993**). "Methods to evaluate synthetic speech quality," J. Acoust. Soc. Jpn. **49**, 866–870 (in Japanese).

Kato, H., and Tsuzaki, M. (**1994**). "Intensity effect on discrimination of auditory duration flanked by preceding and succeeding tones," J. Acoust. Soc. Jpn. (E) **15**, 349–351 (also Section 6.3).

Kato, H., and Tsuzaki, M. (**1998**). "Evidence for functional differences between rise and fall markers in discrimination of auditory filled durations," J. Acoust. Soc. Jpn. (E) **19**, 73–76 (also Section 6.5).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**a**). "Effects of phonetic quality and duration on the acceptability of modifications in speech portions," J. Acoust. Soc. Am. (also Chapter 3; in review).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**b**). "Functional differences between vowel onsets and offsets in temporal perception of speech," J. Acoust. Soc. Am. (also Chapter 5; in review).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**c**). "A modeling of the objective evaluation of durational rules based on auditory perceptual characteristics," J. Acoust. Soc. Jpn. (also Chapter 7; in Japanese with English abstract and English figure captions; in review).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1992**). "Acceptability and discrimination threshold for distortion of segmental duration in Japanese words," in *Proceedings of the 2nd International Conference on Spoken Language Processing* (University of Alberta, Edmonton, AB), pp. 507–510 (also partly incorporated in Section 6.1).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1996**). "Evidence for the predominance of vowel onsets to offsets in speaking-rate perception," in *Proceedings of the 3rd ASJ-ASA Joint Meeting* (Acoustical Society of Japan, Tokyo), pp. 1199–1204.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1997**). "Acceptability for temporal modification of consecutive segments in isolated words," J. Acoust. Soc. Am. **101**, 2311–2322 (also Chapter 4 and Section 6.4).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1998a**). "Acceptability for temporal modification of single vowel segments in isolated words," J. Acoust. Soc. Am. **104**, 540–549 (also Chapter 2).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1998b**). "Effects of phonetic quality and duration on perceptual acceptability of temporal changes in speech," in *Proceedings of the 5th International Conference on Spoken Language Processing* (Australian Speech Science & Technology Association Inc., Sydney, Australia), pp. 2171–2174 (also incorporated in Chapter 3).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1998c**). "Functional difference between onsets and offsets in perceiving temporal structure of speech," in *Proceedings of the ATR Workshop on Events and Auditory Temporal Structure* (ATR Human Information Processing Research Labs., Kyoto, Japan), pp. 29–33 (also incorporated in Chapter 5).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1999**). "A modeling of the objective evaluation of durational rules based on auditory perceptual characteristics," in *Proceedings of the 14th International Congress of Phonetic Sciences* (University of California, Berkeley, CA) (to appear).

Kato, M., and Hashimoto, S. (**1992**). "Rhythm rules in Japanese based on the center of energy gravity of vowels," in *Proceedings of the 2nd International Conference on Spoken Language Processing* (University of Alberta, Edmonton, AB), pp. 1139–1142.

Klatt, D. H. (**1976**). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**, 1208–1221.

Klatt, D. H. (**1979**). "Synthesis by rule of segmental durations in English sentences," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhman (Academic, London), pp. 287–299.

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

Kodera, K., Yamane, H., Yamada, O., and Suzuki, J.-I. (**1977**). "The effect of onset, offset and rise-decay times of tone bursts on brain stem response," Scand. Audiol. **6**, 205–210.

Kristofferson, A. B. (**1977**). "A real-time criterion theory of duration discrimination," Percept. Psychophys. **21**, 105–117.

Kubozono, H. (**1999**). "Mora and syllable," in *The Handbook of Japanese Linguistics*, edited by N. Tsujimura (Blackwell, Oxford) (in press).

Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., and Shikano, K. (**1990**). "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Commun. **9**, 357–363.

Lehiste, I. (**1970**). *Suprasegmentals* (MIT, Cambridge).

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

Lehiste, I., and Peterson, G. E. (**1959**). "Vowel amplitude and phonemic stress in American English," J. Acoust. Soc. Am. **31**, 428–435.

Liberman, A. M., and Mattingly, I. G. (**1989**). "A specialization for speech perception," Science **243**, 489–494.

Lieberman, P. (**1960**). "Some acoustic correlates of word stress in American English," J. Acoust. Soc. Am. **32**, 451–454.

Lindblom, B., and Öhman, S. E. G. (**1979**). *Frontiers of Speech Communication Research* (Academic, London).

Luce, P. A., and Charles-Luce, J. (**1985**). "Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production," J. Acoust. Soc. Am. **78**, 1949–1957.

Marcus, S. M. (**1981**). "Acoustic determinants of perceptual center (P-center) location," Percept. Psychophys. **30**, 247–256.

Massaro, D. W., and Cohen, M. M. (**1975**). "Perceptual auditory storage in speech recognition," in *Structure and Process in Speech Perception*, edited by A. Cohen and S. G. Nooteboom (Springer-Verlag, Berlin), pp. 226–245.

McAuley, J. D., and Kidd, G. R. (**1998**). "Effect of deviations from temporal expectations on tempo discrimination of isochronous tone sequences," J. Exp. Psychol.: Hum. Percept. Perform. **24**, 1–15.

McCall, R. B. (**1980**). *Fundamental Statistics for Psychology* (HBJ, New York), 3rd ed.

Mimura, K., Kaiki, N., and Sagisaka, Y. (**1991**). "Statistically derived rules for amplitude and duration control in Japanese speech synthesis," in *Proceedings of the Korea-Japan Joint Workshop on Advanced Technology of Speech Recognition and Synthesis*, pp. 151–156.

Morton, J., Marcus, S., and Frankish, C. (**1976**). "Perceptual centers," Psychol. Rev. **83**, 405–408.

Nakajima, Y. (**1987**). "A model of empty duration perception," Perception **16**, 485–520.

Nakajima, Y., and Sasaki, T. (**1996**). "A simple grammar of auditory stream formation," J. Acoust. Soc. Am. **100**, 2681.

Nusbaum, H. C., Francis, A. L., and Henly, A. S. (**1995**). "Measuring the naturalness of synthetic speech," Int. J. Speech Technol. **1**, 7–19.

Peterson, G. E., and Lehiste, I. (**1960**). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. **32**, 693–703.

Pols, L. C. W. (**1991**). "Speech perception in the next ten years: Technological solutions vs. acquiring actual speech knowledge," in *Proceedings of the XIIth International Congress of Phonetic Sciences* (Universite de Provence, Aix-en-Provence, France), pp. 130–133.

Rabinowitz, W. M., Lim, J. S., Braida, L. D., and Durlach, N. I. (**1976**). "Intensity perception. VI. Summary of recent data on deviations from Weber's law for 1000-Hz tone pulses," J. Acoust. Soc. Am. **59**, 1506–1509.

Rammsayer, T. H. (**1994**). "Effects of practice and signal energy on duration discrimination of brief auditory intervals," Percept. Psychophys. **55**, 454–464.

Rapp, K. (**1971**). "A study of syllable timing," Royal Institute of Technology, Speech Transmission Lab. Q. Prog. Status Report 1/1971, 14-19.

Riesz, R. R. (**1928**). "Differential intensity sensitivity of the ear for pure tones," Phys. Rev. **31**, 867–875.

Riley, M. (**1992**). "Tree-based modelling of segmental durations," in *Talking Machines: Theories, Models, and Designs*, edited by G. Bailly, C. Benoît, and T. R. Sawallis (Elsevier, Amsterdam), pp. 265–273.

Ruhm, H. B., Mencke, E. O., Milburn, B., Cooper, W. A., and Rose, D. E. (**1966**). "Differential sensitivity to duration of acoustic signals," J. Speech Hear. Res. **9**, 371–384.

Sagisaka, Y. (**1998**). "Corpus based speech synthesis," J. Signal Process. **2**, 407–414 (in Japanese with English figure captions).

Sagisaka, Y., Takeda, K., Abe, M., Katagiri, S., Umeda, T., and Kuwabara, H. (**1990**). "A large-scale Japanese speech database," in *Proceedings of the 1st International Conference on Spoken Language Processing* (Acoustical Society of Japan, Tokyo), pp. 1089–1092.

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

SAS Institute Inc. (**1990**). *SAS/STAT User's Guide, Version 6* (SAS Institute Inc., Cary, NC), 4th ed., Vol. 2.

Sato, H. (**1977**). "Segmental duration and timing location in speech," Acoustical Society of Japan, Trans. Tech. Committee Speech S77-31, 1–8 (in Japanese with English abstract and English figure captions).

Schouten, M. E. H. (**1980**). "The case against the speech mode of perception," Acta Psychol. **44**, 71–98.

Schroder, A. C., Viemeister, N. F., and Nelson, D. A. (**1994**). "Intensity discrimination in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **96**, 2683–2693.

Schulze, H.-H. (**1978**). "The detectability of local and global displacements in regular rhythmic patterns," Psychol. Res. **40**, 173–181.

Scott, S. K. (**1993**). "P-centres in speech — An acoustic analysis," doctoral dissertation, University College London, London, UK.

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

Takeda, K., Sagisaka, Y., and Kuwabara, H. (**1989**). "On sentence-level factors governing segmental duration in Japanese," J. Acoust. Soc. Am. **86**, 2081–2087.

Tanaka, M., Tsuzaki, M., and Kato, H. (**1992**). "On the perception of the click sequence simulating moraic structure," Acoustical Society of Japan, Trans. Tech. Committee Physiol. Psychol. Acoust. H-92-63, 1–7 (also partly incorporated in Section 6.2; in Japanese with English figure captions).

Tanaka, M., Tsuzaki, M., and Kato, H. (**1994**). "Discrimination of empty duration in the click sequence simulating a mora structure," J. Acoust. Soc. Jpn. (E) **15**, 191–192 (also incorporated in Section 6.2).

ten Hoopen, G., Beumer, M., and Nakajima, Y. (**1996**). "What differs between the first and the last interval of a click sequence simulating a mora structure, the DT or the PSE?: A replication of Tanaka, Tsuzaki, and Kato (1994)," J. Acoust. Soc. Jpn. (E) **17**, 155–158.

Tohkura, Y., Vatikiotis-Bateson, E., and Sagisaka, Y. (**1992**). *Speech Perception, Production and Linguistic Structure* (IOS, Amsterdam).

Torgerson, W. S. (**1958**). "The law of categorical judgment," in *Theory and Methods of Scaling* (Wiley, New York), pp. 205–246.

Tsujimura, N. (**1996**). *An Introduction to Japanese Linguistics* (Blackwell, Oxford).

Tsuzaki, M., and Kato, H. (**1998**). "Perceptual reconstruction of temporal structure: Duration and timing," in *Proceedings of the ATR Workshop on Events and Auditory Temporal Structure* (ATR Human Information Processing Research Labs., Kyoto, Japan), pp. 23–28.

Tuller, B., and Fowler, C. A. (**1980**). "Some articulatory correlates of perceptual isochrony," Percept. Psychophys. **27**, 277–283.

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

Umeda, N. (**1975**). "Vowel duration in American English," J. Acoust. Soc. Am. **58**, 434–445.

van Santen, J. P. H. (**1992**). "Contextual effects on vowel duration," Speech Commun. **11**, 513–546.

van Santen, J. P. H. (**1994**). "Assignment of segmental duration in text-to-speech synthesis," Comput. Speech Lang. **8**, 95–128.

van Santen, J. P. H., Sproat, R. W., Olive, J. P., and Hirschberg, J. (**1997**). *Progress in Speech Synthesis* (Springer-Verlag, New York).

van Wieringen, A. (**1995**). "Perceiving dynamic speechlike sounds: Psycho-acoustics and speech perception," doctoral dissertation, University of Amsterdam, Amsterdam.

Vance, T. J. (**1987**). *An Introduction to Japanese Phonology* (SUNY Press, New York).

Yokoyama, S. (**1981**). "Occurrence frequency data of a Japanese dictionary," Bull. Electrotechnical Laboratory Jpn. **45**, 395–418.

Yost, W. A., Popper, A. N., and Fay, R. R. (**1993**). *Human Psychophysics* (Springer-Verlag, New York), Vol. 3 of *Springer Handbook of Auditory Research*.

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

List of Publications[1]

Professional journal full papers

[2]Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1997**). "Acceptability for temporal modification of consecutive segments in isolated words," Journal of the Acoustical Society of America **101**, 2311–2322.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1998**). "Acceptability for temporal modification of single vowel segments in isolated words," Journal of the Acoustical Society of America **104**, 540–549.

Kato, H., Tsuzaki, M., and Sagisaka, Y. "Effects of phonetic quality and duration on the acceptability of modifications in speech portions," Journal of the Acoustical Society of America (in review).

Kato, H., Tsuzaki, M., and Sagisaka, Y. "A modeling of the objective evaluation of durational rules based on auditory perceptual characteristics," Journal of the Acoustical Society of Japan (in Japanese with English abstract and English figure captions; in review).

Tsuzaki, M., Kato, H., and Tanaka, M. "Shrinkage in the perceived duration of speech and tone by acoustic replacement," Japan Psychological Research (in review).

Kato, H., Tsuzaki, M., and Sagisaka, Y. "Functional differences between vowel onsets and offsets in temporal perception of speech," Journal of the Acoustical Society of America (in review).

[1]The list includes only papers concerned with this dissertation.
[2]Honored with the 1997 ATR Paper Award.

Professional journal short papers

Kato, H., and Tsuzaki, M. (**1994**). "Intensity effect on discrimination of auditory duration flanked by preceding and succeeding tones," Journal of the Acoustical Society of Japan (E) **15**, 349–351.

Tanaka, M., Tsuzaki, M., and Kato, H. (**1994**). "Discrimination of empty duration in the click sequence simulating a mora structure," Journal of the Acoustical Society of Japan (E) **15**, 191–192.

Tsuzaki, M., Kato, H., and Tanaka, M. (**1994**). "Apparent duration of speech segments interrupted by a noise burst," Journal of the Acoustical Society of Japan (E) **15**, 205–206.

Kato, H., and Tsuzaki, M. (**1998**). "Evidence for functional differences between rise and fall markers in discrimination of auditory filled durations," Journal of the Acoustical Society of Japan (E) **19**, 73–76.

Refereed book chapters

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1997**). "Measuring temporal compensation effect in speech perception," in *Computing Prosody: Computational Models for Processing Spontaneous Speech*, edited by Y. Sagisaka, W. N. Campbell, and N. Higuchi (Springer-Verlag, New-York), pp. 251–270.

International conferences

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1992**). "Acceptability and discrimination threshold for distortion of segmental duration in Japanese words," in *Proceedings of the 2nd International Conference on Spoken Language Processing* (University of Alberta, Edmonton, AB), pp. 507–510.

Tsuzaki, M., Kato, H., and Tanaka, M. (**1993**). "The apparent duration of the restored speech segments," Journal of the Acoustical Society of America **93**, 2403(A).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1994**). "Acceptability of temporal modification in consonant and vowel onsets," in *Proceedings of the 3rd International Conference on Spoken Language Processing* (Acoustical Society of Japan, Tokyo), pp. 1979–1982.

Tsuzaki, M., Kato, H., and Tanaka, M. (**1994**). "Effects of acoustic discontinuity and phonemic deviation on the apparent duration of speech segments," in *Proceedings of the 3rd International Conference on Spoken Language Processing* (Acoustical Society of Japan, Tokyo), pp. 1971–1974.

Tsuzaki, M., and Kato, H. (**1995**). "Shrinking illusion of speech duration caused by noise/gap replacement," Journal of the Acoustical Society of America **97**, 3275(A).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1996**). "Evidence for the predominance of vowel onsets to offsets in speaking-rate perception," Journal of the Acoustical Society of America **100**, 2829.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1996**). "Evidence for the predominance of vowel onsets to offsets in speaking-rate perception," in *Proceedings of the 3rd ASJ-ASA Joint Meeting* (Acoustical Society of Japan, Tokyo), pp. 1199–1204.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1998**). "Effects of phonetic quality and duration on perceptual acceptability of temporal changes in speech," in *Proceedings of the 5th International Conference on Spoken Language Processing* (Australian Speech Science & Technology Association Inc., Sydney, Australia), pp. 2171–2174.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1999**). "A psychoacoustic study on temporal compensation between consonant and vowel segments," in *Proceedings of the 4th Linguistics and Phonetics Conference* (Ohio State University, Columbus, OH) (to appear).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1999**). "A modeling of the objective evaluation of durational rules based on auditory perceptual characteristics," in *Proceedings of the 14th International Congress of Phonetic Sciences* (University of California, Berkeley, CA) (to appear).

International workshops

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1995**). "Measuring temporal compensation effect in speech perception," in *Proceedings of the ATR International Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing* (ATR Interpreting Telecommunications Research Labs., Kyoto, Japan), pp. 3.38–3.49.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1998**). "Functional difference between onsets and offsets in perceiving temporal structure of speech," in *Proceedings of the ATR Workshop on Events and Auditory Temporal Structure* (ATR Human Information Processing Research Labs., Kyoto, Japan), pp. 29–33.

Tsuzaki, M., and Kato, H. (**1998**). "Perceptual reconstruction of temporal structure: Duration and timing," in *Proceedings of the ATR Workshop on Events and Auditory Temporal Structure* (ATR Human Information Processing Research Labs., Kyoto, Japan), pp. 23–28.

Other technical meetings

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1991**). "Discrimination threshold for segmental duration in words: Effects of mora position, F0 contour, and vowel color," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 391–392 (in Japanese).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1991**). "Discrimination threshold for segmental duration in words: Effects of mora position, F0 contour, and vowel color," Acoustical Society of Japan, Trans. Tech. Committee Physiol. Psychol. Acoust. H-91-43, 1–7 (in Japanese).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1992**). "Estimation of acceptability for distortion of segmental duration in words: Effects of vowel color, mora position, and tone accent," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 243–244 (in Japanese with English figure captions).

Tanaka, M., Tsuzaki, M., and Kato, H. (**1992**). "Discrimination of empty durations in the click sequence simulating the moraic structure," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 451–452 (in Japanese with English figure captions).

Tanaka, M., Tsuzaki, M., and Kato, H. (**1992**). "On the perception of the click sequence simulating moraic structure," Acoustical Society of Japan, Trans. Tech. Committee Physiol. Psychol. Acoust. H-92-63, 1–7 (in Japanese with English figure captions).

Kato, H., and Tsuzaki, M. (**1993**). "Relation between duration discrimination and level difference against preceding/succeeding tone," Acoustical Society of Japan, Trans.

Tech. Committee Physiol. Psychol. Acoust. H-93-86, 1–10 (in Japanese with English abstract).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1993**). "Acceptability for durational modification of segments in words," Acoustical Society of Japan, Trans. Tech. Committee Speech SP92-145, 65–72 (in Japanese with English abstract and English figure captions).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1993**). "Estimation of acceptability for distortion of segmental duration in words: Effects of segment with moraic nasal, geminate, or devocalized vowel," in *Proceedings of the Spring Meeting* (Acoustical Society of Japan, Tokyo), pp. 237–238 (in Japanese with English figure captions).

Tanaka, M., Tsuzaki, M., and Kato, H. (**1993**). "Perception of the click sequence simulating moraic structure," in *Proceedings of the Spring Meeting* (Acoustical Society of Japan, Tokyo), pp. 345–346 (in Japanese with English figure captions).

Tsuzaki, M., and Kato, H. (**1993**). "Apparent duration of restored speech segment," Acoustical Society of Japan, Trans. Tech. Committee Physiol. Psychol. Acoust. H-93-68, 1–10 (in Japanese with English abstract).

Tsuzaki, M., Kato, H., and Tanaka, M. (**1993**). "Apparent duration of the restored speech segment: Measurement of discrimination threshold," in *Proceedings of the Spring Meeting* (Acoustical Society of Japan, Tokyo), pp. 365–366 (in Japanese with English figure captions).

Tsuzaki, M., Kato, H., and Tanaka, M. (**1993**). "Apparent duration of the restored speech: Measurement of PSE," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 435–436 (in Japanese).

Kato, H., and Tsuzaki, M. (**1994**). "Relation between intensity effect on durational discrimination and duration of preceding/succeeding sound," in *Proceedings of the Spring Meeting* (Acoustical Society of Japan, Tokyo), pp. 371–372 (in Japanese with English figure captions).

Kato, H., and Tsuzaki, M. (**1994**). "Discrimination of temporal structures with loudness changes as timing markers," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 555–556 (in Japanese with English figure captions).

[3]Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1994**). "Measuring durational compensation effect of speech perception between neighboring segments," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 315–316 (in Japanese with English figure captions).

Tanaka, M., Tsuzaki, M., and Kato, H. (**1994**). "The effect of marker quality on discrimination of empty duration," in *Proceedings of the Spring Meeting* (Acoustical Society of Japan, Tokyo), pp. 373–374 (in Japanese with English figure captions).

Tsuzaki, M., Kato, H., and Tanaka, M. (**1994**). "Apparent duration of the interrupted speech: Comparison between noise and gap," in *Proceedings of the Spring Meeting* (Acoustical Society of Japan, Tokyo), pp. 369–370 (in Japanese with English figure captions).

Tsuzaki, M., Kato, H., and Tanaka, M. (**1994**). "Illusion in apparent duration of auditory events interrupted by noise burst," Acoustical Society of Japan, Trans. Tech. Committee Physiol. Psychol. Acoust. H-94-73, 1–8 (in Japanese with English abstract and English figure captions).

Tsuzaki, M., and Kato, H. (**1994**). "Effects of noise replacement on the apparent duration of spoken words: A study of the stimulus continuum from short-vowel to long-vowel," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 557–558 (in Japanese with English figure captions).

Kato, H., and Tsuzaki, M. (**1995**). "Temporal discrimination of part of tone marked by two amplitude changes: Comparison among on-marker, off-marker, and their combinations," in *Proceedings of the Spring Meeting* (Acoustical Society of Japan, Tokyo), pp. 461–462 (in Japanese with English figure captions).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1995**). "A modeling of subjective evaluation for temporal distortions in segmental duration," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 221–222 (in Japanese with English figure captions).

Tsuzaki, M., and Kato, H. (**1995**). "Apparent duration of a short tone burst and a sequence of tone bursts interrupted by noise," in *Proceedings of the Spring Meeting* (Acoustical Society of Japan, Tokyo), pp. 463–464 (in Japanese with English figure captions).

[3]Honored with the 1994 Awaya Prize by the Acoustical Society of Japan.

Tsuzaki, M., and Kato, H. (**1995**). "Shrinking of perceived duration by acoustic interruption: Effects of temporal position and "continuity"," Acoustical Society of Japan, Trans. Tech. Committee Speech H-95-49/SP95-45, 25–30 (in Japanese with English abstract and English figure captions).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1996**). "Acceptability for temporal modification of two consecutive segments in words," Acoustical Society of Japan, Trans. Tech. Committee Speech H-96-26/SP95-149, 53–60 (in Japanese with English abstract and English figure captions).

Tsuzaki, M., and Kato, H. (**1996**). "Effects of concurrent sounds on duration perception," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 449–450 (in Japanese).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1997**). "Functional differences between vowel onsets and offsets in temporal perception of speech: Detectability of temporal modifications," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 365–366 (in Japanese with English figure captions).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1997**). "Functional differences between vowel onsets and offsets in temporal perception of speech: Speaking rate estimation," in *Proceedings of the Fall Meeting* (Acoustical Society of Japan, Tokyo), pp. 367–368 (in Japanese with English figure captions).

Tsuzaki, M., and Kato, H. (**1997**). "Shrinkage of apparent duration of non-isochronous tone sequence," Acoustical Society of Japan, Trans. Tech. Committee Physiol. Psychol. Acoust. H-97-87, 1–8 (in Japanese with English abstract and English figure captions).

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**1998**). "Acceptability of temporal modifications in segments including special morae," Acoustical Society of Japan, Trans. Tech. Committee Speech H-98-27/SP97-132, 15–22 (in Japanese with English abstract and English figure captions).

Tsuzaki, M., and Kato, H. (**1998**). "Effects of concurrent sounds on apparent duration: An experiment in the non-isochronous context," in *Proceedings of the Spring Meeting* (Acoustical Society of Japan, Tokyo), pp. 457–458 (in Japanese with English figure captions).

Other publications

Kato, H. (**1998**). “How do humans perceive temporal structures in a spoken language?,” ATR Journal (E) (Advanced Telecommunications Research Institute International, Kyoto, Japan), **1**, 42–43.

Kato, H. (**1998**). “How do humans perceive the temporal structure of spoken language?,” ATR Journal (Advanced Telecommunications Research Institute International, Kyoto, Japan), **30**, 6–7 (in Japanese with English abstract).