PERCEPTION OF RHYTHM IN SEQUENCES OF CLICKS AND OF SYLLABLES

by

Janet Mary D'Arcy

B.A., Simon Fraser University, 1983

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in
THE FACULTY OF GRADUATE STUDIES
Faculty of Medicine
School of Audiology and Speech Sciences

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

August 1984

© Janet Mary D'Arcy, 1984
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Audiology and Speech Sciences

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date August 7, 1984
ABSTRACT

The perception of speech rhythm may be affected by two factors, one being a tendency, by speakers, to lengthen syllables as the utterance progresses, the other being a tendency, by listeners, to impose rhythmic structure on speech sequences. Three tests were constructed in which each stimulus consisted of a sequence of six sounds and each test contained seven different time-altered stimuli. The time alteration was of a nonlinear progressive nature. Test I consisted of sequences of six clicks; Tests II and III consisted of sequences of six [ta] and six [na] syllables, respectively. Native speakers of English, French and Japanese were asked to rate the regularity of the sequence as accelerating, regular, or decelerating.

Results indicate that a stimulus may have values of the alteration parameter which cover a large range and yet still be perceived as regular. However, in all cases, the stimulus that would be produced using the values corresponding to a subject's perception of regularity would be acoustically anisochronous: the timing of this stimulus would always be decelerated. No significant difference in performance was found between English, French and Japanese subjects, nor between click, [ta], and [na] stimuli, nor between the six test orders. The DL for irregularity could not be measured accurately, but it was estimated. A further experiment will be necessary to determine this DL more accurately.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td></td>
<td>TABLE OF CONTENTS</td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td></td>
<td>ACKNOWLEDGMENT</td>
<td>vii</td>
</tr>
<tr>
<td></td>
<td>DEDICATION</td>
<td>viii</td>
</tr>
<tr>
<td></td>
<td>Chapter 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Chapter 2 LITERATURE REVIEW</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2.1 Introduction</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2.2 Timing and Rhythm</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>2.3 Timing in Language and Isochrony</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Chapter 3 AIMS OF THE INVESTIGATION</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Chapter 4 EXPERIMENTAL PROCEDURES AND APPARATUS</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>4.1 Pilot Study</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>4.2 Pilot Test I</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>4.21 Stimulus Preparation</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>4.22 Subjects</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>4.23 Test Procedure</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>4.24 Results</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>4.3 Pilot Test II</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>4.31 Stimulus Preparation</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>4.32 Subjects</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>4.33 Test Procedure</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>4.34 Results</td>
<td>51</td>
</tr>
</tbody>
</table>
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Durational Structure (in samples) for the Ten Stimuli of Pilot Test I</td>
<td>37</td>
</tr>
<tr>
<td>II</td>
<td>Durational Structure (in samples) for the [ta] Stimuli of Pilot Test IIa</td>
<td>49</td>
</tr>
<tr>
<td>III</td>
<td>Durational Structure (in samples) for the [na] Stimuli of Pilot Test IIb</td>
<td>49</td>
</tr>
<tr>
<td>IV</td>
<td>Typical Response Matrix for Subject 2 in Pilot Test IIa</td>
<td>53</td>
</tr>
<tr>
<td>V</td>
<td>Typical Averaged Response for All Subjects of Pilot Test IIa</td>
<td>53</td>
</tr>
<tr>
<td>VI</td>
<td>Durational Structure (in samples) for the Click Stimuli of Main Test I</td>
<td>58</td>
</tr>
<tr>
<td>VII</td>
<td>Durational Structure (in samples) for the Speech Stimuli of Main Tests II and III</td>
<td>58</td>
</tr>
<tr>
<td>VIII</td>
<td>Typical Response Matrix for Subject E2 in Main Test II</td>
<td>62</td>
</tr>
<tr>
<td>IX</td>
<td>Typical Averaged Response for All English Subjects of Main Test II</td>
<td>62</td>
</tr>
<tr>
<td>X</td>
<td>Typical Averaged Response for All Tests for Subject E2</td>
<td>63</td>
</tr>
<tr>
<td>XI</td>
<td>Normalized Response Matrix for Subject E2 in Main Test II</td>
<td>68</td>
</tr>
<tr>
<td>XII</td>
<td>Stimulus Number Corresponding to Perception of Regularity</td>
<td>70</td>
</tr>
<tr>
<td>XIII</td>
<td>Results of ANOVA Performed On 18-Subject Group</td>
<td>72</td>
</tr>
<tr>
<td>XIV</td>
<td>Estimated values of Δa for DL Using the Three Best Subjects</td>
<td>83</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Block Diagram for Digital-to-Analog Conversion</td>
<td>39</td>
</tr>
<tr>
<td>2</td>
<td>Block Diagram for Dubbing of Original Tape and Recording of Item Numbers on Test Tape</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>Mean Score and 2-s.d. Range for Each Stimulus of Pilot Test I, Averaged Over Nine Subjects</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>Block Diagram for Recording of CV Syllables</td>
<td>46</td>
</tr>
<tr>
<td>5</td>
<td>Block Diagram for Analog-to-Digital Conversion</td>
<td>47</td>
</tr>
<tr>
<td>6</td>
<td>Standard Deviations for Subject 3 in Pilot Test IIa</td>
<td>54</td>
</tr>
<tr>
<td>7</td>
<td>Curve Produced by Plotting a-values as a Function of their Corresponding Stimulus Numbers</td>
<td>56</td>
</tr>
<tr>
<td>8</td>
<td>Mean Score and 2-s.d. Range for Each Stimulus of Main Test I, Averaged over All Subjects</td>
<td>64</td>
</tr>
<tr>
<td>9</td>
<td>Mean Score and 2-s.d. Range for Each Stimulus of Main Test II, Averaged over All Subjects</td>
<td>65</td>
</tr>
<tr>
<td>10</td>
<td>Mean Score and 2-s.d. Range for Each Stimulus of Main Test III, Averaged over All Subjects</td>
<td>66</td>
</tr>
<tr>
<td>11</td>
<td>Typical Averaged Response Means for Test I</td>
<td>75</td>
</tr>
<tr>
<td>12</td>
<td>Typical Averaged Response Means for Test II</td>
<td>76</td>
</tr>
<tr>
<td>13</td>
<td>Typical Averaged Response Means for Test III</td>
<td>77</td>
</tr>
<tr>
<td>14</td>
<td>Typical Averaged s.d.'s for Test I</td>
<td>78</td>
</tr>
<tr>
<td>15</td>
<td>Typical Averaged s.d.'s for Test II</td>
<td>79</td>
</tr>
<tr>
<td>16</td>
<td>Typical Averaged s.d.'s for Test III</td>
<td>80</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENT

I would like to thank all those who have had a part in this thesis. In particular, I thank:

- André-Pierre Benguerel for giving so freely of his time and advice during the research and writing of this thesis.
- John Gilbert for serving on my committee.
- Malcolm Greig, Chinh Le, Brenda Morrison, and Ronnie Sizto for statistical advice.
- All my subjects for their time and attention.
- Maryalyce McDonald for reviewing a portion of the manuscript and for encouragement.
- Trish Batch and Rebecca D'Arcy for their help in preparation of the manuscript.
- All my friends and family, especially Paul, Rebecca, Jack, Roy, and my Mother for their understanding and support.
This thesis is dedicated to the memory of my father, Owen D'Arcy (1907-1970), who believed in the value of further education.
CHAPTER 1
INTRODUCTION

Time is not a thing that, like an apple, may be perceived. Stimuli and patterns of stimuli occupy physical time; and we react to such stimuli by perceptions, judgments, comparisons, estimates, etc. (Woodrow, 1951)

In speech perception, the ability of the auditory system to determine the order and time of occurrence of serial events is important. It is well known that an acoustic signal has three parameters: frequency, amplitude, and time. These physical parameters are generally highly correlated with various attributes of auditory sensation: pitch, loudness, and perceptual timing; and these, in turn, correlate with the prosodic features of speech: intonation, stress, and rhythm. Traditionally, psychoacousticians have studied pitch and loudness phenomena extensively, but have devoted less effort to the study of temporal phenomena. This neglect may be partially due to the difficulty encountered in measurement: only recently has instrumentation, including computers, provided capabilities to handle all the necessary experimental constraints conveniently. As audition is often considered to be the temporal modality (the ear has a much better time resolution than other sense modalities have), additional attention should be paid to the temporal organization of sound perception and - more importantly for humans - of speech perception.

It has been noted that the suprasegmental patterns of language are always patterns in time (Lehiste, 1971). Suprasegmental features such as stress, intonation, and rhythm depend highly on the acoustic parameters of intensity, fundamental frequency, and duration, respectively. This study is concerned with rhythm, the more or less regular recurrence in time of
auditory or of phonetic events possibly arranged in similar patterns of
duration. One may rate how rhythmic (or regular) a series of sounds is by
assessing the regularity of patterns of durations. If these durations are
equal or near equal, the group of sounds will be perceived as regular or
rhythmic. This has led many authors to propose the existence of some type of
isochrony - that is, the rhythmic organization of speech into more or less
equal intervals (Lehiste, 1977) - for the various languages they have
studied. It appears that many of these unequal (or near equal) intervals are
perceived as equal because their difference from the standard is below
perceptual threshold. A tendency toward isochrony is thus found in the
signal and the signal is perceived as rhythmic, despite certain
irregularities in temporal intervals. However, if one or several of these
intervals differ by an amount exceeding this perceptual threshold, the series
will be heard as irregular.

It must be recalled that duration, intensity, and fundamental
frequency interact in audition. For example, experiments dealing with
minimal audible duration have indicated that the detection of a brief tone
may depend on its intensity: the more intense the signal, the shorter the
signal duration required to reach threshold. Other studies have indicated
that a sound with higher intensity or a sound with higher pitch will appear
longer than one with lower intensity or with lower pitch; this effect
decreases as duration increases (Fraisse, 1982). Furthermore, it has been
shown that when two tonal stimuli are presented with the same intensity level
but one with a greater duration, the stimulus with the greater duration may
be perceived as being louder. The frequency of the tone affects the period
over which the loudness increases: the lower the frequency of the tone, the
greater the period of loudness summation (Fraisse, 1982).

Depending on the particular language, linguistic stress may be coded into any of the acoustic prosodic parameters (duration, intensity, and fundamental frequency). Fry (1955) carried out an investigation to assess the influence of the physical cues of duration and intensity on the perception of linguistic stress in English. His results showed that both duration and intensity ratios are cues for stress; however, the former is a more efficient cue than the latter. Allen (1975) noted that fundamental frequency is even more important for stress perception. He also noted that in English the occurrence of duration, intensity, and fundamental frequency is highly correlated: the syllables that are higher in fundamental frequency are also usually greater in duration and in intensity.

Rhythm may be inferred either from the durations of individual sounds, or from the intervals between the beats elicited by these sounds, depending on the stimuli. For example, in a train of clicks, the equality or near-equality of the intervals between clicks may result in the perception of a regular rhythm. However, in a sequence of syllables, it is unclear whether it is the equality or near-equality of the durations of the syllables, or the equality or near-equality of the interstress intervals which results in the perception of a regular rhythm. Since speech is a continuous process and not merely a succession of discrete sounds produced by independent motor activities, the acoustic cues in the speech signal overlap in time and change with context as do articulatory activities. It is well known that the auditory system appears to be efficient in detecting change and it would appear that the perception of the rhythm of speech must take place in spite of - or perhaps because of - the constantly changing signal.

There are probably several levels of time control in speech
production. Allen (1973) proposed two levels, one for long-term control, the other for short-term control. Long-term control would monitor speech rate (the average rate at which syllables, words, and phrases are spoken), and short-term control would monitor phonetic duration (segment duration within the syllable and possibly syllable duration within the rhythmic phrase). These two levels are not totally independent, for there is evidence that the rate of an utterance may affect segment and/or syllable durations as well as pause durations. A judgment of the accuracy of either one of these two levels of control must take into account the influence of the other level. For example, when subjects produce speech as stimuli for a segment duration study, they should be instructed to attempt to hold their speech rate as constant as possible in order to reduce variability from an unwanted source (i.e. from long-term control). For the purpose of this study, the level of interest is the one for short-term control.

Since the term 'irregularity' has at least two clearly distinct meanings and uses in the context of rhythm, it is important to clarify the sense in which this term is used in the present study. Whereas the term 'irregularity' can refer to a randomly located timing alteration in an otherwise regular sequence, it can also refer to a timing alteration which would progressively shorten or progressively lengthen intervals or durations along the time dimension. In his experiment, Hibi (1983) used the term 'irregularity' in the former sense; in this study, the term will be used in the latter.

The present study is not primarily concerned with the production ability of subjects, but rather with their perceptual ability, more specifically their ability to perceive whether certain types of stimuli are regular or irregular.
CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

Timing is the organization of events according to a schedule which may or may not involve periodicity. For instance, bus routes are organized according to a schedule. The schedule may be periodic; in such a case, the buses would arrive regularly once per hour. The schedule may also be nonperiodic; in this case, the buses would arrive at irregular times depending on the expected volume of passenger traffic. The events of interest to this study are sounds or, more specifically, sounds which are important, in the sense that they have salience. One can say that accented or lengthened sounds have salience. Important sounds may be said to contain the 'beat' - that is, the emphasis or the main pulse - of the (sequentially) organized events. For example, if the event of interest is a syllable, the important sound may be the 'syllable beat'. If, on the other hand, the event of interest is a metric foot - a group of accented and unaccented syllables forming a rhythmic unit - the important sound may be the 'stress beat'.

The perception of timing implies an organization (or a mental construction) of the events. The events may be perceived as a number of elements (i.e. component parts) which are organized into one or more groups (or units).

Four possible types of organization can be envisaged:

1.a periodic, with one salient element (or beat), and one element only per period.

1.b periodic, with one salient element and one or more non-salient elements per period.
2. scheduling (pre-established, but nonperiodic timing).

3. random (no pre-established timing).

Interested in the temporal organization of speech events, Ohala (1973) investigated three hypotheses (or models) which correspond to the organizational types listed above:

1. Speech units are uttered according to an underlying regular rhythm. This model has periodicity (thus scheduling) and corresponds to types (1.a) and (1.b) above, depending on whether all or only some of the elements are salient.

2. Speech units are produced according to an underlying pre-established time schedule which is basically nonperiodic. This model - called the comb model - has scheduling but no periodicity and corresponds to type (2) above.

3. There is no underlying time programme (or rhythm). Speech elements would simply be strung together, one event being initiated only once the previous one has been completed. This model - called the chain model - has neither periodicity nor scheduling and corresponds to type (3) above.

In his study, Ohala mainly discussed the latter two hypotheses; his results will be discussed later. The present study investigates the former hypothesis, namely types (1.a) and (1.b).

There are several levels of organization in speech at which one can speak of rhythmic organization. Two such levels are:

- the discourse level, where the beats of rhythm are associated with the pauses which separate the successive breath groups. This is a higher level which may also be referred to as 'long-term'.

- the phrase level, where the beats of rhythm are the beats associated with each foot or each syllable. This is a lower level which may also be referred to as 'short-term'.

The discourse level of rhythmic organization is not within the scope of this study.

At the phrase level of organization, language rhythm has been said to be determined by the recurrence, succession and coordination of the stress and syllable beats (Abercrombie, 1967). The two earliest types of speech rhythm proposed (Pike, 1945; Abercrombie, 1967) were:

1. syllable-timed language rhythm: the intervals between the syllable beats are isochronous.

2. stress-timed language rhythm: the intervals between the stress beats are isochronous.

If one of the beat sequences is isochronous, the other cannot be: in a stress-timed language, the syllable beats are unevenly spaced and in a syllable-timed language, there may be no stress beats (e.g. in French) and if they exist, the stress beats are unevenly spaced.

Other types of language rhythm are also possible and have been proposed (e.g. tone-timing, mora-timing), but will not be elaborated on here.

Since the major interest of the present investigation is the rhythmic aspect of the perception of 'short-term' sound sequences, this study deals with the phrase level of organization and endeavours to investigate various aspects related to both types of periodic organization mentioned above, that is, types (1.a) and (1.b). These two types of periodic organization correspond to syllable-timed and stress-timed languages.

As previously stated, events are organized into temporal patterns,
or groups, and these patterns may be perceived as having rhythm. In a
discussion of rhythm, Fraisse (1982) stated

The task of those who study rhythm is a difficult one, because a precise,
generally accepted definition of rhythm does not exist. This difficulty
derives from the fact that rhythm refers to a complex reality in which
several variables [or aspects] are fused. (Fraisse, 1982, p.149)

In a prior study, Fraisse (1963) discussed two important aspects of
rhythm, which are also of interest to the present study. One such aspect,
which has already been discussed to some extent, is that of periodicity: the
recurrence of a beat at regular intervals of time. The second aspect is that
of structure: the recurring units all have a pattern with an analogous
structure.

The rhythm of some phenomena appears to be construed primarily
through an awareness of one or the other of these two aspects. For example,
the perception of the rhythm of day and night alternation or of a military
march is through an awareness of the periodicity of these phenomena. On the
other hand, the perception of the rhythm of seasons during the year, of a
waltz, or of the feet of a verse is through an awareness of the previously
learned structure of recurring structures.

The perception of periodicity and of structure implies the
perception of duration. Fraisse (1963, p.76) noted that "The perception of
duration is that of the duration of an organization." Duration perception
requires (among other things) the comparison of 'segments' or of
'intervals'. However,

...a succession is characterized not by its elements, either stimuli or
intervals, but by the scheme of its durations:...the identity of the
temporal patterns was what made it possible for rhythms composed of
different 'elements' to appear similar. (Fraisse, 1963, p.77)

The durations of segments or of intervals may be manipulated in
order to investigate any effects on the perception of timing. For example, experimenters have used duration manipulation in various experimental designs in order to assess the 'psychological present',

...the physical time over which stimuli may be spread and yet all be perceived as present. This time has sometimes been termed the temporal span of attention. (Woodrow, 1951, p.1230),

and/or the 'indifference interval', a time span within which temporal discrimination is most accurate. Similar manipulations and their effects on the perception of timing in language are currently being investigated extensively. Discussion of the above effects occurs in the remainder of this chapter.

2.2 Timing and Rhythm

In an article on rhythm, Fraisse (1982, p.150) explained that "...rhythm is a perceptual quality specifically linked to certain successions." That is to say, one imposes a rhythm on various elements, but such an imposition may be influenced by certain objective factors. Indeed, subjective factors, individually or in conjunction with objective ones, may intentionally or involuntarily determine the nature of the perceived rhythm.

The perception of the most simple rhythm may occur when the subjective factors entirely determine the grouping.

The most easily perceived rhythm is one that is produced by the simple repetition of the same stimulus at a constant frequency. In the rest of this article, we will call this a 'cadence'. The simplest examples are the beating of a clock or of a metronome. But the most important fact is that these rhythms are characteristic of some very fundamental activities such as walking, swimming, and flying. Both animals and people move about with rhythmic movements characteristic of their species. The first rhythmic movement found in the human newborn is sucking, with periods that follow at intervals of from 600 to 1200 msec. This regularity is interrupted by spontaneous pauses, but sucking movements occur at a cadence that seems to be characteristic for each infant. (Fraisse, 1982, p.151)
The fundamental activities mentioned above, including those of man, are periodic activities which have their own 'spontaneous' rate of sequence, the natural or voluntary rhythm that results when we are allowed to perform an activity at our own rate. The range of spontaneous rates is 0.5 to 2.0 per second: this frequency depends on the muscular mass concerned (Fraisse, 1982). Spontaneous rate is often measured by the natural speed of tapping. The interval length between two taps varies from 200 to 1400 msec with the most representative interval being 600 msec. There is great interindividual variability of this rate but little intra-individual variability: the intra-individual variability is only 3 to 5%, which is in the range of the differential threshold for durations of this type. The small amount of intra-individual variability may indicate that spontaneous rate is characteristic of the individual (Fraisse, 1982).

Spontaneous rate is often compared with 'preferred' rate, a regular succession judged as being neither too slow nor too fast. If a sequence of sounds, such as clicks uniformly spaced and equally loud, is presented, the preferred rate of sequence increases with the number of elements in the group. The range of preferred rates, indicated by the time between successive clicks, is 0.2 to 1.3 seconds (Woodrow, 1951). The upper limit of our capacity for perception of succession is five to six elements. However, if these identical sounds are grouped or form a unit of significance (e.g. in two's, three's, four's, or five's), far more elements can be perceived: one may perceive four or five groups of these sounds (Fraisse, 1963). This grouping of groups is most likely related to the 7±2 of Miller (1967) and to the notion of 'chunking'.

Spontaneous rate, which pertains to a motoric phenomenon, and
preferred rate, which pertains to a perceptual phenomenon, have comparable frequencies and are often associated; a subject generally finds it fairly easy to accompany a regular series of sounds with a motor act.

This accompaniment tends to be a synchronization between the sound and a tap - that is to say, that the stimulus and the response occur simultaneously. [...] A similar behavior is possible only if the motor command is anticipated in regard to the moment when the stimulus is produced. More precisely, the signal for the response is not the sound stimulus but the temporal interval between successive signals. (Fraisse, 1982, p.154)

The perception of a more complex rhythm may occur when objective characteristics determine the grouping. The characteristics of the stimulus sequence may largely determine the nature of the perceived grouping: the relative intensities of the group members, their absolute and relative durations, and their temporal spacing are probably the most important characteristics. Experimenters may manipulate any of these three characteristics and then attempt to answer two types of questions: (a) What is the nature of the effects principally produced by manipulating these characteristics? (b) What are the possible durations of rhythmic groups (Fraisse, 1982)?

The first of these two areas experimenters have investigated is the nature of the effects produced by manipulation of various objective characteristics. Although any differentiation in an isochronous sequence of identical elements may serve as a basis for grouping, each one does not have the same effects as far as the organization of the temporal sequence is concerned.

An element of greater intensity, for example, may determine the nature of a grouping. The objective accent is situated most spontaneously at the beginning of the pattern: a regular succession of a strong and of a weak
sound of equal duration is perceived as a succession of trochees (strong-weak) 60% of the time and as a succession of iambs (weak-strong) 40% of the time (Fraisse, 1982).

An element of greater duration may also determine the nature of a grouping.

In general, a noteworthy lengthening of the duration of a sound or of the interval between two sounds determines the end of a group; this longer duration allows one to distinguish between two successive patterns. (Fraisse, 1982, p.159)

Such a lengthening creates a rupture between two groups.

Woodrow (1951) summed up the effects of these two characteristics: a regularly recurring, relatively greater intensity exerts a group-beginning effect, while a regularly recurring, relatively greater duration exerts a group-ending effect (if the temporal spacing is equal). This latter effect is particularly important to the discussion of the present study.

The objective characteristic of temporal spacing may be best investigated in conjunction with intensity. For example, when the temporal spacing and the relative intensities (referred to as loudness in the following quotation) are considered; Woodrow explains:

It is possible to determine at how slow a rate certain factors producing rhythmical grouping lose their force. One such factor is a difference in the relative loudness of sounds composing a rhythmical series. For example, if every second sound in a series of equally spaced sounds is louder than the others, and if the rate of sequence of the sounds is moderately slow, the sounds tend to be perceived in trochaic rhythm, that is, in groups of two with the louder sound the first member of the group. However, when the sounds succeed each other at a very slow rate, this segregating effect of loudness may vanish: the subject no longer reports any grouping at all. (Woodrow, 1951, p.1230)

According to Woodrow, this grouping seemed to disappear if the temporal spacing - the interval between the sounds - was greater than 3.5 seconds. Fraisse (1963) also observed a limiting value for the temporal
spacing of two stimuli, but stated that the rhythmic structure disappeared when the interval between sounds was approximately 2 seconds. This value would be expected to vary depending on the stimuli and the conditions.

Again, experimenters may manipulate objective characteristics in order to investigate a second area: the possible durations of rhythmic groups. As already alluded to above, there are certain limits to the durations of the perceived groups. According to Woodrow (1951), the study of rhythm - the main interest of this paper - affords an indirect way of attempting to determine the upper limit of the psychological present.

When a subject perceives a sequence of elements as a sequence of groups of elements (e.g. trochees), the perception is described as what is meant by rhythm in the psychological sense.

The successive groups are ordinarily of similar pattern and experienced as repetitive. Each group is perceived as a whole and therefore has a length lying within the psychological present. (Woodrow, 1951, p.1232)

Experiments on the perception of timing have attempted to estimate the limits (both the upper limit - maximum duration - and the lower limit - minimum duration) of the psychological present.

The question may be better phrased, perhaps, as that of the maximal physical time over which may extend a temporal stimulus pattern, the successive parts of which are perceived as a whole possessing a unitary property of duration. There is also a minimal time which constitutes the lower duration threshold. (Woodrow, 1951, p.1230)

As already stated, the nature of the stimuli may affect the size of this time span, the upper limit of which was found to lie between 2.3 and 12 seconds (Kastenholz, 1922, cited in Woodrow, 1951). At the lower limit of this time span, a stimulus of short duration results in a unitary experience but since the stimulus is perceived as instantaneous, it is felt to lack the quality of duration. Therefore, there must exist a minimum time limit for duration:
this minimum amount of time appears to vary between 0.01 and 0.12 seconds, again depending on the stimulus and the conditions (Woodrow, 1951).

If the temporal length of the stimulus lies within the psychological present, then the timing may be said to correspond to the short-term level. It is the perception of this short-term level in which this study is interested.

It has been noted that a rhythmical measure (e.g. syllable, foot) is assumed to constitute a perceptual unity and that this notion seems to be in agreement with results obtained through direct introspections of temporal unity (Woodrow, 1951).

In addition to investigating the limits to the durations of rhythmic groups, one may also investigate the accuracy with which a duration is estimated. The perception of time has been studied by many researchers attempting to discover the size of the 'indifference' interval. The indifference interval was investigated by Vierordt in 1868 (cited by Woodrow, 1951, p.1225) who first noted that "...short intervals are overestimated and long ones are underestimated." A large range of indifference intervals has been reported, with the intervals of 0.5 to 0.7 seconds perhaps being reported most frequently.

Since it is for the indifference interval that the most accurate temporal discriminations are found, the use of this temporal interval appears most appropriate when one attempts to measure the just noticeable difference (jnd), or difference limen (DL) for duration, namely the smallest temporal difference one is able to detect. In 1933, Blakely (cited in Woodrow, 1951) used empty intervals bounded by clicks to measure relative jnd's of interval length and found that discrimination was most accurate at intervals of 0.6 to
0.8 seconds, with the relative jnd being just under 8% of the standard. After reviewing the literature, Woodrow (1951) noted that the relative jnd for durations was almost the same as for intervals. He stated that the differential threshold for sound interval and sound duration had been determined:

In general, the just noticeable difference is smallest at some relatively short magnitude or range of magnitudes and increases both above and below this middle region. (Woodrow, 1951, p.1224)

The relative jnd was slightly greater with shorter and longer standards; however, from 0.2 to 1.5 seconds, it was still less than 10%. Using even longer standards, from 2 to 30 seconds, the jnd rose from 16% to 30%. The accuracy for reproduction of these intervals and durations (for instance, reproduction of an interval by taps on a reaction key) was similar to the accuracy for discrimination, with the greatest reproduction accuracy extending from 0.2 to 2.0 seconds (Woodrow, 1951).

The perception of temporal phenomena is not fully understood. However, any experimental design must consider the two types of factors discussed above, which affect temporal judgment of intervals and of durations. Some of the objective factors are intensity, quality, length of preparatory interval (e.g. amount of time between item number and stimulus), length of bounding elements (in the case of intervals), range of times used, and practice effects. Some of the subjective factors are judgment strategies, attention, interest, motivation, and emotional state. Since time perception is affected by many factors, some of which (e.g. emotional state) are not easily kept under control and often make the study of time estimation using a strict experimental approach difficult, care must be taken in the experiments proposed for this study in order to minimize any undesired
effects caused by the above factors.

Temporal stimuli are ordinarily of two sorts: (1) what are called empty intervals, for example, two flashes of light or two short sounds separated by a period of time; and (2) sounds or lights lasting continuously over a period of time. (Woodrow, 1951, p.1224)

Both sorts of stimuli are of interest in this study. Hereafter, the first type is referred to as an 'interval' and the second type is referred to as a 'duration'. When reference is made to either type in the present paper, it is to be assumed that the stimuli are composed of sounds unless stated otherwise. Furthermore, the types of sounds to be used as stimuli in the experiments carried out for this study are nonspeech sounds for intervals and speech sounds for durations.

Studies dealing with the perception of timing have used both 'multiple' and 'single' temporal stimulus paradigms; however, it is the 'single' stimulus (composed of six elements) paradigm which has particular importance for this paper since it is the type to be used in the present experiments.

In a 'multiple' stimulus (which may be single- or multi-element) paradigm, two time intervals or durations are presented before a subject makes a judgment. For example, the experimenter presents two intervals; the subject then compares them and makes a judgment as to their relative length.

In a 'single' stimulus (which also may be single- or multi-element) paradigm, one time interval or duration is presented without the presentation of a second interval or duration with which to compare the first. For example, the experimenter presents a single interval and the subject makes a judgment as to its length. It should be noted that Woodrow (1951) reviewed various studies and concluded that when judging a single temporal stimulus
...the terms short and long probably always carry the implication of a comparison with a standard even though the subject might be unaware of it. This standard is largely a function of recent temporal experiences, or of a composite trace left by them. (Woodrow, 1951, p.1229)

As already stated, the experiments for this paper will use the single stimulus paradigm; however, the subject will be given the terms 'regular' and 'irregular' for comparison (instead of 'short' and 'long'), with 'irregular' being subdivided into 'accelerating' and 'decelerating'. It may be assumed that these terms also carry the implication of a comparison with a standard whether the subject is aware of it or not.

In a similar study on the rhythmic perception of repetitive sound sequences, Hibi (1983) dealt with the occurrence of one or two random irregularities in a group of clicks. He investigated the amount of temporal distortion which must intervene in a sequence of uniformly spaced sounds in order for a listener to be able to report an irregularity:

...the ability of listeners to report whether there was a distortion in a sequence as a function of the rate of succession did not vary in the regions of from 1 to 2.5 times per second and from 4 to 7 times per second. (Hibi, 1983)

It should be noted, however, that the difference in detection values between these two regions was significant.

The picture emerging from these results, then, is of the subjects' accommodating to the temporal sequences which have different rates of succession ranging from 0.7 to 7.0 times per second. (Hibi, 1983)

As indicated in Chapter 1, this random type of irregularity must be differentiated from the nonrandom type of irregularity such as the gradual lengthening or shortening of intervals that may occur in a grouping of clicks or other sounds. Again, it is this latter type of irregularity which is investigated in the current study, as such irregularities seem to be important and observable in most types of speech. It is well known that a
trend toward lengthening of syllables as the utterance progresses has been found in most languages. Despite the presence of this nonrandom type of irregularity, and despite hesitations and other pauses which tend at times to conceal or to disguise the rhythm, speech rhythm has important implications for perception. However, Hibi's finding regarding the difference in detection values between a low and a high rate of succession is interesting in that the finding led him to propose two different types of processes which may occur in rhythmic perception.

One type of process appears to be an 'ongoing' or 'one-by-one' processing which takes place when the rate of succession is low (slower than three times per second). In this type of processing, a listener would expect the next element at a predictable time and then compare the perception of this next element with his expectation; each element is processed separately. This mechanism would account for the fact that negative compensation among neighbouring time intervals results in the perception of a regular rhythm.

The other type of process appears to be a 'holistic', 'global', or 'Gestalt-like' processing which takes place when the rate of succession is high (faster than three times per second). In this sort of processing, there would not be sufficient time to process in a one-by-one manner, thus the regularity of the timing pattern would be evaluated in two steps:

First, it is necessary to postulate a regular temporal pattern, that is, the pattern of time intervals that would yield an even tempo, and then to detect a departure from these regular values in the observed intervals by pattern-matching or similar routine. (Hibi, 1983)

In this way, the sequence is processed as a whole.

Both of these processes could occur at the short-term level. However, it is the latter process which is of particular interest to the
present study since our stimuli are based on a syllabic rate which is greater than three times per second.

2.3 **Timing in Language and Isochrony**

Since language appears to have the same rhythmic constraints as other motor behaviours, limits on the types of language rhythms to be expected are set by these constraints. Allen (1975) suggested that:

...[language rhythms] should be simple in structure, confined largely to successions and alternations, depending on the relationship between syllables and stress accent in the language; the rate of succession of syllables and rhythmic groups should be in or near the range of 0.2 to 1.0 per second; and the variability with which the time program of the rhythm is realized should conform to the variability of other skilled motor acts. (Allen, 1975)

In addition, these constraints impose more subtle, unspecified limitations for a particular language rhythm, which may be due to certain aspects of articulatory timing which must be learned.

A claim made by some authors (Allen, 1975; Fraisse, 1982) is that the natural rhythms of our movements reinforce the tendency to impose rhythm on language and other human acts. These authors often cite a study by Miyake (1902) which looked at the rhythmic structure of various motor behaviours. Miyake came to two conclusions:

1. it is impossible not to act rhythmically and
2. simple successions and alternations are most prevalent to our movements.

Allen (1975) and Fraisse (1982) further cited studies which showed that when we are performing a motor task at our spontaneous rate for speech and other motor activities, the average limits are approximately 0.2 and 2.0 seconds between acts. The difference in these spontaneous rates of
succession depends on who we are, what we are doing, and whether we are reproducing a given interval or producing our own interval (variability is less if the rate is our own). These figures would be comparable to the preferred rate discussed in Section 2.2.

Classe (1939) was one of the first to propose that isochrony was present in the rhythm of English prose. He claimed that isochrony was probably the essential characteristic of rhythm. However, Classe appreciated that perfect isochrony could be realized only under very special and definite conditions and that these conditions were seldom fulfilled in ordinary speech.

Due to the variation in interstress intervals, many studies have rejected the idea of strict isochrony at the acoustic level, or they have simply stated that perfect isochrony cannot be found (Classe, 1939; Shen and Peterson, 1962; Allen, 1968; Lea, 1974; etc.). Hoequist (1983a) claimed that, due to other constraints, speakers do not usually attain their goal of altering production in order to move toward isochrony. Indeed, strict isochrony in the acoustic signal does not appear to hold for any of the languages examined. The results of Hoequist's study support a weaker version of isochrony: a tendency toward isochrony seems apparent, but isochrony is affected by other processes such as prepausal lengthening (lengthening before a pause, for instance, word, phrase, utterance), changes in rate, etc.

In 1969, Crystal proposed a theory of 'subjective isochrony', namely that rhythmic regularity is 'read into' the perceived utterance. Lehiste (1977), in her review of the literature on isochrony, reached the same basic conclusion. She concluded that isochrony is primarily a perceptual phenomenon imposed on acoustic anisochrony. Lehiste further concluded that isochrony is integrated into the grammar of English at the syntactic level:
large enough deviations from the expected isochrony may serve to signal a syntactic boundary. These conclusions would agree with Fraisse's (1982) explanation of rhythm, in general, being a perceptual quality (see Section 2.2).

Lehiste (1977) writes that many durational differences in interstress intervals may be below perceptual threshold: furthermore, listeners tend to impose a rhythmic structure on speech and therefore may perceive isochrony even in sequences with differences above perceptual threshold.

Ladefoged (1975) noted that different languages had different rhythms due to their variation in use of stress. For example, the rhythm in French has a great evenness as only the phrase-final syllable has a stress different from the other syllables. From this viewpoint, French is a syllable-timed language with syllables tending to recur at regular intervals. However, in English, stress has a dual role: it carries grammatical meaning and organizes rhythm. Therefore, the syntactic and rhythmic patterns interact, often making the stress patterns difficult to describe.

With regard to these different rhythms, the durational characteristics of speech are often proposed as a basis for classifying any language into a particular rhythm or timing category. Three of the categories proposed most often are stress-timing, syllable-timing, and mora-timing, which would have a trend toward equal durations of feet, syllables, and morae, respectively. Much has been written regarding the possible existence of these categories and it is often assumed that English, French, and Japanese are examples of stress-, syllable-, and mora-timed languages, respectively.
When discussing the temporal patterns of speech, some authors (e.g. Lehiste, 1977) use the term isochrony to refer to stress-timed languages only, while others (e.g. Hoequist, 1983a; 1983b) use the term to refer to all languages, regardless of the timing category. This paper intends to use the term isochrony to refer to all language timing categories.

In a study on the location of rhythmic stress beats in English, Allen (1972a; 1972b) gave evidence for English being a stress-timed language: both the actual beat locations and the correlation of accuracy with degree of stress supported this notion.

As noted earlier in this section, Lea (1974) found evidence against strict isochrony in English: stressed syllables show a tendency to occur at intervals of 400 to 500 ms; however, the number of unstressed syllables between stresses very much affects these intervals and causes large interval variation. In his study on prosody, he concluded that

...the concept of English being a stress-timed language is not simply exhibited by exact equality of interstress intervals, or even by an unquestionable 'tendency toward equality' of interstress intervals regardless of other factors. We found that, contrary to several published hypotheses, the average interstress interval increases about linearly with the number of unstressed syllables between the stresses. A tendency toward stressed-unstressed alternation was exhibited, and it is probably this tendency, plus the somewhat uniform durations expected for unstressed syllables, that yields the tendency for interstress intervals to cluster somewhat near an average of 400 ms or so. (Lea, 1974)

This stressed-unstressed alternation may be supported by evidence from speech production. Since an economical model of speech predicts a parallel rhythm pattern for perception and production, Stone (1981) measured jaw movement during speech production of nonsense syllables to see if a rhythm pattern was present. Her attempt to find a physiological correlate of stress level in the jaw displacement signal was not successful, as
displacement was affected by nonrhythmic variables. However, the opening velocity of vowels gave cues for three levels of stress: unstressed, stressed, and 'beat' syllables. Furthermore, a rhythm pattern due to the alternation of these stress levels (i.e. alternating beat and non-beat syllables) was present.

The presence of the 'compression effect' (or of temporal compensation) at the foot level - the shortening of syllable durations as the number of syllables in a foot increases - has been suggested by various investigators (Allen, 1968; Lehiste, 1970; Huggins, 1972; and Klatt, 1976). In a cross-linguistic study, Hoequist (1983a) investigated whether duration was altered differentially depending on the timing category to which a particular language belonged. He suggested that a stress foot in a stress-timed language consists of a foot-initial syllable which carries the stress 'beat', followed by zero, one, or more, less-stressed syllables, up to but not including the next stress 'beat'. He referred to the compression effect at the foot level as evidence for the existence of stress feet. This compression effect allows maintenance of even or near-even foot durations.

In a study on hierarchical (rhythmic) versus serial structure in behaviour, Martin (1972) noted that syllables in a string are timed relative to the whole pattern rather than simply concatenated. This would provide evidence for preset timing and would favour central timing mechanisms during production, rather than peripheral timing mechanisms. He hypothesized that the accented syllables in English are planned first, since the accented elements dominated the temporal organization of an utterance. Lower level, less-stressed syllables are planned later. The accented syllables would appear to be the main targets in such a programme. This view would tend to
support the existence of the compression effect in that the timing of the stressed syllables is most important and therefore takes priority. The timing of the less-stressed syllables is less important and, therefore, susceptible to compression in order to maintain the timing of the stressed syllables.

The compression effect should not occur at the foot level in a syllable-timed language, as presumably no foot structure is present and syllables will therefore tend toward equal durations. Nor should this compression effect occur in a mora-timed language as the morae will also tend toward equal durations.

In another study, Hoequist (1983b) had subjects produce sentences containing multisyllable target words (with stress occurring in most positions in these target words) in order to test durational characteristics of these same three timing categories. His conclusion was that morae and stress feet influenced syllable duration in Japanese and English, respectively; however, Spanish showed no language-specific durational characteristics resulting from syllable-timing (assuming that Spanish is a syllable-timed language). In addition, stress-timed languages showed a small trend toward medial syllable shortening - perhaps due to stress context. These results led Hoequist to suggest that Japanese may be a duration-controlling language; English may be a duration-compensating language; and Spanish may be neither. He noted that other results supported these differential durational characteristics according to the timing category of the particular language involved (e.g. Delattre, 1966, cited in Hoequist, 1983a).

In a discussion of unemphatic stress in French - a prominence of
some kind which falls on the last syllable of an isolated word or the last syllable of each group of words in a longer utterance - Benguerel (1970) noted that the prevailing factor was vowel duration. Referring to these groups of words as rhythmic groups, sense groups, or breath groups, he showed that in his data, obtained for French subjects, the last vowel of the rhythmic group had an average length 50% greater than that of the next-to-last vowel. In a study of English, Klatt (1975) found this phrase-final lengthening phenomenon to be an average of 30%.

This unemphatic stress, or prepausal lengthening (a tendency toward longer durations as one gets to the end of an utterance, e.g. word, phrase, sentence), does not appear to be an exclusive characteristic of a particular timing-type. However, the amount of final lengthening that occurs in a particular timing-type could indicate different degrees of durational control. This may produce indirect evidence with regard to the timing category of a particular language.

Allen (1975) indicated that different rhythmic patterns result in differences in phonetic 'shape' between accented syllables and unaccented ones. In English, unstressed syllables lose their phonetic shape as they are 'reduced' in both quality and quantity. However, in French and Japanese, the unstressed syllables or morae do not lose their phonetic shape. This would indicate that the rhythmic pattern is tied to syllables, morae, and accents.

The claim has been made that English appears to have mainly a rhythm of alternation since it usually has one or more unaccented syllables per foot (Allen, 1975). On the other hand, French and Japanese appear to have mainly a rhythm of succession since stress plays a weaker role in these two languages and they tend to group the syllables or morae into sequences of equals (Allen, 1975).
Abercrombie (1967) noted that it was the way stressed and unstressed syllables succeeded each other that produced the rhythm of a language. This rhythm is one of the most fundamental aspects of the language for it is one of the earliest learned by an infant, one of the last to disappear in various forms of aphasia, and one of the most difficult for an adult to modify when learning a foreign language.

It has been suggested that identical rhythmic sequences are perceived differently by speakers of different languages as a consequence of the rhythmic pattern of their particular language (Jakobson, Fant, and Halle, 1963). Allen (1975) added that different experimental situations might result in a subject's language having different effects on his perceptions, perhaps as a function of whether or not the subject was listening in the 'speech mode'.

As discussed in Section 2.1, it appears that when attempting to decide whether a particular language has a rhythm of succession or a rhythm of alternation, one must specify the hierarchical rhythmic level being considered, for the type of rhythm may be a function of this level. For instance, at the foot level, English may be said to have a rhythm of succession; however, at the syllable level, English may be said to have a rhythm of alternation. These two levels (foot and syllable) appear to be sub-levels of the phrase level of organization discussed earlier.

It must be emphasized that there is evidence that the phrase (or short-term) level and the long-term level, which is not within the scope of the present paper, interact: languages may differ in the way in which rate affects segment duration (Lehiste, 1970). In some languages (e.g. English), an increase in speech rate may be achieved by shortening unstressed
syllables, while in other languages (e.g. French), an increase in rate may be achieved by shortening each syllable more or less proportionately over the entire speeded up utterance.

In addition to the two hierarchical, central components, namely short-term and long-term control, Allen's (1973) and Ohala's (1973) models for durational control had peripheral sources of imprecision: neuromuscular transmission, articulator and aerodynamic constraints, and measurement error. Allen noted that the effects of the peripheral sources of imprecision on speech duration could account for much of the individual variation that is encountered in the measurement of the various language rhythm categories.

To compensate for individual variation, the listener appears to impose a rhythmic structure on the speech signal, as long as the timing is not too distorted. The goal of the speaker appears to be the production of a sound pattern which fits an auditory perceptual target. The fulfillment of this goal results in speech being perceived as regular by the listener.

Since it is the 'important' sounds or 'beats' of the organized speech events which listeners appear to use in order to impose a rhythm on language, various studies, which endeavour to determine rhythmic beat location are discussed below.

In an attempt to validate various findings related to 'units of perceptual processing' (e.g. segment, syllable, word, and foot level), Barry (1983) used a judgment task in which subjects were required to indicate the location of a click in a string of nonsense syllables. With particular regard to stress-timed languages, he sought evidence for perceptual units below the major syntactic group level. His working hypothesis was that
Clicks located objectively 'within' a certain unit of perceptual processing will tend to be placed subjectively within that unit, but will not be differentiated as to their location within the unit. (Barry, 1983)

Since the location of click placement was random for segments, he concluded that segments were unlikely to be perceptual processing units. However, localization within the syllable and the foot resulted in fewer placement errors, supporting the claim that both the syllable and the foot were possible processing units, with the syllable being perhaps the basic unit.

In an investigation of the perception of timing in conversational speech (English sentences), Huggins (1972) previously had reached a conclusion similar to Barry's. Huggins proposed a 'vowel onset' hypothesis where compensation occurs between the onset of stressed vowels and, therefore, these stressed vowel onsets tend toward invariance.

Huggins' vowel onset hypothesis would be compatible with the 'syllable beat' hypothesis proposed by Allen (1972a; 1972b). To examine the perceived location of rhythmic stress beats in English, Allen used click placement tasks and English sentences. Subjects performed three different tasks:

1. tapped their fingers in time to the beat of a given syllable in a sentence.
2. moved a click to match the beat of a given syllable in a sentence.
3. judged whether a click 'hit the beat' of a given syllable in a sentence.

Allen concluded that the beats were closely associated with stressed-syllable vowel onset but preceded it by an amount positively correlated with the degree of stress.
In a paper on speech perception, Morton, Marcus, and Frankish (1976) introduced the concept of the 'perceptual center' (P-center) of a word as "its psychological moment of occurrence", that is, a point which must be temporally equidistant from corresponding points in surrounding words in order for the sequence to be perceptually isochronous. They realized that this point would not necessarily be the acoustic onset of the word (or of the vowel). By presenting pairs of digits, they found that sequences which were acoustically isochronous were not perceptually isochronous (i.e. sequences with regular acoustic onsets were not perceptually regular). When subjects adjusted the sequence timing until they perceived isochrony, they consistently introduced acoustic anisochrony. In a follow-up study on the P-center, Fowler (1979) found "...that listeners judge isochrony based on acoustic information about 'articulatory' timing rather than on some articulation-free acoustic basis." The manner class of the initial consonant had acoustic consequences on the articulatory onset (Fowler, 1979).

The 'P-center' of a syllable appears to correspond to the 'beat' of a syllable and both of these locations are affected by the syllable-initial consonant. It would seem that the P-center and the beat are measurements of the same phenomenon.

According to Allen (1975), the perceived rhythm of English may be due to two tendencies:

1. a tendency to impose a rhythmic structure on any sequence of intervals.
2. a tendency to adjust the perception of interstress intervals toward some average or central duration.

These two tendencies would lead to perceptual isochrony, despite the existence of acoustic anisochrony.
In his paper on temporal regulation of speech, Ohala (1973) looked for support for either the chain or the comb model in speech timing. These models appear to correspond to Kozhevnikov and Chistovich's (1966) closed and open loop models, respectively. Ohala concluded that the speaker must adhere to a pre-programmed time schedule for short spans of speech (e.g. one or two syllables) but that such is not the case for longer stretches. This would correspond to his proposed hybrid model: the comb model for short-term timing and the chain model for long-term timing.

Speech rhythm is often said to be experienced as a rhythm of motor movement (Abercrombie, 1967). In as much as the present study is primarily concerned with the perception of a regular rhythm, not the production, it must be partially assumed the two can be separated, or at least observed independently. If rhythm is perceived through neural interactions between areas of the brain which govern production and perception, perhaps there is a common mechanism underlying both the production and the perception of a temporal sequence. Such an idea is very akin to the motor theory of speech perception and most theories seem to be based on this premise. The motor theory hypothesizes that auditory patterns are interpreted in terms of the articulatory patterns that would elicit them. The process is not presumed to be a conscious process, but a rapid, computer-like cross-check performed subconsciously by the sections of the nervous system that deal with auditory, speech-motor, and linguistic functions.

Two phenomena (already described in this chapter) have important implications for the experiments to be carried out in this study:

1. the trend toward lengthening of syllables as an utterance progresses.
2. the tendency of acoustic isochrony of speech sequences not to result in perceptual isochrony.
In other terms, speech may have timing irregularities which are perceived as regular. Since this study is interested in the difference limen (DL) for timing irregularity, both of the above phenomena must be taken into consideration in order to discover: (a) what degree of irregularity is required for a stimulus, which contains elements of progressive irregularity, to be perceived as regular; (b) what increment or decrement in the degree of irregularity (for accelerating and decelerating stimuli) is required for the detection of an irregularity.
CHAPTER 3
AIMS OF THE INVESTIGATION

This thesis proposes to study the ability of an individual to perceive a progressive irregularity in the rhythm of a repeated nonspeech or speech sound sequence. For this purpose, two types of experiments will be carried out. In the first type, each stimulus will consist of a train of clicks. In the second type, each stimulus will consist of a sequence of concatenated CV (consonant-vowel) syllables. These speech stimuli will use a basic CV syllable, monotone and of constant amplitude, in an attempt to control some of the variables (mainly fundamental frequency and intensity) which can affect the perceived quality of speech sounds. Since this study deals with aspects of rhythmic structure that do not involve pitch and loudness, we are manipulating only the time dimension, and although this type of manipulation may affect pitch and/or loudness, any such effect is assumed to be minimal.

Experiments have been carried out (Hibi, 1983) to assess the degree of temporal distortion that is required in a repetitive sound sequence (trains of clicks) in order to perceive a randomly located irregularity. In the experiments proposed here, stimuli will be altered to varying degrees by either shortening or lengthening gradually the intervals between clicks or between syllable beats. These experiments have two main objectives, based primarily on two tendencies discussed in Chapter 2, namely (1) the tendency, by speakers of most languages, to lengthen durations (of segments, syllables, words) as their utterance (word, phrase, sentence) progresses; (2) the tendency, by listeners, to impose a rhythmic structure on sound sequences.
The first objective of the study is to determine the degree of irregularity required for the listener to perceive a repetitive sequence (of clicks or of syllables) as regular. Since there is a trend toward lengthening of syllables as the utterance progresses and since listeners tend to impose a rhythmic structure on sound sequences, the stimulus which is perceived as most regular is not expected to be the one which is acoustically regular, but one which is decelerated. Hence, the intention is to discover the degree of deceleration that a subject perceives as regular.

The second objective of the study is to determine, with respect to the above perceived regularity, how large an increment or decrement in the degree of irregularity is required for the listener to perceive that a repetitive sequence (of clicks or of syllables) is no longer regular.

In other terms, we are interested in a measure of the degree of irregularity required for the perception of regularity and also in a measure of the DL for timing irregularity.

In addition to the two main objectives of these experiments, there are two others. A third one is to determine if the nature of the stimulus has an effect on either of the two measures of rhythm discussed above.

A fourth objective is to determine if the language background of the listener has an effect on either of the two measures of rhythm discussed above. This study will use native speakers of English, French, and Japanese as subjects in both types of experiments. Since these three languages are alleged to differ by having, respectively, stress-timing, syllable-timing, and mora-timing, in analyzing the results, an attempt will be made to discover any correlation that may exist between the type of timing (and the particular language) and either of the two measures of rhythm discussed above.
The subjects in this study will, in most cases, have some background in linguistics and hence will not be considered linguistically naive. This may make them more aware of durational patterns in speech than non-naive subjects would be, but should help rather than prejudice the study. In a paper on the rhythmic perception of seven languages, Miller (1984), using both linguistically naive and non-naive subjects, found that the non-naive subjects demonstrated greater discrimination ability than the naive subjects did. This supports the decision to utilize linguistically non-naive subjects as it is desirable to obtain the best possible discrimination ability, thereby, the difference limen for irregularity.
4.1 Pilot Study

In order to determine the stimulus structure and the parameter values of the stimuli for the main study, a pilot study was carried out. The stimuli used in Pilot Test I were sequences of clicks, while those used in Pilot Test II were sequences of CV syllables. The two syllable types used in Pilot Test II were [ta] and [na]; however, only one syllable type was heard during any particular sitting.

Three test tapes were produced in order to estimate the values of timing distortion necessary for subjects to detect that the rhythm of a sound sequence was not regular. It was decided that the values with the greater amount of distortion should produce close to (but slightly less than) 100% detection so that the intermediate steps would be small enough to estimate the DL for irregularity. The goal, therefore, was to produce stimuli with extreme accelerated and decelerated parameter values which could correspond to anchor points between which the DL's for rhythmic irregularity would lie.

The pilot tapes were also used to verify that the quality of recording was adequate and to make any necessary improvements.

4.2 Pilot Test I

4.21 Stimulus Preparation

This test used only click stimuli, each stimulus consisting of six 1-ms clicks and each click being separated

\[ \text{In the remainder of this paper, the term 'stimulus' will be used interchangeably with 'stimulus type', whereas the term 'test item' will be used interchangeably with 'stimulus token'.} \]
from its neighbour by a specific silent interval. Each click consisted of 10 samples (1 ms) or one cycle of a 1-kHz sinewave; this cycle always had the same (zero) starting phase. The length of the intervals depended on the timing of the series, which was either speeded up, regular, or slowed down: ten different temporal sequencing patterns were used. Table I shows the number of samples for each stimulus and illustrates the timing alterations. Stimulus 5 was acoustically regular in timing: each interval consisted of 1270 samples (127 ms). Stimuli 1 to 4 and 6 to 9 also had the first four intervals equal, while the last interval was either shorter or longer than the previous intervals. Stimulus 10 had each interval (after the first) longer than the previous interval by a constant amount of 256 samples (25.6 ms).

A distinction should be made here between three types of irregularity that can occur in a sequence of sounds: (1) A single departure from regularity, where just one interval is altered, as is the case for stimuli 1 to 4 and 6 to 9 of Pilot Test I. (2) A progressive irregularity, increasing or decreasing linearly in duration, as is the case for stimulus 10 in Pilot Test I. (3) A progressive irregularity, increasing or decreasing according to some nonlinear function, as will be the case in Pilot Test II and in the main study.

The stimuli were prepared on a PDP-12 digital computer, using a set of computer programmes written by Lloyd Rice at the U.C.L.A. Phonetics Laboratory. A 1-kHz tone from a sine wave generator was sampled into the computer at a rate of 10 kHz.

Fifty test items (click sequences) were recorded in a pseudo-random order ensuring that there were five repetitions of each stimulus type, with
<table>
<thead>
<tr>
<th>Stimulus No.</th>
<th>C1</th>
<th>I1</th>
<th>C2</th>
<th>I2</th>
<th>C3</th>
<th>I3</th>
<th>C4</th>
<th>I4</th>
<th>C5</th>
<th>I5</th>
<th>C6</th>
<th>Length of Final Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>246</td>
<td>10</td>
<td>24.6 ms</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>502</td>
<td>10</td>
<td>50.2 ms</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>758</td>
<td>10</td>
<td>75.8 ms</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1014</td>
<td>10</td>
<td>101.4 ms</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>127.0 ms</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1526</td>
<td>10</td>
<td>152.6 ms</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1782</td>
<td>10</td>
<td>178.2 ms</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>2038</td>
<td>10</td>
<td>203.8 ms</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>2294</td>
<td>10</td>
<td>229.4 ms</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>1270</td>
<td>10</td>
<td>1526</td>
<td>10</td>
<td>1782</td>
<td>10</td>
<td>2038</td>
<td>10</td>
<td>2294</td>
<td>10</td>
<td>229.4 ms</td>
</tr>
</tbody>
</table>

\[ C_n = \text{Number of samples per } n^{\text{th}} \text{ click} \]

\[ I_n = \text{Number of samples per } n^{\text{th}} \text{ interval} \]
no stimulus appearing twice in succession. The test items were first run through the D/A converter, lowpass filtered at the Nyquist frequency (5 kHz) with an RC lowpass filter having a 48 dB/octave skirt slope, and recorded onto the first track of an audio tape, using a two track Revox A77 tape recorder, as shown in Figure 1. In order to allow the first five items on the tape to serve as practice items, these items were repeated at the end of the tape. Only items 6 to 55 were considered in the analysis.

The actual test tape was obtained by dubbing the original tape onto the first track of a second audio tape (with another Revox A77 tape recorder) as shown in Figure 2. Item numbers were recorded onto the second track of the test tape and the two tracks were electronically mixed for the actual testing.

4.22 Subjects Nine adults (7 females and 2 males), none of whom were linguistically naive, were used as subjects for Pilot Test I. Eight were native speakers of English and one was a native speaker of French. None of the subjects had any known speech or hearing problems.

4.23 Test Procedure Subjects were seated individually in a quiet room and informed that they would hear trains of clicks. They were told they would perhaps notice that the timing was altered in some of the trains: in some of the stimuli, the timing may be speeded up, while in others it may be slowed down. They responded on a five-point scale, as suggested by the columns on the response form: column 1 was for 'speeded up', column 3 was for 'regular', and column 5 was for 'slowed down'. Subjects were further encouraged to use columns 2 and 4, whenever they felt the timing was intermediate.
Figure 1: Block Diagram for Digital-to-Analog Conversion
Figure 2: Block Diagram for Dubbing of Original Tape and Recording of Item Numbers on Test Tapes
The first 10 items were presented to the subject, the tape stopped, and any questions the subject had were answered. The tape was then restarted at item 1. The test was presented through high-fidelity headphones at a comfortable listening level (approximately 70 dB SPL).

4.24 Results To analyze the data, each response was coded with a number from 1 to 5, as suggested above. With the aid of a computer programme, a 'response matrix' was computed for each of the subjects. Table IV in Section 4.34 shows a typical example of one of these matrices. In the table, the stimulus types are listed in a column; the response data, after sorting, are listed in a 7 x 5 matrix; and the mean and its standard deviation (s.d.), for each stimulus type, are listed in the two rightmost columns.

Table V in Section 4.34 shows a typical example of the averaged response for all subjects in a particular test. These data were obtained by averaging all of the subjects' means and s.d.'s for each stimulus type. These data are displayed in Table V, by stimulus type.

Figure 3 shows the mean and two s.d.'s, averaged for nine subjects, for each stimulus type of Pilot Test I. This figure gives an indication of the variability in response for each stimulus type.

The results indicate that stimulus 5, which had equal inter-click intervals (cf. Table I), was perceived as having regular timing by all but one subject, who twice reported it as speeded up. However, stimulus 6, which was slowed by 25.6 ms was perceived as regular by all but one subject. In addition, responses to stimulus 6 had a smaller mean (over nine subjects) standard deviation than did those to neutral stimulus 5, as can be seen in Figure 3.
Figure 3: Mean Score and 2-s.d. Range for Each Stimulus of Pilot Test I, Averaged over Nine Subjects
This figure also shows that the variability is far greater for the decelerated stimuli (6 to 9) than for accelerated stimuli (1 to 4). The reduced variability of the accelerated stimuli indicates that less of an irregularity in the sequence of clicks (and therefore smaller changes in timing structure) may be required for the shortened stimuli to produce small enough changes to best estimate the DL for accelerated irregularity. These changes in the amount of irregularity would be necessary as most of the corresponding timing structures used in Pilot Test I were too easily detected and therefore may be too extreme to be close to the DL.

The large variability of the decelerated stimuli indicates that more of an irregularity in the sequence of clicks (and therefore larger changes in timing structure) may be required for the lengthened stimuli to produce large enough changes to best estimate the DL for decelerated irregularity. These changes in the amount of irregularity would be necessary as most of the corresponding timing structures used in Pilot Test I were too difficult to detect and therefore may not include the DL.

Stimulus 10, which underwent progressive lengthening, was easily detected as irregular by most subjects. Therefore, less of an irregularity in the sequence of clicks would be necessary to estimate the DL for irregularity in a linearly progressive lengthening situation.

4.3 Pilot Test II

4.3.1 Stimulus Preparation The stimuli for this test consisted of manipulated natural speech, each stimulus consisting of a six-syllable sequence where the syllables were phonetically identical, of same amplitude, and monotone, but where, as in Pilot Test I, the durations were varied. In
Pilot Test IIa, the syllable used was [ta]; in Pilot Test IIb, the syllable was [na]. As alluded to above and explained below, the stimuli used in Pilot Test IIa and IIb differed from those used in Pilot Test I in that they all underwent nonlinear progressive timing alteration, instead of being identical in all but the last interval (stimuli 1 to 9).

A trained phonetician (native speaker of French), seated in a sound-treated booth, produced the desired syllables which were recorded onto a Revox A77 tape recorder. In order to help the speaker to produce these syllables on a monotone, he heard a 100-Hz pure tone while producing the syllables, as shown in the schematic diagram in Figure 4. He also endeavoured to produce them in such a way that both of the segments (C and V) would be long enough for the intended purpose. The syllables were then sampled at a frequency of 5 kHz, playing the tape at half-speed (9.5 cm/s) through a lowpass filter at the Nyquist frequency (in this case, 2.5 kHz) into a PDP-12 digital computer, using the same computer programmes as before. The instrumental layout is diagrammed in Figure 5.

In order to estimate typical values of timing irregularities encountered in natural speech and likely to be perceived as regular, the production data of Oiler (1973) were recompiled, plotted, and values were obtained for various segment and syllable duration ratios. The ratios utilized are shown below. These values were also compared to the data of Nooteboom (1973) and Klatt (1975; 1976) and found to be fairly similar.

It was decided to use a function which would generate syllable duration values that would most closely match the corresponding median values obtained from Oiler's (1973) data. The following three ratios were considered for the match:
\[ R_1 = \frac{C_5 + V_5}{C_4 + V_4} \]
\[ R_2 = \frac{C_5 + V_5}{C_{123} + V_{123}} \]
\[ R_3 = \frac{C_4 + V_4}{C_{123} + V_{123}} \]

In these formulae, \( C_j \) and \( V_j \) are the durations of the \( j \)th consonant and \( j \)th vowel, respectively.

The function finally arrived at to compute the successive durations of the six repeated syllables was:

\[ D = N_s \cdot \exp(a \cdot x^{n \cdot z}) \]

where
- \( D \) = syllable duration in samples
- \( x \) = a parameter depending on the position of the syllable in the utterance which varies linearly and discretely between 0 (first syllable) and 1 (last syllable)
- \( N_s \) = number of samples in first syllable
- \( a \) = scale factor for the exponent of the exponential
- \( n \) = exponent of \( x \) (when \( n = 2 \), \( D \) has the form of the Gaussian function)
- \( z \) = adjusts the value of the exponential in such a way that, for \( x = 1 \) (i.e. last syllable), it will always have the same value for a given \( a \)-value, regardless of the value of \( n \)

For negative values of \( a \), the value \( D'(a) = 1 - [D(-a) - 1] \) was used instead of \( D(a) \).

According to Gerber (1974), the average syllable rate of conversational speech is slightly above seven syllables per second. Thus, the number of samples, \( N_s \), in the first syllable of each stimulus was obtained by dividing the rate at which the digitized sound sequences were sampled (10 kHz) by seven; hence, \( N_s = 1430 \) samples. As stated previously, the median values obtained from Oller (1973) for the three syllable duration ratios defined above were used to estimate a 'neutral' timing sequence. Minimum and maximum values (that is the extreme accelerated and decelerated
Figure 4: Block Diagram for Recording of CV Syllables
Figure 5: Block Diagram for Analog-to-Digital Conversion
stimulus, respectively) for parameter \( a \) were then computed in an attempt to find the stimulus extremes which are perceived as irregular most of the time, the idea being that the DL for progressive irregularity should be between the perceived neutral timing sequence and the anchor points.

For Pilot Test IIa, using the [ta] syllable, the best fit to the median \( R_1 \), \( R_2 \), and \( R_3 \), estimated from Oller's data, was found for the values \( n = 4 \) and \( a = 0.53 \). These figures were chosen for TA-4, the 'neutral' stimulus (cf. Table II). The two extreme values chosen for parameter \( a \) were \(-0.20\) and \(1.20\); they yielded the ratio values used for stimuli TA-1 and TA-7 and shown in Table II. Two more values of \( a \) were selected between the 'neutral' and each of the two extreme values, corresponding to stimuli TA-2, TA-3, TA-5, and TA-6. The proportion used for the duration of the consonantal part of all of the [ta] stimuli was chosen as 47.5% of the total syllable duration.

The stimuli were prepared, edited, and recorded as described in Section 4.21. However, as Pilot Test IIa contained only seven different stimuli, the test tape contained only forty test items, the first five serving as practice items and items 6 to 40 being those used in the analysis.

In Pilot Test IIb, the [na] stimuli underwent similar preparation to the [ta] stimuli in Pilot Test IIa. However, due to the results of the Pilot Test IIa data analysis (discussed in the following section), only four parameter values were used, mainly in order to determine the values of \( a \) corresponding to stimuli which are perceived as irregular close to 100% of the time. As shown in Table III, the \( a \)-values varied between \(-0.20\) and \(0.88\).

When preparing the [na] stimuli, some difficulty was encountered: in attempting to make the nasal consonant long enough for the syllable
### TABLE II

Durational Structure (in samples) for the [ta] Stimuli of Pilot Test IIa

<table>
<thead>
<tr>
<th>Stimulus No.</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA-1</td>
<td>1430</td>
<td>1430</td>
<td>1425</td>
<td>1404</td>
<td>1346</td>
<td>1217</td>
<td>-0.20</td>
</tr>
<tr>
<td>TA-2</td>
<td>1430</td>
<td>1430</td>
<td>1431</td>
<td>1435</td>
<td>1446</td>
<td>1470</td>
<td>0.04</td>
</tr>
<tr>
<td>TA-3</td>
<td>1430</td>
<td>1430</td>
<td>1437</td>
<td>1466</td>
<td>1548</td>
<td>1736</td>
<td>0.28</td>
</tr>
<tr>
<td>TA-4</td>
<td>1430</td>
<td>1431</td>
<td>1444</td>
<td>1500</td>
<td>1662</td>
<td>2065</td>
<td>0.53</td>
</tr>
<tr>
<td>TA-5</td>
<td>1430</td>
<td>1431</td>
<td>1449</td>
<td>1530</td>
<td>1769</td>
<td>2405</td>
<td>0.75</td>
</tr>
<tr>
<td>TA-6</td>
<td>1430</td>
<td>1432</td>
<td>1455</td>
<td>1562</td>
<td>1889</td>
<td>2820</td>
<td>0.98</td>
</tr>
<tr>
<td>TA-7</td>
<td>1430</td>
<td>1432</td>
<td>1461</td>
<td>1593</td>
<td>2010</td>
<td>3285</td>
<td>1.20</td>
</tr>
</tbody>
</table>

### TABLE III

Durational Structure (in samples) for the [na] Stimuli of Pilot Test IIb

<table>
<thead>
<tr>
<th>Stimulus No.</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA-1</td>
<td>1430</td>
<td>1430</td>
<td>1425</td>
<td>1404</td>
<td>1346</td>
<td>1217</td>
<td>-0.20</td>
</tr>
<tr>
<td>NA-2</td>
<td>1430</td>
<td>1430</td>
<td>1428</td>
<td>1420</td>
<td>1397</td>
<td>1348</td>
<td>-0.08</td>
</tr>
<tr>
<td>NA-3</td>
<td>1430</td>
<td>1431</td>
<td>1449</td>
<td>1531</td>
<td>1774</td>
<td>2421</td>
<td>0.76</td>
</tr>
<tr>
<td>NA-4</td>
<td>1430</td>
<td>1431</td>
<td>1452</td>
<td>1548</td>
<td>1836</td>
<td>2631</td>
<td>0.88</td>
</tr>
</tbody>
</table>

$Sn =$ Number of samples per $n^{th}$ syllable
duration required, the consonant amplitude was too great in comparison to the vowel amplitude. Hence, several new recordings were necessary for this syllable. The most satisfactory way of overcoming this difficulty was to omit, prior to the preparation of the test stimuli, the two periods of the CV-transition which were immediately before the first period of the vowel. The syllable with shortened transition produced a more natural sounding stimulus than the syllable with full transition. The proportion finally selected for the consonant segment of the [na] stimuli was 40% of the total syllable duration.

Five repetitions of each of these stimuli were recorded in a similar manner to that in Pilot Test IIa with an additional five items acting as practice at the beginning of the tape and items 6 to 25 being those used for the analysis.

For both Pilot Test IIa and Pilot Test IIb, the actual test tape was obtained by dubbing the original tape onto the first track of a second audio tape (with another Revox A77 tape recorder) as shown in Figure 2. Item numbers were recorded onto the second track of the test tape and the two tracks were electronically mixed.

4.32 Subjects Three adults (two females and one male) served as subjects for Pilot Test IIa. None of them were linguistically naive. The two female subjects were native speakers of English and the male subject was a native speaker of French. He was also the speaker for the speech material. For Pilot Test IIb, only the male subject and one of the female subjects participated. The subjects had no known speech or hearing problems.
4.33 **Test Procedure**  The subjects were seated individually in a quiet room. They were informed that they would hear stimuli consisting of sequences of six syllables and that they might notice that the timing was altered (either speeded up or slowed down) in some of the sequences. They were asked to put a check in the appropriate column of the seven-point, forced-choice response form, after listening to each test item. A seven-point scale was used in order to give the possibility of a more graded response; it also allowed the experimenter to potentially split the responses for the intermediate columns more conveniently, for example with a grouping 1-2/3-4-5/6-7: column 1, 4, and 7 were labeled 'speeded up', 'regular', and 'slowed down', respectively. Regarding the intermediate columns, the same instructions were given as in Pilot Test I.

The first seven items were presented to the subject, the tape stopped, and any questions the subject had were answered. The tape was then restarted at item 1. The test was presented through high-fidelity headphones at a comfortable listening level (approximately 70 dB SPL).

4.34 **Results**  In order to analyze the results of Pilot Test IIa (and of IIb), each response was coded with a number from one to seven, according to the column used for the response.

The responses and the computed means and s.d.'s were arranged in 7 x 7 matrix form, as described earlier (Section 4.24). Typical results are shown in Tables IV and V.

By examining Table V, it can be noted that stimulus TA-7, the most slowed down [ta] stimulus, was perceived as irregular by all subjects in all presentations. Therefore, it was decided that this stimulus type (with the
a-parameter value of 1.20) would not be used in the main study since it resulted in 100% detection of irregularity. In as much as the s.d. of the averaged (for all subjects) mean (cf. Table V) was small for stimuli TA-1 (one s.d. = 0.30) and TA-6 (one s.d. = 0.78), the parameter values for these two stimuli (a = -0.20 and a = 0.98) were chosen as the two anchor points in the main study. The results of Pilot Test IIb confirmed the choice of these parameter values for the two anchor points.

As can be observed in Table V, stimulus TA-3 had an averaged (for all subjects) mean of 4.00 and a corresponding s.d. of 0.24: this stimulus appeared to be perceived as regular most of the time in the pilot study. Hence, a stimulus with the same parameter values as TA-3 (a = 0.28 and n = 4, corresponding to values of $R_1$, $R_2$, and $R_3$ of 1.121, 1.214, and 1.082, respectively) was chosen as the 'neutral' stimulus in the main study.

For each subject, the s.d. values were plotted on a linear ordinate scale as a function of their corresponding stimulus number, itself plotted on a linear abscissa scale, as exemplified in Figure 6. When examining this figure, a 'camel-back' shape may be seen to emerge. In order to obtain this 'camel-back' shape, there should be minimum variability for three stimuli (the most accelerating, the neutral, and the most decelerating) and larger variability for the stimuli between these three. It is felt that the values of a corresponding to the DL for irregularity should fall between the three minimum variability points, that is, in the areas of maximum variability.

4.4 **Main Study**

The stimulus structure and the parameter values of the stimuli to be used for the main study were chosen on the basis of the results of the pilot
### TABLE IV

**Typical Response Matrix for Subject 2 in Pilot Test IIa**

<table>
<thead>
<tr>
<th>Subject 2</th>
<th>Stimulus</th>
<th>Responses</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA-1</td>
<td>1 1 1 2 1</td>
<td>1.20</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>TA-2</td>
<td>2 3 2 3 2</td>
<td>2.40</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>TA-3</td>
<td>3 4 4 5 4</td>
<td>4.00</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>TA-4</td>
<td>5 6 5 6 5</td>
<td>5.40</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>TA-5</td>
<td>6 6 6 6 6</td>
<td>6.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>TA-6</td>
<td>6 6 7 6 7</td>
<td>6.40</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>TA-7</td>
<td>7 7 7 7 7</td>
<td>7.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE V

**Typical Averaged Response for All Subjects of Pilot Test IIa**

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA-1</td>
<td>1.13</td>
<td>0.30</td>
</tr>
<tr>
<td>TA-2</td>
<td>2.60</td>
<td>0.91</td>
</tr>
<tr>
<td>TA-3</td>
<td>4.00</td>
<td>0.24</td>
</tr>
<tr>
<td>TA-4</td>
<td>5.13</td>
<td>0.59</td>
</tr>
<tr>
<td>TA-5</td>
<td>6.27</td>
<td>0.48</td>
</tr>
<tr>
<td>TA-6</td>
<td>6.53</td>
<td>0.78</td>
</tr>
<tr>
<td>TA-7</td>
<td>7.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Figure 6. Standard Deviations for Subject 3 in Pilot Test IIa.
study (cf. Sections 4.24 and 4.34). Editing and recording of the stimuli was performed as described in detail in Section 4.2.

4.41 Stimulus Preparation

The timing for all of the stimuli in the main study were computed with the same function (D) used for the speech stimuli in the pilot study. In addition, the intervals between the clicks or between the syllable beats were the same for each set of stimuli. Since the click was placed where the syllable beat was thought to occur, all click stimuli were shorter than their corresponding speech stimuli by approximately one syllable (more precisely 45% of the first syllable, plus 55% of the last one).

Based on the results of the pilot study, the values of \( a \) to be used for the most accelerating, the neutral, and the most decelerating stimuli were selected as -0.20, 0.28 and 0.98, respectively. These three values were then plotted as a function of their corresponding stimulus number, itself plotted on a linear abscissa scale, as displayed in Figure 7. The values of \( a \) for the intermediate points were obtained by quadratic interpolation between these three points. This produced the values for the three sets of tests to be used in the main study, namely \( a = -0.20, -0.07, 0.10, 0.28, 0.50, 0.72, \) and 0.98. It is expected that the DL values (measured by \( a \)) for the accelerated and decelerated irregularities will be between the neutral stimulus and each extreme. As in the pilot study, the first syllable of each stimulus had 1430 samples, based on a syllabic rate of 7 per second.

Three sets of stimuli, each using seven stimulus-types, were prepared: the set for Test I consisted of click stimuli, whereas the sets for Test II and Test III consisted of [ta] and [na] speech stimuli, respectively. It should be recalled that all stimuli in the main study
Figure 7. Curve Produced by Plotting $a$-values as a Function of Their Corresponding Stimulus Numbers.
underwent nonlinear progressive timing alteration.

Each stimulus in Test I consisted of a sequence of six clicks, separated by a specific interval calculated with the same function D (cf. Section 4.2) and the clicks were identical to those in the pilot study. The same computer programmes that computed the syllable durations for Pilot Test II computed the interval durations for Test I. The place of vowel onset was computed (i.e. consonant proportion = 45%, vowel proportion = 55%) and the click was placed at this location. The durational structure for the click stimuli was as shown in Table VI.

The stimuli for Test II and Test III were prepared in the same manner, and from the same [ta] and [na] syllables, as those used in Pilot Test II. For a detailed description, see Section 4.3. The durational structure for the speech stimuli was as shown in Table VII.

The stimuli were then prepared, edited, and recorded as described in Section 4.2. As each of the three tests in the main study contained five repetitions of each of the seven stimulus-types, plus seven practice items at the beginning and three buffer items at the end of each test, each of the tests contained 45 items, with items 8 to 42 being used for the analysis. The pseudo-random order imposed on the test items of the main study included the constraint that no stimulus-type could be repeated until at least three other stimulus-types had intervened.

A thirty second, 1-kHz calibration tone was recorded on the actual test tape in order to assure presentation of the stimuli at a repeatable level (approximately 70 dB SPL). Then Tests I, II, and III were dubbed from the original tape onto the first track of a second audio tape (see Section 4.2 for details).
### TABLE VI

**Durational Structure (in samples) for the Click Stimuli of Main Test I**

<table>
<thead>
<tr>
<th>Stimulus No.</th>
<th>C1</th>
<th>I1</th>
<th>C2</th>
<th>I2</th>
<th>C3</th>
<th>I3</th>
<th>C4</th>
<th>I4</th>
<th>C5</th>
<th>I5</th>
<th>C6</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLICK1</td>
<td>10</td>
<td>1430</td>
<td>10</td>
<td>1428</td>
<td>10</td>
<td>1416</td>
<td>10</td>
<td>1378</td>
<td>10</td>
<td>1288</td>
<td>10</td>
<td>-0.20</td>
</tr>
<tr>
<td>CLICK2</td>
<td>10</td>
<td>1430</td>
<td>10</td>
<td>1429</td>
<td>10</td>
<td>1425</td>
<td>10</td>
<td>1412</td>
<td>10</td>
<td>1382</td>
<td>10</td>
<td>-0.07</td>
</tr>
<tr>
<td>CLICK3</td>
<td>10</td>
<td>1430</td>
<td>10</td>
<td>1431</td>
<td>10</td>
<td>1437</td>
<td>10</td>
<td>1456</td>
<td>10</td>
<td>1499</td>
<td>10</td>
<td>0.10</td>
</tr>
<tr>
<td>CLICK4</td>
<td>10</td>
<td>1430</td>
<td>10</td>
<td>1433</td>
<td>10</td>
<td>1450</td>
<td>10</td>
<td>1503</td>
<td>10</td>
<td>1633</td>
<td>10</td>
<td>0.28</td>
</tr>
<tr>
<td>CLICK5</td>
<td>10</td>
<td>1430</td>
<td>10</td>
<td>1436</td>
<td>10</td>
<td>1467</td>
<td>10</td>
<td>1564</td>
<td>10</td>
<td>1816</td>
<td>10</td>
<td>0.50</td>
</tr>
<tr>
<td>CLICK6</td>
<td>10</td>
<td>1431</td>
<td>10</td>
<td>1439</td>
<td>10</td>
<td>1483</td>
<td>10</td>
<td>1628</td>
<td>10</td>
<td>2025</td>
<td>10</td>
<td>0.72</td>
</tr>
<tr>
<td>CLICK7</td>
<td>10</td>
<td>1431</td>
<td>10</td>
<td>1442</td>
<td>10</td>
<td>1503</td>
<td>10</td>
<td>1709</td>
<td>10</td>
<td>2308</td>
<td>10</td>
<td>0.98</td>
</tr>
</tbody>
</table>

### TABLE VII

**Durational Structure (in samples) for the Speech Stimuli of Main Test II and III**

<table>
<thead>
<tr>
<th>Stimulus No.</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA1, NA1</td>
<td>1430</td>
<td>1430</td>
<td>1425</td>
<td>1404</td>
<td>1346</td>
<td>1217</td>
<td>-0.20</td>
</tr>
<tr>
<td>TA2, NA2</td>
<td>1430</td>
<td>1430</td>
<td>1428</td>
<td>1421</td>
<td>1401</td>
<td>1359</td>
<td>-0.07</td>
</tr>
<tr>
<td>TA3, NA3</td>
<td>1430</td>
<td>1430</td>
<td>1433</td>
<td>1443</td>
<td>1471</td>
<td>1533</td>
<td>0.10</td>
</tr>
<tr>
<td>TA4, NA4</td>
<td>1430</td>
<td>1430</td>
<td>1437</td>
<td>1466</td>
<td>1548</td>
<td>1736</td>
<td>0.28</td>
</tr>
<tr>
<td>TA5, NA5</td>
<td>1430</td>
<td>1431</td>
<td>1443</td>
<td>1496</td>
<td>1648</td>
<td>2022</td>
<td>0.50</td>
</tr>
<tr>
<td>TA6, NA6</td>
<td>1430</td>
<td>1431</td>
<td>1448</td>
<td>1526</td>
<td>1754</td>
<td>2355</td>
<td>0.72</td>
</tr>
<tr>
<td>TA7, NA7</td>
<td>1430</td>
<td>1432</td>
<td>1455</td>
<td>1562</td>
<td>1889</td>
<td>2820</td>
<td>0.98</td>
</tr>
</tbody>
</table>
4.42 **Subjects**  
Twenty-four adults, none of whom were linguistically naive, were used as subjects for the main study. These subjects were divided into three groups according to their language background as indicated below. Group I consisted of twelve subjects (6 females and 6 males) who were native speakers of English and whose age varied between 23 and 43 years. Group II consisted of six subjects (3 females and 3 males) who were native speakers of French and whose age varied between 25 and 50 years. Group III consisted of six subjects (3 females and 3 males) who were native speakers of Japanese and whose age varied between 24 and 37 years. All of the subjects had at least a working knowledge of English. None of the subjects had any known speech or hearing problems: any of the subjects who had not undergone audiometric testing in the recent past had their hearing screened prior to the commencement of the experiment. All of the subjects who underwent audiometric screening had hearing within normal limits in the frequencies considered most important for speech.

4.43 **Test Procedure**  
Subjects were seated individually in a quiet room which met the standards required for a hearing screening situation. They were informed that they would hear sequences of nonspeech and speech sounds: in one test (Test I), they would hear stimuli consisting of sequences of clicks; in another test (Test II), they would hear stimuli consisting of sequences of [ta] syllables; and in another test (Test III), they would hear stimuli consisting of sequences of [na] syllables.

They were further informed that they might notice that the timing was altered (either speeding up or slowing down) in some of the sequences. They were asked to put a check in the appropriate column of the seven-point,
forced-choice response form, after listening to each test item. A seven-point scale was used for similar reasons as given in Section 4.33. Column 1, 4, and 7 were labeled 'speeded up', 'regular', and 'slowed down', respectively. Regarding the intermediate columns, the same instructions were given as in Pilot Test I.

The test tape was played on a Revox A77 tape recorder and presented over TDH-39 headphones at a level of 60-70 dB SPL as measured with the calibration tone on a Bruel and Kjaer 2203 precision sound level meter with a Bruel and Kjaer 6 cc 4152 artificial ear.

Each of the subjects took the tests in a particular order as explained below. The order of test presentation varied within each group: the three main test orders were: I, II, III; II, III, I; and III, I, II. One female and one male in each group were arbitrarily assigned to take the tests in one of each of these three orders. Since there were twice as many subjects in Group I as in Groups II and III, one female and one male from Group I were arbitrarily assigned to take the tests in one of each of the following three orders: I, III, II; II, I, III; and III, II, I.

For each of the tests, the first seven items were presented to the subject, the tape stopped, and any questions the subject had were answered. The tape was then restarted at item 1.

The subject was allowed to rest for 1 to 3 minutes between tests while the experimenter set up the equipment for the next test.
CHAPTER 5
RESULTS

5.1 Data Storage and Sorting

In order to analyze the results of the perception tests, each listener response was assigned a number from 1 to 7, corresponding to the number of the column in which the listener had registered his answer. For instance, an answer of 'speeded up' was assigned a 1; an answer of 'regular' was assigned a 4; and an answer of 'slowed down' was assigned a 7. Listener responses were then stored on computer tape so that the figures would be available for data manipulation. Responses to items 8 to 42 were sorted in matrix form, one matrix per test and per subject. This sorting produced three matrices per subject, each matrix containing 35 items. Each 7 x 5 response matrix was incorporated into a table which included three additional columns: one for the stimulus number, one for the response mean, and one for the corresponding standard deviation (s.d.). Table VIII presents an example of one of these tables for Test II and Subject E2. The response matrices and corresponding tables are described in more detail in Section 4.24.

In addition, the responses were cumulated by stimulus type and the corresponding means and s.d.'s were computed and displayed in table form as exemplified in Table IX (all English subjects for Test II) and Table X (all tests for Subject E2).

Figures 8, 9, and 10 show for Test I, II, and III, respectively, the response mean and two-s.d. range for each stimulus type, averaged for all subjects. The two-s.d. ranges indicate the variability in response for each stimulus type in each test. It can be seen see that the variability,
### TABLE VIII

**Typical Response Matrix for Subject E2 in Main Test II**

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Responses</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA-1</td>
<td>1 1 1 1 1</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TA-2</td>
<td>1 1 1 1 4</td>
<td>1.60</td>
<td>1.34</td>
</tr>
<tr>
<td>TA-3</td>
<td>4 4 4 4 4</td>
<td>4.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TA-4</td>
<td>4 4 4 4 5</td>
<td>4.20</td>
<td>0.45</td>
</tr>
<tr>
<td>TA-5</td>
<td>5 4 6 6 6</td>
<td>5.40</td>
<td>0.89</td>
</tr>
<tr>
<td>TA-6</td>
<td>7 7 6 6 6</td>
<td>6.40</td>
<td>0.55</td>
</tr>
<tr>
<td>TA-7</td>
<td>7 7 7 7 7</td>
<td>7.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### TABLE IX

**Typical Averaged Response for All English Subjects of Main Test II**

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA1</td>
<td>1.67</td>
<td>0.80</td>
</tr>
<tr>
<td>TA2</td>
<td>2.33</td>
<td>0.90</td>
</tr>
<tr>
<td>TA3</td>
<td>3.63</td>
<td>0.74</td>
</tr>
<tr>
<td>TA4</td>
<td>4.07</td>
<td>0.76</td>
</tr>
<tr>
<td>TA5</td>
<td>5.37</td>
<td>0.88</td>
</tr>
<tr>
<td>TA6</td>
<td>5.90</td>
<td>0.80</td>
</tr>
<tr>
<td>TA7</td>
<td>6.68</td>
<td>0.47</td>
</tr>
</tbody>
</table>
**TABLE X**

Typical Averaged Response of All Tests for Subject E2

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLICK1, TA1, NA1</td>
<td>1.07</td>
<td>0.26</td>
</tr>
<tr>
<td>CLICK2, TA2, NA2</td>
<td>1.73</td>
<td>1.10</td>
</tr>
<tr>
<td>CLICK3, TA3, NA3</td>
<td>3.73</td>
<td>0.70</td>
</tr>
<tr>
<td>CLICK4, TA4, NA4</td>
<td>4.07</td>
<td>0.26</td>
</tr>
<tr>
<td>CLICK5, TA5, NA5</td>
<td>5.33</td>
<td>0.72</td>
</tr>
<tr>
<td>CLICK6, TA6, NA6</td>
<td>6.33</td>
<td>0.49</td>
</tr>
<tr>
<td>CLICK7, TA7, NA7</td>
<td>7.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Figure 8. Mean Score and 2-s.d. Range for Each Stimulus of Main Test I, Averaged Over All Subjects.
Figure 9. Mean Score and 2-s.d. Range for Each Stimulus of Main Test II, Averaged over All Subjects.
Figure 10. Mean Score and 2-s.d. Range for Each Stimulus of Main Test III, Averaged over All Subjects.
does not show any areas of minimum variability. Absence of areas of minimum variability for the whole group may be due to the fact that the large intersubject variability obscures the location of these areas when the entire subject population is considered. That is, the location of the areas of minimum variability may differ from subject to subject; hence, when the s.d.'s are averaged over all subjects, the areas of minimum and of maximum variability may get smoothed out.

5.2 Pseudo-normalization

Pseudo-normalization of the data was attempted since some subjects consistently did not utilize the extremes of the seven-point scale, while others consistently did not utilize the intermediate steps between 'regular' and the two extremes. All the data was transformed by mapping the responses of the three response groups 1-2, 3-4-5, and 6-7 onto the pseudo-normalized responses of 1, 4, and 7, respectively.

Table XI shows the transformed responses, the corresponding response mean, and s.d. for the Test II responses of Subject E2. The s.d.'s for the transformed scores of the constructed stimuli could be smaller or larger than the s.d.'s for the non-transformed scores since the transformation used may reduce or increase the variability depending on whether the non-transformed scores varied within one of the three response groups or between the response groups. Such manipulation of the data did not, in the end, prove to be useful.

5.3 Perception of Regularity

In order to estimate typical values of timing irregularities likely
<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Responses</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA-1</td>
<td>1 1 1 1 1</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TA-2</td>
<td>1 1 1 1 4</td>
<td>1.60</td>
<td>1.34</td>
</tr>
<tr>
<td>TA-3</td>
<td>4 4 4 4 4</td>
<td>4.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TA-4</td>
<td>4 4 4 4 4</td>
<td>4.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TA-5</td>
<td>4 4 7 7 7</td>
<td>5.80</td>
<td>1.64</td>
</tr>
<tr>
<td>TA-6</td>
<td>7 7 7 7 7</td>
<td>7.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TA-7</td>
<td>7 7 7 7 7</td>
<td>7.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
to be perceived consistently as regular by a particular subject, the perception data of each subject were plotted separately for each test with the stimulus number plotted linearly on the abscissa and the response mean plotted linearly on the ordinate. A value for the stimulus number which would correspond to a response of four was then obtained by fourth-degree polynomial interpolation on these particular data points. Table XII displays stimulus number values thus obtained for each subject and each test.

A number of factors influenced the choice of the degree of the polynomial used for interpolation purposes. The largest degree that could be justified was the sixth, since there were only seven data points and since there could have been, at most, four inflection points; and the smallest degree that could be selected was the second, since a straight-line fit was not desirable. In a low degree fit, a large error on one response mean could be detrimental even to remote points which themselves may not be in error. In a high degree fit, a close fit may be illusory. The number of inflection points on a curve allowed by a polynomial is equal to the degree of the polynomial, minus two. Both fourth- and sixth-degree polynomial fits were calculated. The results obtained led to similar conclusions and only those of the fourth-degree fit are reported here. The figures obtained from the fourth-degree interpolation were used in the analysis of variance (ANOVA) performed for this study and described later in this section.

The values of $a$ for the stimulus that would correspond to a response of four were then determined by quadratic interpolation, using the same curve used to obtain the stimulus parameter values in the main study. This curve is displayed in Figure 7 and described in Section 4.41.

A three-way analysis of variance (ANOVA) was performed, with native
<table>
<thead>
<tr>
<th>Subject</th>
<th>Test No.</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group I</td>
<td>Group II</td>
<td>Group III</td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td>3.44</td>
<td>3.73</td>
<td>3.42</td>
<td></td>
</tr>
<tr>
<td>E2</td>
<td>3.53</td>
<td>3.56</td>
<td>4.68</td>
<td></td>
</tr>
<tr>
<td>E3</td>
<td>4.02</td>
<td>3.91</td>
<td>3.52</td>
<td></td>
</tr>
<tr>
<td>E4</td>
<td>3.89</td>
<td>4.15</td>
<td>4.52</td>
<td></td>
</tr>
<tr>
<td>E5</td>
<td>4.52</td>
<td>4.69</td>
<td>5.84</td>
<td></td>
</tr>
<tr>
<td>E6</td>
<td>6.04</td>
<td>4.15</td>
<td>3.42</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E7</td>
<td>3.67</td>
<td>4.36</td>
<td>3.90</td>
</tr>
<tr>
<td>E8</td>
<td>3.59</td>
<td>3.90</td>
<td>3.33</td>
</tr>
<tr>
<td>E9</td>
<td>4.15</td>
<td>4.51</td>
<td>5.41</td>
</tr>
<tr>
<td>E10</td>
<td>3.84</td>
<td>6.04</td>
<td>3.27</td>
</tr>
</tbody>
</table>

*TABLE XII*

*Stimulus Number Corresponding to Perception of Regularity*
language of subject (L), test order (O), and test type (T) as the factors (test order was not completely balanced). In this design, sex of subject was nested with language and order, and was crossed with type: a repeated measures design in that all subjects respond to all three test types. Table XIII displays a summary of an ANOVA performed on a balanced subset of 18 subjects. It was decided not to utilize the unbalanced design as it produced results which may have been an artifact of the design.

The data displayed in Table XII were used as input for this ANOVA. They were the values estimated for each test as corresponding to each subject's perception of regularity.

This ANOVA indicates that none of the results reached statistical significance. Specifically, the language background of a subject did not have an effect on the results, although this may have been expected to be the case. However, the absence of this effect may have been due to the large intra- and intersubject variabilities; the degree of irregularity required for a perception of regularity clearly differs from subject to subject. Similarly, neither the nature of the stimulus nor the test order had an effect on the results. This effect may have been due to the use of nonsense speech stimuli rather than meaningful speech, although the subjects may have been listening in the 'nonspeech mode' rather than in the 'speech mode' when listening to the speech stimuli.

5.4 Subject Evaluation Criteria

Each subject's responses were tabulated as in Tables VIII, X, and XI and examined individually. Since some of the subjects appeared to have performed better and more consistently than others, it was decided to try to
### TABLE XIII

**Results of ANOVA Performed on 18-Subject Group**

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>DF</th>
<th>Mean Square</th>
<th>F-ratio</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>2</td>
<td>1.0297</td>
<td>2.2641</td>
<td>n.s.</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0.26109</td>
<td>0.57407</td>
<td>n.s.</td>
</tr>
<tr>
<td>L x 0</td>
<td>4</td>
<td>0.76431</td>
<td>1.6805</td>
<td>n.s.</td>
</tr>
<tr>
<td>S(L,0)</td>
<td>9</td>
<td>0.45480</td>
<td>1.7084</td>
<td>n.s.</td>
</tr>
<tr>
<td>T</td>
<td>2</td>
<td>0.042572</td>
<td>0.15992</td>
<td>n.s.</td>
</tr>
<tr>
<td>L x T</td>
<td>4</td>
<td>0.12544</td>
<td>0.47121</td>
<td>n.s.</td>
</tr>
<tr>
<td>O x T</td>
<td>4</td>
<td>0.15484</td>
<td>0.58165</td>
<td>n.s.</td>
</tr>
<tr>
<td>L x O x T</td>
<td>8</td>
<td>0.30472</td>
<td>1.1446</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>18</td>
<td>0.26622</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

L = language group  
O = test order  
S = sex of subject  
T = test type
determine criteria which would select the best subjects and to compare their results with the results of the entire subject population. This decision seemed particularly appropriate in the context of the present experiments due to the fact that, in order to obtain the DL for irregularity, the intention was to determine the best performance possible.

Utilizing the responses to all five repetitions of each stimulus as well as the corresponding pseudo-normalized responses, the following criteria based on the ability to discriminate between accelerating, neutral, and decelerating sound sequences, were established.

Criterion 1: For each of the three tests of a particular subject, the response means contain no reversals (i.e. as the stimulus number increases, so does the response mean) with the response mean to stimulus \( i \) never being less than the response mean to stimulus \( i-1 \), or more than the response mean to stimulus \( i+1 \).

Criterion 2: For each test, the s.d.'s of the response means were first averaged over all stimulus-types in that test, then rank ordered using all 24 subjects. Any subject must have all three of these s.d.'s in the bottom half of the distribution.

Various other criteria were considered, but they were eventually discarded as unsuitable. Four subjects passed both Criteria 1 and 2, namely subjects E1, E2, E4, and E5. These four subjects were then grouped together and their responses as a group were compared with the responses of the entire subject population as a group. The responses were 'averaged' (in the remainder of this paper, when 'averaged' is enclosed in single quotation marks, it means averaged over all subjects in a particular group) in the 4-subject group and in the 24-subject group separately, and the corresponding
s.d.'s were computed. The results of these computations are displayed in Figures 11 to 16.

Figures 11 to 13 show for Test I, II, and III, respectively, the 'averaged' response means for both the 4-subject and for the 24-subject group. Similarly, Figures 14 to 16 show for Test I, II, and III, respectively, the 'averaged' s.d.'s for both the 4-subject and for the 24-subject group. Each graph utilizes a dashed line to represent the 4-subject group and a solid line to represent the 24-subject group.

In Figures 11 to 13, the curves representing the 'averaged' responses for the 4-subject group and for the 24-subject group appear to have basically the same shape. The main difference is that the curve for the 24-subject group covers a slightly smaller range of response means as compared to the 4-subject group.

In Figures 14 to 16, the curves representing the 'averaged' s.d.'s for the 4-subject group and for the 24-subject group do differ in their basic shape. As would be expected, the 24-subject group shows greater variability than the 4-subject group. However, the main difference is that the 24-subject group tends not to exhibit a camel-back shape, whereas the 4-subject group does exhibit this shape. As already explained in Chapter 4, it is felt that the DL for irregularity should be in the areas of maximum variability.

5.5 Estimation of DL for Irregularity

In order to estimate typical values of timing irregularities likely to be perceived 50% of the time as accelerating or decelerating, two curves were plotted for each one of the better subjects.
Figure 11. Typical 'Averaged' Response for Test I.
Figure 12. Typical 'Averaged' Response for Test II.
Figure 13. Typical 'Averaged' Response for Test III.
Figure 14. Typical 'Averaged' s.d.'s for Test I.
Figure 15. Typical 'Averaged' s.d.'s for Test II.
Figure 16. Typical 'Averaged' s.d.'s for Test III.
The first curve was obtained by plotting the response means for each subject, as a function of the stimulus number, for each test on a separate graph. Both ordinates and abscissae were plotted linearly. Stimulus-number values were then obtained for response-mean values of 2.5, 4, and 5.5, using the same fourth-degree interpolation function as previously. The stimulus-number value corresponding to a response of 4 was chosen as corresponding to a stimulus a subject would perceive as regular, as explained in Section 5.3. The hypothesis was presented earlier that the responses 1, 4, and 7 constitute, by design, anchor points. The stimulus-number values corresponding to a response of 2.5 and 5.5 were chosen as the stimuli thought to be in the area of maximum uncertainty, i.e. the stimuli with the smallest degree of irregularity a subject would perceive 50% of the time as accelerating and decelerating, respectively.

The second curve was obtained by plotting the s.d.'s corresponding to the above response means for each subject and for each test, as a function of the stimulus number using linear abscissa and ordinate scales. The stimulus-number value corresponding to a given s.d. in the area of minimum variability (that is, in the trough of some fitted curve) was then obtained by visual inspection. This value was assumed to correspond to the stimulus the subject would perceive as most regular. Two stimulus-number values corresponding to ordinates in the two areas of maximum variability (that is, at the peaks of some fitted curve) were then obtained by visual inspection. The lower stimulus-number value was taken to represent the stimulus corresponding to the DL for accelerated rhythm, the upper one corresponding to the DL for decelerated rhythm.

For all of the stimulus numbers just described, the values of a were
then determined by using the same curve used to obtain the stimulus parameter values in the main study. This curve is displayed in Figure 7 and described in Section 4.41. When the values of $\alpha$, which correspond to maximum uncertainty of the subject, are subtracted from the value of $\alpha$, which corresponds to a perceptually regular stimulus, the resulting (absolute) $\Delta \alpha$ values represent the best estimate of DL for accelerating ($\Delta \alpha_1$) irregularity or for decelerating ($\Delta \alpha_2$) irregularity. Changes in the degree of irregularity required for perception of irregularity were investigated in both directions since there was no reason to expect symmetry around the neutral stimulus, especially as measured by $\alpha$.

The values of $\Delta \alpha$ thus obtained do not appear significantly different for direction of change, nor for the two curves described above. Therefore, $\Delta \alpha_1$ and $\Delta \alpha_2$, the measures used here to estimate the DL for acceleration and deceleration, have values distributed over a fairly narrow range. Since the second curve (described above) for subject E1 did not exhibit a camel-back shape, his data were not considered further; henceforth, only subjects E2, E4, and E5 are considered.

The values of $\Delta \alpha$ obtained for the three remaining subjects, listed in Table XIV, can be seen to range from approximately 0.20 to 0.50, with both the mean and median $\Delta \alpha$ at approximately 0.30. Hence, for our purpose of obtaining a first estimate for progressive irregularity, the mean $\Delta \alpha$ of 0.30 will be satisfactory.

Since the $\alpha$-values corresponding to the perception of regularity has a mean of $\alpha=0.20$ for these subjects, the $\alpha$-values corresponding to the most uncertain stimuli above and below regularity are at $\alpha=-0.10$ and $\alpha=0.50$, for accelerating and decelerating stimuli, respectively. The durations for
TABLE XIV

Estimated Values of $\Delta a$ for DL

Using The Three Best Subjects

<table>
<thead>
<tr>
<th></th>
<th>Based on Mean</th>
<th>Based on s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accelerating</td>
<td>Decelerating</td>
</tr>
<tr>
<td>0.64</td>
<td>0.40</td>
<td>0.30</td>
</tr>
<tr>
<td>0.46</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td>0.30</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td>0.27</td>
<td>0.34</td>
<td>0.25</td>
</tr>
<tr>
<td>0.27</td>
<td>0.31</td>
<td>0.22</td>
</tr>
<tr>
<td>0.27</td>
<td>0.30</td>
<td>0.22</td>
</tr>
<tr>
<td>0.25</td>
<td>0.29</td>
<td>0.18</td>
</tr>
<tr>
<td>0.21</td>
<td>0.27</td>
<td>0.16</td>
</tr>
<tr>
<td>0.20</td>
<td>0.23</td>
<td>0.16</td>
</tr>
</tbody>
</table>
the six successive syllables, which would correspond to the three a-values given above, are as follows: 143.0, 143.0, 142.7, 141.7, 138.9, and 132.7 ms for the accelerated case, 143.0, 143.0, 143.5, 145.6, 151.4, and 164.3 ms for the perceptually regular case, and 143.0, 143.1, 144.3, 149.6, 164.8, and 202.2 ms for the decelerated case.
The main purpose of this study was to attempt to determine two important factors pertaining to the perception of rhythm:

1. The degree of irregularity required for the listener to perceive a repetitive sequence of elements as regular.
2. The increment or decrement in the degree of irregularity required for the listener to detect that a repetitive sequence of elements is no longer perceptually regular.

In addition, the following questions were of interest:

a. Does the nature of the stimulus (speech versus nonspeech) affect the above two factors?

b. Does the language background of a subject (in particular, English, French, or Japanese) affect the above two factors?

Three listening tests including clicks or CV syllables were prepared and presented to native speakers of English, French, and Japanese. Stimuli consisted of sequences of six sounds (clicks or syllables) which underwent progressive timing alterations according to some predetermined function (D) utilized in all three tests.

Test I contained five repetitions of seven time-altered sequences of clicks as well as seven practice and three buffer items.

Test II contained five repetitions of seven time-altered sequences of [ta] syllables as well as seven practice and three buffer items.

Test III contained five repetitions of seven time-altered sequences of [na] syllables as well as seven practice and three buffer items.
Subjects recorded their answers to each stimulus on a seven-point scale with 1 corresponding to 'speeded up', 4 corresponding to 'regular', and 7 corresponding to 'slowed down'. Subjects' responses were sorted into tables according to stimulus type and order of a particular repetition (see Tables VIII, X, and XI). The subjects who performed best were selected according to criteria based on consistency and low variability. Four subjects passed both of the criteria and the responses of these subjects were considered separately.

The 'averaged' responses of the 4-subject group were compared to those of the entire subject population. For each test type (cf. Figures 11 to 13), curves representing response mean (judgment of regularity) versus stimulus number have the same basic shape, regardless of group size. However, when the variability of these judgments of regularity is considered (cf. Figures 14 to 16), the curves for each one of the two groups do not appear to have the same shape. As expected, the larger group exhibits greater variability than the smaller group, but, more interestingly, the larger group does not tend to exhibit a camel-back shape, whereas the smaller group does. It may be assumed that the 'regular' stimulus should occur in the area of minimum variability and should correspond to a response mean of 4 and that the DL for irregularity should be determined by the interval between this last point and a point in the area of maximum variability (i.e. between the trough and the peak of the variability curve). This appears to be approximately so for the smaller group, but not for the other.

Before discussing the DL for irregularity, it is necessary to comment on the perception of regularity and to examine the results of the ANOVA performed on the data which corresponds to the subjects' perception of
regularity. Table XII displays, for the 24 subjects and for the three tests, the predicted stimulus numbers which correspond to timing irregularities likely to be perceived as regular by each subject and which cover a large, quite evenly distributed range. Such a large range in the stimulus-number values corresponding to a perception of regularity is a possible reason for the absence of a camel-back shape for the 24-subject group in Figures 14 to 16: acoustic isochrony is not identical to perceptual isochrony and, in addition, there is large intersubject variability in the perception of regularity. Although there is a large range of values corresponding to 'regular', there may not be any best place, as measured by $a$, at which to perceive regular.

The ANOVA performed for the balanced, 18-subject group produced no significant results for either the main effects or the interactions (cf. Table XIII). There are several possible reasons for the absence of significant results.

The first possible reason is that the large intra- and intersubject variabilities may mask the significance of any one factor in an ANOVA. Perhaps, a future experiment could attempt to use more homogeneous and more strictly selected subgroups.

A second possible reason for not obtaining any significant differences in this ANOVA is that the level of linguistic knowledge differed between language groups. For instance, when the experimenter was recruiting subjects, there was a large subject population from which to recruit English subjects; a smaller population from which to recruit French subjects; and an even smaller population from which to recruit Japanese subjects. The English subjects were recruited partly for their linguistic knowledge and, therefore,
greater linguistic awareness may be attributed to the English group than to the other two language groups. Analysis of only the English group's data gives results in which the factor 'T' (test type) tends toward significance. This tendency may be due to the greater linguistic awareness of the English group. A future experiment could try to include a larger population base from which to recruit subjects; in this way, language groups could be matched for linguistic knowledge.

A third potential reason for no factor reaching significance in the ANOVA is familiarity with the materials and with the experimental situation: the English subjects were all involved to varying degrees with an audiology and speech science programme and most had experience with materials and situations similar to those of the experiment, whereas few of the French and probably none of the Japanese subjects had had this same experience. Any future experiment should ensure a more homogeneous subject population in this respect.

The perception of regularity may also be compared to the production data of Oiler (1973), referred to in Chapter 4. Differences in the syllable duration ratios for the Oiler data and that of the current study were apparent: the median ratio of the last to next-to-last syllable of Oiler's production data was 1.245, whereas the same ratio corresponding to the perception of regularity in the current study was 1.112 for all subjects and 1.085 for the best three subjects. Together, the two sets of data indicate that acoustic isochrony is not present in the speech signal and that acoustic isochrony is not equal to perceptual isochrony. However, the amount of departure from acoustic isochrony which resulted in perceptual isochrony in the present experiments does not agree closely with the corresponding amount.
found in the production data of Oller. Several variables in the experiments undertaken for the present study may have contributed to these disagreements.

The first variable is the listening mode the subjects are in: the subjects may not have been listening in the speech mode and, hence, may not have judged these stimuli as they would have if they had been listening in the speech mode. There is no adequate way to ensure that subjects listen in a particular mode; however, the use of meaningful speech is more likely to result in this objective than is the use of only clicks and nonsense speech stimuli.

The second variable is the linguistic naivete of the subjects. For example, some of the less linguistically naive subjects were aware of the phenomenon of prepausal lengthening in speech and may have consciously or subconsciously tried to take this phenomenon into account during the test situation. A future experiment could ensure that one subgroup knew and understood such phenomena, while ensuring that another subgroup was linguistically naive, and then compare the results of the two subgroups.

After considering the degree of irregularity which would correspond to a subject's perception of regularity, the estimation of a DL for progressive irregularity may now be considered. The characterization of such a DL is very complex: various parameters can be used to estimate this DL.

When preparing the stimuli, using the function (D), all but one parameter was held constant and only the parameter \( a \), which measures the degree of acceleration or of deceleration, was manipulated. Therefore, this parameter is used to estimate the DL, or, more specifically, the least noticeable increment or decrement in the degree of irregularity. Since there is a tendency for speakers to lengthen syllables as their utterance
progresses and for listeners to impose rhythmic structure on sequences of speech sounds and since there is no reason to expect symmetry, especially as measured by $a$, it was thought necessary to measure separately the DL for acceleration and that for deceleration.

As indicated in Chapter 5 (cf. Table XIV), the estimated DL's for acceleration and deceleration ranged approximately from $\Delta a=0.20$ to $\Delta a=0.50$, with the median being approximately $\Delta a=0.30$. The DL's for acceleration were not significantly different from those for deceleration. However, most subjects commented that the accelerating stimuli were much more difficult to detect than the decelerating stimuli. This observation could be due to the fact that the degree of irregularity in the most accelerated stimulus was not as distant (as measured with $a$) from the degree of irregularity in the neutral stimulus as was the degree of irregularity in the most decelerated stimulus.

The results of this experiment are to be used in a future experiment currently being designed. The purpose will be to obtain a DL for accelerating and decelerating irregularity, using the figures obtained here which correspond to the perception of regularity and to the estimated DL's. The best subjects, namely E2, E4, and E5, will also be subjects in the future experiment, should they be available.

The discovery of the DL for irregularity may have important practical considerations, since rhythm is so important in speech intelligibility.

It is obvious that temporal factors play a vital part in speech. Speech whose temporal pattern is grossly abnormal may be almost incomprehensible, as is often true of the speech of the deaf. (Huggins, 1968)

Rhythm may be a more important factor in aural rehabilitation than
generally acknowledged, for instance, in rehabilitation with cochlear implants, speech compressors, and deaf speech, as well as in strategies used in foreign language teaching. In 1952, Guberina (cited in Northern and Downs, 1978), for example, developed the Verbotonal Method, designed to improve foreign language learning by emphasizing the spoken rhythm of the foreign language. He later applied this method to the teaching of speech to the deaf population - still emphasizing the rhythm of the spoken language.

One of the main conclusions to be drawn from the current experiments is that acoustic isochrony does not result in perceptual isochrony, whether the stimulus consists of speech or nonspeech sounds. The difference between physical and perceptual isochrony, in speech, may be due to two tendencies discussed in Chapters 2 and 3, namely the tendency, by speakers, to lengthen syllables as their utterance progresses, and the tendency, by listeners, to impose a rhythmic structure on speech sequences. Nevertheless, these tendencies do not explain such a phenomenon in nonspeech sounds. Since there were no significant differences between the different types of stimuli, perhaps one may assume the existence of an underlying regular rhythm which governs the listener's perception of regularity, at least at the short-term level. However, this assumption should be further assessed in future experiments, particularly since a review of the literature indicates a (claimed) tendency toward isochrony in the languages of the world and concludes that more of a timing alteration may be necessary for detection of an irregularity in the speech stimuli than in the click stimuli (Lehiste, 1973; 1979). As neither the degree of acoustic irregularity necessary for a subject to perceive regularity, nor the estimated DL for timing irregularity is greater for speech stimuli than for nonspeech stimuli, these results do
not lend support to the hypothesis that the listener's perception of rhythm is altered by the expectancy of isochrony in speech. If correct, this hypothesis would imply that the acoustic signal is transmitted and perceived as only quasi-regular by the peripheral nervous system, but that at some higher or more central level of processing, due to the nature of the stimuli (namely speech stimuli), the signal would be interpreted as regular, whereas it would not be expected to be the case for nonspeech stimuli.

Summary

The results of this investigation provided information regarding to the following:

1. The degree of irregularity; when measured by the parameter \( a \), there is a relatively large range of \( a \)-values perceived as 'regular'; there is also great intersubject variability. For all subjects, the perceptually regular stimulus was decelerated.

2. An estimate of DL, measured by \( \Delta a \), for accelerating and decelerating irregularity; the required just noticeable increment or decrement in the degree of irregularity was estimated utilizing only the results of the best three subjects. There is a relatively narrow range of \( \Delta a \) values for this estimate.

Due to the intra- and intersubject variabilities, it was not possible to ascertain whether the language background of a subject or the nature of the sound stimulus affect the perception of rhythm. Similarly, it was not possible to ascertain whether the direction of irregularity, increment versus decrement, affects the DL for irregularity.

The present research is seen as a first step in obtaining the DL for
progressive temporal irregularity at the short-term level in speech and
nonspeech auditory events. A continuation of this experiment has been
proposed in order to refine the results of the present experiments, as well
as to control possible confounding variables.

Previous studies have demonstrated that rhythm at the short-term
level may have important implications for speech perception; the present
study is a step toward discovering more about the perception of rhythm.


Delattre, P. (1966) A comparison of syllable length conditioning among languages. IRAL, IV/3


APPENDIX A

SUBJECT INFORMATION FORM

Identities of all subjects will remain confidential.

Name: ________________________________

Phone: ________________________________

Age: ________________________________

Sex: ________________________________

Native Language: ________________________________

Additional languages spoken: ________________________________ Fluency: ________

_________________________ Fluency: ________

_________________________ Fluency: ________

Have you completed any linguistics courses? ________________________________

If yes, approximately how many? ________________________________

Do you have any hearing problems? ________________________________

If yes, describe problem: ________________________________
APPENDIX B

INSTRUCTIONS

You are going to hear sound stimuli, each one of which consists of a sequence of six clicks, or of six repeated syllables. I would like you to rate their degree of regularity or of irregularity.

If the sequence of clicks (or syllables) sounds speeding up, check the left column; if it sounds regular, check the center column; and if it sounds slowing down, check the right column. If you feel it is in-between these categories, feel free to use the intermediate columns. For example, if it is just slightly speeding up, you may want to check one of the columns between regular and speeded up.

Please mark only one column for each item, but give an answer for each item.

In these tests, there is no right or wrong answer; your subjective evaluation of the items is what matters, and it is all I am interested in.

Before starting to answer, you will be able to listen to the first seven items, including some examples of the most irregular stimuli you will hear. I will then stop the tape to see if you have any questions. When you are ready, the tape will be restarted at item one and you should then start recording your answers.