MICHAEL O´DELL

Intrinsic Timing and Quantity in Finnish

■

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Humanities of the University of Tampere,
for public discussion in the Pinni auditorium B1100
of the University, Kanslerinrinne 1, Tampere,
on January 9th, 2004, at 12 o'clock.

Acta Universitatis Tamperensis 979
University of Tampere
Tampere 2003

ACADEMIC DISSERTATION
University of Tampere, The School of Languages and Translation Studies
Finland

Distribution

University of Tampere
Bookshop TAJU
P.O. Box 617
33014 University of Tampere
Finland

Tel. +358 3 215 6055
Fax +358 3 215 7685
taju@uta.fi
http://granum.uta.fi

Cover design by
Juha Siro

Printed dissertation
Acta Universitatis Tamperensis 979
ISBN 951-44-5837-0
ISSN 1455-1616

Electronic dissertation
Acta Electronica Universitatis Tamperensis 307
ISBN 951-44-5838-9
ISSN 1456-954X
http://acta.uta.fi

Tampereen yliopistopaino Oy Juvenes Print
Tampere 2003

Contents

List of Figures

List of Tables

Acknowledgements

> We can't return, we can only look behind
> From where we came
> And go round and round and round
> In the circle game
>
> —Joni Mitchell

In scientific circles as elsewhere, "No man is an island." So many people have contributed to my intellectual development in numerous ways over the years, that it would be impossible to thank them all individually. I therefore hereby collectively thank you all. Naturally any mistakes I may have made I claim as my own.

I do wish, however, to express a special thanks to a few select individuals. To start at the beginning, I thank my wonderful parents Doyal and Phoebe for providing me an environment as I grew up that was at once loving and intellectually stimulating. I simply cannot imagine having better parents, and I dedicate this work to them.

I thank Bob Port at Indiana University for opening my eyes (and ears!) to the joys of phonetics research in the first place.

My first "phonetic" acquaintance in Finland, Hannele Dufva, I thank for so much fruitful discussion and intellectual support over the years as well as enduring friendship.

To my good friend and colleague at the University of Tampere Urho Määttä I owe a debt of gratitude which can scarcely be repaid. In addition to all the hearty discussions on subjects from linguistic metatheory to philosophy of science, it is no exaggeration to say that without the support and encouragement of Urho, who acted as my advisor, the present work would hardly have seen the light of day.

For comradery and numerous conversations on all manner of weighty subjects I thank the rest of the "Huuhaa Club" as well, Pekka Pälli, Esa Lehtinen, and especially my fellow phonetician Tommi Nieminen, with whom I have had the good fortune to collaborate on several articles. I thank also the other members, past and present, of the Department of Finnish and General Linguistics at the Universty of Tampere for a supportive working environment.

I warmly thank the two reviewers of this dissertation, Elliot Saltzman and Stefan Werner, for their helpful comments and support of my work which went well beyond the call of duty. I hope we will find ways in the future to continue the fruitful dialogue we have started.

To my wife Sirpa I extend my deepest gratitude for her unfailing support and encouragement, and her unselfish willingness to share the burdon with me. My sons Ville and Pekka I thank for all the joy they have given me.

Valkeakoski
December, 2003 Michael L. O'Dell

Chapter 1

Background

> T.T.T.
>
> Put up in a place
> where it's easy to see
> the cryptic admonishment
> T.T.T.
>
> When you feel how depressingly
> slowly you climb,
> it's well to remember that
> Things Take Time.
>
> —Piet Hein

The two terms in the title of this report, *quantity* and *intrinsic timing*, present an apparent conflict which has provided the main impetus for the research reported herein. The intrinsic timing concept represents a theoretically attractive model, but almost no research within such a framework has been carried out on a quantity language. This report presents research on some aspects of phonetic timing in Finnish, a language often mentioned as a prototypical example of a quantity language.

In the literature of phonetics the terms *intrinsic timing* and *extrinsic timing* introduced by Fowler (1977, 1980) have been used to distinguish theories of speech timing. *Extrinsic timing* refers to theories in which the speaker forces his speech to conform to a certain temporal mold regardless of the articulation involved. Thus in such a theory timing and articulation are essentially independent. On the other hand the term *intrinsic timing* has been applied to theories in which measured durations, for instance, are the result of various dynamic properties of the coordinated articulations themselves.

The term *quantity language* is generally taken to refer to a language in which durations (often restricted to durations of segments) are utilized to distinguish words. For instance, in her well known book *Suprasegmentals*, Lehiste says "The term *quantity* will be applied to duration when it functions as an independent variable in the phonological system of a language." (Lehiste, 1970, pp. 42–43). Catford writes "... articulatory duration is much under conscious control: within the limits sketched above one can vary duration as desired. Consequently, many languages make phonological use of duration differences: the phonological use of duration is known as 'quantity'." (Catford, 1977, p. 197). In such a language durational differences are evidently intentional, and thus it seems that they must be directly controlled by the speaker (and attended to by the listener). Therefore the existence of quantity languages, of which Finnish is an often cited example, would seem to support the notion of extrinsic timing, at least for quantity languages, and to form a barrier to acceptance of intrinsic timing theories.

Some confusion results from the fact that the term *duration*[1] itself is used in various ways. It is customary in modern phonetics to use separate terms to distinguish different varieties of a quantity depending on whether the discussion is about "physical" (purely mechanical or acoustic) phenomena, "psychological" (perceptual or psychophysical) phenomena, or "phonological" (linguistic, functional) phenomena. Also different scales are used to measure related phenomena at different levels of the so-called speech chain (at the phonological level it is often presumed there is not a continuous scale, but rather a set of discrete categories). For instance, (fundamental) *frequency* measured in *Hertz* on the physical level is correlated with *pitch*[2] on the psychological level, which can be measured using a psychophysical *Mel* scale, and with *tone* and *intonation* on the phonological level. Likewise, physical *intensity* (*decibel* scale) is associated with psychological *loudness* (*sone* scale) and phonological *stress* (cf. eg. Lehiste, 1970; Ladefoged, 1975; Catford, 1977; Borden and Harris, 1984).

Interestingly, physical *duration* (measured in *milliseconds* or *centiseconds*) is associated in textbooks with phonological *length* (ie. discrete levels of quantity such as short vs. long), but rarely is a corresponding psychological phenomenon discussed, and duration is often used indiscriminately for both. In analogy with the other dimensions, Fry does systematically distinguish a *perceptual dimension of length*, but notes that "we know practically nothing about its relation with duration." (Fry, 1968, p. 386).

Lehtonen (1969) called attention to another related aspect of quantity, in a sense intermediate between physical duration (Lehtonen's *objective quantity*) and phonological length (Lehtonen's *subjective quantity*). He refers to nondistinctive (subphonemic) quantity related differences perceived (and transcribed "by ear") by trained phoneticians. It is perhaps this level which could most justifiably be called "phonetic" quantity. Lehtonen points out that the problem for phonetic description, however, is that the phonetician's transcription is strongly dependent on his own previous arbitrary language experience, especially his native dialect. In response to this problem, Lehtonen

[1] Also, in the older literature, the term *quantity* is often used as a synonym for physical duration, whether distinctive or not.

[2] In the literature on speech signal processing, however, the term *pitch* is often used synonymously with fundamental frequency; in particular the term *pitch period* is used to refer to the number of sample points in a single fundamental period and is not (directly) associated with psychological phenomena.

used perception experiments to investigate how varying the measurable durations of phonetic segments affected classification by native speakers of Finnish into phonological quantity categories (Lehtonen, 1969, 1970).

The problems involved are to a large extent parallel to those delineated in Ladefoged's classic discussion of vowel quality (Ladefoged, 1967, Chapter 2, to which Lehtonen also refers). Ladefoged divides vowel quality into *personal quality* conveying personal information about the speaker and *phonetic quality*, which conveys both linguistic information and social or accentual information (cf. Ladefoged's Figure 22, p. 61 and Figure 43, p. 104).

Ladefoged notes that there are various theories available, which allow (phonetic) classification of vowels so that

> It is often possible for a phonetician to describe a vowel in such a way that another phonetician who has not heard the vowel is nevertheless aware of the particular vowel quality that has been specified. However, the assumptions underlying the theories of vowel classification are seldom explicitly stated." (Ladefoged, 1967, p. 52)

Would it be possible to distinguish "personal" quantity and "phonetic" quantity? The theory of phonetic quantity and timing is much less developed than theories of vowel classification, even today. We don't really understand what quantity distinctions are, except "fairly short" opposed to "fairly long".

For many purposes, it would be advantageous to have some instrumental measure corresponding to the production and perception of speech timing and quantity patterns, eg. for cross-linguistic and sociolinguistic research (for instance quantity considered as a Labovian variable). However, quantity variation is extremely difficult to measure. It is true that some instrumental work on dialect differences in Finnish quantity has been carried out fairly successfully using relative durations (cf. Wiik, 1975, 1985). However, we are certainly not in a position to characterize individual utterances instrumentally in terms of phonetic quantity or timing features. For instance, we don't know how to separate quantity effects from speaking rate, let alone differentiating different varieties of quantity from each other. In a sociolinguistic study of the so-called primary gemination of Finnish dialects, Nahkola noted

> In general it can be said that the phonetic variation of segmental durations is so great, that the "significant" variability which an incipient gemination causes in duration is relatively difficult to separate from the overall variation in duration. The difference is especially difficult in experimental analysis. (Nahkola, 1987, p. 16, my translation)

The present report investigates some aspects of the production and perception of quantity in Finnish, a language well known for its extensive use of phonological length oppositions. It is hoped that this work will go some small way to increase understanding of quantity production and perception and thus to clarify the theoretical basis underlying phonetic vocabulary for the description of quantity patterns. The work reported herein is limited in scope to what might be called phonetic aspects of quantity. This means that questions of speech processing in a larger sense (for instance from a psycholinguistic point of view) are excluded, although the author is well aware of the fact that production and perception of speech are affected by many diverse aspects of the context, both linguistic and extralinguistic. In a sense the present work can be considered a continuation of Lehtonen's endeavor to elucidate the relations between phonological and physical aspects of quantity in Finnish. Lehtonen found a significant correlation between phonological length and the whole pattern of durations within at least a two syllable foot. The present work extends consideration to other, nondurational aspects.

1.1 Philosophical preliminaries

The author of the present work has been greatly influenced by the philosophical writings of A. N. Whitehead. He has endeavored to the best of his abilities to adhere to a Whiteheadian metaphysics in the course of writing this report. A general discussion of Whitehead's philosophy (called the philosophy of organism by Whitehead himself) is obviously beyond the scope of this work, but a few of the metaphysical assumptions of the present author are perhaps worth mentioning because of their immediate relevance and/or their divergence from views which are widely (though often implicitly) assumed in phonetics.

1.1.1 Direct realism

The view of human perception termed *direct realism* by Fowler (cf. Fowler, 1986, 1983, 1994), as well as Gibson's "ecological" theory of perception on which it is largely based (cf. eg. Gibson, 1966), is by and large quite compatible with Whitehead's philosophy of organism. To begin with, both Fowler and Whitehead reject an absolute Cartesian dualism of mind vs. matter: "... the idea that speech production involves a transition from a mental domain into a physical, non-mental domain such as the vocal tract must be discarded." (Fowler, 1986, p. 10); "The philosophy of organism abolishes the detached mind. Mental activity is one of the modes of feeling belonging to all actual entities in some degree, but only amounting to conscious intellectuality in some actual entities." (Whitehead, 1929, p. 56). It should be obvious from this quote, however, that Whitehead does not propose a reduction of mental phenomena to physical. The point is rather that the mental and physical poles are opposing aspects of each final actuality in varying degree, not elements in separate, incompatible worlds.

Fowler's direct realism paradigm postulates direct perception of relevant events in the environment of the perceiver (called *distal events*) by means of an "informational medium" capable of transmitting structural information about the distal source event. In addition the medium itself is basically transparent to the perceiver. For Fowler, in speech perception the "moving vocal tract" is the distal source which is directly perceived.

The first property of Fowler's informational medium, that it be able to transmit information (or *feeling* in Whitehead's terminology), is a basic characteristic of all entities in Whitehead's cosmology.[3]

Fowler's second requirement, that perceivers are by and large blind to the medium itself, can be interpreted in Whiteheadian terms to mean that each of the events constituting the medium is of a low grade character, experienced as merely conforming to its own past, merely repeating the patterns transmitted through it.[4] Such transmission of physical energy (or *pure physical feeling*) opens the possibility of direct knowledge of the environment. But Whitehead warns:

> There is, however, always this limitation to the security of direct knowledge, based on direct physical feeling, namely, that the creative emergence can import into the physical feelings of the actual world pseudo-determinants which arise from the concepts entertained in that actual world, and not from the physical feelings in that world. (Whitehead, 1929, p. 264)

Fowler also allowed for this possibility of error or "mirage" in speech perception in some of her later articles (eg. Fowler, 1990), and in this way avoided criticism based on cases of speech perception with no actual "moving vocal tract" source (such as synthetic speech).

1.1.2 Invariance

Much discussion has been devoted in phonetics to the issue of variability in speech and the search for invariants (cf. eg. Perkell and Klatt, 1986; see also discussion in Hawkins, 1999a). On a theoretical level, invariance is typically taken as unproblematic, and variability something that needs to be explained. Thus very often it is assumed or expected that phonological categories will correspond to necessary and sufficient properties of

[3] One expression of this is his *principal of relativity*: "That the potentiality for being an element in a real concrescence of many entities into one actuality is the one general metaphysical character attaching to all entities, actual and non-actual; and that every item in its universe is involved in each concrescence. In other words, it belongs to the nature of a 'being' that it is a potential for every 'becoming.' "(Whitehead, 1929, p. 22)

[4] Particularly relevant here is Whitehead's chapter *Organisms and Environment*, for instance the following discussion of transmission through a chain of events A, B, C, and D to an ultimate subject M : "Some of the line, A and C for instance, may stand out with distinctness by reason of some peculiar feat of original supplementation which retains its undimmed importance in subsequent transmission. Other members of the chain may sink into oblivion. For example, in touch there is a reference to the stone in contact with the hand, and a reference to the hand; but in normal, healthy bodily operations the chain of occasions along the arm sinks into the background, almost into complete oblivion. Thus M, which has some analytic consciousness of its datum, is conscious of the feeling in its hand as the hand touches the stone." (Whitehead, 1929, p. 120)

utterances that will be discovered by measurement. On the other hand, there has been very little empirical success in the search for invariance.

From a Whiteheadian point of view, there are two reasons to expect such failure. The first may be called the limitation of perspective. Scientific understanding proceeds by concentrating on some details and omitting others. However, we should not assume that the details omitted are completely irrelevant. On the contrary, "Since all things are connected, any system which omits some things must necessarily suffer from such limitations." (Whitehead, 1938, p. 74). One example of this principle is the expectation that the more speechlike detail is included in stimuli for perception experiments, the more speechlike will be the perception of the stimuli.

The second reason stems from the rejection of absolute determinism. For Whitehead, although every actual occasion must conform to the restrictions imposed by what has happened earlier (its own actual world), there is always some room for novelty to enter.[5] Such flashes of novelty are especially characteristic of living events, and will thus be especially significant for the biological, psychological and social sciences.

The lesson to be drawn for phonetics is that while interesting aspects of speech have often been uncovered during the search for invariance, we should not expect to find invariance, at least in an *absolute* sense. The search for "invariants" can be taken to be a search for measurement which models the perception process, but such measurement should not be considered "more real" than the actual perception itself. There is no reason to expect to find a final, finite set of properties which are both necessary and sufficient for the perception of a certain phonological category. Instead we may hope, as researchers, to illuminate the most *important* characteristics of utterances.

In the past, exploration of intrinsic timing theories has often been guided by the hope of finding invariant timing properties. Despite some initial success in this endeavor, in depth research of speech production within the intrinsic timing model has failed to produce measures with reduced variability or insensitivity to contextual factors (cf. eg. Keller, 1987; Nittrouer *et al.*, 1988). It should be pointed out, however, that there is nothing inherent in the concept of intrinsic timing which would require the notion of invariant timing (for general discussion of invariance in relation to Direct Realism and other recent theories, see Hawkins, 1999b).

[5] In philosophical terms, invariance corresponds to *essential qualities* (*universals*) and variability to *accidental qualities* inhering in *particular substances*. Whitehead rejects the idea, which he calls the *subjectivist principle*, "that the datum in the act of experience can be adequately analysed purely in terms of universals." (Whitehead, 1929, p. 157). This is to say that every actual occasion retains some degree of individuality over and above its "exemplification" of any categories. "It is a complete mistake to ask how concrete particular fact can be built up out of universals. The answer is, 'In no way.' The true philosophical question is, How can concrete fact exhibit entities abstract from itself and yet participated in by its own nature? . . . Each fact is more than its forms, and each form 'participates' throughout the world of facts." (Whitehead, 1929, p. 20)

1.1.3 Physical time

Time, however that term is to be interpreted, is a basic notion for phonetics. The objects of phonetic science, utterances, are patterns in time whose substantiations lack endurance. This is one of the basic differences between spoken language and written language, which uses enduring objects to convey meaning. However, this difference leads easily to much confusion in thinking about the nature of phonetic objects. On the one hand we tend to think of phonetic objects as concrete events because of the fleeting nature of their substantiations. On the other hand we tend to "spatialize time" as Bergson put it,[6] and treat speech as though time was merely one more spatial dimension, to be measured with a millisecond ruler, as indeed it appears to be in a spectrographic representation.

The concept of time which has come to dominate western thought in the past few centuries is one which, as Whitehead put it, "enables us to abstract from change and to conceive of the full reality of nature at an instant, in abstraction from any temporal duration and characterized as to its interrelations solely by the instantaneous distribution of matter in space." (Whitehead, 1938, p. 145). It can perhaps be reasonably called the Newtonian view of time, though of course Newton was not its sole architect. In this view process is not essential but accidental. Instead matter in space at an instant is basic and time merely the steady succession of instantaneous facts. In Newtonian physics absolute time is completely independent of actual events which may be happening.[7] Of course this view has had its critics, one of the earliest being Newton's contemporary Leibniz, who regarded the possibility of "empty time" (during which nothing happens) as one of the inherent contradictions of the theory (cf. Alexander, 1956). If Newtonian time is taken not as a descriptive model, but as absolute matter of fact, then here Whitehead must agree with Leibniz, because for Whitehead, events (process) comprise the most concrete layer of reality, whereas (physical) time is a derivative or more abstract notion. Thus, there can be no "absolute time" in a concrete sense, rather physical time is an abstraction of the becoming of events: the temporality of events without considering the events themselves.[8]

In themselves abstractions are not evil, indeed thinking, understanding, and even recognition all necessarily involve abstraction. There is, however, the danger that a

[6] "... in a word, we project time into space, we express duration in terms of extensity, and succession thus takes the form of a continuous line or a chain, the parts of which touch without penetrating one another." (Bergson, 1910, p. 101)

[7] "Absolute, true, and mechanical time, of itself, and from its own nature, flows equably without relation to anything external" (Newton, 1974, p. 6)

[8] "This genetic passage from phase to phase is not in physical time: the exactly converse point of view expresses the relationship of concrescence [ie. 'becoming concrete' or 'happening'] to physical time. It can be put shortly by saying, that physical time expresses some features of the growth, but *not* the growth of the features." (Whitehead, 1929, p. 283). Interestingly, Newton himself characterized his absolute time as abstraction from experienced time: "But in philosophical disquisitions we ought to abstract from our senses, and consider things in themselves, distinct from what are only sensible measures of them" (Newton, 1974, p. 8)

particular abstraction may come to be mistaken for concrete fact and thus regarded as infallible (this is what Whitehead has called the *Fallacy of Misplaced Concreteness*, cf. eg. Whitehead, 1925, p. 51 ff.).

In phonetics, the Newtonian view has two consequences. First, it predisposes us to think of speech as consisting of mere sequences of instants, each of which exhibits a total pattern (eg. a vocal tract configuration, and the corresponding sound quality) all at once. Then timing information is reduced to the temporal "location" of instantaneous segment boundaries, or equivalently to the physical time elapsed between selected recognizable instants. Second, it easily leads to neglect of the question as to how speakers and listeners gain knowledge of the amount of physical time elapsed, since absolute time is implicitly assumed to be universally available.

It would seem to be almost universally accepted by phoneticians, at least in theory, that timing (eg. quantity) is *not* absolute. For instance Lehtonen (1970, p. 50) was explicit in denying more than practical status to segmentation procedures in his research on Finnish quantity:

> Details of segmentation, such as placing the burst of a plosive and the following noise with the duration of the plosive or with the duration of the following vowel, are only of academic interest in this sense. The purpose of this sort of "microscopic" definition of segment boundaries is not to fix any functional boundaries at certain points on curves, but to find some fixed points which appear as regularly as possible to be used systematically in phonetic measurements.

However, in the practice of phonetics research, it is often implicitly assumed that the segment boundaries the researcher finds and lapses of time the researcher measures between them are precisely what speakers produce and listeners perceive. Even when this assumption is challenged, the assumption is still generally retained that there is *some* instantaneous moment or "onset" attended to by speaker and listener in their production and perception of rhythm (cf. the discussion of the so-called p-center phenomenon on page 33).

The bias toward segmental duration can be observed throughout the literature, for instance in the title of Klatt's famous article, *Linguistic uses of segmental duration in English: Acoustic and perceptual evidence* (Klatt, 1976). It is generally taken for granted that what is produced, perceived, used linguistically, is duration, ie. the amount of physical time elapsed between segment boundaries. Indeed, it would seem that few investigators have contemplated any other alternative.

Given a reduction of timing to segments and boundaries, variation in durational data will be interpreted to mean that segments exhibit "elasticity". It is obvious that there is a great amount of variability in the durations of segments measured in absolute time—what amounts to noise in the transmission if durations are indeed intended to be conveyed from speaker to listener. Accordingly it is assumed that this can only be due to the defective nature (or imprecision) of the mechanism producing speech, since time itself is unconsciously taken to be an infallible concrete fact. However, it is not the function of speech to conform to the physicist's atomic clock or even to the phonetician's millisecond ruler.

To sum up, the point of view taken here is that physical time, measured in milliseconds, is an abstraction, not concrete matter of fact. In particular, "instants" or "points in time" are abstractions—all actual entities "take time" to "become". Also, the flow of time is *not* independent of particular events. Thus instantaneous segment boundaries and absolute durations are abstract and the usefulness of these abstractions in particular contexts is essentially an empirical question, not a foregone conclusion. In any case any "lapses of time" that may be relevant in speech must exhibit a psychological or mental aspect. In speech it can only be a feeling of elapsed time which may be relevant in production and the transmission of this feeling to the listener (of course this does not imply that such a feeling must be *conscious*). It is not at all clear that such feelings must exhibit exactly the same metric properties that are generally associated with physical time.

1.2 Extrinsic and intrinsic timing

The majority of timing models in phonetics research can be described as versions of a sort of "rubber band model": The underlying pattern of an utterance may be pictured as inscribed along a strip with various elastic properties, as if a spectrogram were reproduced on a sheet of rubber instead of on paper. Differences in "articulation rate" are then seen as stretching (or possibly shrinking) the strip to various lengths, but the sequence of points along the strip is not disturbed. In this metaphor, (absolute) time is represented by distance along the strip, actual speech events are represented as a series of points along the strip, and the timing variability of these events is represented by the elasticity of the strip.

Various subtypes of the rubber band model may be distinguished. Reliance exclusively on segmental durations or segment-sized durations implies that "elasticity" changes abruptly at segment boundaries and is uniform between segment boundaries, or at least that variable elasticity within segments is unimportant. Indeed, in such a model, there are only a relatively few points (segment boundaries) along the strip that are important. Thus as long as these points are at their proper "positions" along the strip, then any variation in between is irrelevant. Different segments, or segment types, may differ in elasticity as well as "natural length" or duration. For instance, it has long been accepted as fact that in general vowels are "more elastic" than consonants (cf. eg. Jespersen, 1926), and that long consonants and vowels are more elastic than short consonants and vowels (Sievers, 1893).

Kozhevnikov and Chistovich in their classic study (Kozhevnikov and Chistovich, 1965) placed special emphasis on the role of the syllable in their timing model. They interpreted their findings as showing that while individual segments vary considerably in elasticity, syllable durations vary proportionately to each other or to an assumed articulation rate. In terms of the rubber band, this means that syllables have constant overall elasticity.

In automatic speech recognition schemes time alignment to a stored reference pattern generally uses an algorithm which does not attempt to locate segment boundaries at all, but rather considers equally all points in time (up to some degree of accuracy—actually these successive points are represented by relatively short, slightly overlapping segments of the signal). Such an algorithm is generally referred to as Dynamic Time Warping, and its use implies a less restricted rubber band model in which "elasticity" is allowed to vary continuously along the strip. Chapter Three is devoted to the theory of Dynamic Time Warping and its relation to a generalized rubber band model without the assumption of segment boundaries.

How is a rubber band model to account for realization in production and perception? In as much as timing differences can be willfully produced by the speaker and recognized by the listener, these differences must involve contrasts (comparison) of (two or more) simultaneous events. Since in the rubber band model the speech events themselves occur only sequentially and timing variation does not alter that sequence, the rubber band or duration based model must rely on another sequence of events external to the utterance itself. This second event chain may be thought of as a biological clock to which the flow of speech can be compared in order to maintain proper timing. Actually two independent biological clocks are required—one for the speaker to produce correct timing, and one for the listener to interpret the timing so produced. It is no accident that the model which evolves from our thinking in terms of absolute Newtonian time ends up requiring a (biological) clock which mirrors (although imperfectly) that concept of time. Kozhevnikov and Chistovich recognized the need for such a biological clock—they called it a rhythm generator (генератор ритма). Fowler has used the term *extrinsic timing* to characterize these theories, in opposition to theories of *intrinsic timing*.

The terms intrinsic and extrinsic timing were introduced by Fowler (1977, 1980) to describe theories of speech timing. Briefly, Fowler calls a timing theory extrinsic if dynamic properties of speech are excluded from phonetic representations (ie. segments are static, "timeless" entities) and observed timing differences must therefore be externally imposed on sequences of segments. An intrinsic timing theory, on the other hand, assumes that dynamic properties are included in phonetic representations. Fowler suggests that phonetic "representations" may be coordinative structures, or self-executing articulatory plans. Kelso and his coworkers have also supported the idea of intrinsic timing (cf. eg. Kelso, 1995). For Kelso and his coworkers, intrinsic timing means that temporal stability in speech is "achieved without reference to an extrinsic clocking device, but rather in terms of the dynamic topology of the system's behavior." (Kelso and Tuller, 1987, p. 203).

Whereas extrinsic timing theories postulate that durations of speech gestures are directly controlled, an intrinsic timing theory such as Fowler's postulates that durations of individual gestures are not controlled by the speaker and are not significant to the listener, therefore no clock or rhythm generator is necessary. Instead, allowing that speech gestures may "overlap in time" rather than being strictly sequential, Fowler suggests that the speaker can directly control the amount of coarticulation or coupling between gestures, that is, when a gesture begins relative to other gestures. That is, the phase relations (to use Kelso's terminology) between (overlapping) gestures can be controlled. It is assumed that listeners attend to these events indicated by the speech signal and their phase relations, rather than to duration, i.e. time elapsed between instantaneous "landmarks" in the signal. The appeal of such an approach, if it can be shown to fit with empirical data, is obvious. The speaker's task is considerably simplified if he synchronizes gestures by means of internal relations between the gestures themselves rather than forcing them to synchronize with an external rhythm regardless of the dynamic properties of the gestures. The listener's task is also easier and will give more reliable results if the timing contrasts in question are in the speech signal itself, rather than requiring that the listener compare with a "clock" of his own, which may be only poorly synchronized with the speaker's "clock".

Phase portrait analysis

In trying to illuminate "the dynamic topology of the system's behavior," or the natural rhythms of various articulations in speech, Kelso and his coworkers have borrowed the concept of *phase portrait* from dynamic systems theory and applied it to timing in phonetics. A complete phase portrait for a (complex) dynamic system is a model including the 'phase space' consisting of all possible states the system may be in, as well as a vector for each state indicating transition to other states. Time itself is not a state the system can be in, therefore, strictly speaking time as such is absent from such phase portraits. The dynamics of mechanical systems can be succinctly described, at least qualitatively, by the topology of the phase portrait. Often there are certain states (*stable equilibrium points*) or chains of states (*orbits* in the phase space) which the system tends to settle in regardless of what state it starts out in. In the terminology of dynamic systems theory these sets of states are called *attractors* or *limit sets*. A common attractor which can be found in many mechanical systems is the so-called limit cycle: a *periodic orbit* or cyclic chain of events to which the system tends, often very rapidly.

Kelso and his coworkers have attempted to characterize the (potentially) cyclical movements in speech production, for instance opening and closing movements of the jaw, in terms of a limit cycle in an empirical phase portrait analysis. Of course such an empirical analysis must use some set of *a priori* measurements of the movements involved, and thus it necessarily represents a *partial* phase portrait (indeed, if we take seriously the rejection of a deterministic universe, only an ideal system could ever be characterized completely by a finite number of dimensions). One can only hope that the dimensions chosen will be "important" ones for the understanding of the system under consideration. Kelso and Tuller used articulator displacement and velocity, both normalized over a single cycle, presumably to emphasize the topological properties of the system (cf. Kelso and Tuller, 1987).

Points in such a phase portrait can be taken to represent not absolute time but rather a sort of relative time or “phase” of a speech gesture as it unfolds. Unlike absolute time, phase is dependent on the particular system (gestures) involved, and includes within itself a reference to past and future. Thus the phase portrait characterization of Kelso and his coworkers can be seen as an attempt to get away from scientific time as an absolute number line underlying the universe. In the words of Kelso *et al.*, “a given dynamical system—at whatever stage of its development—generates its own intrinsic time” and “Time is indeed with us, only not in the way it is traditionally defined, ie. as conventional or mechanical time (in seconds, hours, etc.) imposed on a system *regardless* [their emphasis] of its particular dynamics.” (Kelso *et al.*, 1986b, p. 192).

1.3 Intrinsic timing and quantity languages

It would seem that a major obstacle to acceptance of intrinsic timing models is the existence of so-called quantity languages. These are languages which have been interpreted as including phonological oppositions conveyed entirely by differences in the duration of various segments. Thus in contrast to other languages, in which durational differences are presumed to be “accidental” consequences of other factors or irrelevant (ie. not distinctive), in a quantity language such timing changes (corresponding to different quantity patterns) are presumably *intentional* (distinctive). Under the intrinsic timing model, what can it mean to say that two words differ only in regard to the duration of segments (eg. short vs. long vowels)? Lubker raised this point in his reply to Kelso *et al.*, 1986a:

> There are too many languages for which timing is a very crucial aspect. In Swedish, for example, duration information can be critical in differentiating between otherwise identical words. I find it difficult to accept that such critical aspects of production are not carefully controlled by the speaker … (Lubker, 1986, p. 137)

If this is true for Swedish, it should be even the more so for Finnish, the language under consideration in this research, since the phonological contexts allowing the opposition of long and short are much less restricted in Finnish.

1.3.1 Quantity in Finnish

A good overall view of the Finnish quantity system is provided in Lehtonen 1970. We give here a brief summary of the basic phonological system and results of previous research.

Long vs. short opposition

The following formula sums up the basic opportunities for the long vs. short opposition to occur in a Finnish word:[9]

$$(\mathrm{C})\begin{Bmatrix}\mathrm{V}\\ \mathrm{VV}\end{Bmatrix}\left(\left\{\begin{matrix}.\\ .\mathrm{C}\\ \mathrm{C.C}\\ \mathrm{CC.C}\end{matrix}\right\}\begin{Bmatrix}\mathrm{V}\\ \mathrm{VV}\end{Bmatrix}\right)^{*}(\mathrm{C})$$

where "VV" indicates a diphthong or a long vowel, " . " indicates a syllable boundary, "C.C" indicates a consonant cluster or geminate (long) consonant, and "( )*" indicates zero or more repetitions. There are further restrictions on which vowels can come together to form a diphthong, and on possible consonant clusters. For instance, a long consonant (geminate) always includes a syllable boundary (ie. closes the previous syllable). This excludes cases such as /kks/. Also, in the case of three consecutive consonants (ie. two syllable final consonants), the first is normally a sonorant (ie. /l r m n ŋ/), the second can normally only be an obstruent (ie. /p t k s/). This allows cases such as /lst/ and /mpp/ but not eg. /rnn/ or /skk/. In addition to these restrictions, it is worth mentioning that the consonant /ŋ/ does not occur short between vowels, and the consonants /h j v d/ do not occur as geminates (at least in the standard language). Therefore these consonants are exempt from the phonemic long vs. short opposition. Basically then, taking into account the above restrictions, the quantity system is very simple from a structural point of view: there is one possible vowel quantity opposition for every syllable, and one possible consonant quantity opposition between every pair of adjacent syllables, with any combination of short and long allowed regardless of other factors such as stress. It is thus quite possible, for instance, to have words with sequences of long segments, as in /(ei) rääkkääkkään/ '(doesn't) maltreat after all'.

Factors affecting measured durations of long and short segments

The obvious result that long segments are longer in duration than short segments in corresponding positions has been confirmed, at least for carefully elicited speech, in several extensive studies which have measured durations of sounds in several dialects of Finnish[10] (eg. Donner, 1912; Laurosela, 1922; Sovijärvi, 1944; Palomaa, 1946; Lehtonen, 1970). In addition these studies have uncovered numerous other factors which have a systematic effect on segmental durations.

[9] Less common or marginal structures include word initial and final consonant clusters, and word internal consonant clusters with more than three consonants. Also the structure ...VVCCC... allowed in the formula exists but is relatively rare.

[10] The Salmi dialect investigated by Donner belongs to group which is often considered to constitute a language distinct from Finnish called Carelian. In any case the two are very closely related and for the most part mutually intelligible.

Number of syllables. All relevant studies of Finnish have shown a strong tendency for durations of component sounds to decrease as more syllables (or segments) are included in an utterance (Donner, 1912; Laurosela, 1922; Sovijärvi, 1944; Palomaa, 1946; Lehtonen, 1970, 1974; Iivonen, 1974a,b; Marjomaa, 1982). Indeed some tendency of this sort would appear to be almost universal in the world's languages (cf. eg. Lehiste, 1970). Sievers called attention to this phenomenon more than a century ago, using the term *rhythmische Abstufung* (Sievers, 1893).

Related to this is the observation that segment, syllable and word durations are longer when words are pronounced by themselves as compared to words in a sentence or phrase (Donner, 1912; Palomaa, 1946).

Sentence-stress. Sentence-stress, or utterance-level prominence, has been reported to have a (small) lengthening effect on segments (Laurosela, 1922; Suomi *et al.*, 2003).

Speech tempo. Although earlier reports have not explicitly examined the effects of tempo on measured durations, in some cases the published durations themselves indicate the strong effect that tempo has evidently had. For instance the recordings used by Palomaa (1946) were evidently spoken at a very slow tempo, judging from the extremely long durations he obtained throughout. Marjomaa (1982) measured durations of vowels in Finnish and English in one and two syllable test words imbedded in a carrier sentence spoken at three rates, normal, fast and slow. In all cases long vowels at the fast rate were considerably shorter in mean duration than the corresponding short vowels at the slow rate, and even very close in value to the short vowels at the normal rate (cf. Fig. 2, p. 126 and Table 5, p. 130).

Inherent durations of segments. Closer vowels in Finnish have generally been reported to be shorter than more open vowels (Donner, 1912; Laurosela, 1922; Sovijärvi, 1944; Palomaa, 1946; Wiik, 1965; Lehtonen, 1970),[11] reflecting a well known universal tendency in languages of the world (cf. eg. Lehiste, 1970).

As to consonants there are various partly conflicting reports of durational differences, but a fairly large and robust effect reported is that voiceless consonants (obstruents) are longer than voiced (sonorant) consonants (Laurosela, 1922; Sovijärvi, 1944; Palomaa, 1946; Lehtonen, 1970). This again would also seem to reflect a universal tendency (cf. eg. Lehiste, 1970). Lehtonen (1970, p. 97) also reported that these two groups of consonants differ in the ratio of long consonant duration to short consonant duration: about 1 : 2 for the obstruents and about 1 : 2½ for the sonorants.

Compensatory effects. On the other hand, these differences in consonant durations have been found to correlate with opposite differences in neighboring vowels, es-

[11] This general pattern may be slightly disturbed for rounded vowels; Laurosela (1922) reported shorter durations while Lehtonen (1970) reported longer durations for rounded vowels in the vicinity of labial consonants.

pecially preceding vowels. That is, vowels are generally longer before the voiced (sonorant) consonants (Laurosela, 1922; Palomaa, 1946; Wiik, 1965; Lehtonen, 1970). According to Lehtonen (1970, p. 79), "the longer the intrinsic duration of a consonant, the more it shortens the preceding vowel, and vice versa." Lehtonen (1970, p. 81) also reported that shorter consonants (especially /r/ and /d/) lengthened following consonants as well.[12]

Quantity patterns. If the Finnish quantity system is very simple from a structural point of view, it is not so straightforward from the point of view of measured durations. As Lehtonen (1970) has explicitly pointed out, quantity related effects on segmental duration are not limited to whether the segment itself is phonologically long or short, but include the whole quantity pattern within at least a two syllable unit. Vihanta (1987, p. 108) has succinctly summed up Lehtonen's major findings as follows:

1. A short vowel in the second syllable is relatively long when the first syllable is short. (Sovijärvi, 1944; Lehtonen, 1970).

2. A long consonant is shorter after a long vowel (Laurosela, 1922; Lehtonen, 1970). Laurosela (1922) also reported that a long vowel or diphthong shortens a following word final consonant, Lehtonen (1970, p. 100) also reported that a long vowel shortened a following consonant cluster, especially the first component.

3. A consonant (long or short) is longer immediately preceding a long vowel Lehtonen (1970). Donner (1912) noted this effect for word initial consonants, Laurosela (1922) noted the effect was largest for intervocalic consonants, especially after a short vowel.[13]

4. A consonant is shorter before a long consonant than it is before a short consonant (Laurosela, 1922; Lehtonen, 1970).

Note that all of the above tendencies can be regarded as types of compensatory effects, with the exception of number 3, which might well be called "anticompensation".

Regional variation

Among other features which distinguish various dialects of Finnish are differences in the quantity system and its realization. The most important of these can be regarded

[12] An exception to the general rule is the finding (Lehtonen, 1970) that vowels are longer in the vicinity of /s/, which is itself quite long.

[13] Palomaa (1946, p. 293) seems to state that the contrary trend holds, based on the principle that all segmental durations diminish the more phonetic material is added, but the majority of his actual measurements support the conclusion of lengthened consonants before long vowels, as he himself states on page 200.

as resulting from varying degrees of the tendencies noted above. The so-called half-long vowel distinction (Wiik, 1975, 1985), has to do with the duration of a short vowel in the second syllable of a foot following a short first syllable (compare tendency number 1 above), often expressed as an average ratio of second vowel duration divided by first vowel duration. On the basis of extensive measurements of fairly spontaneous speech in many dialects, Wiik (1975, 1985) distinguished five major areas, which he called the South-West area, with overall ratio 174%, the Häme area, with ratio 106%, the Savo area, ratio 155%, the South-East area, ratio 130%, and the North-East area, with ratio 138%.

The so-called primary gemination, an expansive feature of a great many dialects, refers to the "exaggeration" of the tendency number 3 above (present to some degree in all dialects) for a consonant to be longer when preceding a long vowel. As noted, this tendency is strongest after a short first syllable (ie. in foot structure CVCVV), and the term primary gemination means that the intervocalic consonant in this position is lengthened to the point that it is no longer in opposition with an original phonologically long consonant.

The so-called secondary gemination (of which two subtypes are generally distinguished) refers to the generalization of this feature (of consonant gemination before a long vowel) to other environments, including after long first syllable and in other syllables (cf. eg. Palander, 1987).

1.4 Statistical methods

1.4.1 Markov Chain Monte Carlo

In the past decade or so the use of so-called Markov Chain Monte Carlo (MCMC) techniques, used for many years in statistical physics, has been dramatically increasing in popularity for statistical modeling in general, and especially for Bayesian inference. No doubt this rise in popularity is partially due to increased computational power and the availability of software tools such as the text based BUGS program (Spiegelhalter *et al.*, 1994) and its successor WinBUGS using a graphical interface (Spiegelhalter *et al.*, 2003). An excellent introduction to the subject is Gilks *et al.* 1996b as well as Carlin and Louis 2000. Because of its great flexibility the technique can be used in analyzing a wide range of complex statistical models, opening the possibility of using more realistic models rather than being forced to use oversimplified classic models with test statistics whose use may not be justified. Also MCMC methods provide a unifying framework for analysis of diverse complex situations. Because extensive use is made of MCMC based Bayesian inference in the present work, and because it differs considerably from classic methods which may be more familiar to readers, a very brief exposition follows, based primarily on Gilks *et al.* 1996a.

In order to make statistical inferences it is necessary to integrate over probability distributions. In simple cases there exist closed form mathematical expressions for the required integrations. However, more complex situations easily arise, for instance in Bayesian inference where it is necessary to integrate over the (high dimensional) posterior distribution of model parameters given the data. In these cases the analytic evaluation is typically impossible, a fact which has traditionally limited the usefulness of Bayesian methods in applied statistics. The computational technique of MCMC is essentially a way to approximate the required integrations to any required degree of accuracy. *Monte Carlo* refers to integration by drawing (a large number of) samples from the required distribution and calculating sample averages to approximate expectations. *Markov chain* Monte Carlo refers to a clever computationally efficient way of generating the samples needed using Markov chains. The alternative in classical statistical inference is to choose a simplified, analytically tractable model in which relevant parameters can be calculated exactly based on the data. The problem is that the validity of such "exactly calculated approximations" is often questionable and can be improved (in the case of asymptotic parameters) only by increasing the amount of data on which inferences are based. In MCMC the situation is reversed: instead of exactly calculating approximate parameters, exact parameters are approximated in calculation. To increase accuracy one simply draws more samples.

In addition, Bayesian inference itself represents quite a different perspective from classical statistical hypothesis testing. Basically, in the classical approach, a null hypothesis or model is set up and it is desired to find the probability of the observed data (or more often of a test statistic summarizing the data), given the hypothesis. In the Bayesian approach, there is no fundamental distinction between observed data and parameters of the model. Both are considered to be random quantities, some of which we are less certain about (parameters and missing data), others of which we are more certain about (observed data). The goal of Bayesian inference is not to calculate the probability of the *known* data given the model and the parameter values of the null hypothesis, but rather the joint probability distribution of the *unknown* parameters and data, given the model and known data. This is known as the *posterior* distribution, in opposition to the *prior* distribution(s) characterizing the uncertainty associated with the various unknowns in the model.

1.4.2 Signal Detection Theory

So-called Signal Detection Theory, or ROC analysis, developed extensively since the 1960's (Green and Swets, 1966; Egan, 1975), deals with the statistical theory of making a decision between alternatives in the face of uncertainty, whether due to insufficient information from the past or the inherent indeterminacy of the future. For instance putting duration measurements of a spoken utterance into a statistical frame embodies simultaneously two uncertainties: we are uncertain as to the speaker's original intention(s) (limitation of perspective on the past) and at the same time uncertain as to a potential listener's interpretation (indeterminacy of the future).

The backbone of the theory is the so-called receiver operating characteristic (ROC) curve first utilized to assess how well radar equipment distinguished real signals from random noise. The ROC curve plots *true positive* (or “hit”) probabilities against *false positive* (or “false alarm”) probabilities, providing a characterization of the possibility of discriminating alternatives based on available information regardless of the particular decision threshold chosen, which will in any case vary depending on the relative costs of making mistakes in one direction or the other.

Originally a summary statistic for the ROC curve called d' (*discriminability* or simply *d-prime*) was proposed which assumed that the underlying statistical distributions of information associated with each alternative were two normal distributions with identical variances but (possibly) differing means, allowing the detection to be performed with better than chance probability. The definition of d' is simply the distance in standard deviations between the two means, which can in principle be estimated with the help of empirical ROC curves.

Area Under the Receiver Operating Characteristic

Because the assumption of underlying normal distributions, and especially the assumption of identical variances, has turned out to be too restrictive in practice, a less restrictive summary of the ROC, the area under the ROC (also called *accuracy*, and variously denoted AUROC, θ, A, or A_z, especially when based on two normal distributions), is now used routinely in place of d' (Hanley and McNeil, 1982). The AUROC, which ranges from 0.5 for chance discrimination to a theoretical maximum of 1 for certain discrimination, given the available information, also has an intuitive interpretation as the probability that a pair of cases to be discriminated, one from each of the two alternatives, is in the right order for some threshold to distinguish them correctly. For instance, if we consider the discriminability of long and short segments in Finnish on the basis of measured durations, the AUROC would be interpretable as the probability that a random long segment is in fact longer in duration than a random short segment. The AUROC statistic is used extensively in the following chapters for the purpose of quantitatively summarizing the difference between probability distributions.

If durational patterns are indeed controlled directly (at least for so-called quantity languages), if there is a rhythm generator of some sort, we would expect that rhythm is to a large degree independent of other features of speech production and perception. Chapter Two is devoted to this question. Chapter Three reviews the theory of Dynamic Time Warping and its relation to a “continuous rubber band model”. It also develops the tools used for the perception experiments of Chapter Four, which address the question of the effect of nontiming differences such as vowel quality and pitch contour on quantity perception.

Chapter 2

Independence of rhythm and segments

> El reloj que en el campo se tendió sobre el musgo
> y golpeó una cadera con su eléctrica forma
> corre desvencijado y herido bajo el agua temible
> que ondula palpitando de corrientes centrales.
>
> —Pablo Neruda

Kozhevnikov and Chistovich (1965) take the position that there is an articulatory program that is executed. This program includes, among other instructions or commands to the motor system, instructions for regulating the rhythm of an utterance (which for them is operationalized as the relative durations of speech segments). They also assume that rate of speaking is independent of this articulatory program, and therefore relative durations which remain constant (or vary randomly) over changes in rate can be ascribed to the articulatory program, while relative durations which vary systematically with speaking rate are due to other factors. With this assumption Kozhevnikov and Chistovich conclude that syllables are the smallest units whose rhythm is controlled by the program. In their Russian language data relative syllable durations were independent of speaking rate, but the relative duration of units smaller than the syllable (consonants and vowels) varied systematically with rate, consonants receiving a smaller proportion of total duration as speaking rate slowed.

Furthermore, they found that when speakers were instructed to read test words "with a closed mouth and with the tongue pressed against the palate," words belonging to the same rhythmic category (ie. words with the same number of syllables and stress on the same syllable) had equal average durations regardless of segmental differences. The segmental differences included number of consonants (e.g. пасу [paˈsu] vs. посту [paˈstu]) and type of consonant (sometimes leading to different syllabification, e.g. посту [paˈstu] vs. коньку [kanʲˈku]). Although in natural speech these differences caused differences in average duration, the closed mouth pronunciations did not differ in duration of the total word. In addition all closed mouth pronunciations consisted of a voicing–silence–voicing sequence (even when the target word had a voiced consonant between the vowels, e.g. коню [kaˈn^{j}u]) whose parts had equal average durations. On the other hand test words with different stress patterns (e.g. пасту [ˈpastu] vs. посту [paˈstu]) exhibited differences in duration of the component parts (stressed "syllable" was longer) in closed mouth pronunciation. From this Kozhevnikov and Chistovich conclude that "the rhythmic figure of a word exists as a separate independent part of the articulatory program." (Kozhevnikov and Chistovich, 1965, p. 115). In order to assess the validity of these arguments, a replication of the "closed mouth" experiment of Kozhevnikov and Chistovich was carried out for Finnish.

2.1 Experiment 1: Closed mouth production

2.1.1 Methods

Twelve two-syllable test words were selected, all of which were real Finnish words starting with [tu] and ending with [li] or [ki]:

tuli	'fire'	*tuki*	'support'	*tulkki*	'interpreter'
tulli	'customs'	*tukki*	'log'	*tunkki*	'jack'
tuuli	'wind'	*tuuki*	'tablecloth'	*turkki*	'fur'
tuoli	'chair'	*tuiki*	'extremely'	*turski*	'gruff'

A list was compiled with ten groups of 14 words, such that each group contained all 12 test words in random order, with an additional test word (chosen at random from the same words) at the beginning and end to avoid the effects of list intonation. These additional words were not considered in the following analyses. Two native speakers of Finnish, both female, were instructed to read the entire list of 140 test words first normally, and then a second time with closed mouth. For the closed mouth condition, subjects were instructed to keep the mouth entirely closed with the tongue against the roof of the mouth. Both readings were recorded on tape using a throat contact microphone. The signals were later low-pass filtered at 400 Hz and digitized using 8-bit digitization at a sampling rate of 5013 Hz for analysis by computer. Automatic segmentation of the signals into silence vs. non-silence was carried out by computer, considering silence to be a stretch of at least 30 msec during which signal amplitude remains less than 5% of peak signal amplitude. Each segmentation was displayed on screen and interactively corrected by the researcher in the case of obvious errors (due to extraneous noise in the middle of otherwise silent periods). In this way measurements were collected for each test word token including total word duration not including initial [t] (ie. from end of pre-word silence to beginning of post-word silence), and, for those tokens with a word-internal silent portion (corresponding to intervocalic

test word	normal reading	closed mouth
tuli	0	4
tulli	0	6
tuuli	0	0
tuoli	0	0
tuki	20	1
tukki	20	19
tuuki	20	0
tuiki	20	1
tulkki	20	19
tunkki	20	17
turkki	20	20
turski	20	17

Table 2.1: Number of productions with word internal silence

consonant or consonant cluster), also the beginning and end of that silent portion measured from the beginning of the word. Measurements were rounded to the nearest millisecond.

Analysis of variance was used to analyze the significance of measurements. Since each test word occurred exactly once in each of the ten blocks of the word list, a Randomized Complete Block Design with fixed effects was used for the analyses, with interaction of blocks and treatments assumed zero. Since there were small but significant interactions between subject and the other independent variables, and especially between task (normal vs. closed mouth) and the other variables, analysis of variance was carried out for each subject and each task separately.

2.1.2 Results and Discussion

One of the striking differences in the present data compared with the results of Kozhevnikov and Chistovich was the non-occurrence of word-internal silence for many tokens. Kozhevnikov and Chistovich reported that in closed mouth speaking, for "all two-syllable words there was a required cessation of phonation in the middle of the word, although in natural speech this does not occur (for example, "konyu")." (Kozhevnikov and Chistovich, 1965, p. 112). In the present study, however, many of the words spoken with closed mouth had voicing throughout with no "intervocalic" silence. The tokens that did show a cessation of phonation in the middle, moreover, were by no means distributed randomly, as can be seen in Table 2.1 showing occurrences of word internal silence (pause) by word type.

There was a strong tendency for voicing to persist throughout the closed mouth versions, except for the words with [kk] or [sk]. This difference may indicate that the task performed was not the same as the task carried out by the subjects of Kozhevnikov and Chistovich, in spite of attempts to replicate their conditions. The difference may be due to a difference between Finnish and Russian, or simply to individual differences between speakers. An explanation is possible in terms of the normal production of Finnish stops. In his contrastive study of Finnish and English stops, Suomi suggests that in Finnish single (short) intervocalic stops are voiceless due completely to aerodynamic reasons, with no adjustment of the vocal folds (Suomi, 1980). In other words, there is no active change in the vocal folds from vowel through stop to following vowel, but the cease in air flow due to the occlusion in the supraglottal vocal tract leads to a cessation of vocal fold vibration. In this conclusion, Suomi relies on a study of vocal fold behavior during consonants carried out by Iivonen (1975). Iivonen's study showed no abduction of the vocal folds for single intervocalic stops. If we assume that the subjects in the present experiment have activated their vocal folds normally, even in the closed mouth condition, then we could expect voicing to continue throughout the words with single stop [k], because in the closed mouth condition there is a continual flow of air (through the nasal cavity). Iivonen's study also showed a clear abduction of the vocal folds for geminate (long) stops, eg. [pukkiɑ], consonant clusters consisting of two stops, eg. [putki], as well as combinations of [s] plus stop, eg. [puski]. This would explain the observation in the present experiment of devoicing in the words containing [kk] or [sk], provided that this "normal" behavior was used by the subjects even in the closed mouth condition. Why there should be abduction for long stops but not for short stops is not clear. It does however suggest that the glottal mechanism for producing stops in Finnish may be different for short and long stops. Interestingly, in a study of geminate consonants in Estonian, Lehiste *et al.* found evidence of rearticulation for both long (geminate) and overlong consonants (Lehiste *et al.*, 1973).

In what follows the word internal measurements are taken into account only for those word types which systematically exhibited word internal pause. Mean durations for each subject in each of the two conditions are shown graphically in Figure 2.1 a–d.

Total durations

The total durations measured for the test words for both subjects in both tasks (normal reading and closed mouth reading) are shown in Table 2.2 (see also Figure 2.1 a–d). It may be noted first of all that the measurements for the normal reading productions varied considerably according to word type. Not only were there differences for words with identical segmental content but differing quantity (eg. *tuli* vs. *tulli*), but also for words which differed in segmental content only (eg. *tuli* vs. *tuki*). This is not surprising, and is parallel to the results of Kozhevnikov and Chistovich, and indeed to results of many other studies of various languages. What is interesting to note, is which differences show up in the closed mouth productions as well.

In the data of Kozhevnikov and Chistovich, all durational differences disappeared in the closed mouth reading task, with the exception of those due to differing stress patterns. In the present data, it is clear that some differences disappear in the closed

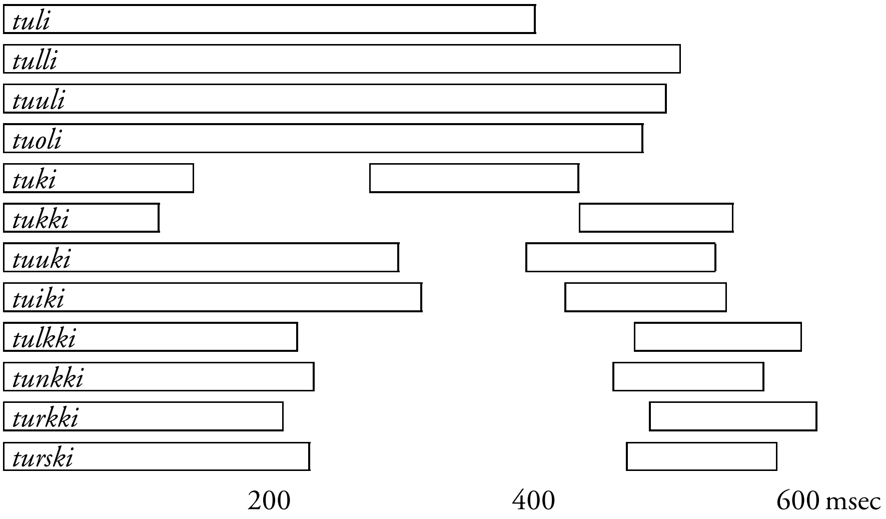


Figure 2.1: Mean total durations for normal and closed mouth readings: (a) Subject SU, normal reading

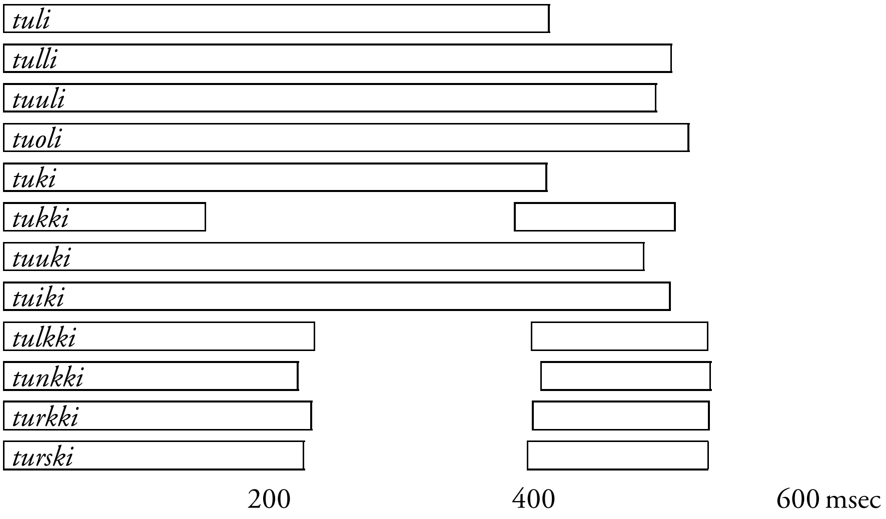


Figure 2.1: (b) Subject SU, closed mouth

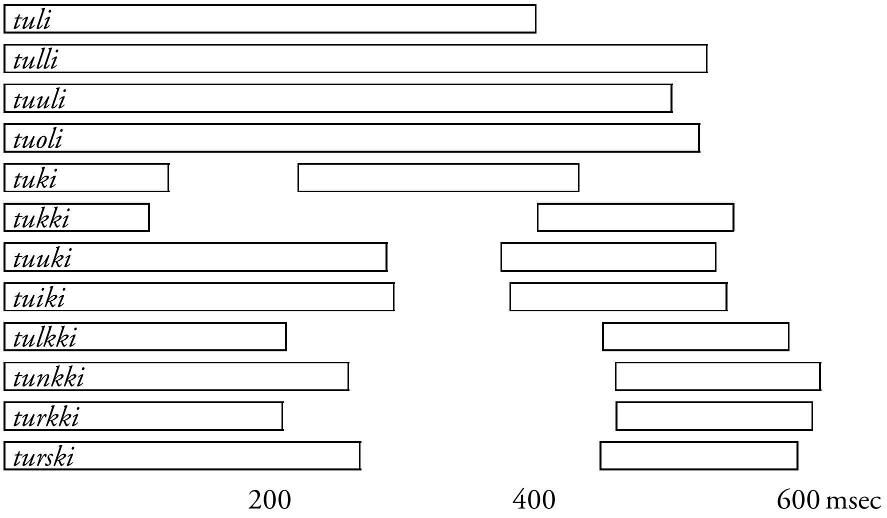


Figure 2.1: (c) Subject KV, normal reading

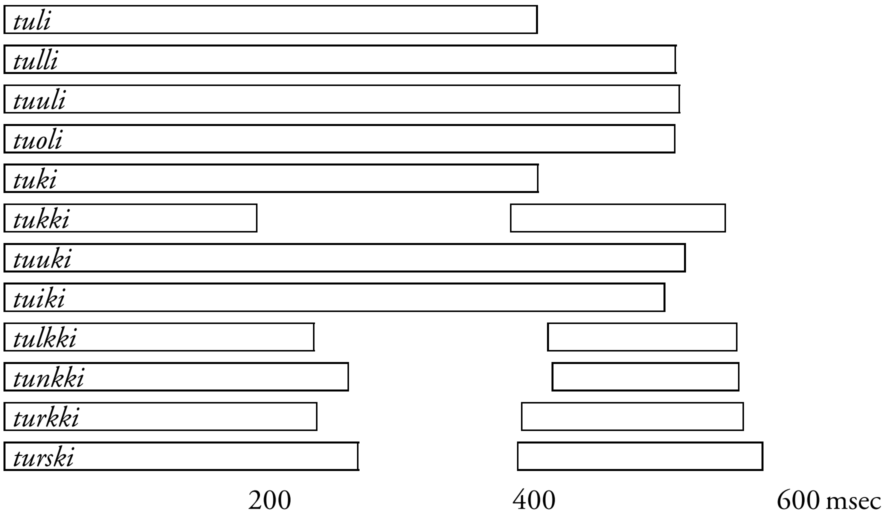


Figure 2.1: (d) Subject KV, closed mouth

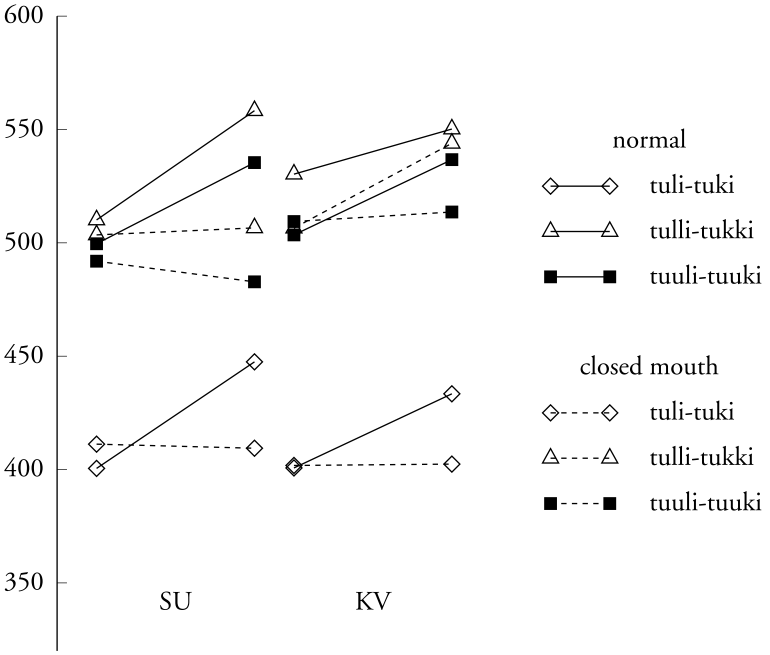


Figure 2.2: Mean total duration of test words for subjects SU and KV

mouth task (statistically speaking), but not all, in spite of the fact that stress in Finnish words is always on the first syllable.

Quantity pattern vs. segmental differences. It is possible to examine the effect of (a) medial consonant and (b) quantity pattern on total duration for the two experimental conditions by looking at the subset of the test words that differ only in these two factors:

	CVCV	CVCCV	CVVCV
medial-*l*	*tuli*	*tulli*	*tuuli*
medial-*k*	*tuki*	*tukki*	*tuuki*

The results for this subset of the test words are shown graphically in Figure 2.2. It would appear from the figure that the difference in total duration between words with different consonant type is greatly diminished in the closed mouth condition, while differences due to quantity type remain. This result is born out by analysis of variance as well. The model used was

$$y_{ijk} = \mu + B_i + \delta_{(i)} + Q_j + C_k + QC_{jk} + \epsilon_{(ijk)} \qquad (2.1)$$

where y_{ijk} is the observed total duration of the test word, μ is the overall mean, B_i is the BLOCK effect, $\delta_{(i)}$ is the restriction error within the ith block, Q_j is the QUANTITY effect, C_k is the CONSONANT effect, QC_{jk} is the QUANTITY BY CONSONANT interaction, and $\epsilon_{(ijk)}$ is the (residual) error term (cf. eg. Anderson and McLean, 1974, Chapter 5). The results of the analysis of variance are presented in Table 2.3. The QUANTITY BY CONSONANT interaction was in no case significant.

As can be seen in this table, the QUANTITY effect is very significant in both conditions (tasks) for both subjects, while the CONSONANT effect is very significant for both subjects in the normal condition, but is not significant for either subject in the closed mouth condition.

The implication may be drawn that the closed mouth condition "filters out" or neutralizes segmental effects, leaving only the effects of the "rhythmic program". In other words, timing differences due to differing segments are unintentional, mechanical consequences of the articulations involved, and thus are neutralized when the articulatory differences are removed. Those due to differing quantity patterns (different rhythmic programs), on the other hand, are controlled directly, and thus remain even in the closed mouth condition.

Remaining test words. How do the other test words fit into this scheme? To gain some insight into this question, individual mean durations were compared for all test words using a Newman-Keuls test (cf. eg. Anderson and McLean, 1974) based on a model similar to the one in Equation 2.1, but including all test words:

$$y_{ij} = \mu + B_i + \delta_{(i)} + W_j + \epsilon_{(ij)} \tag{2.2}$$

where W_j is the test word effect. Results for risk level $\alpha = 0.05$, separately for each subject and each task, are shown in graphic form in Figure 2.3 a–d. In these figures the mean total duration of each test word is shown with a dot, and means which do not differ enough to be significant at the $\alpha = 0.05$ level are connected (underlined) with a single double-headed arrow.

First of all, the results of the analysis of variance carried out above are not contradicted by the Newman-Keuls test: For both subjects, *tuli* and *tuki* differ significantly from all other test words in both conditions (QUANTITY effect), but differ from each other only in the normal condition, not in the closed mouth condition (CONSONANT effect). In addition, for subject SU, *tuuli* is significantly shorter than *tuuki* and *tulli* is significantly shorter than *tukki* in the normal condition, but not in the closed mouth condition.

In general, it is easy to see in these figures that the mean durations of the test words are not as evenly "spread out" in the closed mouth task. Counting the minimum number of groups possible using the information from the Newman-Keuls tests, it can be seen that for subject SU the means fall into at least eight distinct groups in the normal condition, whereas four are sufficient in the closed mouth condition. For subject KV the means fall into at least five distinct groups in the normal condition but only three are needed in the closed mouth condition. This situation is parallel to that reported by

	SU		KV	
test word	normal	closed	normal	closed
tuli	400.5	411.2	400.7	401.8
tulli	510.1	503.5	530.3	506.5
tuuli	499.6	491.9	503.5	509.4
tuoli	481.8	516.6	524.3	505.8
tuki	447.5	409.4	433.4	402.4
tukki	558.3	506.6	550.2	543.9
tuuki	535.4	482.8	536.7	513.6
tuiki	548.2	502.4	545.3	498.1
tulkki	601.7	531.1	591.8	552.8
tunkki	573.1	533.1	615.8	554.4
turkki	613.5	532.2	610.0	557.7
turski	583.3	531.4	598.6	572.1

Table 2.2: Mean durations (msec) of test words in normal and closed mouth reading tasks for subjects SU and KV

		QUANTITY		CONSONANT	
subject	task	$F(2,45)$	p	$F(1,45)$	p
SU	normal	142.73	< 0.001	57.86	< 0.001
	closed	157.14	< 0.001	0.31	0.578
KV	normal	98.50	< 0.001	13.84	0.001
	closed	66.42	< 0.001	2.17	0.148

Table 2.3: Analysis of variance for test words differing in medial consonant and quantity pattern

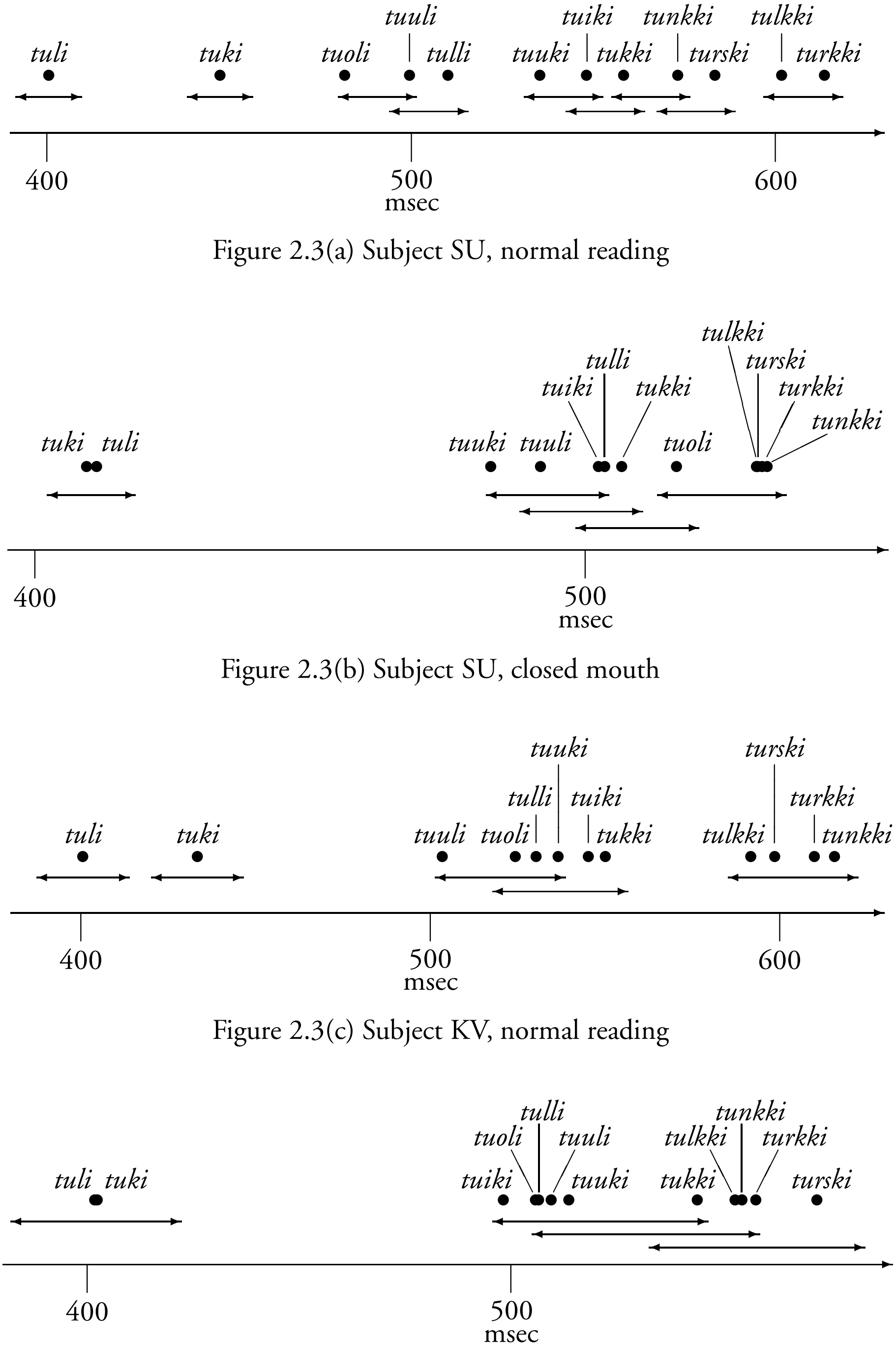


Figure 2.3(a) Subject SU, normal reading

Figure 2.3(b) Subject SU, closed mouth

Figure 2.3(c) Subject KV, normal reading

Figure 2.3(d) Subject KV, closed mouth

Figure 2.3: Mean durations of test words grouped into nonsignificant ranges.

Kozhevnikov and Chistovich. Assuming, with Kozhevnikov and Chistovich, that the closed mouth condition represents the separate "rhythmic program" of the test words with segmental effects neutralized, it will evidently be necessary to postulate at least three or four different rhythmic structures for the test words included in this experiment. A possible grouping consistent with the Newman-Keuls tests for total duration in the closed mouth condition for both subjects, and with the hypothesis that test words differing only in segmental content belong to the same group is as follows:

1. *tuli*, *tuki*.

2. *tuuli*, *tuuki*.

3. *tulli*, *tukki*, *tuoli*, *tuiki*.

4. *tulkki*, *tunkki*, *turkki*, *turski*.

Test word *tuiki* could just as well be placed in the second group. Test word *tuoli*, however, can not be placed in the second group, at least for subject SU, since for that subject, *tuoli* was significantly longer than both *tuuli* and *tuuki* in the closed mouth condition.

Word internal silence

Thus far only the total durations of the test words have been taken into account. The presence of word internal silence for some words in the closed mouth condition (as well as the normal reading where silence is expected as a concomitant of voiceless consonants), does not in general alter the conclusions reached in the previous section. However, the pair *tulli* – *tukki*, which was not distinguished by total duration in the closed mouth task, differs in that only *tukki* consistently showed a period of word internal silence (see Table 2.1). It would thus appear that some segmental differences have indeed survived in the closed mouth task. Of course the closed mouth task does not restrict the functioning of the glottis, so this is, perhaps, not so surprising. This does not explain, however, why silent periods did not show up in cases of the short voiceless consonant [k]. Perhaps voicing is (partly) dependent on rhythmic structure (consonant quantity) as well as segmental differences.

The last four test words (*tulkki*, *tunkki*, *turkki*, *turski*) consistently showed word internal silence in both tasks. In addition, they represent the same quantity pattern with segmental differences only (with the possible exception of *turski*). It is thus of interest to see whether the timing of these words is compatible with the conclusions based on total duration alone.

		UX		K		I	
subject	task	$F(3,27)$	p	$F(3,27)$	p	$F(3,27)$	p
SU	normal	4.13	0.016	4.97	0.007	10.48	< 0.001
	closed	1.99	0.140	0.47	0.706	0.04	0.989
KV	normal	13.88	< 0.001	0.86	0.475	1.36	0.277
	closed*	1.10	0.371	1.32	0.296	1.22	0.329

* $F(3,20)$

Table 2.4: Analysis of variance for *tulkki*, *tunkki*, *turkki*, *turski*

Analyses of variance were carried out for these test words using the following three dependent variables:

1. UX: Duration from end of pre-word silence to beginning of word internal silence.

2. K: Duration from end of pre-word silence to end of word internal silence.

3. I: Duration from end of pre-word silence to beginning of post-word silence (ie. total duration).

The model for each analysis was identical to the model on page 26 (Equation 2.2), restricted to the last four test words. Results are shown in Table 2.4 for each speaker and each task separately. All in all, these results show the same pattern as before: at least some of the variables for each subject showed significant differences between test words in the normal condition, indicating expected timing differences due to segmental differences, but no variable showed a significant difference for either subject in the closed mouth condition. A multivariate analysis of variance taking all three variables simultaneously into account produced the same result.

2.2 Independence of rhythm in perception

The question for perception is: How is quantity perceived? That is, how are patterns of long and short discriminated? One hypothesis compatible with extrinsic control of quantity pattern production according to a "biological clock" or rhythm generator, is that the decision as to short or long is based only on the absolute duration of the segment in question. For this to work, the listener must have (a) some way to detect segment boundaries and (b) some sort of "clock" to tick off the time while waiting for the next boundary to occur.

One difficulty with this hypothesis is that segment durations (as normally measured) for long and short quantities are typically not unique but overlap considerably. It is implausible that boundary detection criteria could be found which would guarantee (near) identical absolute durations regardless of what segment is being measured.

At the very least, then, the model should include a separate "reference duration" (corresponding to what is often termed *inherent duration*) for each type of segment.

But even the absolute durations for the "same" segment vary considerably so that, e.g. a "long" [kk] in one instance may be shorter in absolute duration than a "short" [k] in another instance. Any algorithm based on absolute duration alone will be subject to considerable error, since very few if any absolute duration values will be guaranteed to represent only short, or only long. Of course, so-called top down processing should be able to compensate for some of these errors.

Duration normalization

A more plausible model, which is still based on absolute durations and thus still requires boundary detection and clocking, may be called the moving average hypothesis. In this model, the decision point between long and short is not fixed, but varies with each successive segment. That is, after identification of a segment as short or long, the expected durations for long and short are adjusted to comply with the perceived duration of that segment. (Or alternatively, one may say that segment length is timed by a "clock" whose rate is affected by the measured length of the previous segment(s).) Such an algorithm allows more flexibility—provided speaking rate does not change radically from one segment to the next, fluctuations can be handled. To see how this might work, let us follow a simplified example of successive durations.

In this example many effects are ignored for simplicity, such as differences in inherent duration for different segments, and phonotactic ("syntagmatic") constraints, but these could readily be incorporated. Suppose the system starts out with the assumption that a 150 msec duration represents the cutoff point between short and long segments. This value is adjusted after each quantity judgment by a simple rule such as: "The cutoff point is 150% of a short duration and 75% of a long duration." (This rule is consistent with a ratio of 1 : 2 between short and long. Instead of being determined solely by the previous duration, the new boundary value could be a weighted average of the old cutoff value and the value based on the previous duration.) The system then encounters the sequence of durations shown in Table 2.5 (in msec).

Note that interpretation of a particular duration depends on the immediate context and not just on an absolute value. In this example for instance, durations of 115 msec and 130 msec are interpreted as short towards the beginning of the sequence, but as long at the end. According to this model, information in the signal is neatly partitioned: everything between segment boundaries is utilized for identification of the segment in question (quality information), while the time elapsed between boundaries supplies quantity information. Success in recognizing the quantity patterns involved depends directly on the accuracy of the assumed biological clock. Here *accuracy* means approximation to a universal timekeeper.

It is possible to link this type of model with the practice sometimes used in phonetic timing research of reporting relative durations (ie. ratios of durations measured

Observed duration	Quantity interpretation	Adjusted boundary value
		(150.0)
115	short	172.5
210	long	157.5
130	short	195.0
220	long	165.0
110	short	165.0
210	long	157.5
190	long	142.5
95	short	142.5
180	long	135.0
80	short	120.0
150	long	112.5
65	short	97.5
50	short	75.0
40	short	60.0
115	long	86.25
130	long	97.5

Table 2.5: Example of sequence of durations

locally from the same utterance) instead of absolute durations, for given a model of this type, ratios of adjacent durations will reflect length differences with less influence of global fluctuations in tempo.

For Kozhevnikov and Chistovich there is an articulatory program that includes instructions for regulating the timing of successive syllables. However, there does not seem to be any straight forward correspondence to timing of acoustic parameters. More recently the proposal has been made that a syllable has a *psychological center* (or *p-center*) serving as the instant of alignment rather than any particular acoustic boundary. This proposal is prompted by data from experiments of two types—production experiments in which speakers are asked to produce series of isochronous syllables, and perception experiments in which subjects are asked to adjust the spacing between syllables in series until they sound isochronous. The syllables obtained in such production experiments are not isochronous as measured by acoustic onset, but they do correspond to the spacing perceived by subjects as isochronous in the perception experiments. Cooper *et al.* conclude that

> These findings indicate that the event that listeners attend to when judging relative timing is opaque to conventional measurement techniques. (Cooper *et al.*, 1986, p. 187)

Of course it is also possible that the perceived timing pattern itself is distorted by conventional measurement techniques.

The results which prompted the notion of p-centers have received two opposing explanations in the literature. What one may call the "auditory explanation" (e.g. Howell, 1984, 1988) gives a general account in terms of human hearing—the p-center is a function of the amplitude envelope of the signal and not peculiar to speech. Under this explanation, speakers try to produce a signal which will sound isochronous. Opposed to this is the "articulatory explanation" (e.g. Fowler, 1979; Tuller and Fowler, 1980; Cooper *et al.*, 1986; Fowler *et al.*, 1988) which says basically that it is muscle activity which is isochronous, and this is what listeners attend to. The acoustic signal merely transmits this information (ie. perceivers extract information about a "distal source" which in the case of speech is activity of the vocal tract).

Both explanations share the basic assumption of the p-center concept, that the "synchronizing" events that listeners pick out are *points* in physical time (see also Hoequist, 1983), and would thus seem to require some form of extrinsic timing model in order for the proposed isochrony to be produced and perceived.

The two explanations differ in their predictions as to what will happen if segmental (quality) information in the speech signal is deliberately degraded so that listeners can no longer tell which articulatory gestures were used in production, whereas intensity variation is preserved. According to the articulatory explanation, we would not expect listeners to be able to extract quantity information (or other timing information) independently of segmental information (ie. without knowledge of the gestures involved), since for different segments the timing relationship between muscular activity and acoustic effect is radically different. The auditory explanation on the other hand would predict that quantity and segmental content can be perceived independently.

2.3 Experiment 2: Quantity perception based on intensity variation

Part of the argumentation for and against intrinsic timing in speech perception has centered around so-called rate-dependent processing of speech. Summerfield (1979, 1981) noted that at first blush the fact that some aspects of speech timing (eg. VOT and closure duration of voiceless stops) are perceived differently depending on the articulation rate of an immediately preceding phrase would seem to support an extrinsic model of perception—previous articulation rate could "adjust" the biological "clock" by means of which durations are judged. Summerfield found, however, that voicing judgments (based on VOT) are affected not only by preceding syllables but also by the duration of the vowel following a stop. It would seem that effects are (mainly) centered around the stop articulation itself within about 400 msec, about the time required to execute the stop articulation at normal rate. Furthermore, when the precursor phrase was replaced by a tune, fast and slow variants of the tune had no affect on the test word. Diehl *et al.* (1980) found that changing the fundamental frequency of the precursor phrase, thus giving the impression of a different speaker, also reduced the influence of the precursor. Summerfield interprets these results as supporting an intrinsic timing account: "the precursive influence of articulatory rate, earlier interpreted as reflecting the influence of extrinsic timing, may be better described as the result of manipulations applied to the intrinsic time course of the acoustical concomitants of the event of stop consonant production itself." (Summerfield, 1981, p. 1089).

On the other hand, Gordon (1988) found that when precursor sentences spoken at various rates were replaced with non-speech signals that matched the amplitude envelope of the original carrier sentences, enough rhythm information remained to affect segmental judgments about test words (English *rapid* vs. *rabid*) in the same way as the real carrier sentences. Gordon concludes that "the speech signal contains information about rate of articulation, independent of the identity of phonetic segments, and that listeners can exploit this information in rate-dependent processing." (Gordon, 1988, p. 144).

If independent rhythm information is available at the phrase level in English, the question arises whether it might not be independently available at word level in a quantity language such as Finnish. Experiment Two was inspired by Gordon's work, and was intended to test whether the amplitude envelope of the speech signal alone provides enough rhythm information to allow listeners to make quantity judgments independent of segmental content. Of several different non-speech signals used by Gordon, sine waves at a steady-state frequency close to the original fundamental frequency produced the best performance. For this reason amplitude matched sine waves were used in the present experiment.

2.3.1 Methods

Several tokens each of the sentences

Mitä sana tuuli tarkoittaa?	'What does the word wind mean?'
Mitä sana tulli tarkoittaa?	'What does the word customs mean?'
Mitä sana tuli tarkoittaa?	'What does the word fire mean?'

spoken by a male native speaker of Finnish living in Jyväskylä were recorded on DAT tape. One token was selected for each sentence and transferred in digital form from DAT tape to computer and the test words *tuuli*, *tulli*, and *tuli* were replaced with sine waves at the mean fundamental frequency of the original test words (108 Hz). The amplitude of the sine wave was varied one period at a time to match the (digital) root mean square (RMS) of the original test word. Replacement of the test words was done by aligning the abrupt onset of high energy in the sine wave which corresponded to the release of the [t] in the original. Counting backwards from this point an integral number of (very low energy) sine periods were used corresponding to the closure of the [t] itself. The total length of the replaced word was also an integral number of sine periods so that the test "word" started and ended at a zero crossing. The carrier sentence was adjusted so that the end of the pre-word signal and the beginning of the post-word signal came as close to zero as possible to avoid clicks. In all cases, the adjustment amounted to less than a millisecond. The durations of the replaced "words" were: *tuuli* – 370 msec, *tulli* – 398 msec, and *tuli* – 296 msec.

A stimulus tape was made by transferring the sentences in digital form to a DAT tape in random order at a rate of one sentence every four seconds. The interstimulus interval was approximately two seconds. (In a pilot test using an interstimulus interval of approximately 300 msec subjects complained that they had no time to think.) Each sentence appeared on the stimulus tape ten times for a total of 30 stimuli, and the total length of the tape was thus 118 seconds. A second stimulus tape, used as control, consisted of the original sentences in random order. Again there were ten tokens of each of the three sentences for a total of 30. Half of the subjects ($N = 5$) heard the control tape first, and were then told that the test tape contained versions of the same sentences with the test word severely "muffled" (Finnish *vaimennettu*). In other words, these subjects were aware of the segmental content (t–u–l–i) of the original test words. In addition, their answer sheets had the three choices marked as **tuli**, **tulli** and **tuuli**. The other subject group ($N = 5$) heard the test sentences first, and were not told what the original test word was. They were asked to decide whether the "muffled" word consisted of consonant—vowel—consonant—vowel, for example *papa*, consonant—vowel—double consonant—vowel, for example *pappa*, or consonant—double vowel—consonant—vowel, for example *paapa*. Choices on the answer sheets for this group were marked as **"papa"**, **"pappa"** and **"paapa"**. The instructions emphasized that the test words were not necessarily composed of p and a, but were of the same rhythmic type as the example words.

2.3.2 Results

The first result, which is not surprising, except perhaps in its consistency, is that all subjects scored perfectly on the control sentences.

The scores for the sine-replaced test words were considerably poorer. The analysis may be simplified by lumping all mistakes together and considering only the number of correct answers. The "no knowledge" group (those that heard the sine-replaced versions of the test words first, and were not aware of the segmental content of the words) scored 46 correct answers out of 135 responses (14 missing responses and one ambiguous response were left out of consideration), or 34.1% correct. Since there were three possible responses, the chance level was 33.3%. Indeed, the number of correct responses does not differ significantly from chance as calculated with the half-integer normal approximation to the binomial distribution ($z = 0.09$, $p > 0.46$). The "prior knowledge" group (those that heard the control sentences first, and were aware of the segmental content of the "muffled" test words) scored 57 correct answers out of 150 responses, or 38% correct. While this overall score is somewhat better than for the no knowledge condition, the number of correct responses still does not significantly differ from chance ($z = 1.13$, $p > 0.12$). Combining results for the two groups gives 103 correct answers out of 285 responses, or 36.1% correct ($z = 0.94$, $p > 0.17$). Table 2.6 shows the observed number of responses to each stimulus, given separately for the two subject groups.

When we look at the responses of individual subjects, an interesting fact emerges. Taken individually, two subjects, both in the prior knowledge group, performed significantly better than chance. Subject number 3 had 15 correct responses out of 30 (50% correct, $z = 1.74$, $p < 0.05$), and subject number 4 had 18 correct out of 30 (60% correct, $z = 2.905$, $p < 0.01$). No other subjects in either group performed better than chance at the 0.05 level of significance. Of course there is bound to be chance fluctuation in the number correct from subject to subject. How likely is it that two subjects out of ten get at least 15 out of 30 correct due merely to chance fluctuation? Since $p \approx 0.041$, we would expect one such occurrence in a group of about 24 subjects, two such occurrences in a group of about 49 subjects. The probability of at least two such subjects in one group of ten is approximately 0.06. Likewise, how likely is it that one subject get at least 18 out of 30 correct by chance? Since $p \approx 0.002$, we would expect one occurrence in a group of about 500 subjects. The probability of at least one such subject in one group of ten is less than 0.02. This suggests that the performance of these two subjects was not entirely due to chance. The fact that both subjects were in the prior knowledge group suggests there may have been a slight ad-

	prior knowledge			no knowledge		
response:	**tuli**	**tulli**	**tuuli**	**"papa"**	**"pappa"**	**"paapa"**
stimulus						
"tuli"	19 38%	24 48%	7 14%	16 35.6%	21 46.7%	8 17.8%
"tulli"	14 28%	20 40%	16 32%	12 28.6%	19 45.2%	11 26.2%
"tuuli"	8 16%	24 48%	18 36%	16 33.3%	21 43.8%	11 22.9%
total	41 27.3%	68 45.3%	41 27.3%	44 32.6%	61 45.2%	30 22.2%

Table 2.6: Response frequencies for Experiment 2

vantage in knowing ahead of time the segmental content of the test words. However, it is of course possible that these particular subjects would have performed as well had they been assigned to the no knowledge condition. Given two "talented" subjects in a group of ten, the probability of assigning them both to the prior knowledge condition by chance, $p = 2/9$ (0.222), is small, but not very small.

2.3.3 MCMC analysis

Another statistical analysis was carried out using a general model formally identical to one called the *Mover-Stayer Model* in Upton 1978 (cf. also the model in Goodman 1975). In this model it is assumed that there is a certain probability of recognition (to be estimated) for each stimulus type (say p_i^S for stimulus i), otherwise the subject guesses (thus the probability of guessing for stimulus i will be $(1 - p_i^S)$). Guessed responses are assumed to be independent of stimuli. All wrong responses are considered the result of guessing only, but correct responses are the result of either recognition or guessing. Letting p_j^R be the probability of response j when guessing, the general model is thus

$$p_{j \cdot i}^{R \cdot S} = \begin{cases} (1 - p_i^S)p_j^R, & \text{for } i \neq j; \\ (1 - p_i^S)p_j^R + p_i^S, & \text{for } i = j. \end{cases} \tag{2.3}$$

where $p_{j \cdot i}^{R \cdot S}$ is the conditional probability of response j given stimulus i.

Posterior distributions for these probabilities, separately for each subject, given the response data, were calculated using the WinBUGS program (Spiegelhalter *et al.*, 2003). The model is shown schematically as a directed acyclical graph in Figure 2.4. Each node in the graph represents a quantity in the model, whether a constant fixed

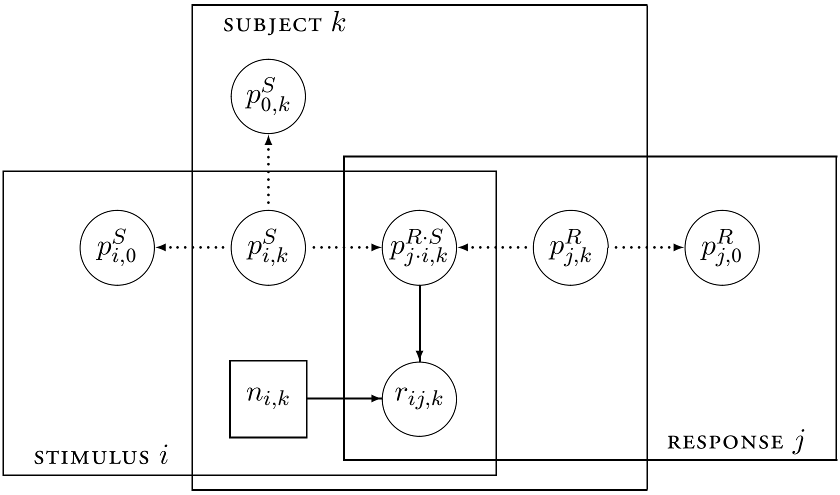


Figure 2.4: Directed acyclical graph of model for Experiment 2

by design (represented by rectangles), a stochastic variable associated with a probability distribution, or a deterministic variable which is merely a logical function of other quantities used for convenience. Solid arrows represent stochastic dependence, while dotted arrows represent logical functions. In addition, the graph is greatly simplified by reducing sets of repeated variables (for instance the subject related variables of the present study) to single subscripted nodes enclosed in a large rectangle labeled with the repeated factor.

The quantity $r_{ij,k}$, denoting the number of responses in category j to stimulus i by subject k is assumed to be distributed as a multinomial random variable

$$(r_{i,1,k}, r_{i,2,k}, r_{i,3,k}) \sim \text{Multinomial}(p^{R\cdot S}_{1\cdot i,k}, p^{R\cdot S}_{2\cdot i,k}, p^{R\cdot S}_{3\cdot i,k}; n_{i,k}) \tag{2.4}$$

where $n_{i,k}$ indicates the number of stimuli in category i responded to by subject k and the probabilities probabilities $p^{R\cdot S}_{j\cdot i,k}$ are given by Equation 2.3. The two "founder" nodes (nodes with no parent nodes) in the model were given uninformative prior distributions:

$$\begin{aligned} p^S_{i,k} &\sim \text{Beta}(1,1) \\ (p^R_{1,k}, p^R_{2,k}, p^R_{3,k}) &\sim \text{Dirichlet}(1,1,1) \end{aligned} \tag{2.5}$$

where Dirichlet$(\alpha_1, \alpha_2, \ldots)$ denotes a Dirichlet distribution[1] and Beta(a, b) denotes a beta distribution with parameters a and b, which in the present case is equivalent to a uniform distribution over the interval from 0 to 1.

[1] Probability density function $f(p_1, p_2, \ldots, p_n) = \left(\Gamma\left(\sum_i \alpha_i\right) / \prod_i \Gamma(\alpha_i)\right) \prod_i p_i^{\alpha_i - 1}$, with $\sum_i p_i = 1$.

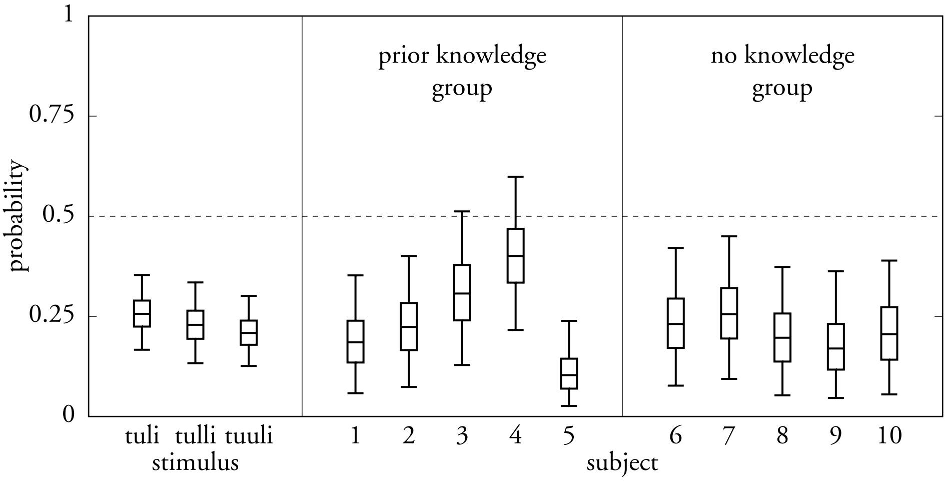


Figure 2.5: Estimated probabilities of recognition

In addition three sets of probabilities were computed during the MCMC run for summary purposes: $p^S_{i,0} = \frac{1}{10}\sum_k p^S_{i,k}$, indicating the probability of recognizing stimulus i by random subject in the group, $p^S_{0,k} = \frac{1}{3}\sum_i p^S_{i,k}$, indicating the probability for subject k of recognizing a random stimulus, and $p^R_{j,0} = \frac{1}{10}\sum_j p^R_{j,k}$, indicating the probability of guessing response j for random subject in the group.

Convergence of the MCMC run was monitored visually using two chains starting form overdispersed initial values. Convergence appeared to be quite rapid. After an initial "burn-in" of 5000 samples in each chain convergence was checked using the Gelman-Rubin statistic provided by the WinBUGS program. Since convergence was satisfactory, the burn-in period was discarded and an additional 50 000 samples were generated to represent the posterior distribution of parameters. The adequacy of the 50 000 sample run was monitored by checking that the MC error automatically computed by the WinBUGS program for each parameter was less than 5% of the estimated parameter standard deviation.

Model fit was monitored by computing an independent set of $\hat{r}_{ij,k}$ during the MCMC run and visually comparing the results with the empirical $r_{ij,k}$ using the WinBUGS model fit comparison tool. All empirical $r_{ij,k}$ fell within the 95% credibility intervals (CI) for the $\hat{r}_{ij,k}$, indicating an excellent fit.

Results for the recognition probabilities are shown in Figure 2.5, first by stimulus, then by subject and divided into the two subject groups. In each case the posterior distributions are reduced to five points: the top and bottom crossbars indicate the 95% CI, the central box includes the 50% CI, and the center crossbar indicates the median of the distribution.

The values obtained for probability of recognition were median $p^S_{1,0}$ (for stimulus *tuli*) $= 0.2564$, 95% CI $(0.1667, 0.353)$, median $p^S_{2,0}$ (for stimulus *tulli*) $= 0.229$, 95% CI $(0.1331, 0.3348)$, median $p^S_{3,0}$ (for stimulus *tuuli*) $= 0.2086$, 95% CI $(0.1262, 0.3013)$. It can be seen in Figure 2.5 that the two subjects with the best recognition records (subjects 3 and 4) both belong to the prior knowledge group as noted before. However, the difference is not great, and it is perhaps indicative that subject number 5 with the least recognition also belongs to this group.

The values obtained for probability of responding in each category when guessing were median $p^R_{1,0}$ (for response CVCV) $= 0.2839$, 95% CI $(0.229, 0.3418)$, median $p^R_{2,0}$ (for response CVCCV) $= 0.4673$, 95% CI $(0.4048, 0.5301)$, median $p^R_{3,0}$ (for response CVVCV) $= 0.2474$, 95% CI $(0.1958, 0.3054)$. It is apparent that the category CVCCV (**tulli** for the prior knowledge condition, **"pappa"** for the no knowledge condition) was favored over the other two. Since this response category was located between the other two on the response forms, this response bias probably reflects a tendency to favor the central category when guessing.[2]

2.3.4 Discussion

The basic result of this experiment was that the signal intensity (RMS) information alone was not sufficient for perception of quantity. Even the "best" subjects who answered at better than chance levels performed very poorly.

It would appear that perception of quantity depends on recognition of segments. Provided that timing perception for quantity patterns is parallel to perception of "isochronous" syllables, this might be taken to support an articulatory interpretation of quantity perception. However, it is important to note that mere conscious knowledge of the segmental content of the test words did not allow subjects to identify the quantity pattern from intensity fluctuation, as was seen in the prior knowledge condition. Thus the difficulty the listeners encountered was not due to ambiguities in the "inherent durations" involved. That is, it cannot be assumed that listeners were successful in calculating the *durations* of the test word segments, but were unable to identify them as long or short only because they didn't know the appropriate "reference durations" to compare them with.

Gordon interpreted his results as support for extrinsic timing—intensity fluctuation of the speech signal (even if reduced to fluctuating sinusoids) drives or synchronizes an internal clock by which rate-dependent judgments are made.

[2] O'Dell (1995) estimated the probabilities of the model presented here, by fitting various log-linear models of quasi-independence (ignoring the diagonal) to the frequency data of Table 2.6. The estimates obtained for the response probabilities when guessing (p^R_j) were very similar to the present results, but the estimates for recognition probabilities (p^S_i) were much lower: 0.125, 0, and 0.167 for the prior knowledge condition only. This discrepancy is probably due to the fact that in the log-linear model fitting individual subjects were lumped together to avoid low frequency cells, and also to the fact that the prior distribution used here is more conservative, lessening the chance of extreme estimates near zero or one.

> It is apparent that some aspects of the timing of speech are extrinsic in the sense that they can be extracted and utilized by listeners in situations in which the specific underlying articulatory movements cannot be identified. (Gordon, 1988, p. 142)

Gordon suggested that the failure of earlier experiments (Summerfield, 1979, 1981; Diehl *et al.*, 1980) to exhibit rate effects of non-speech precursor "sentences" was due to discontinuity of fundamental frequency between precursor and test words, or to the fact that in the earlier experiments the precursors were identifiable sources which conflicted with the test word, while in Gordon's sinusoid experiment the source of the precursor was unclear to listeners. This suggestion seems to imply that the listener may have several internal clocks running at the same time—one for each identifiable source. In any case, this cannot explain the failure of the present experiment, since the transformations used were the same as Gordon's. It would appear that intensity fluctuation of the speech signal was not able to drive or synchronize a proposed clock by which quantity judgments are made.

A series of experiments conducted by Remez and Rubin (1990) suggests a possible explanation. In their first two experiments Remez and Rubin found that "the perception of segmental and suprasegmental linguistic attributes" (as measured in a transcription plus syllable counting task) was not impaired by using a steady-state or "misleading" amplitude envelope in a sinusoidal replica of the sentence *Where were you a year ago?*, but that the natural amplitude envelope proved essential for a sinusoidal replica of the sentence *My T.V. has a twelve-inch screen*. They hypothesized that the difference was due to the fact that in contrast to the first sentence, the (original) second sentence had many stretches of silence corresponding to (voiceless) stops. Their third experiment confirmed this by using versions of the second sentence in which the silent portions corresponding to the stops were left intact. With this modification, the use of (otherwise) steady-state or misleading amplitude envelopes did not impair perception.

The carrier sentence used as a model for the precursor in Gordon's experiments was "*I'm trying to say …*". The fact that this sentence includes two voiceless stops suggests the possibility that utilization of intensity fluctuation for perception of rhythm depends on the presence of stops.

Accordingly, the task in the present experiment might have been easier for subjects had the test words contained an intervocalic stop (for instance *kato*, *katto*, *kaato*) instead of a lateral. Of course the appearance of silence would mean that considerable segmental information would remain in the sinusoid versions and would thus weaken any conclusion that quantity could be perceived independently of the particular segments involved. There is some circumstantial evidence that Finnish quantity patterns may be more easily perceived for voiceless obstruents (basically [p t k s]) than for other consonants. Some dialects of Finnish lack the phonological long vs. short opposition for sonorant consonants (such as [m n l r]) in many positions where the obstruent opposition is maintained (especially after long vowels or unstressed syllables, cf. Uusivirta, 1971; Rapola, 1966, pp. 273–285). However, modern standard Finnish *does* distinguish quantity patterns for sonorant consonants as well as obstruents, thus the abrupt intensity changes of voiceless obstruents cannot be *essential* to quantity perception in real speech.

Also, anticipating the results of Experiments 3 and 4 below, it is possible that the listeners' task would have been easier had the original F_0 contour been preserved in place of the constant 108 Hz frequency. In fact, it is not unreasonable to assume that in general, the more speechlike stimuli are, the more likely they are to be perceived in a speechlike manner.

Златоустова (1981, p. 49–52) reported a comparatively high rate of identification of rhythmic structure (ie. number of syllables and location of primary stress) for listeners presented with words masked by noise (above 90% correct overall, with a maximum of 93% for words containing three syllables or less, only the vowel [a] and no consonant clusters). In spite of the noise masking, however, there was evidently much segmental information in the signal, since the identity of the vowels and consonants was observed to have a large effect on the results.

Chapter 3

Model based on continuously varying articulation rate

> "If you knew Time as well as I do," said the Hatter, "you wouldn't talk about wasting *it*. It's *him*."
>
> "I don't know what you mean," said Alice.
>
> "Of course you don't!" the Hatter said, tossing his head contemptuously. "I dare say you never even spoke to Time!"
>
> "Perhaps not," Alice cautiously replied: "but I know I have to beat time when I learn music."
>
> "Ah! that accounts for it," said the Hatter. "He won't stand beating. Now, if you only kept on good terms with him, he'd do almost anything you liked with the clock. For instance, suppose it were nine o'clock in the morning, just time to begin lessons: you'd only have to whisper a hint to Time, and round goes the clock in a twinkling! Half past one, time for dinner!"
>
> —Lewis Carroll

The duration model of quantity assumes that segment durations are what is produced and perceived in timing oppositions. In terms of a rubber band model of timing, this is equivalent to saying that each segment has a constant "elasticity". An alternative to the duration model is a more general version of the rubber band model which allows "elasticity" of the timing pattern to vary continuously rather than changing only at segment boundaries. Another way to think of this is to assume that there is a "master control" which (continuously) regulates the articulation rate. Thus intentional timing differences such as quantity distinctions would be produced not by regulating the instant when a segment boundary occurs but by varying the speed of articulation. When a long segment is called for, the articulation rate is slowed to prolong the speech gesture involved. If a short segment is to follow the long segment articulation rate is speeded up again. According to a theory such as this, it might be more appropriate to speak of "slow" and "fast" segments instead of long and short. Similar assumptions were made in O'Dell, 1987. There it was assumed that a "fundamental rhythm" (or "base rhythm," analogous to the carrier frequency in FM radio) is created and maintained between speaker and hearer, and that quantity information is transmitted as deviations from this.

Naturally, the acceleration or change of rate will not be instantaneous, but will itself take time to be accomplished. This would presumably mean that by measuring durations, we might under some conditions find that phonological length would correlate not only with the duration of the phonetic segment in question, but also the duration of adjacent segments. This would happen, for instance, if slowing down for a long segment started somewhat before the onset of the boundary we have chosen to record. In this way it is possible to explain several well known phenomena in Finnish noted by Lehtonen, such as that "In all the comparisons a double second syllable vowel has lengthened the preceding consonant." (Lehtonen, 1970, p. 123) and that "In all of the cases the vowel preceding a geminate consonant is longer than the vowel before a single consonant." (Lehtonen, 1970, p. 124). Similar results have also been reported for Lappish (Magga, 1984, p. 134). Under a strictly durational model, effects such as these must be handled in some other manner, such as complex adjustment rules. Lehtonen (1979) reported another interesting effect which may be interpreted as supporting a model of quantity differences as rate control. In that study he found for speakers of Finnish that duration of lip movement into and out of rounded vowels was the same regardless of whether a single consonant intervened or not (eg. /ui/ vs. /uki/, or /au/ vs. /aku/), but that duration of the transition was significantly longer when a long (geminate) consonant intervened between the vowels (eg. /uti/ vs. /utti/, or /itu/ vs. /ittu/). If a long consonant means a slowly articulated consonant, this is just what would be expected.

Quantity considered as a feature which is controlled continuously may also provide insight into the behavior of non-native speakers. For instance it is instructive to consider some of the difficulties that speakers of French, a non-quantity language, encounter in learning Finnish quantity patterns (Vihanta, 1987). In production experiments Vihanta compared French learners' and native speakers' productions of various minimal pairs embedded in sentences. Average durations for a typical example are shown in Figure 3.1. Vihanta notes that in general the French subjects do not exhibit the "compensation phenomena" typical for Finnish. Here, for instance, whereas [i] in *tuuli* is much shorter than [i] in *tuli* for native speakers, for the French subjects just the opposite was the case: after successfully lengthening the [u], the French subjects tended to stretch out the following segments as well. In terms of continuous control of quantity, this makes sense if we assume that the French learners of Finnish have not yet learned to "shut length on and off" (slow down and speed up) as quickly as native speakers. This extends to many different structures in Vihanta's study, for instance in cases of sonorant consonant plus short/long stop (eg. *korpi* vs. *korppi*), the French subjects who were successful in making the stop longer also lengthened the preceding sonorant consonant, just the opposite of native Finnish speakers. We can describe such results by saying that for these learners, length often "begins too soon" and/or "stays on too long".

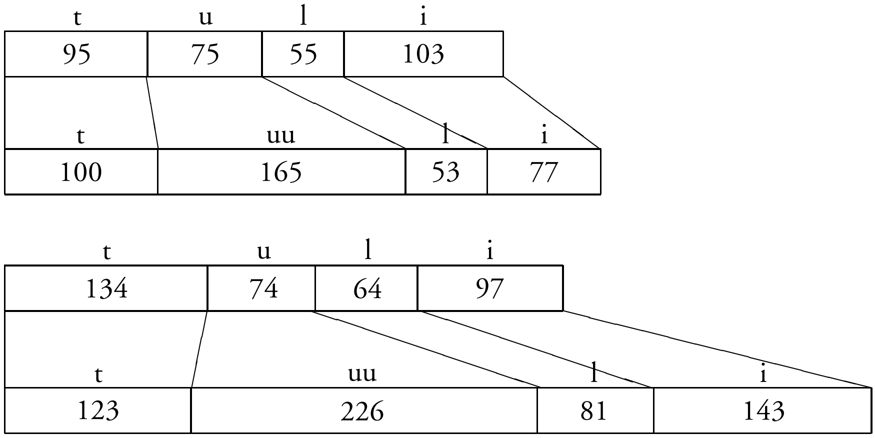


Figure 3.1: Example of average durations for native speakers (top) and French learners of Finnish (bottom) (Vihanta, 1987, p. 116)

Vihanta's perception data for the French subjects listening to native productions also show signs of "localization" difficulties. Often "a word was apparently perceived as containing a long degree of quantity, but listeners were unsure as to how many long sounds the word contained and what these sounds were." (Vihanta, 1987, p. 107, my translation). If quantity perception is based on "measuring time between segment boundaries," it is difficult to see how the French subjects could recognize length without being able to assign it to the proper segment(s).

3.1 Dynamic Time Warping

The idea of "temporal elasticity" which may vary continuously has been incorporated into a body of theory and a general algorithm known as Dynamic Time Warping (DTW). Because this algorithm has found wide application in the field of automatic speech recognition, the theory is well developed. We might say that time warping is the rubber band model given a mathematical foundation. It therefore is useful to consider this theory in more detail. Also, DTW may provide a valuable tool for timing research, especially if the assumptions of the rubber band model prove valid (cf. O'Dell, 1991).

3.1.1 Origins and theory of DTW

DTW has been in use for several years in the field of automatic speech (word) recognition (cf. Dixon and Martin, 1979; Levinson and Liberman, 1981). This is due to the fact that there is great timing variation in speech, due to, among other things, sentence stress, speaking rate, etc. (cf. Klatt, 1976). In a typical automatic speech recognition application using DTW, the unknown word to be recognized is compared with a set of templates for various words the system is capable of "recognizing". The DTW algorithm is responsible for stretching and compressing the word in an "optimal" way to find the best fit before measuring the overall fit between the word and a template. The template giving the best fit is then considered the winner and the input signal is assumed to represent a token of the word corresponding to that template. It has been suggested that timing information may actually be helpful in distinguishing various words and that therefore this type of procedure (DTW), while eliminating some unwanted timing variation in speech, may also eliminate important timing differences as well (Port *et al.*, 1986). This is certainly likely to be true in so called quantity languages such as Finnish, where lexical differences may presumably be signaled entirely by timing.

Kruskal and Liberman (1983, p. 129) suggested that symmetric time warping could be used for "comparison of related utterances so as to study timing variability of normal speech." DTW produces both a measure of overall fit as well as the stretching and compressing needed to achieve that fit. In a typical recognition application only the measure of overall fit is utilized, whereas in a timing study the warp itself (ie. how one signal needs to be stretched to match another) is of most importance. While the traditional method of measuring durations gives us one reference point for each pair of adjacent phonemes (that is, the "boundary" between them), with time warping correspondences between signals are computed almost continually (eg. using reference frames updated every 5 msec).

Terminology

A clear and thorough presentation of the theory of symmetric time warping can be found in Kruskal and Liberman, 1983. Here we provide a short review by defining terms used in the theory and discussion of time warping.

Trajectory. A trajectory can be defined to be "a continuous function of time in multidimensional space, i.e., a time-labelled curve in multidimensional space." (Kruskal and Liberman, 1983, p. 125). This multidimensional space ("feature space" or "parameter space") is meant to characterize the "instantaneous" quality of the speech signal, while the curve traced out in feature space represents the happening or development of the speech event. Timing is then represented by the "speed" with which the trajectory is traced out. In other words, the trajectory is nothing more than a "spatialized" interpretation of the speech signal itself. It is possible for two different trajectories to trace out the same points in feature space and in the same order, but at different local rates. Normally in practical applications of time warping a sampled version of the trajectory is used, ie. the trajectory is approximated by a discrete sequence of points in feature space rather than a continuous function. Each such point corresponds to (and is calculated from) a small section or *frame* of the original signal.

Links. We will call a correspondence between a point of one trajectory and a point of another trajectory (or between points of many trajectories) a *link*. The traditional practice in phonetics of segmenting the speech signal establishes such links between the corresponding segment boundaries of the tokens which are segmented. It is the purpose of time warping algorithms to come up with a set of links between trajectories. In this sense traditional segmenting procedures constitute a sort of course grain time warping. In the case of a discrete sequence, each point represents a (small) time span and thus we must allow links to several points in the other sequence.

Distortion measure. This is a measure intended to assess how different (how "far apart") points of trajectories are to each other. (Sometimes the terms *distance function*, *distance measure* or even *distance metric* have been used, but this is misleading since in general the measures used do not fulfill the mathematical requirements for a *distance*. Cf. Gray *et al.*, 1980) The distortion measure is zero if and only if the points are identical, and should increase the more "different" the points in question are. In practice, the distinction between the distortion measure and the feature space it is based on is somewhat arbitrary. For instance, we may use a set of m autocorrelations as an m-dimensional feature space, and compute distortions by conversion of the autocorrelations into m cepstral coefficients and computing a Euclidean distance ($\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_m - b_m)^2}$). However, if we consider the feature space itself to consist of cepstral coefficients, then the distortion measure will just amount to a Euclidean distance. The end result, the distortion, is the same, and this is all that matters. In comparing distortion measures we are really comparing feature space–distortion measure combinations. In what follows we will use $w(i, j)$ to symbolize the measure of distortion between the ith point in one sequence and the jth point in the other.

Time warping. According to Kruskal and Liberman (1983, p. 125), time warping "refers to comparison of trajectories, or to comparison of sequences derived from them by time-sampling, when each trajectory is subject not only to the usual additive random error but also to variations in speed from one portion to another." An (approximate, discrete) time warping (or simply *warp* for short) is a set of links connecting two (or more) sequences of points in feature space such that each point in each sequence is linked to a point or points (approximately) equal to it in the other sequence(s) without crossing links (ie. the order of points is not violated). *Symmetric time warping* refers to comparison of sequences in such a way that comparing sequence **a** to sequence **b** is guaranteed to give the same result as comparing **b** to **a** (only "in reverse," of course). Some distortion measures, such as the Itakura measure used widely in automatic speech recognition applications (Itakura, 1975), do not treat the two signals being compared identically, and thus give rise to *asymmetrical time warping*. This may be appropriate for comparing an unidentified signal to a known template. For application to speech timing research, however, it is probably best to stick with symmetric time warping, as there is generally no *a priori* reason to value different tokens differently.

Dynamic time warping (DTW). This refers to any of a number of related computational techniques for finding an optimum warp between tokens (sequences). These techniques have in common the fact that they incorporate some form of dynamic programming algorithm (whence the name *dynamic* time warping). The heart of these algorithms is a recursive minimization of the cumulative distortion (D_{ij}) from the beginning of the sequences being compared up to a tentative link between points i of one sequence and j of the other. When the algorithm gets to the link connecting the end points of the two sequences, it has produced the minimum total cumulative distortion. In cases where the links themselves are of interest, they can be recorded along the way and "collected" and the end. Various weights (or "penalties") for compression/expansion have been used as well as various constraints on how much compression/expansion is allowed.

All dynamic time warping in the present research was done using a minimum of constraints on links—in addition to the requirement that links weakly preserve order (ie. they are not allowed to cross), the only constraint was that if a point is multiply linked, the points at the other ends of these links must not be multiply linked. Put another way, only expansions from 1 to k frames, or compressions from k frames to 1 were allowed (cf. Kruskal and Liberman, 1983, p. 139). Following Kruskal and Liberman (1983, p. 149,150), weights for compression/expansion were taken to be proportional to the average time spent in the corresponding regions of the two original trajectories. Since each trajectory was sampled at a constant rate, all frames represent equal time, and the average time spent in a correspondence between 1 frame and k frames is proportional to $(k + 1)/2$. Dividing this by the k links involved, the appropriate weight for each is $(k + 1)/2k$ and the recurrence equation is then:

$$D_{ij} = \min \begin{cases} \min\limits_{2\le k<i} \left[D_{i-k,j-1} + \dfrac{k+1}{2k} \sum\limits_{r=0}^{k-1} w(i-r,j) \right], \\ D_{i-1,j-1} + w(i,j), \\ \min\limits_{2\le k<j} \left[D_{i-1,j-k} + \dfrac{k+1}{2k} \sum\limits_{r=0}^{k-1} w(i,j-r) \right]. \end{cases} \tag{3.1}$$

Kruskal and Liberman note that "a time-warping between two trajectories may be seriously misleading when the interval at which the trajectories are sampled is large in comparison to the differences between them," (Kruskal and Liberman, 1983, p. 129) and they suggest an algorithm to interpolate between sample points during the computation of the warp. When comparing tokens of the same or related utterances, as in the present study, the trajectories will indeed generally be close. An alternative to

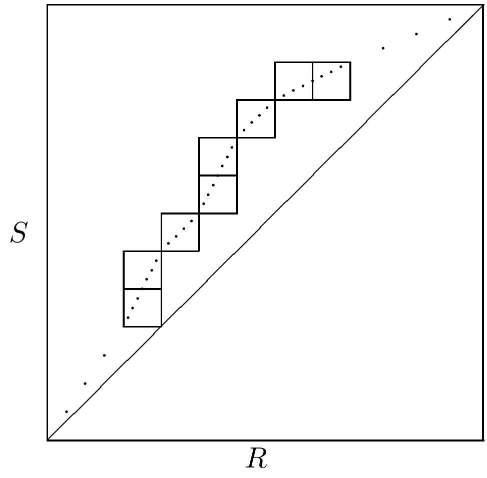


Figure 3.2: Schematic time warp

interpolation is to sample the trajectories at a relatively high rate, thus introducing a high degree of redundancy between adjacent samples along the trajectory.

Representing warps graphically

The results of measuring durations in speech timing research have often been presented by showing a diagram with stylized segments whose lengths represent averages of the segment durations measured. Often two (or more) diagrams for utterances to be compared are placed side by side with lines connecting corresponding points as in Figure 3.1.

This type of diagram displays a time warp of sorts, albeit a very course-grained one. The corresponding points (or segment boundaries) of the two tokens are the links of our time warping terminology. We could present a fine-grained time warp between two tokens in similar fashion, with many lines connecting linked frames of the two tokens, but with a frame perhaps every 5 msec, the links are so numerous that interpretation becomes difficult, especially if we want to compare different time warps of the same two tokens. In the literature on dynamic time warping a time warp between two tokens is often presented with the x-axis as the time line for one token (often the reference token), the y-axis as the time line for the other (often the test token), and a curve joining the (x, y) pairs of frames connected by the warp. Part of a hypothetical warp between tokens R and S is shown schematically in Figure 3.2.

If the tokens progress at equal rates according to the warp, the curve will follow the $x = y$ diagonal shown in the figure. Departure from the diagonal can be taken to indicate that one token is spending more time in a certain region of its trajectory

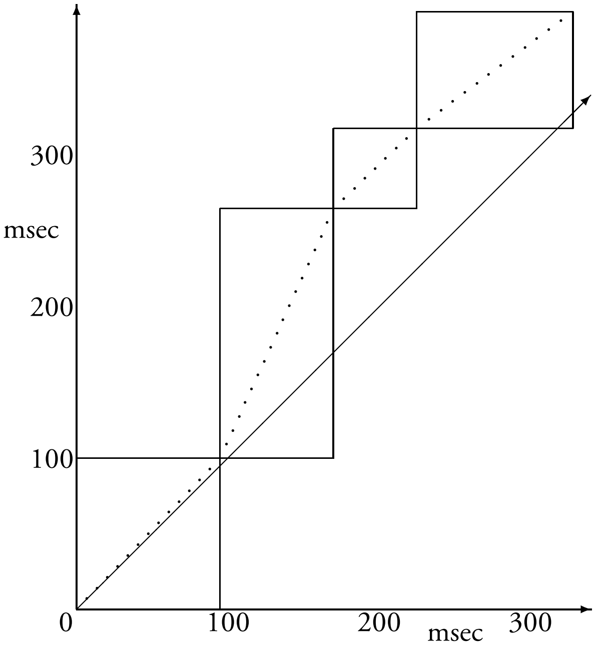


Figure 3.3: Vihanta's duration data shown as a time warp

than the other, ie., there is a "stretching" or "slowing down" relative to the other token. Although theoretically a warp can be thought of as a continuous curve, in practice warps are discrete. In Figure 3.2 this is shown by the small "boxes". The height of a box shows the duration of a frame of the S token and its width shows the duration of the corresponding R frame. Of course, traditional comparisons based on duration measurements can also be shown in this manner, and it is instructive to compare the two. Figure 3.3 displays the data from the upper *tuli*/*tuuli* pair of Figure 3.1 in this format.

If we take into account only segment boundaries, we can determine the relative timing of only a few points in the speech signal. On the other hand DTW enables us (nearly) to follow the continual change in speed of articulation. The finer grained time warp reveals also the relative speed of articulation within segments.

Another representation of a time warp equivalent to the earlier one, but perhaps easier to interpret, shows each of the tokens as a separate curve with measured time elapsed shown on the y-axis for both tokens. The x-axis is then a sort of quasi-time or "phase" axis which is neutral between the tokens being compared. Points of the tokens linked by the time warp are assigned the same x value. This representation of a time warp, in addition to displaying each token as a separate curve, has the advantage of being easily generalized to warps (links) between several tokens—each token is indicated by an additional curve.

The actual x values used are essentially arbitrary, but perhaps the best way to compute position along the x-axis is to take the average of times which correspond to each other in the tokens correlated by the time warp. For instance, if the 42nd frame of one token corresponds to the 54th frame in the other, then the y coordinates of the two curves at these corresponding points will be 42 and 54 (or the equivalent in milliseconds), and the x coordinate for both will be $(42 + 54)/2 = 48$. For convenience sake it is also possible to draw in several vertical reference lines which correspond to familiar landmarks as determined, eg. by inspection of spectrograms. Instead of averaging the time of a point in a token with the time of the corresponding point in the other token to determine the x coordinate of that point, we can just as well average the times of the corresponding points in many speech tokens, provided we have somehow determined all the appropriate links.[1] For m links between n tokens, and letting t_{ij} be the time of the ith token at the jth link, we have

$$x_j = \frac{1}{n}\sum_{i=1}^{n} t_{ij}, \qquad 1 \leq j \leq m \tag{3.2}$$

as the x-axis parameter for the jth link. The advantage is that a time warp, or correspondence, between n tokens can be shown as n curves in two dimensions rather than a single curve in n dimensions. Instead of showing elapsed time (t_{ij}) itself on the y-axis, some space can be saved by showing only the deviation from the average, $y_{ij} = t_{ij} - x_j$. Also, if desired, average curves can be calculated for various groups of tokens as well.

For a time warp between just two tokens, the two curves will be mirror images of each other about the x-axis, and thus for clarity it is preferable to show only one of them. With appropriate scaling, this transformation then amounts to rotating the type of representation shown in Figure 3.2 clockwise by 45 degrees. This is shown in Figure 3.4.

Here the x-axis represents "average time" while the y-axis represents deviation of the R token from the average. For sake of comparison, the same transformation of the segment measurements from Figures 3.1 and 3.3 is shown in Figure 3.5.

[1] How do we determine all the appropriate links between the tokens? At the very least, this means we must calculate $n - 1$ warps for n speech tokens (to insure that each token is compared at least indirectly to all other tokens). However, the more tokens are involved, the less reliable this arrangement will be, since it will depend more and more on the particular pairs of tokens chosen to be compared directly. A more reliable method would be to directly calculate a warp between each token and all other tokens. This of course involves much more computation, since it entails $n(n-1)/2$ time warps for n tokens. This is still much less than the computation required for a single n-dimensional warp. For example, if there are, say, 20 tokens with 60 frames each, then the general 20-dimensional time warp will require calculation of $60^{20} \approx 3.656 \times 10^{35}$ cells, whereas calculating a 2-dimensional warp for each token pair will require calculating $60^2 \times 190 = 684000$ cells. By comparison, ordering tokens and computing warps only for adjacent pairs would involve calculation of $60^2 \times 19 = 68400$ cells.

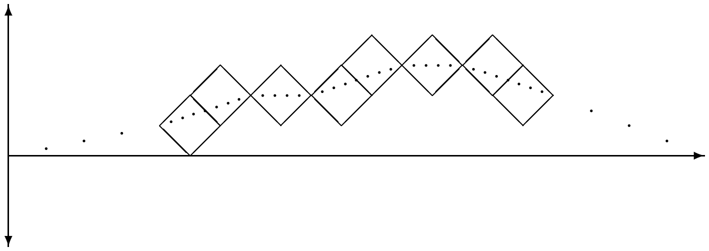


Figure 3.4: Schematic time warp rotated 45 degrees

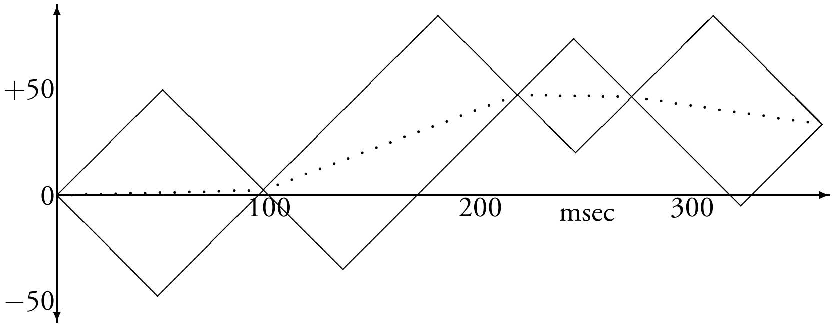


Figure 3.5: Vihanta data in average vs. deviation format

3.1.2 DTW for signal averaging and stimulus synthesis

In spite of several obstacles, dynamic time warping holds much promise for speech timing research. It has the potential of freeing timing analysis from considerations of segment boundary determination since it is, in effect, "applied all over". This means that segment internal timing relations may be taken into consideration in addition to differences in "segment durations".

An important application for DTW in speech timing research is for computing an "average" signal which doesn't merely average parts of the original signals directly, but takes temporal differences into consideration. In other words, by computing a time warp first, we can determine which parts of the original signals best correspond to each other, and average parameters (eg. LPC parameters) according to these correspondences. For instance Strik and Boves suggested using DTW (dynamic programming, in their terminology) for averaging measurements (physiological signals) from multiple repetitions of an utterance (Strik and Boves, 1991).

Dynamic time warping also opens up the possibility of synthesizing a stimulus series whose timing ranges between two real speech tokens. It is possible, for example, to start with two natural speech tokens, for instance *tuli* and *tuuli*, and synthesize a series of stimuli whose timing gradually changes from one to the other, but whose qualitative features represent either of the two original tokens, or some weighted average thereof. In the past such stimulus series have generally been constructed by varying the duration of a single "near steady-state" portion of a speech signal. One advantage when synthesis of stimuli is carried out "along the warp" is that stimuli are generally closer to natural speech since there is no need to prolong steady-state portions. On a theoretical level, there is no need to assume that we can identify *a priori* point landmarks in the speech signal (ie. segment boundaries) which correspond in number to the number of phonemes as revealed by abstract analysis and whose timing alone accounts for the totality of quantity perception. This technique for synthesis of stimuli was used extensively in the experiments reported in the following chapter.

3.2 Effect of various distortion measures

All links established by the DTW procedure are ultimately dependent on the distortion measure used to judge the similarity of frames from the tokens compared. What distortion measure should be used in calculating time warps? We may note in passing that this problem is not unique to DTW. It corresponds to the consideration in traditional segment duration measurement as to what criteria are to be used to identify segment boundaries.

The distortion measure used as the basis for a time warp between two signals can be computed in many different ways, and the difference between distortion measures used will to some extent affect the outcome of the warp itself. To get an idea of the

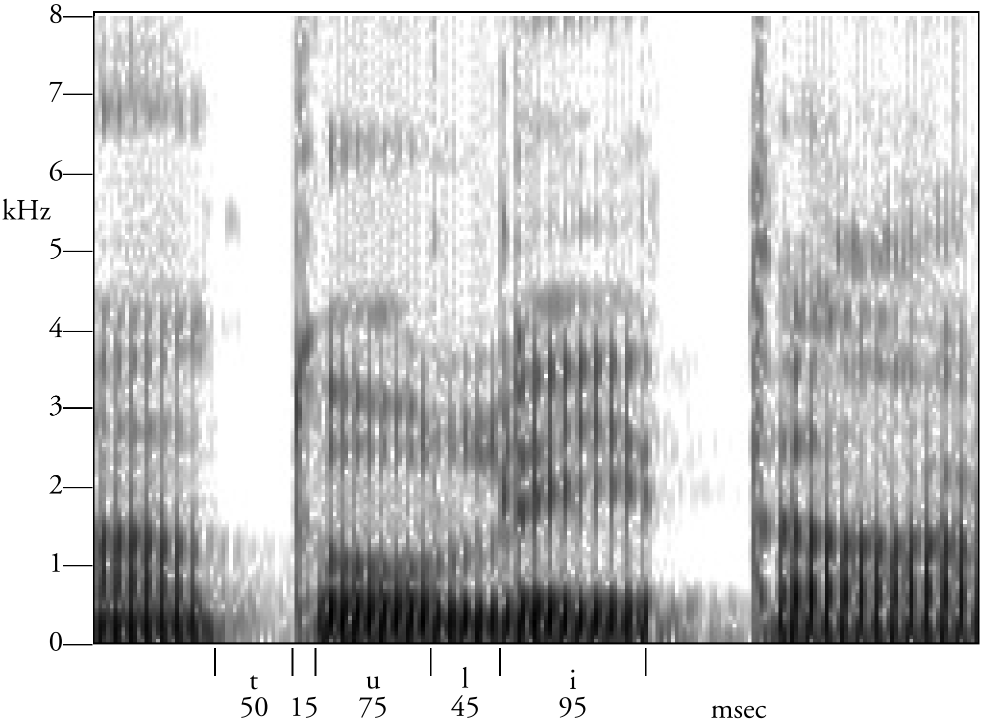


Figure 3.6: Spectrogram of the test word *tuli*

magnitude of these effects, a time warp was computed for tokens of the words *tuli* and *tuuli*, using a variety of different distortion measures.

3.2.1 Methods

One token each of the sentences *Mitä sana tuli tarkoittaa?* and *Mitä sana tuuli tarkoittaa?* recorded for Experiment 2 was transferred in digital form from DAT tape to computer for processing. Stretches of signal centering on the test words *tuli* and *tuuli* (roughly [ɑˈtulitɑ] and [ɑˈtuːlitɑ]) were excised and subjected to dynamic time warping using various distortion measures. Spectrograms of the excised signals are shown in Figures 3.6 and 3.7.

3.2.2 Results

Figure 3.8 shows various warps computed between the *tuli* and *tuuli* tokens superimposed on each other for comparison. The warps shown are as follows:

Normalized energy: The highest point of the energy curve for the each test word was set to 1.0 by multiplying the entire curve by a constant, thus giving a normalized

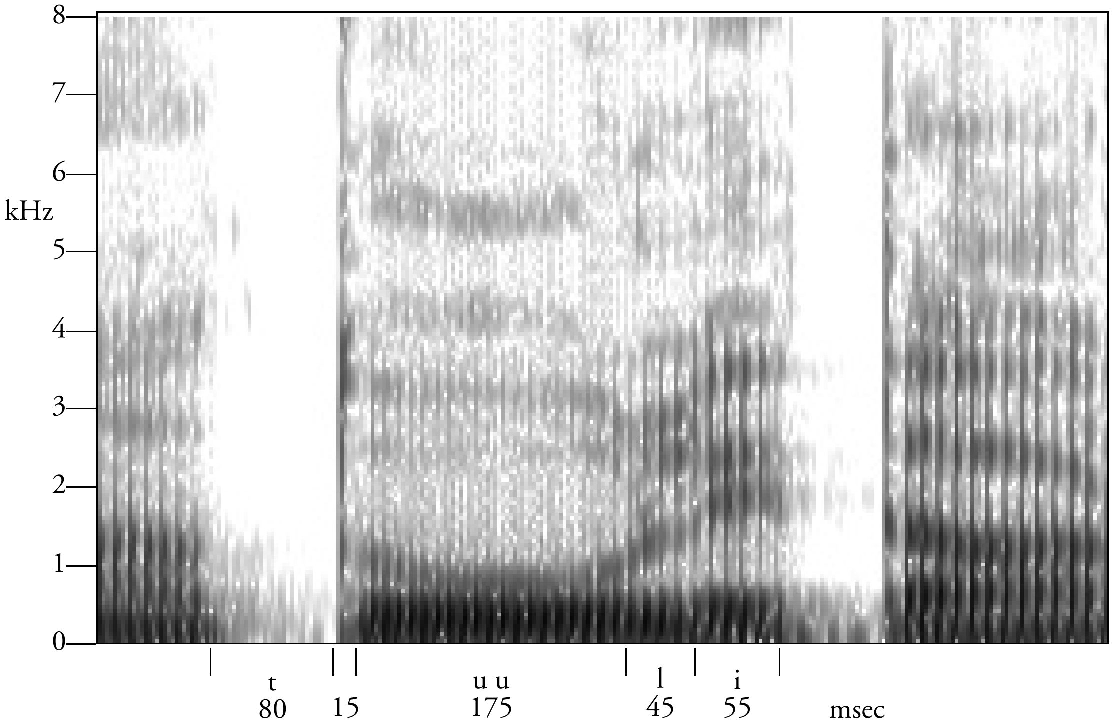


Figure 3.7: Spectrogram of the test word *tuuli*

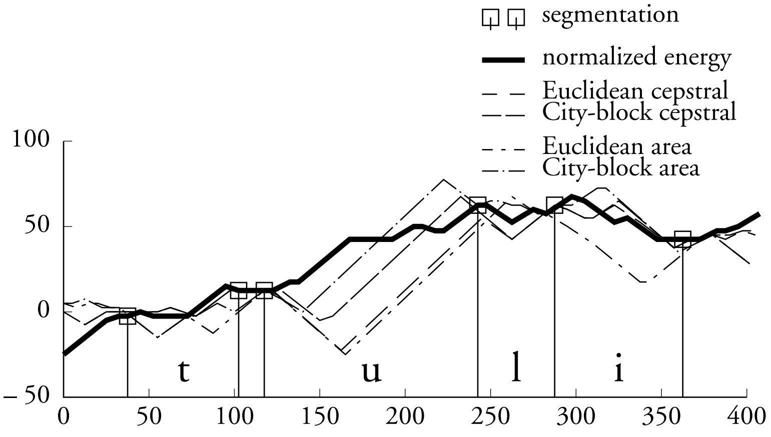


Figure 3.8: Various warps between *tuli* and *tuuli*

energy curve for each token. The absolute difference between normalized energy levels was then used as the distortion measure for the time warp,

$$w(i,j) = |e_i - e_j|\,.$$

Cepstral coefficients: Euclidean:

$$w(i,j) = \sqrt{\sum_{r=1}^{m} (c_{ir} - c_{jr})^2}$$

and so-called "city block", which is just the sum of the absolute differences:

$$w(i,j) = \sum_{r=1}^{m} |c_{ir} - c_{jr}|$$

Area coefficients: Euclidean:

$$w(i,j) = \sqrt{\sum_{r=1}^{m} (\mathcal{A}_{ir} - \mathcal{A}_{jr})^2};$$

city block:

$$w(i,j) = \sum_{r=1}^{m} |\mathcal{A}_{ir} - \mathcal{A}_{jr}|$$

Figure 3.8 also shows the approximate boundaries of the segments [t u l i] for each token as determined by traditional spectrogram inspection techniques (this segmentation is shown in Figures 3.6 and 3.7). There is a great deal of variation in the temporal alignment of the two tokens indicated by time warps calculated using various parameters. Two features of the various warps computed stand out: first of all, most of the differences between various warps occur within the stretch from [u] to [l]. This is perhaps not too surprising, since [u] and [l] are acoustically very similar, and thus different warp calculations are not in agreement as to how (when) [u] changes into [l]. A second result is that the warp based solely on (absolute) difference in normalized energy corresponds best to the segment boundaries found by spectrogram inspection. This warp was used as the basis for synthesis of the stimuli used in the experiments described in the next chapter.

Chapter 4

Perception of quantity differences

> Tule jo, kesäki, kerran,
> talvi, siirräite sivuitse,
> kule, päivä, viere, viikko,
> alene, Jumalan aika.
> Mene, aika, miel'aloissa,
> hoilattele huolissaki.
>
> —Kanteletar

It is well known that by manipulating the measured duration of a segment in a Finnish word it is possible to change the perception of the segment from short to long or vice versa (cf. eg. Lehtonen, 1970). On the other hand, it is also known that the difference between words of different quantity types (eg. *tuli* vs. *tuuli*) is not limited to the duration of a single segment. It is often assumed, however, that the durations of all the segments of the word (or of a one to three syllable foot) taken together form a necessary and sufficient criterion for perception of quantity. This can be regarded as the perceptual side of the duration model (or discrete rubber band model). For the less restrictive rate control model (or continuous rubber band model), the "location in time" of any part of the speech signal (not just segment boundaries) might conceivably affect quantity perception. In any case, such an arrangement requires the postulation of extrinsic timing: since timing is independent of local properties of the speech signal, the listener must "measure time" on his own for purposes of comparison with the speech signal.

4.1 Experiment 3: Effect of quality differences on quantity perception

Assuming some version of rubber band model, it should be possible by suitably distorting the time course of one word to produce a word of a different quantity type. At least for Finnish, this will involve adjusting the timing of the entire word (thus changing the durations of more than one segment, cf. the perception experiments in Lehtonen, 1970, p. 185). It should not, however, require changing the local properties ("quality") of the speech signal.

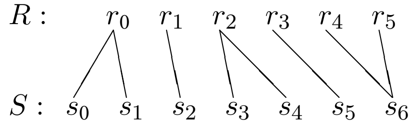
4.1.1 Methods

Stimuli

Two series of stimuli were constructed, based on the natural speech tokens of *tuli* and *tuuli* described in section 3.2.1 (see Figures 3.6 and 3.7). A time warp based on relative signal energy (see section 3.2.2) was used to align the two tokens, and stimuli were (re)synthesized gradually varying the timing in steps from that of the original *tuli* token to the that of the original *tuuli* token.

The computation of parameters from the original tokens used a Hamming window of length 10 msec (441 samples) which was moved ahead 5 msec (220 samples) for each frame (5 msec frame length). The LPC parameters consisted of 70 reflection coefficients and a gain coefficient (alpha) calculated using the autocorrelation method of linear prediction (see Markel and Gray, 1976, pp. 217–219). Pitch period was calculated using a rectangle window and autocorrelation with 3-level center clipping of the signal (see Rabiner and Schafer, 1978, pp. 150–157).

Synthesis was accomplished using a computer program which updates parameters not at a set rate, but at a variable rate according to an additional rate parameter for each parameter frame indicating how long (how many samples) that frame is to last. (Except for the addition of the variable frame rate, the synthesis program was based on the linear prediction synthesizer of Markel and Gray (1976, p. 243). In particular, parameters were continuously interpolated, but updated only at the beginning of a new pitch period for voiced frames.) Once a temporal correspondence between tokens is established (ie. a time warp), the original set of parameter frames can easily be reorganized so that the two tokens have an equal number of frames (of varying duration). For instance, if two 5 msec frames of token A correspond to one 5 msec frame of token B, the 5 msec frame of token B can be split into two (identical) 2.5 msec frames. When the two tokens have an equal number of parameter frames (and timing is expressed as a rate parameter included in each frame), it is an easy matter to form a weighted average between the tokens by taking a weighted average of corresponding frames. To take a simple example, suppose two sets of LPC parameter frames R and S have six and seven frames respectively, and a warp has been computed between them as follows:


R :
r_0 r_1 r_2 r_3 r_4 r_5
S :
s_0 s_1 s_2 s_3 s_4 s_5 s_6

Assume further that the original frame length is 100 samples. In a sense, reparameterization just splits those frames that are multiply connected, adding frame length as a parameter:

$$\begin{array}{lcccccccccl}
R': & r_0 & r_0 & r_1 & r_2 & r_2 & r_3 & r_4 & r_5 & \\
 & 50 & 50 & 100 & 50 & 50 & 100 & 100 & 100 & \text{(frame length)} \\
 & | & | & | & | & | & | & | & | & \\
S': & s_0 & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_6 & \\
 & 100 & 100 & 100 & 100 & 100 & 100 & 50 & 50 & \text{(frame length)}
\end{array}$$

Now if we want to average these two timing schemes, we operate on corresponding frames, for instance the midpoint interpolation for timing (frame length) would be (letting t_{ij} signify the average of frames r_i and s_j):

$$\begin{array}{lcccccccccl}
T: & t_{00} & t_{01} & t_{12} & t_{23} & t_{24} & t_{35} & t_{46} & t_{56} & \\
 & 75 & 75 & 100 & 75 & 75 & 100 & 75 & 75 & \text{(frame length)}
\end{array}$$

This is shown graphically in Figure 4.1. Figure 4.2 shows the corresponding diagram for the stimuli synthesized for the present experiment. In this figure the parameter frames at each end of the continuum are divided into segments for reference based on casual identification of each frame as belonging to one of the phonetic segments assumed to exist in the signal.

Two series of stimuli were thus synthesized for this experiment. Both series contained eleven stimuli ranging in timing from *tuli* (= stimulus 0) to *tuuli* (= stimulus 10). The two series differed, however, in quality (including fundamental frequency and intensity): one series was synthesized using the qualitative parameters calculated from original *tuli*, while the other used the qualitative parameters from original *tuuli*.

After synthesis, it was noticed that the first voiced period of the test word was too loud. This was presumably due to forcing the LPC analysis (and synthesis) to make a binary decision between voiced and voiceless frames, whereas the beginning of voicing may have overlapped with the voiceless burst (V.O.T.) of the initial consonant. Therefore the first period of voicing of all stimuli was systematically reduced in amplitude to 25%. In the future this problem could be avoided by using a more sophisticated analysis and synthesis model such as multi-pulse excitation (cf. eg. Picone *et al.*, 1986) and possibly variable window and frame length in the analysis.

A stimulus tape was prepared by transferring tokens of the synthesized digital signals to DAT tape in random order. The interval between stimuli was set at two seconds, with a longer pause (six seconds) after every tenth stimulus. Each stimulus type appeared on the tape five times giving a total of $5 \times 2 \times 11 = 110$ tokens.

Subjects

The subjects for the present experiment were 12 students at the University of Jyväskylä (subjects 1 and 12 were the subjects SU and KV of Experiment 1, all others had not

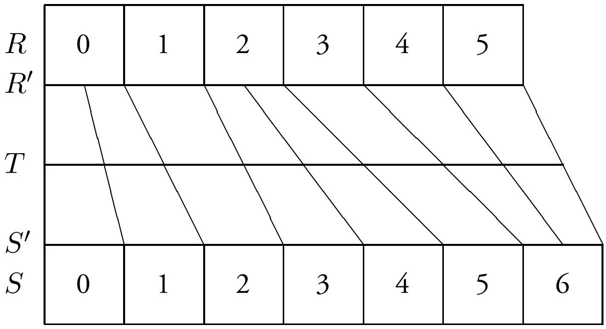


Figure 4.1: Example of interpolation by warping

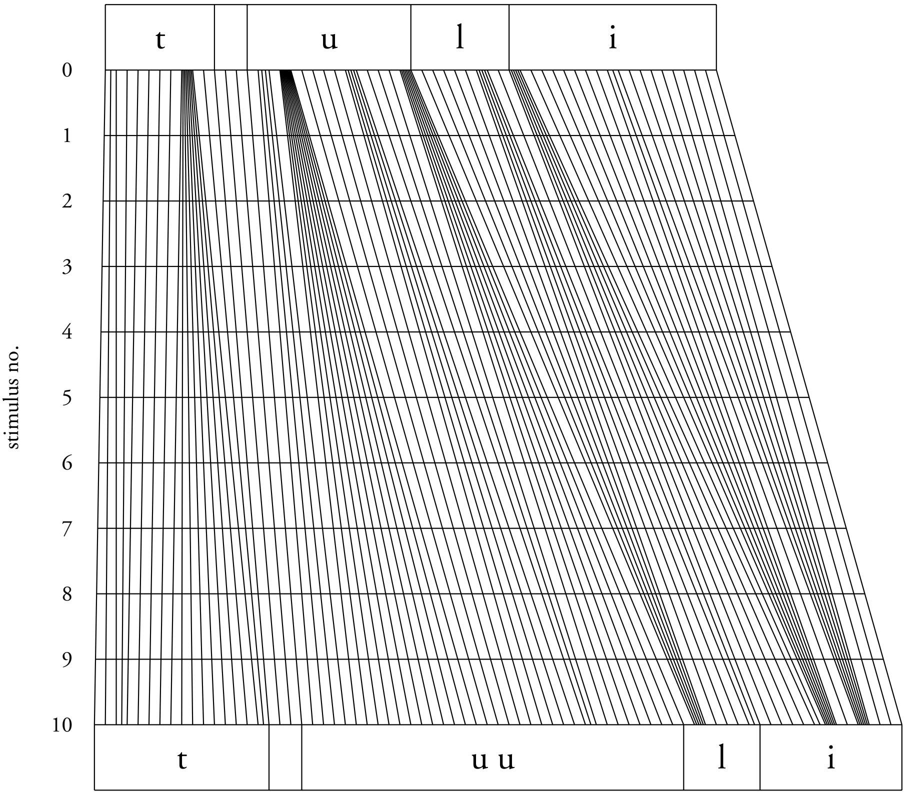


Figure 4.2: Synthesis of stimulus series

participated in the other experiments). All subjects reported being right-handed with normal hearing. Their ages ranged from 23 to 38 years with median age 30 years.

Informants also provided information as to the location in Finland they had spent their childhood. Using Wiik's broad classification of Finnish dialect areas (Wiik, 1975) on the basis of $\mathrm{V}_1/\mathrm{V}_2$ ratios in CVCV type words, five subjects had spent their childhood in the Savo area (subjects 2, 3, 8, 9, 10), three in the Häme area (subjects 4, 11, 12), two in the South-West area (subjects 5, 7), one in both Savo and Häme areas (subject 1), and one in both South-West and North-East areas (subject 6).

4.1.2 Statistical analysis

The results were analyzed using MCMC to estimate the posterior distributions of model parameters, given the data (in what follows these distributions will be summarized using means or median values and credibility intervals).

The model used to analyze the perception results is shown in the form of a directed acyclic graph in Figure 4.3. Each node in this graph represents a quantity in the model, whether a constant fixed by design (represented by rectangles), a stochastic variable associated with a probability distribution, or a deterministic variable which is merely a logical function of other quantities used for convenience. Solid arrows represent stochastic dependence, while dotted arrows represent logical functions. In addition, the graph is greatly simplified by reducing sets of repeated variables (for instance the subject related variables of the present study) to single subscripted nodes enclosed in a large rectangle labeled with the repeated factor.

In the present model (cf. Figure 4.3) n_{ijk} is the number of stimuli with stimulus number i (TIMING, $0 \leq i \leq 10$) in stimulus series j (QUALITY, $j = 1$ representing the *tuli*-series and $j = 2$ the *tuuli*-series) responded to by subject k (SUBJECT, $1 \leq k \leq 12$). Thus for this experiment n_{ijk} was always 5 (with the exception of one case in which a response was left blank giving $n_{ijk} = 4$). On the other hand r_{ijk} is the number of times (out of n_{ijk}) that subject k responded **tuuli** to that same stimulus. According to the model, r_{ijk} is distributed as a binomial random variable with probability p_{ijk}:

$$r_{ijk} \sim \text{Binomial}(p_{ijk}, n_{ijk}) \tag{4.1}$$

that is, it is modeled as the number of successes in n_{ijk} trials, each of which has probability p_{ijk} of succeeding. The probability p_{ijk} itself is modeled as a standard Gaussian (or normal) cumulative distribution function along the TIMING axis (x_i) with mean μ_{jk} and standard deviation σ_{jk}:

$$p_{ijk} = \Phi\left(\frac{x_i - \mu_{jk}}{\sigma_{jk}}\right) \tag{4.2}$$

where Φ here stands for the normal cumulative distribution function. Thus μ_{jk} expresses the "location" of the ogive curve (ie. the 50% cross-over point) along the x-axis

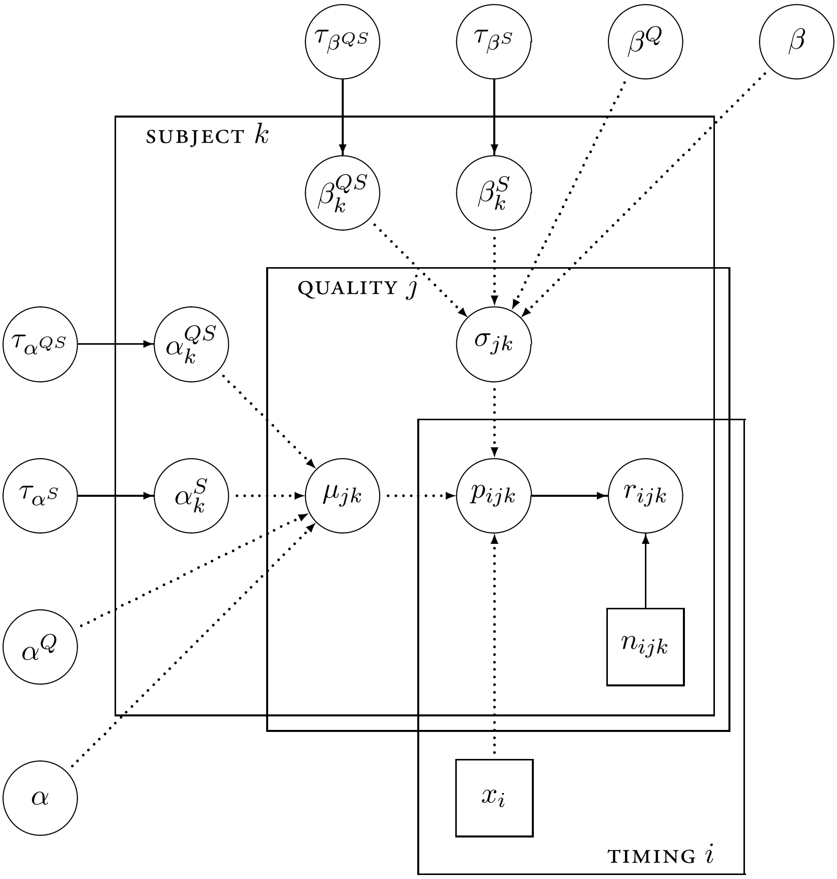


Figure 4.3: Directed acyclical graph of model for Experiment 3

(TIMING) for each subject and quality series, while σ_{jk} relates to the "slope" of the curve, with "steeper" curves corresponding to smaller σ_{jk}. Both μ_{jk} and σ_{jk} are allowed to vary between subjects (SUBJECT parameters α_k^S and β_k^S) and between stimulus series as a whole (QUALITY parameters $\pm\alpha^Q$ and $\pm\beta^Q$, added for the *tuli*-series $j = 1$, subtracted for the *tuuli*-series $j = 2$) as well as for each subject (QUALITY BY SUBJECT interaction parameters $\pm\alpha_k^{QS}$ and $\pm\beta_k^{QS}$):

$$\begin{aligned}\mu_{jk} &= \alpha + \alpha_k^S \pm (\alpha^Q + \alpha_k^{QS})\\ \sigma_{jk} &= \exp(\beta + \beta_k^S \pm (\beta^Q + \beta_k^{QS}))\end{aligned} \tag{4.3}$$

The parameters related to subject are in turn taken to be normally distributed with means of zero (since they are expressed as deviations from the grand population means) and precision parameters τ_{α^S}, $\tau_{\alpha^{QS}}$, τ_{β^S} and $\tau_{\beta^{QS}}$ (*precision* is defined as the inverse of variance):

$$\begin{aligned}\alpha^S &\sim \text{Normal}\,(0, 1/\tau_{\alpha^S})\\ \alpha^{QS} &\sim \text{Normal}\,(0, 1/\tau_{\alpha^{QS}})\\ \beta^S &\sim \text{Normal}\left(0, 1/\tau_{\beta^S}\right)\\ \beta^{QS} &\sim \text{Normal}\left(0, 1/\tau_{\beta^{QS}}\right)\end{aligned} \tag{4.4}$$

where Normal (a, b) denotes a normal distribution with mean a and variance b.

Finally, the founder nodes (nodes with no parents) in the model were assigned vague prior distributions, which were broad enough to support all reasonable values and fairly flat in the region of support:

$$\begin{aligned}\alpha &\sim \text{Normal}\,(5, 10\,000)\\ \alpha^Q &\sim \text{Normal}\,(0, 10\,000)\\ \beta &\sim \text{Normal}\,(0.5, 50)\\ \beta^Q &\sim \text{Normal}\,(0.5, 50)\\ \tau_{\alpha^S} &\sim \text{Gamma}\,(0.01, 0.01)\\ \tau_{\alpha^{QS}} &\sim \text{Gamma}\,(0.01, 0.01)\\ \tau_{\beta^S} &\sim \text{Gamma}\,(1.0, 0.01)\\ \tau_{\beta^{QS}} &\sim \text{Gamma}\,(1.0, 0.01)\end{aligned} \tag{4.5}$$

where Gamma (a, b) denotes a gamma distribution with mean a/b and variance a/b^2.

Area under the ROC. An AUROC was calculated for each subject as

$$A_{z\,k} = \Phi\left(\frac{\mu_{1k} - \mu_{2k}}{\sqrt{\sigma_{1k}^2 + \sigma_{2k}^2}}\right) \tag{4.6}$$

(cf. Hellmich *et al.*, 1998, 1999). This statistic summarizes the difference in the distributions underlying the ogive response curve for the two stimulus series. If these underlying distributions are interpreted as specifying the varying location of a decision threshold along the stimulus axis, then the AUROC can be thought of as giving the probability that the decision threshold for a random perception in the *tuli*-series is further to the right than for a random perception in the *tuuli*-series. That is, it characterizes the discriminability of the two series, even though this was not part of the subjects' task.

Also two values relating to the overall performance of the subjects were calculated: first, an "average" subject AUROC calculated by using the mean values of the subject related parameters, ie.

$$\overline{A_z} = \Phi\left(\frac{\overline{\mu}_1 - \overline{\mu}_2}{\sqrt{\overline{\sigma}_1^2 + \overline{\sigma}_2^2}}\right), \tag{4.7}$$

where $\overline{\mu}_j = \alpha \pm \alpha^Q$ and $\overline{\sigma}_j = \exp(\beta \pm \beta^Q)$, and second, a "total" AUROC A_z^* for this group (calculated as the mean of the individual $A_{z\,k}$) expressing the discrimination performance of the subjects as a group, or the probability of decision thresholds being in the proper order for a random subject in the group.

Convergence and adequacy of MCMC simulation. Actual estimation of parameter distributions was carried out using the WinBUGS program as follows: a 5000 sample "burn-in" was generated with two chains starting from widely differing initial values, after which convergence was checked both visually and using the Gelman-Rubin statistic provided by the WinBUGS program. In the present case convergence appeared to be rather rapid. After this the burn-in period was discarded and an additional 50 000 samples were generated to represent the posterior distribution of parameters. The adequacy of the 50 000 sample run was monitored by checking that the MC error automatically computed by the WinBUGS program for each parameter was less than 5% of the estimated parameter standard deviation. These summary statistics can be used to evaluate the probability of various conditions (or hypotheses) given the model and the data, for instance the hypothesis that a certain parameter differs from zero (significance).

4.1.3 Results

Figure 4.4 shows the percentage of **tuuli** responses for stimuli 0 through 10 for both the *tuli*-series (broken line and triangles in Figure 4.4) and the *tuuli*-series (solid line

stimulus	% **tuuli** responses	
	tuli-series	*tuuli*-series
0	0.0	0.0
1	0.0	1.67
2	0.0	1.67
3	1.67	18.33
4	3.33	58.33
5	31.67	83.33
6	36.67	98.33
7	61.67	98.33
8	66.1	100.0
9	80.0	100.0
10	78.33	100.0

Table 4.1: Responses to stimuli for Experiment 3

and diamonds in Figure 4.4), pooled for all subjects. Table 4.1 shows the same data in numeric form.

Main effects

QUALITY. The mean value for the overall cross-over point (α in the model) was 5.4776 with a 95% credibility interval (CI) of $(4.7234, 6.263)$. More interestingly, the overall QUALITY effect α^Q (which indicates how many units along the TIMING axis the *tuli*-series deviated in the positive direction and the *tuuli*-series in the negative direction from the overall cross-over) had a mean of 1.568 with 95% CI $(0.9363, 2.237)$. The total difference between the two series at the cross-over points was thus twice this amount, ie. the *tuli*-series was roughly three stimulus steps to the right of the *tuuli*-series. Since the CI does not include zero (the value equivalent to no effect), we can conclude that the significance of this effect is $p < 0.05$ (or $p < 0.025$ assuming a one-tailed test).

TIMING. The mean value for the TIMING effect β (relating to the overall slope of the curves) was 0.1311, with 95% CI $(-0.05632, 0.2972)$. This value of β corresponds to a standard deviation for the underlying normal distribution of $\sigma = \exp(\beta) = 1.140$, 95% CI (0.9452, 1.346), ie. roughly one stimulus step.

TIMING BY QUALITY. The TIMING BY QUALITY effect β^Q (expressing the effect of stimulus series on slope) had a mean of 0.3033, 95% CI $(0.1538, 0.4518)$. This means the standard deviation for the *tuli*-series was slightly less than twice the standard deviation for the *tuuli*-series ($\exp\left(2\beta^Q\right) = 1.834$). Here again no effect at all corresponds to a value of zero, which is not included in the CI, so we conclude that the slopes were significantly different ($p < 0.05$) in the two stimulus series.

Average response. Figure 4.5 sums up these results graphically, showing two curves corresponding to mean values for all SUBJECT related parameters (ie. setting $\alpha_k^S = \alpha_k^{QS} = \beta_k^S = \beta_k^{QS} = 0$). It thus represents a sort of "average" subject response. This type of "averaging" is in some ways preferable to pooling results (as in Figure 4.4) because pooling may obscure effects of "location" for different subjects with effects of "slope" for different subjects. For instance if all subjects have steeply sloping curves but widely different cross-over points the pooled curve will rise very slowly, a feature not characterizing any individual subject.

The equations for the two fitted curves of Figure 4.5 are

$$\begin{aligned}
p_j(x) &= \Phi\left(\frac{x-\mu_j}{\sigma_j}\right) \\
&= \Phi\left(\frac{x-(5.4776 \pm 1.568)}{\exp(0.1311 \pm 0.3033)}\right), \quad \text{or} && (4.8) \\
p_1(x) &= \Phi\left(\frac{x-7.0456}{1.544}\right) \quad (\textit{tuli}\text{-series}) && (4.9) \\
p_2(x) &= \Phi\left(\frac{x-3.9096}{0.8418}\right) \quad (\textit{tuuli}\text{-series}). && (4.10)
\end{aligned}$$

The difference between the two curves in Figure 4.4 or between the two curves in Figure 4.5 is very striking.[1] Indeed, the estimated size of the effect of the quality difference (between original *tuli* and *tuuli*) on response (α^Q in the model) is quite large.

Subject related effects

SUBJECT. The model parameters α_k^S, normally distributed about zero with precision τ_{α^S}, allow for different overall cross-over points on the stimulus axis for different subjects. Inspecting the 95% CI for the twelve subject parameters shows four subjects with CI not including zero (subjects 7 and 8 in the positive direction, subjects 10 and 11 in the negative direction), suggesting that the SUBJECT effect was quite important. This is confirmed by the precision parameter, representing the variation in the population from which the subjects were drawn. The precision estimate had mean $\tau_{\alpha^S} = 0.7693$ with 95% CI $(0.2446, 1.631)$ corresponding to a population standard deviation of 1.140 stimulus steps, 95% CI $(0.7830, 2.022)$.

[1] O'Dell (1995) used logistic regression to analyze the data presented here, with likelihood ratio chi-square (L^2) as test statistic. In spite of the serious shortcomings of using this asymptotic statistic applied to the present data, which contains many empirical zeroes, identical conclusions were reached as to the significance of effects.

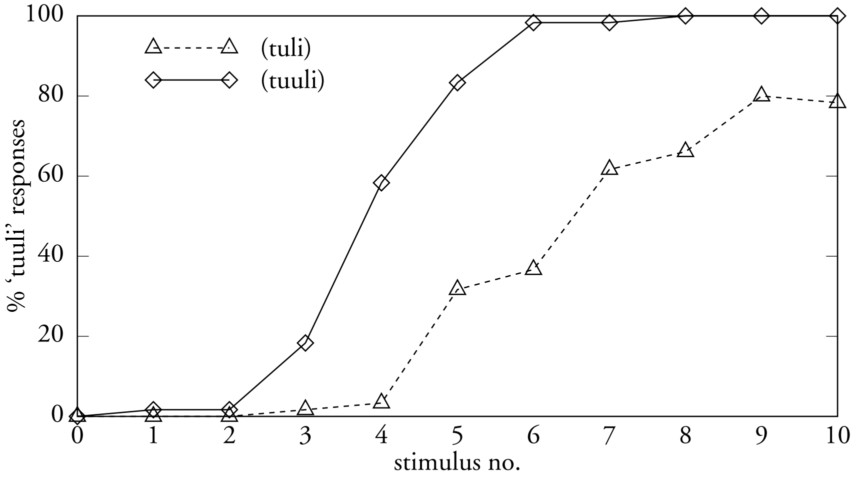


Figure 4.4: Percent **tuuli** responses for Experiment 3

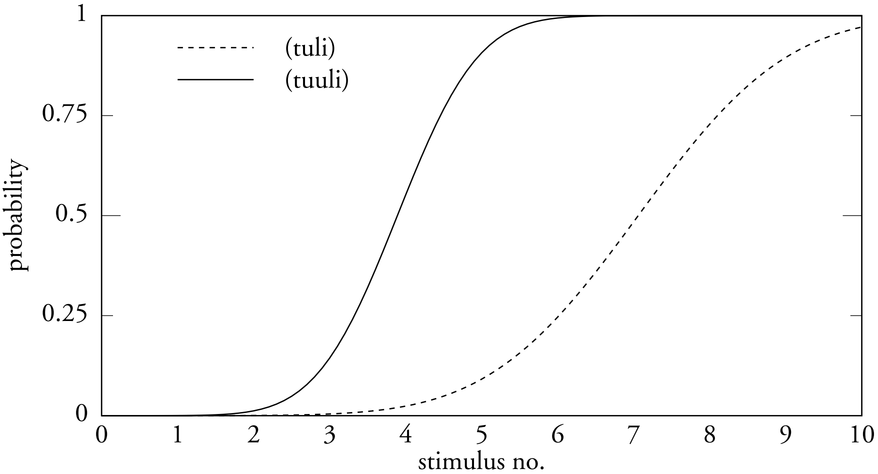


Figure 4.5: "Average" response curves for Experiment 3

TIMING BY SUBJECT. The model parameters β_k^S, normally distributed about zero with precision $\tau_{\beta S}$, modify the slope of the overall ogive curve (or equivalently the variance of the underlying normal distribution) for each subject k. In this case no subject had a 95% CI which excluded zero, indicating that overall intersubject differences in how quickly perception changed from *tuli* to *tuuli* were relatively small. The fairly high precision estimate obtained, mean $\tau_{\beta S} = 64.53$, 95% CI $(6.532, 283.0)$, also indicates relatively little variation of slope in the population. This value corresponds to a population standard deviation of only 0.1245, 95% CI $(0.05944, 0.3913)$, which may be also expressed as a factor $\exp(0.1245) = 1.133$, 95% CI $(1.061, 1.479)$, multiplying or dividing the standard deviation of the overall ogive curve.

QUALITY BY SUBJECT. Model parameters α_k^{QS}, normally distributed with precision $\tau_{\alpha QS}$, express the extent that the difference in the cross-over points for the two stimulus series for subject k deviates from the overall QUALITY effect. The 95% CI for four of the subjects exclude zero (for subject 7 in the positive direction, ie. greater difference between the stimulus series, and for subjects 9, 10 and 11 in the negative direction, ie. smaller difference between stimulus series). It appears that there were important differences between subjects in this respect. This is confirmed by the precision parameter, mean $\tau_{\alpha QS} = 1.112$, 95% CI $(0.3378, 2.450)$, equivalent to a standard deviation of 0.9483 stimulus steps, 95% CI $(0.6389, 1.721)$. Since this parameter expresses the amount each stimulus series deviates in opposite directions from the mean, the effect for the total difference is twice this amount.

QUALITY BY TIMING BY SUBJECT. The model parameters β_k^{QS}, normally distributed about zero with precision $\tau_{\beta QS}$, modify for each subject k the differences in slopes of the ogive curves for the two stimulus series (or equivalently the variances of the underlying normal distributions). In this case no subject had a 95% CI which excluded zero, indicating that the intersubject differences were relatively small. The high precision estimate obtained, mean $\tau_{\beta QS} = 112.8$, 95% CI $(13.91, 371.0)$, also indicates relatively little variation in the population. This value corresponds to a population standard deviation of 0.09416, 95% CI $(0.05192, 0.2681)$, or expressed as a multiplicative factor $\exp(0.09416) = 1.099$, 95% CI $(1.053, 1.308)$.

Evaluating the difference between stimulus series

The median value for the posterior distribution for average subject AUROC was $\overline{A_z} = 0.9614$, with 95% CI $(0.848, 0.9958)$, while the median value for the total AUROC for this subject group was $A_z^* = 0.8914$, with 95% CI $(0.8516, 0.9261)$. These results indicate a very large difference for the two stimulus series, a fact which is already clear from Figure 4.5.

It appears there were significant subject related effects, and in particular the QUALITY BY SUBJECT interaction (α_k^{QS}) was significant, meaning that the size of the QUAL-

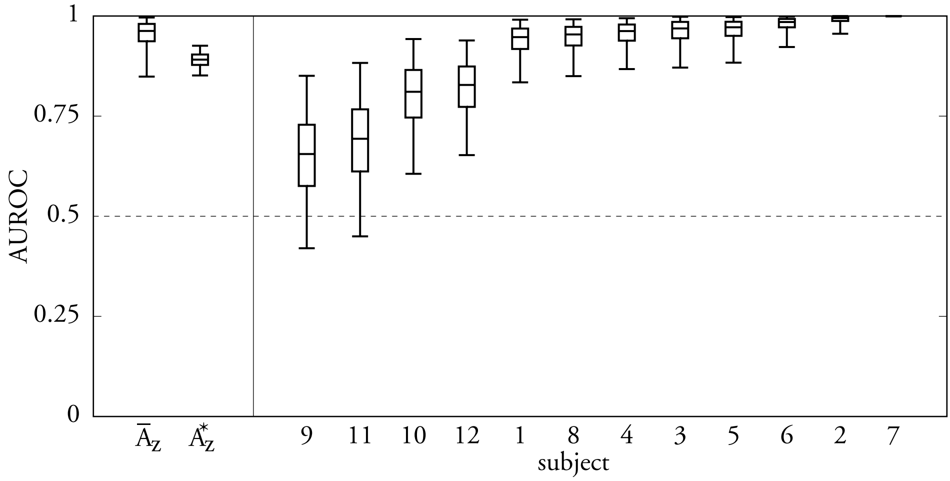


Figure 4.6: AUROC credibility intervals by subject for Experiment 3

ITY effect differed between subjects. Since the question of the extent to which the two stimulus series differed is of major concern in the present study, the difference between stimulus series for individual subjects must be examined more closely. To this end an AUROC value for each subject was monitored during the MCMC run.

Figure 4.6 shows $\overline{A_z}$ and A_z^* as well as the estimated AUROC values $A_{z\,k}$ for individual subjects arranged in order from smallest to largest median values. For clarity the posterior distributions are reduced to five points: the top and bottom crossbars indicate the 95% CI, the central box includes the 50% CI, and the center crossbar indicates the median of the distribution.

It is obvious that differences do indeed exist between subjects. An extreme case was Subject 7, whose entire 95% CI is so close to 100% that the box plot cannot be distinguished in Figure 4.6 (This is a result of the fact that Subject 7 consistently reported hearing *tuli* for the entire *tuli*-series, even though for the *tuuli*-series perception ranged normally from *tuli* at one end to *tuuli* at the other). However it is also noteworthy that for all subjects the median AUROC values are well above the chance level of 50%. In other words, stimuli synthesized from original *tuli* were more likely to be heard as *tuli*, not only on average, but for individual subjects as well.

4.1.4 Discussion

On the basis of these results, it would appear that the rubber band model of quantity timing differences (including both discrete and continuous versions) should be rejected. If quantity distinctions are only a matter of timing differences, then variation in quality (including fundamental frequency and intensity) should be random from one token to the next, and also from one quantity type to the next. In particular, quality differences should have no effect on quantity perception, as they clearly did in the present experiment. Something about the original *tuli* made the stimuli that were based on it more "*tuli*-like" in spite of temporal stretching.

Intensity differences

It is of interest to investigate what the main remaining differences were between the two series of stimuli based on single tokens of *tuli* and *tuuli*. It is logical to start with differences in (relative) intensity, or energy, since this was the basis of the time warp. Figure 4.7 shows the normalized energy for the original tokens on which the stimuli for this experiment were based.

Both curves show evidence of three energy peaks, which we can associate with [u], [l] and [i]. To get a better idea of how corresponding pairs in the two stimulus series compare, we need to align these curves in the same way the time warp does. This will also indicate how well the warp was able to fit these curves together, and show us what differences, if any, are still left. In order to use a time axis which is not entirely arbitrary, we choose the time scale of the "midpoint" stimulus pair *tuli*(5)–*tuuli*(5). This is a convenient choice also because in the perception experiment, a majority (68.33% of responses) heard *tuli*(5) as *tuli*, while a majority (83.33% of responses) heard *tuuli*(5) as *tuuli*. Figure 4.8 shows this time warped version of the normalized energy curves. Of course the time course varies gradually throughout each stimulus series, but the correspondence between the two series remains the same as that shown in Figure 4.8.

There is a very close fit between the curves in Figure 4.8, with very little "residue" or left over differences. This is to be expected since the warp itself was computed on the basis of these curves. Perhaps the main differences still remaining are a higher energy peak associated with [l] for the *tuuli*-series, a lower energy trough between [l] and [i] for the *tuuli*-series, and a slightly higher energy peak associated with [i] for the *tuli*-series. However, it is perhaps unlikely that these small local differences in relative intensity had a significant effect on listeners' perceptions.

Spectral differences

It is well known that Finnish long and short vowels can also exhibit systematic spectral differences. On average, short high and mid vowels are somewhat more centralized than the corresponding long vowels, although the differences in spectral quality between corresponding long and short vowels are not as great as between neighboring vowels (cf. Wiik, 1965, pp. 56–79). The traditional view, however, has been that such qualitative differences have no effect on perception:

> In both languages [ie. Finnish and Estonian] the quality differences between shorter and longer vowels should be taken as an automatic con-

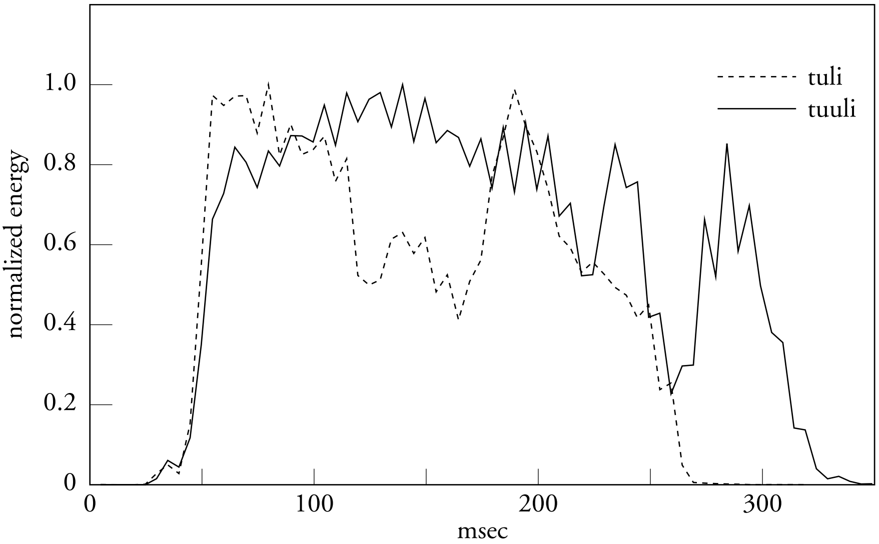


Figure 4.7: Normalized energy for *tuli* (broken line) and *tuuli* (solid line)

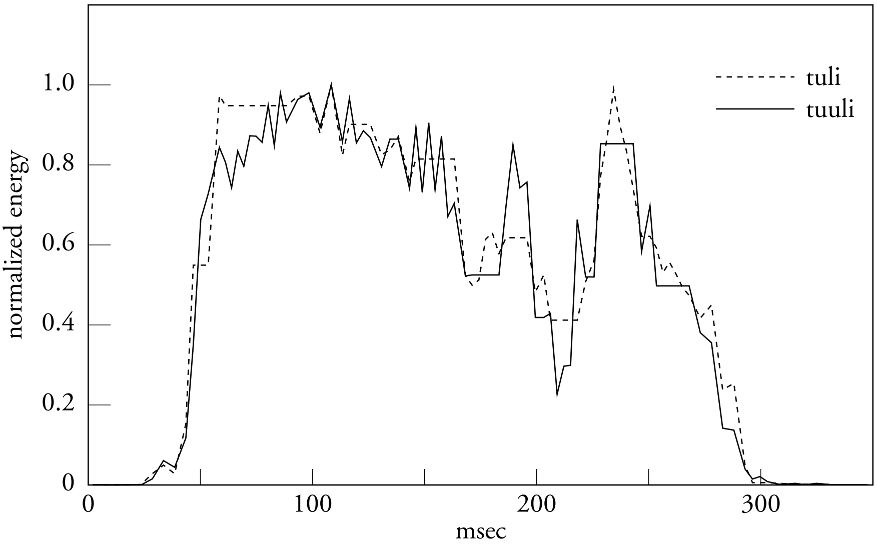


Figure 4.8: Normalized energy aligned according to time warp

> comitant of phonetic characteristics, which plays no decisive role in the perception of quantities (Lehtonen, 1970, pp. 21–22),
>
> . . . a Finnish listener does not make use of the quality difference between vowels as a cue to the phonemic length of the vowel segment as an English listener does. (Lehtonen, 1970, p. 87)

Ravila also notes that

> Of course there are small phonetic differences in addition to purely quantitative ones. Presumably the contact [Finnish *liittymä*] is different in the case of *tuli*, *tulli*, *tuuli*, and there are probably differences in pitch contour as well, but the only thing a speaker of Finnish is clearly aware of is the quantity difference. (Ravila, 1961, my translation)

The so called *close* vs. *open contact* which Ravila refers to (German *fester Anschluss* vs. *loser Anschluss*, Finnish *luja liittymä* vs. *höllä liittymä*) is a concept, which, as Lehtonen pointed out (Lehtonen, 1970, p. 91), has sometimes been invoked in discussions of quantity. The intuitive idea of contact is that a vowel is more "closely joined" to a following consonant when that consonant is geminate (or more generally when that consonant is in the same syllable). Attempts have been made to give contact a concrete interpretation as vowel intensity at the point of (acoustic) consonant onset: close contact would then mean consonant onset before vowel intensity has abated (cf. Sievers, 1893, p. 204). This interpretation, however, has been criticized as only a subjective impression without substance (cf. also the discussion in Sadeniemi, 1949, p. 29, 30).

The traditional position that quality differences are irrelevant for quantity perception is given some support by perception tests administered to Finnish students of English, who, in contrast to native English listeners, had considerable difficulty distinguishing the vowels in synthetic tokens of English *seat* and *sit* when duration was held constant (Raimo and Suomi, 1976).

This traditional view of Finnish researchers on the status of quantity oppositions in Finnish is expressed also by Hakulinen (1968, p. 23, my translation):

> The difference in duration of short and long *vowels* [emphasis in the original] in Finnish is just as great as that of single and geminate consonants: the ratio is usually about 1 : 2 or 1 : 2½, but may be as much as 1 : 3. On the other hand, there is no use whatsoever made of *quality* [emphasis in the original] differences to differentiate short and long vowels in the pronunciation of the educated standard language.

and in a footnote on the same page:

> Robert Harms' remark that e.g. Finnish short *e* is considerably lower than the corresponding long vowel *ee* and that the high long vowels *ii*, *yy*, *uu*

> are somewhat higher than the corresponding short *i*, *y*, *u*, has no phonological significance (*Word*, Vol. 20, 1964). The quality difference of corresponding long and short vowels is clearly greater e.g. in Hungarian, Swedish Swedish, German and English than in Finnish. Cf. Bakó and Sovijärvi *Virittäjä* 1939 p. 386–.

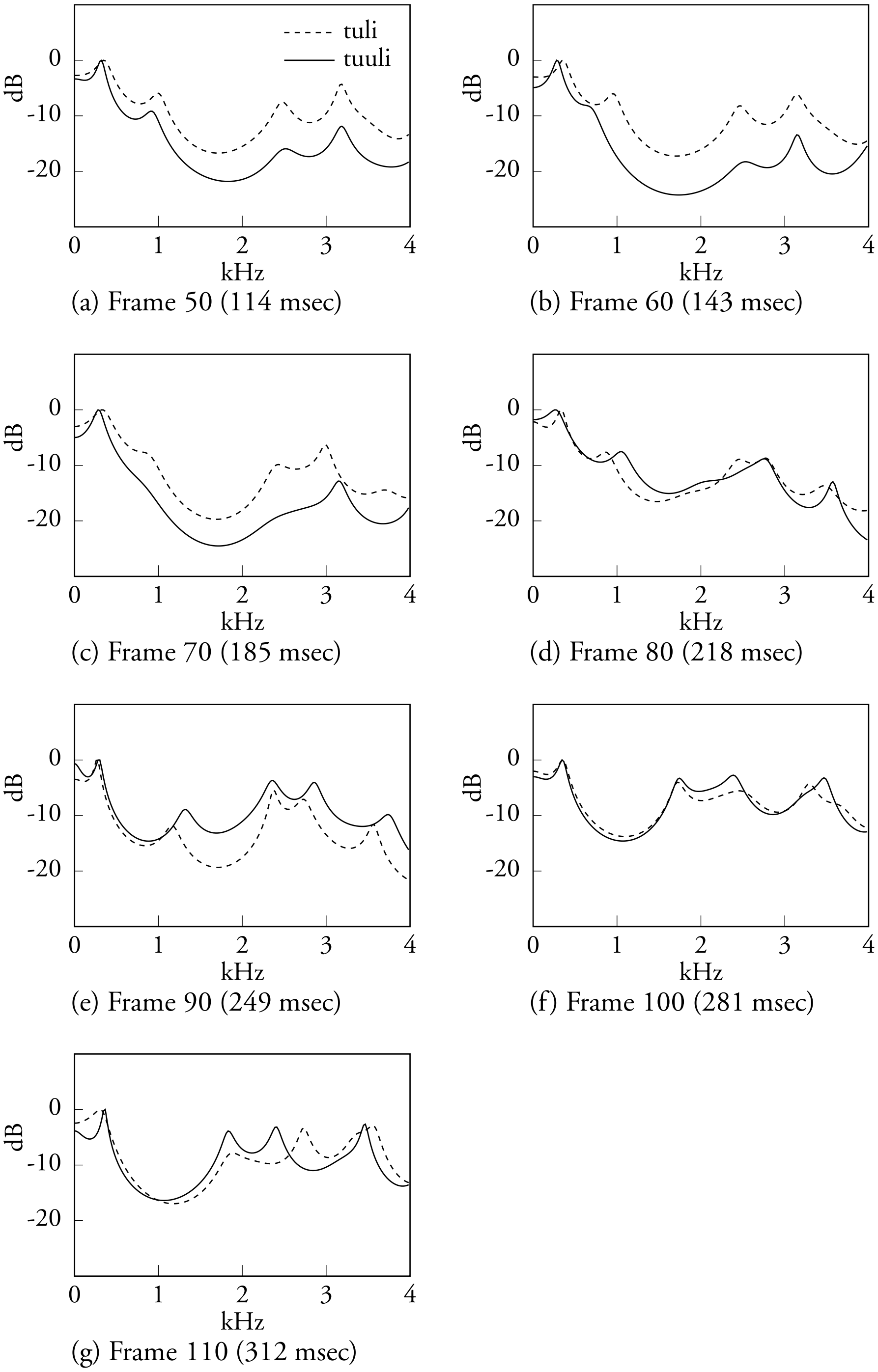
Bakó and Sovijärvi in the study referred to in this footnote compared Hungarian and Finnish high vowels on the basis of X-rays, palatograms and (manually computed!) spectra. They concluded that while there were articulatory differences between short and long high vowels in Finnish, the differences were much smaller than in Hungarian. In Finnish the vowel formants were found to be in the same position, *practically speaking*, for long and short vowels (Bakó and Sovijärvi, 1939).

To answer the question as to whether there were systematic spectral differences in the two stimulus series of the present experiment, we can easily compute the log-spectra of various corresponding synthesis frames as shown in Figure 4.9 a–g[2]. In Figure 4.9 the scale on the vertical axis is in decibels, and for easier comparison, all spectra have been placed so that the highest point is at 0 dB. Each box in Figure 4.9 contains a pair of corresponding spectra for one point along the time warp. The solid line spectra is for the *tuuli*-series, and the broken line spectra for the *tuli*-series. The frame numbers refer to the variable frame rate representation common to both series and correspond to various frames within the voiced portion of the test words. The corresponding duration in milliseconds from the beginning of the word for the midpoint stimuli *tuli*(5) and *tuuli*(5) is given in parentheses.

The most obvious observation to be made from the spectra in Figure 4.9 is that F_1 and F_2 for the *tuli*-series are systematically higher in the earlier spectra (Figure 4.9 a–c). This is perhaps clearest in Figure 4.9 b. In other words, the spectra show a more centralized first vowel (the vowel [u]) for the *tuli*-series, just as expected for a short vowel. The opposite effect, although not as dramatic, can be seen in the last pair of spectra, Figure 4.9 g: here F_1 is higher and F_2 slightly lower, that is, the vowel [i] is more centralized, for the *tuuli*-series. Although not connected in any obvious way with phonological length ([i] in both words represents a phonologically short vowel), this difference is logical if we assume that vowel centralization is directly tied to phonetic duration, since the second syllable vowel in a CVVCV structure ([i] in *tuuli*) is shorter than the second syllable vowel in the corresponding CVCV structure ([i] in *tuli*) by about 50% (cf. Lehtonen, 1970, pp. 107,116). While there are other differences to be seen in Figure 4.9, they are not as easily interpreted.

By solving for the complex roots of the filter coefficients for each frame, the exact locations of the poles of the corresponding spectra can be found. The location of the

[2] These spectra were derived from the reflection coefficients by first converting them into linear prediction filter coefficients using the *step-up* algorithm of Markel and Gray (1976, pp. 94–95), then inverting the all zero spectra obtained from the inverse filter (cf. Markel and Gray, 1976, Chapter 6). The spectra themselves were computed using the Hartley transform (Bracewell, 1986), giving a result equivalent to the Fourier transform for real-valued signals.



(a) Frame 50 (114 msec)

(b) Frame 60 (143 msec)

(c) Frame 70 (185 msec)

(d) Frame 80 (218 msec)

(e) Frame 90 (249 msec)

(f) Frame 100 (281 msec)

(g) Frame 110 (312 msec)

Figure 4.9: Log-spectra for various frames of the test words

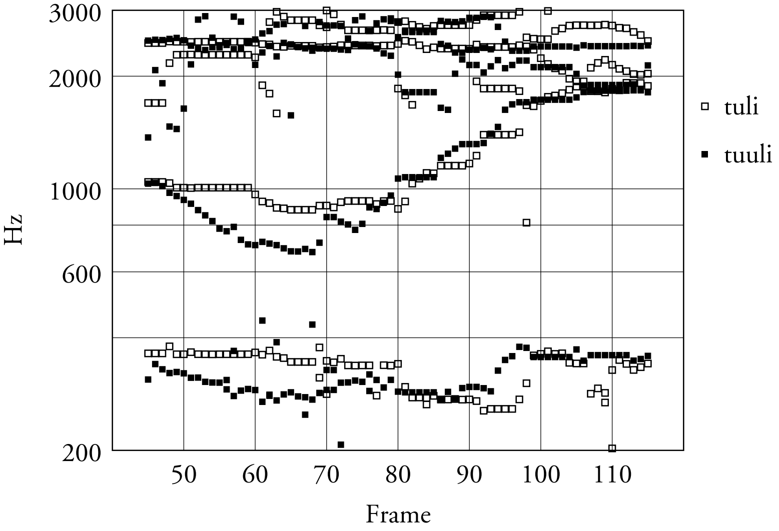


Figure 4.10: Locations of spectrum poles for the two stimulus series

lowest poles for the voiced frames of both the *tuli* and *tuuli* stimuli is shown in Figure 4.10. The frame numbers on the x-axis are the same as in Figure 4.9. Figure 4.11 shows the same data for F_1 and F_2 only connected into a trajectory in the $F_1 \times F_2$ plane showing the changes from [u] through [l] to [i]. In both figures the more extreme values for the first vowel in *tuuli* can be seen clearly. The values of F_1 and F_2 in Hz for the frames shown in Figure 4.9 are presented in Table 4.2.

Figure 4.12 shows spectrograms of the two tokens *tuli*(5) and *tuuli*(5), making it easy to compare the spectral differences between the two stimulus series. As can be seen, in addition to the differences in F_1 and F_2 observed above, there are also marked differences in higher formant structure as well, especially in the during the vowel [u].

Frame no.	(msec)	*tuli*-series		*tuuli*-series	
		F_1	F_2	F_1	F_2
50	(114)	345	991	323	904
60	(143)	366	947	280	624
70	(185)	323	818	280	754
80	(218)	345	861	258	1055
90	(249)	280	1163	301	1314
100	(281)	345	1723	345	1744
110	(312)	301	1895	366	1830

Table 4.2: Formant values for *tuli*- and *tuuli*-series

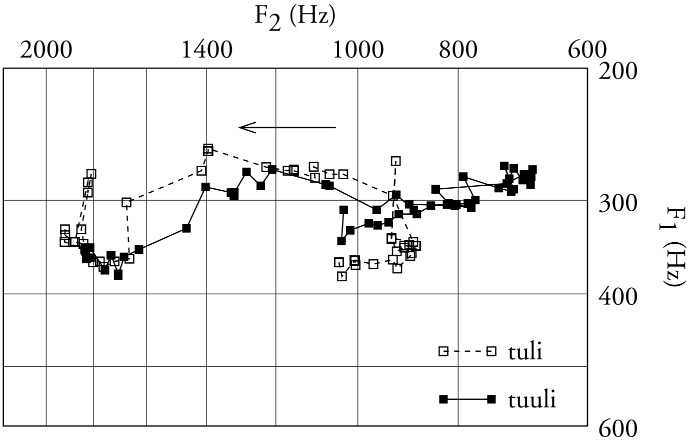


Figure 4.11: Formant trajectories for the two stimulus series

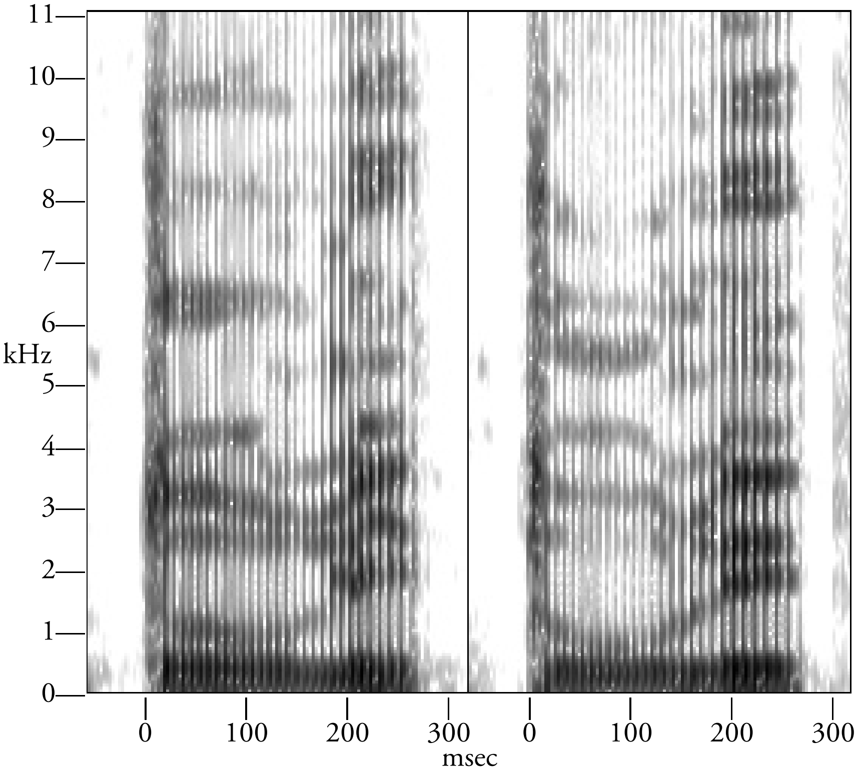


Figure 4.12: Spectrograms of the two tokens *tuli*(5) and *tuuli*(5)

It would appear that F_3 for [u] is relatively stronger in *tuli* than in *tuuli*. Also F_5 drops from [u] to [l] more abruptly and sooner relative to the other formants in *tuli* than in *tuuli*.

Fundamental frequency differences

A further possibility is that a difference in the pitch contours of the two series was responsible for the differences in perception. As Lehtonen points out (Lehtonen, 1970, p. 22), there is some evidence for quantity related differences in pitch pattern for both Finnish and Estonian. For instance Malmberg found that long vowels in Finnish had a falling F_0, while short vowels had a level F_0 after an initial rise (Malmberg, 1949, p. 43–45). However, Lehtonen's measurements of a large number of sentences failed to confirm these results. He did note that pitch patterns for stressed syllables are largely independent of segments and therefore a tonal peak might occur during the stressed vowel if that vowel is long, but not until after the vowel if that vowel is short (Lehtonen, 1970, p. 23). This relationship is in fact confirmed in Wiik's measurements (Wiik, 1988), at least for speakers of some dialects of Finnish.

Figure 4.13 shows the pitch period parameter (converted to Hz) calculated frame by frame from the original tokens of *tuli* and *tuuli*. The time scale shown here is that of the original tokens. These curves also represent the extreme values for the two stimulus series, ie. *tuli*(0) and *tuuli*(10). The fundamental frequency contour appears remarkably similar up to the point where voicing ends in the token *tuli*, after which the token *tuuli* continues on for about 55 msec. Here again, to get a better idea of how corresponding pairs of stimuli compare, it is necessary to show which frames correspond to each other along the time warp used to generate the stimuli. This is shown in Figure 4.14, again using the time scale of the stimulus pair *tuli*(5)–*tuuli*(5). Though timing differs, the correspondence between the two fundamental frequency contours throughout the stimulus series remains the same as that shown in Figure 4.14.

It is easy to see that the time warp used here resulted in a substantial difference in fundamental frequency contour for corresponding stimuli of the two series: Fundamental frequency drops more rapidly in the first part of the stimulus for the *tuuli*-series. This is what we would expect given that change in fundamental frequency is temporally fairly independent of other parameters of the signal as shown in Figure 4.13. With a falling contour this means, for instance, that during a long segment F_0 will have more time to fall, and will thus fall a greater amount, than during a short segment. If this represents a general feature of Finnish, the relation of pitch movement to other (spectral) changes may provide listeners with a fairly reliable cue to quantity differences. In that case, the more pitch changes during a vowel, for instance, the more perception should be biased toward hearing a long vowel, since it would seem that more time has passed. Put another way, it may be conjectured that pitch movement helps to provide the listener with an indication of the speaker's "time line" (ie. helps synchronize listener

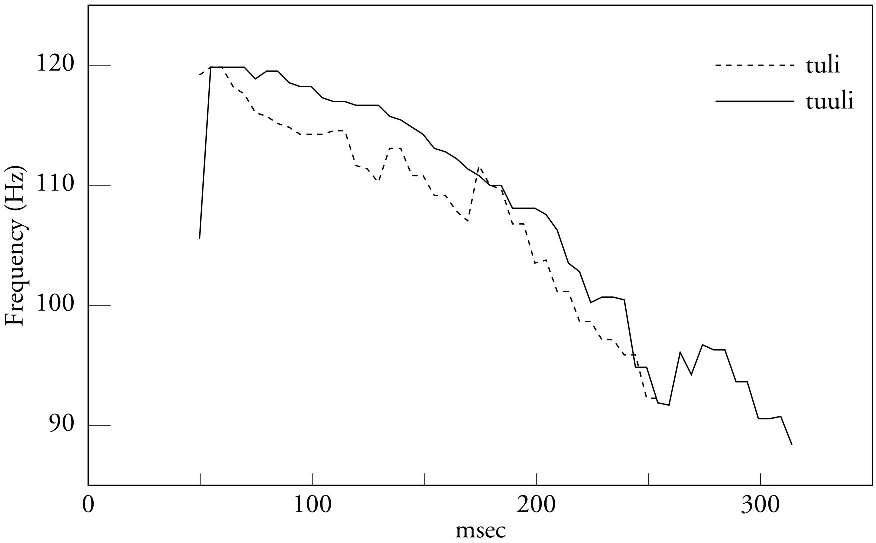


Figure 4.13: Pitch parameter for *tuli* and *tuuli*

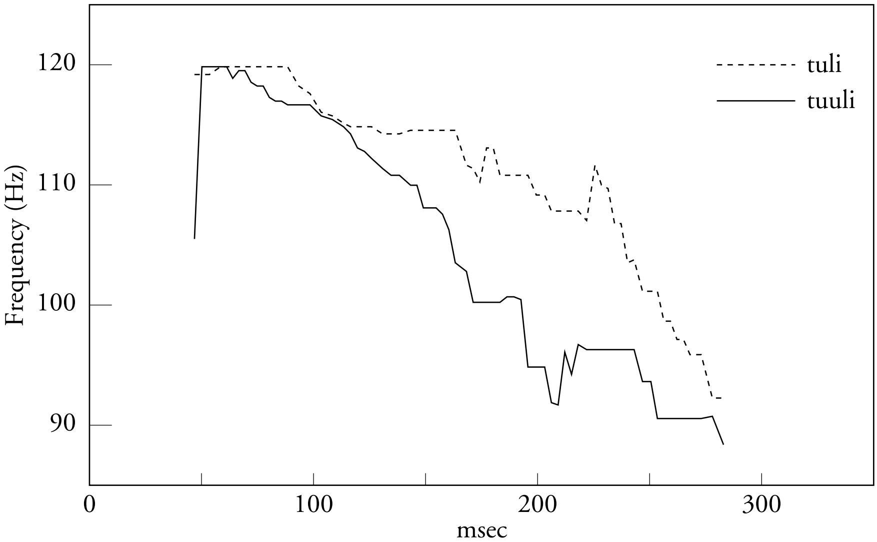


Figure 4.14: Pitch parameter aligned according to time warp

and speaker) for the purpose of making quantity judgments. The results of the present perception experiment are certainly consistent with such a hypothesis.

Vihanta (1988) provides evidence that this state of affairs is indeed quite general in Finnish. He measured fundamental frequency at various points in Finnish test words differing only in phonological length. Although different minimal pairs occurred in different positions within the carrier sentences, and therefore manifest slightly different fundamental frequency contours, for each minimal pair, it would appear that to a first approximation fundamental frequency contour was temporally fairly constant, ie. independent of the quantity difference. Vihanta concludes

> It would seem that a change in F_0, which is determined by factors of sentence prosody, progresses further during a long segment. More often than not this change in F_0 is downward in accordance with typical Finnish intonation. (Vihanta, 1988, p. 34, my translation)

Although Vihanta did not perform perception tests, he does suggest that

> Even though systematic differences in pitch found in conjunction with a quantity opposition could be interpreted as originating in word structure or sentence intonation, this does not prevent F_0 from acting as a cue to recognizing quantity class; ... Similarly F_0 may be an important factor for instance when the durational difference realized is for some reason small (Vihanta, 1988, p. 34, my translation)

Aulanko also noted a difference in the range of F_0 covered by long and short vowels and commented “This might be a feature that is used unconsciously by Finns in perceiving linguistic quantity degrees: a vowel with a relatively short duration but with a great F_0 movement might tend to be perceived as long.” (Aulanko, 1985, p. 48)

Of course these speculations are perfectly consistent with the results of the present experiment. They also accord with the finding of Lehiste (1976) that a changing F_0 in a vowel-like synthetic monosyllable tended to attract more judgments of “longer” compared to monotonic syllables. Naturally it is possible that other differences between the two stimulus series used in the present experiment also influenced perception, either alone or in conjunction with the differences noted here.

4.2 Experiment 4: Effect of fundamental frequency alone

Since both spectral differences and fundamental frequency contour were manipulated in the previous experiment, the question naturally arises whether differences in fundamental frequency alone might also affect perception of quantity types. It is conceivable that F_0 is tied more closely to a different level of timing (eg. sentence intonation), and may thus provide a reference for perception of quantity.

4.2.1 Methods

Stimuli

Two series of stimuli were constructed for this experiment in exactly the same manner as those for the previous experiment, but with one exception. The gain parameter and reflection coefficients used for both series consisted of values averaged between the original *tuli* and *tuuli* parameters. Thus the only difference which remained between the two series was in fundamental frequency (pitch period). The pitch parameter for each series was identical to that of the previous experiment (cf. Figure 4.14 on page 78). Each series contained eleven stimuli ranging in timing from *tuli* (= stimulus 0) to *tuuli* (= stimulus 10).

A stimulus tape was prepared with five tokens of each stimulus in random order giving a total of $5 \times 2 \times 11 = 110$ tokens. The interval from the beginning of one stimulus to the beginning of the next was set at 3.5 sec, with an additional pause of 3.5 sec after every tenth stimulus.

Subjects

The subjects for the present experiment were 93 students of an introductory phonetics class at the University of Jyväskylä, none of whom had participated in the previous experiments. Nine subjects were excluded because they reported that they were not monolingual speakers of Finnish, or had spent their childhood outside of Finland. Two were excluded because they reported not having normal hearing. Data from two subjects were excluded on the basis of their actual responses: one answered in apparently completely random fashion while the other responded with *tuli* to all stimuli. All together responses of 80 subjects were taken into consideration. The ages of these subjects ranged from 18 to 31 with median age 20.

According to the information provided by the subjects themselves and using Wiik's broad classification of Finnish dialect areas (Wiik, 1975), 44 subjects had spent their childhood in the Savo area (11 of these in Jyväskylä), 22 in the Häme area, 5 in the South-West area, 2 in the South-East area, 2 in both Savo and Häme areas, and 1 in both Savo and South-East areas (4 subjects responded with a location term too broad to allow classification).

4.2.2 Results

Figure 4.15 shows the percentage of **tuuli** responses for stimuli 0 through 10 for both the *tuli*-series (broken line and triangles in Figure 4.15) and the *tuuli*-series(solid line and diamonds in Figure 4.15), pooled for all subjects. Table 4.3 shows the same data in numeric form.

Statistical analysis was carried out in the same manner as for the previous experiment using MCMC techniques and the model described in section 4.1.2 and illustrated in Figure 4.3.

stimulus	% **tuuli** responses	
	tuli-series	*tuuli*-series
0	0.0	0.0
1	0.0	0.25
2	0.5	0.5
3	2.5	16.25
4	11.5	40.0
5	44.0	67.25
6	79.25	87.0
7	93.23	98.5
8	97.24	94.75
9	99.0	98.5
10	98.75	98.75

Table 4.3: Responses to stimuli for Experiment 4

Main effects

QUALITY. The mean value for the overall cross-over point was $\alpha = 4.8606$ with a 95% CI of $(4.6993, 5.02751)$. The overall QUALITY effect[3], indicating how many units along the TIMING axis (using the scale shown in Figure 4.15) the *tuli*-series deviated in the positive direction and the *tuuli*-series in the negative direction from the overall cross-over, had a mean of $\alpha^Q = 0.3795$ with 95% CI $(0.2931, 0.4674)$. The total difference between the two series at the cross-over points was thus twice this amount, ie. the *tuli*-series was roughly 0.8 unit to the right of the *tuuli*-series. Since this CI does not include zero (the value equivalent to no effect), we conclude that this effect is significant ($p < 0.05$), or significantly greater than zero ($p < 0.025$) using a one-tailed test.

TIMING. The mean value for the TIMING effect, indicating the overall slope of the curves, was $\beta = -0.01911$, with 95% CI $(-0.1183, 0.07819)$. This value of β corresponds to a standard deviation for the underlying normal distribution of $\sigma = \exp(\beta) = 0.9811$, 95% CI (0.8884, 1.0813), ie. roughly one stimulus unit.

TIMING BY QUALITY. The TIMING BY QUALITY effect, expressing the effect of stimulus series on slope, had a mean of $\beta^Q = -0.123$, 95% CI $(-0.1883, -0.05954)$. This means the standard deviation for the *tuli*-series was approximately 4/5 the standard deviation for the *tuuli*-series ($\exp\left(2\beta^Q\right) = 0.7819$). Here again no effect at all corresponds to a value of zero, which is not included in the CI, so we conclude that the

[3] We continue to refer to this as the QUALITY effect for compatibility with the previous experiment. Note however, that here this pertains only to differences in the F_0 contour.

slopes were significantly different ($p < 0.05$) in the two stimulus series, even though the magnitude of the effect is fairly small.

Average response. Figure 4.16 sums up graphically the results for the main effects, showing two curves corresponding to mean values for all SUBJECT related parameters. The equations for the two fitted curves of Figure 4.16 are

$$\begin{aligned}
p_j(x) &= \Phi\left(\frac{x-\mu_j}{\sigma_j}\right) \\
&= \Phi\left(\frac{x-(4.8606 \pm 0.3795)}{\exp(-0.01911 \mp 0.123)}\right), \quad \text{or} \qquad (4.11) \\
p_1(x) &= \Phi\left(\frac{x-5.2401}{0.8675}\right) \quad (\textit{tuli}\text{-series}) \qquad (4.12) \\
p_2(x) &= \Phi\left(\frac{x-4.4811}{1.1095}\right) \quad (\textit{tuuli}\text{-series}). \qquad (4.13)
\end{aligned}$$

Subject related effects

SUBJECT. The model parameters α_k^S, normally distributed about zero with precision τ_{α^S}, allow for different overall cross-over points on the stimulus axis for different subjects. Inspecting the 95% CI for the 80 subject parameters shows 37 subjects with CI not including zero (16 subjects in the positive direction, 21 in the negative direction), suggesting that the SUBJECT effect was quite important. This is confirmed by the precision parameter, representing the variation in the population from which the subjects were drawn. The precision estimate had mean $\tau_{\alpha^S} = 1.981$ with 95% CI $(1.35, 2.77)$ corresponding to a population standard deviation of 0.7105 stimulus units, 95% CI $(0.6008, 0.8607)$.

TIMING BY SUBJECT. The model parameters β_k^S, normally distributed about zero with precision τ_{β^S}, modify the slope of the overall ogive curve (or equivalently the variance of the underlying normal distribution) for each subject k. In this case 12 subjects had a 95% CI which excluded zero (1 subject in the positive direction, 11 in the positive direction), suggesting the existence of moderate overall intersubject differences in how quickly perception changed from *tuli* to *tuuli*.

The mean precision estimate obtained for the TIMING BY SUBJECT effect was $\tau_{\beta^S} = 8.199$, 95% CI $(4.91, 12.98)$, also indicating moderate variation of slope in the population. This value corresponds to a population standard deviation of 0.3492, 95% CI $(0.2776, 0.4513)$, which may be also expressed as a factor $\exp(0.3492) = 1.418$, 95% CI $(1.320, 1.570)$, multiplying or dividing the standard deviation of the overall ogive curve.

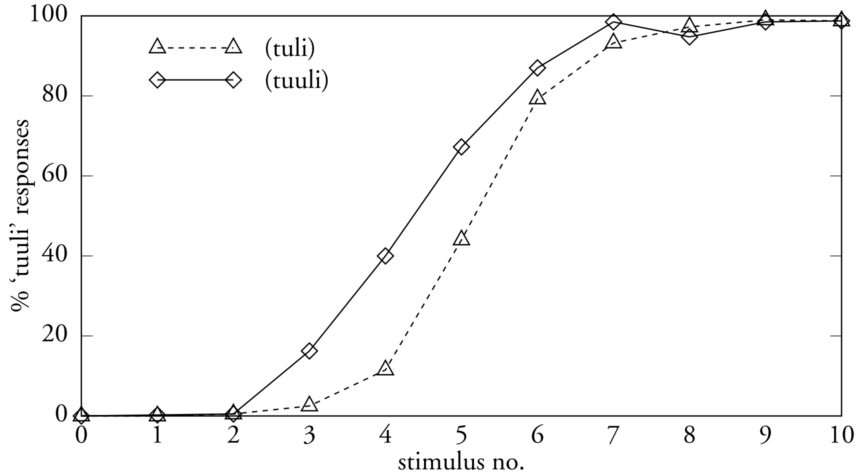


Figure 4.15: Percent **tuuli** responses for Experiment 4

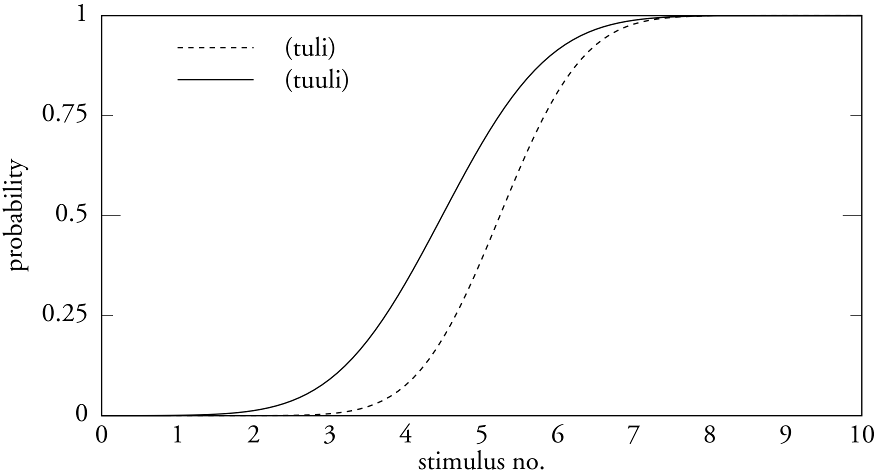


Figure 4.16: "Average" response curves for Experiment 4

QUALITY BY SUBJECT. Model parameters α_k^{QS}, normally distributed with precision $\tau_{\alpha^{QS}}$, express the extent that the difference in the cross-over points for the two stimulus series for subject k deviates from the overall QUALITY effect. A total of 7 subjects had a 95% CI excluding zero, (2 in the positive direction, 5 in the positive direction), indicating only small intersubject variation as to how far apart the cross-over points were for the two series.

This is confirmed by the precision estimate obtained, with mean $\tau_{\alpha^{QS}} = 10.88$, 95% CI $(6.338, 18.0)$, equivalent to a standard deviation of 0.3032 stimulus units, 95% CI $(0.2357, 0.3972)$.

QUALITY BY TIMING BY SUBJECT. The model parameters β_k^{QS}, normally distributed about zero with precision $\tau_{\beta^{QS}}$, modify for each subject k the differences in slopes of the ogive curves for the two stimulus series (or equivalently the variances of the underlying normal distributions). In the present case only one subject had a 95% CI which excluded zero (in the positive direction), indicating that the intersubject differences were relatively small. The high precision estimate obtained, mean $\tau_{\beta^{QS}} = 43.0$, 95% CI $(15.59, 122.1)$, also indicates relatively little variation in the population. This value corresponds to a population standard deviation of 0.1525, 95% CI $(0.09050, 0.2533)$, or expressed as a multiplicative factor $\exp(0.1525) = 1.165$, 95% CI $(1.095, 1.288)$.

Comparison of stimulus series

The difference between the curves in Figure 4.15 or between the two curves in Figure 4.16 is not as large as for the previous experiment (compare Figures 4.4 and 4.5), but it seems quite clear nonetheless. In order to assess the extent to which the two stimulus series differed for individual subjects, as well as over all, an AUROC value was monitored for each subject in the MCMC run, along with two summary statistics, average subject AUROC $\overline{A_z}$ and total AUROC A_z^*, as in the previous experiment.

Figure 4.17 shows the estimated AUROC values $A_{z\,k}$ for individual subjects arranged in order from smallest to largest median values. For clarity, in this figure the quantiles of the posterior distributions for separate subjects are joined by lines instead of showing a separate box for each subject: the top and bottom broken lines show the 2.5% and 97.5% quantiles (indicating the 95% CI), the thinner solid lines show the 25% and 75% quantiles (indicating the 50% CI), while the thicker solid line in the middle indicates the median of the distribution.

As in the case of the previous experiment, is obvious that differences do indeed exist between subjects. For several subjects the difference between the two series is minimal. In fact three subjects (31, 34, 44) actually show a posterior AUROC distribution with median less than 0.5; for several others the CI includes 0.5 even though their median is greater than 0.5. Nonetheless, median $\overline{A_z} = 0.7045$, 95% CI $(0.6589, 0.751)$, median $A_z^* = 0.6876$, 95% CI $(0.6615, 0.713)$, indicating that a stimulus from the

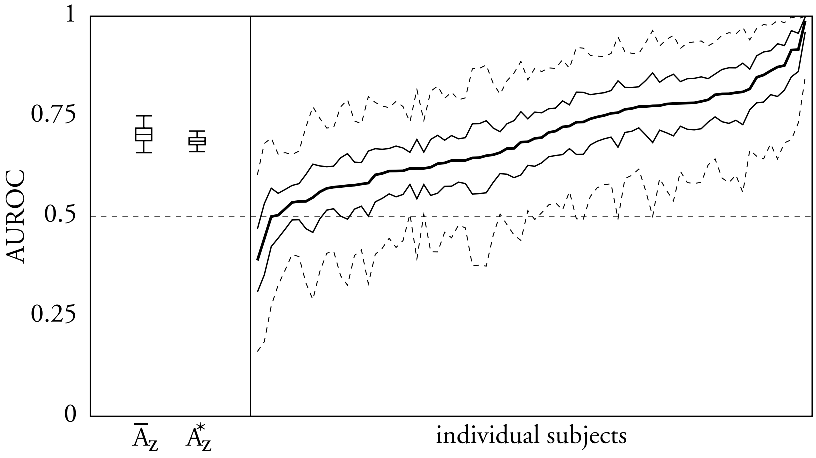


Figure 4.17: AUROC credibility intervals by subject for Experiment 4

tuuli-series was more likely to be perceived as *tuuli* than the corresponding stimulus from the *tuli*-series, both for an average subject, and for the subject group as a whole.

Since there were considerably more subjects in the present experiment, an attempt was made to correlate the intersubject differences with information on the response forms about the location where the subjects spent their childhood. However, differences did not appear to fall into any geographical pattern.[4]

4.2.3 Discussion

The previous experiment left open the question as to whether the differences in perception were due to pitch differences or to spectral differences, both of which were found to exist in the two stimulus series used for that experiment. In the present experiment, F_0 was the only difference between the stimulus series, spectral differences having been averaged out. The fact that the remaining difference in F_0 had a significant effect on perception supports the conclusion that a simple rubber band model is not adequate. If quantity was only a matter of stretching or compressing the time axis, any differences in F_0 should be random and thus would not affect quantity perception.

For the previous experiment there exists the theoretical possibility that the systematic difference in perception reflected a failure of the time warp to correlate proper sections of the original tokens. Such a failure could result in stimulus series whose timing, as indicated by the progression of spectral quality, was not identical. No such interpretation is possible, however, for the present results since all spectral differences have been averaged out. A poor time warp could only result in degradation of the spectral

[4] In the future, however, it might be worthwhile to attempt to assess possible regional effects using various spatial distribution methods available in GeoBUGS, an additional module for the WinBUGS program (cf. Thomas *et al.*, 2002, and references therein).

quality of both series. If this were the case, and quantity perception was based solely on timing of spectral progression, we would expect more uncertainty in responses for the present experiment. In fact no such effect was observed, as can be seen, for example, by comparing the “slopes” of the response curves in Figures 4.5 and 4.16.

That the difference between the *tuli*-series and the *tuuli*-series was considerably smaller than in the previous experiment suggests that pitch alone does not totally account for the difference in perception observed for the two stimulus series in that experiment. One would then expect that averaging out the F_0 differences and keeping the spectral differences should also produce a substantial effect on quantity perception.

A further difference compared with the previous experiment was that the F_0 difference appeared to affect a majority of the listeners but not all of them. Two factors should be considered here. First of all, it may be simply that since the pitch difference alone was a much weaker effect, a much larger number of perceptions is needed for each listener before a statistical effect is apparent for all. On the other hand, if the previous experiment had included as many subjects as the present experiment, there might well have been subjects failing to show a statistical difference. In both experiments differences between subjects were extremely significant. It must be remembered that real language communities are far from being homogeneous and there can be no necessary or sufficient phonetic requirements for belonging to one.

4.3 Experiment 5: Perception of Finnish stop quantity

In the previous two experiments evidence was found that perception of quantity type was affected by various other factors such as vowel quality and fundamental frequency in addition to “pure” timing. The next experiment is a replication changing as many parameters as possible in order to begin to get an idea of the generality of this phenomenon.[5] For the previous experiments a quantity minimal pair was purposefully chosen which was known to exhibit relatively large quality differences (/uu/ vs. /u/). The minimal pair chosen for the present experiment was, on the contrary, expected to show little if any qualitative differences: the words contained no high vowels and the quantity opposition involved a short vs. geminate voiceless stop.

4.3.1 Methods

Stimuli

The stimuli for the perception test were based on real speech tokens of the test words *katoa* ‘crop failure (partitive case)’ and *kattoa* ‘roof (partitive case)’ excised from sentences *Ei ollut sellaista katoa / kattoa täällä ennen nähty*. ‘Such a crop failure / roof had not been seen here before.’ spoken by a female native speaker of Finnish living

[5] This experiment was reported on previously in O’Dell (1999). In the following the statistical analysis has been redone using MCMC techniques as in the previous sections.

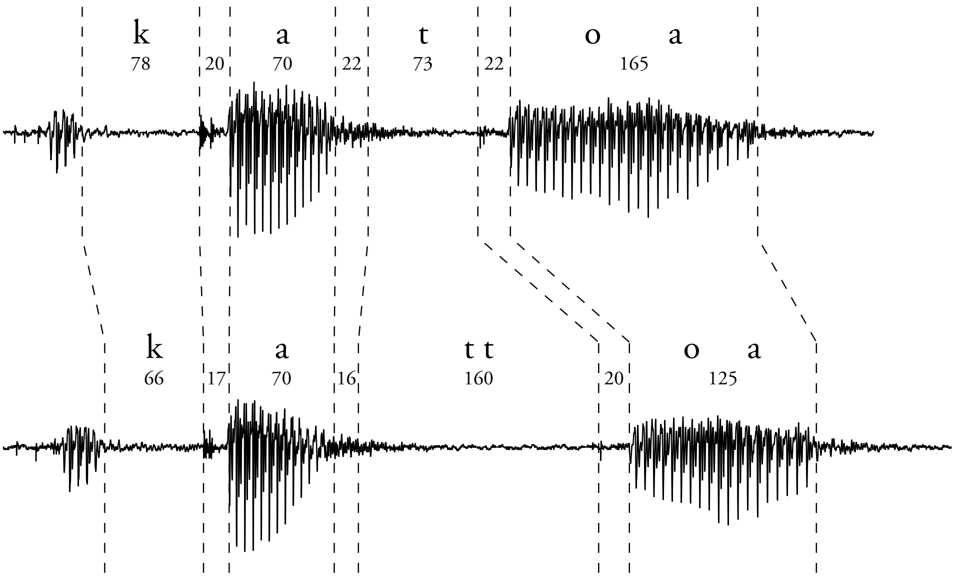


Figure 4.18: Waveforms and segmentation for the tokens used as the basis for construction of stimuli

in the Tampere region. One token each of the test words was selected at random and digitized and analyzed using LPC methods. A dynamic time warp was then computed between the two tokens for the purpose of constructing stimulus series.

One difference in relation to the earlier experiments, and indeed to DTW in general, was that the time warp was computed not for fixed intervals, but by dividing the signal into the individual pitch periods for voiced portions (pitch synchronous analysis) and into equal sections for unvoiced sections, always respecting the *a priori* segmentation shown in Figure 4.18, which was carried out on the basis of the waveforms and spectrograms. The time warp itself was computed for each segment pair using a distortion measure based on simple cumulative absolute differences ("city-block") in the autocorrelation coefficients.

Once a time warp was calculated, various versions of the test words were (re)synthesized using LPC synthesis. Copies of the original test words were synthesized as well as stimuli intermediate between the two by using appropriately weighted averages of the variable frame lengths for the LPC parameters. Two stimulus series were synthesized in this manner, one using the LPC parameters from the original word *katoa*, the other using parameters from the original word *kattoa*. These will be referred to as the *katoa*-series and the *kattoa*-series, respectively. Within each series only the timing (variable frame length) was gradually changed to correspond to original *katoa* at one extreme and to *kattoa* at the other. It can thus be reasonably asserted that one series represented *katoa* "qualitatively" while the other represented *kattoa* "qualitatively".

stimulus	% **kattoa** responses	
	katoa-series	*kattoa*-series
0	0.0	2.5
2	5.833	10.084
4	39.496	50.0
6	90.0	90.833
8	96.667	99.167
10	99.167	100.0

Table 4.4: Responses to stimuli for Experiment 5

If linear timing (or durational) differences formed the only distinguishing feature for these words, it would thus be expected that perceptions within both series should change from *katoa* to *kattoa*, and furthermore that there should be no systematic differences between the two series. Each series was composed of six stimuli (numbered 0, 2, 4, 6, 8, 10 for compatibility with experiments 3 and 4) in equal steps and each stimulus was presented 10 times giving a total of 120 stimuli, which were presented to subjects in random order.

Subjects

The listeners for the present experiment were 11 students and one teacher at the University of Tampere, none of which had participated in the other experiments. All subjects reported being right-handed with normal hearing. Their ages ranged from 20 to 44 years with median age 22½ years.

Classified according to Wiik's broad Finnish dialect areas (Wiik, 1975), eight subjects reported having spent their childhood in the Häme area (subjects 1, 2, 4, 5, 6, 7, 10, 12), three in the Savo area (subjects 8, 9, 11), and one in the South-West area (subject 3).

4.3.2 Results

Figure 4.19 shows the percent **kattoa** responses pooled for all subjects. The broken line represents responses to the *katoa*-series and the solid line responses to the *kattoa*-series. Table 4.4 shows the same data in numeric form.

Although differences between individual subjects are not visible in Figure 4.19, it appears that the two series do differ systematically: a stimulus in the *katoa*-series was heard as **kattoa** less often than the corresponding stimulus in the *kattoa*-series.

Statistical analysis was carried out in the same manner as for the previous experiments using MCMC techniques and the model described in section 4.1.2 and illustrated in Figure 4.3.

Main effects

QUALITY. The mean value for the overall cross-over point was $\alpha = 4.1545$ with a 95% CI of (3.612, 4.7218). The overall QUALITY effect, indicating how many units along the TIMING axis (using the scale shown in Figure 4.4) the *katoa*-series deviated in the positive direction and the *kattoa*-series in the negative direction from the overall cross-over, had a mean of $\alpha^Q = 0.1542$ with 95% CI $(-0.02069, 0.3388)$. The total difference between the two series at the cross-over points was thus twice this amount, ie. the *katoa*-series was roughly 0.3 unit to the right of the *kattoa*-series. Since this CI does include zero (the value equivalent to no effect), this effect is not significantly different from zero ($p > 0.05$) using a two-tailed test, although it is quite close to significance at the 5% level. However if we wish to test the expectation that the effect is *greater* than zero, we examine the 5% quantile of the posterior distribution, which in the present case was slightly greater than zero (the 90% CI was $(0.007813, 0.3058)$), indicating a significant effect $p < 0.05$ assuming a one-tailed test.

TIMING. The mean value for the TIMING effect, indicating the overall slope of the curves, was $\beta = 0.1606$, with 95% CI $(-0.1647, 0.4493)$. This value of β corresponds to a standard deviation for the underlying normal distribution of $\sigma = \exp(\beta) = 1.175$, 95% CI $(0.8481, 1.567)$, ie. roughly one stimulus unit.

TIMING BY QUALITY. The TIMING BY QUALITY effect, expressing the effect of stimulus series on slope, had a mean of $\beta^Q = -0.04203$, 95% CI $(-0.1905, 0.104)$. This means that the slopes for the two series (standard deviations) were practically identical. No effect at all of stimulus series on slope corresponds to a value of zero, which is almost centered in the CI, so we may conclude that the slopes were not significantly different ($p > 0.5$) in the two stimulus series.

Average response. Figure 4.20 sums up graphically the results for the main effects, showing two curves corresponding to mean values for all SUBJECT related parameters. The equations for the two fitted curves of Figure 4.20 are

$$\begin{aligned}
p_j(x) &= \Phi\left(\frac{x-\mu_j}{\sigma_j}\right) \\
&= \Phi\left(\frac{x-(4.1545 \pm 0.1542)}{\exp(0.1606 \mp 0.04203)}\right), \quad \text{or} && (4.14) \\
p_1(x) &= \Phi\left(\frac{x-4.3087}{1.1259}\right) \quad (\textit{katoa}\text{-series}) && (4.15) \\
p_2(x) &= \Phi\left(\frac{x-4.0003}{1.2246}\right) \quad (\textit{kattoa}\text{-series}). && (4.16)
\end{aligned}$$

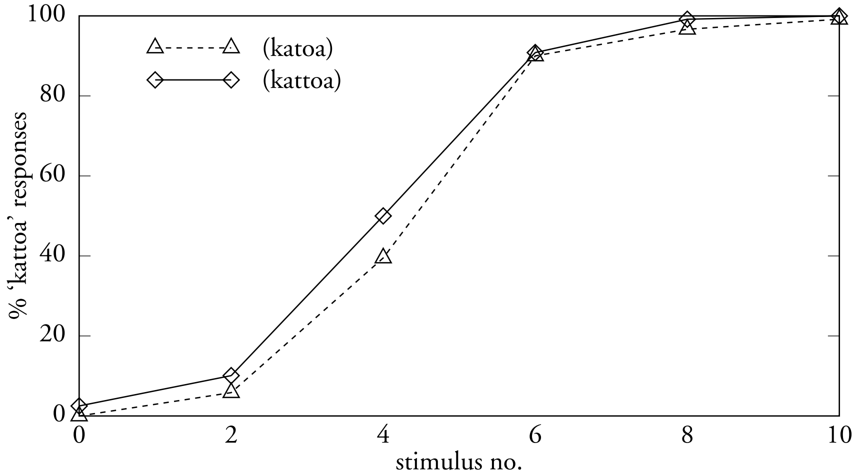


Figure 4.19: Raw pooled percentages of **kattoa** responses

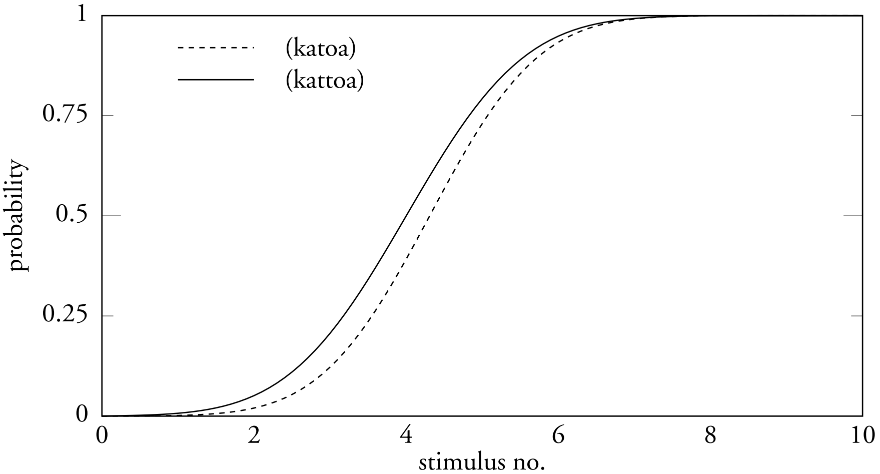


Figure 4.20: "Average" response curves for *katoa*-series and *kattoa*-series

Subject related effects

SUBJECT. The model parameters α_k^S, normally distributed about zero with precision τ_{α^S}, allow for different overall cross-over points on the stimulus axis for different subjects. Inspecting the 95% CI for the twelve subject parameters shows five subjects with CI not including zero (subjects 3, 7 and 9 in the positive direction, subjects 4 and 5 in the negative direction), suggesting that the SUBJECT effect was quite important. This is confirmed by the precision parameter, representing the variation in the population from which the subjects were drawn. The precision estimate had mean $\tau_{\alpha^S} = 1.52$ with 95% CI $(0.4722, 3.365)$ corresponding to a population standard deviation of 0.8111 stimulus units, 95% CI $(0.5451, 1.455)$.

TIMING BY SUBJECT. The model parameters β_k^S, normally distributed about zero with precision τ_{β^S}, modify the slope of the overall ogive curve (or equivalently the variance of the underlying normal distribution) for each subject k. In this case 2 subjects had a 95% CI which excluded zero (subjects 7 and 8, both in the positive direction, corresponding to less steep ogive), suggesting the existence of moderate overall intersubject differences in how quickly perception changed from *katoa* to *kattoa*.

The mean precision estimate obtained for the TIMING BY SUBJECT effect was $\tau_{\beta^S} = 6.764$, 95% CI $(1.94, 16.48)$, also indicating moderate variation of slope in the population. This value corresponds to a population standard deviation of 0.3845, 95% CI $(0.2463, 0.7180)$, which may be also expressed as a factor $\exp(0.3845) = 1.469$, 95% CI $(1.279, 2.050)$, multiplying or dividing the standard deviation of the overall ogive curve.

QUALITY BY SUBJECT. Model parameters α_k^{QS}, normally distributed with precision $\tau_{\alpha^{QS}}$, express the extent that the difference in the cross-over points for the two stimulus series for subject k deviates from the overall QUALITY effect. No subject had a 95% CI excluding zero, indicating negligible intersubject variation as to how far apart the cross-over points were for the two series.

This is confirmed by the fairly high precision estimate obtained, mean $\tau_{\alpha^{QS}} = 66.17$, 95% CI $(7.668, 251.4)$, equivalent to a standard deviation of only 0.1229 stimulus units, 95% CI $(0.06307, 0.3611)$.

QUALITY BY TIMING BY SUBJECT. The model parameters β_k^{QS}, normally distributed about zero with precision $\tau_{\beta^{QS}}$, modify for each subject k the differences in slopes of the ogive curves for the two stimulus series (or equivalently the variances of the underlying normal distributions). In this case no subject had a 95% CI which excluded zero, indicating that the intersubject differences were relatively small. The high precision estimate obtained, mean $\tau_{\beta^{QS}} = 111.1$, 95% CI $(13.34, 377.0)$, also indicates relatively little variation in the population. This value corresponds to a pop-

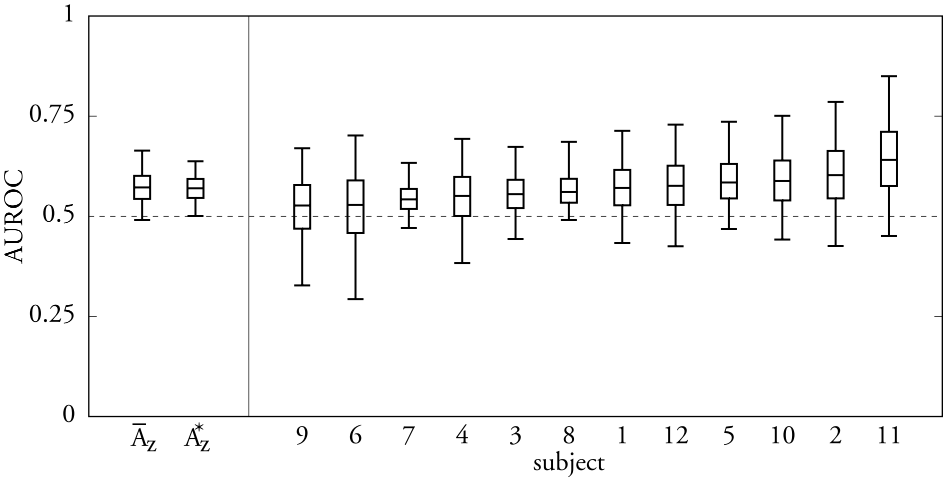


Figure 4.21: AUROC credibility intervals by subject for Experiment 5

ulation standard deviation of 0.09487, 95% CI $(0.05150, 0.2738)$, or expressed as a multiplicative factor $\exp(0.09416) = 1.100$, 95% CI $(1.053, 1.315)$.

Evaluating the difference between stimulus series

As for the previous experiments the difference between stimulus series was evaluated with the help of AUROC values monitored during the MCMC run. Figure 4.21 shows the average subject AUROC $\overline{A_z}$ and total AUROC A^*_z along with the estimated AUROC values $A_{z\,k}$ for individual subjects arranged in order from smallest to largest median values. As in the previous figures, the top and bottom crossbars indicate the 95% CI, the central box includes the 50% CI, and the center crossbar indicates the median of the posterior distribution.

Average subject AUROC had a median value of $\overline{A_z} = 0.572$ with a 95% CI $(0.4902, 0.6641)$, and total AUROC had a median value $A^*_z = 0.5697$ with 95% CI $(0.5001, 0.6372)$. The 95% CI for $\overline{A_z}$ includes the chance value 0.5, but the 5% quantile for $\overline{A_z}$ was 0.5037, which means that the probability that $\overline{A_z} < 0.5$ is $0.025 < p < 0.05$. In the case of A^*_z, this probability is $p < 0.025$, since the 2.5% quantile is (just slightly) above 0.5. It is clear that these values are much lower than those obtained for the previous two experiments as expected, on average as well as for subjects taken individually. What is unexpected is the appearance of a (small) trend even in this “worst case” situation.

4.3.3 Discussion

Evidently there were other factors besides “pure timing” which affected at least a majority of perceptions for a majority of listeners. Something about the original word

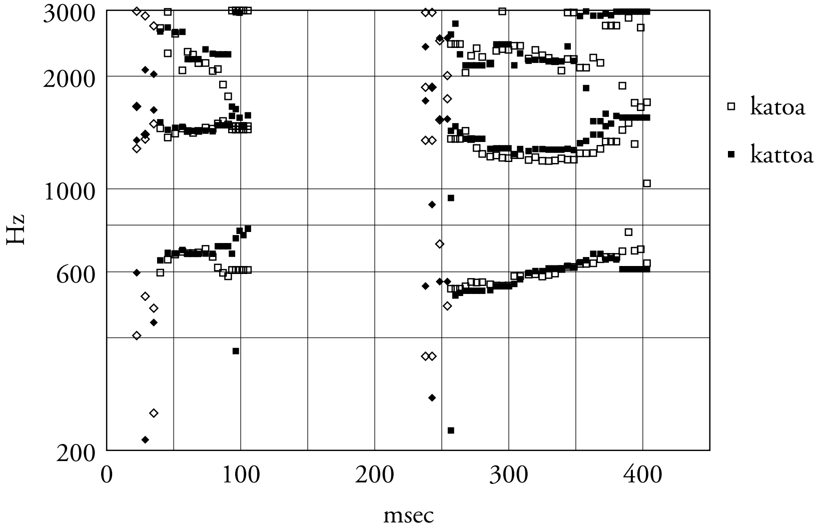


Figure 4.22: Locations of spectrum poles for *katoa*-series and *kattoa*-series

katoa made the stimuli which were based on it sound a bit more like *katoa*, in spite of the manipulation of the time axis. It is therefore of interest to examine what differences remained between the two stimulus series after dynamic time warping has leveled the timing differences.

We first examine the formant structure of the stimuli. Figure 4.22 shows the lower spectrum poles of each series, computed directly from the LPC parameters. There is at least one clear difference visible here in the F_2 of the [oɑ] at the end of the word—F_2 dips a little further down in the *katoa*-series before it rises. This is probably related to the fact that [oɑ] in the original word *katoa* is relatively long compared to [oɑ] in the word *kattoa*. Since this is likely to be true quite generally in Finnish, the greater movement of the formant could act as a cue to the word's quantity type. There are also slight differences in the vowel [ɑ] of the first syllable, particularly in F_1 and F_3 at the end. These differences are more difficult to interpret, but they might be related to a different transition or linking from vowel to consonant in the two quantity types (calling to mind the traditional concept *contact* / *Anschluss* / *liittymä*, cf. eg. Ravila, 1961; Lehtonen, 1970; cf. the discussion in section 4.1.4, p. 72). Another possible explanation is that the [ɑ] of the first syllable is simply closer to the upcoming vowel [o] in the case of single [t]. In this case there could conceivably be more overlapping of the labial gesture for the upcoming rounded vowel, causing the formants to drop slightly at the end of the first [ɑ] of *katoa*. This is consistent with Lehtonen's (1979) measurements of lip movements going from unround to round vowels. According to Lehtonen's Figure 3, average onset of lip movement occurred somewhat before stop occlusion in cases with a single medial stop (eg. [itu]) as opposed to being approximately simultaneous with stop onset for geminate stops (eg. [ittu]). Needless to say, this could also provide a partial perceptual cue to stop quantity, provided the phenomenon is general enough in Finnish speech.

We next examine variation in F_0. In Figures 4.23 and 4.24 the dotted line shows the movement of fundamental frequency for the *katoa*-series, while the solid line shows the *kattoa*-series. The timing of the original tokens is used in Figure 4.23, while Figure 4.24 shows the curves aligned according to the time warp as in corresponding stimuli of the two series. On the basis of previous studies (eg. Vihanta, 1988) one might expect there to be a greater fall in F_0 across a geminate stop compared with a short stop simply because the end of a geminate consonant would correspond to a later phase in the independent (falling) intonation curve. In that case such a difference could act as a cue to the quantity of the stop. However, no such difference is visible in the present case—with the exception of the first period, the F_0 curves after the stops are almost identical in the two series as seen in Figure 4.24. Of course there is a marked difference in the first period of voicing after the stop—it is much shorter (ie. higher F_0) for the word *kattoa*. If this is a general state of affairs, it could conceivably provide a cue to quantity type. There is some evidence that the glottis may be more open during voiceless geminate stops in Finnish (Iivonen, 1975). A raised F_0 could well be a consequence of this difference. For instance many researchers have pointed out differences in F_0 corresponding to stop voicing in many diverse languages to the effect that voicelessness (open glottis) tends to raise F_0 (cf. eg. Lehiste and Peterson, 1961; Hombert *et al.*, 1979).

A possible difference in glottal state might also explain a small difference in intensity visible in Figure 4.25, which shows the changes in the gain parameter of the LPC analysis for the two series, aligned according to the time warp as in Figure 4.24. It appears the explosion after the geminate [tt] is somewhat weaker than after single [t]. However it is unclear just how a difference in glottal opening could result in a weaker explosion while raising F_0. No other clear differences in intensity are apparent.

Obviously much more research is required to ascertain which of the differences observed between the two stimulus series are general enough in Finnish speech that they could serve as perceptual cues for stop quantity. Preliminary investigations of other tokens of the test words suggest that perhaps the intensity difference observed in the stop burst is not generally characteristic, but the frequency difference, though certainly not observable in every token, may be part of a general trend. It does appear that the existence of non-timing differences in the production of quantity oppositions and their influence in quantity perception may be a fairly ubiquitous phenomenon. In any case, results such as these raise interesting questions about the exact nature of distinctive use of timing in so called quantity languages of which Finnish has often been taken to be a prime example.

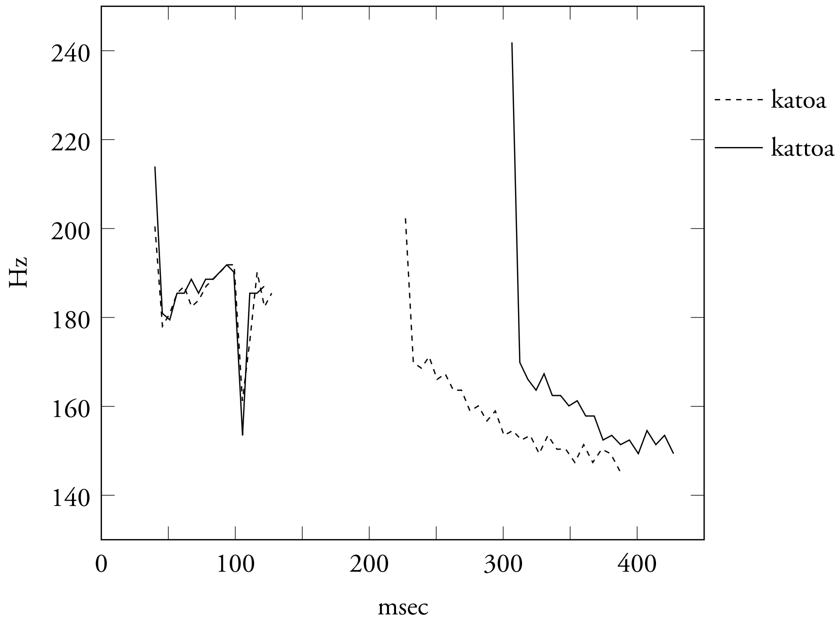


Figure 4.23: Fundamental frequency parameter for *katoa* and *kattoa*

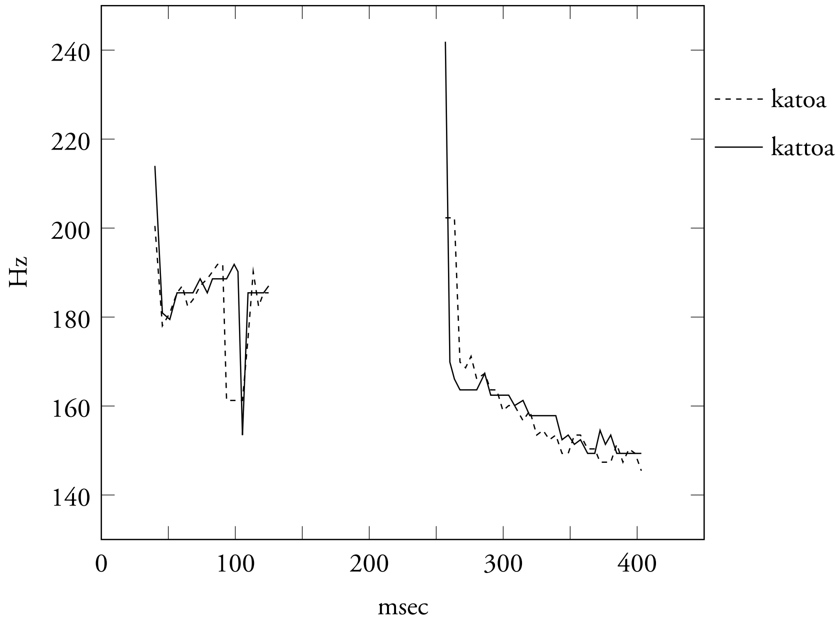


Figure 4.24: Fundamental frequency aligned according to the time warp

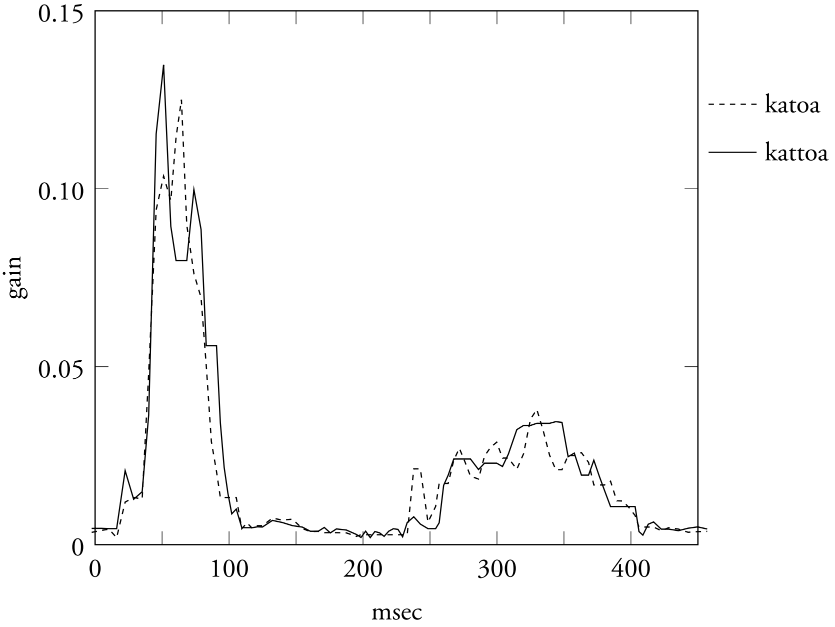


Figure 4.25: Gain parameter aligned according to the time warp

4.4 Previous perception experiments

Abramson and Ren (1990) conducted perception tests of long and short vowels in Thai similar to Experiment 3 of this chapter. In their study, timing was varied by duplicating or deleting pitch periods from near steady state portions of natural speech one syllable words. Thus, just as in the present experiments two series of stimuli were created, although by a different method: one series was created by lengthening original short vowels, the other by shortening original long vowels. The two series were matched with each other using measured vowel duration. They found that although changing vowel duration was always sufficient to change perception from short to long vowel, the boundary between short and long varied significantly depending on whether the original test word contained a short or a long vowel. The difference was in the expected direction, that is, stimuli derived from original short vowel words were more likely to be perceived as short vowel words and vice versa. Since the (spectral) quality of vowels in Thai also typically varies slightly with phonological length (as shown by Abramson and Ren's formant measurements), they conclude that "a major secondary cue to the distinction [ie. phonological vowel length] lies in spectral differences between the long and short members of the vowel pairs." Of course, in addition to the spectral cues there may have been other (perhaps subtle) differences in the two stimulus series which also biased perception.

Hankamer *et al.* (1989) performed perception experiments for geminate voiceless stops in Turkish and Bengali. By lengthening an original short stop and shortening an original geminate they constructed matching stimulus series with equal stop durations. In spite of this matching, listeners in both languages heard geminates more readily when the stimulus was fashioned from a original geminate.

Lehtonen also used the technique of deleting or adding portions of natural speech tokens to create stimuli with varying durations for perception tests in his study of quantity in Finnish (Lehtonen, 1970). Although his vowel duration series were all constructed by deleting portions of original long vowels, included in his data are several cases where stimulus series for varying consonant durations were created both by shortening original long consonants and (producing a separate but related series) by lengthening original short consonants. For instance, there are two series with the segments *t–a–k–a*, one from original *taka* with lengthening of the [k], one from original *takka* with shortening of the [kk]. If we compare the perception of these series by matching measured consonant durations, we may note that equal consonant durations systematically produce a greater percentage of long consonant perceptions when created from a word with an original long consonant. While Lehtonen does not make these comparisons directly, his discussion would seem to imply that the effect can be attributed not to quality differences between the two series in question, but to timing (duration) differences in preceding and following vowels characteristic of the quantity types involved. Thus one would expect that this effect would disappear if the timing of the entire test word were adjusted, not just the duration of a single segment. This is exactly what was done in the present experiments, and indeed, it was part of the motivation for these experiments. The present results show, however, that the effect did not disappear: differences other than strict timing differences had a profound effect on perceived (phonological) quantity, even when the timing of the entire test word was adjusted.

The experiments of Thorsen (1984) on the effect of F_0 timing on perception of vowel length in Danish are directly relevant to the question of the interaction of pitch contour and quantity perception. She found that when stimuli were synthesized so that (spectral) durational cues were ambiguous between *kugle* [kʰuːl̩] and *kulde* [kʰull̩], an earlier rise in F_0 favored long vowel judgments. The explanation proposed is that the F_0 pattern of Danish prosodic stress groups can be characterized "in terms of an essentially invariant F_0 wave upon which the segments and syllables are superimposed." (Thorsen, 1984, p. 17). Thus the longer a vowel persists in relation to the characteristic rising F_0 contour the more likely it is to be perceived as long instead of short. Put another way, "length" is not based on absolute duration, but is relative to other events in the stream of speech, in this case to fluctuations in F_0. This is directly parallel to the results of Experiment 4.

We must not forget, however, that in all these studies as well as in the three experiments of this chapter timing differences did, in fact, "win out" in the long run. In the present experiments this was clearest when the timing of original *tuuli* was changed to that of original *tuli* (stimulus 0 in the *tuuli*-series for both experiments). Listeners unanimously identified these stimuli as *tuli*, even though the “quality” of the signal was presumably more appropriate for *tuuli*.

Nagano-Madsen (1992, pp. 106–115) conducted experiments on perception of a minimal pair in Japanese differentiated by phonological vowel length, *beru* vs. *beeru*, which were resynthesized with the original fall in F_0 shifted by steps. In both original words there was a fall between first and second mora, as appropriate for the pitch accent of these words in the Tokyo and Osaka dialects of Japanese. As the F_0 fall was shifted, forward in the case of original *beeru*, backwards for *beru*, the changed actually “overruled” the durational cue, at least for listeners from the Tokyo and Osaka dialect areas. Listeners from an “accentless” dialect area were not affected. Here again, it appears that perception of durational cues to quantity differences is not independent of other features of the speech signal.

Actually this situation is not peculiar to quantity oppositions, but can also be found in cases of so-called segmental oppositions. For instance, Port (1979) found that shortening the interval corresponding to /b/ in *rabid* produced the perception *ratted*. In other words, a timing difference “won out” over the transitional cues in the surrounding vowels. In effect, the listener hears an alveolar tap instead of a bilabial stop in spite of quality cues to the contrary when the event happens too quickly to have been a stop articulation. In the same article Port discussed findings of other investigators that shortening the intervocalic interval corresponding to a cluster of two stops between vowels VC_1C_2V leads to perception as VC_2V, eg. [ɛbdɛ] → [ɛdɛ], presumably because it would be impossible to articulate two stops so quickly.

It may thus be briefly suggested that the distinction between quantity and quality is not absolute, but rather a matter of degree. The quality/quantity distinction in perception corresponds to local (eg. spectral) properties of the speech signal vs. global timing. Determination of local properties depends in turn on the length of “analysis window” used or the duration of what has been called the “subjective present”. What at one level appears as a sequence of different “qualities” occurring in a certain timing pattern, from another point of view, with a longer subjective present, amounts to a single unified quality. What we see as elapsed time and what we interpret as quality at a single point in time is not predetermined but open to interpretation. This explains how it is that historically quantity can be swapped for quality as languages change (or, for example, for word tone cf. Niemi and Niemi, 1984), or how quality and quantity can be “swapped” in the process of lexical borrowing (O’Dell *et al.*, 1990), as well as how it is possible that varying “degrees” of quantity languages can exist. Perhaps the safest conclusion is that while there is some correspondence between the milliseconds measured by the phonetician and the perception of quantity, that perception cannot be adequately described only in terms of physical linear time.

Chapter 5

Conclusions

> It is impossible to meditate on time and the mystery of the creative passage of nature without an overwhelming emotion at the limitations of human intelligence.
>
> —A. N. Whitehead

5.1 Experimental evidence

Let us now sum up the main findings of the experiments. In Experiment 1 it was found that measured durational differences due to different segments (e.g. *k* vs. *l*) disappear (or are reduced to statistical insignificance) in the "closed mouth" condition, while differences due to differing quantity types (e.g. CVCCV vs. CVCV) remain. This parallels the results of Kozhevnikov and Chistovich's closed mouth experiment for Russian in which segmental differences also disappeared but differences due to "rhythmic figure" (stress on first or second syllable) remained. For Kozhevnikov and Chistovich (Russian) words with the same pattern of stressed and unstressed syllables belong to the same rhythmic figure. It seems reasonable that for Finnish, which lacks phonemic word stress but has a quantity opposition for both vowels and consonants, rhythmic figure should be defined by quantity pattern, not stress. This of course would mean that at least for Finnish, rhythmic figure is not realized at the syllable level as Kozhevnikov and Chistovich propose. Kozhevnikov and Chistovich conclude from their results

> that words determined to coincide according to a rhythmic figure actually have the same rhythmic program. ... it is possible to conclude that the rhythmic figure actually exists as some independent sign of a word (phrase) and that consequently it is necessary to assume the presence in the nervous system of special setups which provide for the generation of complex rhythmic sequences. (Kozhevnikov and Chistovich, 1965, p. 115)

We also conclude from our closed mouth experiment that there is a rhythmic unity underlying a quantity category in Finnish, i.e. speakers have an idea of the rhythm of a quantity category (e.g. CVCV) which is independent of the segments involved. Measured timing differences corresponding to segmental differences can then be thought to be a reflection of various peripheral physiological factors such as differences in "sluggishness" of various articulators (cf. MacNeilage, 1972). However, these differences also reflect the failure of traditional techniques (e.g. measuring durations using *a priori* boundary criteria assumed to be equivalent for different types of phonetic segment) to measure the rhythm transmitted.

If quantity enjoys a certain independence in speech production, the question naturally arises whether it might not also be possible to perceive quantity patterns independently of the segmental content of words. Experiments by Gordon (1988) have shown that the rhythmic basis of speech rate perception in English can be transmitted independently of segmental content. In Gordon's experiments carrier sentences deprived of segmental clues triggered rate-dependent responses to test words in just the way the original carrier sentences did. Experiment 2 was designed to see whether test words deprived of segmental cues could still convey the rhythmic basis of quantity perception. Sine wave replacements preserving the intensity variations and the average frequency of the original test words were used, since Gordon obtained the best results for just such a transformation. Also, time warps based on intensity alone correspond quite well to traditional "spectrogram segmentation," probably since abrupt changes in intensity are salient in spectrograms and thus form easily identifiable boundary markers.

The main result of Experiment 2 was that listeners were *not* able to discriminate quantity types when the original test words were replaced with intensity matched sine waves. This was true even for the group that was informed ahead of time of the segmental content of the original words (although two subjects in this group performed slightly better than chance). Why this negative result? What was the crucial difference between this experiment and Gordon's? One possibility that suggests itself is that quantity rhythm (this experiment) and speech rate rhythm (Gordon's experiment) may be perceived in quite different ways. Another explanation, not necessarily incompatible with this, is that in Gordon's experiments the greater length of the carrier sentence was crucial. In other words, it may be that there was a sufficient sample of the impoverished signal to provide generalized rhythm information despite the signal's "poor quality," whereas in Experiment 2 the impoverished test word was too short for most subjects to form an idea of the rhythm involved.

In the light of the results from Experiment 3 and especially Experiment 4, it is certainly possible that retaining the original F_0 contour would have helped listeners distinguish the two quantity patterns in spite of the lack of quality information. An interesting follow-up experiment would be to reverse the transformation and modify the original test words by leveling the intensity to see if subjects are able to identify quantity patterns without the information provided by intensity fluctuation.

Just how important to quantity perception are the small qualitative differences that exist between quantity types but were missing from the "paapa" stimuli? The experiments of Chapter Four were designed to throw light on this question. Using LPC analysis, dynamic time warping (based on intensity variation) and (re)synthesis, it was possible to match intensity changes in the word pair *tuli*/*tuuli*, and produce two series of stimuli whose timing varied gradually from *tuli* to *tuuli*: In Experiment 3 one series retained the spectral progression and pitch contour of original *tuli*, the other series retained the spectral properties and pitch contour of original *tuuli*. In Experiment 4 only the pitch contour differed between stimulus series. Results showed that these differences did indeed have a very significant influence on perception. That this is likely to be a very general effect is shown by the results of Experiment 5, substituting a pair of test words *katoa*/*kattoa* which *a priori* was not expected to exhibit systematic quality differences at all and which included voiceless stops providing clearer intensity differences than for the *tuli*/*tuuli* pair.

Such a result is disconcerting for all varieties of rubber band timing, since variation in quality (the "stuff" of the rubber band) should be random and thus irrelevant as long as the segmental content remains the same as was the case in these experiments. By proper stretching, we should arrive at just as good a token of *tuuli* regardless of whether the starting point is *tuuli* or *tuli*, or at just as good a token of *kattoa* regardless of whether the starting point is *kattoa* or *katoa*. Since such was not the case, we may reasonably conclude either that the stretching (and hence the time warp) was incorrect, or that the rubber band model is inadequate.

The warp used for the Experiment 3 corresponded well with traditional segmentation of spectrograms. In Experiment 5 the warp was "hand constrained" to conform to traditional segmentation. Thus, at the very least we can reject the discrete, segmental rubber band or duration model of timing, which would require that any warp should do the job, as long as the "segment boundaries" were in the right places. The possibility remains that the two stimulus series in Experiment 3 (or even in Experiment 5) failed to give (statistically) identical results because the time warp used was not "correct". Under this account, listeners actually did perceive the stimuli as merely "stretched in time" relative to each other, but in a way radically different from the stretching used to synthesize the stimuli. Indeed, there is no guarantee that the warps used in these experiments were the best possible ones for minimizing the "perceptual difference" between the two stimulus series. Whether "better" warps (in this sense) could be found, and whether or not remaining spectral differences between tokens would still influence perception is open to speculation.

In Experiment 4 spectral differences between the stimulus series were eliminated by averaging of parameters. The fact that the remaining pitch differences also had a significant effect on perception provides further evidence against a rubber band model, evidence that is not open to the same criticism as in Experiment 3. The evidence would seem to indicate that the difference between (phonologically) *long* and *short* is more complex (for listeners as well as speakers) than just the amount of time spent at each stage (or in each segment) of the word. Appropriate temporal stretching of the acoustic form of a Finnish word (eg. *tuli*) will eventually change perception to a different quantity type (eg. *tuuli*). The end result will not however be as "good" perceptually as an original version of that quantity type. Apparently a genuine change in quantity type always requires reorganization of gestures in production, and this in turn is reflected in the speech signal itself. This could be construed as an increase in complexity compared to a simple difference in absolute time. It must be remembered, however, that computation of absolute time itself is not an unproblematic concept for production and perception. If the results are interpreted to mean that quantity oppositions are indicated in the signal itself, then it may be argued that the task for speaker and listener is actually simplified.

5.2 Speculation

How is the (partial) reorganization of component gestures involved in different quantity patterns to be characterized? Fowler has suggested a model of coarticulation of segments to characterize timing phenomena such as shortening of stressed vowels when unstressed syllables are added (eg. Fowler, 1981a,b). On this account segments (or component gestures) themselves remain relatively invariant, but overlap to varying degrees causing differences in measured durations. For instance, the more a segment is overlapped by neighboring segments, the shorter its own measured duration becomes and the more it affects the quality of those neighboring segments. Applying this to quantity, we would predict that qualitative differences between quantity types, eg. between *tuli* and *tuuli*, should be greatest around segment transitions and in adjacent segments (ie. [l] and [i]) since presumably the differences are due to different amounts of segment overlap or coarticulation. The present study was not designed to consider this question directly, but the differences between the tokens of *tuli* and *tuuli* used in Experiment 3 which were discussed in section 4.1.4 (p. 69 ff.; see especially Figure 4.10, p. 75) would seem to indicate that the opposite relation is perhaps more accurate: qualitative differences for these tokens appear to be at least as great if not greater in the vicinity of the center of the vowel in question ([u]). Perhaps a different type of "coarticulation" is called for.

5.2.1 Bifurcation of rhythmic behavior

Jumping rope at different speeds can be used to illustrate the concept of bifurcation in rhythmic behavior. When a person jumps rope fairly rapidly, hopping and rope twirling quickly become synchronized so that there is one hop for each twirl of the rope. It is possible to slow down both hopping and rope twirling to some extent, but at some point the pattern of jumping and twirling changes to a sort of double hop for each single twirl. In other words, there is a natural boundary between slow jumping and fast jumping, but one which can not be strictly described using absolute time.

Likewise we may conjecture that a bifurcation in the behavior of rhythmic speech gestures is exploited (created?) by so called quantity languages. This would mean that differences between quantity types do not involve *just* lengthening (slowing down) or shortening (speeding up), but that the organization of component gestures would be different for different quantity types.

Much success in mathematical modeling of biological rhythms has been achieved in recent years utilizing groups of coupled oscillators (cf. Rand *et al.*, 1988; Kopell, 1988, and many references therein). The idea is to assume that subsystems can be postulated which would exhibit (simple) oscillatory behavior in isolation, but may exhibit more complex behavior when allowed to influence each other normally. A mathematical technique known as *Average Phase Difference Theory* (APD) can be used for cases of weak coupling between oscillators with unique limit cycles (Kopell, 1988). This technique ignores details of the oscillators involved by using only phase measured along the limit cycle to characterize the oscillators. For instance, general detailed equations for a pair of coupled oscillators might be characterized as

$$\dot{X}_1 = F_1(X_1) + G_1(X_1, X_2) \tag{5.1}$$
$$\dot{X}_2 = F_2(X_2) + G_2(X_1, X_2) \tag{5.2}$$

where X_1 and X_2 are state vectors (most likely of many dimensions each), $\dot{X}_1$ and $\dot{X}_2$ indicate the change in state, $\dot{X}_1 = F_1(X_1)$ and $\dot{X}_2 = F_2(X_2)$ are limit cycle oscillators and the (multidimensional) functions G_1 and G_2 represent the coupling (influences) between the two oscillators. Using APD, on the other hand, equations can be used describing only the interactions of two phases θ_1 and θ_2:

$$\dot{\theta}_1 = \omega_1 + H_1(\theta_1, \theta_2) \tag{5.3}$$
$$\dot{\theta}_2 = \omega_2 + H_2(\theta_1, \theta_2) \tag{5.4}$$

where ω_1 and ω_2 are scalars indicating the uncoupled speed of the component oscillators, while H_1 and H_2 are scalar functions showing the contribution of the coupling. Such modeling is in general quite “robust” in the sense that the behavior of the system as a whole is relatively insensitive to details of the oscillators or the coupling. This is very important since very often the detailed behavior of the component oscillators is not known and they cannot in fact be studied in isolation. It also means that topological behavior of the system can be modeled without assuming that the phase of the oscillators is to be strictly interpreted as physical time.

A simple example of coupling allowing bifurcation between $1:1$ and $2:1$ phase entrainment (ie. synchronization) is given by Keith and Rand (1984, p. 137), (and discussed in Rand *et al.*, 1988, p. 359 ff.):

$$\dot{\theta}_1 = \omega_1 - \alpha \sin(\theta_1 - \theta_2) - \beta \sin(\theta_1 - 2\theta_2) \tag{5.5}$$
$$\dot{\theta}_2 = \omega_2 + \alpha \sin(\theta_1 - \theta_2) + \beta \sin(\theta_1 - 2\theta_2) \tag{5.6}$$

This system can switch between $1:1$ and $2:1$ entrainment for a variety of values of the coupling parameters (for instance $\alpha = \beta = 0.5$) by changing the (uncoupled) frequency (or "speed") of either of the oscillators (ω_1 or ω_2). In this respect it illustrates bifurcation behavior similar to the rope jumping described above. A similar hierarchical model could be proposed for quantity, assuming there are many speech rhythms going simultaneously, with cycles corresponding for instance to something like mora, syllable or foot. With suitable coupling, changes could easily lead to bifurcation in entrainment (and concomitant reorganization of gestures).

While both Fowler and Kelso have mainly tried to characterize phase relations between sequential gestures, quantity may perhaps be better understood as involving coupling of hierarchical gestures.

According to a model of weakly coupled oscillators, paradigmatic timing differences corresponding to different segments such as those found in Experiment 1 could be brought about by a change in the "uncoupled" behavior of one of the oscillators (corresponding to change of articulator), which through the effects of coupling would normally also affect the behavior of the other oscillator(s) in the system. This other oscillator could then be associated with quantity patterns. However, while an oscillator model makes it possible to "uncouple" quantity effects and segmental effects conceptually, it is not possible to directly observe their uncoupled behavior. In particular the closed mouth condition of Experiment 1 should not be interpreted as "removing" the "segmental" oscillator to let quantity "run free". To the extent that segmentally related differences are eliminated in the closed mouth condition, it may be postulated that the normal "segmental articulation" oscillators involved have been replaced with other oscillators that are practically identical.

In this account, qualitative differences are to be expected along with timing differences, because of the different phase relations between the component oscillations. Using Fowler's terminology, we could say the two rhythms are coarticulated.

Recognition of rhythmic patterns by the listener in this account could be interpreted as tracking of the various oscillator cycles *and attending to their interrelations*. Naturally the qualitative differences due to differences in coarticulation relations would then be expected to greatly influence recognition as was seen to be the case in Experiment 3. It may be that some fluctuations in the speech signal, such as pitch contour, are more closely associated with larger rhythms. Wiik has suggested that at least for some dialects of Finnish pitch contour is directly associated with a mora rhythm (Wiik, 1988). If this is the case, the significant effect of pitch contour on quantity perception in Experiment 4 finds a natural explanation.

Postulation of hierarchical units of speech timing is certainly not new (for Finnish cf. Lehtonen, 1970, pp. 148–152). The appeal of characterization as oscillator coupling is first of all that it is not necessary to assume instantaneous boundaries in speech. Secondly, because the oscillators will in general exert mutual influence (eg. both H_1 and H_2 will be non-zero in Equations 5.3 and 5.4), we would not expect strict isochrony (or periodicity) at any level, although various "compensatory" trends are likely.

5.2.2 *Rhythmische Abstufung* revisited

O'Dell and Nieminen (1998, 1999, 2001, 2002a,b) found that an abstract model of interacting hierarchically coupled oscillators proved very helpful in understanding the interaction of syllable and stress rhythms. The ubiquitous phenomenon in speech, well known for more than a century (called *rhythmische Abstufung* by Sievers, 1893), that smaller units such as segments and syllables tend to become shorter in duration as more of them are incorporated into a single higher level unit, finds a reasonable explanation in this model by assuming that one cycle of a higher level rhythm such as a stress group (one oscillator in the model) is synchronized with a variable number of cycles of a lower level rhythm such as a syllable. Furthermore, application of Kopell's APD theory to the hierarchically coupled oscillator model allows prediction of approximate relations between these rhythms which have been empirically verified for many languages. One such prediction involves the tendency for the period of a higher level oscillator to be a linear function of the number of subordinate level cycles to which it is synchronized. In the case of a two level model, this function can be expressed as

$$\begin{aligned} T_1 &= c_1 + c_2 N_{1,2} \\ &= q(r_{1,2} + N_{1,2}), \\ q &= 1/(r_{1,2}\omega_1 + \omega_2) \end{aligned} \tag{5.7}$$

where T_1 is the period of the level one oscillator at equilibrium, $N_{1,2}$ refers to the number of level two oscillator periods synchronized with (contained in) each level one period, $r_{1,2}$ is the relative strength of coupling between the two oscillators (ie. the ratio of influences in opposite directions: $r_{1,2} > 1$ means the level one oscillator dominates), ω_1 and ω_2 are eigenfrequencies as in Equations 5.3 and 5.4. In a two oscillator model only "compression" effects can be observed, where adding an extra unit at the lower level will simultaneously increase the overall duration while decreasing the average duration of the components. For instance, modeling stress and syllable rhythms this way means that adding a syllable can only increase stress group duration while decreasing average syllable duration.

In a model with three or more hierarchical oscillators, more complex patterns can emerge. For instance the formula for preferred period of the top oscillator in a model with three hierarchically coupled oscillators is

$$\begin{aligned} T_1 &= c_1 + c_2 N_{1,2} + c_3 N_{1,3} \\ &= q(r_{1,2}r_{2,3} + r_{2,3}N_{1,2} + N_{1,3}), \\ q &= 1/(r_{1,2}r_{2,3}\omega_1 + r_{2,3}\omega_2 + \omega_3) \end{aligned} \tag{5.8}$$

Now adding a level 3 unit (ie. increasing $N_{1,3}$ by one) may or may not entail adding a level 2 unit. If it does involve adding a level 2 unit, it can easily happen that the addition will actually *increase* the average duration of level 3 units, given that the level 2 oscillator is fairly dominant (ie. $r_{1,2}$ small so that $r_{1,2} < N_{1,3} - N_{1,2}$).

To put this in intuitive terms, suppose the rhythmic units so modeled are (phonological) word, mora and segment. Then adding a single segment which simultaneously adds a mora could cause an increase in segment durations rather than "compression," provided the mora rhythm is dominant enough. This is a possible consequence of allowing something like *rhytmische Abstufung* to work at more than one level simultaneously.

This type of "anticompensation" is demonstrated and discussed at length by Port *et al.* (1987) for Japanese. On the basis of their experiments they tentatively conclude that "the kind of model proposed here for Japanese ... postulates an abstract timing unit that must be specified independently of the actual syllabic gestures that implement it. That is, this model seems to require extrinsically specified timing." (Port *et al.*, 1987, p. 1584). Since anticompensation can be a result of coupled oscillators, this is not a necessary conclusion.

The hierarchically coupled oscillator model in its present form deals only with average effects, since it depends on averaging out phase differences over the component periods. This is indeed one of its strengths and the reason it is possible to draw very general conclusions. However this means it cannot in principle describe how these average effects (such as anticompensation) are "spread out" over the various cycles. For this a different approach is needed, one which focuses on characterizing the coupling between individual speech gestures, that is, intergestural coordination (Fowler and Saltzman, 1993). Recent work within the so-called Task Dynamics paradigm (Saltzman and Munhall, 1989; Saltzman, 1995) has been exploring ways to characterize various levels of speech timing from a dynamic systems point of view. Among other levels, a level of intergestural timing is postulated, referring to coordination (or coupling) between gestures, as well as a level of *transgestural timing*, referring to "modulations of the timing properties of all gestures active during a localized portion of an utterance" (Byrd and Saltzman, 2003, p. 156), associated with so-called π-gestures (or prosodic gestures). It remains to be seen whether quantity differences can be accounted for in terms of intergestural timing, transgestural timing, or perhaps some level intermediate between these two. An exciting possibility for investigating the dynamics of quantity production would be to apply perturbation analysis (cf. eg. Saltzman *et al.*, 1998) or rate induced intergestural phase transitions (Kelso *et al.*, 1986a,b; Kelso, 1995).

5.2.3 Intrinsic vs. extrinsic timing

What has happened to the distinction between intrinsic and extrinsic timing in relation to a model of coupled oscillators? It can be reasonably associated with two different theoretical types of influence or *forcing* of oscillators. Kopell calls these types *temporal* forcing and *phasic* forcing:

> By temporal forcing, we mean periodic influences on the system which depend explicitly on time, and which may be independent of the state of the system. ... By contrast, phasic forcing means an influence that is

> exerted when the system is in a particular state, independent of what time that state (i.e., set of phases) is reached. (Kopell, 1988, p. 383)

It should be clear, then, that the proposal made here is for an intrinsic timing model. The mutual influence of the postulated oscillators is sensitive to the relative phases of the oscillators, not to absolute clock time. In addition, it should be emphasized that neither oscillator can be characterized as an independent timekeeper which provides timing information to the other without itself being affected by the association.

In Chapter One Lubker's remark was quoted as an example of the criticism faced by intrinsic timing models when dealing with so-called quantity languages. An interesting point is raised in the reply made by Kelso *et al.* to Lubker's critique:

> The mistaken assumption that what the experimenter chooses to measure must be what the speaker controls is also apparent in Lubker's assertion that speakers directly control timing because "In Swedish, for example, duration information can be critical in differentiating between otherwise identical words". Tellingly, phonologically long and short vowel pairs in, for example, Estonian also differ slightly in formant frequency (cf. Lehiste, 1970). To our knowledge investigators have yet to characterize acoustic formant transitions to and from vowels that differ "only" in duration. (Kelso *et al.*, 1986b, p. 184)

This "mistaken assumption" is, unfortunately, very easy to make and quite common in phonetics. It calls to mind Whitehead's warning that

> The only logical conclusion to be drawn, when a contradiction issues from a train of reasoning, is that at least one of the premises involved in the inference is false. It is rashly assumed without further question that the peccant premise can at once be located. (Whitehead, 1929, p. 8)

The present study, in a sense, has been an attempt to "characterize acoustic formant transitions to and from vowels that differ 'only' in duration," that is, to investigate both temporal and qualitative differences in a quantity language. In a later article Vatikiotis-Bateson and Kelso (1993, p.239) stated that consonant quantity in Japanese provides an exception to the claim that time need not be a controlled variable. In light of the research and speculation presented here, this retreat in position may be premature.

5.3 Summary

Time (absolute, scientific or physical time) is not a primary aspect of utterances, but rather the result of measuring events. Models of speech timing which directly make use of physical time imply an underlying mechanism whereby the human perceptual system measures "time" independently of speech and then utilizes that measurement in phonetic judgment.

However, natural speech itself is rich in "timing" information, that is, there is an indeterminate number of cues which help the listener track the unfolding of the articulatory events reflected in the speech signal. Probably one such cue which is often prominent is closely correlated with fluctuation of F_0 in relation to other features of the signal. On the other hand, it is apparent that manipulating timing "alone" in the synthesis of stimuli may override other information, at least in extreme cases.

An analogous case can be imagined of a video recording of the rope jumping used above as an example of bifurcation in rhythmic activity. At near normal viewing speeds it should be possible for a viewer to distinguish "single hop" and "double hop" jumping, but if the video is speeded up, at some point the entire action will take place too quickly for an observer to distinguish them. Here "too quickly" can be taken to mean within the (psychological) present of the observer. Note that in this account no recurrence need be made to absolute time or to segment boundaries.

The model proposed here may be considered basically an intrinsic model, though in one sense an "extrinsic" aspect is readily admitted: gross influences of events both in the speaker and in the listener which are not directly related to speech production will influence the timing of utterances to some extent. The sum of these "external" influences amounts to a biological clock of sorts, but one which is not accurate as demonstrated by various studies of rhythm perception using both speechlike and non-speech stimuli (cf. eg. Fox and Lehiste, 1989; Povel, 1981). Quantity differences necessarily involve some reorganization of speech gestures and consequently have acoustic effects which cannot be characterized merely as "stretching of the time axis". Of course, it must be remembered that this mathematical model is only a possible metaphor for the process of quantity production, but perhaps it has some potential for illuminating that process.

It seems that a "pure" quantity language in an absolute sense would require some form of extrinsic timing. On the other hand Finnish, at least, though often taken as a prime example of a quantity language, does not appear to be such a pure quantity language. Associated with phonological quantity differences are diverse effects which are not merely temporal in nature. How then do so-called quantity languages differ from other languages? The present work was not designed for comparison of language types, but perhaps the answer lies in the ubiquitous use of hierarchical rhythms. In any case the simple answer that quantity languages make phonological use of duration is obviously inadequate.

What about the usefulness of the dynamic time warping techniques discussed in Chapter 3? Consider the following succession of timing models and the associated measuring techniques: a duration based or discrete interval timing model lends itself to traditional segmentation and measurement of elapsed time between segment boundaries. A continuous rubber band model of timing is compatible with the computational technique of dynamic time warping. These may perhaps be termed "pure linear timing" models. If instead of these a model is proposed which is based on phase relations of (simultaneous) interacting systems, hierarchical as well as (roughly) sequential, what then is the utility of the observational methods associated with the other models?

Human speech perception does *not* involve dynamic time warping or similar computational method—it doesn't have to, because perception does not proceed on the basis of a series of static "snapshots" whose dynamic relations must then be "computed". Rather it may be more accurate to say that identification is based on "tracking" or following the signal. However, as an analysis technique DTW may still prove quite useful, as indeed it has in the present investigation. The important point to keep in mind in the application of this technique is that speech will in general exhibit multiple simultaneous rhythms, so that no one time warp will be universally adequate.

Likewise traditional studies presenting measured durations provide useful information, if used with caution. The technique of searching for recognizable points in the speech signal and measuring the intervening lapses of time is roughly analogous to the so called Poincaré section of dynamic system analysis (cf. eg. Guckenheimer and Holmes, 1996, pp. 22–32). The difficulty is in finding the "same" point in different tokens of speech, especially if the component gestures are systematically different or restructured.

Yhteenveto

Tämän työn nimessä, *Intrinsic Timing and Quantity in Finnish*, on kaksi avaintermiä, *kvantiteetti* (*quantity*) ja *sisäinen ajoitus* (*intrinsic timing*), joiden näennäinen ristiriita on ollut tutkimuksen lähtökohtana.

Termiä *ulkoinen ajoitus* (*extrinsic timing*) on käytetty fonetiikan kirjallisuudessa luonnehtimaan teorioita, joiden mukaan puhuja pakottaa puheensa tiettyihin ajallisiin muotteihin artikulaatioista riippumatta. Ajoituskaava ja artikulaatio, joka sen täyttää, ovat tällöin toisistaan riippumattomia. Sen sijaan sisäinen ajoitus viittaa teorioihin, joissa ajoitus ei ole artikulaatioista riippumaton, vaan suorastaan emergoituu niistä ja niiden vuorovaikutuksista. Tällöin esim. mitatut kestot ovat seurausta artikulaatioiden erilaisista dynaamisista piirteistä.

Termi *kvantiteettikieli*, jollaisena myös suomen kieltä yleisesti pidetään, tarkoittaa perinteisen selityksen mukaan sellaista kieltä, jossa äänteiden kestoja käytetään erottamaan eri sanoja toisistaan. Tällaisessa kielessä kestoerot ovat ilmeisesti tarkoituksellisia ja niinpä tuntuu ilmeiseltä, että puhuja säätää niitä suoraan ja kuulija laskee niitä saadakseen sanoista selvää. Kvantiteettikielien olemassaolo siis tuntuu tukevan ulkoisen ajoituksen käsitettä, ja muodostavan esteen sisäisen ajoituksen hyväksymiselle.

Tutkimuksen tuloksia voidaan tiivistää seuraavasti. Toisaalta kokeissa, joissa pyydettiin puhujia lausumaan sanoja siten, että suu oli kiinni ja kieli liikkumatta, äänteelliset ajoituserot hävisivät, mutta kvantiteettiin liityvät erot jäivät. Vaikuttaa siltä, että suomen kielen puhujalla on kvantiteettihahmo tai rytmillinen malli, joka on äänne-eroista suhteellisen riippumaton. Voidaan siis väittää, että hän käyttää samaa rytmiä lausuessaan esim. sanat *tulli* ja *tukki*, vaikka mitatut kestot ovatkin systemaattisesti eri. Sen sijaan havaintokokeet, joiden ärsykkeistä poistettiin äänteellinen informaatio, viittasivat siihen, että suomen kielen kuulija ei havaitse kvantiteettia riippumattomasti, vaan tarvitsee myös äänteellistä informaatiota kvantiteettihahmon tunnistamista varten. Tämä tulos on sopusoinnussa sen olettamuksen kanssa, että puheen havaitsija kuuntelee todellisia artikulaatioita, eikä irrallisia piirteitä, joista sanan identiteettiä sitten laskettaisiin.

Toisissa kokeissa todettiin, että venyttämällä keinotekoisesti sanan kestoja toisen sanan kestojen mukaisiksi on mahdollista muuttaa sanan tunnistus odotettuun suuntaan. Pelkästään ajoitusta muuttamalla jää kuitenkin aina muita tekijöitä, esim. sävelkulku ja äänen laatu, jotka sopivat paremmin alkuperäiseen sanaan, ja näin ”jarruttavat" tunnistuksen muuttumista. Esim. sana *tuli* saadaan venyttämällä kuulostamaan sanalta *tuuli*, mutta lopputulos ei ole yhtä hyvä kuin alkuperäinen *tuuli*. Tämä ilmeisesti johtuu siitä, että todellinen muutos sanan rytmissä edellyttää aina puheliikkeiden erilaista organisaatiota, ja nämä erot heijastuvat myös akustiseen puhesignaaliin tarjoten kuulijalle monia vihjeitä sanan identiteetistä.

Työssäni spekuloin, että erilaiset kvantiteettihahmot voisivat selittyä samantyyppiseksi toiminnan *bifurkaatioksi* eli *haaraumaksi*, jonka voi nähdä kun ihminen hyppii narua. Siinä kaksi rytmiä, hyppiminen ja narun pyörittäminen, synkronoituvat eli kytkeytyvät toisiinsa ajallisesti. Kun ihminen aloittaa esim. nopeasti ja sitten hyppii yhä hitaammin, jossain vaiheessa tapahtuu laadullinen muutos rytmien kytkennässä: alkuperäinen yksittäinen hyppy jokaista narun pyöritystä kohti muuttuu kaksoishypyksi. Näin ollen on olemassa luonnollinen raja hitaan ja nopean hyppimisen välillä, mutta ero *ei* perustu absoluuttiseen aikaan. Samoin on kuviteltavissa, että niin sanottu kvantiteettikieli käyttää hyväksi vastaavia rytmillisiä eroja liikkeiden kytkennöissä puheen tuottamisessa. Silloin informaatio kvantiteettihahmosta olisi itse puheessa, eikä edellyttäisi ulkoista "kelloa" kestojen tuottamista tai havaitsemista varten.

Voi olla, että ajatus kvantitteettikielestä perinteisessä mielessä (absoluuttiset kestot erottamassa merkityksiä) on yhteensopimaton sisäisen ajoituksen periaatteen kanssa. Suomi ei ole kuitenkaan osoittautunut tällaiseksi kieleksi. Voimme silti sanoa, että kvantiteettikielessä ajoituserot ovat distinktiivisiä, mutta on muistettava, että puhujan ja kuulijan kokema ajoitus ei ole sama kuin tutkijan mittaama kesto.

Bibliography

Aaltonen, O. and Hulkko, T., editors (1985). *XIII Meeting of Finnish Phoneticians – Turku 1985*. Department of Finnish and General Linguistics of the University of Turku.

Abramson, A. S. and Ren, N. (1990). Distinctive vowel length: Duration vs. spectrum in Thai. *Journal of Phonetics*, **18**, 79–92.

Alexander, H. G., editor (1956). *Leibniz-Clarke Correspondence*. Manchester University Press.

Anderson, V. L. and McLean, R. A. (1974). *Design of Experiments: A Realistic Approach*. New York: Marcel Dekker, Inc.

Aulanko, R. (1985). Microprosodic features in speech: Experiments on Finnish. In Aaltonen and Hulkko (1985), pages 33–54.

Bakó, E. and Sovijärvi, A. (1939). Unkarin ja suomen lyhyiden ja pitkien *i*-, *ü*- ja *u*-vokaalien fysiologis-akustista vertailua. *Virittäjä*, **43**, 386–400.

Bergson, H. (1910). *Time and Free Will: An Essay on the Immediate Data of Consciousness*. London: George Allen and Unwin. Translation by F. L. Pogson of *Essai sur les données immédiates de la conscience*, 1889. Quotations and page references in the text are from the English translation.

Borden, G. J. and Harris, K. S. (1984). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. London: Williams and Wilkins, second edition.

Bracewell, R. N. (1986). *The Hartley Transform*. Oxford University Press.

Byrd, D. and Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, **31**, 149–180.

Carlin, B. P. and Louis, T. A., editors (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman & Hall, 2nd edition.

Catford, J. C. (1977). *Fundamental Problems in Phonetics*. Edinburgh: Edinburgh University Press.

Cohen, A. H., Rossignol, S., and Grillner, S., editors (1988). *Neural Control of Rhythmic Movements in Vertebrates*. New York: John Wiley and Sons.

Cooper, A. M., Whalen, D. H., and Fowler, C. A. (1986). P-centers are unaffected by phonetic categorization. *Perception & Psychophysics*, **39**, 187–196.

Diehl, R. L., Souther, A. F., and Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psychophysics*, **27**, 435–443.

Dixon, N. R. and Martin, T. B., editors (1979). *Automatic Speech and Speaker Recognition*. New York: IEEE Press.

Donner, K. (1912). Salmin murteen kvantiteettisuhteista. *Suomi*, **IV**(9).

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press.

Fowler, C. (1977). *Timing Control in Speech Production*. Bloomington, Indiana: Indiana University Linguistics Club.

Fowler, C. (1979). 'Perceptual centers' in speech production and perception. *Perception & Psychophysics*, **25**, 375–388.

Fowler, C. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, **8**, 113–133.

Fowler, C. (1981a). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research*, **24**, 127–139.

Fowler, C. (1981b). A relationship between coarticulation and compensatory shortening. *Phonetica*, **38**, 35–50.

Fowler, C. (1983). Realism and unrealism: A reply. *Journal of Phonetics*, **11**, 303–322.

Fowler, C. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, **14**, 3–28.

Fowler, C. (1990). Calling a mirage a mirage: Direct perception of speech produced without a tongue. *Journal of Phonetics*, **18**, 529–541.

Fowler, C. (1994). Speech perception: Direct realist theory. In R. E. Asher, editor, *The Encyclopedia of Language and Linguistics*, pages 4199–4203. Oxford:Pergamon.

Fowler, C. and Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech*, **36**, 171–195.

Fowler, C. A., Whalen, D. H., and Cooper, A. M. (1988). Perceived timing is produced timing: A reply to Howell. *Perception & Psychophysics*, **43**, 94–98.

Fox, R. A. and Lehiste, I. (1989). Discrimination of duration ratios in bisyllabic tokens by native English and Estonian listeners. *Journal of Phonetics*, **17**, 167–174.

Fry, D. B. (1968). Prosodic phenomena. In B. Malmberg, editor, *Manual of Phonetics*, pages 365–410. Amsterdam: North-Holland.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton-Mifflin.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996a). Introducing Markov chain Monte Carlo. In Gilks *et al.* (1996b), pages 1–19.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996b). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association*, **70**, 755–768. Reprinted as Chapter 9 in Goodman (1978).

Goodman, L. A. (1978). *Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent-Structure Analysis*. London: Addison-Wesley.

Gordon, P. S. (1988). Induction of rate-dependent processing of coarse-grained aspects of speech. *Perception & Psychophysics*, **43**, 137–146.

Gray, R. M., Buzo, A., Gray, A. H., J., and Matsuyama, Y. (1980). Distortion measures for speech processing. *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-28**, 367–376.

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

Guckenheimer, J. and Holmes, P. (1996). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verlag, 5th edition.

Hakulinen, L. (1968). *Suomen kielen rakenne ja kehitys*. Helsinki: Otava, 3rd edition.

Hankamer, J., Lahiri, A., and Koreman, J. (1989). Perception of consonant length: voiceless stops in Turkish and Bengali. *Journal of Phonetics*, **17**, 283–298.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Harms, R. (1964). Review of Lauri Hakulinen, The Structure and Development of the Finnish Language. *Word*, **20**(1), 105–114.

Hawkins, S. (1999a). Looking for invariant correlates of linguistic units: Two classical theories of speech perception. In Pickett (1999), pages 198–231.

Hawkins, S. (1999b). Re-evaluating assumptions about speech perception: Interactive and integrative theories. In Pickett (1999), pages 232–288.

Hellmich, M., Abrams, K. R., Jones, D. R., and Lambert, P. C. (1998). A Bayesian approach to a general regression model for ROC curves. *Medical Decision Making*, **18**(4), 436–443.

Hellmich, M., Abrams, K. R., and Sutton, A. J. (1999). Bayesian approaches to meta-analysis of ROC curves. *Medical Decision Making*, **19**(3), 252–264.

Hoequist, C. E. (1983). The perceptual center and rhythmic categories. *Language and Speech*, **26**, 367–376.

Hombert, J.-M., Ohala, J. J., and Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, **55**, 37–58.

Howell, P. (1984). An acoustic determinant of perceived and produced anisochrony. In M. P. R. van den Broeck and A. Cohen, editors, *Proceedings of the 10th International Congress of Phonetic Sciences*, pages 429–433. Dordrecht, Holland: Foris Publications.

Howell, P. (1988). Prediction of P-center location from the distribution of energy in the amplitude envelope: I & II. *Perception & Psychophysics*, **43**, 90–93, 99.

Hurme, P. and Dufva, H., editors (1987). *Papers from the 14th Meeting of Finnish Phoneticians*. Department of Communication, University of Jyväskylä.

Iivonen, A. (1974a). Äännekeston riippuvuus sanan pituudesta irrallaan äännetyissä sanoissa. *Virittäjä*, **78**, 134–151.

Iivonen, A. (1974b). Äännekestojen riippuvuus ilmauksen pituudesta. *Virittäjä*, **78**, 399–402.

Iivonen, A. (1975). Ääniraon avauma-asteen suuruudesta suomen konsonanteilla. In *Fonetiikan paperit – Helsinki 1975*, pages 43–61. Helsinki: Publications of the Institute of Phonetics, University of Helsinki.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-23**, 62–72.

Jespersen, O. (1926). *Lehrbuch der Phonetik*. Leipzig.

Karjalainen, M. and Laine, U. K., editors (1988). *Papers from the 15th Meeting of Finnish Phoneticians*. Helsinki University of Technology, Faculty of Electrical Engineering, Acoustics Laboratory.

Keith, W. L. and Rand, R. H. (1984). 1 : 1 and 2 : 1 phase entrainment in a system of two coupled limit cycle oscillators. *Journal of Mathematical Biology*, **20**, 133–152.

Keller, E. (1987). The variation of absolute and relative measures of speech activity. *Journal of Phonetics*, **15**, 335–347.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT Press.

Kelso, J. A. S. and Tuller, B. (1987). Intrinsic time in speech production: Theory, methodology, and preliminary observations. In E. Keller and M. Gopnik, editors, *Motor and Sensory Processes of Language*, pages 203–222. Hillsdale, New Jersey: Lawrence Erlbaum.

Kelso, J. A. S., Saltzman, E. L., and Tuller, B. (1986a). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, **14**, 29–59.

Kelso, J. A. S., Saltzman, E. L., and Tuller, B. (1986b). Intentional contents, communicative context, and task dynamics: A reply to the commentators. *Journal of Phonetics*, **14**, 171–196.

Klatt, D. K. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1208–1221.

Kopell, N. (1988). Toward a theory of modelling central pattern generators. In Cohen *et al.* (1988), pages 369–413.

Kozhevnikov, V. A. and Chistovich, L. A. (1965). *Speech, Articulation, and Perception*. Washington, DC: Joint Publications Research Service. Translation of Речь. Артикуляция и восприятие. Москва-Ленинград, «Наука». Quotations and page references in the text are from the English translation.

Kruskal, J. B. and Liberman, M. (1983). The symmetric time-warping problem: From continuous to discrete. In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 125–161. Reading, Massachusetts: Addison-Wesley.

Ladefoged, P. (1967). *Three Areas of Experimental Phonetics*. London: Oxford University Press.

Ladefoged, P. (1975). *A Course in Phonetics*. New York: Harcourt Brace Jovanovich.

Laurosela, J. (1922). *Foneettinen tutkimus Etelä-Pohjanmaan murteesta*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, Massachusetts: MIT Press.

Lehiste, I. (1976). Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics*, **4**, 113–117.

Lehiste, I. and Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America*, **33**, 419–425.

Lehiste, I., Morton, K., and Tatham, M. A. A. (1973). An instrumental study of consonant gemination. *Journal of Phonetics*, **1**, 131–148.

Lehtonen, J. (1969). Huomioita kvantiteettien foneemirajoista ja subjektiivisista kestohavainnoista. *Virittäjä*, **73**, 363–370. Summary: Phoneme boundaries and subjective quantity observations of perception.

Lehtonen, J. (1970). *Aspects of Quantity in Standard Finnish*. Number VI in Studia Philologica Jyväskyläensia. Jyväskylä: University of Jyväskylä.

Lehtonen, J. (1974). Sanan pituus ja äännekestot. *Virittäjä*, **78**, 152–160.

Lehtonen, J. (1979). On labial co-articulation. In P. Hurme, editor, *Papers from the Eighth Meeting of Finnish Phoneticians*, pages 99–106. Jyväskylä: Institute of Finnish Language and Communication, University of Jyväskylä.

Levinson, S. E. and Liberman, M. Y. (1981). Speech recognition by computer. *Scientific American*, **244**, 56–68.

Lubker, J. (1986). Articulatory timing and the concept of phase. *Journal of Phonetics*, **14**, 133–137.

MacNeilage, P. (1972). Speech physiology. In J. Gilbert, editor, *Speech and Cortical Functioning*, pages 1–72. New York: Academic Press.

Magga, T. (1984). *Duration in the Quantity of Bisyllabics in the Guovdageaidnu Dialect of North Lappish*. Acta Universitatis Ouluensis, Series B Philologica No. 4. Oulu: University of Oulu.

Malmberg, B. (1949). Review of Marguerite Durand, Voyelles longues et voyelles brèves. *Studia Linguistica*, **3**, 39–59.

Marjomaa, I. (1982). Englannin ja suomen vokaalien kestoista puhenopeuden vaihdellessa. In A. Iivonen and H. Kaskinen, editors, *11th Meeting of Finnish Phoneticians — Helsinki 1982*, pages 119–137. Helsinki: Department of Phonetics, University of Helsinki.

Markel, J. D. and Gray, A. H. J. (1976). *Linear Prediction of Speech*. New York: Springer-Verlag.

Nagano-Madsen, Y. (1992). *Mora and Prosodic Coordination: A Phonetic Study of Japanese, Eskimo and Yoruba*. Lund University Press.

Nahkola, K. (1987). *Yleisgeminaatio*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Newton, I. (1974). *Mathematical Principles of Natural Philosophy*. Berkeley: University of California Press. Translation of *Philosophiæ naturalis principia mathematica*, 1687. Quotations and page references in the text are from the English translation.

Niemi, J. and Niemi, S. (1984). Word tone and related matters in the Finnish Southwest. In C.-C. Elert, I. Johansson, and E. Strangert, editors, *Nordic Prosody III: Papers from a Symposium*, pages 187–200. Acta Universitatis Umensis, Umeå Studies in the Humanities 59.

Nittrouer, S., Munhall, K., Kelso, J. A. S., Tuller, B., and Harris, K. S. (1988). Patterns of interarticulator phasing and their relation to linguistic structure. *Journal of the Acoustical Society of America*, **84**, 1653–1661.

O'Dell, M. (1987). Rytmin modulaatio kvantiteetin tutkimuksessa. In Hurme and Dufva (1987), pages 69–81. Abstract: Rhythm distortion in quantity research.

O'Dell, M. (1991). Dynamic time warping as a tool in speech timing research. In K. Suomi, editor, *Papers from the 16th Meeting of Finnish Phoneticians*, pages 131–137. University of Oulu.

O'Dell, M. (1995). Intrinsic timing in a quantity language. Unpublished licentiate thesis, University of Jyväskylä.

O'Dell, M. (1999). Some factors affecting perception of stop quantity in Finnish. In J. Järvikivi and J. Heikkinen, editors, *Out Loud: Papers from the 19th Meeting of Finnish Phoneticians*, number 33 in Studies in Languages, pages 76–85. University of Joensuu, Faculty of Humanities.

O'Dell, M. and Nieminen, T. (1998). Reasons for an underlying unity in rhythm dichotomy. *Linguistica Uralica*, **3**, 178–185.

O'Dell, M. and Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. Bailey, editors, *Proceedings of The XIVth International Congress of Phonetic Sciences*, volume 2, pages 1075–1078. University of California, Berkeley.

O'Dell, M. and Nieminen, T. (2001). Speech rhythms as cyclical activity. In S. Ojala and J. Tuomainen, editors, *21. Fonetiikan päivät Turku 4.-5.1.2001*, number 67 in Publications of the Department of Finnish and General Linguistics of the University of Turku, pages 159–168.

O'Dell, M. and Nieminen, T. (2002a). How long is a stress group? *Cadernos de Estudos Lingüísticos*, **43**, 93–108.

O'Dell, M. and Nieminen, T. (2002b). Rytmijakson pituus oskillaattorimallissa. In P. Korhonen, editor, *Fonetiikan päivät 2002 / The Phonetics Symposium 2002*, number 67 in Laboratory of Acoustics and Audio Signal Processing, pages 195–204. Helsinki University of Technology.

O'Dell, M., Hurme, P., Dufva, H., and Raimo, I. (1990). Prosody of foreign words in Finnish. In K. Wiik and I. Raimo, editors, *Nordic Prosody V*, pages 266–281. University of Turku.

Palander, M. (1987). *Suomen itämurteiden erikoisgeminaatio*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Palomaa, J. (1946). *Suomen kielen äännekestoista puhumaan oppineen kuuromykän ja kuulevan henkilön ääntämisessä*. Department of Phonetics, University of Helsinki.

Perkell, J. S. and Klatt, D. H., editors (1986). *Invariance and Variability in Speech Processes*. Hillsdale, New Jersey: Lawrence Erlbaum.

Pickett, J. M., editor (1999). *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology*. Needham Heights, MA: Allyn and Bacon.

Picone, J., Prezas, D. P., Hartwell, W. T., and Locicero, J. L. (1986). Joint estimation of the LPC parameters and the multi-pulse excitation. *Speech Communication*, **5**, 253–260.

Port, R. F. (1979). The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics*, **7**, 45–56.

Port, R. F., Reilly, W., and Maki, D. (1986). Use of syllable scale timing to discriminate words. *Journal of the Acoustical Society of America*, **83**(1), 265–273.

Port, R. F., Dalby, J., and O'Dell, M. (1987). Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America*, **81**(5), 1574–1585.

Povel, D.-J. (1981). Internal representation of simple temporal patterns. *Journal of Experimental Psychology: Human Perception and Performance*, **7**(1), 3–18.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey: Prentice-Hall.

Raimo, I. and Suomi, K. (1976). Pari kontrastiivista kuuntelukoetta. In T. Tikka, editor, *VI Fonetiikan päivät*, pages 63–68. Joensuu.

Rand, R. H., Cohen, A. H., and Holmes, P. J. (1988). Systems of coupled oscillators as models of central pattern generators. In Cohen *et al.* (1988), pages 333–367.

Rapola, M. (1966). *Suomen kielen äännehistorian luennot.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Ravila, P. (1961). Kvantiteetti distinktiivisenä tekijänä. *Virittäjä*, **65**, 345–350. Résumé: De la quantité comme facteur distinctif.

Remez, R. E. and Rubin, P. E. (1990). On the perception of speech from time-varying acoustic information: Contributions of amplitude variation. *Perception & Psychophysics*, **48**, 313–325.

Sadeniemi, M. (1949). *Metriikkamme perusteet.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Saltzman, E. (1995). Dynamics and coordinate systems in skilled sensorimotor activity. In R. Port and T. van Gelder, editors, *Mind as Motion*, pages 149–172. MIT Press.

Saltzman, E. and Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, **1**, 333–382.

Saltzman, E., Löfqvist, A., Kay, B., Kinsella-Shaw, J., and Rubin, P. (1998). Dynamics of intergestural timing: A perturbation study of lip-larynx coordination. *Experimental Brain Research*, **123**, 412–424.

Sievers, E. (1893). *Grundzüge der Phonetik.* Leipzig.

Sovijärvi, A. (1944). *Foneettis-äännehistoriallinen tutkimus Soikkolan inkeroismurteesta.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1994). *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.30.* Cambridge: Medical Research Council Biostatistics Unit.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). *WinBUGS User Manual, Version 1.4.* Cambridge: Medical Research Council Biostatistics Unit.

Strik, H. and Boves, L. (1991). A dynamic programming algorithm for time-aligning and averaging physiological signals related to speech. *Journal of Phonetics*, **19**, 367–378.

Summerfield, Q. (1979). Timing in phonetic perception: Extrinsic or intrinsic? In W. J. Barry and K. J. Kohler, editors, *"Time" in the Production and Perception of Speech*, pages 169–204. Institut für Phonetik, Universität Kiel.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 1074–1095.

Suomi, K. (1980). *Voicing in English and Finnish Stops.* Turku: Publications of the Department of Finnish and General Linguistics of the University of Turku.

Suomi, K., Toivanen, J., and Ylitalo, R. (2003). Durational and tonal correlates of accent in Finnish. *Journal of Phonetics*, **31**, 113–138.

Thomas, A., Best, N., Arnold, R., and Spiegelhalter, D. (2002). *GeoBUGS User Manual, Version 1.1 Beta.* Cambridge: Medical Research Council Biostatistics Unit.

Thorsen, N. (1984). F_0 timing in Danish word peception. *Phonetica*, **41**, 17–30.

Tuller, B. and Fowler, C. A. (1980). Some articulatory correlates of perceptual isochrony. *Perception & Psychophysics*, **27**, 277–283.

Upton, G. J. G. (1978). *The Analysis of Cross-Tabulated Data.* Chichester: John Wiley & Sons.

Uusivirta, P. (1971). Suomen kielen resonanttien kvantiteettikorrelaation neutraloituminen. Unpublished licentiate thesis, University of Helsinki.

Vatikiotis-Bateson, E. and Kelso, J. A. S. (1993). Rhythm type and articulatory dynamics in English, French and Japanese. *Journal of Phonetics*, **21**, 231–265.

Vihanta, V. V. (1987). Suomen äännekestot ranskalaisen suomenoppijan kannalta. In Hurme and Dufva (1987), pages 101–122.

Vihanta, V. V. (1988). F_0:n osuudesta suomen kvantiteettioppositiossa. In Karjalainen and Laine (1988), pages 13–37. Résumé: Sur le rôle de la F_0 dans l'opposition de quantité en finnois.

Whitehead, A. N. (1925). *Science and the Modern World.* New York: Macmillan. Free Press Paperback Edition 1968.

Whitehead, A. N. (1929). *Process and Reality.* New York: Macmillan. Free Press Paperback Corrected Edition 1978, D. R. Griffin and D. W. Sherburne, eds. Quotations and page numbers in the text follow the Corrected Edition.

Whitehead, A. N. (1938). *Modes of Thought.* New York: Macmillan. Free Press Paperback Edition 1968.

Wiik, K. (1965). *Finnish and English Vowels: A Comparison with Special Reference to the Learning Problems Met by the Native Speaker of Finnish Learning English.* Annales Universitatis Turkuensis. Series B Tom. 94. Turku.

Wiik, K. (1975). On vowel duration in Finnish dialects. In V. Hallap, editor, *Congressus Tertius Internationalis Fenno-Ugristarum, Pars I*, pages 415–424. Tallinn.

Wiik, K. (1985). Suomen murteiden vokaalien kestoista. In Aaltonen and Hulkko (1985), pages 253–317. (On the duration of vowels in Finnish dialects).

Wiik, K. (1988). F_0:n huipun sijainti suomessa. In Karjalainen and Laine (1988), pages 215–229. Abstract: On the location of the fundamental frequency peak in Finnish.

Златоустова, Л. В. (1981). Фонетические единицы русской речи. Москва: Издательство Московского университета.

Index of Citations

Appendix A

WinBUGS model specifications

A.1 Experiment 2

```
model
{
# summary for stimulus recognition and response probabilities
   for (i in 1:K) {
      pp.S[i] <- sum(p.S[1:M,i])/M
      pp.R[i] <- sum(p.R[1:M,i])/M
   }
     for (h in 1:M) {
# summary for subject recognition probabilities
      subj[h] <- sum(p.S[h,1:K])/K
      for (i in 1:K) {
         n[h,i] <- sum(r[h,i,])

         p.R[h,i] <- phi[h,i] / sum(phi[h,])
         phi[h,i] ~ dgamma(alpha[i],1)
# ie. Dirichlet(alpha[ ])
         p.S[h,i] ~ dbeta(1,1)
         r[h,i,1:K] ~ dmulti(p[h,i,1:K], n[h,i])
         rhat[h,i,1:K] ~ dmulti(p[h,i,1:K], n[h,i])
         for(j in 1:K) {
            p[h,i,j] <- p.R[h,j] * (1 - p.S[h,i])
            + equals(i,j) * p.S[h,i]
         }
      }
   }
}
```

A.2 Experiments 3, 4, 5

```
model
{
   for(k in 1:K) {
      for(j in 1:2) {
         mu[k,j] <- alpha+alpha.q*quality[j]
            + alpha.s[k]
            + alpha.qs[k]*quality[j]
         sigma[k,j] <- exp(beta+beta.q*quality[j]
            + beta.s[k] + beta.qs[k]*quality[j])
         for(i in 1:N) {
            rr[k,j,i] <- n[k,j,i]-r[k,j,i]
            rr[k,j,i] ~ dbin(p[k,j,i],n[k,j,i])
            p[k,j,i] <- phi((timing[i] - mu[k,j])/sigma[k,j])
         }
      }
      alpha.s[k] ~ dnorm(0.0,tau.as)
      alpha.qs[k] ~ dnorm(0.0,tau.aqs)
      beta.s[k] ~ dnorm(0.0,tau.bs)
      beta.qs[k] ~ dnorm(0.0,tau.bqs)
      az[k] <- phi((mu[k,1]-mu[k,2])
         / sqrt(sigma[k,1]*sigma[k,1]+sigma[k,2]*sigma[k,2]))
   }
   alpha ~ dnorm(0.0,0.0001)
   alpha.q ~ dnorm(0.0,0.0001)
   beta ~ dnorm(0.5,0.02)
   beta.q ~ dnorm(0.0,0.02)
   tau.as ~ dgamma(0.01,0.01)
   tau.aqs ~ dgamma(0.01,0.01)
   tau.bs ~ dgamma(1.0,0.01)
   tau.bqs ~ dgamma(1.0,0.01)
   az.tot <- mean(az[])
   az.bar <- phi(2*alpha.q
      / sqrt(exp(2*(beta-beta.q)) + exp(2*(beta+beta.q))))
}
```