A Characterization of American English Intonation

by

Shinji Maeda

S.B., The University of Electro-Communications (1966)
S.M., The University of Electro-Communications (1968)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology

January, 1976

Signature of Author
Department of Electrical Engineering, January 30, 1976

Certified by ——— Thesis Supervisor

Accepted by Chairman, Departmental Committee on Graduate Students
Eng. APR 30 1976

APR 2 1976.
A CHARACTERIZATION OF AMERICAN ENGLISH INTONATION

BY

Shinji Maeda

Submitted to the Department of Electrical Engineering on January 30, 1976 in partial fulfillment of the requirements for the degree of Doctor of Philosophy

ABSTRACT

We have investigated how American English intonation can be represented in a meaningful manner acoustically, physiologically, perceptually, and linguistically. Chapter 1 describes briefly our approach for studying intonation.

In Chapter 2, the fundamental frequency contours of 39 isolated sentences and 14 sentences in a text read by several speakers are matched visually with piecewise-linear patterns. These "schematized fundamental frequency patterns" are specified by a set of symbols (attributes). The attributes rise (R) and lowering (L) characterize upward and downward rapid movements of the fundamental frequency contours, respectively, and baseline (BL) represents a gradual fall of the fundamental frequency along a sentence. The attributes R and L often appear as a pair, and thus the schematized fundamental frequency pattern exhibits a so-called "hat-pattern". In addition to these basic attributes, two more attributes are postulated: a fundamental frequency peak (P) which often occurs with R, and a rise (R1), with a relatively slow rate of rise, on the plateau of the hat-pattern. The fundamental frequency contours of the sentences are, thus, characterized by sequences of these attributes (attribute patterns). The reset of BL signals the onset of a major constituent of a sentence (often the sentence itself), and R and L mark lexical stresses in the words. It was found that the attribute patterns contain certain information concerning the structure of a sentence. This information is reflected in groupings and subgroupings of the words. A set of rules relating these groupings and subgroupings to the observed attribute patterns is proposed. The groupings and subgroupings often correspond to constituents and to the internal structure of constituents, respectively. However, whether any particular constituent is actually manifested in the attribute pattern seems to vary from one speaker to another, and perhaps from time to time. At least
two factors, besides constituent structures, seem to affect the groupings: emphasis placed on certain words in a sentence, and economy in the manner in which stresses are marked on word sequences.

Chapter 3 reports on an investigation of the physiological correlates of the attributes. The movements of the larynx and the change of the laryngeal ventricle length, which corresponds to the vocal-fold length, were measured in a cineradiographic experiment for four sentences read by one speaker. It was found that attribute L was accompanied by a fall in the laryngeal height, and BL seemed to be related to a gradual shortening of the ventricle length throughout the sentence. The electromyographic activities of certain intrinsic and extrinsic laryngeal muscles for 24 sentences read by two speakers were investigated. The temporal relation between the cricothyroid, and the sternothyroid or the sternohyoid activities seemed to distinguish L from R and P. An active fundamental frequency lowering mechanism is proposed in this chapter.

In Chapter 4, a simple transformation in which the attribute patterns of sentences are mapped (coded) into the fundamental frequency contours is postulated. A set of utterances was synthesized using the rule-generated fundamental frequency contours derived from a variety of attribute patterns for one sentence. Listeners interpreted those utterances in a consistent manner depending on the attribute patterns. For instance, the attribute patterns which cannot be specified by the rules were often rejected as not having American English intonation. The results of the perceptual experiment suggest that the rules characterize a certain aspect of American English intonation, but more research is needed in this area.

THESIS SUPERVISOR: Kenneth N. Stevens

TITLE: Professor of Electrical Engineering
ACKNOWLEDGEMENTS

I am extremely grateful to Professor Kenneth N. Stevens for allowing me to join the Speech Communication group of the Research Laboratory of Electronics at M.I.T. Without his constant encouragement and support, this thesis would not have been possible. I thank him for his courage in being a subject for both the electromyographic experiment and the cineradiographic experiment described in this thesis.

I have benefited from assistance given me by my thesis readers, Prof. Jonathan Allen, Prof. Morris Halle, and Dr. Dennis H. Klatt. Dr. Osamu Fujimura at Bell Laboratories has generously given his time in providing valuable comments, especially at the final stage of this thesis work. I can only express my regret that their constructive criticisms and insightful suggestions are not fully reflected in this thesis.

I have profited from various discussions with those in the Speech Communication group, in particular, Dr. Thomas Baer and Dr. Joseph S. Perkell. Dr. Thomas Baer has also served as a subject for the electromyographic experiment reported here. Dr. William Huggins, Mr. Bernard Mezrich, and Mr. Michael Portnoff have been extremely helpful in numerous computer-related problems. Mr. Keith North has provided technical support.

I would like to thank Mr. Robert Donahue at the Cardiac Catheterization Laboratory of the Massachusetts General Hospital for kindly giving his time in the cineradiographic experiment.

I would like to thank Dr. Tatsuiro Ushijima, Dr. Seiji Niimi, Prof. Katherine S. Harris, and Dr. Franklin S. Cooper, at Haskins Laboratories for their kindness and their help in the electromyographic experiments. The electromyographic facilities at Haskins Laboratories are supported by a grant from the National Institute of Dental Research.

I thank Dr. Max Mathews and Dr. Peter Denes both at Bell Laboratories for helping me to enter M.I.T.

To my wife, Jacqueline, I owe special thanks for her help in all stages of the preparation of this work, from productive discussions to typing of the first manuscript of this thesis.

This research was supported in part by a grant from the National Institutes of Health (grant NS04332).
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>2</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>4</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>5</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>8</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>9</td>
</tr>
<tr>
<td>Chapter I Introduction</td>
<td>12</td>
</tr>
<tr>
<td>Chapter II A Schematic Analysis of the Fundamental Frequency Contours of Speech Signals</td>
<td>16</td>
</tr>
<tr>
<td>2.1 A Schematic Analysis of F₀ Contours of Speech: Background</td>
<td>16</td>
</tr>
<tr>
<td>2.2 Experimental Procedure: Corpus and F₀ Detection Program</td>
<td>29</td>
</tr>
<tr>
<td>2.3 Attributes</td>
<td>38</td>
</tr>
<tr>
<td>2.3.1 Structure of the Schematized F₀ Patterns</td>
<td>39</td>
</tr>
<tr>
<td>2.3.2 Baseline (BL)</td>
<td>46</td>
</tr>
<tr>
<td>2.3.3 Rise (R), Peak (P), and Lowering (L)</td>
<td>68</td>
</tr>
<tr>
<td>2.3.4 Effect of the Consonant on the F₀ Contour</td>
<td>85</td>
</tr>
<tr>
<td>2.3.5 Rise on the Plateau (RI)</td>
<td>93</td>
</tr>
<tr>
<td>2.3.6 A Function of the Basic Attributes: Stress-marking</td>
<td>96</td>
</tr>
<tr>
<td>2.4 The Attribute Patterns and Constituents of Sentences</td>
<td>99</td>
</tr>
<tr>
<td>2.4.1 Empirical Hypothesis Concerning Attribute Patterns</td>
<td>100</td>
</tr>
<tr>
<td>2.4.2 The Attribute Patterns Associated with Noun Phrases and Compound Words Composed of Two Lexical Words</td>
<td>103</td>
</tr>
<tr>
<td>2.4.3 Attribute Patterns in Noun Phrases with Various Constituents Structures</td>
<td>117</td>
</tr>
<tr>
<td>2.4.3.1 Noun Phrases with Right-Branched Structure</td>
<td>117</td>
</tr>
<tr>
<td>2.4.3.2 Noun Phrases with Left-Branched Structure</td>
<td>132</td>
</tr>
<tr>
<td>2.4.3.3 Words Containing More Than One Pair of the Basic Attributes R and L</td>
<td>145</td>
</tr>
</tbody>
</table>
### Chapter III Physiological Correlates of the Attribute Patterns

#### 3.1 Studies on the $F_0$ control in Speech: Background
- Page 196

#### 3.2 Laryngeal Dynamics During Speech
- 3.2.1 Procedure
- Page 206
- 3.2.2 The Vertical Movements of the Larynx
- Page 211
- 3.2.3 Variations in the Ventricle Length
- Page 222

#### 3.3 EMG Activities of the Laryngeal Muscles During Speech
- 3.3.1 Procedure
- Page 232
- 3.3.2 The Attributes and the EMG activities
- Page 234
- 3.3.3 Emphasis, and Intraspeaker Differences in the Manner of its Generation
- Page 246
- 3.3.4 Influence of Voiced and Voiceless Stops upon the $F_0$ contours
- Page 258

#### 3.4 Speculation of the $F_0$ Control Mechanisms
- 3.4.1 The Mechanism Generating the Baseline
- Page 263
- 3.4.2 A Simple Laryngeal Model Interpreting the Properties of the Localized $F_0$ movements
- Page 270
- 3.4.3 A Speculation on the Active $F_0$ Lowering Mechanism
- Page 283

#### 3.5 Summary of This Chapter
- Page 286

### Chapter IV Some Speech Synthesis Experiments on Attribute Patterns: A Preliminary Study

#### 4.1 Synthesis of Stimuli: A Transformation from Attribute Patterns to the $F_0$ Contours
- Page 289
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2 Perceptual Adequacy of the Piecewise-Linear Approximation to the Rule-Generated $F_o$ Contours</td>
<td>293</td>
</tr>
<tr>
<td>4.3 Some Linguistic and Perceptual Effects Due to the Variation of Attribute Patterns</td>
<td>298</td>
</tr>
<tr>
<td>4.4 Summary of This Chapter</td>
<td>311</td>
</tr>
<tr>
<td>Chapter V Conclusions and Some Remarks</td>
<td>314</td>
</tr>
<tr>
<td>FOOTNOTES</td>
<td>321</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>322</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 The first part of the corpus (S1-S30)</td>
<td>30</td>
</tr>
<tr>
<td>2.2 The second part of the corpus (S31-S44)</td>
<td>30</td>
</tr>
<tr>
<td>2.3 The third part of the corpus (S45-S53)</td>
<td>30</td>
</tr>
<tr>
<td>2.4 Measurements of the falling rate of the baseline</td>
<td>53</td>
</tr>
<tr>
<td>2.5 Average ($\bar{F}$), standard deviation ($\sigma$), coefficient of variation (c.o.v.) and the range of the measured $F_0$ values at terminal points</td>
<td>60</td>
</tr>
<tr>
<td>2.6 The average magnitude and duration of the localized $F_0$ movements</td>
<td>77</td>
</tr>
<tr>
<td>2.7 A summary of the rules</td>
<td>154</td>
</tr>
<tr>
<td>3.1 Sentences used in the physiological experiments (S54-S77)</td>
<td>207</td>
</tr>
<tr>
<td>4.1 The list of attribute patterns and the corresponding stylized $F_0$ patterns for S15</td>
<td>299</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The $F_0$ contours of S31 read by KN, JP, and KS</td>
<td>24</td>
</tr>
<tr>
<td>2.2</td>
<td>An idealized description of the schematized $F_0$ patterns</td>
<td>41</td>
</tr>
<tr>
<td>2.3</td>
<td>An example of the $F_0$ contour and the corresponding schematized $F_0$ pattern</td>
<td>44</td>
</tr>
<tr>
<td>2.4</td>
<td>Superimposed $F_0$ contours of the 30 sentences</td>
<td>47</td>
</tr>
<tr>
<td>2.5</td>
<td>Relation between the falling rate ($r$) of the baseline and the duration ($t$) of the 30 sentences</td>
<td>56</td>
</tr>
<tr>
<td>2.6</td>
<td>Definition of the terminal point</td>
<td>56</td>
</tr>
<tr>
<td>2.7</td>
<td>The $F_0$ contours and the corresponding baselines</td>
<td>64</td>
</tr>
<tr>
<td>2.8</td>
<td>The $F_0$ contours and the corresponding amplitude envelope</td>
<td>71</td>
</tr>
<tr>
<td>2.9</td>
<td>The superposition of the localized $F_0$ movements</td>
<td>73</td>
</tr>
<tr>
<td>2.10</td>
<td>The relationship between the duration and the amplitude of the $F_0$ movements</td>
<td>79</td>
</tr>
<tr>
<td>2.11</td>
<td>The $F_0$ contours and the amplitude envelopes</td>
<td>86</td>
</tr>
<tr>
<td>2.12</td>
<td>The sound spectrographs of the beginning of the sentences S8 and S1</td>
<td>89</td>
</tr>
<tr>
<td>2.13</td>
<td>The $F_0$ contours of noun phrases</td>
<td>94</td>
</tr>
<tr>
<td>2.14</td>
<td>The $F_0$ contours of noun phrases</td>
<td>104</td>
</tr>
<tr>
<td>2.15</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for S45</td>
<td>110</td>
</tr>
<tr>
<td>2.16</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for S29</td>
<td>113</td>
</tr>
<tr>
<td>2.17</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for S47</td>
<td>122</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>2.18</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for $S_{11}$</td>
<td>127</td>
</tr>
<tr>
<td>2.19</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for the noun phrase &quot;the small black fat cat&quot;</td>
<td>133</td>
</tr>
<tr>
<td>2.20</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for $S_{46}$</td>
<td>135</td>
</tr>
<tr>
<td>2.21</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for $S_{48}$</td>
<td>140</td>
</tr>
<tr>
<td>2.22</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for $S_{53}$</td>
<td>143</td>
</tr>
<tr>
<td>2.23</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for $S_{50}$</td>
<td>147</td>
</tr>
<tr>
<td>2.24</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for $S_{13}$ and $S_{12}$</td>
<td>158</td>
</tr>
<tr>
<td>2.25</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for $S_{52}$ and $S_{51}$</td>
<td>162</td>
</tr>
<tr>
<td>2.26</td>
<td>The generation of the possible attribute patterns for the noun phrase, Adj+N+N, with left-branched and with right-branched structure</td>
<td>167</td>
</tr>
<tr>
<td>2.27</td>
<td>The generation of the possible attribute patterns for the noun phrase, Adj+Adj+N, with left-branched and with right-branched structure</td>
<td>167</td>
</tr>
<tr>
<td>2.28</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for $S_{1}$</td>
<td>174</td>
</tr>
<tr>
<td>2.29</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for $S_{3}$</td>
<td>177</td>
</tr>
<tr>
<td>2.30</td>
<td>The $F_0$ contours and the corresponding schematized $F_0$ patterns for $S_{7}$</td>
<td>181</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>2.31</td>
<td>The F&lt;sub&gt;0&lt;/sub&gt; contours and the corresponding schematized F&lt;sub&gt;0&lt;/sub&gt; patterns for 4 sentences read by KS.</td>
<td>184</td>
</tr>
<tr>
<td>2.32</td>
<td>A state transition network accepting any observed sequence of the attributes.</td>
<td>191</td>
</tr>
<tr>
<td>3.1</td>
<td>Schematic drawings of the laryngeal cartilages and the muscles.</td>
<td>198</td>
</tr>
<tr>
<td>3.2</td>
<td>An example of the lateral x-ray tracing.</td>
<td>209</td>
</tr>
<tr>
<td>3.3</td>
<td>The movements of the mandible, the hyoid bone, and the thyroid cartilage.</td>
<td>212</td>
</tr>
<tr>
<td>3.4</td>
<td>The fluctuation of the laryngeal ventricle length.</td>
<td>223</td>
</tr>
<tr>
<td>3.5</td>
<td>The vocal-fold length vs. F&lt;sub&gt;0&lt;/sub&gt; relationship.</td>
<td>230</td>
</tr>
<tr>
<td>3.6</td>
<td>The F&lt;sub&gt;0&lt;/sub&gt; contours and the corresponding EMG activities of the laryngeal muscles.</td>
<td>235</td>
</tr>
<tr>
<td>3.7</td>
<td>The EMG peak level and the temporal relation between CT peak and the corresponding ST peak.</td>
<td>242</td>
</tr>
<tr>
<td>3.8</td>
<td>Influence of emphasis on the F&lt;sub&gt;0&lt;/sub&gt; contours and on the corresponding EMG activities for speaker KS.</td>
<td>247</td>
</tr>
<tr>
<td>3.9</td>
<td>Influence of emphasis on the F&lt;sub&gt;0&lt;/sub&gt; contours and on the corresponding EMG activities for speaker TB.</td>
<td>247</td>
</tr>
<tr>
<td>3.10</td>
<td>Influence of the voiced/voiceless contrast in an initial consonant cluster.</td>
<td>260</td>
</tr>
<tr>
<td>3.11</td>
<td>A schematic representation of the force acting on the cricoid cartilage.</td>
<td>264</td>
</tr>
<tr>
<td>3.12</td>
<td>A visco-elastic model of a skeletal muscle.</td>
<td>272</td>
</tr>
<tr>
<td>3.13</td>
<td>The impulse response of the F&lt;sub&gt;0&lt;/sub&gt; control model.</td>
<td>279</td>
</tr>
<tr>
<td>4.1</td>
<td>An example of the rule-generated F&lt;sub&gt;0&lt;/sub&gt; contour.</td>
<td>295</td>
</tr>
<tr>
<td>4.2</td>
<td>One of the rule-generated F&lt;sub&gt;0&lt;/sub&gt; contours for S15.</td>
<td>299</td>
</tr>
</tbody>
</table>
Chapter I  Introduction

The objective of this study is to find a meaningful representation of American English intonation. Our analysis of intonation is based on physical data: the fundamental frequency (F₀) contours of the speech signals, electromyographic measurements of the muscle activities that control pitch, and cineradiographic data of the larynx movement. Although we use a small set of simple declarative sentences, the approach we have taken is directed toward universal aspects of intonation. We have attempted to analyze intonation in terms of a limited number of attributes that characterize the F₀ contours. The five attributes, rising, lowering, peak, and so on, are postulated on the basis of a schematic analysis of the F₀ contours by using a visual abstraction procedure. The abstraction of schematized patterns is an essential part of the study. It is recognized that some aspects of linguistic messages are coded into speech signals by means of intonation, in particular F₀ contours. However, we do not know the mechanisms of the coding, and our aim is to abstract common attributes that characterize the F₀ contours of the set of sentences. A speaker must expend specific physiological effort to realize each of the F₀ movements that are characterized by the common attributes. The sequence of the attributes, therefore, may be regarded as a discrete signal which transmits information about linguistic messages.
Although the five attributes seem to describe adequately all the $F_0$ contours analyzed in this study, the description using the attributes might be considered to merely provide an arbitrary approximation to representative curves. An important question, therefore, is whether these attributes are meaningful both linguistically and physiologically; in other words, whether they reflect the underlying mechanisms of the generation of intonation. We have attempted to show that people actually use these attributes of intonation for sending a linguistic message. We have also tried to show, in the electromyographic and cineradiographic experiments, that each of the five attributes is related to a specific coordination of the physiological gestures.

We are aware of the fact that intonation signals distinctions in sentence modes, such as declarative, interrogative, imperative and so on. Only declarative sentences, occasionally with emphasis, are studied here. Therefore, we do not know whether the five attributes adequately specify the distinction of such sentence modes. We also avoid consideration of the emotional aspects of intonation. It is well known that intonation carries information concerning the emotional state of a speaker, but this problem is beyond the scope of our study.
In the course of this study, a number of interesting insights into the phenomena of intonation have been obtained. In most cases, the location of attributes such as rising and lowering correspond to stressed syllables of certain lexical words in a sentence (as expected). The sequences of attributes, which we shall call attribute patterns, associated with sentences seem to be organized in such a manner that each stress is efficiently represented on the $F_0$ contours. It is suggested that a principle of economy in physiology is one of the factors which constrain the attribute patterns (that we consider as a discrete representation of intonation). Other factors that constrain the attribute patterns are the local grammatical structures and emphasis of one or more words in each sentence. We have also found, on the basis of speech synthesis experiments, that a set of simple rules is sufficient to transform given attribute sequences into the corresponding $F_0$ contours.

As a final introductory comment, we should explain why we choose $F_0$ contours as material for studying intonation. As mentioned before, our study aims toward deeper understanding of the mechanisms underlying intonational phenomena in speech. In such a study, one must have in mind a model of the generation of intonation. A basic concept of such a model is a notion of phonological features, which are defined as
the minimal linguistic message units. We recognize the features as control signals to the speech peripheral mechanisms, such as the larynx and the respiratory system, when we deal with intonation phenomena. Unfortunately, the behavior of the physiological mechanisms is poorly understood. We postulate, however, that the controlled elements are the states of the vocal folds, because the states primarily govern the oscillatory patterns of the vocal folds. Since there is no way of studying directly both the control signals to the peripheral structures and the states of the vocal folds such as the vocal-fold stiffness and the glottal opening, it seems to be quite natural to investigate the $F_0$ contours in which the states are directly reflected. Also, $F_0$ contours are the signals that must be interpreted by listeners. It is well known that intonation is primarily correlated with the $F_0$ variation of speech. In other words, we cannot discuss intonation without the $F_0$ contours of speech signals.
Chapter II A Schematic Analysis of the Fundamental Frequency Contours of Speech Signals

In this chapter, we shall show that the fundamental frequency ($F_0$) contours of declarative sentences can be characterized by using the five attributes: baseline $BL$, rise $R$, peak $P$, lowering $L$, and a rise on the plateau (We call 'plateau' the portion between the rise $R$ and the lowering $L$.) we will then try to show how speakers use a particular attribute pattern for signaling a certain aspect of a linguistic message. It will be postulated that such attribute patterns are determined by the combination of the relevant aspect of the linguistic message and a principle of economy in physiology of speech production.

2.1 A Schematic Analysis of $F_0$ Contours of Speech: Background

In the past, many noteworthy studies of intonation have been undertaken by linguists. Intonation is considered as pitch movements, and the perceptual impression of linguists is represented by a pattern drawing, such as in the studies by Armstrong and Ward (1926) and Jones (1932). Investigation of intonation and meaning of sentences has led linguists to consider intonation comparable with morphemes, and pitch with phonemes (Bloomfield, 1933). Wells (1945) specifies four different levels of pitch as distinctive pitch phonemes. The transcription of intonation by the
levels of pitch was developed further by Pike (1945), and Trager and Smith (1951). Trager and Smith (1951) introduced different types of pauses and lexical stress levels, which together the four pitch levels were used to describe the entire prosody of sentences. Probably due to the peculiarity of analysis using the auditory perception, these linguists considered intensity as the acoustic correlate of lexical stress. In short, lexical stress is related to loudness (intensity), and intonation to pitch (fundamental frequency of the voice). The level representation of stress is widely used among American linguists. More recently with the development of theory of generative phonology, prediction of stress patterns in sentences has been highly elaborated. Chomsky and Halle (1968), and Halle and Keyser (1971) has proposed a procedure for predicting stress patterns of sentences depending on their surface structure. Bresnan (1971) shows that the deep structure determines more generally the stress assignment in a predictable way. It should be noticed, however, that a prosodic contour represented by the sequence of stress levels may not be necessary to correlate directly with the $F_0$ contour of a sentence.

Another group of linguists uses configurations rather than levels to describe intonation. Pitch contours are transcribed by using distinctive pitch movements, or so-called tones. In the tonetic representation of intonation,
no clear distinction between lexical stress and pitch movements is made. Pitch movements corresponding to stressed syllables are represented by such tones as level, rise and fall. Stress and pitch movements are treated as related elements, and not entirely independent. Interestingly, the linguists who prefer to use the tonetic representation think that intonation is constrained by the meaning of a sentence instead of by the syntax. (Bolinger, 1972; Crystal, 1969; Stockwell, 1962; Halliday, 1967). Although the intonation system proposed by Halliday (1967) takes into account the major syntactic constituents, it is meaning (specifically the information focus) rather than syntax that determines intonation within each constituent.

The transcription system based on perceptual impressions is criticized by Lieberman (1965). He points out that the transcriptions of intonation in terms of Trager-Smith notation by two linguists are not always consistently related to the physical reality of speech, specifically the fundamental frequency ($F_0$) of the voice. However, in his experiment, one of the two linguists showed better performance in describing pitch movements in terms of configuration than in terms of the levels. But probably the more serious disadvantage of the studies based on auditory impressions of speech sounds is the lack of experimental means for checking whether or not the analysis is correct.
Recently, researchers have become interested in the study of super segmental phenomena, including stress, and intonation with respect to acoustic sound, such as $F_0$ values, intensity, and segmental durations. Using new tools, such as the sound spectograph, the vocoder, and the speech synthesizer, a number of researchers have tackled the problem of the acoustical correlates of lexical stress. Fry (1955) tested the relative importance of duration and intensity as acoustic correlates of stress, in an analysis experiment. He claims that duration is more important than intensity. He further showed, in a perceptual experiment (1958), that $F_0$ is the primary cue for the perception of stress. In the experiment, the stimuli, noun-verb word pairs, such as the pair 'OBJECT' versus 'obJECT', were synthesized varying the $F_0$ contour, the duration and the intensity independently. Morton and Jassem (1965) also undertook a perceptual experiment using synthetic nonsense syllables. They concluded that $F_0$ is the primary cue to stress, and that a rising $F_0$ contour is more effective in stress-marking than a falling one. Denes (1959), and Denes and Milton-Williams (1962) have demonstrated that $F_0$ contours are dominant cues to the perception of tones, although they suggested the existence of a complex interaction among $F_0$ contour, segmental duration and intensity. Lieberman (1960) also noted that there is some blending effect among these acoustic cues.
These studies have shown that lexical stress is highly correlated with $F_0$ values, probably with specific $F_0$ patterns such as rise and fall. We expect that the $F_0$ contours of sentences are strongly influenced by lexical stresses. Rather surprisingly, few studies have been undertaken to investigate how stress-marking is made at the level of the sentence, or how lexical stress is distinguished from emphatic stress. Bolinger (1958) claims, based on his analysis and synthesis experiments, that $F_0$ is the primary acoustic correlate of stress. He further notes that lexical stress only has the potential to receive pitch accent in a sentence. Lexical stress in a word that contains important information is manifested as rapid $F_0$ movement in the $F_0$ contours.

Lieberman (1967) has tried to characterize intonation of sentences in terms of two features: Prominence and Breath-group. Absence or presence of a $F_0$ peak is regarded as a manifestation of the opposition [- Prominence] vs. [+ Prominence]. The opposition [- Breath-group] vs. [+ Breath-group] corresponds to absence or presence of a final rise in a basic rise-fall $F_0$ pattern. In this respect, the two feature system may be regarded as a pattern representation of the $F_0$ contours. Incidentally, the two basic patterns, Tune-1 and Tune-2 in Armstrong and Ward (1926) correspond to the manifestations of [- Breath-group] and [+ Breath-group], respectively. These two hypothetical features have been
investigated in further detail in terms of acoustic and physiological studies (Atkinson, 1973). Although the reports of Lieberman (1967) and Atkinson (1973) have shown insights into the underlying mechanisms of intonational phenomena, little attention has been paid to the F₀ movements inside the breath-groups, except one F₀ peak which is created, in their claim, by a momentary increase in sub-glottal pressure on a vowel. We feel, as discussed below, that these two features are not sufficient to specify intonation of American English.

Cohen and t'Hart (1967), and t'Hart and Cohen (1973). These two researchers studied Dutch sentence intonation by using an analysis-by-synthesis technique in which a piecewise-linear trapezoidal representation is used to approximate the F₀ contours. The artificial F₀ contour is manipulated so that the perceived intonation of the synthetic speech is perceived to be identical to that of the original speech. According to their results, Dutch intonation is well characterized by the so-called "hat-pattern", which is composed of a rise and a plateau followed by a rapid fall. The F₀ contour of a sentence can be approximated by the superposition of the "hat-patterns" and a gradual fall along the entire sentence, which they call "declination line". A search for the inventory of distinctive patterns was further undertaken, and other
patterns such as the "valley pattern" and the "cap pattern" were found (Collier and t'Hart, 1972). In these studies, the authors emphasize that the analysis of $F_0$ contours is not effective for studying intonation, since $F_0$ movements which are relevant to the perception of intonation are not easy to determine from the actual $F_0$ contours of speech. However, our preliminary investigation of $F_0$ contours has indicated that a characteristic pattern like the "hat-pattern", could be rather easily abstracted from the $F_0$ contours of American English sentences.

Hatori (1961) characterized the pitch contours of Japanese words in terms of two types of features, an accent kernel and prosodemes, which apparently correspond to a prosodic feature and configurational features defined in Jakobson, Fant and Halle (1969). The pitch contours of the words are represented by basic configurational patterns, specified by prosodemes, each of which is modulated depending on the location of the accent kernel in the word generating a distinctive pitch pattern. Fujimura (1972) formulated a mathematical model for generating the corresponding $F_0$ contours from the pitch patterns specified by these features.

In the case of American English, it is no doubt that lexical stresses play an important role in the specification of the $F_0$ contours, especially in the localized $F_0$ movements.
In this regard, it is not necessary to be suitable to represent the $F_0$ contours in terms of patterns, although we often observe regular patterns in the $F_0$ contours, such as a "hat-pattern".

We shall attempt to describe the $F_0$ contours using a limited number of elements which characterize localized and non-localized $F_0$ movements separately. It may be appropriate to show some examples at this point, to explain our purpose. In Fig. 2.1, we show the $F_0$ contours and the amplitude envelopes of the beginning of a long sentence "In the jungle of Asia, there is a large bird with brilliant colors, red feathers on the wings...", read by three speakers. Even though these contours differ quantitatively from each other, we can see some similarities between the curves. These similarities may become clear when described qualitatively as follows: Each contour is raised during the first stressed syllable, in the word 'jungle', and the rise is associated with a large peak, indicated by the letter 'P'. Then the contour is lowered during the second syllable of "Asia". The $F_0$ dips observed during the voiced consonants /v/ in the word 'of', and /ʒ/ in the word "Asia" are ignored, since these dips are considered to be an acoustical effect due to the manner of the production of the voiced consonants (Vaissiere 1971, Lea 1973) upon the $F_0$ values. Notice the $F_0$ dips correspond to valleys in the amplitude envelope. During the first phrase "In the
Figure 2.1

$F_0$ contours of the sentence S31 read by three speakers, KN in (a), JP in (b) and KS in (c), and the corresponding schematized patterns. A.E. in each figure represents the amplitude envelope.
In the jungles of Asia there is a large bird with brilliant colors, red feathers on the wings, yellow....

\[ F_o (\text{Hz}) \]

\[ 170 \]
\[ 130 \]
\[ 90 \]

TIME (sec)

\( (100\%) \)

Fig. 2.1 (a) KS S31
In the jungles of Asia, there is a large bird with brilliant colors, red feathers on the wings...

Fig. 2.1 (b) JP S31
In the jungles of Asia there is a large bird with brilliant colors, red feathers on the wings, yellow...
jungle of Asia", the whole contour may be regarded as gradually falling (indicated by the dashed lines), and the gradual fall is raised at the beginning of the next phrase, except for the speaker KN. The contour is raised again in the word "large" and this rise is associated with a smaller peak than that of the first phrase (on the word "jungle"). Then the contour is lowered on the word 'bird', and so on. This description may be represented by the schematic drawing shown in Fig. 2.1, in which a piecewise-linear approximation is superimposed on each original contour. The schematic F₀ pattern may differ distinctively from one speaker to another (see, for instance, the difference in the pattern for the word "brilliant" in the second phrase of the sentence for the speaker JP and the two other speakers). The schematic patterns can be specified by common configurational elements, such as the baseline, represented by the dashed lines; a rise; a plateau that is parallel to the baseline; a lowering and a peak that in these examples occurs with the initial rise. Since any contour can be characterized by using a small numbers of common elements, the F₀ contours for the three speakers seem somewhat similar, even though they may differ in the combination of the elements.

The main objectives in this chapter are first to study the physical properties of the configurational elements (in terms of the F₀ values and the duration of the F₀ contours),
and second, to find out what factors govern the organization of the sequence of attributes. It has to be kept in mind that the attributes are abstracted categories of the configurational elements: the latter have different physical properties depending on the individual speakers, while the attributes (such as Rise, Lowering and so on) are common symbols used for describing the $F_0$ contours of all speakers, and are assumed to manifest a certain linguistic unit.

2.2 Experimental Procedure: Corpus and $F_0$ Detection Program

The corpus to be analyzed has three parts. One is a set of thirty isolated sentences composed of mostly multisyllabic words, which are listed in Table 2.1. This part of the corpus was designed based on the results of a preliminary study, in which we analyzed sixty isolated sentences spoken by one speaker. The structure of the sentences listed in Table 2.1 is quite simple: a noun phrase as the subject followed by a simple verb and a noun phrase or prepositional phrase(s). The length of the sentences is systematically manipulated in terms of the number of syllables in the words and the number of syllables in the noun or prepositional phrases. The second part is a text, entitled 'Chickens' (by Amorosi and Bowles, 1971) and it consists of fourteen sentences (the text is given in Table 2.2.). The second part is used primarily to investigate how far the results obtained
Table 2.1

The first part of the corpus.

Table 2.2

The second part of the corpus.

Table 2.3

The third part of the corpus.
Table 2.1

S 1. The cat likes the dog in the mud.
S 2. The cat likes the alligator in the mud.
S 3. The cat likes the dog in the mud in the park.
S 4. The cat likes the alligator in the mud in the park.
S 5. The cat likes the yellow dog in the mud.
S 6. The cat likes the yellow alligator in the mud.
S 7. The cat likes the dog in the yellow mud.
S 8. The cat likes the alligator in the yellow mud.
S 9. The white dog likes the small black boy.
S 10. The big white dog likes the small black boy.
S 11. The big white dog likes the small black cat.
S 12. The dog likes the (small) (school boy).
S 13. The dog likes the (small school) (boy).
S 14. The dog likes the enormous monkey.
S 15. The dog likes the enormous gorilla.
S 16. The dog likes the enormous kangaroo.
S 17. The dog likes the paralyzed monkey.
S 18. The dog likes the paralyzed kangaroo.
S 20. The dog likes the magnificent monkey.
S 21. The dog likes the magnificent gorilla.
S 22. The dog likes the magnificent kangaroo.
S 23. The big white dog likes the yellow monkey.
S 24. The big white dog likes the big yellow monkey.
S 25. The big white dog likes the enormous yellow monkey.
S 26. The big white dog likes the magnificent yellow monkey.
S 27. The big white dog likes the magnificent yellow cat.
S 28. The big white dog likes the magnificent yellow alligator.
S 29. The big white dog likes the big yellow alligator.
S 30. The cat likes the dog in the puddle.
In the jungles of Asia, there is a large bird with brilliant colors - red feathers on the wings, yellow on the neck and head, black on the tail. It is hard to believe that our common chicken is related to this jungle bird, but it is the same animal, except that the chicken is tamed.

Almost all farmers raise some chickens, and some raise nothing else. Imagine the noise they must make, all cackling at once. Chickens provide delicious meat and billions of eggs every year.

Chickens cannot fly very high or very far. They eat by picking at their feed with their strong beaks. The little chickens are very lively.

Most American chickens are kept for both meat and eggs. The best egg chicken is the Leghorn, from Italy. If its ear lobes are white, the eggs it lays will be white too, but if the ear lobes are red, it will lay brown eggs. Both are good to eat.

A chicken farm is very noisy, because chickens make all kinds of sounds. The happiest sound comes after an egg has been laid, and the loudest comes early.
in the morning, when the roosters (male chickens) wake everyone with their crowing.

(After Amoroso and Bowles, 1971)
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>45.</td>
<td>My labor union</td>
</tr>
<tr>
<td>46.</td>
<td>My labor union president</td>
</tr>
<tr>
<td>47.</td>
<td>My lazy union president</td>
</tr>
<tr>
<td>48.</td>
<td>My labor union president election</td>
</tr>
<tr>
<td>49.</td>
<td>My community center building council</td>
</tr>
<tr>
<td>50.</td>
<td>My morning computer course</td>
</tr>
<tr>
<td>51.</td>
<td>My (light [yellow bus])</td>
</tr>
<tr>
<td>52.</td>
<td>My ([light yellow] bus)</td>
</tr>
<tr>
<td>53.</td>
<td>My father's mother's sister's dog</td>
</tr>
</tbody>
</table>
from the study of the first part can be generalized to sentences with various grammatical structures. The third part includes nine noun phrases composed of adjectives and compound nouns (the nine noun phrases are given in Table 2.3.) This part is designed for the study of the relationship between the schematized patterns and the detailed structure of the noun phrases.

In the recording sessions, three native speakers of American English were asked to read the three-part corpus. Each of the thirty sentences was written on an individual card, and each card was presented to the speaker after he had completed reading the preceding one. The fourteen sentences of the text were written on a single page.

The speech signals and the glottal signals (detected by using an accelerometer located on the trachea-notch of each speaker) were recorded on two-channel tapes. The third part of the corpus (composed of the nine noun phrases) was read by four speakers, and during the sessions, only the speech signals were recorded.

For calculating the $F_0$ contours of the sentences, a detection program based on an absolute difference sum algorithm (ADSA) was used (Meo and Gignini, 1971; Shaffer, Ross and Cohen, 1973). This technique is quite similar to an autocorrelation method (Sugimoto and Hashimoto, 1962;
Cheng, 1975) ADSA uses subtraction instead of the multiplication used in the autocorrelation method. Further, ADSA determines a fundamental period of speech signals as the reciprocal of the delay time at which the value of the absolute difference sum indicates the minimum (instead of the maximum in the autocorrelation method). The program thus can skip the computation whenever the value of the absolute difference sum exceeds a certain threshold value. Because of these properties, ADSA is particularly suitable for implementation in a $F_0$ detection program which runs on a small computer, since integer arithmetic can be used without fear of overflow. The $F_0$ detection program works sufficiently well for our purpose, both on the glottal signals and the lowpassed speech signals.

The amplitude envelopes are calculated by taking the absolute sum of successive segments (typically, 10msec of the duration) of the speech signals (with a bandwidth of 4.8 kHz). Hard copies of the $F_0$ contours displayed in parallel with the amplitude envelopes have been used for the schematic analysis of the $F_0$ contours.
2.3 Attributes

The fact that many factors in addition to intonation are involved in the specification of the $F_0$ contours makes it necessary to use visual inspection for determining the attributes. However, during the subjective analysis, we often face the problem of deciding whether or not a $F_0$ movement should be characterized by one of the attributes. In order to improve consistency in the subjective judgements, it is appealing to postulate a certain structure for the schematized patterns and physical properties which the schematized patterns must satisfy. The attributes correspond to the elements that compose such schematized patterns. By taking this approach, we are able to reduce freedom in the subjective analysis, since the original $F_0$ contours must be matched with the schematized patterns, which are constrained in their structure and must have certain physical properties. The comparison of a original $F_0$ contour and the corresponding schematized pattern will provide some insight into errors in the analysis. If a set of $F_0$ movements is distinctively different from any element of the schematized patterns, then we may introduce another attribute to describe them. In this study, however, we use only the five attributes already mentioned. We shall propose first the framework of the schematized patterns, and then describe the physical properties of
the attributes BL (baseline), R (rise), P (peak) and L (lowering). These properties were obtained by measurements of typical \( F_0 \) movements corresponding to the attributes. Finally, we will describe the relationship between the basic attributes and lexical stresses.

2.3.1 Structure of the Schematized Patterns

In a number of previous studies, \( F_0 \) contours have been decomposed into two major components, a gradual \( F_0 \) fall along the entire sentence, and localized movements (such as a rapid \( F_0 \) rise and a lowering). The gradual fall can be regarded as a reference line relative to which the localized \( F_0 \) movements occur. In their study of Dutch intonation, Cohen and t'Hart (1967) have noted that an averaging procedure for the localized \( F_0 \) movements cannot be established in a meaningful manner without taking into account this gradual falling. In generative models of \( F_0 \) contours for sentences for Swedish (Ohman, 1965, 1967; Carlson and Granstrom, 1973) and for Japanese (Fujisaki and Sudo, 1971), the two components are combined to derive the final \( F_0 \) contour. Vaissière (1971) also analyzed \( F_0 \) contours of French into two components of this type. Bolinger (1958) has predicted the gradual falling to be universal. We postulate a model of the two \( F_0 \) components as the basic structure of the schematic pattern.
The non-localized $F_0$ component, i.e. the gradual fall, is characterized by the attribute baseline (BL). The term 'baseline' was chosen to emphasize its function as the reference (or base) to the localized movements (Cohen and t'Hart, 1967, introduced the term "declination line"). Since the data show that the gradual falling is approximately linear, we use a straight line for its representation, as shown in Fig. 2.2 (a). The baseline will be indicated by a dashed line in the schematic representation of $F_0$ contours. In a representation of contour in terms of a sequence of attributes, the symbol BL is placed where the baseline begins to fall as shown in Fig. 2.2 (d). Each of the other four attributes corresponds to a particular localized $F_0$ movement (Maeda, 1974).

The two basic localized $F_0$ movements, i.e. the rapid rise and lowering, can be approximated by upward and downward straight lines, manifesting the attributes R and L, respectively. In the analyzed utterances, these two attributes often appear successively, R followed by L, generating a trapezoidal shape in the corresponding $F_0$ contour (so-called "hat-pattern") as represented in Fig. 2.2 (b). We assume that the height of the plateau of the trapezoidal pattern is constant, but when the pattern is combined with the baseline, the plateau becomes parallel to the baseline. (see Fig. 2.2 (c)).
An idealized description of the schematized patterns. The schematized pattern in (c) is the addition of the non-localized $F_0$ movement (i.e. baseline) in (a) and the localized $F_0$ movements (combination of rises and lowerings) in (b). (d) represents the corresponding sequence of the attributes.
Fig. 2.2
Instead of using schematized \( F_0 \) patterns, intonation can also be represented by a sequence of symbols representing the attributes, each successive symbol being associated with the time value specifying the point where the corresponding \( F_0 \) movement occurs during the utterance. A schematic pattern and the corresponding sequence of attributes are shown in Fig. 2.2 (c) and (d), respectively. Such a sequence of attributes (i.e. an attribute pattern) can be recognized as a discrete representation of intonation.

In addition to the two basic attributes, \( R \) and \( L \), we introduce two more attributes: peak (P) and rise on the plateau (RI). In the actual \( F_0 \) contours, the \( F_0 \) rise is often associated with a \( F_0 \) peak, in which the height is considerably greater than that of the plateau. Occasionally, the plateau portion is not flat, but contains a rise. This rise on the plateau is distinctively less steep than the rise \( R \), and it is more than one syllable in duration. An example of rise on the plateau is shown at the beginning of the sentence represented in Fig. 2.3. In this figure, the \( F_0 \) contour and the basic schematic pattern are superimposed. The amplitude envelope displayed next to the \( F_0 \) contour is useful identifying phonetic units such as syllables and words. Since we found that both the peak and the rise on the plateau seem to be linguistically significant, we
Figure 2.3

An example of the $F_o$ contour for a sentence read by KS, and the corresponding schematized pattern and the attribute sequence. A.E. represents the amplitude envelope.
The small black cat on the tree likes the dog.

Fig. 2.3
assign the attributes P and R1 (In chapter four, we will discuss the perceptual relevance of the attribute P.) However, we do not intend to schematize the F₀ contours corresponding to these two attributes with a specific form; we only mark the peak and the rising contour on the schematized patterns by using the symbols "P" and "R1", respectively, as shown in Fig. 2.3. The attribute pattern is assigned to the string of words in the sentence as shown at the bottom of Fig. 2.3. Note that the rise R and the peak P are located at the same position in the sentence, on the word "small". We have also found examples in which the F₀ peak occurs in the middle of the plateau or in front of the lowering. Hence, we recognize the two attributes P and R as independent attributes.

Although we regard the trapezoidal pattern as the most common and basic pattern in American English, we often find that the pattern is not fully realized in specific phonetic environments as can be seen in Fig. 2.3, on the word "tree". (This sentence was analyzed in the preliminary study, and it is not listed in Table 2.1.) In this example, the rise occurs during a voiceless consonant. Less frequently in our analysis, the lowering part of the trapezoidal pattern is not found. Also for one of the three speakers (JP), the rise R and/or the lowering L can sometimes be missing, since F₀
changes sometime occur during the initial and the final consonant of the word, to which the trapezoidal pattern is assigned. For such cases, we assume that the speakers raise (or lower) $F_0$ during the unvoiced portions, and hence the $F_0$ movements are not manifested directly in the contour. We assign the attributes R and L within parentheses ('(R)' and '(L)') when the $F_0$ movements are judged to occur during unvoiced portions of speech. Thus, the discrete representation contains sufficient information to specify the corresponding schematic patterns.

2.3.2 Baseline (BL)

In order to make the gradual fall more visible, we classify the thirty sentences listed in Table 2.1 into five or six groups for each speaker, depending on their length. The $F_0$ contours of the sentences belonging to the same group are then superimposed, by lining them up at the onset of each sentence. We show such superpositions of $F_0$ contours of the sentences spoken by one of the three speakers, KS, in Fig. 2.4, from (a) to (e). The sentence numbers indicated on the top of each figure refer to the numbers used in Table 2.1.

It is evident that the superposition of the $F_0$ contours in each group indicates a zone in which $F_0$ values gradually fall along the sentences. This phenomenon is particularly
Figure 2.4

Superimposed F₀ contours of the 30 sentences (listed in Table 2.1), read by speaker KS. The contours are classified into five groups depending on their length. The straight line in each figure from (a) to (e), represents the gradual fall of the F₀ contours, which is determined visually.
Fig. 2.4 (a) and (b)
Group III
KS(S3, S4, S6, S7, S8, S9, S16, S19, S20, S22, S23)

Fig. 2.4 (c)
Group IV $KS(S10, S11, S24, S25, S27, S29)$

$F_0$ (Hz)

TIME (sec)

Fig. 2.4 (d)
Group V  KS(S26, S28)

Fig. 2.4 (e)
clear in Fig. 2.4 (c). The falling rate at the upper edge of the zone seems to be greater than that of the lower edge, reflecting the fact that the localized movements, in particular, peak P, decrease in magnitude from the beginning to the end of the sentences. Since we assumed that any contour is constructed by the simple addition of the baseline and the localized $F_0$ movements, the falling rate of the lower edge of each zone must correspond to that of the baseline. The straight lines in Fig. 2.4 were drawn subjectively, by visual inspection of all the superimposed contours, such that each line represents the falling of the lower edge of the corresponding zone. It should be noticed that we do not take into account the $F_0$ values at the onset of the sentences, which are usually far below the straight lines. A study of the electromyographic activities of the laryngeal muscles (reported in Chapter 3) suggests that the low $F_0$ values at the onset of the sentences are due to an active lowering. Consequently, we consider these low values to be the result of a localized movement and we do not take them into account in the decision concerning location of the baseline.

In Table 2.4 (c), we show the average duration ($\bar{\tau}$) and the falling rate ($r$) ($\bar{\tau}$ and $r$ are obtained by measurements) and the magnitude of the falling ($\Delta F$), which is calculated by using the equation $\Delta F = r \cdot \bar{\tau}$, for each group. We applied
Table 2.4

Measurement of the falling rate (r) of the baseline for the 30 sentences listed in Table 2.1. The sentences are classified into 5 to 6 groups depending on the length. The data for the three speakers are given separately: KN in (a), JP in (b) and KS in (c).

$\bar{t}$: average duration of the sentences in each group (in sec)

r: falling rate of the baseline (in Hz/sec)

$\Delta F$: average magnitude of the fall calculated by using

$$F = r \cdot \bar{t}$$

$\overline{\Delta F}$: average of $\Delta F$'s over all groups

\(\sigma\): standard deviation of $\Delta F$'s

$$c.o.v. = \frac{\sigma}{\overline{\Delta F}}$$
<table>
<thead>
<tr>
<th>Group</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) KN</td>
<td>1.70</td>
<td>1.90</td>
<td>2.00</td>
<td>2.21</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td>12.4</td>
<td>10.9</td>
<td>9.00</td>
<td>8.55</td>
<td>10.4</td>
</tr>
<tr>
<td>ΔF</td>
<td>21.0</td>
<td>21.0</td>
<td>18.0</td>
<td>18.5</td>
<td>25.4</td>
</tr>
</tbody>
</table>

\[ \bar{ΔF} = 21.0 \text{ Hz, } \sigma = 2.4 \text{ Hz, c.o.v.} = 0.11 \]

<table>
<thead>
<tr>
<th>(b) JP</th>
<th>t 1.7</th>
<th>1.91</th>
<th>2.14</th>
<th>2.4</th>
<th>2.62</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>17.8</td>
<td>15.6</td>
<td>14.4</td>
<td>10.7</td>
<td>9.7</td>
</tr>
<tr>
<td>ΔF</td>
<td>30.4</td>
<td>29.9</td>
<td>30.8</td>
<td>25.7</td>
<td>25.4</td>
</tr>
</tbody>
</table>

\[ \bar{ΔF} = 28.4 \text{ Hz, } \sigma = 2.51 \text{ Hz, c.o.v.} = 0.09 \]

<table>
<thead>
<tr>
<th>(c) KS</th>
<th>t 1.74</th>
<th>1.86</th>
<th>2.1</th>
<th>2.44</th>
<th>2.82</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19.5</td>
<td>18.2</td>
<td>14.5</td>
<td>13.5</td>
<td>10.3</td>
</tr>
<tr>
<td>ΔF</td>
<td>33.2</td>
<td>34.2</td>
<td>32.3</td>
<td>32.3</td>
<td>28.8</td>
</tr>
</tbody>
</table>

\[ \bar{ΔF} = 32.2 \text{ Hz, } \sigma = 2.3 \text{ Hz, c.o.v.} = 0.07 \]
the same procedure for the remaining two speakers, with the results shown in Tables 2.4 (a) and (b), respectively. It must be noticed that the values of $\Delta F$ for each speaker is very close to each other, as indicated by the small values of the standard deviation ($\sigma$) as well as by the coefficient of variation (c.o.v), in Table 2.4. Therefore, as far as the thirty sentences are concerned, the value of $\Delta F$ may be regarded as constant in spite of the varying length of the sentences, for individual speakers. This property leads us to establish the equation that predicts the rate of the fall $r$, when the duration of the sentences, $t$ is given. $r = \frac{\Delta F}{t}$, where $\Delta F$ is the average value of $\Delta F$ for the individual speaker. The value of $\Delta F$ is listed in Table 2.4. The measured values of $r$ and the prediction are shown in Fig. 2.5, using dots and curves, for each speaker. Observe that the dots are distributed around the predicted curve.

The F$_0$ values at the offsets of the sentences are also approximately constant. This property is useful in the determination of the baseline. We have measured the F$_0$ values at the terminal point of the lowering contour located at the end of each sentence (all the sentences are declarative sentences). The definition of the terminal point is illustrated in Fig. 2.6. Two typical forms of the idealized lowering contour are found in the analysis of the thirty sentences,
Relationship between the falling rate \((r)\) of baseline and the duration \((t)\) for the thirty sentences (listed in Table 2.1), read by speakers KN in (a), JP in (b) and KS in (c). The dots represent the measured falling rate, and the curve in each figure is calculated by the equation \(r = \bar{F} \cdot t\), where \(\bar{F}\) is defined in Table 2.4.

Definition of the terminal point. The curves represent different types of the final \(F_0\) lowering in breath-groups.
Fig. 2.5

(a) KN
\[ \Delta F = 21.0 \text{ Hz} \]

(b) JP
\[ \Delta F = 28.4 \text{ Hz} \]

(c) KS
\[ \Delta F = 32.2 \text{ Hz} \]

DURATION OF SENTENCE (sec)
Fig. 2.6
as shown in Fig. 2.6 (a) and (b). In Fig. 2.6 (a), the terminal point is defined as the lower edge of the contour. In Fig. 2.6 (b), the terminal point (indicated by the arrow) is defined as the point where the first large lowering is terminated and where the contour starts to fall rapidly again. We consider the second fall at the very end of the sentence to be an artifact due to the termination of the utterances. The last contour shape, indicated in Fig. 2.6 (c), represents a falling contour terminated by a slight continuation rise. Only a few examples of such a contour have been found in the analysis of the corpus. For such sentences, the terminal point is defined as the lowest point of the contour.

The average $F_0$ value, the standard deviation, and the range of the measured $F_0$ values at the terminal point for the thirty sentences are shown in Table 2.5. The quite small value of c.o.v. and the range in $F_0$ values indicate small deviation in the values of the terminal point. This invariance is presumably due to the minimum influence of the localized $F_0$ movements at the offset of the sentence. Therefore, it is reasonable to draw the baseline in such manner that the magnitude of its falling is fixed at $\overline{AF}$, regardless of the length of the sentence; the baseline also intersects with the terminal point. Essentially, any point on $F_0$ contours in which the influence of localized $F_0$ movements is minimum
<table>
<thead>
<tr>
<th>Speaker</th>
<th>$\bar{F}$ (Hz)</th>
<th>$\sigma$(Hz)</th>
<th>c.o.v.=$\sigma/$</th>
<th>range (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KN</td>
<td>83.4</td>
<td>2.8</td>
<td>0.03</td>
<td>79.0~92.1</td>
</tr>
<tr>
<td>JP</td>
<td>81.2</td>
<td>1.7</td>
<td>0.02</td>
<td>77.8~85.6</td>
</tr>
<tr>
<td>KS</td>
<td>78.5</td>
<td>1.9</td>
<td>0.02</td>
<td>74.6~82.0</td>
</tr>
</tbody>
</table>

**Table 2.5**

Average ($\bar{F}$), standard deviation ($\sigma$), coefficient of variation (c.o.v.) and the range of measured $F_o$ values at terminal points (see text and Fig. 2.6 for its definition). The data are given for the 30 sentences listed in Table 2.1, read by the speakers KN, JP and KS.
can be used as the point where the baseline must occur. However, the terminal point seems to be determined most easily and consistently by visual inspection. The baseline shown in Fig. 2.3 (also in Fig. 2.1) was obtained by imposing the above two conditions. It is clear that the \( F_0 \) contour can be approximated reasonably well by a pattern which is the addition of the baseline and schematized local \( F_0 \) movements. We will describe in the following section how the localized \( F_0 \) movements are schematized. It should be noted that the final \( F_0 \) value of the baseline is adjusted to each sentence.

In the case of the isolated sentences, the speakers took a breath after each sentence had been uttered. Consequently, each sentence can be regarded as a breath group. Since the gradual fall of \( F_0 \) could be related to the activities of the respiratory system during speech, we speculate that the baseline generally characterizes the gradual fall inside a breath group rather than inside a sentence. It is a common phenomenon that a long sentence, or a sentence containing major grammatical breaks, is divided into a number of breath groups. We therefore predict that for such sentences the non-localized component of the \( F_0 \) contour is characterized by a certain number of baselines, corresponding to the division of the sentences into breath groups.
To examine the above prediction, we have investigated the fourteen sentences in the text in terms of the baseline. It was found that the average $F_0$ values of the terminal points for the isolated sentences was an excellent cue to decide the termination of a breath group. When the lowering contour in a sentence reaches near that value, for instance 78.5 Hz for speaker KS, and is followed by a relatively long pause, say 300 msec, the sentence can be considered to be divided at that point. By using this method, the text is successfully divided into breath groups, and we postulate that the $F_0$ contours within each breath group are schematized with reasonable accuracy by the additive scheme of the baseline and the localized $F_0$ movements. The breath groups of the sentence represented in Fig. 2.1 (spoken by the three speakers) were determined by using this procedure. Note that the speakers JP and KS divide the sentence (up to the word "colors") into two breath groups, while KN uses a single breath group. Discrepancies between the particular manner of dividing the sentences into breath groups, for individual speakers, were found throughout in the text. Any boundary between two successive breath groups always corresponds to either a sentence boundary or to a major grammatical break; however, it is not necessary for all grammatical breaks to be marked by a breath group boundary. Rather, a major grammatical break
can be considered only as a potential breath group boundary, and whether it will correspond effectively to a breath group juncture depends on the particular speaker. We further speculate that even the same speaker may divide the sentences differently in different readings of the sentences.

To determine the baseline for each breath group, we need to know either the falling rate or the magnitude of the fall. It was found that for the three speakers, the average magnitude of the fall, $\overline{AF}$, calculated from the isolated sentences, can be used also for the sentences consisting of one breath group in the text. However, for the sentences consisting of more than one breath group, the magnitude of the fall had to be reduced to 50% of $\overline{AF}$ to obtain a reasonable approximation of the $F_0$ contours. Such an example has already been shown in Fig. 2.1. The magnitude of the fall in each case is indicated by the percentage values of $\overline{AF}$ at the onset of each baseline.

The straight line approximation of the baseline which characterizes the gradual $F_0$ fall inside a breath group seems to work well for most of the cases. However, for a few sentences of the text spoken by one of the three speakers, the approximation is somewhat poor. Such an example is shown in Fig. 2.7 (a), where the dashed line represents a straight-line
Figure 2.7

$F_0$ contours and the corresponding baselines represented by the dashed lines. The two straight lines in (a) indicate a better approximation of the gradual $F_0$ fall. A.E. represents the amplitude envelope.
Imagine the noise they must make all cackling at once.

Fig. 2.7
approximation of the baseline. The corresponding $F_0$ contour indicates a rather rapid fall during the first 1.3 second, and then the falling rate becomes much smaller (near zero) as represented by the two connected straight lines. Cohen and t'Hart (1967) have noted a similar phenomenon in Dutch intonation. Since the same speaker also utters a longer phrase (longer than 1.3 sec) in which the $F_0$ contour indicates a straight fall, the turning point where the falling rate changes cannot be predicted simply from the duration of the sentence alone. In this particular example, it seems to be related to the particular structure of the sentence. It is clear that the $F_0$ contour of the same sentence, spoken by the speaker KS (shown in Fig. 2.7 [b]) indicates the division of the sentence into two breath groups at that point. Therefore, in the case of the speaker JP, it may be explained in terms of the failure of the recovery reset of the baseline to occur in the second breath-group.

It should be noticed that the peaks in the amplitude envelope in the second breath-group are relatively smaller than the peaks in the first breath-group, in Fig. 2.7 (a). On the other hand, in the case of a recovery of the baseline (see Fig. 2.7 [b]), a simultaneous recovery of the amplitude envelope can be observed. Generally, in the analyzed text, a recovery of the baseline is well correlated with a recovery
of the amplitude envelope (see Fig. 2.1 for other examples). The physiological mechanisms underlying these properties of the baseline will be discussed in some detail in Chapter 3.

A comment about the term "breath group" should be interjected at this point. The term "breath group", as employed here, is not necessarily related to an actual physiological gesture for taking breath. This gesture creates a relatively long pause (Fujisaki and Omura, 1971), which divides sentences into smaller groups of words. However, there is no sure way to tell when the speaker effectively takes a breath from an acoustic analysis, except in the case of very short isolated sentences. We rather define the breath group from a functional point of view, in specifying the baseline to a group of words such that the \( F_0 \) contour is well approximated by the addition of the baseline and the schematized local \( F_0 \) movements.

Since we have analyzed only a small set of material the results should not be considered as generally conclusive. However, the following two properties of the gradual \( F_0 \) fall that is characterized by the baseline are strongly suggestive. First, the \( F_0 \) value at the terminal point of any breath group can be considered roughly constant for an individual speaker, regardless of its length and its position in the sentence of type discussed here. Second, the magnitude of the gradual
fall is also roughly constant for an individual speaker, when the breath group corresponds to an entire sentence. However, when the breath group is located in non-initial position in the sentence, the magnitude is often reduced. In fact, the magnitude of the gradual fall can be as low as zero. There seems to be no simple way to predict the variations of the magnitude of the $F_0$ fall in a breath group in non-initial position. The division of the sentence into breath groups is closely related to the grammatical structure of the sentence. However, locations where the division occurs vary from one speaker to another, and perhaps even from time to time, for an individual speaker.

2.3.3 Rise(R), Peak(P), and Lowering(L).

Before studying the physical properties of the localized $F_0$ movements characterized by the attributes R (rise), P (peak) and L (lowering), we must specify a criterion to determine whether or not a particular $F_0$ movement corresponds to one of these attributes. Observation of the $F_0$ contours leads us to impose the following two conditions that must be satisfied by $F_0$ movements to be considered as a manifestation of one of the attributes. First, the $F_0$ localized movements, rise R (with or without a peak) and lowering L, must be rapid, movements, such that the movements are completed within at most two syllables. Second, the $F_0$ movements must be associated
with relatively high speech intensity during that portion. Since such $F_0$ movements must code a linguistic message, we speculate that the $F_0$ movements are controlled actively by the speaker, and should be carried with relatively high intensity. As an example, Fig. 2.8 (a) represents the $F_0$ contour and the amplitude envelope corresponding to the beginning of sentence S-1: "The dog...". One can observe a rapid fall to the baseline after a rapid rise during the vowel /ɔ/ in the word "dog", the combination of the rise and the successive fall forming a peak. However, we do not consider this fall as the attribute $L$, since the corresponding amplitude envelope also indicates a rapid fall. The point may become clearer, when this is compared with the $F_0$ contour shown in Fig. 2.8 (b), representing the $F_0$ contour for the end of the sentence S-5, "... in the mud". Observe that the amplitude envelope indicates a peak during the $F_0$ lowering immediately after the $F_0$ rise. We thus recognize the falling $F_0$ contour as a manifestation of the attribute $L$. According to the convention described in section 2.3.1, we represent the $F_0$ contour in Fig. 2.1 (a) as $"R(L)"$ while the $F_0$ contour in Fig. 2.8 (b) is indicated as $"R L\"$.

In this connection, a factor which must be taken into account for the determination of the three attributes $P$, $R$, and $L$ is the influence of consonants on the $F_0$ contour. When the
location of voiced consonants, in particular voiced fricatives and voiced stops, corresponds to the plateau of the schematized pattern, a large dip is observed in the $F_0$ contour of that portion. Such examples were already shown in Fig. 2.1 and explained briefly in section 2.1. The rapid fall represented in Fig. 2.8 (a) may be, at least partially, due to the effect of the voiced stop /g/ in the word "dog". This phenomenon is caused presumably by a momentary decrease of the transglottal pressure due to a strong constriction in the vocal tract during the consonant articulation, and possibly also by an increase in vocal-fold slackness for the voiced consonant. We therefore recognize that the dip in the $F_0$ contours is a non-controlled factor in terms of intonation, and does not count as a manifestation of the attributes. This consideration account for at least part for the second condition imposed previously, since we usually observe a dip in the $F_0$ contour, as well as in the amplitude envelope in such circumstances.

We describe now the physical properties of the three attributes. To obtain a perspective for the properties, we have superimposed the portion of the $F_0$ movements corresponding to the attributes for a number of utterances. In Fig. 2.9, we represent such superpositions of the $F_0$ movements, sampled from the $F_0$ contours of the isolated sentences, read
Figure 2.8

$F_0$ contours and the corresponding amplitude envelope (A.E.) The vertical dashed lines in each figure indicate the time when the $F_0$ peak occurs.
(a) KS S14

The dog likes...

(b) KS S5

... in the mud.

Fig 2.8
Figure 2.9

Superposition of the localized \( F_0 \) movements. The \( F_0 \) contours are superimposed upon each other, in reference with the points where each movement starts.
Fig. 2.9

Rise at onset of sentence

Rise

Lowering

Lowering at offset of sentence

KS (S1~S30)
by KS. The $F_0$ movements are categorized into four groups, rise at the initial position in the sentences, rise and lowering in the middle, and the lowering at final position. Since the peak $P$ occurs with the rise, we do not separate the movement into the two components, $R$ and $P$. The superposition is made by fitting every onset point of the rising contours, where the contour begins to rise rapidly, and by fitting every onset of the lowering contours, where the rapid fall starts.

It is evident that the magnitudes of the rising contours, in particular in medial position, vary continuously, and the range of variation is rather large relative to the variations in the lowering contours. This variability is probably due to the fact that we did not separate the simple rise from the rise with peak, and to the fact that this particular speaker reduces the peak height consistently from the beginning to the end of the sentence in non-emphatic mode. There seems to be no reasonable way to separate the two kinds of rise (rise with or without peak) dichotomously. On the other hand, the duration of any rise or lowering seems to be relatively constant, in contrast to greater variation in the $F_0$ magnitude.

To make the above observation more specific, we have measured the duration and the magnitude of each rise and lowering in which the $F_0$ contour of the portion is well defined in the sense that we can mark the points where the rise or the lowering begins and where it ends. For instance, we
define the end of the rise as the maximum point when the rise is associated with a peak, and as the point where the $F_0$ contour reaches the plateau in the case of a simple rise. Since the decisions concerning the onsets and offsets of the movements are subjective, an error of $\pm 10$ msec may be involved in the measurements of the durations. The effect of the baseline contained in the actual magnitude value of the $F_0$ rise and lowering is subtracted from each measurement.

The results of the measurements are listed in Table 2.6, in terms of the average values ($t$ and $F$), standard deviation ($\sigma_t$ and $\sigma_F$), the coefficient of variation (c.o.v.'s), of the durations and of the $F_0$ magnitudes, for the rise in the initial position in the sentences, the rise and lowering in final position, and so on. In Fig. 2.10 from (a) to (c), the measurements of the individual rises and lowerings are plotted for each of the three speakers. The horizontal axis represents the normalized durations and the vertical axis the normalized $F_0$ magnitude. For normalization, the measured durations and the magnitudes of the rising and lowering contours are divided by the average values which are listed in column (iii) for the rise and in column (vi) for the lowering, in Table 2.6. The $F_0$ movements are categorized either as rises (with or without peak) or lowerings in each part of the figure.
Table 2.6

The averaged magnitudes and durations of the localized \( F_0 \) movements, rises and lowerings, for the thirty sentences listed in Table 2.1, read by the three speakers: DK in (a), JP in (b) and KS in (c). Column (i) corresponds to the rises, \( R \), at the initial position in the sentences (s.i.), (ii) to the rises, \( R \), inside the sentences, (iii) to all rises, (iv) to lowerings, \( L \), at the sentence's final position (s.f.), (vi) to all lowerings, \( L \), and (vii) to all rises and lowerings, respectively.

Note: \( N \) = number of samples,
\[ \bar{A} \] = averaged durations,
\( \text{c.o.v.} \) = the corresponding coefficient of variation,
\[ \bar{F} \] = averaged magnitudes of the \( F_0 \) movements,
\[ \sigma_F \] = standard deviation of magnitudes of the \( F_0 \) movements.
### Table 2.6

(a) \( \text{KN} \)

<table>
<thead>
<tr>
<th></th>
<th>(i) ( R )</th>
<th>(ii) ( R )</th>
<th>(iii) Both</th>
<th>(iv) L</th>
<th>(v) L</th>
<th>(vi) Both</th>
<th>(vii) Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>14</td>
<td>13</td>
<td>28</td>
<td>6</td>
<td>16</td>
<td>22</td>
<td>50</td>
</tr>
<tr>
<td>( \bar{t} (\text{ms}) )</td>
<td>131</td>
<td>129</td>
<td>130</td>
<td>143</td>
<td>147</td>
<td>145</td>
<td>15</td>
</tr>
<tr>
<td>( \sigma_t (\text{ms}) )</td>
<td>26</td>
<td>16</td>
<td>22</td>
<td>31</td>
<td>25</td>
<td>29</td>
<td>25</td>
</tr>
<tr>
<td>( \text{c.o.v.} )</td>
<td>0.20</td>
<td>0.13</td>
<td>0.17</td>
<td>0.22</td>
<td>0.17</td>
<td>0.20</td>
<td>0.18</td>
</tr>
<tr>
<td>( \bar{F} (\text{Hz}) )</td>
<td>15</td>
<td>12</td>
<td>14</td>
<td>12</td>
<td>15</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>( \sigma_F (\text{Hz}) )</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>( \text{c.o.v.} )</td>
<td>0.33</td>
<td>0.18</td>
<td>0.30</td>
<td>0.34</td>
<td>0.27</td>
<td>0.28</td>
<td>0.29</td>
</tr>
</tbody>
</table>

(b) \( \text{JP} \)

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>9</th>
<th>13</th>
<th>13</th>
<th>12</th>
<th>25</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \bar{t} (\text{ms}) )</td>
<td>102</td>
<td>117</td>
<td>110</td>
<td>109</td>
<td>119</td>
<td>114</td>
<td>112</td>
</tr>
<tr>
<td>( \sigma_t (\text{ms}) )</td>
<td>21</td>
<td>18</td>
<td>20</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>( \text{c.o.v.} )</td>
<td>0.21</td>
<td>0.15</td>
<td>0.16</td>
<td>0.15</td>
<td>0.14</td>
<td>0.14</td>
<td>0.16</td>
</tr>
<tr>
<td>( \bar{F} (\text{Hz}) )</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>11</td>
<td>13</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>( \sigma_F (\text{Hz}) )</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>( \text{c.o.v.} )</td>
<td>0.30</td>
<td>0.48</td>
<td>0.40</td>
<td>0.38</td>
<td>0.22</td>
<td>0.30</td>
<td>0.32</td>
</tr>
</tbody>
</table>

(c) \( \text{KS} \)

<table>
<thead>
<tr>
<th></th>
<th>14</th>
<th>23</th>
<th>37</th>
<th>13</th>
<th>21</th>
<th>34</th>
<th>71</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \bar{t} (\text{ms}) )</td>
<td>127</td>
<td>116</td>
<td>121</td>
<td>120</td>
<td>117</td>
<td>119</td>
<td>120</td>
</tr>
<tr>
<td>( \sigma_t (\text{ms}) )</td>
<td>18</td>
<td>15</td>
<td>17</td>
<td>12</td>
<td>19</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>( \text{c.o.v.} )</td>
<td>0.14</td>
<td>0.13</td>
<td>0.14</td>
<td>0.10</td>
<td>0.16</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>( \bar{F} (\text{Hz}) )</td>
<td>38</td>
<td>20</td>
<td>29</td>
<td>16</td>
<td>17</td>
<td>17</td>
<td>23</td>
</tr>
<tr>
<td>( \sigma_F (\text{Hz}) )</td>
<td>15</td>
<td>7</td>
<td>12</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>( \text{c.o.v.} )</td>
<td>0.40</td>
<td>0.34</td>
<td>0.41</td>
<td>0.31</td>
<td>0.18</td>
<td>0.23</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Figure 2.10

Relationship between the duration and the amplitude of the $F_0$ movements, sampled from the $F_0$ contours of the thirty sentences (listed in Table 2.1), for three speakers, KN in (a), JP in (b) and KS in (c). The duration and the magnitude are normalized by dividing the measured values by the average values in the column (iii) and (iv) in Table 2.6.
Fig. 2.10 (a)
Fig. 2.10 (b)
Fig. 2.10 (c)
From Fig. 2.10 and Table 2.6, the following three properties come to light. Although a considerable scattering of the dots can be observed in Fig. 2.10, the two speakers KN (in Fig. 2.10 (a)) and JP (in Fig. 2.10 (b)) show a negative correlation between the duration and the magnitude of the rises, while all speakers show a positive correlation for the lowering. For all the speakers, the variation of the $F_0$ magnitude is always greater than that of the duration. This fact is also indicated by the c.o.v.'s. in Table 2.6. The values of the c.o.v.'s. for the $F_0$ magnitude are roughly two times greater than those for the durations. Second, by the difference of the average durations between rise and lowering (shown in columns (iii) and (vi) in Table 2.6) is smaller than their standard deviations. The average durations for the two localized $F_0$ movements, therefore, are not significantly different from each other. Third, the c.o.v. for the lowering is considerably smaller than that for rise for all the speakers.

Taking into account the above properties, we may impose certain constraints on the schematized patterns. As a zero-order approximation, we can regard the durations for both the rise (with or without a peak) and the lowering as constant and the same. It would be quite natural to use the total average value shown in column (viii) of Table 2.6, as
the constant duration for an individual speaker.

In order to complete the schematized pattern, we need to calculate the height of the plateau. Since we have postulated the plateau to be parallel to the baseline, the magnitude of the rise and the lowering in the schematized pattern must be the same. It seems reasonable to use the average magnitude of the lowering (indicated in column (vi) of Table 2.6) as the height of the plateau, since the c.o.v. of the lowering is considerably smaller than that of the rise. Whenever the magnitude of a rapid localized \( F_0 \) movement is markedly higher than the height of the plateau, we recognize the movement as a manifestation of the attribute P, regardless of its location in the plateau.

Although we do not know why the negative and the positive correlations between the magnitudes and the duration exist in the lowering and the rise, respectively, it seems to be suggestive that two different mechanisms are involved in the control of the rise and the lowering, and that the lowering is not simply due to the relaxation of the muscles which were involved in the rise of the \( F_0 \) values. The fact that the average durations of the rise and the lowering are similar gives further support to this supposition. We shall discuss this problem again in Chapter 3.
2.3.4 Effect of the Consonants on the \( F_0 \) contour

A remark should be made concerning the effect of an initial consonant upon the \( F_0 \) contour of the following vowel, in a stressed syllable. After an unvoiced consonant or a voiced stop, we often observe only a lowering contour (which corresponds to the attribute L), although this effect does not always occur. Compare the \( F_0 \) contours shown in Fig. 2.11 (a) and (b), which represent two distinctively different \( F_0 \) movements in an identical context. In Fig. 2.11 (a), only the lowering contour corresponding to the word "cat" can be seen, while in Fig. 2.11 (b), we observe only the rising contour. The discrete representation of the two contours therefore should be "(R)L" for the first case, and "R(L)" for the second case. Figure 2.11 (c) and Fig. 2.8 (a) illustrate a similar pair for a voiced stop, in the word "dog". In our data, when the initial consonant in a stressed syllable is unvoiced, and in particular when it is an unvoiced stop, the \( F_0 \) contour during the following vowel indicates a lowering (as shown in Fig. 2.11 (a)), much more often than a rising (as shown in Fig. 2.11 (b)). A high or a low \( F_0 \) value at the onset of the vowel seems to constitute a secondary perceptual cue for voiceless-voiced distinctions of stop consonants in the initial position (Fujimura, 1961). There must be an intrinsic reason for such a phenomenon.
Figure 2.11

The $F_0$ contours and the amplitude envelope (A.E.)
A.E.

F₀(Hz)

140

120

100

.2 s

(R)L

The cat likes....

KS S8

(a)

KS S1

(b)

KS S7

(c)

Fig. 2.11
We speculate that the laryngeal gesture for the rise may be incorporated with the gesture for the unvoiced consonant. In other words, the gesture for the unvoiced consonant probably corresponds to that of the rise, and consequently, only the lowering is manifested on the $F_0$ contour during the following portion. Halle and Stevens (1971) and Stevens (1975) postulate an increased vocal-fold stiffness for unvoiced stops in comparison with the voiced stops, in their scheme of laryngeal features for the distinction of the consonants: an increased stiffness is presumably also the gesture that is utilized to actualize a rise in the $F_0$ contour.

Although the speaker can raise or lower the $F_0$ contour of the vowel portion regardless of the identity of the preceding consonants as described above, there seems to exist an interaction between the consonant gesture and the $F_0$ control. Such interaction is not unexpected, if one takes account of the fact that both $F_0$ control and consonantal articulation involve adjustment of the state of the larynx. Fig. 2.12 (a) and (b) show spectograms of the words "the cat", corresponding to the $F_0$ contours shown in Fig. 2.11 (a) and (b), respectively. The unvoiced stop /k/, followed by the rising $F_0$ contour (attribute R) shown in Fig. 2.12 (b) is heavily aspirated, while the same consonant followed by the lowering $F_0$ contour (attribute L), shown in Fig. 2.12 (a) is less
Sound spectograms of the beginning of the sentences S8 (in [a]) and S1 (in [b]). The spectograms in (a) and (b) correspond to the $F_0$ contours in Fig. 2.11 (a) and (b), respectively.
aspirated, but it is associated with a relatively strong burst (The burst can be seen as a small peak at the /k/ release in the corresponding amplitude envelope shown in Fig. 2.11 (a)). Since our spectograph (Voiceprint Model 4691A) is equipped with an automatic gain control, energy in the two different spectra cannot be compared using the gray scale.

This phenomenon may be explained in terms of the laryngeal feature scheme (Stevens, 1975) as follows: When the consonant is followed by the attribute $R$, the state of the vocal-fold is presumably non-stiff (since the $F_0$ values at the onset of the following vowel must start low). Thus, in order to prevent vocal cord oscillation, the glottis must be widely spread. This spread glottis perhaps causes the heavy aspiration. On the other hand, when an unvoiced stop is followed by the attribute $L$, the state of the vocal folds must be stiff, because the $F_0$ values must be relatively high at the onset of the following vowel. The condition of stiff vocal folds probably allows the adjustment of the glottal width to be less spread, resulting in less aspiration, than in the case of non-stiff vocal folds and widely spread glottis.

It should be noted that the (less-stiff, widely spread) versus (more stiff, less spread) contrast described above is only relevant to unvoiced consonant pairs such as that as described above. For instance, non-stiff vocal folds in the
unvoiced stops are probably more stiff than in the voiced consonants, assuming all other things are equal. The rising contour after /k/ in Fig. 2.11 (b) starts from far above the baseline (the baseline is determined as described in the previous section), while after /d/ in Fig. 2.8 (a), it begins at about the level of the baseline, indicating a less stiff vocal fold state during the voiced consonant. Similarly, when the stop is followed by the attribute L, the vocal folds are perhaps more stiff for the unvoiced stop than for the voiced stop. A comparative example may be found in Fig. 2.11 (a), where the unvoiced consonant /k/ is followed by L, and in Fig. 2.11 (c), where /d/ is followed by L, although this comparison is less valid, due to the fact that the contexts are different in two cases. However, it seems to be a general phenomenon that the $F_0$ value at the onset of the vowel following an unvoiced stop is higher than that of the vowel after a voiced stop. (This finding has been extensively investigated by Lea (1973)). Therefore, we speculate that the vocal folds are more stiff during an unvoiced stop than during a voiced stop. We shall discuss this problem further in Chapter 3, in terms of the measurement of the activities of the laryngeal muscles.
2.3.5 Rise on the Plateau (RL)

We do not have many examples of the rise on the plateau, which is characterized by the attribute RL, since RL occurs only in phrases composed of no less than three lexical words. Further, since the $F_0$ contours of only one of the speakers, KS, show the rise RL frequently, we shall only describe actual examples in which that kind of rise occurs.

In Fig. 2.13 (a), two typical $F_0$ contours with and without attribute RL are superimposed on each other, together with the schematic pattern. (The $F_0$ contours correspond to the beginning of the sentences S-11 and S-25, respectively.) The $F_0$ contour with RL, represented by the solid line indicates a rising contour during the second word in the noun phrase, the word "white". On the other hand, the contour without RL, shown by the dotted line, is located near the plateau in the portion where RL occurs in the first example. The rate of rising in RL is much less large than that of attribute R. This rising rate could vary over a continuous range, just as the magnitude of the peak P can have a continuous range of values as described before. Observe Fig. 2.13 (b), which represents the $F_0$ contours of the noun phrase "the small black cat" spoken in three different contextual environments. The direction of each of the three $F_0$ contours on the plateau is represented by the arrows: the direction
Figure 2.13

$F_0$ contours for the noun phrases (NP) 'the big white dog' from S 11 and S 25 in (a), and the small blcak cat' in (b), in various contextual environments: (i) '(NP) likes the dogs', (ii) 'The big white dog likes (NP)', (iii) 'The dog likes (NP)'.

The big white dog .......

... the small black cat.

Fig. 2.13
of the arrows changes continuously from falling to rising. 

R1 is differentiated from the attributes R and P by the following two features: first, R1 is distinguished from R by the fact that it occurs on the plateau, like the attribute P; second, the $F_0$ movements corresponding to R1 can occur on an entire word, unlike P.

Although the existence of an attribute R1 is somewhat less evident than the existence of the other attributes, it will be shown later how the introduction of the attribute R1 fits into a theoretical framework dealing with the generation of the attribute patterns associated with phrases composed of more than two lexical words.

2.3.6 A Function of the Basic Attributes: Stress-marking

During the study of the physical properties of the localized $F_0$ movements, it has become apparent that the characteristic $F_0$ movements are highly correlated with lexical stress. Without exception, the $F_0$ rise characterized by the attribute R (with or without P) occurs during a stressed syllable in a word. All the $F_0$ contours shown in this thesis show such examples. The lowering can occur in two different places, either during a stressed syllable or during the immediately following non-stressed syllable, i.e., the attribute L is located either on the stressed syllable or on the following syllable. Therefore, if a word is monosyllabic, or if a
stressed syllable is located at the final position in the word, the lowering always occurs during the stressed syllable. As far as our data are concerned, when a word is polysyllabic, the F₀ lowering occurs much more often during the non-stressed syllable following the stressed syllable than during the stressed syllable itself.

In any case, the attribute R (with or without P) and the attribute L are so well related to lexical stress that it may be safe to state that an important function of the attribute assignment on a specific syllable is stress-marking. In fact, a number of studies (Bolinger, 1958; Morton and Jassem, 1965; Fromkin and Ohala, 1968; Ohala, 1970) have shown that both F₀ rise and F₀ lowering are involved in stress-marking.

In the above description, we did not specify whether the syllable is assigned primary or secondary stress in the word. In our observation, the rise occurs during the primary-stressed syllable. We expect, however, that the rise may occur during a syllable with secondary stress when this stressed syllable is located before the syllable with primary stress in that word. Actually, Vaissière (1976) claims that such examples are found in the F₀ analysis of English utterances. Similarly, although the attribute L is associated in most cases with the syllable with primary stress, L can occur
with the syllable having secondary stress. Such examples will be discussed later, in Section 2.4. In short, a 1-stressed syllable (that is a syllable with primary stress) has more chance to be assigned one (or both) of the two basic attributes, than a 2-stressed syllable (a syllable with secondary stress). We speculate that there is not only a difference in degree of stress between the two stresses, but also the 1-stressed syllable is more likely to receive the attributes (stress-marking) than the 2-stressed syllable. It should be noted that the stressed syllable is not necessarily always associated with a specific attribute. The number of attributes assigned to a lexical word seems to depend on several factors, such as its grammatical context, emphasis on that word or on a neighboring word, and on the speaker's habits. These factors make the prediction of the attribute pattern extremely difficult. In the following section, we shall investigate this problem in some depth.
2.4 The Attribute Patterns and Constituents of Sentences

The primary purpose of this section is to show that the attributes are used for sending information about certain constituents of sentences. We shall describe how the attribute patterns (which are sequences of the attributes) associated with sentences are generated using a set of rules. Application of the rules requires a specification of the grouping of the words in the sentence. There are two classes of the groupings: one corresponds to a chunking of the sentence into smaller units, and the other is related to the structure inside each unit. We shall call the first one 'grouping' and the second one 'subgrouping'.

At least three factors seem to be involved in the specification of the grouping and the subgrouping: the syntactic/semantic constituent structure of the sentence, a principle of physiological economy for manifesting the lexical stresses, and the emphasis of one or more word(s) in the sentence. Even for sentences spoken in the non-emphatic mode, it is difficult, if not impossible, to predict a unique attribute pattern. The phonetic manifestation of the syntactic/semantic constituent structure may be affected by the principle of economy in speech production process. It is, however, possible to generate any attribute pattern observed in our data, using a small number of rules.
It should be noted that we are not concerned with the attribute BL. The study of the baselines (described in Section 2.3.2) has indicated a grammatical function for BL: the marking of the onset of each breath-group. Since we shall investigate in this section isolated sentences each comprising exactly one breath-group, the symbol BL is always assigned at the onset of the sentence. We therefore study only the assignment of the four attributes, R, L, P, and R1, that characterize the localized $F_0$ movements in the sentences. However, it should be remembered that the successive groups to be described are always located inside a single breath-group.

2.4.1 Empirical Hypotheses Concerning Attribute Patterns

Before going into a detailed investigation of the observed attribute patterns, we present first the basic approach that we have taken for interpreting the attribute patterns associated with phrases and sentences.

The grouping of the words in sentences plays an important role in the generation of the patterns, as will be described below. To make the point clear, let us use the observed attribute pattern associated with the sentence shown in Fig. 2.3 in Section 2.3.1 as follows:

\[
P \quad \text{BL} \quad \text{R} \quad \text{R1} \quad \text{L} \quad (R)L \quad \text{R} \quad \text{L}
\]

(2.1) "The small black cat on the tree likes the dog."

(KS)
The above sentence with the attribute pattern may be represented symbolically by the following form:

$$P_{BL} R_{RL} R_{L} L_{RL} R_{L} L_{w} w_{w} w_{w} w_{w} w_{w} w_{w} w_{w} w_{w} \text{,}$$

where 'w' represents a word. When attributes are assigned to a word, the locations of the attributes are determined independently of the context, as described in Section 2.3.6. We therefore need only to specify the ordered attributes assigned to the individual word as represented in (2.2).

Observation of (2.2) suggests that the beginning of each group (the initial word in the group) is marked by R, and the end (the final word in the group) by L. If we group w's in (2.2) according to this suggestion, the following grouped series of w's is obtained.

$$w_{w} (w_{w} w_{w}) w_{w} w_{w} (w_{w}) (w_{w} w_{w})$$

Apparently, in this particular example, the groups correspond to syntactic (semantic) constituents of the sentence. It seems reasonable to establish the following two empirical hypotheses on the basis of an overall observation of the schematized $F_{O}$ patterns. First, the pairs of the attributes, R and L, reflect the underlying groupings of the words in the phrases and in sentences. In other words, the attribute patterns associated with any group must begin with R and
terminate with L. Secondly, R and L must occur alternately; R cannot follow immediately after a previous R. These two hypotheses significantly constrain the form of the grouping. For instance, a structure in which two groups of words are further grouped together cannot be described by using R and L. In short, the attributes R and L can signal only the non-embedded groups of words in a sentence.

The above consideration may lead us to postulate a superficial theory, for instance, 'put R on the initial word in the group and L on the final word'. However, examination of the assignment of R1 and P will suggest that the attribute pattern associated with a group of words depends on the internal structure of the group. Such structure is described in terms of the subgrouping of the words within the group.

A basic approach to be taken for generating the attribute patterns is as follows: we assume that initially any word is grouped with itself, thus indicating the basic ordered attributes, R and L. As shown later, some words receive more than one pair of attributes; but we assume here that only one pair is assigned. When two words are grouped, then the initial pattern (i.e. the sequence of the two pairs of attributes) is transformed to a new pattern, in accordance with the two empirical hypotheses for the groups composed of
two words. The symbolic description of the transformation will have the form shown by the following example:

\[(2.4) \begin{array}{c}
\text{RL} \\
(w) \\
\end{array} \begin{array}{c}
\text{RL} \\
(w) \\
\end{array} \rightarrow \begin{array}{c}
\text{R} \\
(w) \\
\end{array} \begin{array}{c}
\text{L} \\
(w) \\
\end{array},
\]

where the arrow "\(\rightarrow\)" must be read as 'the right item can be generated from the left items'. If the structure of the grouped words is described as multi-embedded parentheses (i.e. subgroupings), we apply rules from the innermost parentheses to the outer ones, erasing the paired parentheses after each application of a rule. Thus we are applying the principle of the transformational cycle proposed by Chomsky and Halle (1968), to the generating process of the attribute pattern within the group.

In the following sections, we shall describe in detail the generation of the attribute patterns associated with compound words, phrases and sentences.

2.4.2 The Attribute Patterns Associated With Noun Phrases and Compound Words Composed of Two Lexical Words

The \(F_o\) contours of simple noun phrases composed of a determiner, an adjective and a noun are shown in Fig. 2.14 (the noun phrases were read by three speakers). Besides the phonetic differences, these phrases differ also in the number of syllables and in the location of the lexical stress on each content word. It is clearly seen in the \(F_o\) contours that a rise in
Figure 2.14

The $F_0$ contours of noun phrases located at the ends of sentences, read by the three speakers, KN in (a), JP in (b) and KS in (c). The dashed line in each figure represents the baseline for the longest phrase, which ends with the word 'kangaroo'.
Fig. 2.14 (a)
Fig. 2.14 (b)
Fig. 2.14 (c)
\( F_0 \) (characterized by the attribute \( R \)) occurs during the adjective, and the \( F_0 \) lowering (characterized by attribute \( L \)) occurs during the noun. Further, the location of the characteristic \( F_0 \) movement inside each word corresponds to the primary stress (which is indicated by the superfix '1' in Fig. 2.14) as described in Section 2.3.5. In every phrase, and for every speaker, the \( F_0 \) rise occurs during the syllable with primary stress in the adjective and the \( F_0 \) lowering occurs at one of two different places in the noun, depending on the location of the primary stress in the noun. In the word 'kangaroo', the lowering occurs during the last syllable, 'roo', with primary stress, while in the words 'monkey' and 'gorilla', the lowering starts from the offset of the stressed syllable, and continues on the following nonstressed syllable, 'key' and 'la', respectively. Between the two stressed syllables in each noun phrase, the \( F_0 \) contour is kept high (roughly 10 Hz to 20 Hz above the baseline), forming the plateau, regardless of the distance from one stressed syllable to the next one. The schematized \( F_0 \) patterns, therefore, are well specified by the location of the two attributes \( R \) and \( L \). In Fig. 2.14, only the baselines corresponding to one of the phrases, the longest one, terminated by the word 'kangaroo', are shown.
The discrete representation of the schematized $F_0$ contours in Fig. 2.14 may be described for each of the three speakers, KN, JP and KS, as follows:

\begin{align*}
(2.5) & \quad \emptyset \quad R \quad L \\
(2.6) & \quad \emptyset \quad R \quad L \\
(2.7) & \quad \emptyset \quad R \quad L \\
(2.8) & \quad \emptyset (R) \quad L \\
(2.9) & \quad \emptyset (R) \quad L \\
(2.10) & \quad \emptyset (R) \quad L
\end{align*}

where '$\emptyset$' represents null attribute.

The attribute $P$ associated with $R$ as observed in KS is not marked in the above description. In our data, any noun phrase composed of an adjective and a noun indicates the sequence of the two basic attributes (i.e. $R$ and $L$). The $F_0$ contour of the noun phrases always forms the 'hat-pattern'. It may be safe to state that the noun phrase composed of two lexical words (an adjective and a noun) has a strong tendency to receive the two consecutive attributes: $R$ on the adjective and $L$ on the noun.

The $F_0$ contours of the compound noun (S45), composed of two nouns, and read by four speakers are shown in Fig. 2.15, in which the schematized $F_0$ patterns are superimposed on the original $F_0$ contours. The attribute pattern for each of four speakers, MB, SB, DK and KS, can be represented as
The $F_0$ contours and the corresponding schematized $F_0$ patterns for the noun phrase S45, 'My labor union,' read by the four speakers, MB in (a), SB in (b), DK in (c) and KS in (d).
Fig. 2.15
follows:

\[(R) \quad L \quad R \quad L\]  
\[(\emptyset) \quad R \quad L \quad \emptyset\]  
\[(\emptyset) \quad R \quad L\]  
\[(\emptyset) \quad R \quad L\]  
\[\text{(MB), S45}\]  
\[\text{(SB and KS), S45}\]  
\[\text{(DK), S45}\]

Two of the speakers (SB and KS) assign both R and L on the first noun, and no attribute (represented by the null attribute \(\emptyset\)) on the second noun; such a combination of attributes is different from the pattern associated with the noun phrases. A different organization can also be seen for the speaker MB.

We shall try to interpret the above observations. The observed attribute pattern for the noun phrases (we consider here only the adjective and the noun, but not the article) can be described symbolically as follows:

\[R \quad L\]  
\[w \quad w\]  
\[(2.14)\]

Recalling the two empirical hypotheses given in the previous section, the pattern (2.14) can be generated by postulating a rule as follows\(^2\):

\[RL \quad RL \quad R \quad L\]  
\[\text{Rule 01}\]  
\[\text{(w w )} \rightarrow w \quad w\]

CONDITION: \((w \quad w)\) must be a noun phrase or a compound noun, where the parentheses ",(,)" indicate the grouping of the word. If we assume the hypotheses to be true,
the right items are to be guaranteed as the immediate derivation of the left items, and there can be no intermediate stage in the derivation. For instance, the noun phrase, 'brilliant colors', in the sentence shown in Fig. 2.1, exhibits the two possible attribute patterns: $\text{RL}_w \text{RL}_w$ for the speaker JP, and $\text{R}_w \text{L}_w$ for the remaining two speakers. For the latter two speakers, the pattern indicates that the two words have been grouped, while they have not been grouped in the first case.

As will be shown later, not only a pair of lexical words composed of an adjective and a noun, but also pairs of lexical words of different decomposition (such as a noun and a verb), and even a lexical word and a function word, may indicate the attribute pattern generated by Rule 01. The condition in Rule 01, therefore, may be omitted having the following rule.

$$\text{RL}_w \text{RL}_w \rightarrow \text{R}_w \text{L}_w$$

Rule 1

There are two types of observed attribute patterns for the compound nouns as shown from (2.11) to (2.13). We may, therefore, postulate the following rule for the compound words:

$$\text{RL}_w \text{RL}_w \rightarrow \text{RL}_w \emptyset$$

Rule 2

$$(w_w) \rightarrow w_w$$
CONDITION: \((w \; w)\) must be a compound word. In Rule 2, the condition is generalized such that the rule can be applied for any compound word instead of only for compound nouns. This generalization is based on a correspondence between the compound rule (CR) proposed by Chomsky and Halle (1968), and by Halle and Keyser (1971), and Rule 2. It should be noted that only compound words can be applied to either Rule 1 or Rule 2, and thus two possible patterns are generated.

We have noted, in Section 2.4.1, that the pair of the basic attributes, \(R\) and \(L\), reflects an underlying grouping of the words. In the case of the compound words, the end of the group is not always marked, as seen in the generated sequence in Rule 2. Generally speaking, the attribute \(R\) signals the beginning (the first word) of the grouped words, while \(L\) does not always mark the end (the last word) of the group, especially in the case of compound words.

The \(F_0\) rises characterized by the attribute \(R\), therefore, are recognized to have the two following functions: stress-marking and the marking of the beginning of the group. None of the noun phrases in our data indicates a pattern such as "\(\emptyset \; RL\)\), and this phenomenon is probably due to the second function of the attribute \(R\). We have noted that
CR puts 1-stress on the first word in the compound word.

One might expect to find the following two attribute patterns: "RL $\emptyset$" and "R L", providing that the $R_o$ rise (R) is a stronger stress-marker than the following (1) (Morton and Jassem, 1965). Such examples have been seen in our data. On the other hand, the nuclear stress rule (NSR), proposed by Chomsky and Halle (1968), and by Halle and Keyser (1971), puts 1-stress on the final word of the noun phrases. Patterns such as "$\emptyset$ RL" or "L R", therefore, may be expected to appear. However, such cases have not been found in the data we analyzed. Our speakers seem to mark consistently the beginning of any grouped words by assigning R (which often occurs simultaneously with P) to the initial word. Presumably, this phenomenon constrains the attribute patterns for the noun phrases to be of one type only: "R L".

We note that the groupings of the words represented by the parentheses are not always related to the constituent structure. For instance, the attribute pattern in (2.7) read by MB indicates the following grouping of the words (using Rule 1):

$$ (R) \ L \ R \ L $$

(2.15) (My labor) (union) (MB) S45

This grouping is not the same as the constituent structure of the noun phrase, which can be represented as follows:

$$ (2.16) \ [\text{My} \ [\text{labor union}]_N]_{NP}, $$
where the brackets "[ ]" represent the grouping based on the constituent structure, and the subscripts represent the grammatical categories: N for a noun, and NP for a noun phrase. Therefore, we must recognize that other factors have to be involved in the determination of the grouping.

Perhaps the assignment of "(R)" on the function word "my", used to emphasize that particular word, influences the organization of the attribute pattern for the following compound word. Since (R) is assigned to 'my', the first noun, "labor", receives the attribute L for stress-marking: because only one word, the word "union" is left in the phrase, R and L are located on that word. Further examples of the disagreement between syntactic and semantic constituent structure and the actual grouping of the words will be seen in the following sections.

In order to derive the observed attribute patterns for the noun phrases shown at the beginning of this section, we must introduce one more rule. In the examples shown in (2.5) to (2.10) and in (2.12) and (2.13), the function words, 'the' and 'my' are regarded as receiving the null attribute, '∅', since the corresponding $F_0$ contours are located near the baselines. We consider that in those cases, the intrinsically assigned attributes are deleted. As shown later, the $F_0$ contours for lexical words which are isolated (i.e. not
grouped with other words) are also sometimes found near the baselines of the sentences. We therefore postulate the following deletion rule for any non-grouped word (both lexical or function words):

$$\text{Rule 03:} \quad w \rightarrow w$$

CONDITION: $w$ is not grouped with other words. For instance, the observed attribute pattern in (2.5) can be generated assuming the following grouping in the phrase:

$$RL \quad R \quad L \quad R \quad L$$

(2.17) ... the (enormous monkey)

Then, the application of Rule 03 to the first word, 'the', and the application of Rule 1 to the grouped words generate the desired pattern.

In the following section we shall describe the attribute patterns and their generation for word groups composed of more than two words, where the structure influences the attribute patterns inside the groups.

2.4.3 Attribute Patterns In Noun Phrases With Various Constituent Structures

2.4.3.1 Noun Phrases With Right-Branched Structure

The actual $F_0$ contours and the corresponding schematized $F_0$ contours for the noun phrase "the fat yellow alligator" in the sentence S29 are shown in Fig. 2.16. All the
The $F_0$ contours and the schematized $F_0$ patterns for the noun phrase located at the final position of the sentence S29, read by the three speakers, Kn in (a), JP in (b), and KS in (c). A.E. in each of the figures represents the amplitude envelope.
Fig. 2.16

A.E.

\( F_\theta (\text{Hz}) \)

(a) KN S29

(b) JP S29

(c) KS S29

the fat yellow alligator
schematized patterns exhibit the "hat-pattern": there is an $F_0$ rise in the first adjective, and an $F_0$ lowering during the noun. The $F_0$ contours of the word in the middle position in the phrase, "yellow" correspond to the plateau of the "hat-pattern" for the three speakers. The attribute pattern associated with the phrase can be therefore described as follows:

$$\emptyset(R) \quad \emptyset \quad L$$

(2.18) "... the fat yellow alligator."

(KN, JP, and KN) S29

In this particular example, the attribute $L$ is not assigned to the second syllable, "li", which is supposed to receive the attribute $L$, since the primary stress occurs in the first syllable of the word, the syllable "al". Perhaps the two first syllables, "al" and "li" are pronounced as if they were a single syllable. The attribute $(R)$ seems to indicate the beginning of the noun phrase, while the attribute $L$ marks its end.

This example, and the examples shown from (2.5) to (2.10) in the previous section, might lead us to postulate a simple theory; for instance, the attribute patterns associated with the noun phrases are generated by assigning $R$ on the initial word and $L$ on the final word. However, the following analysis suggests that a more sophisticated theory,
which takes into account the internal structure, is necessary to generate appropriate attribute patterns for the noun phrases.

In Fig. 2.17, the $F_0$ contours and the schematized patterns for the noun phrases S47, read by the speakers SB, DK and KS are represented. The attribute patterns associated with S47 for each speaker may be described as follows:

\[
\begin{align*}
\emptyset & \quad R \quad (L)R \quad P \quad L & \text{(SB) S47} \\
\emptyset & \quad R \quad R1 \quad L & \text{(DK) S47} \\
\emptyset & \quad RL \quad R \quad L \quad \emptyset & \text{(KS) S47}
\end{align*}
\]

It can be observed in Fig. 2.17 (c) that the valley (formed by the successive attributes L and R) at the boundary between the words "lazy" and "union" does not reach the baseline. If the valley is neglected, then we can derive the following attribute pattern:

\[
\begin{align*}
\emptyset & \quad R \quad L \quad \emptyset & \text{(KS) S47}
\end{align*}
\]

We shall discuss this pattern later, and meanwhile we assume the speaker KS indicates the pattern shown in (2.21).

Let us consider the generation of the patterns associated with noun phrases composed of three lexical words in the above examples (the first word in the noun phrase, "my" is assigned the null attribute $\emptyset$ by application of Rule 03).
Figure 2.17

The $F_0$ contours and the corresponding schematized $F_0$ patterns for the noun phrase S47, 'My lazy union president,' read by three speakers, SB in (a), DK in (b), and KS in (c). A.E. in each figure represents the amplitude envelope.
My lazy union president.
The noun phrase can be assumed to have a right-branched structure:

(2.23) \[ [w [w w] N ] \] NP

If the speakers actually grouped the last two words (by assigning the attribute R on the word "union" and L on the word "president"), then either Rule 1 or Rule 2 can be applied, resulting in the generation of two possible patterns:

(2.24) \[
\begin{array}{cccc}
RL & RL & RL & \text{Rule 1/Rule 2} \\
( w & w & w ) & \rightarrow & w & w & w & / & w & w & w,
\end{array}
\]

where "/" must be read "either the left items or the right items."

The observed pattern shown in (2.21) for KS corresponds to the first generated pattern in (2.24), and the pattern described in (2.19) for SB corresponds to the second pattern in (2.24). However, the attribute pattern for DK (represented in [2.20]) does not fit any of the above generated patterns.

In order to describe the pattern for DK, another transformation must take place: the two last words have to be regarded as forming a subgroup inside the group composed of the three lexical words. The transformational rule may be described as follows:

Rule 004 \[
\begin{array}{c}
RL \ R \ L \\
( w & w & w & w ) \rightarrow w & w & w & w
\end{array}
\]
The observed pattern shown in (2.20) thus can be derived assuming the right-branched structure of the noun phrase to be as follows:

\[
(2.25) \quad \text{RL} \quad \text{RL} \quad \text{RL} \quad \text{Rule 1} \quad \text{RL} \quad \text{RL} \quad \text{Rule 004} \quad \text{RL} \quad \text{RL}
\]

\[
(2.25) \quad (w \quad (w \quad w)) \quad \longrightarrow \quad (w \quad w \quad w) \quad \longrightarrow \quad w \quad w \quad w
\]

RL seems to mark the beginning of the subgroup.

Other examples of \(F_0\) contours corresponding to the \(F_0\) pattern generated by Rule 004 have been shown in Fig. 2.13 in Section 2.3.5. The \(F_0\) contour for the sentence S11 in Fig. 2.13 (a) corresponds to the following sequence of attributes:

\[
(2.26) \quad \text{P} \quad \text{R} \quad \text{Rl} \quad \text{L}
\]

\[
(2.26) \quad "\ldots \text{big white dog} \ldots" \quad (\text{KS}) \quad \text{S11}
\]

However, for the sentence S25, the same noun phrase indicates another sequence of attributes:

\[
(2.27) \quad \text{P} \quad \text{R} \quad \emptyset \quad \text{L}
\]

\[
(2.27) \quad "\ldots \text{big white dog} \ldots" \quad (\text{KS}) \quad \text{S25}
\]

A similar phenomenon can be observed in Fig. 2.13 (b). In the comparison of the sequences represented in (2.26) and in (2.27), we may postulate the following weakening rule of R1:

\[
\text{Rule 05} \quad \text{Rl} \longrightarrow \emptyset
\]

We have noted in Section 2.3.5, that the \(F_0\) contour characterized by the attribute Rl varies continuously from a clear rise to a gradual fall along the plateau. Perhaps this phenomenon is a reflection of the occurrence of the attribute Rl in the intermediate position of the following derivations:
In the above examples, the grouping of the words corresponds to the constituent structure of the noun phrase. The attribute $R$ signals the beginning of the constituent, while $R_1$ marks the beginning of the sub-constituent. If the attribute $R_1$ occurs on the word in the middle position, it indicates precisely that the structure is a right-branched structure (It was unfortunate that we stated in an earlier report that $R_1$ indicates a parallel structure (Maeda, 1974).)

However, we must state again that the grouping of the words does not always correspond to the syntactic or semantic constituents. For instance, the $F_o$ contours shown in Fig. 2.18 represent the following sequences of attributes, for the three speakers KN, KS and JP:

\[
\begin{align*}
(2.29) & \quad " \ldots \text{small black cat} \" \quad (\text{KN and KS}) \ S_{11} \\
(2.30) & \quad \underbrace{R \ \emptyset \ L}_{(R)} \quad (R)L \quad (\text{JP}) \ S_{11}
\end{align*}
\]

The sequence in (2.29) can be generated as shown in (2.28), assuming the right-branched structure. The pattern in (2.30), however, can be derived only if we assume that the first two words are grouped:

\[
\begin{align*}
(2.31) & \quad (R) \ L \ (R) \ L \ (R)L \ \text{Rule 1} \ (R) \ L \ (R)L \\
& \quad (\text{small black}) \text{ cat} \rightarrow \text{small black cat}
\end{align*}
\]
The $F_0$ contours and the schematized patterns for the noun phrase located at the final position in the sentence S11, read by the three speakers, KN in (a), JP in (b), and KS in (c). A.E. represents the amplitude envelope.
...the small black cat.

Fig. 2.18
The generated attribute pattern in (2.31) may not specify the entire left-branched structure in terms of the attributes, but only indicates that the first two words "small black" are grouped. However, this is sufficient for the specification of the structure of the phrase, since it is reasonable to assume that the speaker and the listener interpret the three consecutive words as a constituent by the meaning of the sentence, even though there is no phonetic indication of the constituent. The grouping of the first two words by the attribute R and L is inconsistent with the constituent structure, that may be assumed to be the right-branched structure. Probably, the other factor, say a principle of economy in physiology (which will be described later) for the stress-marking, is a dominant influence in determining the grouping and then the attribute pattern.

Let us now investigate the attribute pattern shown in (2.22) for the phrase S47 "My lazy union president". The pattern can be derived by applying Rule 1 to "lazy union" and Rule 03 to "president". It would appear that a left-branched structure of the phrase must be assumed in this derivation. However, it is more reasonable to assume that the structure is right-branched. In such case, the observed attribute pattern in (2.22) may be recognized as the immediate derivation from the pattern shown in (2.21). The following rule,
therefore, is postulated.

\[
\begin{array}{c}
\text{RL} & \text{RL} & \emptyset \\
\text{Rule 06} & (w & w & w) \rightarrow & w & w & w .
\end{array}
\]

Rule 06 specifies the mapping of the attribute patterns when a word is bonded with the following compound word which is composed of two lexical words.

It may be noticed that Rule 06 is quite similar to Rule 1. If we establish a convention such that any compound word with two lexical words can be regarded as one word, then Rule 1 may be applied to generate the observed pattern. However, such a convention creates some problems. For instance, Rule 2, instead of Rule 1, also can be applied, resulting in a pattern which is not found at all in our entire data such as:

\[
\begin{array}{c}
\text{RL} & \text{RL} & \emptyset & \text{Rule 2} & \text{RL} & \emptyset & \emptyset \\
(2.32) & (w & w & w) \rightarrow & w & w & w ,
\end{array}
\]

where the last two words are considered as one word. Therefore, Rule 06 has to be only used in the rather special case in which the compound word is grouped with the preceding word.

It should be noticed in Rule 004, that only the attributes associated with the first two words are involved in the transformation process. Rule 004 thus may be generalized
to the following form:

\[
\text{Rule 04 } ( \text{w w x} ) \rightarrow \text{w w x},
\]

where \( x \) is a sequence of \( \text{w} \)'s.

To explain how Rule 04 works, it may be instructive to show an idealized example. Suppose that a group of words has a right-branched structure as represented by the subgrouping of the words as follows:

\[
(2.33) \quad ( \text{w ( w ... ) ( w ( w w ) ... )} ),
\]

where \( n \) indicates the number of occurrences of \( \text{w} \)'s. Let us assume that the innermost group corresponds to a noun phrase. The application of the Rule 1 generates the following sequence:

\[
(2.34) \quad ( \text{w ( w ........ ( w w w ) ... )} ),
\]

Then, Rule 04 can be applied cyclically until all parentheses are exhausted, resulting in the following pattern:

\[
(2.35) \quad \text{R Rl Rl Rl L w w ........ w w w}
\]

Evidently, \( \text{Rl} \) indicates the beginning of each subgroup. Therefore, if the subgroups correspond to the constituents \( \text{Rl} \) can be said to indicate the beginning of each constituent inside the group.
An $F_0$ contour indicating such attribute patterns, in the case where $n=4$, is shown in Fig. 2.19 (a), where the corresponding attribute sequence is described as follows:

$$
\begin{array}{cccc}
P & R & R_1 & R_1 & L \\
\end{array}
$$

(2.36) "... small black fat cat. " (KS)

The application of Rule 05 to the first $R_1$ generates the following pattern:

$$
\begin{array}{cccc}
P & R & \emptyset & R_1 & L \\
\end{array}
$$

(2.37) "... small black fat cat ..." (KS)

The $F_0$ contour indicating such a pattern is found in Fig. 2.19 (b).

2.4.3.2 Noun Phrase With A Left-branched Structure

So far, we have described the generation of the attribute sequences of words with a right-branched structure. Let us now investigate the sequences of attributes for the left-branched structure. In Fig. 2.20, we show the schematized $F_0$ contour superimposed on the corresponding $F_0$ contour of the noun phrase $S_{46}$ for the two speakers SB and KS. The attribute sequences for each speaker can be described as follows:

$$
\begin{array}{cccc}
BL & \emptyset & P & L \\
\end{array}
$$

(2.38) "My labor union president" (SB) $S_{46}$

$$
\begin{array}{cccc}
BL & \emptyset & P & \emptyset & (R) & L \\
\end{array}
$$

(2.39) "My labor union president" (KS) $S_{46}$

The speaker DK indicates the same attribute pattern as that
Figure 2.19

The $F_0$ contours and the corresponding schematized patterns and the amplitude envelope (A.E.) for the noun phrase 'the small black fat cat,' read by the speaker KS. The noun phrase (NP) in (a) is taken from the sentence 'The dog likes NP,' and in (b) from the sentence 'NP likes the dog,' which have both been studied in the preliminary analysis and not listed in Table 2.1.
Fig. 2.19

(a) KS

(b) KS

A.E.

$P_0$ (Hz)

120

100

80

A.E.

$P_0$ (Hz)

140

120

100

80

the small black fat cat

R1

P

R1

Fig. 2.19
The $F_0$ contours and the corresponding schematized $F_0$ patterns and the amplitude envelope (A.E.) for the noun phrase S46 'My labor union president,' read by two speakers, SB in (a), and KS in (b).
My labor union president.

Fig. 2.20
The attribute pattern in (2.39) can be generated by applying Rule 03 to "my" and Rule 2 to the words "labor" and "union". The pattern in (2.38) seems to indicate that the last three words are grouped, since R (associated with P) is assigned to the word 'labor' and L to the last word 'president'. Assuming a left-branched structure, the group can be represented symbolically as follows:

\[(2.40) \quad (w \ w \ w) \wedge (w \ w \ w)\]

Then, the application of either Rule 1 or Rule 2 can generate the following two possible patterns:

\[(2.41) \quad (w \ w \ w) / (w \ w \ w)\]

It seems preferable to relate the first pattern in (2.41) to the observed pattern in (2.38), since the mapping involves only the two last words. Then we may postulate the following rule:

\[\text{Rule 07} \quad (w \ w \ w) \rightarrow w \ w \ w\]

As we have noted, R1 signals the beginning of the subgroup. Since the second word in Rule 07 corresponds to the end of the subgroup, R1 should not appear in the patterns for Rule 07. (We implicitly assume that R1 signals only the beginning of a subgroup.) It may be concluded, therefore, that the
right-side items in Rule 07 must be derived directly from the left-side items.

By an argument similar to that made in the case of Rule 04, Rule 07 may be generalized to the following rule:

\[ \text{Rule 07 } (x \ w \ w) \rightarrow x \ w \ w, \]

where \( x \) is a sequence of \( w \)'s.

Apparently, Rule 07 governs the mapping of the attribute patterns, when a word is grouped with the preceding already grouped words. Consider a sequence of words that indicates the left-branched structure described by the subgroupings as follows:

\[ (2.42) \ \underbrace{((w \ w ) w ) \ldots \ldots w ) }_{n}, \]

where \( n \) represents the number of words. Applying Rule 2 and Rule 1 to the above sequence, the following two sequences (2.43) and 2.44) are generated, respectively.

\[ (2.43) \ \underbrace{((w \ w ) w ) \ldots \ldots w ) }_{R L \ \emptyset \ R L \ R L \ R L}, \]

\[ (2.44) \ \underbrace{((w \ w ) w ) \ldots \ldots w ) }_{R \ \emptyset \ L \ R L \ R L \ R L}. \]

In the case of the sequence (2.43), further derivation cannot be made. In the case of the sequence in (2.44), Rule 07 can be applied cyclically until all parentheses are cancelled,
deriving the following sequence:

\[ R \emptyset \emptyset \emptyset L \]
\[ w \ w \ w \ldots \ w \]

(2.45)

The \( F_0 \) contours indicating such sequence of attributes, for \( n=4 \), are shown in fig. 2.21 (a). The corresponding attribute sequence is represented as follows:

\[ \begin{array}{cccc}
BL & \emptyset & P & \emptyset \\
R & \emptyset & (R) & L \\
\end{array} \]

(2.46) " My labor union president election. "

(SB) S48

In the above example, the four nouns are regarded as being grouped. However, this is not the only observed pattern for that phrase. In Fig. 2.21 (b), we show the \( F_0 \) contour and the schematized \( F_0 \) pattern for the same phrase, indicating the following attribute pattern:

\[ \begin{array}{cccc}
BL & \emptyset & P & \emptyset \\
R & \emptyset & (R) & L \\
\end{array} \]

(2.47) ' My labor union president election ' 

(KS) S48

To derive the above pattern, the following grouping of words must be assumed:

(2.48) " My ( labor union ) ( president election )."

Then the applications of Rule 3 to the first isolated word, 'my', and of Rule 2 to each of the following two groups of words, generate the observed pattern in (2.47).
Figure 2.21

The $F_o$ contours and the corresponding schematized $F_o$ patterns and the amplitude envelope (A.E.) for the noun phrase S48, 'My labor union president election' read by two speakers, DK (a) and KS in (b).
My labor union president election.

Fig. 2.21
Another interesting example is shown in Fig. 2.22 for the noun phrase S53, read by DK and by KN. The attribute pattern for DK in Fig. 2.22 (a) may be represented as follows:

\[(2.49) \quad \text{'My father's mother's sister's dog'} \]

This pattern can be derived by assuming a left-branched structure. The pattern for KS in Fig. 2.22 (b) can be described as follows:

\[(2.50) \quad \text{'My father's mother's sister's dog'} \]

Since R1 indicates the beginning of the subgroup, the following subgroup may be postulated:

\[(2.51) \quad \text{My ( ( father's ( mother's sister's ) ) ) dog )} \]

The derivation of the observed pattern is described as follows:

\[(2.52) \quad \text{Rule 1} \]

\[(2.53) \quad \text{Rule 04} \]

\[(2.54) \quad \text{Rule 7} \]

\[(2.55) \quad \text{Rule 3} \]
The $F_0$ contours and the corresponding schematized $F_0$ patterns and the amplitude envelope (A.E.) for the noun phrase S53, 'My father's mother's sister's dog' read by two speakers, DK in (a) and KS in (b).
Fig 2.22

My father's mother's sister's dog.
where each underline indicates the segment where the rule is applied. It should be noticed that the same observed patterns can be derived assuming a different subgrouping of the words as follows:

\[
(2.53) \quad \text{RL RL RL RL RL}
\]

We must state, therefore, that the attribute patterns do not always contain sufficient information to determine uniquely the subgrouping of the words. Furthermore, the attribute RL that can be used to specify the right-branched structure, could be weakened by the application of Rule 05. A group of words with a right-branched structure and with a left-branched structure, therefore, can exhibit the same attribute pattern having R on the initial word and L on the final word.

2.4.3.3 Words Containing More Than One Pair of The Attributes R and L

We have described, so far, the generation of the attribute patterns assuming that a word intrinsically receives only one pair of the basic attributes, R and L. However, this assumption is not always correct. We shall show an example in which a word receives more than two attributes. Applications of the rules already proposed seem to be capable, in a generalized sense, of handling attribute patterns for such words. However, we shall not go much further in this problem, since we have only one such example.
The $F_0$ contours and the corresponding schematized $F_0$ patterns of the noun phrase S50, read by the four speakers, are shown in Fig. 2.23. The discrete representation of the schematized $F_0$ patterns can be represented as follows:

\[(2.54) \quad \text{BL } \emptyset \quad R \quad L(R) \quad L \quad \text{computer course } \quad \text{MB S50}\]

\[(2.55) \quad \text{BL } \emptyset \quad R \quad P \quad L \quad \text{computer course } \quad \text{SB S50}\]

\[(2.56) \quad \text{BL } \emptyset \quad R \quad L \quad (R) \quad L \quad \text{computer course } \quad \text{DK S50}\]

\[(2.57) \quad \text{BL } \emptyset \quad R \quad L \quad (R) \quad (R) \quad L \quad \text{computer course } \quad \text{KS S50}\]

In (2.57), the word 'computer' receives three attributes, and in (2.54) and in (2.56), L is located in front of (R) in that word. It seems to be reasonable to assume therefore, that the word 'computer' may contain initially two pairs of the attributes R and L as follows:

\[(2.58) \quad (R) L \quad (R) \quad L \quad \text{computer}\]

We speculate that the isolated words in which the secondary stress occurs in front of the primary stress can probably exhibit the pattern shown in (2.58). The proposed rules cannot apply directly to such words associated with two pairs of the basic attributes.
Figure 2.23

The $F_o$ contours and the corresponding schematized $F_o$ patterns for the noun phrase S50 'My morning computer course' read by the four speakers, MB in (a), SB in (b), DK in (c) and KS in (d).
Fig. 2.23

MB S50

My morning computer course.

SB S50

DX S50

KS S50

P

P

P
As we have formulated the rules, we have implicitly assumed that any word in a group contains initially only one pair of R and L. What is essential in the mapping of the patterns in the rules (i.e. Rule 1, Rule 2, Rule 04, Rule 06 and Rule 7) is the deletion or transformation of the two attributes, L followed by R, or R followed by L, such that the generated pattern also satisfies the empirical hypothesis described in Section 2.4.1. Thus we may apply the rules by taking into account only the attributes located at the onset and the offset of the pattern assigned on a group or on a word. In other words, we use the rules pretending that the middle portions of the attributes in the patterns are invisible. Let us call this kind of application the application of the rules in the generalized sense.

For instance, if the second word of a compound word contains two pairs of R and L, the application of Rule 1 in the generalized sense can be described as follows:

\[
\begin{align*}
\text{RL RLRL} & \quad \text{Rule 1* R LRL} \\
(w \quad w) & \rightarrow w \quad w,
\end{align*}
\]

where the superscript "*" indicates the generalized application. (We are only concerned here with the case when the compound word is uttered as a noun phrase). It should be noticed that the above transformation derives the attribute pattern corresponding to that of "computer morning" in (2.57).
If the first word in the compound noun contains two pairs of R and L, the application of Rule 1 in the generalized sense is represented as follows:

\[(2.60) \quad (w \quad w) \rightarrow w \quad w\]

Further, when the generated sequence in (2.60) is grouped with a word, say with the previous word, the application of Rule 7 in the generalized sense will generate a new sequence as follows:

\[(2.61) \quad (w \quad w \quad w) \rightarrow w \quad w \quad w\]

The above new sequence corresponds to the observed attribute pattern for "morning computer course" in (2.54) and in (2.56). In the case of the utterance in (2.55), the direct application of the rules can generate the observed attribute patterns, assuming that the word "computer" initially contains only one pair of R and L associated with the syllable with primary stress.

The above analysis has shown that the proposed rules are capable of dealing with the words which can receive initially more than one pair of R and L. However, we do not know when a word exhibits one or two pairs of the attributes R and L. This problem is beyond the scope of this study.

It may be noteworthy that, when a group contains such a word (with more than one pair of attributes), there is no
simple correspondence between the attribute R and L and the underlying group in which R marks the beginning of the group and L signals the end. To obtain such correspondence, the attributes L followed by R which occur within a word must be subtracted.

2.4.3.4 Assignment of the Attribute P

We have not yet discussed the assignment of the attribute P; we remarked only that this attribute is often associated with the attribute R. The examples represented in Fig. 2.23 illustrate where P can occur. In the case of speaker SB, shown in Fig. 2.23 (b), P (associated with R) occurs not only at the beginning of the group, but also on the plateau portion, specifically, during the syllable with 1-stress in the second word, 'computer.' It may be stated, therefore, that P can occur at the beginning of the subgroup, since the compound noun "computer course" is regarded as a subgroup. In the cases of the speaker DK and KS (shown in Fig. 2.23 (c) and (d), respectively), P occurs with (R) during the word "computer". This P associated with R may be recognized as signalling the beginning of the constituent "computer course" although, in the case of KS, the two words are not grouped because each of the words exhibits its own "hat-pattern"). Further examples confirming the fact that P signals the beginning of the subgroup will be described in the following section.
The above observations lead to a generalization of Rule 04 with respect to the assignment of P at the beginning of a subgroup, as follows:

$$\text{RL} \; \begin{array}{c} R \; \begin{array}{c} w \; w \; x \end{array} \end{array} \rightarrow \begin{array}{c} R \; \begin{array}{c} R1/P \; \end{array} \end{array},$$

where "R1/P" must be read as "either R1 or P". In our data, R1 occurs when the subgroup (i.e. 'w x') corresponds to a noun phrase, while P occurs when the subgroup corresponds to a compound word. We must admit, however, that our data are too small to consider the above statement as a conclusive one. We thus maintain Rule 4 as a non-deterministic rule.

For the sake of the consistency, Rule 06 which specifies the mapping of the patterns when a compound word (composed of two words) is grouped with the preceding word, must be modified as follows:

$$\text{RL} \; \begin{array}{c} RL \; \begin{array}{c} \emptyset \end{array} \end{array} \rightarrow \begin{array}{c} R \; \begin{array}{c} PL \; \emptyset \end{array} \end{array},$$

where we assume P to occur during the syllable with primary stress. Thus, P must precede L in Rule 6.

As a consequence of these modifications of the rules, it is necessary to modify the weakening rule, Rule 05, to delete P as well as R1, as follows:

$$\text{Rule 5: } R1 \rightarrow \emptyset \; / \; P \rightarrow \emptyset$$
One more rule must be postulated regarding the assignment of P which often occurs simultaneously with R, as follows:

\[ P \]

Rule 8  \[ R \rightarrow R \]

The attribute P in Rule 8, and P in Rule 4 and Rule 7 are somewhat different in their functions in the sense that the first P signals the beginning of the group with R, while the second P on the plateau marks the beginning of the subgroup. In that sense, perhaps, the first P may be regarded as emphatic, and the second P as grammatical.

We have proposed, so far, eight rules in Sections 2.4.2 and 2.4.3. These rules are summarized in Table 2.7. The rules essentially state that the attribute patterns are determined such that R (with or without P) signals the beginning of the groups, and either R1 or P marks the beginnings of the subgroups. The rules containing the parentheses must be applied cyclically so that all parentheses are cancelled. As we have noted above, a compound word can be assigned two possible patterns by the application of either Rule 1 or Rule 2. Rule 5 can be used anytime, while Rule 8 which locates P on R must be applied after all the parentheses are exhausted. If R is associated with P, then the rules, except Rule 5, can no longer be applied, and thus the cyclic procedure must stop with some of the parentheses still remaining. It should be noticed that the manner of groupings
Rule 1  \[
\begin{array}{cc}
\text{RL} & \text{RL} \\
\text{w} & \text{w} \\
\end{array}
\rightarrow
\begin{array}{c}
w \\
w \\
\end{array}
\]

Rule 2  \[
\begin{array}{cc}
\text{RL} & \text{RL} \\
\text{w} & \text{w} \\
\end{array}
\rightarrow
\begin{array}{c}
w \\
w \\
\end{array}
\]

CONDITION: \((w \quad w)\) must be a compound word

Rule 3  \[
\begin{array}{c}
\text{RL} \\
\text{w} \\
\end{array}
\rightarrow
\begin{array}{c}
w \\
\end{array}
\]

CONDITION: \(w\) is not grouped with another word

Rule 4  \[
\begin{array}{cccc}
\text{RL} & \text{R} & \text{R} & \text{R} \\
\text{w} & \text{w} & \text{x} \\
\end{array}
\rightarrow
\begin{array}{ccc}
w & w & x \\
\end{array}
\]

where \(x\) is a sequence of \(w\)'s, \(\text{R}1/\text{P}\) must be read "either \(\text{R}1\) or \(\text{P}\)."

Rule 5  \[
\begin{array}{c}
\text{R}1 \\
\end{array}
\rightarrow
\begin{array}{c}
\emptyset \\
\end{array}
\]

\[
\begin{array}{c}
\text{P} \\
\end{array}
\rightarrow
\begin{array}{c}
\emptyset \\
\end{array}
\]

Rule 6  \[
\begin{array}{cccc}
\text{RL} & \text{RL} & \emptyset \\
\text{w} & \text{w} & \text{w} \\
\end{array}
\rightarrow
\begin{array}{cc}
w & w \\
w & w \\
\end{array}
\]

Rule 7  \[
\begin{array}{cccc}
\text{L} & \text{RL} & \emptyset & \text{L} \\
x & w & w & \\
\end{array}
\rightarrow
\begin{array}{ccc}
x & w & w \\
\end{array}
\]

Rule 8  \[
\begin{array}{c}
\text{R} \\
\end{array}
\rightarrow
\begin{array}{c}
\text{R} \\
\end{array}
\]

**TABLE 2.7** A summary of the rules proposed in Sections 2.4.2 and 2.4.3.
and subgroupings are not arbitrary. In certain groupings, the parentheses cannot be erased completely. For the groupings, 
\[(w w(w w)w)\]
for instance, there is no way to erase the outside parentheses by the applications of the rules.

The postulation of the rules is based strictly on observation of the attribute patterns. The existence of underlying groups and subgroups is only an empirical hypothesis. It is not necessary that a speaker actually generates the patterns according to the rules somewhere in his brain. It is true, however, that those groups and subgroups which produce the observed patterns using the rules often correspond to constituents in the sentences. The rules proposed here should be considered as prototypes, and they may undergo further refinement on the basis of more extensive data. In the following section, we shall show, however, that these rules are sufficient to generate any observed pattern in our data.

It may be noteworthy to mention that the eight rules are formulated in a form of deletion in the sense that the application of any rule (except Rule 8) causes a decrease in the number of attributes. The exactly same attribute pattern can be generated using a set of rules which are formulated in a generating mode, assuming the same groupings and subgroupings of the words in a sentence. In such a process, any word must be assumed not to have any attribute initially.
However, we feel that the process which uses mainly a deletion process is more consistent with the principle of economy.

2.4.4 Ambiguous Noun Phrases

Noun phrases such as 'light house keeper' and 'American history teacher' can indicate two different constituent structures - a left-branched structure and a right-branched structure - depending on the semantic interpretation. Bolinger and Gerstman (1957) have shown, using a tape-splicing technique, that the utterance '(light house) keeper' is changed into 'light (house keeper)', and vice versa, by varying the length of the interval (i.e. the pause) between the words 'light' and 'house'. Lieberman (1967) has claimed that the speakers actually vary the disjunctures (that is the intervals between the vowels) to differentiate the two constituent structures.

Our primary interest here is to evaluate the extent to which the attribute patterns reflect the two constituent structures. The speakers read the sentences containing the noun phrases of which the structures are specified using parentheses, such as '(small school) boy' in S13, 'small (school boy) in S12, ' (light yellow) bus' in S52, and 'light (yellow bus' in S51. The F₀ contours, and the schematized F₀ patterns and the amplitude envelope (A. E.) of these phrases
are shown in Fig. 2.24 for 'small school boy' and in Fig. 2.25 for 'light yellow bus'. In each figure, the noun phrases with left-branched structure is presented at the top, and with right-branched structure at the bottom. The corresponding discrete representation of the schematized patterns may be described for each speaker as follows:

<table>
<thead>
<tr>
<th>left-branched</th>
<th>right-branched</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>(small school) boy</em></td>
<td><em>(small (school boy)</em></td>
</tr>
<tr>
<td>(R) L Ø</td>
<td>(R) P L</td>
</tr>
<tr>
<td>w w w</td>
<td>w w w</td>
</tr>
<tr>
<td><em>(R)R L Ø Ø</em></td>
<td><em>(R)L (R)P L</em></td>
</tr>
<tr>
<td>w w w</td>
<td>w w w</td>
</tr>
<tr>
<td><em>(R) L (R)R L</em></td>
<td><em>(R) PL Ø</em></td>
</tr>
<tr>
<td>w w w</td>
<td>w w w</td>
</tr>
</tbody>
</table>

*(light yellow) bus* | *light (yellow bus)* |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R Ø L</td>
<td>R Ø L</td>
</tr>
<tr>
<td>w w w</td>
<td>w w w</td>
</tr>
<tr>
<td>R L (R)L</td>
<td>R R1 L</td>
</tr>
<tr>
<td>w w w</td>
<td>w w w</td>
</tr>
<tr>
<td>R Ø L</td>
<td>R R1 L</td>
</tr>
<tr>
<td>w w w</td>
<td>w w w</td>
</tr>
</tbody>
</table>

In order to assess the above observed patterns, it may be instructive to generate the possible attribute patterns for the two noun phrases using the proposed rules listed in Table 2.7. The derivation of the patterns for 'small school boy' is shown in Fig. 2.26, and for 'light yellow bus' in
The F₀ contours, the corresponding schematized patterns, and the amplitude envelope (A.E.) of the phrase 'small school boy,' read by the three speakers KN in (a), JP in (b), and KS in (c). In each figure, the phrase with the left-branched structure (S13) is presented on the top, and the phrase with the right-branched structure (S12) on the bottom.
The small school boy.
Fig. 2.24 (b)

A.E.

\[ F_0 (\text{Hz}) 

.....the small school boy.  JP S13

A.E.

\[ F_0 (\text{Hz}) 

JP S12

---

2 s
A.E.

\[ F_0(\text{Hz}) \]

130
110

.. the small school boy.

KS S13

A.E.

\[ F_0(\text{Hz}) \]

120
100

KS S12

Fig. 2.24 (c)
The $F_0$ contours, the corresponding schematized patterns and the amplitude envelope (A.E.) of the phrase 'my light yellow bus,' read by the three speakers SB in (a), JP in (b), and KS in (c). In each figure, the phrase with the left-branched structure (S52) is presented at the top, and the phrase with the right-branched structure (S51) at the bottom.
Fig. 2.25 (a)
Fig. 2.25 (b)

A.E.

F₀ (Hz)

My light yellow bus.

DK S52

120

100

80

A.E.

F₀ (Hz)

DK S51

130

110

90

70
A.E.

$P_0 (Hz)$

140

120

100

A.E.

$P_0 (Hz)$

140

120

100

Fig. 2.25 (c)
Fig. 2.27. In each of the two figures, (a) represents the left-branched structure, while (b) represents the right-branched structure. To avoid unnecessary complications, Rule 3 and Rule 8 are not used in these derivations (except in Fig. 2.27 [a]), although these rules may be applied. The generated pattern is marked with the speaker's symbols such as (DK) and (KS), when the pattern corresponds to the observed pattern for the speakers.

In Fig. 2.26 and in Fig. 2.27, the pattern \( R \emptyset L \) \( w w w \) is found both for the left-branched structure (in [a]) and the right-branched structure (in [b]). That pattern, therefore, does not specify the internal structure of the two noun phrases, but it indicates that the three words are grouped. Each of the remaining generated patterns contains information about the structure, and thus these patterns can be said to contrast the two different constituent structures. Let us call these patterns "contrastive patterns".

Apparently, only three out of the six speakers indicate the contrastive pattern for the two phrases with the left-branched structure. None of the generated patterns in Fig. 2.26 (a) corresponds to the observed one for the phrase in (2.63) for JP. But it is seen in Fig. 2.27 (a), as the pattern marked by JP*. Presumably, the two first words "small school" are uttered as a compound word, instead of a
The generation of the possible attribute patterns for the noun phrase Adj+N+N, with left-branched structure in (a), and with right-branched structure in (b), using rules which are listed in Table 2.7. Each symbol, such as (KS) indicates the speaker whose schematized pattern (shown in Fig. 2.24) corresponds to the generated attribute pattern marked by the symbol.

The generation of possible attribute patterns for the noun phrase, Adj+ N +N, with left-branched structure in (a) and with right-branched structure in (b), using the rules which are listed in Table 2.7. Each symbol, such as (DK), indicates the speaker whose schematized $F_0$ pattern (shown in Figure 2.25) corresponds to the generated attribute pattern marked by that symbol. For (JP*), see text.
(a) \( ( \text{small school} ) \text{ boy} \)

\[ \text{Rule 1} \]
\[
\begin{array}{ccc}
R & L & RL \\
w & w & w \\
\end{array}
\]
\[ \ldots \ldots \text{(KS)} \]

\[ \text{Rule 7} \]
\[
\begin{array}{ccc}
R & \emptyset & L \\
w & w & w \\
\end{array}
\]
\[ \ldots \ldots \text{(KN)} \]

(b) \( ( \text{small ( school boy )}) \)

\[ \text{Rule 2} \]
\[
\begin{array}{ccc}
RL & RL & \emptyset \\
w & w & w \\
\end{array}
\]

\[ \text{Rule 6} \]
\[
\begin{array}{ccc}
R & PL & \emptyset \\
w & w & w \\
\end{array}
\]
\[ \ldots \ldots \text{(KS)} \]

\[ \text{Rule 5} \]
\[
\begin{array}{ccc}
R & L & \emptyset \\
w & w & w \\
\end{array}
\]

\[ \text{Rule 1} \]
\[
\begin{array}{ccc}
RL & R & L \\
w & w & w \\
\end{array}
\]
\[ \ldots \ldots \text{(JP)} \]

\[ \text{Rule 4} \]
\[
\begin{array}{ccc}
R & Rl/P & L \\
w & w & w \\
\end{array}
\]
\[ \ldots \ldots \text{(KN)} \]

\[ \text{Rule 5} \]
\[
\begin{array}{ccc}
R & \emptyset & L \\
w & w & w \\
\end{array}
\]

Fig. 2.26
(a) 

\[
\begin{align*}
&\text{Rule 2} \\
&\text{Rule 1} \\
&\text{Rule 3} \\
&\text{Rule 7}
\end{align*}
\]

(b) 

\[
\begin{align*}
&\text{Rule 1} \\
&\text{Rule 4} \\
&\text{Rule 5}
\end{align*}
\]

Fig. 2.27
noun phrase. This pattern, however, is certainly contrastive, since such a pattern cannot be generated assuming the right-branched structure. On the other hand, five out of the six speakers indicate the contrastive patterns for the right-branched structure. Only the speaker SB does not show the contrastive pattern for this structure. In fact, the two F₀ contours shown in Fig. 2.25 (a) and (b) exhibit similar curves. Further, notice that the corresponding two amplitude envelopes in the portion of 'light yellow bus' are quite similar, indicating that the disjuncture contrast does not exist either.

In general, the right-branched structure exhibits the contrastive patterns more often than the left-branched one. This is presumably due to the fact that the speakers are in the habit of grouping the three words of such a noun phrase, assigning R on the first word and L on the last word. However, to specify the left-branched structure, the two first words must be grouped, by putting R on the first word and L on the second word, and R and L on the last word, as seen in Fig. 2.26 (a) and in Fig. 2.27 (a).

To interpret this phenomenon, let us postulate a principle of economy in the physiology underlying the specification of the groupings, and then of the attribute patterns. This principle may be regarded as a limited case of a more basic hypothesis: speakers minimize their effort consistent with
providing sufficient information in the messages to their listeners. If we assume that speakers expend roughly equal amounts of physiological effort for maintaining the $F_0$ contours near the baseline and the plateau, then the number of attributes associated with a phrase or a sentence may be regarded as a gross measure of the effort expended for the control of $F_0$. In short, more $F_0$ movements require more effort. The principle of economy, then, is interpreted such that the speakers tend to reduce the number of attributes, and equivalently, tend to group more words in a sentence.

When a phrase is composed of three lexical words, the speakers have a choice for grouping the phrase into either one of the two groups. If the entire phrase corresponds to one group, only the two attributes $R$ and $L$ are needed. In the case of the right-branched structure, the assignment of either $R_1$ or $P$, regarding the two last words (for instance, 'yellow bus' and 'school bus') as a subgroup, is sufficient to obtain the contrastive pattern, costing the speakers three attributes in all. In the case of the left-branched structure, on the other hand, there is no way to specify that structure. Thus, the first two words, for instance 'light yellow' and 'small school' must be grouped (not subgrouped). Such grouping requires at least two pairs of $R$ and $L$ for the phrase, which requires somewhat more effort to produce than
the grouping of the entire words in each phrase. It may be stated, therefore, that the speakers tend to group the three lexical words regardless of whether the right or the left-branched structure is intended.

However, the principle of economy alone does not explain fully why the speakers have developed such a habit. Perhaps, the following two facts must be related to this phenomenon. First, in speech, the meaning of such ambiguous phrases is determined uniquely from the context of speech or even from the environments in which people are speaking. Thus speakers do not need to disambiguate the phrases by intonation, or more generally, the prosodic factors in speech. Lieberman (1967) has noted that the disjunctural contrast is overridden by the context of the entire sentence. Similar phenomena may occur in the case of contrast by attribute patterns. In such circumstances, the specification of the structure is regarded as redundant, and thus the principle of economy must dominate for the determination of the grouping. Secondly, a speaker has the freedom to compose phrases with various degrees of preciseness. For instance, if the context does not contain the necessary information to disambiguate the phrases, he may compose unambiguous phrases, such as 'the bus with light yellow color', instead of 'light yellow bus'. Because of these factors, perhaps speakers are not
used to dividing these phrases into two groups. Therefore, when asked to utter such phrases while specifying the two different structures, they can create two different patterns depending on the structure, but often the observed patterns for the left-branched structure are not contrastive, since the speakers have the habit of grouping all the lexical words in the phrase.

2.4.5 Prepositional Phrases and Short Sentences

In the previous section, we have shown that the attribute patterns of the noun phrases composed of lexical words are specified if the groupings and the subgroupings of the words in the phrases are given. The groups and the subgroups often correspond to the constituents of the phrases. The linguistic factor, therefore, is said to determine primarily the attribute patterns, although the principle of economy in the physiology interacts constantly in the determination of the patterns. In this section, we shall investigate briefly the assignment of the attribute to function words, specifically prepositions, and to single verbs in short sentences.

The schematized \( F_0 \) patterns superimposed on the original \( F_0 \) contours for the phrase, "...the dog in the mud", at the end of S1, are shown in Fig. 2.28. The three speakers KN, JP and KS produce the same attribute patterns, as follows:
Figure 2.28

The $F_0$ contours and the corresponding schematized $F_0$ patterns and the amplitude envelope (A.E.) of the phrase '.... the dog in the mud,' in S1, read by KN in (a), by JP in (b) and by KS in (c).
The dog in the mud.

(a) KN Sl

(b) JP Sl

(c) KS Sl

Fig. 2.28
The function words "in the" correspond to the plateau portion of the "hat-pattern". The pattern can be generated by assuming, for instance, the following subgroupings:

\[(2.69) \ ( \text{the} \ ) \ ( \text{dog} \ ( \text{in} \ ( \text{the} \ )) \ )\]

The \( F_0 \) contours and the corresponding schematized patterns for the phrase "the dog in the mud in the park" are shown in Fig. 2.29 for the three speakers. These patterns are considerably different depending on the individual speaker. The discrete representation of the schematized \( F_0 \) pattern for each of the three speakers may be described as follows:

\[(2.70) \ "\ldots \text{the (dog in the (mud in the park))}" \quad \text{(KN) S3}\]

\[(2.71) \ "\ldots \text{the (dog in) the (mud) in the (park)}" \quad \text{(JP) S3}\]

\[(2.72) \ "\ldots \text{the (dog in) the (mud in the park)}" \quad \text{(KS) S3}\]

The parentheses in the above expressions indicate the groupings and the subgroupings of the phrase based on the observed attribute patterns. In the case of the pattern shown in \[(2.70)\), the corresponding \( F_0 \) contour is shown in Fig. 2.29 (a); the entire phrase corresponds to one group. The assignment of RI on the word "mud" indicates that the phrase is
The $F_0$ contours, the corresponding schematized $F_0$ patterns and the amplitude envelope (A.E.) of the phrase '...the dog in the mud in the park,' in S3, read by KN in (a), by JP in (b) and by KS in (c).
the dog in the mud in the park.

(a) KN S3

(b) JP S3

(c) KS S3

Fig. 2.29
interpreted to have a right-branched structure. The phrase read by KS, shown in (2.72), also exhibits the right-branched structure, since the noun phrase "mud in the park" is grouped by putting R on "mud" and L on "park".

It should be noticed that the major syntactic / semantic boundary should appear between the two words "dog" and "in", as follows:

(2.73) [ The dog [ in the mud [ in the park ]pp]pp]NP,

where PP is the symbol used for designating "prepositional phrase". In the actual grouping in (2.72), however, the preposition "in" is rather grouped with the previous lexical word "dog". A similar phenomenon can be observed in the pattern (2.71) for JP. The word 'dog' received R and L for the above two cases. Probably there are two possible interpretations which would explain this phenomenon. One is that the preposition 'in' is grouped with the previous lexical words so that the preposition receives L for the stress-marking without extra cost in physiological effort. The function words, such as articles and prepositions, do not receive the attributes, except when these words are emphasized. We understand, however, that L can be assigned to these words whenever the realization does not require some extra effort. The second possibility is that the F₀ lowering may be considered to occur
at any place, either in the lexical words, or in the function word, and its purpose is to prepare the next $F_o \text{ rise.}$

In any case, the lexical words play an active role in determining the attribute patterns, while the function words are said to have a passive role. The following examples may be also considered as the manifestation of such roles in the grouping process. In Fig. 2.30, we present the $F_o$ contours and the schematized pattern for the phrase "the dog in the yellow mud" read by the three speakers KN, JP, and KS. The discrete representation may be described as follows:

(R) L R L (2.74) "... the dog in the yellow mud" (KN) S7

(R)L R L (2.75) "... the dog in the yellow mud" (JP, KS) S7

The three speakers indicate a quite similar pattern. The noun phrase "yellow mud" is grouped and the remaining lexical word "dog" forms a group with the following preposition "in" for KN, or the word forms a group by itself for JP and KS. We have noted, in Section 2.4.2, that any noun phrase composed of an adjective and a noun is almost always grouped. This fact explains why the three speakers have shown quite similar patterns for this noun phrase. Since the two last lexical words are grouped, not much choice is left in terms of the grouping of the remaining words. In the phrase in the previous examples, on the other hand, the bond between the
Figure 2.30

The $F_o$ contours and the corresponding schematized patterns and the amplitude envelope (A.E.) of the phrase '...the dog on the yellow mud,' in S7, read by KN in (a), by JP in (b) and by KS in (c).
(a) KN S7

(b) JP S7

(c) KS S7

Fig. 2.30

...the dog in the yellow mud.
lexical words in terms of the meaning is less strong, causing
the various patterns for the individual speakers, as shown
in (2.70), (2.71) and (2.72). There seems to exist an hier-
archical order in the manner of the grouping such that each of
the most closely related words in terms of the meaning is grouped
first, and the grouping of the remaining words depends on the
situation created by the first groupings of the words.

The following observation of the patterns associated
with short sentences may make this point more clear. The
schematized $F_0$ patterns superimposed on the original $F_0$
contours for the four sentences read by KS are shown in Fig. 2.31.
The corresponding attribute patterns may be represented as
follows:

$P_{BL} R L$ (2.76) 'The (dog likes - the elephant)' (KS)

$P$ $BL R (L) (R) L$ (2.77) 'The (dog likes) the (cat on the tree)'

$P_{BL} R R L R L$ (2.78) 'The (small black cat) (likes the dog)' (KS)

$P_{BL} R L R Ø L$ (2.79) 'The (white dog) likes the (small black cat)'

where the parentheses represent the groupings determined on
the basis of the observed attribute patterns. The sentences
The $F_0$ contours, the schematized patterns and the amplitude envelope (A.E.) of the four sentences read by KS. The $F_0$ contour for the verb 'likes' can be located on the middle of the plateau portion (in [a]), at the offset of the plateau portion (in [b]), in the rising portion (in [c]), and near the baseline (in [d]).
The dog likes the elephant.

Fig. 2.31 (a) and (b)
The small black cat likes the dog.

(d) KS S11

The white dog likes the small black cat.

(c) KS

Fig. 2.31 (c) and (d)
in (2.76), (2.77) and (2.78) were used in the preliminary study and are not listed in Table 2.1.

Let us observe how the verb, "likes" is grouped with the neighboring words. In the case of the sentence in (2.76), the entire sentence corresponds to one group. The verb is located on the plateau portion of the schematized pattern. In the second sentence, (2.77), the verb is grouped with the subject, but in the third sentence, (2.78), with the object 'the dog'. In the last example, in (2.79), none of the attributes is assigned to that verb. The $F_o$ contour for the verb in Fig. 2.31 (d) is located near the baseline. How can this observation be interpreted? Obviously, the syntactic structure of the sentence cannot deal with this problem, since, if a grammar analyzes the sentence into a subject noun phrase and a verbal phrase, then the grouping in (2.78) may be specified on the basis of such analysis, but not the grouping in (2.77). We must look, therefore, for another process that determines the groupings of the words in these sentences. As described before, it is probable that the words that are grouped first are most closely related to each other in terms of meaning. Thus "cat on the tree", "small black cat" and "white dog" are grouped first. Then, the remaining words such as "dog" and "likes" in (2.77), and "likes" and "the dog" in (2.78) are grouped, perhaps according to an economical manner of stress-marking in terms of physiological effort. Because
of the economical stress-marking, such grouping may include only two lexical words. In the case of the sentence in (2.79), the verb is isolated, and then the deletion rule, Rule 3, may be applied.

In summary, it may be stated that the grammatical factor, specifically the constituent structure, dominates in determining the attribute patterns for closely related words in sentences (such as the grouping of the words composing a noun phrase, as described previously). However, the principle of economy probably dominates for the generation of attribute patterns for the remaining successive words in sentences. In the generation of the stress patterns using NSR and CR, the rules are applied cyclically until the level of the entire sentence is reached, according to its syntactic structure. In the derivation that we have presented, on the other hand, the cyclical operation of the rules, Rule 4 and Rule 7, is blocked at the level of the groups; for instance, a noun phrase, which is marked by means of R and L, cannot be grouped at a higher level. It should be noticed that the rules can operate cyclically only over the subgroupings. Bierwisch (1968) has pointed out the necessity of the blocking of the operation of NSR and CR. We must recognize that the influence of the constituent structure upon the specification of the attribute patterns is only a localized phenomenon. However,
the constituent which corresponds to the group varies from time to time and from one speaker to another. Probably, this variation is due, at least partially, to the various semantic interpretations of the sentence by the individual speakers.

We have postulated, this far, two factors for the determination of the grouping of the words in the sentence (and consequently the attribute patterns): the localized grammatical structure, and the principle of economy. The emphasis of one or more words in a sentence is still another factor, and this will be investigated in the following chapter.

2.5 Summary of This Chapter: A State Transition Network Representation of the Attribute Pattern

We have shown in this chapter that the $F_0$ contours of the sentences are well characterized by using the five attributes, BL (baseline), R (rise), L (lowering), P (peak) and Rl (a rise on the plateau). BL represents the gradual $F_0$ fall along the entire sentence, specifically the breath-group. BL, therefore, is said to be supersegmental. Since the rest of the baseline occurs at the onset of each breath-group, we associate the symbol BL with the onset of the breath-group. The remaining four attributes characterize the localized $F_0$ movements. The attributes R, P (which often occurs simultaneously with R), and L are assigned on stressed syllables.
These attributes, then, are regarded as segmental. However, seems to be supersegmental, in the sense that the corresponding $F_0$ rising contour spreads over more than one syllable.

The sequence of the attributes associated with sentences seems to be imposed by a strong constraint, which is well described in Fig. 2.2, where the structure of the schematized patterns is defined. In short, the $F_0$ rise (i.e. R) occurs only from the baseline to the plateau, and the lowering (i.e. L) in the reverse manner. This structure can be illustrated in terms of the attributes by using a simple state transition network as shown in Fig. 2.32. The network accepts any observed sequence of attributes. State 1 corresponds to the baseline, while State 2 corresponds to the plateau. The existence of State 0 is not so evident in the $F_0$ contours. We can observe only a rise of the baseline at the onset of each breath-group. Observations of the laryngeal activities described in Chapter 3 suggests that there is transitions between a rest state and a phonetory state. In this respect, State 0 may be assumed to correspond to the rest state. We assume that the state changes automatically from State 1 to State 0 at the offset of the breath-group.

We have demonstrated in Section 2.4 that the eight rules listed in Table 2.7 can generate any observed patterns. In fact, the rules are determined such that any generated pattern
A state transition network accepting any observed sequence of the attributes associated with a sentence. State 0 and State 1 may be regarded as the starting state and the accepting state, respectively, for declarative sentences.
Fig. 2.32
is accepted by the network. The application of the rules requires the groupings and the subgroupings of the words in a sentence. Clearly, the onset and the offset of each group correspond to a change in state: from State 1 to State 2, and from State 2 to State 1, respectively. The onsets of the subgroupings correspond to the self-loops of State 2, associated with $P$, $R_l$, and $\emptyset$. (the $P$ loop can be produced at the onset of the group, since $P$ can occur simultaneously with $R$).

The groups and subgroups which produce the observed attribute patterns using the eight rules often reflect the underlying constituent structures. It is stated, therefore, that the attribute representation of the $F_0$ contours is meaningful linguistically, in the sense that it carries certain linguistic information contained in speech. However, we must recognize that the linguistic factor is not the only factor determining the attribute patterns. The other factor, the principle of economy in physiological efforts, seems to influence the organization of the patterns as well. These factors determine the groups and the subgroups of the words in a sentence. In the generation of the patterns, every word is assumed to have intrinsically at least one pair of the attributes $R$ and $L$, and then the rules are applied depending on the manner of the groupings and subgroupings of the words.
Application of any rule (except Rule 8) reduces the total number of attributes associated with the sentence. Thus, for either linguistic or physiological reasons, the grouping of the words reduces the amount of effort required for the realization of the attribute patterns in the $F_0$ contour. Therefore, it may be said that the principle of economy always underlies the generation of the attribute patterns.

We have noted in Section 2.1 that the traditional level notation system, such as High and Low, is related to our representation. The relation is made more clear in the state transition network shown in Fig. 2.32. Let us assume that the register Low corresponds to State 1 and High to State 2. The attributes $R$ and $L$, then, can be represented as Low-High and High-Low, respectively. The null attribute $\emptyset$ associated with the self-loop for State 1 and for State 2 can be regarded as Low and High, respectively. The attributes $P$ and $R_1$, however, cannot be related to the level representation in a simple manner, since $P$ and $R_1$ are not distinguished by a different level, but by a different shape of the corresponding $F_0$ contours. It will probably be possible to describe $P$ and $R_1$ using the level representation, by increasing the number of levels. However, we speculate that an essential difference exists between the two notational systems for describing American English intonation. Our study has shown that the attribute $R$ (including $(R)$, i.e. an invisible $F_0$ rise) and
attribute L are consistently related to stressed syllables. Thus R and L can be regarded as a phonetic manifestation of stress, that the speaker produces and the listener interprets. The levels, for instance, High, on the other hand, cannot be related directly to stress. If such a case is assumed, then every syllable corresponding to the plateau (which is represented by State 2) must be regarded as stressed, which is unrealistic. Thus, although both notational systems may be equally sufficient for describing the \( F_0 \) contours, the attribute representation seems to reflect more directly the underlying mechanisms of intonational phenomena. Further, in a practical sense, the attribute notation can be said to be more efficient than the level notation for describing \( F_0 \) contours of American English.
Chapter III  Physiological Correlates of the Attributes

The primary objective of this chapter is to investigate the underlying mechanisms that generate the characteristic movements corresponding to the attributes. When we discussed a principle of economy in physiology, the number of attributes specified within a phrase or a sentence was regarded as a gross measure of the physiological effort required for the realization of the $F_0$ pattern. In that discussion, we implicitly assumed that each attribute is related to certain physiological activities. A straightforward step to be taken next, then, is to determine how the attributes are correlated with physiological activities during speech. If the attributes are truly elementary units which specify the intonation, we should find a consistent relationship between the attributes and some activities in physiology.

3.1 Studies on the $F_0$ Control in Speech: Background

In the past, a vast number of studies investigating the manner of the regulation of $F_0$ during speech and during singing, has been undertaken. The primary factors that determine the $F_0$ values are the subglottal air pressure ($P_s$), or more correctly the air pressure drop across the glottis, and the states of the vocal folds. The vocal-fold states may be defined as the mechanical parameters that determine the manner of ascillation of the folds. For instance, the stiffness and
mass of the folds, and the degree of spreading of the glottis may be considered as those states (Stevens, 1975). The vocal-fold states are controlled primarily by participation of the intrinsic muscles, such as the cricothyroid muscles (CT), the vocalis muscles (VOC) and the lateral cricoarytenoid muscles (LCA). Their locations in the larynx are depicted schematically in Fig 3.1 (a). The extrinsic laryngeal muscles, such as the sternohyoid muscles (SH), the sternothyroid muscles (ST) and the thyrohyoid muscles (TH), which suspend the larynx, as shown in Fig. 3.1 (b), are also responsible for controlling the states, although considerable controversy surrounds the issue of the role of these extrinsic muscles in controlling F0.

In order to obtain some perspective into the physical correlates of the attributes, the relative importance of the two factors, P5 and the vocal-fold states, in regulating F0 of voice must be evaluated. A measure of the sensitivity of F0 variation due to a change in P5 is defined as the rate of change in F0 with respect to P5 (r.f.p.). Ladefoged (1963) has shown that the value of r.f.p. is about 5 Hz/cmH2O. In his experiment, the subject attempted to sing a steady note (about 95 Hz), while one of the experimenters pressed against his chest at unpredictable moments, to vary P5. Similar experiments were undertaken by Öhman and Lindqvist (1966), and by Fromkin and Ohala (1968). The first experimenters have estimated the rate of change of fundamental period with P5 to
Figure 3.1

Schematic drawings of the laryngeal cartilages and the intrinsic laryngeal muscles in (a), and the larynx and the extrinsic laryngeal muscles in (b).
be a constant and equal to about \(-0.16 \text{ msec/cmH}_2\text{O}\), independently of the initial values of \(F_0\) and of the pressure. This result indicates that r.f.p. depends on \(F_0\) values: the value of r.f.p. would be greater in higher \(F_0\). In the range from 70 Hz to 150 Hz, the r.f.p. value is about 2.5 Hz/cmH$_2$O. Fromkin and Ohala (1968) noted that r.f.p. is greater in flasetto (roughly 7 Hz/cmH$_2$O) than in normal chest voice (about 2.5 Hz/cmH$_2$O). Liebermean, Knudson and Mead (1969) have measured r.f.p. values by modulating sinusoidally the oral pressure. They estimated the r.f.p. value to be between 3 and 18 Hz/cmH$_2$O, depending on the average \(F_0\) value. Another group has conducted a similar experiment and found for r.f.p. to be about 2 to 4 Hz/cmH$_2$O for chest voice (Hixon, Klatt and Mead, 1971).

The r.f.p. value is commonly assumed to be about 5 Hz/cmH$_2$O, or probably less, for normal speech in chest voice. In non-emphatic utterances, \(P_s\) falls gradually along a sentence, and its magnitude is less than, say 3 cmH$_2$O, as shown, for instance, by Atkinson (1973). The rapid localized \(F_0\) movements corresponding to R, P and L cannot be caused by the \(P_s\) variation. On the other hand, the \(P_s\) fall might account for the baseline fall, i.e., the gradual fall in the \(F_0\) contour along the entire sentence. Collier (1975) claimed that the declination line (i.e., the baseline) is correlated with this \(P_s\) fall. However, our analysis of the \(F_0\) contours described in Chapter 2 indicated that the magnitude of the baseline fall is between 20 Hz to
40 Hz depending on the individual speakers. If the values of Hz/cmH₂O for r.f.p., and 3 cmH₂O for the Pₛ fall are assumed, the Fₒ drop due to Pₛ would be 15 Hz. The value 15 Hz may be considered as an upper bound for the influence of Pₛ, and in fact, the contribution Pₛ is probably less than that value. In any case, it is obvious that the Pₛ fall alone cannot account for the fall of the baseline. We speculate, therefore, that not only Pₛ, but also some factor, which apparently is related to the vocal-fold states, must be involved in the generation of the non-localized Fₒ component.

The regulation of the vocal-fold states is often investigated in terms of the electromyographic (EMG) activities of the laryngeal muscles. In order to interpret such EMG activities meaningfully, basic knowledge concerning the control mechanisms for the states, is needed. Those mechanisms, however, are rather poorly understood (except that of CT), causing a great deal of controversy in the interpretation of the EMG data, particularly for the extrinsic laryngeal muscles.

A rise in Fₒ is known to be accomplished by contraction of CT assisted by VOC and LCA, and so on. The Fₒ rising mechanism based on the CT contraction is rather simple; the contraction of CT causes rotation of the cricoid cartilage at the cricothyroid joints with respect to the thyroid cartilage (See Fig. 3.1 [a], for the geometrical relation of the muscles, the
cartilages and the joints). This rotational movement apparently causes a lengthening of the vocal-folds, which results in an increase in the stiffness and then a rise in $F_0$. A positive correlation between the vocal-fold length and the $F_0$ value was found for singing by Damste, Hollien, Moore and Murry, (1963), by Hollien, Brown and Hollien (1971), and by Hollien (1974). Many EMG experiments, for instance, conducted by Shimada and Hirose (1971), by Ohala (1970), by Atkinson (1973) and by Collier (1975), indicated the CT participates consistently in an $F_0$ rise and an $F_0$ peak.

However, it is not so well understood how $F_0$ is lowered. Lieberman (1970) and Collier (1975) have claimed that an $F_0$ lowering is simply due to relaxation of the muscles for which contraction causes the $F_0$ rise, suggesting that the CT activities are also responsible for $F_0$ lowering. In other words, they postulated a passive control mechanism in the $F_0$ lowering. According to Ohala (1970) and Ohala (1972), however, $F_0$ is lowered both by the passive mechanism and by the active contraction of some laryngeal muscles, especially SH. We shall attempt to show, in a theoretical study using a simple muscular model, that an active lowering mechanism must be assumed to account for the results of measurements of the physical properties of the localized $F_0$ movements; for instance, the duration of the $F_0$ rise and of the $F_0$ lowering are about equal, as described in Section 2.3.3.
It is a common observation that the larynx moves up and down during speech. This movement seems to be related to a large variation of $F_0$. Hollien and Curtis (1962) have suggested a positive correlation between a degree of elevation of the larynx and an increase in $F_0$, for singing. Vanderslice (1967) found a similar correlation in utterances consisting of three consecutive words, using a so-called cri-cothyrometer. Kakita and Hiki (1974) found also, for isolated Japanese words, that a rise and a fall of $F_0$ are well correlated with the upward and downward movements of the larynx, respectively.

It is evident that the laryngeal movements involve participation of the extrinsic laryngeal muscles. However, the reason for the correlation between laryngeal height and the $F_0$ value is not sufficiently understood. Sonninen (1968) has proposed an "external frame function," which means the participation of extrinsic laryngeal musculature in the control of length and, in turn, the vocal-fold stiffness. The coordinated activity of ST, and the thyrohyomandibular muscle chain (which is termed a "functional chain" by Zenker [1960]) produces a resultant force that causes not only a vertical movement of the thyroid cartilage (a change in the laryngeal height), but also a horizontal movement (posterior to anterior) of the thyroid cartilage. In this interpretation, a loose joint connecting the thyroid and cricoid cartilages must be assumed so that the two cartilages can slide as well as rotate.
This horizontal movement of the thyroid with respect to the cricoid cartilage may vary the vocal-fold length, and consequently the $F_0$ value. In such mechanism, horizontal movement (toward anterior) of the cricoid cartilage must be assumed to be prevented. Sonninen (1968) postulated a cricopharyngeal muscle activity for pulling the cricoid cartilage dorsocranially. This external frame function theory, however, has been postulated for pitch regulation during singing, in particular, for explaining the production of extremely high notes. It should be noticed further that the external frame function does not account for the active $F_0$ lowering.

Another possible interpretation of the positive correlation between the vertical laryngeal movement and the $F_0$ variation has been proposed by Ohala (1972) and Stevens (1975). In their speculation, the vertical movement of the larynx causes directly a change in the vertical stiffness, not in the anterior-posterior (horizontal) stiffness, of the vocal folds. Baer (1975) postulated, on the basis of observations of excised larynxes in oscillation, that the vertical stiffness of the surface membrane of the folds can be regarded as one of the states that determine the oscillatory behavior of the vocal folds. A speaker might utilize this mechanism for controlling the $F_0$ rise and the $F_0$ lowering; however, this does not mean that no other active mechanism exists in $F_0$ lowering. For instance, a possible mechanism in which the lowering
of the larynx causes a shortening of the folds, and thus a lowering of $F_o$ will be postulated in this chapter. In such a mechanism, the horizontal stiffness is controlled actively both for $F_o$ rise and for $F_o$ lowering.

Action of the extrinsic laryngeal muscles can be observed in terms of the EMG activities. The correlation between the $F_o$ contour and the EMG activity patterns is rather complicated in comparison with that for the CT activities. This complexity of the relation is presumably due to the fact that the extrinsic muscles participate in controlling $F_o$ as well as in certain segmental speech gestures. Ohala and Hirose (1969) noted that SH are active for jaw lowering and tongue retraction. Data shown in Collier (1975) indicate a consistent EMG peak activity of SH during /k/ in a sentence which is uttered with a variety of intonations. On the basis of these facts, some authors concluded that participation of the extrinsic laryngeal muscles, in particular SH, in $F_o$ control was regarded as negligible (Lieberman, Sawashima, Harris and Gay, 1970; Lieberman, 1970; Collier, 1975). There seems to be no reason to assume, however, that the extrinsic muscles are only active for the segmental speech gestures. In fact, Ohala (1972) has demonstrated participation of the SH in $F_o$ lowering, independent of their activity in other speech gestures.

The problems of $F_o$ control so far may be summarized in
the following three questions; 1) How is the baseline generated? 2) What are the mechanisms for the localized $F_0$ movements, especially of the $F_0$ lowering? 3) What is the role of the extrinsic laryngeal muscles in the $F_0$ control? In order to obtain a deeper understanding of these problems, we have conducted two physiological experiments. The first experiment is a direct measurement of the laryngeal movement during speech using a cineradiographic technique. In the second experiment the EMG activities of the intrinsic and the extrinsic laryngeal muscles are recorded simultaneously.

3.2 Laryngeal Dynamics During Speech

3.2.1 Procedure

The subject was one of the previous speakers, (KS). The cineradiographic data were taken at the Cardiac Catheterization Laboratory of the Massachusetts General Hospital in Boston, Massachusetts. In the experiment, the lateral x-ray motion picture (35 mm) with a frame rate of 35 frames/sec and the speech signals were recorded simultaneously. A simple device was used for synchronization of the movie frame, and the speech signals which were recorded on audiotape. The speaker read only four independent sentences consecutively, S60, S62, and S76, listed in Table 3.1, and S15 listed in Table 2.1, so that total dosage of x-ray radiation to the subject would not exceed 2 Roentgen. In order to minimize movement of the
Table 3.1

Sentences used in the physiological experiments.

(*Note: A speaker is instructed to emphasize the word spelled in capital letters.)

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S54</td>
<td>Ken raises sheep</td>
</tr>
<tr>
<td>S55</td>
<td>The farmer raises sheep</td>
</tr>
<tr>
<td>S56</td>
<td>All farmers raise sheep</td>
</tr>
<tr>
<td>S57</td>
<td>Almost all farmers raise sheep.</td>
</tr>
<tr>
<td>S58</td>
<td>The farmer raises yellow sheep.</td>
</tr>
<tr>
<td>S59</td>
<td>All farmers raise yellow sheep.</td>
</tr>
<tr>
<td>S60</td>
<td>Almost all farmers raise yellow sheep.</td>
</tr>
<tr>
<td>S61</td>
<td>Ken raises the great yellow sheep.</td>
</tr>
<tr>
<td>S62</td>
<td>Ken raises the light yellow sheep.</td>
</tr>
<tr>
<td>S63</td>
<td>In the house, Bill drinks a beer.</td>
</tr>
<tr>
<td>S64</td>
<td>Bill drinks a beer in the house.</td>
</tr>
<tr>
<td>S65</td>
<td>Bill drinks a beer in the box.</td>
</tr>
<tr>
<td>S66</td>
<td>I like the cat in the park on the hill.</td>
</tr>
<tr>
<td>S67</td>
<td>Bill meets Steve.</td>
</tr>
<tr>
<td>S68*</td>
<td>Bill meets Steve.</td>
</tr>
<tr>
<td>S69*</td>
<td>Bill MEETS Steve.</td>
</tr>
<tr>
<td>S70*</td>
<td>Bill meets STEVE.</td>
</tr>
<tr>
<td>S71</td>
<td>You see the bill.</td>
</tr>
<tr>
<td>S72</td>
<td>You see the pill.</td>
</tr>
<tr>
<td>S73</td>
<td>You see the dill.</td>
</tr>
<tr>
<td>S74</td>
<td>You see the till.</td>
</tr>
<tr>
<td>S75</td>
<td>You see the goat.</td>
</tr>
<tr>
<td>S76</td>
<td>You see the coat.</td>
</tr>
<tr>
<td>S77</td>
<td>I like the cat in the tree in the park.</td>
</tr>
</tbody>
</table>

207
speaker's head, a head rest was used during the recording.

In the measurement, each frame of the x-ray movie is projected onto a plain paper using photographic enlarger for tracing of such items as the mandible, anterior portion of the hyoid bone, the laryngeal ventricle and a calcified portion of the thyroid cartilage, as shown in Fig. 3.2. A scale which was fixed to the shadow of the intensifier edge served as reference for the measurements of the three points A, B and C in Fig. 3.2, corresponding to vertical positions of the thyroid cartilage, the hyoid bone, and the mandible (jaw), respectively. Anterior-posterior length of the ventricle, indicated by "$l" in Fig. 3.2, was measured as an indication of the vocal-fold length. The absolute values were calculated by reference to the shadow of the microphone (of known diameter) which was located near the midsagittal plane of the subject. It was quite unfortunate that the cricoid cartilage was totally invisible in the x-ray picture. The measurements of the cricoid and the thyroid movements would have provided useful information concerning the states of the vocal folds and their controlling mechanisms.

A few comments should be made regarding error in the measurements. The distance between the x-ray anode and the median plane of the subject's head was about 300 cm, and the distance between the median plane and the intensifier surface was about 10 cm. Coma distortion of the images is considered to be
Figure 3.2

An example of the lateral x-ray tracing. The three points A, B and C indicate the vertical positions of the thyroid cartilage, of the hyoid bone, and of the mandible, respectively. The scale is fixed with respect to the outline of the frame. The laryngeal ventricle length is indicated by "\( \ell \)."
negligible except near the edge of the circular image. Most of error seemed to occur during the tracings, in particular for the laryngeal ventricle. A jitter seen in sequences of the points which represent the vertical movements of the three points A, B and C was less than ± 1 mm. We speculate, therefore that the error in the measurements is most likely about ± 1 mm. On the other hand, the measurements of the ventricle length may contain more error than the ventricle measurements. This large error is presumably due to the fact that the posterior edge of the ventricle often cannot be seen clearly on the x-ray pictures.

3.2.2 The Vertical Movements of the Larynx

The results for the vertical movements of the thyroid cartilage, of the hyoid bone and of the mandible are shown in Fig. 3.3, for S60 in (a), 262 in (b), S76 in (c) and S15 in (d), respectively. In each figure, the curve from the top to the bottom represents the movement of the mandible marked by "JAW", of the hyoid bone marked "HYD", of the thyroid cartilage marked by "HYD", of the thyroid cartilage marked by "TYD" and the corresponding $F_O$ contour and the amplitude envelope marked by "A.E.", respectively. The curves representing the vertical movements are drawn by smoothing visually each sequence of the data points plotted for frame by frame.

The following remarks can be made concerning those measurements.
The movements of the mandible marked as "JAW," the hyoid bone marked as "HYD" and the thyroid cartilage marked as "TYD," and the corresponding F₀ contour and the amplitude envelope (A.E.). The four sentences, S60 in (a), S62 in (b), S77 in (c) and S15 in (c) were read by the speaker KS.
Almost all farmers raise yellow sheep.

Fig. 3.3 (a) KS S60
Fig. 3.3 (b) KS S62

JAW (in mm)

HYD (in mm)

TYD (in mm)

A.E.

Ken ral ses the light yellow sheep.
Fig. 3.3 (c) KS S77

JAW (in mm)

HYD (in mm)

TYD (in mm)

A.E.

I like the cat in the tree in the park.
The dog likes the enormous gorilla.
1) The movements of the larynx and those of the mandible, shown in each figure of Fig. 3.3, are not correlated significantly with each other. The movements of the mandible and of the thyroid cartilage seem independent of each other.

2) The movement of the thyroid cartilage is correlated but only partially, with that of the hyoid bone, except at the onset and at the offset of each sentence. Generally, the height of the hyoid bone seems to be influenced greatly by the tongue positions, while that of the larynx (actually, the thyroid cartilage) is affected much less from these segmental gestures. The magnitude of the localized movement of the hyoid bone is about 10 mm, while that of the larynx is roughly 3 mm with a few exceptions. It is rather surprising that the movements of the three articulatory structures, the mandible, the hyoid bone and the larynx, are fairly independent of each other, and further that the laryngeal position is well established, in spite of the muscular connection between the mandible and the hyoid bone and between the hyoid bone and the thyroid cartilage that is connected in turn to the sternum by ST (the sternothyroid muscles). The stabilization of the laryngeal height is clearly seen in Fig. 3.3 (b). It must be assumed, therefore, that the extrinsic laryngeal muscles, which suspend the larynx and the hyoid bone, participate in the stabilization of the laryngeal position as well as in
other segmental speech gestures and, perhaps, in the control of \( F_0 \).

Only a partial correlation is found between the vertical laryngeal movements and the corresponding \( F_0 \) contours.

3) Clearly, the initial \( F_0 \) rise corresponding to the attribute R is always accompanied by a rise in the larynx for all four sentences. It should be noticed, at the beginning of each sentence, that the larynx is raised from a rest position, and then raised again from the point where the initial \( F_0 \) rise occurs. This hesitation in rising of the larynx at the beginning of a sentence is seen clearly in the two sentences S60 in Fig. 3.3 (a) and S15 in Fig. 3.3 (d). The first rise in the laryngeal height can be regarded as a transition from the rest position to the phonatory position. This transition presumably corresponds to the reset of the baseline that occurs at the onset of each breath group.

4) The final \( F_0 \) lowering corresponding to L is accompanied by a fall in the laryngeal height, although some segmental influences due to the final consonants, such as /p/ and /k/, are observed in Fig. 3.3. The lowering in the laryngeal height at the sentence's final position, continues beyond the offset of the phonation. This further lowering may be regarded as a transition from the phonatory position to the rest position. The two transitions, the rest to the phonatory position and the phonatory to the rest position, can be regarded
as the state transitions, State 0 to State 1, and State 1 to State 0, in the network shown in Fig. 2.32.

5) Each lowering, L located in the middle portion of a sentence is accompanied by a lowering in the laryngeal height, except for the "invisible" lowering, i.e. (L), in S62 shown in Fig. 3.3 (b). The magnitude of the laryngeal lowering in the middle portion of the sentences is somewhat smaller than that at the final position. It should be noticed that the curve representing the height of the hyoid bone is also lowered during L in some extent.

6) On the other hand, the $F_0$ rise, R in the middle positions in the sentence is not correlated with a specific movement of the larynx.

7) The overall configuration of the laryngeal movement seems to be correlated with that of the amplitude envelope rather than the $F_0$ contour. The amplitude envelopes exhibit peaks for every word. A gross change in the magnitude of these peaks along each sentence is roughly correlated with the vertical movements of the larynx. For instance, the thyroid movement shown in F. 3.3 (b) for S62, is correlated with the amplitude envelope better than the corresponding $F_0$ contour that gradually falls along the sentence. In this example, the magnitudes of the peaks in the amplitude envelope are kept fairly constant until just before the final word, "sheep, as the laryngeal height remains at a constant level.
Let us discuss briefly the mechanism that might create such correlation between the laryngeal height and the amplitude envelope. We speculate that the subglottal air pressure, $P_s$, is probably responsible for this correlation. Bouhuys, Mead, Proctor and Stevens (1968) found experimentally, for singing, that the acoustic intensity is proportional to $P_s$ cubed. If this is true for speaking mode, then the amplitude envelope may be regarded as an indication of $P_s$ variation, since the amplitude envelope is roughly proportional to square root of $P_s$ cubed. Hollien, Brown and Hollien (1971) suggested that vocal intensity regulation in modal register is aerodynamic in nature rather than a consequence of muscular control.

If 4 sq. cm of the cross-sectional area of the lower laryngeal cavity is assumed, than an upward force of $4P_s$ acts on the larynx ($P_s$ in dyens/cm$^2$). Since this upward force is proportional to $P_s$, the laryngeal height depends on $P_s$ assuming that all laryngeal muscular forces are in a state of equilibrium, except for certain segmental gestures. Thus, the amplitude envelope and the laryngeal height can be related to each other through $P_s$.

However, the amplitude envelope and the laryngeal movement are not well correlated at the beginnings of the sentences, for instance, during "almost" in S60 shown in Fig. 3.3 (a). The amplitude envelope indicates that $P_s$ has been built up at the onset of the utterance. Measurements of the $P_s$ variations during
speech reported in publications, for instance in Atkinson (1973), exhibit the build-up of \( P_s \) just before the onset of each utterance. It might be stated, therefore, that there is an opposition to the upward force on the larynx due to \( P_s \) which is about 40 grams). However, as described in Section 3.3.2, ST and SH, whose activities can prevent the rise of the larynx, are not particularly active. We may conclude that some laryngeal muscles are activated during the initial \( F_0 \) rise, resulting in a new equilibrium state corresponding to a higher laryngeal position. One may ask why this happens. Perhaps, the rise of the larynx after the hesitation occurs in order to assist the initial \( F_0 \) rise. This problem, however, remains an open question.

Some authors have found a good correlation between laryngeal height and \( F_0 \) (Vanderslice, 1967; Ohala, 1972; Kakita and Hiki, 1974) while we have seen only a partial correlation between these parameters. Those previous studies were conducted for isolated words, or lists of words. In such cases, both the \( F_0 \) contour and the laryngeal height would most likely exhibit a rise-fall pattern, since each word must correspond to a single breath group. As long as a word or a sentence corresponds to a single breath group and exhibits a single hat pattern, a good correlation between the height and \( F_0 \) will be observed.

In summary, our data (from one speaker) have shown that
F₀ lowering is accompanied by a lowering in the laryngeal height. The F₀ rise, however, is not correlated with the height, except for the initial F₀ rise in each sentence. Surprising to us, the laryngeal height does not always fall along each of the four sentences, as the F₀ contours fall. We had expected that the baseline, BL might be related to the gradual falling in the laryngeal height. However, this was not the case. The baseline turns out to be correlated with a gradual shortening of the laryngeal ventricle toward the end of a sentence, as described in the following section.

3.2.3 Variation in the Ventricle Length.

In Fig. 3.4, the variation of the ventricle length and the corresponding F₀ contours are presented for four sentences, S60 in (a), S62 in (b), S76 in (c) and S15 in (d), respectively. The dashed line in each figure, indicating the baseline, is determined such that its magnitude of fall within the sentence is equal to 32 Hz, and the line intersects with the terminal point of the F₀ contour. It should be noticed that the height of the F₀ plateau, say 30 Hz, and the F₀ value at each terminal point, about 90 Hz, are considerably higher than those (20 Hz and 80 Hz, respectively) described in Chapter 2. This is probably due to the fact that the speaker KS read the corpus loud, perhaps, to override the noise generated by the movie camera which was located near the subject.
Figure 3.4

The fluctuation of the laryngeal ventricle length and the corresponding $F_0$ contours for the four sentences, S60 in (a), S62 in (b), S77 in (c) and S15 in (d), read by speaker KS. The straight line superimposed on the dots in each figure indicates a least square error fit.
Almost all farmers raise yellow sheep.

Fig. 3.4 (a) KS 360
Fig. 3.4 (b) KS S62

General: The light yellow sheep.

Ken raises the light yellow sheep.
I like the cat in the tree in the park.

Fig. 3.4 (c) KS S77
Fig. 3.4 (d) KS SL5

The dog likes the enormous gorilla.
Since the data points contain a considerable amount of jitter, a detailed comparison of the length variation and the $F_0$ contour cannot be made. The general trend of the fluctuation of the ventricle length, however, may be investigated meaningfully.

The straight line superimposed on the dots representing the frame-to-frame variation in the ventricle length in each figure is determined by a least square fitting algorithm. The fitting algorithm is applied only to dots located inside the sentence marked by the two vertical lines. It is apparent that every straight line exhibits a negative gradient, indicating a gradual shortening of the length toward the end of each sentence. Notice that the ventricle length in S62 shown in Fig. 3.4 (b) is shortened gradually, while the corresponding laryngeal height shown in Fig. 3.3 (b) is kept remarkably constant along the sentence. As far as the four sentences are concerned, the baseline BL is correlated more consistently with the change in the ventricle length, and consequently with the vocal-fold length, than with laryngeal height.

It may be worthwhile to evaluate quantitatively the effect of a change in the ventricle length upon the $F_0$ contours. A number of authors have investigated the relationship between the length of the ventricle, or the vocal-fold length, and the $F_0$ values. Hollien and Moore (1960), Hollien, Brown and Hollien (1971). and Hollien (1974) conducted laryngoscopic
experiments for studying the fold-length vs. $F_0$ relationship, in singing notes. An x-ray technique was used by Kitzing and Sonesson (1967), and by Dämste, Hollien, Moore and Murry (1968) for measurement of the ventricle length as speakers sang different notes. Both techniques have provided similar results. A significant difference in the results depending on the two methods, is found only during abduction of the folds, i.e. during inhalation, in which the vocal folds appear relatively short on the x-ray pictures. We expect, therefore, that the actual length of the ventricles may be greater than the measured length during the non-speech portions in Fig. 3.4. More importantly, however, the measured ventricle length may be regarded as that of the vocal folds during phonation.

In general, the sensitivity or the rate of change in $F_0$ with respect to vocal-fold length (r.f.l.), increases rapidly with an elongation of the folds. This may be explained, at least partially, by the fact that the stress-strain relationship of the ligament, and perhaps, of the vocalis muscles exhibit a sigmoid shape (Van den Berg, 1960). The length-$F_0$ relation, however, seems to be roughly linear in a certain $F_0$ range. For instance, in Fig. 3.5, $F_0$ values are plotted as a function of vocal-fold length in the range from 80 Hz to 180 Hz, for two speakers, on the basis of data presented
Figure 3.5

Vocal-fold length vs. $F_0$ relationship; the data points for two male speakers, A (indicated by the closed circles) and B (indicated by the open circles) are provided by Hollien, Brown and Hollien, (1971). Each straight line represents a least square error fit for the individual speakers.
Fig. 3.5

F₀ (Hz)

10.1 Hz/mm

7.2 Hz/mm

VOCAL-FOLD LENGTH (mm)
in Hollien, Brown and Hollien (1971). Each of the two straight lines is determined in terms of least square error criterion for each speaker. The measured length-$F_0$ relation may be said to be a reasonable approximation to a straight line. The values of r.f.l. appear to be 7.2 Hz/mm for speaker A and 10.1 Hz/mm for speaker B. It is recognized that the r.f.l. values vary considerably depending on individual speakers. According to the published data, the r.f.l. varies from 7 Hz/mm to as high as 20 Hz/mm depending on the individual subjects. One of the causes for such intra-speaker variation is presumably individual difference in the vocal-fold length in the rest condition.

The magnitude of the shortening of the folds for the entire sentence, $\Delta l$ can be estimated from the straight line superimposed on the data points in each figure shown in Fig. 3.4. The magnitudes of $\Delta l$ appear to vary from 2.7 mm to 5.4 mm, depending on the sentence, and is 3.8 mm in the average for the four sentences. This large variation in $\Delta l$ is probably due to the noisy data points and, perhaps, due to localized components of the length fluctuation corresponding to the localized $F_0$ movements.

Let us assume $\Delta l$ to be equal to the average value, i.e. to 3.8 mm, and the magnitude of the baseline fall to be 32 Hz. Then, the r.f.l. value for our speaker is calculated as 8.4 Hz/mm, which compares favorably with the estimation data
shown in Fig. 3.5. In this calculation, however, the influence of the $P_s$ fall on the baseline fall is excluded. The actual r.f.l. value for this speaker, KS would be smaller than 8.4 Hz/mm. Since $P_s$ data for KS are not available to us, further examination of these questions must be deferred. It may be stated, however, with reasonable certainty that the gradual shortening of the vocal folds is the primary factor that specifies the baseline. In Section 3.4, we shall investigate a possible mechanism which causes this vocal-fold shortening along individual sentences.

3.3. EMG Activities of the Laryngeal Muscles during Speech

3.3.1 Procedure

The experiments were conducted at Haskins Laboratories, New Haven, Connecticut, for speaker KS and a new speaker TB. A corpus composed of twenty seven isolated sentences was used in the experiments. Data from twenty three of these sentences, listed from S54 to S75 in Table 3.1, are discussed in this chapter. Bipolar hooked-wire electrodes (Hirose, 1971) were inserted into each of the laryngeal muscles for detecting the EMG signals. Two sets of the 27 sentences written on cards were separately randomized, and the subjects read each sentence as the card was shown. The two sets of the cards were presented alternately eight times such that each sentence was read sixteen times in total. For the speaker KS,
the EMG signals from the intrinsic muscles, CT, VOC, and LCA, and the extrinsic muscles, SH and ST, were successfully recorded. For the speaker TB, however, only the signals from the intrinsic muscles, CT, VOC and LCA, were obtained. We shall, therefore, describe primarily the results for KS in this chapter.

The raw EMG data were processed using the Haskins Laboratories EMG Data System (Port, 1971; Port, 1973). Smooth curves representing the EMG activities of each muscle, which apparently correspond approximately to a force generated within the muscle (Bigland and Lippold, 1954), were obtained by integrating the raw EMG signals for each sentence over 10 msec time window and then averaging over 12 to 16 repetitions of the sentence. The same window length was used for each of the laryngeal muscles.

3.3.2 The attributes and the EMG activities

In Fig. 3.6, the $F_0$ contours (with the schematized $F_0$ patterns superimposed) and the corresponding EMG activities of CT, VOC, LCA, SH and ST are presented for the sentences S56 in (a), S58 in (b), and S60 in (c), read by the speaker KS. The $F_0$ contour in each figure is computed from one of the 12 to 16 repetitions of the sentence, and it is not an averaged $F_0$ contour. In order to compensate for the time delay of the effect of the EMG activities upon $F_0$, which is due to
Figure 3.6

$F_0$ contour and the corresponding EMG activities of the laryngeal muscles, CT, VOC, LCA, SH and ST for the three sentences, S56 in (a), S58 in (b) and S60 in (c), read by speaker KS.
Fig. 3.6 (a) KS S56
Fig. 3.6 (c) KS S60
a contraction time of the muscle, the EMG curves are shifted from 20 msec to 100 msec, depending on the identity of the muscle. The vertical lines located on the individual EMG curves represent the amount of such shift from the thick vertical line across all the curves. The amount of the shift, 60 msec for CT, 20 msec for VOC and LCA, and 100 msec for SH and ST, is estimated in reference with measurements of the contraction time for various laryngeal muscles of different animal species (Sawashima, 1970), and then finally determined by comparing the specific F₀ movements and the corresponding EMG curves.

It is observed that the CT trace in each figure in Fig. 3.6 exhibits a large peak for every syllable located within the hat pattern. In particular, CT is distinctively active during the syllable with R associated with P. It is evident that the CT curve does not indicate any peak activity during "mers" in the word "farmers" in Fig. 3.6 (a) and in (c), where the F₀ contour of that portion is located near the baseline. On the other hand, "mer" in (b) corresponding to the F₀ plateau exhibits a peak (although the peak is masked somewhat by a slope of the previous large peak) in the CT activities.

The SH and the ST curves also exhibit peaks in activity, and seem to be related to the F₀ lowering. The F₀ contour for the word "the" in S58 shown in Fig. 3.6 (b) represents very low frequency values which are lower than the baseline; correspond-
ingly large peaks are observed in the SH and ST curves in that syllable. Marked activities in SH and ST are seen during syllable with the $F_o$ lowering, in particular with the final $F_o$ lowering. The SH curves and the corresponding ST curves are quite similar to each other, as pointed out by Atkinson (1973) and by Collier (1975), although the two activities differ considerably for certain phonemes. For example, during /l/, ST is much more active than SH (in terms of EMG) for speaker KS.

One may ask how the attributes are distinguished in terms of the EMG responses. In the syllables with the $F_o$ rise (i.e., R), with the $F_o$ plateau, and with the $F_o$ lowering (i.e., L), for instance, "all" in Fig. 3.6 (a) and (b), "rai-" in the word "raises" in (b), and "sheep" in each sentence, respectively, both CT and the extrinsic muscles, SH and ST are active. It should be noticed, however, that the peaks in the SH and ST curves precede that in CT, or occur at the same time, when a syllable corresponds to the $F_o$ rise or to the $F_o$ plateau. When a syllable has $F_o$ lowering, on the other hand, the temporal relation is reversed; the CT peak precedes the SH and ST peaks.

The brief analysis of the EMG curves has suggested that the magnitude and the temporal relationship of CT peak and either ST or SH peak within a syllable seem to be meaningful measures for distinction of the attributes. A summary of the measurements for 13 sentences, from S54 to S66 listed in
Table 3.1, is shown in Fig. 3.7. The sternothyroid muscles were chosen to make a pair with CT, because action of ST is recognized to lower the larynx more directly than that of SH, on the basis of anatomical consideration. In this figure, the peak values above the noise level of CT and ST curves are plotted as a function of the delay time \( t_{CT-ST} \) in msec, where the positive values indicate precedence of CT peak against ST peak, while the negative values indicate the reverse temporal relation. The dots located on "N" for CT in Fig. 3.7, represent the peak values of the CT curves where the corresponding ST peaks are not seen. The dots are classified into five categories depending on the corresponding attributes, such as R with P, a simple R, and so on. Although we do not consider the \( F_0 \) plateau as one of the attributes, we conducted the measurements for the plateau in order to permit comparison with other attributes. In the measurement of \( t_{CT-ST} \), the compensation for the muscle contraction time was not taken into account.

There is a clear evidence that the temporal relationship between CT peak and ST peak distinguishes the attribute L from the remaining attributes. The triangles representing the peak activity during L are located in the right half of figure, and furthermore, none of the triangles was located on "N." If the difference in the contraction times for the two muscles is taken into account (presumably 40 msec for the
The EMG peak level and temporal relation between CT peak and ST peak within the syllable associated with the attributes such as R with P, R and so on. A positive value of $t_{CT-ST}$ indicates precedence of CT peak to the ST peak, and the negative value corresponds to the reverse temporal relation.
Fig. 3.7

CT

EMG ACTIVITY (µV x 100)

ST

KS

-100 -50 0 50 100

t_{CT-ST} (ms)

○: R with P
○: R
●: Plateau
△: L
△: L(sentence final)
speaker KS), and if the CT peak corresponds to about the onset of the F₀ lowering the maximum effect of the ST peak would appear during the last half of the F₀ lowering. Especially, the ST peak during the F₀ lowering at the sentence's final (shown by the crossed triangles in Fig. 3.7) will occur after the F₀ lowering has occurred, although the F₀ fall is in part consequence of ST activity. This phenomenon, presumably, is explained by the fact that lowering of the thyroid cartilage continues after the offset of the sentence until its rest position as described in the previous section, and that ST participates in the action.

The magnitude of the CT peak seems to distinguish R associated with P from a simple R and the F₀ plateau. The crossed circles indicating R with P are distributed above 250 µV, and other circles indicating the simple R and the F₀ plateau are located below that value.

There is no systematic manner in which the distribution of the two types of the circles, closed and open circles, can be separated from one another. The specification of R, therefore, is presumably the same as that of the F₀ plateau in terms of the CT and the ST activities.

Three more remarks should be made concerning the EMG data shown in Fig. 3.6 and Fig. 3.7.

First, the CT activities seem to be correlated with F₀ contours to a substantial degree. The correspondence between
the CT activity and the $F_0$ contour, however, may be improved by subtracting the baseline component from the observed $F_0$ contour. Except for the large CT peak during R with P, the peak values in the CT curves are about equal to each other regardless of the corresponding attributes. This property is exhibited in Fig. 3.7, in which the dots except those for R with P are located in a certain range of the activity level, say between 100 µV and 250 µV. The baseline, BL, therefore, is considered not to be related to the CT activities.

Second, the VOC and the LCA activities are not correlated significantly with the $F_0$ contours, as far as the speaker KS is concerned. The primary function of these two muscles seems to be the adduction of the vocal folds. A distinctly large peak activity in VOC and LCA, respectively, is observed at the onset of S56 shown in Fig. 3.6 (a). A glottal stop presumably occurs at this point; perhaps the invisible $F_0$ rise is caused by this glottal stop, even though the initial phoneme of the word "all" is a vowel. A similar event is found at the beginning of S60 shown in Fig. 3.6 (c).

Third, we have pointed out, in Section 3.2.2, that the rise of the larynx hesitates until the initial $F_0$ rise, as typically seen during the word, "almost" at the beginning of S60 shown in Fig. 3.3 (a). The corresponding EMG activities in ST and in SH are observed during the word. It is, therefore, more reasonable to assume that participation of some laryngeal muscles raises the larynx during the initial $F_0$ rise,
rather that the rise of the larynx due to the upward force of \( P_s \) being prevented by the activities in the lowering muscles, ST and SH.

3.3.3 Emphasis, and Intraspeaker Differences in the Manner of its Generation

In the previous chapter and in the previous sections, the sentences investigated were pronounced in a non-emphatic mode. More specifically, the speakers were not asked to emphasize a certain word in the sentence. The \( F_0 \) peak, \( P \) was considered, often, to signal the beginning of the group and of the sub-group, indicating a syntactic function in the assignment of \( P \). It is known that an \( F_0 \) peak appears also during an emphasized word. A question may arise as to whether or not the emphatic \( F_0 \) peak and the \( F_0 \) peak characterized by \( P \) are generated by the same physiological process. To obtain some insight into this problem, we included several short sentences, with emphasis, into the corpus. Four types of utterances for the same sentence "Bill meets Steve" were read: without emphasis as S67, emphasis on "Bill" as S68, on "meets" as S69, and on "Steve" as S70 in Table 3.1.

The \( F_0 \) contours and the corresponding EMG activities for the two speakers, KS and TB are shown in Fig. 3.8 and Fig. 3.9, respectively. In the case of TB, only the EMG activity in CT, VOC and LCA is shown. The \( F_0 \) contours for KS indicate the
Influence of emphasis on the $F_0$ contour and the corresponding EMG activities in CT, VOC, LCA, SH and ST, for the sentence, "Bill meets "Steve" without emphasis in (a), with emphasis on Bill" in (b), on "meets" in (c) and on "Steve" in (d), corresponding to S67, S68, S69 and S70, respectively. The four sentences were read by speaker KS.

Influence of emphasis on the $F_0$ contour and the corresponding EMG activities in CT, VOC and LCA for the sentence, "Bill meets Steve" without emphasis in (a), with emphasis on "Bill" in (b), on "meets" in (c) and on "Steve" in (d), corresponding to S67, S68, S69 and S70, respectively. The four sentences were read by speaker TB.
Fig. 3.8 (a) KS,S67
Fig. 3.8 (b) KS S68
Fig. 3.8 (c) KS S69
Fig. 3.8 (d) KS S70
Fig. 3.9  TB S67
Fig. 3.9 (c) TB S69
Fig. 3.9 (d) TB S70
following assignments of the attributes to each of the four utterance types:

\[(3.1) \quad \text{BL(R)} \quad \emptyset \quad L \quad \text{Bill meets Steve.} \quad (\text{KS}, \ S67)\]

\[(3.2) \quad \text{BL(R)P} \quad L \quad \emptyset \quad \emptyset \quad \text{Bill meets Steve.} \quad (\text{KS}, \ S68)\]

\[(3.3) \quad \text{BL(R)} \quad P \quad L \quad \text{Bill MEETS Steve.} \quad (\text{KS}, \ S69)\]

\[(3.4) \quad \text{BL(R)} \quad \emptyset \quad PL \quad \text{Bill meets STEVE.} \quad (\text{KS}, \ S70)\]

where the words spelled with capital letters were emphasized during the utterance. It appears that each emphasized word exhibits a large F₀ peak that can be characterized by the attribute P. Each of these F₀ peaks is accompanied by, markedly, a large peak in CT activity. There is no significant difference in the VOC and in the LCA curves depending on the location of the emphasized word. It may be stated, therefore, that emphasis can be realized by locating P to that word, and that CT activity is primarily responsible for generation of the emphatic F₀ peak, as well as for the generation of the F₀ peak, P, as described in the previous section.

Emphasis, however, is not only characterized by P, but often introduces a contrast in terms of the F₀ contour of the adjacent words. For instance, in S68 listed at (3.2), only the emphasized word "BILL" receives the attribute that includes P,
while the remaining words in the sentence are assigned the null attribute "Ø". This phenomenon, i.e. a deaccentuation of the adjacent words, is more often observed in the case of the speaker TB.

The $F_0$ contours of those four sentences read by TB represent the following attribute patterns:

\[
\begin{align*}
(3.5) & \quad \text{BL R L (R)L} & \text{Bill meets Steve.} & \text{(TB), S67} \\
(3.6) & \quad \text{BL RPL} & \emptyset & \emptyset & \text{Bill meets Steve} & \text{(TB), S68} \\
(3.7) & \quad \text{BL(R)L RP (L)Ø} & \text{Bill MEETS Steve.} & \text{(TB), S69} \\
(3.8) & \quad \text{BL} & \emptyset & \emptyset & \text{(R)PL} & \text{Bill meets STEVE.} & \text{(TB), S70}
\end{align*}
\]

The attribute pattern for the non-emphatic sentence in (3.5) indicates that the first two words are grouped. In the following three emphatic sentences, however, the individual words correspond to the group itself, and often only the word with emphasis receives the attributes including P. It is important to note that emphasis causes the deaccentuation of the adjacent words, yet the attributes are capable of characterizing the $F_0$ contours under the influence of the emphasis.

Although the emphasized word is always assigned the attribute P for both speakers, the mechanism for generating P seems to be different for the two speakers. Speaker KS was found to use enhanced activity of CT for realizing P. For speaker TB, on the other hand, peak activities of CT, VOC and
LCA are observed during the emphasized words, "BILL" and "MEETS" as shown in Fig. 3.9 (b) and (c), respectively. In the case of the emphasized word, "STEVE", only the LCA curve exhibits a distinctive peak during that word, while the remaining two curves, for CT and VOC, do not indicate such peak activities. As far as the speaker TB is concerned, a complex mechanism involving at least the three intrinsic laryngeal muscles is apparently used for controlling $F_0$ contours.

In summary, it may be safe to state that emphasis on a certain word in the sentence is realized by locating the attribute P, and that perhaps in order to achieve a clear contrast with the adjacent words, emphasis may change the local organization of an attribute pattern, in the sense that different groupings of the words in the same sentence can occur depending on the location of the emphasized word. Since the size of our data corpus is small, it cannot be regarded as conclusive. It suggests, however, that an attribute may be realized by using different control of the laryngeal musculature depending on the individual speaker.

3.3.4 Influence of Voiced and Voiceless Stops upon $F_0$ Contours

It was pointed out in Chapter 2 that the $F_0$ rise becomes invisible, i.e. "(R)" when a stressed syllable with a voiceless and a stop consonant at the onset is assigned R. Also, the maximum $F_0$ value during the following vowel tends to be greater after the voiceless stop than the voiced stop. The
cause for this difference in $F_0$ is not understood in detail. In order to obtain some insight into this phenomenon, we investigated word pairs such as "bill" vs. "pill" in S71 and in S72, "dill" vs. "till" in S73 and S74, and "goat" vs. "coat" in S75 and S76, respectively. These word pairs contrast voiced/voiceless consonants in word-initial position. The $F_0$ contour for each of these words is raised during the initial consonant and only the lowering $F_0$ contour can be seen as shown in Fig. 3.10, for the speaker KS. In this figure, the curves representing $F_0$ and EMG activities for each of the word pairs are superimposed, with time alignment at the onset of the vowel.

For each of the three word pairs, we can see systematic differences which seem to be related to the voiced/voiceless contrast. These differences occur in the $F_0$ contours, and the EMG curves for CT, LCA and VOC. The EMG curves for LCA are not shown in Fig. 3.10, since the SCA curves are similar to the VOC curves. The maximum $F_0$ value that occurs at the onset of each vowel is consistently higher after the voiceless stop than after the voiced stop. This phenomenon may be explained by interpreting the corresponding muscle activities.

First, the peak in the CT curve after each voiceless stop is higher than that after the corresponding voiced stop. The reason why the CT curve reaches a higher value after the voiceless consonant may be as follows. It should be noticed
Figure 3.10

Influence of the voiced/voiceless contrast in an initial consonant cluster of a stressed syllable upon the $F_0$ contour and the EMG activities in CT, VOC and SH. The continuous curves correspond to words with voiceless stops, and the dashed curves represent the words with voiced stops.
Fig. 3.10
that the onset point of the CT rising for the voiceless stop always precedes that of the voiced cognate. In detail, in the case of the voiceless stops, the CT curve starts to rise immediately after the offset of the previous word, "the" which is spoken with a lowering F₀ contour. In the case of the voiced stops, on the other hand, the CT rising starts at a middle point between the offset of the previous word and the stop release position, which is indicated by the dashed line in each box at the bottom of Fig. 3.10. The longer duration of the CT rising during the voiceless stops presumably leads to a higher EMG activity, and thus a higher F₀ value at the onset of the following vowel. This is clear in the contrast, /p/ vs. /b/ and /t/ vs. /d/, but is less clear in /k/ vs. /g/, in our examples. The similar timing of the CT rising can be seen in the voiceless fricatives in "farmers" and "sheep" in Fig. 3.6.

Second, the systematic variation due to the voiced/voiceless contrast also can be found in VOC activities. Deeper dips in the VOC curves (corresponding to wider spread of the glottis) are found more consistently during voiceless stops than during the corresponding voiced stops.

It is worthwhile to notice that the results for one speaker described here seem, at least partially, to support a scheme of laryngeal features proposed by Halle and Stevens (1971). Hirose and Gay (1972), on the other hand, suggested that there was no such systematic difference in the muscle activities due
to the voiced/voiceless contrast. We are not in a position to state, however, whether these discrepancies are due to inter-speaker differences or a consequence of the different experimental conditions.

3.4 Speculation of the $F_0$ Control Mechanisms

3.4.1 The Mechanism Generating the Baseline

We have noted that a drop in the subglottal air pressure, $P_s$ during a sentence is only partially responsible for the generation of the baseline. It has been shown in Section 3.2.3 that gradual shortening of the vocal-fold length during the sentence seems to account for a large portion of the baseline fall. What mechanism underlies this shortening of the folds? We have suggested in Section 3.3 that the intrinsic laryngeal muscles, i.e. CT, VOC, and LCA, seem not to participate in the gradual shortening, to any significant degree. Therefore, we must look for the underlying mechanism in the respiratory system.

The mechanism that we shall propose is a rather simple one. The vocal-fold length is primarily determined by the geometrical relation of the thyroid and the cricoid cartilages. Since, the two cartilages are connected by the cricothyroid joint as shown in Fig. 3.11, their angular relation with respect to the joint specifies the vocal-fold length. We assume the cricothyroid joint to be a purely rotational joint instead of a flexible one which can slide, as assumed in the
Figure 3.11

Schematic representation of the forces acting on the cricoid cartilage: the tracheal pull and \( r \) force generated by CT, marked as "CT force." A hypothetical rotation center for the cricoid cartilage is indicated by the closed circle.
Fig. 3.11
external frame function theory (Sonninen, 1968). Since we are not dealing with extremely high $F_0$ values, the assumption of a purely rotational joint is not unreasonable.

Let us assume, as an idealized case, that the position of the thyroid cartilage is fixed. The thyroid movement shown in Fig. 3.3 (b) for S62 is close to this ideal case. In such circumstances, the cricoid movement is directly related to the vocal-fold length, and thus to the $F_0$ value. Two primary forces may be considered to act on the cricoid cartilage: a force generated by CT, and a tracheal pull which is a downward force acting through the trachea as shown schematically in Fig. 3.1. An increase in the tracheal pull would cause a shortening of the vocal-fold length, assuming everything else being equal. In our speculation, a decreasing in the lung volume during speech may cause an increase in the tracheal pull, resulting in a gradual rotation of the cricoid cartilage and then the shortening of the folds.

We do not have direct evidence for the increase of the tracheal pull along a sentence. However, an interpretation of the functions of the respiratory system seems to provide a support to the theory described above. It is well known that during the inhalation phase of breathing, the thoracic cavity (corresponding to the air volume in the lungs) increases by expansion at the base of the thorax, due to the activities of the diaphragm, and by expansion of the rib cage,
primarily caused by the participation of the external intercostals and the intercartilaginous portion of the internal intercostals (A good summary of the breathing mechanisms may be found in Zemlin (1968)). The expansion of the rib cage is said to be a major contributing factor to the increase of the thoracic cavity in forced inspiration. After a deep inspiration, as in preparation for speech, the elastic recoil of the thorax may generate $P_S$ in excess of that required by the larynx for voice production. Draper, Ladefoged and Whitteridge (1959) found that during speech production, inspiratory muscles, typically the external intercostals, may continue to be active and thus counteract the excessive elastic recoil of the inflated thorax. As the volume of air in the lungs decreases, the recoil force becomes less. Then in order to maintain the necessary air pressure for speech, the expiratory muscles, the internal intercostals, and then the abdominal muscles begin to contract. It is important to notice that the compressing action of the expanded rib cage is a primary factor in maintaining $P_S$ during speech, and further that this action is well controlled by the participation of the intercostal muscles.

During the compression of the expanded rib cage, the volume of the thoracic cavity decreases in two directions: the transverse dimension by virtue of the lowering of the curved ribs, and the antero-posterior dimension as exhibited by a
simultaneous backward and **downward** movement of the sternum. This downward movement would cause a steady lowering of the trachea bronchial tree, resulting in a steady increase in the tracheal pull.

One may ask whether or not the trachea, which is composed of cartilaginous rings enclosed in elastic fibrous membranes, can transmit a sufficient force to pull fibrous membranes, can transmit a sufficient force to pull the crivoicd cartilage. Physical properties of the trachea are not available to us. We can show, however, that only very small downward movement of the cricoid is needed to account for the shortening of the vocal-fold length along the sentence. On the basis of measurements of the laryngeal cartilage dimensions (Maue and Dickson, 1971), we estimated the lever ratio between the cricoid movement and the vocal-fold length, $\frac{\ell_1}{\ell_2}$ as shown in Fig. 3.11, to be about 3 or more. We calculated, in Section 3.2.3, that the average shortening of the folds for the 4 sentences should be 3.8 mm for the speaker KS. This means that only 1.3 mm of downward movement near the center of the cricoid cartilage is necessary to achieve the observed shortening of the vocal fold.

In the light of the proposed underlying mechanism for the baseline, such as the recovery (reset) of the baseline without inhalation, and cessation (or bottoming of the baseline fall within a long breath group, as described in Section 2.3.2. The recovery of the baseline must be associated with a
decrease in the trachea pull, with or without inhalation. Perhaps a quick increase in the activities of the abdominal muscles, such as the external obliques, pushes the diaphragm upward, increasing $P_s$. This action might be accompanied by the internal intercostal activity to expand the rib cage. Thus the thorax and then the trachea bronchial tree may be raised upward. These actions could cause a sudden decrease in the tracheal pull, resulting in a reset of the baseline without inhalation.

The bottoming of the baseline within a breath group may be understood to occur when the compression of the rib cage is prevented, and only the abdominal muscles participate in maintaining $P_s$ for speech production.

We have noted that the intercostals perform the checking action to the excessive elastic recoil to maintain a proper $P_s$. In other words, the state of the rib cage, which could govern the tracheal pull, is controlled by the action of the intercostal muscles. This circumstance could be the reason why the magnitude of the baseline fall is kept fairly constant regardless of the length of the breath group.

It is quite important to point out that the proposed mechanism for the baseline supports the linear additive scheme of the specification of the $F_o$ contours, as described in Section 2.3.1. The fluctuation in the vocal-fold length must be governed by the tracheal pull plus some laryngeal muscular control.
(primarily by CT). Since, as shown in Fig. 3.5, the vocal-fold length vs. \( F_0 \) relation is regarded as linear, the \( F_0 \) contour must be analyzed into the addition of the two components: the non-localized component, i.e., the baseline corresponding to the tracheal pull, and the localized \( F_0 \) components corresponding to the laryngeal muscular control. It should be noted, however, that the linear additive scheme is appropriate for only a limited \( F_0 \) range, say 70 Hz to 200 Hz for male speakers, because linearity of the length-\( F_0 \) relation holds only within that range.

3.4.2 A Simple Laryngeal Model Interpreting the Properties of the Localized \( F_0 \) movements

The \( F_0 \) rise \( R \) with and without the peak \( P \) is controlled primarily by the participation of the intrinsic laryngeal muscles. The criocothyroid muscles, CT, lengthen the vocal folds, thus increasing the stiffness of the folds. The vocalis muscles, VOC, tense the muscle body, and adduct the folds. The activity of the lateral cricoarytenoid is said to adduct and slightly lengthen the folds (Hirano, 1975). The mechanism for the \( F_0 \) lowering, \( L \), however, remains as a controversial issue in the problem of the \( F_0 \) control in speech. We shall attempt to show, on the basis of the properties of muscle contraction, that an active lowering mechanism must be assumed to account for the observed phenomena described in Section 2.3.3. i.e., the agreement of the duration of the \( F_0 \) rise with
(a) A visco-elastic model of skeletal muscle: $k_s$ and $k_p$ represent stiffness as of the series and parallel elastic components respectively. The component $Bw$ w account for viscous loss, and $f$ indicates a contractile component.

(b) An idealized model for the $F_0$ control by CT. The element on the left side corresponds to CT and the element on the right side represents an equivalent muscle load, where $k_c$ ($k_c$) and $B_c$ ($B_c$) indicate stiffness and viscous constants of the CT element (the load muscle element). Shortening of the CT element is denoted by "y."

(c) An equivalent network of the model shown in (b).
that of the $F_0$ lowering. We shall postulate an idealized laryngeal $F_0$ control model using a visco-elastic model of muscle. Rather complex dynamic models of the larynx (Kakita and Hiki, 1974), and of the tongue body (Perkell, 1974) have been devised using a visco-elastic model. We shall, however, use a grossly simplified laryngeal model which is composed of only two components: CT and its load. In spite of its simplicity, the two-component models seem to be sufficient to study a certain basic property of the $F_0$ control by excitation of the muscles.

A visco-elastic model of muscle provides a simple and useful representation of muscle contraction behavior (c.f., Akazawa, Fuji and Kasai, 1969; Huxley, 1957; Bahlar, 1968; Huxley and Simmons, 1971), although it should be noted that no model has proven completely adequate in this regard. The simplest form of such a model may be represented by the block diagram shown in Fig. 3.12 (a), in which $f$ represents a contractile component which generates an internal force by nerve stimulation, $k_p$ and $k_s$ indicate the stiffness of the parallel and the series elastic component, respectively, and $B$ corresponds to the viscous constant of the muscle. We assume that the physical properties of the muscle are linear so that those parameters are regarded as constants. The model does not include a mass component. Akazawa, Fuji and Kasai (1969) found that a visco-elastic model can account for the contractile
behavior in a variety of conditions, such as the isometric (length constant) and and the isotonic (tension constant) condition. The force due to acceleration of the mass, therefore, is considered to be small in comparison with forces due to the viscosity and the elasticity.

In order to obtain a complete model, we must characterize the contractile component. This component is altered rapidly to an active state by nerve stimulation, which develops tension. For instance, Zierler (1974) describes the development of tension, in general terms, such that the active state leads very rapidly to the maximum tension development, within only a few milliseconds: the maximum tension is maintained at a constant value for another few milliseconds, and then tension decays slowly to rest. The active state, therefore, may be characterized approximately by one rate constant specifying the slow decay, although it is known that many rate constants are involved in the force-generating process (Huxley, 1957; Huxley and Simmons, 1971; Julian, Sollins and Sollins, 1974). The system function that relates the active state, $f$, to nerve stimulus, $x$ may be described as follows:

\[
(3.9) \quad \frac{f}{x} = \frac{K}{s + a},
\]

where $K$ represents a gain constant, and $a$ corresponds to a constant specifying the rate of the decay in the active state.
Let us now postulate an $F_0$ control model of the larynx. Since the vocal-fold length is related linearly to the length of CT, as described before, the fluctuation of the vocal-fold length (and thus the $F_0$ values) is specified explicitly in the variation of the CT length. It is, therefore, reasonable to characterize the complex laryngeal mechanisms in terms of the two components: the cricothyroid muscles, CT and their load, composed of a number of the laryngeal muscles, ligaments, and other structures that affect the dynamics of CT. We assume that only CT contains the force generating element. The model represents an active $F_0$ rising mechanism based on the contraction of CT and a passive $F_0$ lowering mechanism due to the relaxation of CT. As a first-order approximation, the load may be assumed to be represented by a single equivalent muscle, as shown in Fig. 3.12 (a), but without the force generating element, which is aligned in one dimension with the CT muscle model.

For the sake of computational convenience, let us further simplify the visoelastic representation of the $F_0$ control model. The physical properties of the muscle components vary significantly depending on the state of the muscle. For instance, Akazawa, Fuji and Kasai (1969) have shown for a frog's semitendinosus muscle, that when the muscle is not active, the stiffness of the series elastic component, $k_s$ is considerably greater than that of the parallel elastic component, $k_p$ ($k_s = 4k_p$). Thus most length variation occurs in the parallel elastic component. The load element is thus specified by two components: stiffness of the parallel elastic component,
\( k_k \) and the viscous constant, \( B_k \) as shown in Fig 2.12 (b).

When the muscle is active, the stiffness of the series component increases exponentially with its stretching. Since the stiffness of the parallel component of the passive (load) muscle, \( k_k \) is regarded as a constant, the series elastic component in the active (CT) muscle in the \( F_0 \) control model may be omitted if in a right range. Thus, CT element is also specified by two constants, the stiffness of the parallel component, \( k_c \) and the viscous constant, \( B_c \). The final configuration of the one-force \( F_0 \) control model of the larynx is shown in Fig. 3.12 (b).

An equivalent network of the idealized laryngeal \( F_0 \) control model is presented in Fig. 3.12 (c). The system function that specifies the relation between the active state, \( f \) and the velocity of the displacement, \( \dot{y} \), may be described as follows:

\[
\frac{\dot{y}}{F} = \frac{1}{B + B_s} \cdot \frac{s}{s + \left( k_c + k_k \right) / \left( B_c + B_k \right)}
\]

Using Eq. (3.9) and Eq. (3.10), we have the following system function for the laryngeal model.

\[
\frac{y}{x} = \frac{K}{B_c + B_k} \cdot \frac{s}{(s + a) \cdot (s + ma)}
\]

where \( m = (1/a)/(B_c + B_k) \), the constant \( m \) can be regarded as a ratio of the two time constants: one for the force generating process and the other for the mechanical process. The
impulse response of the system, \( y_0(t) \), (displacement, and not velocity) can be described by the following equation.

\[
(3.12) \quad y_0(t) = \frac{K}{B_c + B_\lambda} \cdot \frac{1}{a(1 - m)} \left( e^{-mt} - e^{-at} \right) u_{-1}(t)
\]

where \( u_{-1}(t) = 1(t > 0) \), and \( u_{-1}(t) = 0 \) \((t < 0)\).

The maximum displacement (corresponding to the maximum shortening of CT) of \( y_0(t) \) occurs at the moment \( t_{\text{max}} \) as defined by:

\[
(3.13) \quad t_{\text{max}} = \frac{1}{a} \left( \frac{\ln m}{m - 1} \right)
\]

As expected, the response is faster with larger value of the rate constant \( a \) and of the constant \( m \) and thus with smaller time constants for the force generating process and the mechanical process, respectively.

In the case of \( m = 1 \), the impulse response is described by the equation:

\[
(3.14) \quad y_0(t) = \frac{K}{B_c + B_\lambda} \cdot \frac{1}{a} t e^{-at} u_{-1}(t)
\]

The peak deviation occurs at the moment \( t_{\text{max}} \), given by:

\[
(3.15) \quad t_{\text{max}} = \frac{1}{a}
\]

The actual values of the constants in the above equations are not directly available to us. The values, however, may be estimated roughly. Atkinson (1973) calculated the mean
response time (MRT), which he defined as the lag time at which the cross-correlation of a EMG curve and the corresponding $F_0$ contour becomes a maximum. He found that MRT is correlated well with the contraction time in a twitch (Sawashima, 1971), which is the tension response of the muscle excited by a single nerve impulse in the isometric condition.

Mannard and Stein (1973) measured the frequency response of a nerve-muscle preparation in the isometric condition by exciting it by a simulated random nerve pulse train. They found that the response of the preparation corresponds to that of an optimum second-order system for a wide range of values of the average firing rate of the pulse train. This result may be interpreted in terms of the visco-elastic model so that the time constants for the force generating process and for the mechanical process are about equal to each other, within a single muscle, i.e. $1/a = B_C/k_c$. We do not know very much about the physical properties of the equivalent load muscle. Let us, however, assume that the time constant of the load muscle is equal to that of CT, i.e. $m = 1$.

Any error in the value of $m$ will not significantly affect our qualitative analysis. Thus, we shall continue our investigation assuming $m = 1$. Apparently, the time, $t_{\text{max}}$ in (3.15), becomes about equal to MRT, i.e. $a = 1/MRT$.

The impulse response specified in Eq. (3.14) is shown in Fig. 3.13 (a). The curve is normalized so that the maximum
Figure 3.13

(a) The impulse response of the $F_0$ control model shown in Fig. 3.12 (b).

(b) The response of the same model for a sequence of two impulses 50 msec apart, assuming MRT (mean response time) = 60 msec ($=1/a$).
Fig. 3.13

1/a = 60 ms
value of the amplitude corresponds to unity. It should be recalled that the $F_0$ value is linearly related to vocal-fold length as shown in Fig. 3.5. The response curve in Fig. 3.13 (a), therefore, is regarded as the $F_0$ response to an impulse excitation.

In Section 3.3, we estimated MRT visually for the speaker KS to be about msec. It is evident in the impulse response that the rise time, i.e. $1/a=60$ msec for KS, is much shorter than the average duration of the $F_0$ rise, which is about 120 msec for KS. This means that the neuro-muscular system responds fast enough to manipulate the duration of the $F_0$ rise by varying the nerve excitation. For instance, the response for the two consecutive pulses (50 msec apart) is shown in Fig. 3.13 (b). The rise time is about 100 msec, which is close to the measured duration of the $F_0$ rise.

The lowering time, i.e. the time needed to reach 90% relaxation, on the other hand, takes roughly 4 times as long as the rise time. In the case of the speaker KS, the lowering time becomes 240 msec, which is much longer than the observed duration of the $F_0$ lowering, i.e., 120 msec. Since there is no way to shorten the lowering time by manipulating the input signals, we must recognize that the single force model for the laryngeal $F_0$ control cannot account for the duration of the $F_0$ lowering. This explanation is the basis of our claim that an active lowering mechanism must exist in the $F_0$
control in speech. For instance, the load element in Fig. 3.12 (b) must contain a force generating component so that CT and the load muscle composes, in effect, an antagonistic pair.

One may ask how the degree of the contraction, and of the \( F_0 \) movement is controlled. The firing rate of the nerve impulses can change the magnitude of the response. It is known, however, that the firing rate for the nerve impulses is typically 10 to 20 pulses/sec. Thus, in our model only one or two pulses can excite the muscle during an \( F_0 \) rise or \( F_0 \) lowering (Note that the peak response occurs at time \( 1/\tau \) after the final pulse, as shown in Fig. 3.13 [b]). Therefore, the magnitude of the \( F_0 \) movements cannot be controlled in terms of the firing rate.

The actual muscles are regarded as a bundle of muscle fibers. A number of the fibers are connected to the motor-neuron, composing a so-called motor unit that the visco-elastic model actually represents. (Results of studies concerning the behavior of the motor units are described in MacNeil-age, 1973.) Since, roughly speaking the motor units are combined in parallel, a fine control on the generation of force can be achieved by manipulating the population of the motor units activated. Further, because of this parallel composition, the time constant for the entire muscle is about equal to the individual motor unit.

Since the motor units are not fired synchronously, the
firing pattern, which is defined as a function of the population of the activated motor units with respect to time, presumably exhibits a peak. In fact, such peak in the firing pattern seems to be reflected as the peak activity in EMG curve. It is understood that the impulse, $y_0(t)$ represents a special case in which the motor units are fired at the same time, corresponding to the fastest contraction. In fact, the rise time depends on the firing pattern rather than the firing rate of the individual motor units. It is known that when a large force is needed (for instance, to generate a large $f_0$ peak), the motor units are fired more synchronously, resulting in a narrower firing pattern which may be closer to an ideal impulse. The narrower pattern presumably produces a faster contraction of the muscle, and thus a faster $F_0$ rise. This mechanism may explain why the magnitude of the $F_0$ rise (R with P) and the duration are negatively correlated with each other as described in Section 2.3.3. In the case of the lowering, however, a positive correlation was found between the $F_0$ lowering magnitude and the duration. The reason why such a positive correlation between lowering magnitude and duration should exist cannot be given until more is known about the $F_0$ lowering mechanisms.

3.4.3 A Speculation on the Active $F_0$ Lowering Mechanism

As described in Section 3.3.2, the extrinsic laryngeal muscles, ST and SH are active during $F_0$ lowering. Furthermore,
as an effect of the EMG activity of these muscles, a lowering in the height of the thyroid cartilage is observed in the x-ray data. It seems, therefore, to be natural to assume that the extrinsic muscles act antagonistically to CT during F₀ lowering.

Our speculation regarding the active lowering mechanism involving the extrinsic muscles is based on the observation of the movements of the entire laryngeal ventricle provided by Kitzing and Sonesson (1967). To explain the movement of the ventricles these authors postulated that the cricoid cartilage is not simply raised upward with the rise of the thyroid cartilage, but also rotates toward the posterior direction so that the vocal folds are lengthened. Their interpretation essentially states that a hypothetical rotation center exists somewhere on a plane posterior to the cricothyroid joint, (for instance, as represented by the closed circle in Fig. 3.11). This center, of course, is not fixed, but moves with the vertical movement of the larynx. If a force acts on the cricothyroid joint, then a moment with respect to the rotation center is created. The location of the rotation center is, therefore, very critical; if the center is located on the plane posterior to the joint, rising and lowering of the thyroid cartilage leads to a lengthening (and thus an F₀ rise) and a shortening (and then F₀ lowering) of the vocal folds, respectively; if the center is located on the anterior-
plane, the effect is reverse. (This generalization of the interpretation by Kitzing and Sonesson [1967] was developed during a discussion with Dr. Thomas Baer).

The laryngeal height and the vocal-fold length, therefore, can be correlated with each other in certain circumstances. It is noted, for instance, that the thyroid movement for S77 in Fig. 3.3 (c) and the corresponding variation in the ventricle length shown in Fig. 3.4 (c) are grossly correlated. Other sentences, in particular S62, in Fig. 3.3 (b) and in Fig. 3.4 (b), on the other hand, do not indicate such correlation. Perhaps, a mechanism affecting the location of the hypothetical rotation center must exist. In our speculation, the inferior constrictor that supports the thyroid and cricoid cartilages from posterior direction may participate in the control of the location of the rotation center, although there is no direct evidence to support this speculation.

In summary, there seems to exist a floating rotation center for the cricoid cartilage, which is located in a plane posterior to the cricothyroid joint under certain conditions. Because of this rotation center, whenever the thyroid cartilage is pulled downward on SH or/and ST activities, the cricoid cartilage tilts so that the vocal folds are shortened. We must admit, however, that the supportive data are too meager for the proposed active lowering mechanism to be regarded as a conclusive one.
3.5 Summary of This Chapter

The baseline BL is related primarily to a gradual shortening of the vocal-fold length along a sentence. An increase in the tracheal pull due to the compression in the lungs during speech, tilts the cricoid cartilage and thus shortens the vocal-fold length gradually. The contribution of the subglottal air pressure $P_s$ to the baseline is probably less than 30% of the magnitude of the $F_0$ fall.

The $f_0$ rise corresponding to the attributes, R and P is related to the peak activity in CT for one speaker, and in LCA, in CT, and in VOC for the other speaker. It is suggested that the laryngeal maneuver for the $F_0$ control may differ depending on the individual speakers. The attributes, P and R are distinguished only by the magnitude of the EMG activities for the same muscles, as far as the two speakers are concerned.

The $F_0$ lowering, L, is accompanied by a lowering in the laryngeal (specifically, thyroid) height. Peak EMG activity of ST and of SH, preceded by that of CT was observed consistently during the syllable to which L was assigned. We have investigated a response of the vocal-fold length due to an activation of CT by an impulse, using an idealized model of the larynx. According to that investigation, an active $F_0$ lowering mechanism, instead of a passive mechanism based on the relaxation of the muscle that raises $F_0$, is essential for realization of the attribute L. We postulated an active
$F_0$ lowering mechanism which involves the lowering of the larynx.

The overall form of the laryngeal vertical movement is correlated more consistently with that of the amplitude envelope (intensity) of the speech than with that of the $F_0$ contour. We speculated that $P_s$ is responsible for this correlation between the laryngeal height and the intensity of speech. Finally, it should be emphasized that our investigation was based on a small body of data, and further, that some of the characteristic $F_0$ movements are not fully explained. For instance, problems exist such as why the $F_0$ plateau exhibits a steady contour, while the corresponding EMG curve for CT indicates a peak in activity. Perhaps a nonlinear property of the muscle may account for such a phenomenon. Those problems, however, were not touched in this chapter. The proposed $F_0$ control mechanisms must undergo either further refinement or a generalization of the basis of a larger body of physiological data.
Chapter IV Some Speech Synthesis Experiments on Attribute
Patterns: A Preliminary Study

In this Chapter, we shall investigate the perceptual adequacy of the schematized $F_0$ patterns, and the psycholinguistic effect of imposing certain variations on the attribute patterns. The schematized $F_0$ patterns were used extensively in Chapter 2 for analyzing the $F_0$ contours. The $F_0$ patterns may be regarded as a piecewise-linear approximation of the $F_0$ contours. The validity of the approximation must be evaluated in a perceptual experiment.

We have postulated, in Section 2.4, a set of the rules which generate the attribute patterns of any declarative sentence provided that groupings of the words in the sentence and subgroupings which specify a syntactic structure within each group are given. It is of interest to test how listeners judge the patterns generated by the rules, and those which are not accepted by the network shown in Fig. 2.23 and thus cannot be generated by the rules. If the attribute patterns generated by the rules are accepted as American English intonation, while the other patterns are rejected, then it is reasonable to state that the rules effectively characterize a certain aspect of American English intonation.
4.1 Synthesis of Stimuli: A Transformation From Attribute Pattern to $F_0$ Contour

In order to evaluate the perceptual and linguistic effect due to variations in the $F_0$ patterns, it is necessary to vary only the $F_0$ contours of sentences without changing any other speech parameter. In this experiment, we shall use a linear prediction vocoder to synthesize the stimuli, because of its high quality encoding and decoding of speech signals. Combining the $F_0$ detection method described in Section 2.2 and the already established technique for the linear prediction of speech (Itakura and Saito, 1968; Atal and Hanauer, 1971; Markel and Gray, 1974), we have implemented on our laboratory computer a program simulating the linear prediction vocoder. The synthesis part of this vocoder was designed for our particular purpose; an arbitrary $F_0$ contour can be specified manually through a graphic input device (Rand tablet), and then used for speech synthesis together with other speech parameters consisting of linear prediction coefficients and short-term speech energy. A dichotomous voiced-voiceless decision based on presence or absence of $F_0$ was used for generating the excitation source of the vocoder.

We describe now the procedure for generating the $F_0$ contours which form the piecewise-linear patterns for speech
synthesis. We assume that the groupings of the words in a sentence are given and that the $F_0$ peak $P$ can occur only with $R$, and the rise on the plateau $R_1$ does not appear. In other words, the internal structure of the grouped words is not manifested on the $F_0$ contour. Applying the rules described in Section 2.4, the attributes are assigned to each word in a sentence, depending on the groupings. We noted, in Chapter 2, that once the attributes are assigned to a word, they are located on specific syllables depending on the (lexical) stress pattern inside the word. However, special care must be taken, since the specification of the attributes for a specific syllable in the word is not always unique. For instance, the lowering $L$ can occur either during a syllable with 1-stress or during that with 2-stress. Furthermore, the correspondence between the attribute and the $F_0$ movement is not always unique. For example, when an initial consonant of the syllable associated with $R$ is stop or voiceless, the $F_0$ rise can occur either during the consonant (consequently the contour is "invisible") or during the following vowel.

In order to generate $F_0$ contour from the attribute patterns, we shall define a unique procedure for assigning attributes to syllables within the word when the attributes are assigned to that word. 1) If only one of the two basic attributes, $R$ and $L$ is assigned to a word, then the attribute
is located on the syllable with 1-stress in that word. 2) If both of the basic attributes $R$ and $L$ are assigned to a word, then for a polysyllabic word $R$ is located on the syllable with 1-stress, and $L$ on that with the strongest stress after the syllable with $R$; for a monosyllabic word, both $R$ and $L$ are located on the syllable.

A unique transformation from each attribute to the corresponding $F_0$ movement is defined as follows: 3) The baseline $BL$: the magnitude of the baseline fall is constant, and the baseline is terminated at a fixed $F_0$ value, independently of the length of the sentence. The average values listed in Tables 2.4 and 2.5 are used for the magnitude and the terminal point of the baseline, for instance, 30 Hz and 78 Hz, respectively, for KS. 4) The $F_0$ rise $R$: if the initial consonant is voiceless, the $F_0$ rise occurs during the consonant, and thus the $F_0$ contour starts from the plateau level at the onset of the following vowel, and can move in one of three different directions: upward, parallel to the plateau, or downward depending on the remaining attribute in the syllable, either $P$ or $\emptyset$ or $L$, otherwise the $F_0$ contour starts to rise from the baseline at the onset of the consonant. The duration of the $F_0$ rise is about equal to the average value of the localized $F_0$ movements listed in Table 2.6, for instance, 120 msec for speaker KS. When the duration of the syllable
is shorter than that of the F₀ rise, the F₀ contour starts from above the baseline so that the contour reaches to the plateau at the offset of the last voiced phoneme in the syllable. The magnitude of the F₀ plateau may be set to be about equal to the average magnitude of the F₀ lowerings, for instance, 20 Hz for speaker KS. 5) The F₀ rise R with an F₀ peak P: if the initial consonant is voiceless, then the rising portion of the F₀ peak starts from the F₀ plateau at the onset of the following vowel and its magnitude is the same as that of the plateau, otherwise the F₀ rising is the same as that for R in terms of the onset time and the duration, except the magnitude of the rising is twice the height of the plateau. Although, in the original F₀ contours, the magnitude of the F₀ peak varies considerably, a constant magnitude is used as a zero-order approximation. The falling portion of the F₀ peak starts immediately after the peak with the same duration as that of the F₀ rise. 6) The F₀ lowering: if the syllable is located at the final position of a word (always the case for the syllable in a monosyllabic word), the lowering starts to fall from the plateau at the onset of the vowel, otherwise it occurs at the offset of the syllable associated with L. The duration is set to be equal to that of the F₀ rise.
In order to use the above rules, a segmental information, such as the onset time of a syllable, the identity of a consonant, and so on, is needed. We obtain these from the amplitude envelope, and sometimes from a spectrogram of the original speech. The assignment of the attribute BL is not specified in the above definition. We shall determine this on the basis of observation of the original F₀ contour of the sentence. Examples of the generated F₀ contours will be shown in the following sections.

4.2 Perceptual Adequacy of the Piecewise-Linear Approximation to the Rule-Generated F₀ Contours

The sentences in the text, listed in Table 2.1, were used for a perceptual experiment. Speech signals corresponding to the first paragraph (S31 and S32) read by KS, the second paragraph (S33, S34, and S35) read by DK, and the third paragraph (S36, S37, and S38) read by JP, were analyzed to determine the linear prediction coefficients, the short term speech energy, and the F₀ contours. In the synthesis procedure, two types of speech signals were calculated using the F₀ contours without any modification (let us call this type of synthetic speech "vocoder speech"), and using the generated F₀ contours (let us call this type "speech with rule-generated F₀"). In order to generate the F₀ contours, the groupings of the words in the sentences were abstracted, and then the
attribute patterns were determined, using rules described in Chapter 2, and transformed to the F₀ contour by applying the rules defined in the previous section. This procedure was performed manually, and then the generated F₀ contours were fed into the computer through the Rand tablet. An example of the original F₀ contour and the corresponding generated F₀ contour are shown in Fig. 4.1 for speaker JP.

An informal listening test was conducted in a conference room using a tape recorder with speakers. Listeners, consisting of nineteen native Americans, were instructed to evaluate the quality of the intonation of the synthetic speech on an absolute scale in which the maximum point, 10 corresponds to a perfect or an ideal intonation, and the minimum point, 0 to a poor or an unacceptable intonation. The listeners assigning a rating after each stimulus consisting of the three paragraphs had been presented. The speech with the rule-generated F₀ was played back two times, at first and at last, while the vocoder speech was presented only once in between these two stimuli. The evaluations for the vocoder speech and the second speech with the rule-generated F₀ were used as data.

The absolute evaluation for the vocoder speech, averaged over the nineteen listeners, was 8.3 with a standard deviation of 1.1. On the other hand, the average value for the
An example of the rule-generated $F_o$ contour superimposed on the original $F_o$ contour for sentence S37 read by JP. The piecewise-linear pattern is transformed from the attribute pattern using the rules described in Section 4.1.
They eat by picking at their feed with their strong beaks.
speech with the rule-generated $F_0$ was 6.2 with standard deviation 1.8. The quality loss due to the rule-generated $F_0$ contours, therefore, is 2.1, which is roughly comparable to the quality loss of the vocoder speech from ideal speech (i.e. 1.7). A relatively large value of the standard deviation for the speech with rule-generated $F_0$ indicates a large variation in the ratings depending on the individual listeners. Some of the listeners commented that the loss of quality due to the rule-generated $F_0$ patterns varies depending on the paragraph i.e., on the speaker (since each speaker produced a different paragraph). Perhaps, this interspeaker difference in the quality loss may cause the large variation in the listener's ratings. In fact, the distribution of the ratings by the nineteen listeners exhibits two broad peaks, and the average value (i.e., 6.2) is located about the middle of the two peaks. Perhaps, one group of the listeners evaluated the speech with the rule-generated $F_0$ with more weight on the speaker for whom the quality was relatively close to that of the vocoder speech. The other group, on the other hand, rated the quality with more weight on the speaker for whom the quality was relatively low in comparison with the vocoder speech.

Considering the fact that the loss of quality due to the rule-generated $F_0$ contours compares roughly to that of vocoder speech in relation to ideal speech, it is, perhaps
safe to state that the speech quality is not significantly reduced by the rule-generated $F_0$ contours.

4.3 Some Linguistic and Perceptual Effects Due to the Variation of Attribute Patterns

The sentence S15, "The dog likes the enormous gorilla", read by KS, was used as a base sentence. The $F_0$ contour and amplitude envelope, and the corresponding rule-generated $F_0$ contour superimposed on the original contour, are shown in Fig. 4.2. Using the timing information provided by the amplitude envelope, a variety of the attribute patterns are transformed into the $F_0$ contours, and then are used for synthesizing speech.

The attribute patterns tested and the corresponding stylized $F_0$ contours are listed in Table 4.1. In the table, the symbols $w_1$, $w_2$, $w_3$, and $w_4$ represent the four lexical words in that sentence. The symbol $w_1$ corresponds to the first lexical word "dog", and $w_2$ to the second word "likes", and so one. We assume, in this experiment, that only the lexical (content) words may receive the attributes R, L, or P, and that P can occur only simultaneously with R.

The attribute patterns may be categorized into the following five classes. 1) The patterns are generated on the basis of proper groupings of the words. The term "proper"
The rule-generated $F_0$ contour superimposed on the original $F_0$ contour for sentence S15 read by KS. The rule-generated $F_0$ is transformed from the attribute pattern P1 shown in Table 4.1.

**Table 4.1**

The list of attribute patterns and the corresponding stylized $F_0$ patterns for sentence S15, "The dog likes the enormous gorilla." Listener's judgments as to whether or not the utterance is accepted as having American English intonation, and which word in the utterance is perceived to be emphasized, are summarized in the right side of the table for each utterance.
The dog likes the enormous gorilla.

**Fig. 4.2**
### Table 4.1

**Class I) Patterns generated from proper groupings.**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Patterns</th>
<th>Number of Rejections</th>
<th>Votes for Emphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1: BL</td>
<td>P R(L) φ P R L</td>
<td>0</td>
<td>w₁ (2) w₃ (3)</td>
</tr>
<tr>
<td>P2: BL</td>
<td>P R(L) φ R R L</td>
<td>0</td>
<td>w₁ (5) w₃ (1)</td>
</tr>
<tr>
<td>P3: BL</td>
<td>P R L R R L</td>
<td>0</td>
<td>w₁ (2) w₃ (1)</td>
</tr>
<tr>
<td>P4: BL</td>
<td>P R(L) R φ L</td>
<td>0</td>
<td>w₁ (2)</td>
</tr>
<tr>
<td>P5: BL</td>
<td>P R L φ R R L</td>
<td>0</td>
<td>w₁ (1) w₄ (2)</td>
</tr>
<tr>
<td>P6: BL</td>
<td>P R(L) φ φ φ</td>
<td>0</td>
<td>w₁ (6)</td>
</tr>
<tr>
<td>P7: BL</td>
<td>φ R (L) φ φ</td>
<td>0</td>
<td>w₂ (9)</td>
</tr>
</tbody>
</table>
(Table 4.1 cont.)

<table>
<thead>
<tr>
<th>w₁</th>
<th>w₂</th>
<th>w₃</th>
<th>w₄</th>
<th>number of rejections</th>
<th>votes for emphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>P₈: BL Ø Ø P R L Ø</td>
<td>0</td>
<td>w₃ (8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₉: BL Ø Ø Ø RL</td>
<td>0</td>
<td>w₁ (1)</td>
<td>w₂ (1)</td>
<td>w₄ (2)</td>
<td></td>
</tr>
<tr>
<td>P₁₀: BL Ø Ø Ø Ø</td>
<td>1</td>
<td>w₁ (2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₁₁: BL R Ø Ø L</td>
<td>3</td>
<td>w₁ (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₁₂: BL Ø R Ø L</td>
<td>2</td>
<td>w₁ (1)</td>
<td>w₂ (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₁₃: BL R L R L Ø</td>
<td>4</td>
<td>w₃ (3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₁₄: BL R L R L RL</td>
<td>3</td>
<td>w₃ (3)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(Table 4.1 cont.)

Class II) Mislocation of R

<table>
<thead>
<tr>
<th>w_1</th>
<th>w_2</th>
<th>w_3</th>
<th>w_4</th>
<th>number of rejections</th>
<th>votes for emphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>P15: BL R (L)Ø R L</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P16: BL R (L)Ø R L</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Class III) Patterns generated from improper groupings.

<table>
<thead>
<tr>
<th>w_1</th>
<th>w_2</th>
<th>w_3</th>
<th>w_4</th>
<th>number of rejections</th>
<th>votes for emphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>P17: BL R Ø L RL</td>
<td>4</td>
<td>w_1 (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P18: BL R (L)R L RL</td>
<td>2</td>
<td>w_1 (1)</td>
<td>w_3 (1)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Class IV) Patterns generated from improper groupings and miss-location of L

<table>
<thead>
<tr>
<th>w_1</th>
<th>w_2</th>
<th>w_3</th>
<th>w_4</th>
<th>number of rejections</th>
<th>votes for emphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>P19: BL R Ø L RL</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(Table 4.1 cont.)

<table>
<thead>
<tr>
<th>$w_1$</th>
<th>$w_2$</th>
<th>$w_3$</th>
<th>$w_4$</th>
<th>number of rejections</th>
<th>votes for emphasis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P20: BL R (L)R L RL 5

Class V) Patterns which cannot be generated by the rules described in Section 2.4.

P21: BL R L L RL 6

P22: BL R ø R L 4 $w_1 (2)$

P23: BL R ø RL L 3 $w_3 (1)$

P24: BL R R L L 7

P25: ø R (L) ø R L 2

P26: ø ø ø ø ø ø 8
means that the groupings do not violate the semantic constituents of the sentence. For instance, the grouping, ("enormous gorilla") is proper, but the grouping, ("likes enormous") is improper. The patterns from P1 to P14 in Table 4.1 belong to this class. (Note that the generated $F_0$ contour for P1 is shown in Fig. 4.2). 2) The attributes are located on the wrong syllables, although the attributes are assigned to the words on the basis of a proper grouping of the words. The patterns P15 and P16 are in this class. 3) The patterns, P17 and P18, are generated from improper grouping. 4) The patterns are generated from improper groupings, and furthermore, the attributes are located on the wrong syllables, i.e., a combination of 2) and 3). 5) The attribute patterns cannot be accepted by the network shown in Fig. 2.32, assuming that State 0 to be the starting state and State 1 to be accepting state. Such patterns, of course, cannot be generated by the rules proposed in Section 2.4. The patterns from P19 to P26 are in this class. Two sets of the twenty-six utterance types for the same sentence, S15, were synthesized in random order, and then used as the stimuli in the experiment.

Five native Americans participated in the listening test. The listeners were instructed to perform two tasks after each utterance had been presented. First, each listener judges whether or not the utterance is accepted as having
American intonation. Second, if the answer is "Yes", then the listener marks a word which he interprets to be emphasized. The first task is forced judgement, while the second one is not a forced choice, that is, listeners are free to say that none of the words is emphasized. Since each utterance type was presented two times, ten judgements were made for each utterance in total.

The results are summarized in Table 4.1, where the number of rejections (i.e., the utterance was not accepted as having American English intonation), and the number of the votes for each word which was interpreted to be emphasized are listed for each attribute pattern.

Let us examine, first, the patterns which are accepted 100%. It should be noticed that such patterns are found only in Class I in Table 4.1. It is quite interesting to see how the listeners interpret the emphasis depending on the variation of the attribute patterns. Almost without exception, only words associated with R (with or without P) perceived to be emphasized. The four patterns from P6 to P9 were intended to emphasize only one word in the sentence by assigning the attribute only to that word. This intention was quite successful, except for P9 in which the final word was intended to be emphasized. This failure is probably due to the fact that the duration and the amplitude envelope of
the original utterance are not long and strong enough to permit emphasis to be placed on this word by varying only the $F_0$ contour.

In the utterances with the patterns from P6 to P8, the word not only collects high votes for emphasis but also only that word in the utterance is perceived to be emphasized. This result is significant in comparison with the patterns which contains more than two R's (with or without P), such as P1, P2, and P3. In these patterns, the votes are divided between the two words which receive the attribute R, or R with P. In order to emphasize only one word in the sentence, the assignment of P to that word is not sufficient, but the attribute on the remaining words, or perhaps only some adjacent words in the case of a long sentence, must be suppressed. In other words, the remaining words must be deaccentuated. This emphasis by deaccentuation seems to have a stronger perceptual effect than by the assignment of P to the word to be emphasized. For instance, in the pattern P7, the word $w_2$, "likes" is associated with R without P, yet the word is perceived nine times out of ten to be emphasized.

Comparison of the patterns, P1, P2, and P3 may provide some insight into the effect of the attribute P. The first two patterns are distinguished from each other by the presence or absence of P on the third lexical word $w_3$. In the case
of P1, where both \( w_1 \) and \( w_3 \), are assigned \( R \) with \( P \), the votes are divided between the two words: two votes for \( w_1 \) and three votes for \( w_3 \). In the case of P2, where \( P \) is assigned to only \( w_1 \), on the other hand, the votes are shifted to \( w_1 \), as five votes for \( w_1 \) and one for \( w_3 \).

In the pattern P3, the first two words \( w_1 \) and \( w_2 \) are grouped, and thus \( w_2 \) receives the attribute \( L \). Note that in the case of P1 and P2, the lowering occurs during the word boundary between \( w_1 \) and \( w_2 \). There are only two votes for \( w_1 \) in P3, in comparison with five in P2. Probably, the absence of \( L \) and the lack of an \( F_0 \) plateau during \( w_2 \) in P2 is understood to enhance the effect of the \( F_0 \) peak in \( w_1 \), and thus \( w_1 \) is interpreted as more likely to be emphasized.

It is suggested that the listeners interpret the variation of the attribute pattern in a consistent manner. The decisive factor in creating emphasis is the location of \( R \) (with or without \( P \)) and \( L \) on the word to be emphasized, and suppression of the attributes for the remaining words in a sentence. When there is no such suppression, i.e., deaccentuation, a word assigned \( R \) with \( P \) is more likely to be perceived as emphasized, in comparison with a word with simple \( R \). But if two words in the sentence are assigned either \( R \) or \( R \) with \( P \), then the interpretation of the listeners regarding which word
is emphasized seems to be a matter of chance. This is, perhaps, particularly true for an isolated sentence in which no other information is supplied from the context.

Five patterns in Class I, where the attribute patterns are generated from the proper groupings, were rejected at least once. Surprisingly, P10, which contains only the baseline, was rejected only once. The patterns P11 and P12 are rejected, perhaps, because each contour is kept too long on the $F_0$ plateau level. Our impression was that the utterances sounded tense and mechanical.

In the patterns P13 and P14, the noun phrase "enormous gorilla" is divided into two groups. This division of the noun phrase seems to cause a significant number of rejections. We have, however, observed such division in natural speech. For instance, the noun phrase, "brilliant color", shown in Fig. 2.1 (b), is divided into two groups. Perhaps, the division of a noun phrase which originally corresponded to a single group may affect the rhythm of the utterance in an undesirable manner.

In the patterns categorized into Class II, the attribute $R$ is located on the wrong syllable instead of the syllable with 1-stress. In P15, $R$ is assigned on "e" in the word $w_3$, "enormous"; and in P16, $R$ is on "mous". These
patterns were rejected only two times each, suggesting that the timing of the \( F_0 \) rise may not be so crucial for certain words in a sentential environment.

The patterns in Class III are generated from improper groupings. The pattern P17 was rejected four times, while P18 was rejected only two times. Presumably, in the case of P17, at least two different factors which may cause the rejections (let us call such a factor an "inferior factor") seem to be involved: the improper grouping and the lengthy \( F_0 \) plateau as seen in P11 and P12. The effect of an increase of the number of inferior factors upon the rate of rejection is seen more clearly in the patterns P19 and P20 in Class IV. These two patterns are generated from the same improper groupings as P17 and P18, respectively. But P19 and P20 contain another inferior factor locating the attribute L on the wrong syllable, and in this respect are similar to P15 and P16 in Class II. The number of rejections increases from four for P17 to five for P19, and from two for P18 to five for P20, respectively.

The attribute patterns in Class V cannot be generated by the rules described in Section 2.4. These patterns were determined arbitrarily. The patterns from P21 to P24 exhibit more than one level of the \( F_0 \) plateau. These patterns suffer from a large number of rejections. This result may suggest
that only two states, the baseline and the plateau, exist in the system of American English intonation.

The patterns P25 and P26 lack the baseline fall. Both patterns, in particular P26, are rejected, suggesting that the baseline fall is necessary for generating an acceptable intonation, as far as declarative sentences are concerned. It is our impression that the utterance with P25 is somewhere between declaratory and interrogatory.

4.4 Summary of This Chapter

Since the experiments were conducted using a small number of sentences, and a small number of subjects participated in the listening tests, the investigation described here must be regarded as a preliminary study, and thus firm conclusions should not be drawn. However, the following remarks may be noted.

The result of the perceptual experiments seems to support the method of schematic analysis of the $F_o$ contours, which was used extensively in Chapter 2. A transformation in which attribute patterns associated with sentences are mapped to the corresponding piecewise-linear representation of $F_o$ contours was postulated. The utterances synthesized using the rule-generated $F_o$ contours seemed to produce utterances with reasonable quality, in spite of the crude repre-
sentation of the $F_0$ contours.

The listeners interpreted the variation of the attribute patterns coded into utterances in a consistent manner. The patterns which cannot be generated by the rules proposed in Section 2.4 were often interpreted to be unacceptable as American English intonation, suggesting that the rules characterize a certain aspect of American English intonation. The patterns which can be generated by applying the rules, yet based on improper groupings, which are against the semantic constituent structure of a sentence, were not accepted, but in less degree in comparison with the former patterns. The $F_0$ contours generated with a violation of mapping rules in the transformation, for instance locating the attribute on a wrong syllable, were judged sometimes to be unacceptable. The degree of rejection appeared to be positively correlated with the number of inferior factors in an utterance. In other words, an increase in the number of inferior factors involved during the generation of the $F_0$ contour causes an increase in the number of the listeners who interpret the utterance to be unacceptable as American English.

Manipulating the attribute patterns, it was possible to create an effect that placed emphasis on certain words in an utterance. The strongest and decisive effect for emphasis was obtained by assigning R with or without P to the word
to be emphasized, and suppressing the attributes on the adjacent words (i.e., deaccentuation). The assignment of P to the word was also effective to some extent, but the effectiveness was not as strong as that of deaccentuation. When deaccentuation was not present in the sentence, the interpretation of the listeners as to which word was emphasized was often divided among the words assigned R with or without P, although the word with both R and P was more often perceived to be emphasized.

The perceptual experiments have provided some interesting insights into the psycholinguistic aspect of intonation in terms of the attribute patterns. We feel that more extensive studies should be undertaken in this area.
Chapter V: Conclusions and Some Remarks

Acoustic, physiological, and perceptual aspects of American English intonation have been investigated experimentally and theoretically in this thesis. Intonation of declarative sentences was studied by examining the fundamental frequency ($F_0$) of speech. If there is one prevailing point to be made by this thesis, it is that the $F_0$ contours can be characterized using a limited number of configurational elements, or attributes. We have postulated five attributes: baseline BL, rise R, lowering L, peak P, and a rise on the $F_0$ plateau Rl. The baseline BL characterizes a gradually falling component of the $F_0$ contour along a sentence (more specifically, a breath-group). The baseline, thus, may be regarded as suprasegmental. The remaining attributes characterize localized $F_0$ movements, and in this respect are considered to be segmental. The basic attributes are R and L, which characterize a rapid $F_0$ rise and a rapid $F_0$ fall, respectively, in the $F_0$ contours. The other two attributes P and Rl correspond to a $F_0$ peak which often occurs simultaneously with R, and a gradual $F_0$ rise which occurs on the $F_0$ plateau that appears between the $F_0$ rise and the $F_0$ lowering, respectively.
The F<sub>0</sub> contours of sentences are represented by sequences of these five attributes (i.e., by attribute patterns). The attribute patterns appear to be structured as indicated in Fig. 2.32 in terms of a simple state transition network. In other words, the manner of forming the sequences is not arbitrary, but rather is constrained by a number of factors. Since the F<sub>0</sub> contours are produced by a physiological device, configuration of the contours must be limited by the capability of the vocal organs. It may not be necessary, however, for speakers to use the full capability of their vocal organs in individual languages. The physiological factors which constrain the F<sub>0</sub> movements are probably inherent to a specific language.

In the case of American English, the basic attribute pattern appears to be an alternation of the attributes R and L. Correspondingly, the F<sub>0</sub> contours exhibit up-and-down movements, forming a series of "hat-patterns" which are superimposed on the baseline. Such F<sub>0</sub> movements seem to be common to many languages, for instance, Dutch, French, Japanese, and others. The specification of this basic pattern for a sentence is influenced by two linguistic factors - the constituent structures of the sentence and the stress pattern of each lexical word - and by a physiological factor. The pair of basic attributes seems to mark a group of the words
which often correspond to a constituent of a sentence. The attribute R (often associated with P) always occurs during the first lexical word in the group, and L occurs often during the final word. Inside the word to which R or L or both is assigned locations of the attributes are determined depending on the (lexical) stress pattern of that word. It is recognized that R and L can play two linguistic functions, signaling of the constituents and of the lexical stresses. Attributes P and R1 also indicate a linguistic function as markers of the internal structure inside the grouped words.

However, the attribute patterns are not governed entirely by linguistic factors. We have postulated a principle of economy in physiology, whereby a speaker composes the patterns such that least effort is needed for signaling the messages which he wants to send. Some trade-off relation was found between the least physiological effort principle and the signaling of constituents. This trade-off relation seems to make the prediction of the attribute patterns very difficult, if it is not impossible, on the basis of only the constituent structure of a sentence.

In Chapter 2, we have postulated a set of rules for generating possible attribute patterns, provided that the groupings of the words in a sentence and subgroupings which
specify the internal structure of each of the grouped words are given. The groupings and subgroupings of a sentence are determined by the constituent structure of the sentence, the least effort principle, and, in addition, emphasis that may be placed on certain words in the sentence. Once groupings and subgroupings are determined, the attribute patterns are generated applying the rules.

The physiological studies described in Chapter 3 provided some insight into the underlying mechanisms that are used to produce the attribute patterns into the $F_0$ contours. The most interesting finding concerns the physiological correlates of the baseline BL. The cineradiographic experiment as well as a theoretical investigation have suggested that a decrease in the lung volume during speech causes a shortening of the vocal-fold length. We speculated that tracheal pull acting on the cricoid cartilage is responsible for this phenomenon. The shortening of the vocal folds apparently results in the gradual falling of the $F_0$ baseline. According to our estimation, the contribution of the decrease in the lung volume to the $F_0$ fall is significantly greater than that of the subglottal air pressure during speech.

In Chapter 4, we have studied, to a limited extent, the perceptual effect of variations in the attribute patterns
for an isolated sentence. We postulated a transformation in which the attribute patterns are mapped or coded into the corresponding $F_o$ contours (which are represented by piecewise-linear patterns). The rule-generated $F_o$ contours were used for synthesizing a series of sentence-type utterances. In a perceptual experiment, listeners interpreted the variation of the attribute patterns in these utterances in a consistent manner. For instance, the listeners often judged the utterances with the attribute patterns that are in violation of the rules to be unacceptable as having American English intonation. It is, therefore, safe to state that the rules proposed in this thesis characterize a certain aspect of American English intonation.

Finally, we consider the question of whether or not the basic attributes $R$ and $L$ should be regarded as prosodic features. Wang (1967) broke down thirteen tones found in tone languages into seven phonological features, such as High, Mid, Low, Rising, Falling, and so on. Klatt (1973) has postulated a multidimensional encoding of cues in the $F_o$ contours for distinguishing four tones of Mandarin Chinese, to explain the listener's striking ability in the identification of the tones. The $F_o$ contour corresponding to each tone contains several contrasting elements that
constitute the phonological features, such as high vs. mid vs. low, rising vs. steady vs. falling, fast rate-of-change vs. slow rate-of-change, and so on. In the case of American English described in this thesis, however, the $F_0$ contour are structured such that certain of those oppositions appear to be redundant. Specification of R (i.e., rising), for instance, creates also the opposition of low vs. high to the contour adjacent to R, and vice versa. We seem to have, therefore, a choice between a level (or static) representation and a transitional (or kinetic) representation. The kinetic representation, as described by the attributes, is more appealing to us for the following two reasons. First, the transitions occur during the syllable with lexical stress. Speakers expend certain physiological effort to produce them, and listeners apparently interpret them. A second practical advantage is that the kinetic representation is more efficient than the static one for characterizing the $F_0$ contours.

Some authors, however, use rather pattern representation (Armstrong and Ward, 1926; Lieberman, 1967, 1970; Atkinson, 1973). Lieberman (1967, 1970) has postulated two prosodic features, "Breath-Group" and "Prominence". We suggest, however, that these two features are only sufficient for characterizing the $F_0$ contours of short sentences consisting of, say, three lexical words or less. The attribute
representation of intonation has indicated that two different levels of constituents of sentences can be signaled by the $F_0$ contours. The recovery (or reset) of the baseline signals the onset of a major constituent, and as described above, the basic attributes R and L can mark lower constituents inside the major constituent. The two-feature system, on the other hand, can specify only a single level of constituents. A rise-fall $F_0$ pattern followed by a rise, which presumably characterizes [+Breath-Group], might be capable of marking the offset of a major constituent. However, in our observation, such a rise, i.e., a continuation rise, occurs only occasionally in English, although the marking of the major constituents by the continuation rise seems to be essential in some languages, for instance French (Vaissiere, 1971, 1975).

It is clear that many questions remain unanswered. We have investigated a small set of declarative sentences read by a limited number of speakers. The five attributes proposed in this thesis are probably sufficient for characterizing only a subset of American English intonation. The physiological correlates of the attributes still are not fully understood. In this respect, it is hoped that this study has made a small contribution toward an understanding of the speech communication mechanisms.
FOOTNOTES

1. The grouping and subgrouping should be regarded as a structure of a sentence indicated by the attribute pattern, and not the surface structure of the sentence as conventionally defined. A set of rules which generate the attribute patterns when the groupings and subgroupings are given, is postulated in a deductive manner on the basis of observation of the patterns associated with simple phrases. The groupings and subgroupings of arbitrary sentences, when the attribute patterns are given, may be determined by an ad hoc analysis-by-synthesis procedure using these rules. By applying these procedures, we can then examine, for instance, the relationship between the surface structure of a sentence and the structure indicated by the groupings and subgroupings.

2. The first digit "0" in Rule 01 indicates that the rule will be the subject of a generalization.

3. I do not intend to evaluate the principle of economy in any quantitative way. Rather, it is assumed intuitively that elimination of the attributes gives a greater economy, or less effort in the speech production.
REFERENCES


Bloomfield, L. (1933), Language, Holt, New York


Collier, R. and t' Hart, J. (1972), "Perceptual Experiments on


Fujimura, O. (1972), "Accent," Sec. 1.3.3. in Phonetic Science,


Status Report on Speech Research, SR-25/26, 73-86.


Stockwell, R.P., (1972), "The Role of Intonation: Reconsiderations and other Considerations," Intonation, Bolinger


Vaissiere, J. (1976), "A Study in Comparative Prosody for Four Languages (English, French, German, and Spanish)," (in preparation).


