USE OF THE ANALYSIS BY SYNTHESIS MODEL OF SPEECH PERCEPTION BY CHILDREN ACQUIRING THE SOUND SYSTEM OF LANGUAGE

by

CHRISTINE ANN REDDY
B.Sc., University of British Columbia, 1974

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

in
the Division of Audiology and Speech Sciences

in
the Department of Paediatrics

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
March, 1977
© Christine Ann Reddy, 1977
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date May 2, 1977
ACKNOWLEDGEMENTS

I should like to express sincere gratitude to Dr. John Gilbert for his continued advice and criticism throughout the writing of this thesis. Many thanks are also due to my friends and members of my family for the much-appreciated understanding and the support that they have offered.
ABSTRACT

During the time when a child learns the sound system of his language, there is much evidence that the child can perceive phonological distinctions and therefore detect phonetic differences before he can produce these distinctions. This evidence is often provided to disprove the hypothesis that the child could be using an "active" model of speech perception. One such model, the analysis by synthesis model of speech perception, supposes that decoding of the acoustic signal employs the articulatory representation that would be required to produce the hypothesized identity of the incoming signal. The model proposes that while the human auditory system is innately equipped to handle the segments contained in speech, that the correlations between the acoustic information and articulation are learned with experience and form the basis for the division of the continuous acoustic signal into discrete categories of speech sounds.

This thesis reviews recent research into the speech perception process and revises the analysis by synthesis model. It reveals that the human auditory system is innately equipped to divide stimuli (both speech and non-speech) that vary along certain acoustic dimensions into discrete classes. The unique processing that results for speech stimuli, occurs when the stimuli is recognized as having a function in the system of language. Hence the requirements for phonetic processing involve the psychological realization that stimulus originated in the human vocal tract.

This investigation then reviewed the available literature on the perception and production of children acquiring language to determine
whether there is support for their use of the revised analysis by synthesis model. The results favoured that children do use such a model. When resolving the various acoustic cues that combine to form a stimulus complex, the child does refer to his articulatory abilities. Lacking full articulatory knowledge, the perceptual errors that typify children's language, occur. It was shown that the child need not have the precise adult articulatory configuration in order to utilize this model. The model is operative during the child's perception of both himself and the adult. In both instances the comparator performs the function of matching the child's articulatory representation with his perceived representation of a form. The results serve to improve his knowledge of acoustic-articulatory correlations. In this manner the processes of perception and production are closely integrated and as understanding of their fine interrelationship improves, production becomes more accurate and perception is simplified.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.0 INTRODUCTION</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Review of the Literature</td>
<td>6</td>
</tr>
<tr>
<td>1.2.1 The Status of Perception with Respect to Production in the Child Developing Language</td>
<td>6</td>
</tr>
<tr>
<td>1.2.2 Summary</td>
<td>17</td>
</tr>
<tr>
<td>1.3 Hypothesis</td>
<td>18</td>
</tr>
<tr>
<td><strong>2.0 THE ANALYSIS BY SYNTHESIS MODEL OF SPEECH PERCEPTION</strong></td>
<td>20</td>
</tr>
<tr>
<td>2.1 History</td>
<td>20</td>
</tr>
<tr>
<td>2.2 The Existing Model</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Recent Research into the Speech Perception Process</td>
<td>30</td>
</tr>
<tr>
<td>2.4 A Revised Model</td>
<td>50</td>
</tr>
<tr>
<td><strong>3.0 THE USE OF THE REVISED MODEL BY THE CHILD</strong></td>
<td>71</td>
</tr>
<tr>
<td>3.1 The Literature on Children's Perception and Production, Re-reviewed</td>
<td>71</td>
</tr>
<tr>
<td>3.1.1 Literature Evaluating the Relationship of Perception to Production</td>
<td>71</td>
</tr>
<tr>
<td>3.1.2 Literature Examining the Child's Perception of Speech</td>
<td>84</td>
</tr>
<tr>
<td>3.2 Conclusions</td>
<td>99</td>
</tr>
</tbody>
</table>
4.0 DISCUSSION ............................................. 109
4.1 The Importance of Babbling ............................. 109
4.2 Implications for Research ................................. 112

SELECTED BIBLIOGRAPHY .................................. 114
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>29</td>
</tr>
</tbody>
</table>
1.1 Introduction

When the infant is approximately four months of age he "babbles" and begins to experience the sensory-motor connections that are so important for the development of speech. The child progresses from experiencing simple sensations and discriminating between auditory stimuli, to associating the different movements in the vocal tract with the auditory sensations that they produce (by comparison of kinesthetic feedback with auditory feedback). When the child learns to attach linguistic meaning to the discriminated stimuli he is perceiving phonological distinctions. At the same time, as the child begins to interact with his environment, his experiences accrue and he begins to form concepts about the world surrounding him. These concepts, formed from the child's perceptions about objects, are the "meanings" of the words spoken by the child. Hence development of the semantic system as well as the articulatory and perceptual systems is responsible for development of the child's language system. Semantics form the fundamental structure of language and as the child interacts with his environment in a more complex manner, his concepts become more abstract. In a sense, language begets language, i.e. language can be utilized to form new concepts and hence new language.

Not only does the child learn to recognize phonological contrasts,
but he learns to disregard allophonic differences not relevant in his
language, to choose the correct allophone required in a specific
phonetic context, and to use morphophonemic rules. Phonological
systems describe a language that is both spoken and heard, so the
child acquires the ability to produce and perceive these aspects of the
system. It is likely though, that perception and production do not
develop simultaneously. In fact, due to the fine motor co-ordination
required to articulate the sounds of a language, it is likely that a
child has knowledge of his phonological system before his productions
reveal this knowledge.

Research into phonological or phonetic development has for the
large part, studied the order of appearance of children's phonemes and
looked for recurring patterns that may reveal what constitutes "ease
of production". As the child has only limited experience with language,
such research may yield information as to which abilities are innate,
which are acquired and what develops in order to promote acquisition.
Theories of phonological development use results of such research and
attempt to incorporate them into a theory, usually consistent with the
author's linguistic background.

At any particular time during this process the child actually
has his own system, yet he must remain flexible enough to allow changes,
usually towards the adult's system. It might be expected that the
child's system is a "subset" of the complete adult system, yet certain
behaviours are frequently reported that contradict this assumption.
These behaviours are apparent conflicts with use of the adult's system
or with use of the child's system as it is revealed by his production.
In order to explain such behaviours it becomes necessary to research
the state of the child's perceptual abilities and to hypothesize about the role of perception in the acquisition of phonology. An adequate theory of phonological development must not only be complete and account for all aspects of the phonological system, and be consistent with other aspects of language acquisition, but it must also account for behaviours that apparently contradict use of a system resembling the adult's. At the same time it must allow for flexibility and progression towards the adult system. To do this, knowledge of perceptual development and the interaction between perception and production is mandatory. No component of the child's developing system is static and it is this interaction of developing systems that allows for change and at the same time produces the occasional contradiction.

Perception obviously plays a crucial role in the development of production, yet it has been scarcely researched. While diary studies provide extensive data on a child's longitudinal production, virtually no research has examined perceptual development over a period of time. There are inherent problems with such an undertaking. The adult observer will be biased in his interpretation of the meaning associated with a child's production, and in his phonetic interpretation of that production. Research in this area often assumes the format of examining the common features of the child's production and the adult's production (usually in terms of distinctive feature analysis) to determine what contrasts are maintained and therefore, assumed to be perceived. Two assumptions underlie such a rationale. First, that the child can produce what he perceives and second, that the child is perceiving by using an incomplete subset of the adult system.

The child is in a unique position for he not only perceives the
adult but presumably perceives himself, even though the two forms are not identical. One of three possible explanations is usually cited to explain this. The child may perceive some constituents that are common to both forms and he would not discriminate between the forms. His perception of a speech event would be a partial subset of the adult perception and he would produce the same form that he perceived. The rest of the perceptual information would be "noise" to the child and not yet utilized in his system.

Alternatively, the child may perceive as the adult does, but his productive abilities are not mature enough to permit him to produce the forms that he can perceive. In order to be able to perceive himself, the child must have some knowledge of the correspondence between his own and the adult forms. He would develop his production by using this knowledge to reduce the difference between forms. A third possibility exists; the child can perceive using two independent systems, one for the adult and one for himself. His mode of perception would depend upon the speaker and he would not have knowledge of the direct correspondence between forms.

To learn more about what the child perceives it is appropriate to investigate how the child perceives. Is there evidence that perception relies upon production abilities (or vice versa) and therefore the two systems develop concurrently? Is the child first able to discriminate acoustic information to which the auditory system responds well, and then he progresses to more complex forms of analysis? With some insight into the perceptual process, one can better understand the processes of phonetic and phonological acquisition. One can begin to
speculate on what is innate, what is acquired and how the process is perpetuated.

This thesis is concerned with understanding the role of perception in the acquisition of phonology by seeking evidence for the use of a speech perception model by the child. The analysis by synthesis model of speech perception, proposed in its latest form by Stevens and House (1972) has been favourably (though not unanimously) received to account for speech perception by the adult. The model proposed that while some acoustic attributes of the speech signal may be simply converted to linguistic data, for the large part some reference is made to the articulatory mechanisms during speech perception.

The acoustic signal undergoes auditory analysis, then a hypothesis is formed concerning its phonetic identity. This hypothesis is "synthesized" to yield an articulatory description (at some neurophysiological level) that is compared with the results of preliminary analysis. The hypothesis is either accepted or rejected. Processing at different linguistic levels occurs in parallel so that information from other levels is available when forming a hypothesis about a segment's phonetic identity. Perception and production are not independent processes, but two aspects of the same system. In order to use this model, the adult must have developed a strong knowledge of acoustic-articulatory correlations.

When this model was proposed it was supported by extensive research into various aspects of the perception process. This thesis will review that literature and then re-assess the model with respect
to the research that has since been published. The revised analysis by synthesis model will then be evaluated in terms of the literature that is available concerning the perception process in the child. It can then be determined whether there exists support for the use of this "active" model of speech perception by the child. Phonetic development, phonological development and their interaction can then be better understood.

1.2 Review of the Literature

1.2.1 The Status of Perception with Respect to Production in the Child Developing Language

This section will review literature that has evaluated the ability of the child to perceive and produce phonetic and phonological distinctions.

In Russia, Shvachkin (1973) performed an experiment to determine the order that Russian children perceive the vowel and consonant distinctions in their language that differentiate the meanings of words. Eighteen children ranging in age from ten months to a year and one-half, were tested over a period of time ranging from one to eight months.

The children were first taught monosyllabic nonsense names for novel objects until the examiner was satisfied that the child could correctly relate the object to its name. Shvachkin did not clearly state the exact method used to teach the names nor the criteria established to accept a name as "taught". When the subject had learned several names, he was required to select one object from a pair of
objects and then from a group of three objects. An initial evaluation was made of the phonological distinctions that the child could successfully perceive.

Additional experimentation was carried out on the distinctions that the children could not yet perceive in order to determine the sequence in which children acquire the ability to make these contrasts. The children were again taught words and a set of six methods were used to determine when the distinctions could be made. The methods required that the subjects follow various instructions containing the words being tested. Not all children were tested with the complete set of methods and Shvachkin again did not specify the criterion used to determine the subjects' successful performance.

The results from the experiment indicated that the distinctions Shvachkin tested could be ordered into a sequence of twelve stages and that each child acquired the distinctions with little variation from this order. Briefly, vowels were found to be phonemically discriminated before consonants, then the presence or absence of the consonant in CVC-VC pairs was perceived. Finally consonants were differentiated in a series of ten stages.

Although this study was apparently unrelated to the work being done by Jakobson (1968) the results paralleled Jakobson's remarkably well. Jakobson hypothesized that the child learns phonological oppositions in an invariant order and that this order is universally valid. Certain distinctions must appear in a child's speech before later distinctions may be acquired. The sequence of stages is based upon the hypothesis that children learn to distinguish consonants and vowels in
a manner beginning with maximal contrasts and progressing to minimal contrasts. The complexity of the contrast is determined by the number of oppositions that a sound participates in, within the child's system at a specific time. The child's system at first consists of an optimal vowel plus an optimal consonant, such as /pa/ and then the distinctive features are differentiated along two axes — the sonority axis, which includes secondary consonantal source features and the tonality axis, which includes the oral resonance features. Jakobson did not explicitly state the order of appearance of specific phonological contrasts, but he stated that certain oppositions will appear before or after other oppositions. Attempts to verify this theory have led to its support and its rejection. There is no evidence from acoustic-physiological data to support this theory. Shvachkin's work may appear to offer support but in view of the methodological weaknesses and lack of quantitative data in this study, it is wise to reject the existence of a universal invariant order for the acquisition of the perception or production of phonological oppositions.

In addition to determining a sequence for phonological development, Shvachkin made the observation that the children in his experiment were often able to perceive an opposition yet not produce the contrast. Also, those oppositions that the child could produce were perceived from the outset of the experiment and the oppositions that were acquired during the experiment were not usually produced by the children for the duration of testing. Shvachkin therefore concluded that the role of articulation must not be overrated in the process of phonological development. He stated that articulation and hearing influence
the phonemic development of child language. In some cases the child's hearing may serve to distinguish phonemes. In other cases some approximation to correct pronunciation will be sufficient to discriminate phonemes. But phonemes not clearly pronounced by the child will be discriminated at a later stage than those which the child already pronounces clearly.

Garnica (1973) performed an experiment similar to Shvachkin's with English-speaking children. A sequence of stages was tentatively established using Shvachkin's results as a guide, but changing some of the specific oppositions to be consistent with the English language. Following a pilot study the sequence was re-ordered.

Sixteen children ranging in age from 1;5 to 1;10 served as subjects (older children were found to be able to perceive most of the oppositions). The children were first trained to perform the task of choosing only one object named from a pair. The stimuli consisted of nonsense CVC syllables that when paired differed only in the initial consonant sound. The stimuli were presented to the children as "names" for novel objects.

On each trial on a testing session the subject was required to perform an activity with the object that was named from a pair of objects. The criterion for success was seven or more correct trials out of ten trials. Testing began at the stage where a child was judged to be performing, then more advanced oppositions were tested. Later the child was re-tested on oppositions that he failed to make in the earlier stage of the experiment.
Garnica's results do not support the invariant sequence postulated by Shvachkin. The subjects in her study displayed a great deal of variation in the order that they acquired the oppositions. However, she observed that general trends did exist in her data, "which suggest that the order of acquisition is simply more variable than Shvachkin's data would indicate [p.221]."

For many of the subjects in Garnica's study, all of the oppositions were discriminated by the end of testing. At this time the children were less than two-years old and were therefore, at an age when it is very unlikely that they could produce such distinctions in their speech. Although Garnica made no mention of her subjects' productions, it appears again that there is evidence that children are capable of perceiving phonological distinctions before producing them.

In a recent article, Barton (1975a) criticized the criterion used in experiments investigating the speech-sound discrimination abilities in children. The criterion used by Garnica will allow the child to be successful by using random choice in 17.2% of the sessions. In order to reduce the chance of success by random choice (and therefore produce reliable data) Barton raised the criterion to fifteen correct responses in twenty trials. In this case the subject could reach the criterion by chance in only about 2% of the sessions. Alternatively if the subject got the first five trials correct, this could occur by chance only 3% of the time, and testing could be discontinued.

Barton (1975b) sought to determine "whether most phonologically relevant discriminations can be made by a child at an early stage of his or her phonological development and whether they are acquired as
pronunciation develops [p.1]." He carried on the Shvachkin-Garnica experiments but made some changes in methodology and criterion.

Twenty subjects ranging in age from 2;3 to 2;11 were tested with phonological contrasts that are generally considered to be acquired the latest. The stimuli were pairs of monosyllabic words that differed from each other in one phonological feature of one segment. The words were the names of objects that the children were likely to know.

Testing for several pairs of words took place in one session. For each child the testing took place over ten days. The first stage of the experiment consisted of presenting to the child each stimulus to be used in the session and requesting that he identify it. The word was recorded as named if the child named the object without the tester naming it, as prompted if the experimenter was required to name the object once, or as taught if the experimenter had to name the object more than once and give some explanation about its function, etc. The first part of the session was complete when the child could identify each of the stimuli by discriminating it from words differing in all segments.

In the second portion of the experiment the child was required to discriminate the word-pairs that differed only in one phonetic segment. If the child responded correctly in each of the first five trials, testing was discontinued. Otherwise twenty trials were administered for each word-pair, with the criterion of fifteen correct responses in twenty trials.
Analysis of seven pairs of words testing the voicing contrast showed that the children's errors increased when the words were prompted, and increased even more so when the words were taught.

Considering the results for named words only, Barton found that nearly all pairs were discriminated by the subjects consistently either in the first five trials or in fifteen out of twenty trials. Only in three out of seventy cases did the subject respond randomly or in a biased manner, choosing one of the pair consistently. While the children differed in their overall pattern of results, all of them could make some voicing distinctions. It was not found to be the case that some children made the voicing distinction and others did not.

For pairs where one or both of the words were taught, most of the subjects again discriminated the pairs (twenty-five out of forty cases) but did not do so as well as when they could name the words. In Shvachkin's and Garnica's experiments the stimuli were invented words, so the subjects could not have performed better than the subjects did with the prompted or taught words in Barton's study. Some of the errors found in these experiments may have resulted from the children not knowing the words well enough. The children may have been able to make the discriminations that they were reported to not make, if real words had been used.

Barton obtained similar results using pairs of words testing coronality and nasality. The level of accuracy was high, suggesting that children of this age have already acquired these discriminations.

Barton also cautioned that generalizations should not be made about the acquisition of a certain class of oppositions based on just
one or some members of that class. For example, although the subjects discriminated the g/k pair well, these were words that the children were likely to know. He suggested that although the order of acquisition and the order of perceptual difficulty are not necessarily the same, difficulties in discrimination may limit the acquisition of some voicing pairs.

Kornfeld (1971) hypothesized that children's production is based predominantly upon their perception of language, with only some influence from motor constraints and knowledge of the language system. She sought evidence revealing the nature of the representational system that the child has internalized, in order to determine in what way, if any, the child's encoder differs from the adult's. Kornfeld rejected the hypothesis that the child's system of phonological distinctions is determined by that of the adult, i.e. that the child perceives as the adult does, or by using some proper subset of the adult's system, but that he cannot produce the distinctions that he perceives due to motor constraints.

Kornfeld studied thirteen children ranging in age from 1;6 to 2;6 over a period of time (not stated). The children's productions of cluster words were compared with their productions of non-cluster words. The observed pairs differed in the initial segment. Some words began with a singleton consonant, while the others began with a cluster, e.g. 'truck' and 'tuck'. The samples were obtained from the spontaneous speech of children in play sessions. The utterances were transcribed immediately and were also tape recorded. The words were then spliced out and analyzed spectrographically.
Spectrographic analysis revealed that the cluster words were consistently identifiable from the singleton words even though the transcriber had failed to observe any difference. The children did simplify the clusters, but the resulting singleton consonant differed from the consonant produced in the singleton words.

When the clusters contained liquids and glides, the children were judged as having produced a /w/, but spectrographic analysis revealed that these were not "real" "w's". Moreover, the "w" produced for /l/ and the "w" produced for /r/ differed in the mean frequency on their second formant locus. This difference does not correspond to the acoustic difference produced by the adult for /l/ and /r/.

Disregarding whether these results support Kornfeld's hypothesis, they do reveal that children produce distinctions between sequences of phonemes that the adult does not perceive. In order to produce an acoustic variant the child must have perceived that a difference exists. The child may in fact produce the form that he perceives as Kornfeld proposed, or he may not yet have mastered the articulation necessary to replicate the representation that he has perceived. In either instance the ability of the child to perceive differences is at least equal to, if not more advanced than, his ability to produce the difference.

In a further study Kornfeld (1976) presented evidence to show that language learners may attend to unique principles when classifying phonological distinctions. Kornfeld argued that the child must depend more upon phonetic data than the adult, due to his poor
knowledge of morpheme structure constraints and that the child may interpret phonetic contexts differently when there exists a conflict of acoustic cues. She presented acoustic-phonetic data and showed that ambiguous acoustic cues do exist in just the environments where children simplify initial consonant clusters. Again, Kornfeld supported the notion that the child's production is based on the perceptual distinctions that he is capable of making. The state of the child's perception is at most equivalent to the state of his production, and most likely precedes his production by some margin.

In a longitudinal study of his son, A's, phonological development, Smith (1973) presented some interesting observations about his son's behaviour that indicate the relative status of his perception with respect to his production.

Smith analyzed his son's production data in two ways — first as though A's language was a language in its own right bearing no necessary connection to the adult language, and second, as though A's language was derived from the adult system, i.e., the child perceived in a manner equivalent to the adult, but he could not yet produce these forms. Smith failed to find evidence supporting the hypothesis that his son's language existed as a system in its own right. He supported the viewpoint that A had internalized the adult system in order to represent the forms that he perceived. The child would be unable to produce the perceived, underlying forms due to the inability of his articulatory mechanism to perform the required gestures and as a result of the child not yet having made certain hypotheses about the language he is learning.
Three examples led Smith to believe that A could perceive distinctions and do so in the adult manner, before he could produce them. As reported in other diary studies (Velten, 1943; Weir, 1962) the child's productions revealed extensive homonymy, yet when A was required to demonstrate that he could phonologically distinguish the adult forms he did so. Before A could speak at all, Smith tested him with word-pairs such as 'mouth' and 'mouse' and found that he could indicate the correct object from the pair. A was still unable to produce both members of such a pair at the time Smith wrote the book.

A was able to understand his own productions provided that he was still at the stage where he produced these forms. However, if his "incorrect" production of a word was equivalent to the adult form for a different lexical item and A was required to identify the word, he would invariably give the meaning expressed by the adult. Smith quotes the following example:

```
NVS What does [maus] mean?
A Like a cat.
NVS Yes: what else?
A Nothing else.
NVS It's part of you.
A [disbelief]
NVS It's part of your head.
A [fascinated]
NVS [touching A's mouth] What's this?
A [maus]
```

[p. 137]

A must perceive that there is a phonological distinction between the two forms, or he would have responded without the extensive prodding. But did A realize that his production of 'mouth' was not
the same as the adult's form? This would seem to imply that he perceived the distinctions made by the adult but not the lack of this distinction in his own speech. Perhaps a more likely solution is that A was very aware of his "incorrect" pronunciation and was avoiding saying the word.

Avoidance behaviours have been reported in other research (Drachman, 1971; Engel, 1973) as well as in Smith (1973). Often the child will deliberately avoid producing a word that he knows does not make a necessary phonological distinction. Therefore the child is able to perceive the phonological distinctions that are made both by the adult and by himself. Smith reported that on several occasions upon mastering a new word, A would comment on his newly acquired ability. Again this behaviour would appear to indicate that the child perceived that his forms did or did not contain certain phonological distinctions.

These observations made by Smith support the notion that a child perceives distinctions made in his language before he is able to produce them, and furthermore that he is aware of the state of his productive abilities.

1.2.2. Summary

The research presented here indicates that a child can perceive distinctions made in segments of the adult's speech and that he utilizes this information to differentiate the meaning of lexical items. The child can definitely perceive the distinctions that he
produces and it is likely that he can perceive the lack of certain distinctions in his system. Therefore there exists evidence that the child is able to perceive phonological oppositions contained in his language before he can produce those oppositions.

1.3 Hypothesis

The analysis by synthesis model of speech perception (Stevens & House, 1972) "is based on the premise that there exists close ties between the processes of speech production and speech perception, and that there are components or operations that are common to both processes [p. 51]". Once the acoustic attributes have been analyzed to produce a partial specification of the features and a hypothesis concerning the input has been made, the hypothesis is generated in terms of the instructions to the articulatory mechanisms that are necessary to actualize the hypothesis. Then the acoustic information is compared with the articulatory instructions to determine whether the hypothesis can be accepted. To utilize this system the subject must be able to generate, at some level, the articulatory equivalent of the signal undergoing processing and he must have knowledge of the match between auditory patterns and articulatory instructions that give rise to these patterns.

The research presented above supports the hypothesis that children are capable of making perceptual distinctions before they are capable of producing them. In order to discriminate two phonemes the child must be able to detect phonetic differences in the acoustic signal. The implication is that children must be able to perceive phonetic information in a segment without relying upon their ability
to articulate the same segments. Therefore the null hypothesis is that the analysis by synthesis model of speech perception cannot be valid for the child.


Chapter 2

THE ANALYSIS BY SYNTHESIS MODEL OF SPEECH PERCEPTION

2.1 History

The most recently published form of the analysis by synthesis model of speech perception (Stevens & House, 1972) was derived from updating previous "active" models for this process proposed by Stevens and various other researchers.

In 1960, Stevens proposed the design for a machine that would both recognize speech by accepting the speech wave and generating a sequence of phonetic symbols, and synthesize speech by accepting a sequence of symbols and generating a speech wave. Two peripheral units, an analog filter and a model of the vocal tract achieve coupling between the acoustic speech signal and the machine.

Briefly, the speech wave first undergoes "peripheral analysis" where it is reduced by a set of analog filters to display its short-time spectra and periodicities. Thus some information about the segment's phonetic identity and pitch is extracted. The central unit then makes a tentative decision as to the articulatory instructions that could give rise to the information about this segment, using its knowledge of previous analyses, its knowledge about analysis of adjacent spectral samples, and previous scores for the sample under analysis. The "model" then generates speech spectra from the decision made by the
control unit and the comparator compares these spectra with the acoustic attributes of the input that are residing in temporary store. If the comparator accepts the match then the output at this level is an articulatory description of the input. Stevens states that this description may be at the acoustical, anatomical or neurophysiological level, but he does not commit the model to any of these. In the second stage of processing the articulatory description is converted to a sequence of phonetic symbols, again by actively synthesizing signals to be compared with the input that is under analysis. The rules for synthesizing possible signals at each stage of conversion are contained in the machine and the order in which the hypothesized trial signals are generated during analysis is determined by strategies that are also built into the machine or can be evolved as the analysis proceeds (contained in the control unit). The machine can synthesize speech by converting an input of symbols into the articulatory descriptions and then activating the peripheral unit, the vocal tract.

Bell, Fujisaki, Heinz, Stevens and House (1961) described the procedures used in an analysis by synthesis technique to reduce a speech wave to time-varying vocal tract resonance and source characteristics. The analysis by synthesis model described above was programmed on a digital computer to perform the first stage of conversion of the speech wave, i.e. into an articulatory description. Comparison spectra were generated using rules based on an acoustical theory of speech production. The control unit provided information on the poles and zeros of the transfer function and on the type of vocal tract excitation. The experimenters were concerned with establishing rules or strategies that the control component may utilize to convert the
results from peripheral analysis into such information with maximum speed and accuracy. Two methods were used to implement the operations in the control component. Either the experimenter performed the control function or a rudimentary strategy that permitted automatic analysis was employed. When the experimenter controlled the strategies the values for spectra poles and zeros were found to be more accurately matched than by automatic analysis, however the time required was far greater than by automatic analysis. The spectral match determined by automatic analysis resulted in consistent small errors in the formant locations. These errors could be eliminated if more elaborate and time consuming strategies were used to match the spectra.

The authors concluded that such a "feedback" model is advantageous to other models because it ensures that adequate representation of the input is being analyzed, whereas in open-loop analysis a simple attribute analysis is performed and may omit important data or permit errors to occur. Also in this model, once variations due to speaker differences have been resolved, the strategy can be simplified. The strategy is then only concerned with extracting the dynamic features of speech.

Halle and Stevens (1964) presented a speech perception model that utilized an active or feedback process. The model was proposed to account for the conversion of a continuous speech wave into a set of discrete phonemes without segmentation. The dynamic speech wave containing transitions and overlap of information results from the interaction of independent articulatory structures, each with its
inertia and limitations in neural and muscular control. Therefore in this model reference is made to a set of phonetic parameters that describe the independent articulator's behaviours. Stevens and House acknowledged that the phonetic parameters may be akin to distinctive feature systems or the more traditional classificatory systems. They did not commit the model to either but stated that "the vocal-tract behaviour must be described by specifying a set of quasi-independent phonetic parameters [p. 605]" in order that the speech wave preserve its non-segmentable and time-varying characteristics.

The first step in the model consists of preliminary analysis — where vocal-tract information is extracted and tentative identification by special attributes is made, thereby limiting the number of possible sequences that may be generated later. The generative rules are capable of synthesizing spectra for all sets of phonetic parameters that it may receive, but does not need to do so because the control unit orders likely representations to achieve rapid convergence to a match.

The analysis takes place in two stages. The first stage reduces the spectral representation "to a set of parameters which describe the pertinent motions and excitations of the vocal tract, i.e. the phonetic parameters [p. 609]". Here variations between different speakers and a specific speaker under different conditions are resolved. In the second stage, the output from stage I, the phonetic parameters, are transformed to a sequence of phonemes. Here the variations in speech signals due to rate of speech, linguistic background, dialect, and the contextual variation of phonemes are resolved.
The authors do not expand upon how the transformation from phonetic parameters to phonetic segments takes place nor upon the form that phonetic description takes. The model is also capable of transforming an input of discrete phonemes into a continuous speech wave, by using the same set of phonetic parameters in the intermediate stage and then generating speech at a peripheral vocal tract.

2.2 The Existing Model

During the next decade, the extensive research that was carried out on various aspects of the speech perception process, led Stevens and House to refine existing forms of the analysis by synthesis model and to propose a revised version (Stevens & House, 1972). Two areas of research were particularly important for the model's development.

First of all, investigation was carried out to support the existence of a set of "features" that would describe each linguistic unit, the phoneme, in articulatory terms. (The best known such system being that of Chomsky & Halle (1968).) At the linguistic level these twenty to thirty features are capable of describing, uniquely, the phonemes contained in all languages by specifying the presence or absence of attributes that exist at some level in the speech event. They describe such characteristics of the segment as place of articulation, secondary vocal-tract constrictions, the manner of articulation and the type of acoustic source of vocal-tract excitation.

The results from experiments that analyzed errors made in short-term memory, metathetical errors and the perceptual confusions
that result when the speech signal is degraded, indicate that there is some validity for the existence of the distinctive feature as an entity that is analyzed at some point in both speech production and perception. Furthermore the authors stated that because these features occur repeatedly in all languages and because "children seem to learn the rules for manipulating these features from exposure to a relatively small number of samples, this constitutes strong evidence that the nervous system of man is predisposed not only to encode sounds as segments and features, but also to decode sounds in the same ways [p. 13]."

If such features are analyzed during perception of speech then there is strong support for reference being made to articulatory knowledge during the conversion of the speech wave to phonemes, as is required by the analysis of synthesis model.

Researchers then turned to examining the acoustic waveform in cases where some feature was known to exist, to determine the acoustic correlates of that particular feature. Such analysis has not been entirely successful because while some features are represented by invariant acoustic information, the acoustic manifestations for many vary according to the phonetic context and the particular speaker. Therefore research was unable to define particular acoustic attributes to which the speech-perception mechanism is predisposed to detect or recognize, and there can exist no simple acoustic to articulatory, to phonetic translation. The authors proposed that the role of synthesis during perception is still necessary to resolve the non-invariant information carried in the speechwave.
Data that indicated there was a perceptual dichotomy between speech and nonspeech discovered the phenomenon known as "categorical perception". Studies in this area form the second class of experiments that influenced the theory behind the most recent analysis by synthesis model. These experiments examine the identification versus the discrimination of synthetically produced speech sounds. The stimuli vary in one feature such as 'voice' or 'place of articulation' by varying the VOT or formant characteristics in equal steps along a continuum. It was found that for the stimuli of syllabic, dynamic structure, the listener could discriminate them only if he could identify them as being different. Discrimination was excellent near a phoneme boundary yet poor within a phoneme region. The authors disregarded that categorical perception may result from a type of "filtering" mechanism in the auditory system that separates acoustic data into classes, due to the lack of data to support that the auditory system performs such complex analysis. They proposed that the listener uses his knowledge of linguistic categories to separate the auditory patterns into categories and that the listener is assisted in doing so, because as a generator of speech he has knowledge of the articulatory instructions that produce speech, available for his use. They exemplify this proposal with the fact that when producing a stop consonant, the tongue has a discrete target in mind and that when perceiving such, the listener learns to hear these articulatory plateaus or targets as discrete classes.

Therefore distinctive features were considered to be partly innate and partly learned. The ability to handle features is innate
in the human nervous system, but resolving the various forms of a phoneme into a phonetic class, requires knowledge of articulatory-auditory relationships. Also the listener learns to disregard intraphonemic information that is not relevant in his language system. Stevens and House state that such knowledge is probably built up from an early age in the child when he begins to form acoustic-articulatory connections.

In order for a signal to be processed as speech in this model, it is necessary that it evoke the "speech perception mode". When a signal contains the syllabic, dynamic characteristics that identify it as a stimulus that can be categorized in an articulatory class, this mode is achieved. The listener learns to perceive signals that may have originated in the vocal tract as speech, due to the characteristics of the signal that evoke reference to articulatory mechanisms.

In this model, the two-loop synthesis system from previous models can be collapsed to one loop. The existence of distinctive features is a natural explanation for the acoustic to articulatory, to phonetic transformations. These features are assumed to be represented at all levels and once initial analysis has derived the more invariant acoustic information for some features, the remaining features are resolved using knowledge of the correlations between linguistic categories and the articulatory patterns that produce classes of sounds.

The components of this model are shown in Figure 2.1. The acoustic signal undergoes peripheral processing whether it be speech or not. It
is at this stage that the presence of the dynamic characteristics of speech are sought and if found, the signal is processed accordingly. The peripheral processing transforms the acoustic wave-form into a neural time-space pattern that will be used at later stages of analysis. It is known that at this level more than simple time-frequency analysis takes place and probably extraction of some of the more easily discriminated information relevant to description of certain features, also occurs. Some normalization of the signal must also take place, so that the output is a set of normalized attributes that contribute to later identification of linguistic units, (A). These auditory patterns are placed in temporary store to await further processing.

The preliminary analysis component derives the features that are not strongly context dependent from the signal (A). These features are available after conversion of the results from peripheral processing and result in a partial specification of a feature matrix of the utterance, (B).

The control component has access to the results of preliminary processing as well as the results of analysis of previous parts of the utterance, the lexicon, and the output of the comparator. With this information, the control unit makes a hypothesis concerning the representation of the utterance in terms of morphemes. The features are an abstract quantity that underlie, but are not necessarily identified with, the acoustic attributes or the production of the signal.

This hypothesized representation (B) travels to the generative rules where it is transformed into a representation of the articulatory instructions that would be necessary to generate such an utterance.
Figure 2.1 A proposed model of the speech perception and production processes.
(from Stevens and House, 1972)
The articulatory instructions (V) that result, could be used to control the articulatory mechanisms and produce speech output.)

The generated patterns (V) are compared by the comparator with the attributes of the analyzed utterance residing in temporary store and judged as to their closeness of match. This information is relayed to the control component and the hypothesized sequence is either accepted or a new hypothesis is made using the error detected in the comparator. This loop is transversed until the message has been successfully identified.

The model relies on the comparison of auditory patterns (A) with articulatory instructions (V) that potentially are able to produce such patterns. The comparator must contain a catalogue of such relations. The articulatory gesture is represented in terms of tactile, proprioceptive sensations and motor-commands. Stevens and House state that the catalogue is built up as the child begins to utter sounds, hear the auditory result and form their association. Therefore the child is aided in learning to perceive speech, by being able to articulate it.

2.3 Recent Research into the Speech Perception Process

In view of the studies of categorical perception that have taken place since the Stevens and House model was proposed, the theory underlying this model must be revised. Results of various studies on infant perception reveal that the young infant is capable of discriminating sounds varying in acoustic dimensions that are common to
all languages. Eimas, Siqueland, Jusczyk and Vigorito (1971) presented to one- and four-month old infants, the syllables /pa/ and /ba/ that varied in only one acoustic dimension, voice onset time. The stimuli were synthesized so that the VOT varied from -20 to +80 m sec in 20 msec steps. The infant's responses were measured by determining the rate of non-nutritive sucking. After a baseline rate was established, an auditory stimulus was presented which caused the rate to increase. When the infant became habituated to this stimulus and his rate of sucking decreased, a second stimuli was presented. Typically, the sucking rate then increased again. In the control group, the second stimulus was identical to the first stimulus. An infant was assumed to have discriminated the stimuli if his sucking rate increased after a change in stimulation, or if its decrease was of smaller magnitude than that shown by the control subjects. The results indicated that the infants discriminated /p/ from "non-/p/" and that they did so at the same phoneme boundary determined for adult subjects by Lisker and Abramson (1970). In addition, the infants more often discriminated the stimuli when they were from different adult phonemic categories than when they were from the same category. Eimas et al. took this as evidence for categorical perception and stated

"that the means by which the categorical perception of speech, that is, perception in a linguistic mode, is accomplished may well be part of the biological makeup of the organism and, moreover, that these means must be operative at an unexpectedly early age [p. 306]."

Morse (1972) designed an experiment using the same method as Eimas et al. for infants between one and two months of age. The
stimuli were the syllables /ba/ and /ga/ varying only in one acoustic characteristic, their formant frequency pattern. The second and third formant transitions varied in direction and rate of change in order to vary the feature, 'place of articulation'. Morse was also concerned as to whether children can discriminate intonation patterns before they can discriminate phonemic differences, so the fundamental frequency contour in the syllable /ba/ was varied so that intonation ranged from rising to falling, (ba+) to (ba−). The results showed that the infants categorically perceived 'place of articulation' at the same boundary and in the same manner as the adults. The infants also discriminated in the acoustic cues that determine intonation patterns. Morse concluded that if infants of this age can utilize acoustic information in a linguistically relevant manner, their experience in producing these contrasts is not necessary for perceiving these distinctions. He suggested that the results support the theory that the speech code is structured in terms of articulatory invariants that are extracted from the speech signal during perception, (proposed by Liberman (1970)), and that the articulatory basis for this code must be primarily a phylogenetic acquisition.

Kuhl and Miller (1975) performed an experiment to determine whether the chinchilla, an animal without a phylogenetic history of phonetic knowledge, could be trained to differentiate classes of speech sounds differing in the feature, 'voice'. In Experiment I, the caged animals were presented with one of two auditory stimuli and were trained to cross a barrier whenever the "negative" stimulus was presented. They were positively reinforced if they did so, and
negatively reinforced if they failed to do so. The stimuli were /ti, ta, tu/ and /di, da, du/ produced by four different speakers. The chinchillas were successfully trained to discriminate each pair into voiced-voiceless classes and in addition, without further training they generalized this learning to discriminate pairs selected from /te, tæ, to/ and /de, dæ, do/.

In Experiment II, learning was kept to a minimum by controlling the order of stimulus presentation and rewarding the animal after each trial. The stimuli consisted of /ta/ and /da/ tokens with the VOT value ranging in 10-msec steps from 0 to 80 msec. Four English-speaking adults identified the same stimuli as /ta/ or /da/. The animals identified the stimuli in a manner strikingly similar to the adults. The adults and the chinchillas determined a very similar phonetic boundary, 35.2 and 33.5 msec respectively. Both sets of subjects were further tested with labial and velar VOT series and again they both determined the same location for phonetic boundaries and these values agreed with those reported by Lisker and Abramson (1970). The authors concluded that if speech is "special" because reference is made to articulatory representations during perception, or because "special phonetic feature detectors" (review following) are involved in this process, then the chinchillas should not have performed as they did. They also suggested that the articulatory basis for speech classes cannot be considered to be phylogentic. Kuhl and Miller proposed that it is necessary to reveal the "special" status of speech and that it is more likely that these perceptual
abilities result because speech-sound oppositions are selected to be highly distinctive to auditory systems, both human and animal.

These experiments provide evidence that the interpretation given to categorical perception by Stevens and House must be reconsidered. To summarize the revised information available on categorical perception: 1. The ability to perceive certain auditory stimuli in categories (at least along the dimensions of 'place of articulation' and 'voicing') does not require previous experience in articulating these sequences, nor is learning required to form phonetic boundaries. 2. Categorical perception may reflect a general characteristic of the auditory system and speech-sound oppositions may have been selectively chosen to coincide with the system's resolving powers. (In this case, categorical perception does not reflect "perception in a linguistic mode" as stated by Eimas et al. (1971). 3. Consequently, theories that consider speech as "special" must not consider categorical perception as evidence for this "specialness", because the chinchilla also perceives speech sounds categorically. Therefore, neither reference to articulatory representations nor the use of phonetic feature decoders, are required by a listener to divide auditory stimuli changing along one acoustic dimension, into specific classes.

In order to determine whether there exist general acoustic dimensions to which the auditory system may respond in a discontinuous fashion, researchers investigated the nature of stimuli that would evoke categorical perception. It was necessary to show that these dimensions evoked categorical perception in nonspeech stimuli as well
as speech stimuli. At the same time researchers were interested in determining what differentiates speech from nonspeech in order that processing is performed in a different manner. House, Stevens, Sandel and Arnold (1962) showed that listeners do categorize stimuli as either speech or nonspeech and that there are no degrees mediating the two judgments. Dichotic-listening experiments, such as those performed by Shankweiler and Studdert-Kennedy (1967), Studdert-Kennedy and Shankweiler (1970), Kimura (1964) and Cutting (1974) and experiments on humans with brain lesions (Kimura, 1961) support the fact that speech (or signals resembling speech) and nonspeech stimuli are processed in different areas of the brain. Gilbert and Climan (1974) found this to be true for children as young as two and one-half years of age. If this speech-nonspeech distinction is not made in the auditory system, signalled by the presence of certain acoustic dimensions, and if resolving these acoustic dimensions into specific categories can be considered as forming a kind of general "auditory concept" then the question may be proposed, "When does an auditory concept become a phonetic concept?". The following experiments were directed towards answering these questions.

Burdick and Miller (1975) performed an experiment to determine whether chinchillas were capable of distinguishing /a/ and /i/. The method was the same as that used by Kuhl and Miller. Four chinchillas were successfully trained to discriminate the stimuli /a/ and /i/ produced by one speaker, for the same token at constant pitch and sound level. The token, speaker and pitch level were then varied
until 48 stimuli were presented for identification. The chinchillas successfully generalized their training to the new stimuli and thereby ignored variations in pitch level, sound level and voice quality when making the phonetic judgments. Synthetic /a/ and /i/ tokens varying in formant structure and pitch contour were then presented to the subjects. The chinchillas correctly observed the relevant formant differences and ignored the pitch differences when differentiating these stimuli. In these experiments, the chinchillas were required to do more than simply differentiate between two stimuli differing by one variable. They were required to isolate the essential differences, disregard nonessential variables and disregard differences with the same nature as those that were essential, but were still within a vowel category. These requirements meet those considered in general psychology as being essential for concept formation. The authors proposed that in this case the animals defined the stimuli as members of a psychological category, in this case an acoustic category. The chinchillas do this in the same way as they form concepts for the meaningful sounds of nature that occur in their environment. The authors suggested that speech sounds whose category membership can be perceived, should be described as auditory concepts, for no cognitive rule is employed when judging membership in these categories. The chinchillas perceived the category membership and therefore formed auditory concepts, so Burdick and Miller concluded that this ability can be accounted for by general psychoacoustical processing.

Sinnott, Beecher, Moody, and Stebbins (1976) studied the ability of Old World monkeys to discriminate the speech sounds /ba/
and /da/. Four monkeys were successfully trained to discriminate /ba/ and /da/ stimuli spoken by an adult male. Two monkeys carried this training over to synthetic /ba/ and /da/ stimuli, while the other two required retraining but did achieve the transfer. In the second experiment human and monkey subjects discriminated synthetic /ba/ and /da/ stimuli that varied in second and third formant frequencies (80 Hz steps between the endpoint stimuli used for initial training). Latencies from stimulus onset to response time were recorded on a PDP-8 computer and the interstimulus interval (ISI) was one of 0.5, 2.0 or 3.8 sec. The task was a version of an AX discrimination task, where the subject heard two stimuli and released a key if the second was different from the first.

The human subjects identified a phoneme boundary, yet they were able to detect intraphonemic stimulus differences and were not completely "categorical" in their discrimination. The difference threshold for formant transitions was 160 Hz for human subjects and 320 Hz for monkey subjects.

The latency functions for the monkeys were essentially linear with some slight increase as the two stimuli approached the same value. For the humans, latencies were constant when the stimuli were from distinct categories, yet there was a discrete latency increase as the second stimuli approached the value of the first. The nature of the AX task permits a direct comparison of the auditory stimuli using "echoic memory" whereas the ABX task usually used in categorical perception experiments requires that the stimuli be coded and stored in short-term memory. The authors suggested that the humans first sought inter-
phonemic differences and if they were found the response was immediate. Otherwise the response was delayed to search for differences in timbre. The monkeys relied solely on timbre information, producing constant latencies. Further support for this proposal occurred when ISI increased. The humans' performance became more categorical. In this case the stimuli would be needed to be coded into short-term memory and the task more closely resembled an ABX paradigm. Monkeys showed no such signs of latency differences even after extensive training. The authors interpreted this finding as evidence that the monkeys perceived the stimuli in a continuous manner and the humans in a categorical manner, and that there exists some difference in the underlying perceptual mechanism. Sinnott et al. proposed that while auditory systems may detect acoustic discontinuities that occur along a continuum, a particular species will form concepts about such distinctions that are meaningful for him in his environment and communication system.

Cutting and Rosner (1974) first demonstrated that categories and boundaries occur for both speech and nonspeech stimuli that differ in rise time. In Experiment I subjects were presented with four sets of stimuli that varied in rise time in nine steps, from 0 to 80 msec. The nonspeech stimuli were sawtooth waves at 294 Hz and 440 Hz. The speech stimuli were the affricate/fricative pair /ʃʃ/ combined with one of the two vowels, /a/ or /æ/. The rise varied in 10 msec intervals. The subjects performed an identification task and then an ABX comparison task.
The results of both arrays of nonspeech stimuli were combined as they revealed no differences. Similarly, the two sets of speech stimuli were analyzed as one set. Subjects identified the sawtooth waves either "plucked" or "bowed" sounds from a string instrument and identified a boundary at 40 msec. The discrimination function showed a sharp peak at this boundary and discrimination was significantly poorer for stimuli within a category. The speech stimuli were also perceived in a categorical manner. Therefore categorical perception takes place for the affricate/fricative distinction as well as for distinctions among stop consonants.

In Experiment II, the order of the tasks was reversed to rule out the possibility that the identification task had influenced the discrimination task. The same two sawtooth series as well as two series of sine waves (at 294 and 440 Hz) varying in rise time from 0 to 70 msec were presented. Reversing the order of tasks had no effect on the results and categorical perception occurred for sine waves as well as the sawtooth waves. Moreover, the category boundary determined by the peak in the discrimination function occurred at the same location, 40 msec.

In the ABX discrimination task it is necessary that some type of coding take place, for echoic memory could not store all three stimuli during analysis. Within-category information may be lost during the coding of A and B because these differences did not remain long enough in short-term memory. "The quick loss of within-category information is the crux of categorical perception [p. 569]." Therefore the ABX paradigm is very sensitive to identification of perceptual categories in audition.
For stop consonants, it is obvious that the listener perceives the event almost immediately, as a phonetic class. But for the "plucked" and "bowed" notes the encoding cannot be phonetic.

"This fact, coupled with the result of the first experiment, which demonstrated that rise time can cue perceptual categories in both speech and music, suggests that certain aspects of phonetic coding may be intimately related to the coding of naturally occurring nonlinguistic sounds [p. 569]."

The authors concluded that the perception of speech has developed around the existing properties of the auditory system and that rise time may be one of the dimensions by which different categories are detected.

Miller, Weir, Pastore, Kelly and Dooling (1976) investigated whether noise-buzz sequences would be categorically perceived in a similar manner to plosive consonants differing in voice onset time. A 100 Hz buzz of constant duration (500 msec) was presented with thermal noise that led the buzz in 10 msec steps along a continuum from -10 msec to +80 msec. In the discrimination test the stimuli were presented in groups of three. Two were identical while the third differed, and the subject was required to identify the different stimulus. For the identification test, the subject labelled the stimulus as "noise" or "no-noise", labels chosen from the subjects' observations. The stimuli were discriminated and identified in a manner that met the requirements for categorical perception proposed by Studdert-Kennedy, Liberman, Harris and Cooper (1970). The authors interpreted categorical perception as the detectability of one single component in the stimulus complex that is judged relative to the
constant part of the stimulus context. There are psychophysical boundaries or thresholds for perceptual effects that are encountered as the one component changes, and at the threshold the effects undergo rapid changes in discriminability, clarity or perceived magnitude. Pairs of stimuli that straddle the threshold will be discriminated well, but pairs selected from within a category will follow Weber's law. Therefore if constant differences are tested they will be discriminated at a constant Weber fraction (the ratio of ΔX to X). Miller et al. added that training, attention and memory may serve to sharpen the boundary and enhance discrimination of items from either side of the boundary, but that these effects are only a matter of degree. Also, when discrimination involves comparisons of categories that are named, as opposed to comparing sensory traces, then the results will more likely be categorical. Therefore while categorical perception may occur in general sensory psychoacoustical experiments, it will occur in ideal form only when "certain combinations of stimulus configurations as well as training and perceptual-cognitive factors obtain [p. 416]".

The boundary determined by Miller et al. for noise-lead (20 msec) was also determined in an experiment by Stevens and Klatt (1974). In Experiment I, the stimuli consisted of a 5 msec burst of noise followed by a variable interval of silence (from 0 to 40 msec) and the onset of a synthetic vowel with fixed formants. The stimuli were not judged to be speech. Listeners could not detect a silent interval for a VOT up to 15 msec. After a VOT of 25 msec, listeners reported a silent interval. Fifty percent detection occurred at a VOT of 20 msec.
In Experiment II the stimuli were /ta/ or /da/ tokens that varied independently in both voice onset time and transition duration. The authors sought evidence for listeners perceiving the voiced–voiceless distinction, not only by detecting VOT differences but by detecting the presence or absence of a significant formant transition following voicing onset (a characteristic of natural English stop consonants). The results showed that the VOT was increased when the duration of the formant increased. The average VOT at the phoneme boundary moved from 26 to 39 msec for a 30 msec change in formant transition duration.

These findings indicate that the various components contained in the initial 20–25 msec of a stimulus are integrated and perceived as a unit. The presence of an $F_1$ transition after the onset of voicing is a cue for a voiced consonant. The authors postulate

"first that the cue for the presence of a consonantal segment is a rapid change in the acoustic spectrum occurring at a point where there is an abrupt or discontinuous increase in intensity in some frequency range (Stevens 1971). This rapid change in the acoustic spectrum is in the frequency range above about 1000 Hz, and occurs over a brief time interval of 20–30 msec. The characteristics of this transient spectrum shift provide some of the cues for place of articulation for the consonant [p. 657]."

The results of Experiment II can be explained by this hypothesis. When VOT is delayed, a second discontinuity arises in the spectrum and the two onsets are perceived as separate events. If the VOT is less than 20 msec then the two onsets are perceived as simultaneous and the stimulus is identified as voiced. In voiced consonants the transition is of significant duration and occurs at the onset of voicing, but in the voiceless consonants the transition is short and is virtually
complete by the beginning of VOT. The listener determines whether the onsets are simultaneous or discrete and whether there exists a presence or an absence of rapid spectrum change at the onset of voicing, as cues to the voiced-voiceless distinction.

These results clearly explain the results found by Eimas et al.. Stimuli were varied along the VOT continuum in 20 msec steps. The infants perceived the 20 msec VOT as voiced and the 40 msec VOT as voiceless. The present results would predict such a division based on the VOT values chosen.

Stevens and Klatt offered a tempting explanation for the longer VOTs that occur for velar consonants than dentals, and for dental consonants than labials. "It is known that the duration of the movement of the articulator that forms the closure is greatest for the tongue body, less for the tongue tip, and least for the lips [p. 658]". The formant transitions are manifested acoustically during this time and therefore their rate of change follows the same order. When the rate of the transition is the slowest, the voice onset time for voiceless stops must also increase so that the transition is complete before voicing begins and a conflicting cue for a voiced stop does not result. Hence the authors posited an articulatory basis for some of the acoustic cues that exist in a stimulus and the suggestion is apparent — that perceptual distinctions are based on distinctions performed necessarily during production.

While the above studies clearly reveal that the auditory system is capable of distinguishing auditory stimuli into classes when one acoustic dimension is varied about specific "thresholds" for that
dimension, it is also clear that the simple notion of fixed regions of auditory sensitivity will not suffice to explain this ability.

Lisker and Abramson (1970) in their studies of categorical perception, revealed that speakers of languages with characteristically different VOT values will also determine categorical boundaries at these different values. Experience then must play some role in shaping these boundaries and furthermore the experience involves articulating the classes of speech sounds. Burdick and Miller (1975), Miller et al. (1976), and Sinnott et al. (1976), have all formed conclusions in this direction. Burdick and Miller claim that the categorical perception experiments reveal no more than the formation of an auditory concept and that this concept is likely "perceived". They suggest that these concepts are then relevant to the perception of the constituents of language, allophones, phonemes and syllables, but elaborate no further. Miller et al. concluded that training, attention and memory are important factors in establishing categories and they point out the parallels that exist with studies on training and generalization (Mostofsky, 1965). Most important is the authors' observation that "overlearned responses or labels for the stimuli.....probably increase the chances that the subjects will compare labels or names of the stimuli rather than their sensory traces during discrimination tasks [p. 416]". Therefore knowledge of the speech sounds in a language will contribute to the ability of the mechanism to identify the speech stimuli as "named" sounds (that name themselves). Sinnott et al. proposed that a specific species will perceive stimuli that vary along an acoustic continuum, as distinct categories depending upon the concepts
they have formed about such stimuli in their communication system, and therefore, if these categories are meaningful to that species in its environment. A human may perceive a series of animal calls as falling along an acoustic continuum, yet the animal may respond categorically. The boundaries and classes that exist in a communication system reveal the relevant distinctions that the users require.

The paradigms used to determine discrimination functions reveal that coding into short-term memory is necessary to reveal the characteristics of discrimination found in categorical perception. If such coding is not required, the sensory traces stores in "echoic memory" can be directly compared and perception becomes more continuous. The above studies propose that poor discrimination for stimuli within a class results because the items are identified as members of the same class and enter phonetic store. This information is analyzed during discrimination because the intraphonemic differences are lost due to poor auditory memory. Discrimination is good for stimuli from different classes because the representations in phonetic store are sufficient for accurate discrimination. The latency functions reported by Sinnott et al. support this idea and indicate that the listener will first process phonetic information and then (and probably not normally) process intraphonemic information. Pisoni (1973a), using an AX same-different task, varied the interval from A to X from zero to two seconds, for both vowel and stop consonant continua. When the stimuli were from different categories, discrimination was high for both consonants and vowels did not vary with the interstimulus interval. Presumably the phonetic store was utilized; it having a slower rate of decay
than the auditory store. When the stimuli were from the same category the discrimination was poor for consonants and independent of stimulus delay. For vowels, discrimination was high but declined with an increase in the delay interval. Presumably in this case, the information was coded into auditory store; it having a rapid rate of decay. If the hypothesis of the role of auditory memory in categorical perception is to be accepted, then it has been shown that auditory memory is strong for vowels and weak for consonants. It follows that when vowels are degraded so that they are poorly represented in auditory memory, they are perceived more categorically (Pisoni, 1973b; Lane, 1965; Sachs, 1969).

Consonants and vowels are distinguished in these experiments not by the processes of assignment to their phonetic class, but by their characteristics and the duration of their auditory stores. These considerations and their implications for the revision of the analysis by synthesis model proposed by Stevens and House will be presented in the following section. First it is necessary to make some mention of the possible existence of feature detectors.

Neurophysiological evidence for the existence of feature detectors in the auditory system has been found for species other than the human. Single cells at the cortical level have been found to respond to a specific stimulus or a specific characteristic contained in stimuli. For example, Wollberg and Newman (1972) described single cells in the cortex of the squirrel monkey that responded only to the species' 'isolation peep'. In humans, the adaptation paradigm is used to evidence the existence of such feature analyzing systems. In short,
the rationale is that if such detectors exist and a stimuli is presented continuously, then it will adapt or fatigue the detector being utilized and sensitize an adjacent detector (or system). Therefore, if perceptual shifts can be demonstrated to occur after prolonged stimulation by one feature, a feature detector is said to exist for that feature. If such systems exist, then several questions need to be answered—"Are they peripheral or central?", "Are there separate detectors for the presence or absence of a feature, or is there only one detector?" and "Are the detectors auditory or phonetic?".

Adaptation studies have shown that perceptual boundaries are shifted towards the adapting stimulus, for the features 'voicing' and 'place of articulation' (Ades, 1974a; Cooper, 1974a; Eimas, Cooper, & Corbit, 1973; Eimas & Corbit, 1973). Ades (1974b) and Eimas et al. (1973) demonstrated that the shift occurs as strongly when the adapting stimulus is presented to the ear contralateral to the ear being tested, as when both stimuli are presented binaurally. They construed this as evidence that the detectors are central. Support for the existence of separate detectors for the two values of a feature comes from evidence that voiced stops are more resistant to adaptation and yield smaller boundary shifts than do voiceless stops (Eimas & Corbit, 1973; Eimas et al., 1973).

Whether such systems are phonetic or auditory in nature has not been determined. Eimas et al. (1973) demonstrated that the first 50 msec of /da/ (an acoustic segment that contains voicing information but is not heard as speech) was ineffective as an adapting stimulus. Eimas and Corbit (1973) showed that when using the continua /ba/ to /pa/ or
/da/ to /ta/ and the extreme values from either set as an adapting stimulus, cross-series adaptation took place. Also the discrimination function shifted to coincide with the adapted phonetic boundary. Eimas et al. (1973) showed that adaptation with labial stop produced boundary shifts on alveolar and velar stop consonant VOT continua.

If the effect were acoustic, differences in second and third formant transition directions might have ruled out the effect. The researchers interpreted this as evidence that the detectors are phonetic. But the feature tested was voice onset time, a feature identified by complex and relational cues (Stevens & Klatt, 1974). Cooper (1974a) and Ades (1974b) produced cross-series adaptation for /b/-/d/ continua with different vowels. Therefore the formant transitions could rise for a token with one vowel and fall for a token with another vowel. Ades and Cooper interpreted the results as showing that the detectors are phonetic in nature.

Bailey (1973) constructed two /ba/-/da/ series. In one, $F_2$ was fixed and place cues were in $F_3$. In the other, there was no $F_3$ and all place cues were in $F_2$. The experiment yielded cross-adaptation from the $F_2$ cue series to the fixed $F_2$, but none from the $F_3$ cue series to the non-existent $F_3$. These results strongly favour that adaptation is auditory in nature.

The controversy over phonetic or auditory detectors has not been resolved. At this stage it is probably wise not to rule out auditory, phonetic or both auditory and phonetic systems. Cooper (1974b) showed that adaptation on the /bi/-/pi/ continuum produced not only a perceptual boundary shift, but that speakers also shifted this boundary during production. Here is strong support for the often
postulated perception-production link. Studdert-Kennedy (1976) suggested that the origin of this link be sought in studies of language acquisition. In the following chapter, literature concerning child perception and production will be examined to seek support for this production-perception link.

Recent research has succeeded in demonstrating that the auditory system performs analysis more complex than was previously believed to be the case. However it has not demonstrated that results of this analysis produce a complete specification of features that can be converted "dictionary-style" into a sequence of phonemes. Invariance between acoustic attributes and linguistic features has not been successfully resolved and there remains the need for a mediating step, such as the synthesis of the signal at some abstract articulatory level. Results from studies of categorical perception (Lisker & Abramson, 1970) and feature analyzing systems (Cooper, 1974b) also support the existence of a link with articulation during perception.

The studies reviewed in the previous section demonstrated that the type of processing previously believed to be reserved for speech, also occurs for nonspeech that contains specific acoustic dimensions. Hence this analysis occurs in the auditory system for all stimuli. Knowledge that speech perception requires processing beyond that used for nonspeech, and that unique cortical sites are employed for such, led research to re-investigate the requirements for phonetic processing.

Miller et al. (1976), Burdick and Miller (1975), Cutting and Rosner (1974), and Sinnott et al. (1976) hypothesized that phonetic processing requires that the stimuli be recognized as part of a larger system of sound stimuli that interact and are hierarchically organized.
The transformation occurs at the acoustic to psychological level. Stimuli are processed accordingly when they have some meaning in this system and represent a concept. Such systems may be linguistic or nonlinguistic (music) but when they are linguistic, the components are sounds that form auditory concepts, their own "name".

The current research has led to only a surface understanding of the generalities involved during speech perception. The specifics are buried under discrepancies in data and authors' interpretations of their data. This is certainly the case concerning the nature of speech processing that extends beyond that used for nonspeech.

In the following section the analysis by synthesis model will be revised as specifically as possible, in view of current research.

2.4 A Revised Model

Peripheral Analysis — Huggins (1964) and Cherry and Taylor (1954) presented continuous speech to subjects while alternating the signal between ears. The subjects were required to "shadow" the speech by repeating it aloud as they heard it. At a critical rate of alternation — about 3 to 4 alternations per second — the intelligibility of the speech was sharply reduced. Huggins also demonstrated that this effect was not the result of an "attention-switching" factor but that processing at the peripheral level was interrupted. This rate corresponds approximately to the rate of syllable occurrence. Therefore it was suggested that a partial extraction of cues occurs at the peripheral level and that such analysis takes place over the approximate length of a syllable.
Stevens and House (1972) proposed that this peripheral analysis is more than simple frequency analysis and that all signals undergo such processing. Acoustic differences that result from different speakers are normalized at this level and the result is a set of normalized acoustic attributes, some of which may correspond directly to features that are not strongly context-dependent.

Results from categorical perception experiments show that stimuli not heard as speech can be perceived categorically if they contain certain acoustic information that varies over a continuum. Therefore such analysis is not linguistically relevant and is a general characteristic of the auditory system. All acoustic stimuli must undergo the same auditory analysis. This finding also indicates that the differentiation between speech and nonspeech does not occur at this level, due simply to the detection of these acoustic dimensions.

Stevens and House cited neurophysiological studies of animals "such as [those of] Kiang, Watanbe, Thomas, and Clark (1965) on the auditory system of the cat, of Lettvin, Maturana, McCulloch, and Pitts (1959) on the visual system of the frog, and of Frishkopf and Goldstein (1963) on the frog's auditory system as evidence that fairly complex processing takes place peripherally [p. 48]". Current research examining the acoustic correlates of the more invariant features support that complex analysis is performed in the auditory system. Stevens and Klatt (1974) report that the voiced-voiceless distinction is indicated by the presence or absence of rapid spectrum changes at the onset of voicing (also found in nonspeech stimuli). The presence of a stop consonant is detected by a rapid intensity increase in the
spectrum (Stevens, 1971) and rise time indicates the fricative-affricate distinction as well as differentiating classes of nonspeech stimuli (Cutting & Rosner, 1974).

Recently, evidence has been produced to show that some of the more strongly context-dependent features can be derived directly from acoustic analysis of the speech signal (apart from the possible existence of feature analyzing systems). Kuhn (1975) studied spectrographic samples from fricative and normal speech to determine whether the front cavity resonance frequency could be directly calculated. A variable frequency component in the fricative speech was found to specify the quarter-wave resonance of the front cavity. In normal speech this resonance was calculated from the second or third formant transition depending on whether the vowel was formed in the back or front of the mouth. Kuhn also determined that stop consonant bursts yield information that can be used to estimate the cavity resonance and hence place of articulation data. Further research may verify that such information is derived directly by acoustic analysis and that the same may occur for other non-invariant features.

Stevens (1960) required that the model perform linguistic and nonlinguistic analysis simultaneously, yet independently. Stevens and House cite examples from neurophysiological studies of animals that show the periodicity of the vowel, and analysis of the vowel's spectrum envelope are analyzed separately in the auditory system (Davis, Silverman, & McAuliffe, 1951; Schouten, Rijsma, & Lopes Cardozo, 1962).
Burdick and Miller showed that chinchillas could discriminate two categories of speech sounds without responding to the irrelevant variation in sound level, pitch level, pitch contour and voice quality. Therefore it seems that the animals processed the information relevant for the phonetic decision independently of the irrelevant information.

Wood (1975) reported evidence for parallel extraction of pitch and segmental information bearing on phonetic classification. In a speeded classification task the reaction times increased when subjects were required to make decisions on 'place of articulation' when pitch was also varied independently (as opposed to the control condition where only the target dimension varied). When the pitch was the target dimension and 'place' also varied the reaction times did not increase.

Wood (1974) varied fundamental frequency and phonetic class in a correlated manner rather than independently and found that reaction times were significantly shorter than on the two-dimensional test. He concluded that linguistic and non-linguistic information are extracted separately and simultaneously.

The Stevens and House model proposed that the acoustic attributes arising from peripheral analysis are placed in temporary store while the synthesis "loop" is transversed and phonetic analysis is completed. Hence the model implicitly assumed that auditory analysis is performed at the peripheral level and phonetic analysis beyond. The temporary store component must be an auditory store as no phonetic coding has yet taken place. The nature of auditory memory (to be discussed below) presents problems if such is assumed to exist at such
an early stage in the speech perception process. It is more appropriate to divide the study of the processes involved into auditory and phonetic processes with their respective stores, rather than peripheral and central processes. The results considered above, under the heading "peripheral analysis", include the complex analysis that is performed in the auditory system regardless of site.

**Temporary Store** — One of the alternatives that a model of speech perception must specify, is whether processing is carried out in a series of stages, i.e. auditory analysis, phonetic analysis, morphological analysis and so forth, or whether processing of these various levels occurs in parallel. The Stevens and House model allows for parallel processing and therefore some form of store must be utilized during analysis. As mentioned above the temporary store must be auditory in nature.

"The auditory store, or trace, is usually assumed to be rather like an echo: a faint simulacrum, if not of the waveform, at least of its neural correlates at an early stage of processing [p. 262]", (Studdert-Kennedy, 1976). The echo is an analog of the original and decays rapidly. If another sound arrives before decay is complete, the trace is immediately displaced. As revealed by the results of categorical perception experiments, the auditory store is utilized when discriminating stimuli that are phonetically members of the same class yet acoustically different. The resulting discrimination functions revealed that this memory is strong for vowels yet weak for consonants.

Auditory memory can be divided into two stores. Store I is very brief, while Store II can last several seconds. Decoding of a CV syllable requires analysis of information spread throughout the
syllable. Therefore syllable information (consonant and vowel information) must be processed in parallel and Store I must last over syllable length, probably 200–300 msec. (Pisoni & Tash, 1974; Liberman, 1970). However studies that sought to specify the duration of this store by determining the time required to free a target CV from interruption by a masking syllable, produced widely varied results probably due to the variety of masking conditions imposed. Syllable processing time probably varies with attentional control, speaking rate and other factors.

Stage II of auditory memory probably lasts several seconds. It was studied in detail by Crowder and Morton (1973), under the term "precategorical acoustic stage" (PAS). Three effects are typically evidenced in PAS: 1. Error increases from the beginning to the end of a list that is recalled, with a slight drop on terminal items (recency effect). 2. The terminal drop is increased if the list is presented by ear rather than eye (modality effect). 3. The recency effect is reduced if the auditory list is followed by a redundant spoken suffix that indicates the subject should begin recall (suffix effect). These three effects were shown to occur when CV lists consist of members differing in vowel alone or vowel and consonant. The effects do not occur for CV or VC syllable lists where members differ in voiced stop consonants (Cole, 1973). Crowder (1971a) concluded that vowels receive representation in PAS, while voiced stop consonants do not. However in a study by Pisoni and Tash (1974) subjects were required to make same different judgments for pairs of stimuli drawn from the /ba/-/pa/ continuum. "Same" reaction times were faster when the pairs
were identical than when they were acoustically distinct. "Different" reaction times decreased as the acoustic differences between items in a pair from different categories increased. Hence it appears that at least some trace of consonants must reside in the store. In addition, Darwin and Baddeley (1974) demonstrated the recency effect for tokens of a stop CV, /ga/, and the two CV syllables /fa/ and /ma/. They also eliminated the recency effect for vowels by reducing their duration (30 msec of a 60 msec CV syllable). They concluded that items in PAS are not reliably accessed when they are similar acoustically and that the consonant-vowel distinction is for the large part, irrelevant.

During recall of an eight-item list, the degree of interference will decrease as time between items decreases and, if the suffix is presented two seconds after the last item, the suffix effect virtually disappears. Crowder (1971b) postulated that an active rehearsal process at the articulatory level is operative and therefore that performance will improve as time allowed for PAS decay increases. Crowder stated then, that decay does not occur in PAS because the subject can check rehearsal of items against his auditory store. If another stimuli arrives before this verification is complete the error probability will increase. "A preliminary articulatory, if not phonetic, decision must be made before PAS is lost if rehearsal is to permit cross-check with the store [p. 267]" (Studdert-Kennedy, 1976).

Although Crowder's hypothesis may appear to support the temporary store component in the Stevens and House model there exists
one important difference. Precisely, the information that would need to be represented in PAS in order to resolve invariance (i.e. stop consonant data) appears to be poorly represented in auditory memory. Consonantal auditory memory is probably much less than a second. In order to allow the synthesis loop to operate, the store would need to be of longer duration. The PAS rehearsal loop may go into operation at an early stage and prolong the auditory memory. Consonantal information would still decay though, and there is no evidence of early operation of this loop.

Studies therefore began to reconsider the serial-parallel distinction and began to distinguish the types of information processed auditorally and phonetically and their interaction. Wood (1975) performed four experiments to distinguish between auditory and phonetic levels of processing. Two techniques were utilized; the first measured reaction times on a speeded classification test and the second measured average evoked potentials during the same tests. Experiment I measured reaction times for the identification of a phonetic dimension (place of articulation for voiced stop consonants /ba/ and /ga/) and an auditory dimension (fundamental frequency). The evoked potentials for each cerebral hemisphere were also recorded. The stimuli varied between two values for each dimension and subjects performed a two-choice identification task. In the control condition only the target dimension varied, in the orthogonal condition the target dimension and the irrelevant nontarget dimension varied orthogonally. Results showed there was a substantial increase in reaction time from the control condition to the orthogonal condition for place, but there was minimal
difference for pitch. There was no difference in the average evoked potentials for processing of pitch and place at right hemisphere locations, but significant differences occurred at every electrode location of the left hemisphere. The conclusions were that an auditory-phonetic distinction can be made, that the phonetic dimension required processing beyond that used for the auditory dimension, and that such processing took place in the specialized left hemisphere.

In Experiment II two auditory dimensions, pitch and intensity, were compared to determine whether the results from Experiment I did indeed occur due to different levels of processing required for phonetic and auditory processing. The reaction times for both auditory dimensions increased significantly from the control to orthogonal condition. The evoked potentials for pitch and intensity did not vary at any electrode location. These results support that phonetic processing took place for the phonetic dimension, 'place', in Experiment I and indicate that the auditory dimensions, pitch and intensity can interfere with each other's processing when irrelevant variations occur.

Experiment III was designed to determine whether the processes involved in speech perception that extend beyond those involved for nonspeech perception were auditory or phonetic in nature. That is, were mechanisms responding to specific acoustic events contained in the speech signal or were abstract phonetic features extracted from the speech signal? The stimuli varied again on the auditory dimension of pitch and also varied on the isolated second formant transitions from the stimuli used in Experiment I (/ba/ and /ga/). If the results
from Experiment I were due to processing of second formant transitions alone, then the results for Experiment III should have been identical. The reaction times increased for both dimensions from the control to orthogonal condition. The average evoked potentials revealed no significant differences at any electrode location. These results resemble those from Experiment II. Hence the perception of second formant transitions in isolation is closely related to perception of nonlinguistic dimensions. In order to evoke phonetic processing the phonetic context is required.

Experiment IV investigated the dimensions of pitch contour and pitch. Pitch contour was chosen because it is considered to cue linguistic distinctions but it is not under context-conditioned variation. The reaction times and error rates increased substantially from the control to orthogonal condition. In addition the reaction times and error rates increased substantially from the control to orthogonal condition. In addition the reaction time was significantly larger for pitch contour than for pitch under both conditions. The evoked potential revealed no differences between dimensions at any location. Reaction times were faster for pitch than for place in Experiment I, or for pitch contour in Experiment IV. However, in Experiment IV this occurred in both conditions, whereas in Experiment I, only in the orthogonal condition. Hence this difference reflects different temporal requirements for the judgments, rather than interference between dimensions.

Wood reviewed that the auditory and phonetic distinction had been noted in dichotic listening results, the "phoneme boundary effect",
speeded classification tasks, average evoked potential recordings, and adaptation studies. He believed that there exists an empirical basis for this distinction and that it is now necessary to investigate the nature of phonetic processing. The results from Experiments I and III indicated that phonetic processing consists of more than simple detection of certain acoustic features. The isolated second formant transitions were not perceived in the same manner as when they are imbedded in a phonetic context and serve as cues for a linguistic distinction. In view of recent evidence that certain acoustic features have been isolated that do correspond with linguistic features, it seems plausible that phonetic processing involves mechanisms that extract relatively invariant acoustic properties from the speech signal and more abstract mechanisms that resolve context-variable acoustic cues.

Wood cites Studdert-Kennedy (1976) on the subject of adaptation studies:

"We should not discount the possibility that the auditory-phonetic distinction is misleading in this context, and that the adapted systems are both auditory and phonetic. If, for example, the output from the auditory analyzers tuned to speech funneled directly into phonetic processors so that adaptation of one set entailed adaptation of the other, a convincing separation of the two would be difficult to demonstrate [p. 278]."

(Studdert-Kennedy then reported the study by Cooper (1974b) mentioned above as support for this theory).

In addition to distinguishing between auditory and phonetic processing levels, these experiments yield information on the relationship of phonetic processes with the general auditory system. Wood rules
out the possibility that processing follows a strict serial order, with auditory information processed at one level followed by phonetic processing at another level. Wood (1974) proposed a serial-parallel model with three components. A peripheral component performs preliminary analysis of all signals, a central auditory component processes nonlinguistic auditory information and a central phonetic component extracts phonetic features from the results of preliminary analysis. The two central components perform in parallel but are dependent upon input from the peripheral component. The auditory level of processing then consists of two parts, peripheral analysis and a central component working in parallel with the phonetic component.

Pastore, Ahroon, and Puleo (1975) disagreed with the basic assumption underlying Wood's (1974) research into the auditory-phonetic distinction. The assumption is that because certain aspects of auditory stimuli are found to be important when discriminating different phonetic stimuli, the use of these aspects in auditory stimuli as variables in a study implies a phonetic causality in the results. The authors therefore believed that Wood did not observe any phonetic processing and the results could be explained in auditory terms. Pastore et al. replicated Wood's experiment using two tone pips varying in pitch and duration. The reaction times and error rates were determined. The results followed the same pattern as Wood's results. Reaction times for correlated conditions were faster than those for the control condition and only the reaction times for the orthogonal tone pip condition were slower.
In retrospect, a comparison of the results from Experiments I and III in Wood (1975) and the significant differences in average evoked potential values that resulted between these experiments, show that Wood revealed that a type of processing more complex than auditory processing and involving the left hemisphere, took place. Furthermore, the results indicated the importance of the acoustic dimension existing in a phonetic context, for the processing to take place in the left hemisphere. Therefore the results of Pastore et al. may indicate that certain acoustic features may be processed in a manner similar to speech, just as nonspeech was shown to be categorically perceived, yet something more is involved to evoke phonetic processing. Wood has shown that phonetic processing exists but in order that it is utilized, the signal must be judged as having sufficient speech-like features.

Blechner, Day and Cutting (1976) varied nonspeech stimuli along two dimensions — rise time and intensity. Having shown that sawtooth waves differing in rise time (not heard as speech) could be categorically perceived (Cutting & Rosner, 1974) and could be selectively adapted (Cutting, Rosner & Foard, 1976), both characteristics of speech stimuli, the authors were interested in determining whether these stimuli would produce similar results as speech on a speeded classification task. The stimuli were the same as those used by Cutting and Rosner (1974) only they varied in two levels of intensity. The results were identical to those of Wood (1975). Reaction times increased significantly for the rise time dimension in the orthogonal condition. Thus there was an asymmetrical pattern of interference; intensity variation
interfered with the processing of rise time, while the reverse was not true. The authors proposed that neither a strict serial nor a strict parallel processing model could account for their results and that factors such as stimuli discriminability and task characteristics may affect the mode of processing. In this experiment nonspeech stimuli varying in rise time and intensity produced the same result as speech stimuli. Therefore two auditory dimensions produced asymmetric interference and the authors felt that the auditory system should be considered to possess various processing levels. The linguistic-nonlinguistic distinction cannot be overlooked as there is evidence for left hemisphere processing. But Blechner et al. proposed that less emphasis should be given to "special" speech processing and that the linguistic-nonlinguistic dimension is not an accurate way of describing the non-acoustic factors that determine certain perceptual processes. Instead, they proposed that speech is an important hierarchically coded system of sound where sounds can be recoded into higher order linguistic units. The "plucks" and "bows" used in this experiment are also lower level components in a highly structured sound system — music. The need for certain perceptual processing may be determined by non-acoustic factors and therefore studying the linguistic-nonlinguistic difference may not be an accurate method of studying these factors. The type of processing results from the interaction of the acoustic nature of a sound and the manner in which it may be coded in a hierarchically organized sound system.

Preliminary Analysis — At this stage in the Stevens and House model the acoustic attributes that arise from peripheral processing undergo
analysis and some of those that correspond more invariantly with features are decoded into such. This results in a partial feature specification of the segment.

Support for this stage of analysis comes from several sources. The categorical perception studies that revealed certain acoustic dimensions produce this type of processing whether they be contained in speech or nonspeech, support that there can exist acoustic detection of some features. Such dimensions as rise-time (Cutting & Rosner, 1974), noise lead (Miller et al., 1976) and the presence or absence of sharp spectral change at voice onset time (Stevens & Klatt, 1974) all produce categorical perception in nonspeech and cue respectively the features of affricate/fricative distinction, voicing, and the voiced-voiceless distinction in speech. In addition, Stevens (1971) showed that a sharp increase in intensity at a specific frequency range in the spectrum revealed the presence of a stop consonant. As mentioned earlier, Kuhn (1975) may have found evidence that a more context-dependent feature, place of articulation, can be acoustically analyzed directly from the speech signal.

The studies on the auditory-phonetic processing distinction support the view that there are various levels of auditory processing. Central levels of auditory processing perform more complex analysis on the results from preliminary analysis. The preliminary analysis component in the analysis by synthesis model would correspond to a central auditory component, such as described by Wood (1974). Forming the distinction between a central auditory and a central phonetic com-
ponent will explain the results of Pastore et al. and Klechner et al. that seemed to indicate only auditory processing took place. In these two studies the stimuli underwent extraction of acoustic attributes that were decoded into "feature" information, but in the absence of phonetic context they were not judged as "speech" and the psychological transformation from sound into "named" sound did not take place, engaging the central phonetic component. That phonetic processing exists cannot be denied. It involves the more abstract transformation of the invariant acoustic attributes into features.

Control and Generative Rules — These two components are the "hypothesis" of this model and by their very nature defy investigation into their existence. The control component uses the partial feature specification that results from preliminary analysis and forms a hypothesis as to its identity. It has available for its use information from analysis of adjacent segments, analysis of earlier parts of the utterance, the lexicon and results from the comparator. Knowledge of the lexicon provides information about what sequences may be expected in the particular language. Little is known about the form that the lexicon assumes, except that it would be unreasonable to expect that every possible sequence for every speaker under every condition could be realized, let alone stored. Therefore a model with recourse to a generative rule system is required, rather than one where direct matching occurs between the input and a lexicon.

The hypothesized representation is transmitted to the generative rules in the form of morphemes, segments and features. The generative rules yield a representation of instructions to the arti-
calculators that would be necessary to generate the utterance. Stevens (1960) recognized that the representation could exist at one of three levels - acoustic, anatomical or neurophysiological - but did not commit to model to one. The generative rules could be used to generate actual speech, given a set of phonemes as input and actualizing the peripheral oral musculature.

The control unit and generative rules form the phonetic processing component in this model. In order for a signal to be processed in this manner the distinction between speech or nonspeech must have been made prior to the activation of the synthesis loop. It is necessary that peripheral auditory analysis be performed, before preliminary analysis, (or central auditory analysis) and the phonetic processing. The model operates in serial order with some parallel recourse between central auditory and phonetic processing, in much the same manner as that proposed by Wood (1974) and Blechner et al. (1976).

The Comparator — The comparator calculates the degree of match between the articulatory representation and the acoustic attributes of the signal under analysis, that reside in temporary store. The degree of match is communicated to the control component and the hypothesis is either accepted or rejected. If the hypothesis is rejected the control component uses the error detected in the comparator to form a new hypothesis. The loop is transversed until a match is successfully achieved.

The authors proposed that because much acoustic information is already in a form that is related to features, the initial hypothesis will usually be correct, yet the matching process is always employed as a check.
Stevens and House provided essentially two direct sources of evidence to support comparison with articulation during perception. When a subject is required to repeat a CV nonsense syllable from an ensemble of varying size, Saslow (1958) found that the reaction time is independent of the size of the ensemble and therefore postulated that the mechanism required does not depend on anticipated features but that a direct recoding is made from the stimulus pattern to the motor response.

Kozhevnikov and Chistovich (1965) examined the courses of events that take place as a speaker repeats an utterance originated by another speaker. They measured tongue, palate and lip contacts and nasal closure as they occurred in the original production and the repetition. The experiments used VCV and CV syllables with the consonant changing values. The repeater was found to constantly correct and refine his articulatory position for the consonant as more information became available to him from the speaker. The repeater approximates the consonant he is to repeat, but does not release it until the speaker has effected closure and release, and the repeater's articulation correctly approximates the position of closure. (About 100 msec later the repeater releases his consonant.) Kozhevnikov and Chistovich hypothesized that the states that the producer goes through in this case parallel the perceptual mechanisms that are taking place. Each new state depends on previous analyses and modifies the next state. For this to be the case, they argued there must be a common underlying feature set and these features form the substructure of phonemes, in articulatory
terms anyway. Kozhevnikov and Chistovich proposed that the speaker refers to an "inner state" that is reflective of the information received to a specific time and that is continually reviewed as more information is added.

The research reviewed above also supports reference to articulation during perception. The different values that speakers of different languages produce for phonetic classes is also reflected in the exact boundary position that they determine in perceiving phonetic classes (Lisker & Abramson, 1970). Similarly after adaptation with a token from one phonetic class, subjects do not only shift their perceptual boundary, but a shift occurs in the value that they produce also (Cooper, 1974b). Hence the reference to articulation during perception is ongoing. Certain relationships are not established and then left to be used. The link is actively reinforced with each production and perception.

The research of Kuhn (1975) also supported that there exists reference to an articulatory representation during perception but the inference here is that synthesis is not required to accomplish the link as the information can be directly extracted through auditory analysis. Stevens (1972) studied spectrographic data and computations from a model of the vocal tract to describe possible acoustic correlates of phonetic features. He found that spectral patterns associated with place of articulation do not change continuously but in quantal steps, within which a change in the point of constriction produces little acoustic effect. These plateaus are bounded by abrupt acoustic discontin-
uities. These plateaus correspond to place of articulation in many languages and therefore the origin of phonetic categories may lie in the human vocal tract.

Stevens and House stated that a natural explanation for the "speech perception mode" arises if the model contains a component that establishes a correspondence between auditory and articulatory patterns.

"After processing by peripheral structures, some attributes of an incoming auditory pattern are then, as it were, looked up in the dictionary of auditory-articulatory correspondences. If a correspondence is found - i.e., if it is established that the stimulus is of a class that could have been produced by the human articulatory mechanism - then the speech perception mode is brought into play, and in subsequent processing, use is made of articulatory information derived from the dictionary [p. 54]."

The requirements that the signal need meet to enter "speech perception mode" must equal those needed for a signal to be processed phonetically. From the foregoing discussion on phonetic processing it is obvious that detection of a specific acoustic dimension "peculiar" to speech is not sufficient for phonetic processing to occur. That is to say, that nonspeech containing an acoustic dimension found in speech will not be phonetically processed, yet it will undergo complex auditory analysis and reveal some characteristics similar to speech. The acoustic dimensions that characterize speech need to exist in phonetic context, and hence in their natural state before they are processed as speech. The distinction is a psychological one, a division of stimuli into natural classes, like those found in general psychology. In the case of speech the natural classes form auditory concepts, or the names of sounds. These names exist in the form of symbols (phonèmes)
or in the form of acoustic patterns. It follows that division of acoustic stimuli into these classes relies upon identification of the stimuli as a distinct product of the vocal tract. Stevens and House quote Hayek (1962) as writing [p. 362] that "identification" means that

"some movement (or posture, etc.) of our own which is perceived through one sense is recognized as being of the same kind as the movements of other people which we perceive through another sense [p. 54]."
Chapter 3

THE USE OF THE REVISED MODEL BY THE CHILD

3.1 The Literature on Children's Perception and Production, Re-reviewed

3.1.1 Literature Evaluating the Relationship of Perception to Production

Experiments that have been performed using adults as subjects to investigate the processes involved in speech perception, have been paralleled using infants and children as subjects. Such categorical perception studies revealed that by one-month of age the human can perceive categorically, acoustic stimuli that vary along the acoustic dimensions that specify the features, voicing and place of articulation (Eimas et al., 1971; Morse, 1972). The infants identify the stimuli about the same boundary as determined by the adult and their discrimination functions reveal sharp peaks at this boundary and troughs within a phonetic class. In view of the total knowledge about categorical perception, it is obvious that the infant utilizes an innate ability of his auditory system to perform in this manner. No more than auditory analysis, the same as that used for nonspeech stimuli containing similar information, is carried out. At the age of one-month or in the case of the chinchilla, no knowledge of language as a system can be expected. Until the child has experience in producing sound combinations and can attend to the language system, he has not realized his potential as a user of that language and the phonetic percept cannot be realized.

Having seen that adults and preverbal infants perceived a series
of synthetic speech stimuli that varied in VOT categorically, Wolf (1973) assessed the ability of kindergarten and second-grade children to do the same. The two groups of subjects performed an identification task and a same-different AX discrimination task. The stimuli were nine tokens of /ba/ or /pa/ with VOTs ranging from -10 msec to 70 msec.

Both groups determined a phoneme boundary that agreed with adult values reported by Lisker and Abramson (1970). The obtained discrimination functions followed the predicted discrimination functions based on absolute categorical perception, but the obtained discrimination was somewhat lower at the peaks and higher at the troughs.

A second experiment was performed to test the generality of the findings from Experiment I and to determine whether the task order had affected discrimination of the speech stimuli. The stimuli were a set of /da/ and /ta/ tokens varying in VOT and the discrimination task was presented before the identification task. Again the groups determined phoneme boundaries that did not differ significantly from each other or from the values determined by adults. The discrimination functions were very nearly categorical but again discrimination was poorer than predicted in the region of the phoneme boundary and better than predicted at the end of the continuum. The order in which the tasks were presented did not affect discrimination.

Comparison of the children's identification and discrimination functions with those of the adult revealed that children identify the stimuli only slightly less consistently than the adult and at very similar phoneme boundaries. Their discrimination though is consistently poorer than the adults' at peak levels. Wolf explained that the children's discrimination was poorer due to extra-linguistic aspects of the
task. The subject was required to identify two stimuli and remember the two successive stimuli before making a same-different judgment. Therefore effects of memory, attention and the cognitive same-different judgment may have affected the discrimination results.

The nature of the discrimination task may explain the children's improved discrimination of items within a phonemic class. The AX same-different discrimination task permits direct comparison of the stimuli from auditory store without coding into phonetic classes. Supposing that at least some traces of consonants are represented in this store, it follows that their discrimination would be better in such a task than an ABX discrimination task. Regardless of these minor variations, the results of this experiment demonstrated that like infants and adults, children also perceive VOT continuum categorically and therefore the mechanisms involved are biologically innate.

Chapter 2 reviewed studies of children's acquisition of their phonological/phonetic knowledge. These studies indicated that children's perceptual abilities are more advanced than their productive abilities and hence implied that the analysis by synthesis model of speech perception could not be operative during this stage of language acquisition. The evidence will now be reviewed again and reconsidered in view of the revised analysis by synthesis model to determine whether the children could be using such a model for speech perception.

In the studies by Shvachkin (1973), Garnica (1973), and Barton (1975b) the child was required to make a distinction between phonological contrasts that resulted in a meaningful difference between stimulus items. Hence the child was making a phonetic discrimination in a
linguistically relevant manner and was not just listening to the "music" of language. The results of adaptation and categorical perception studies indicate that the child can make discriminations from a very young age based on the results of auditory analysis. It depends upon the nature of the task and the nature of the distinctions that the child must attend to, to reveal the type of processing that the child is performing. In order for the child to make a distinction that forms a phonological contrast, that distinction must have been conceptualized in the child's system and hence have achieved the phonetic percept. It may be expected from the outset, that as a child gains knowledge about production, this knowledge will aid in his conceptualization of sounds and their function in the system and hence, aid in phonetic processing also.

It has been shown that Shvachkin's results do not have the universal validity that he supposed they had. Nor can this piece of research be reliably assessed due to its methodological weaknesses and lack of quantitative data. It produces two interesting facts though — the children did perceive contrasts before they could produce them and although the results were not invariant and universal, some general trends appeared when compared to Garnica's work.

The fact that the children could differentiate stimuli before they could produce these differences can be explained by analyzing the task they performed. The children all had some articulatory skills, they were equipped with the ability to make auditory discriminations and they indicated that they were aware of the role of these distinctions in the language system. From Shvachkin's vague account it is probable that in many instances the children were taught the phonological dis-
tinction before being tested. The children learned the association between the stimulus and the corresponding article and in the task relied upon this learned association and the abilities of the auditory system. Subjects were successful as long as the opposition being tested involved acoustic differences that could be assessed either relatively directly by auditory analysis or from the child's existing phonetic system. In this manner, the child underwent a "forced" and probably premature learning experience. During the normal course of learning such distinctions, it is likely that matching of a perceived distinction with different lexical items takes place. The child learns the function of the distinction within his language system. When the child has achieved the ability to produce a distinction, the perceived phonetic distinction is matched with the articulatory distinction and the information required by the comparator unit of the analysis by synthesis model is established. It is expected that the ability to perceive distinctions is available to the child before the ability to produce them because of the innate characteristics of the auditory system. When the corresponding articulatory data is also available, phonetic processing no longer occurs in a "trial and error" manner but is more consistent. This notion will be discussed further with reference to Barton's study (1975b).

That there exist differences in the order of phonetic acquisition between languages is not surprising. The phonemes contained in different languages are composed of different combinations of features. The features that differentiate two phonemes must be discriminated with respect to the rest of the phonetic information in the stimulus complex. Certain combinations of acoustic attributes may result in a stimulus
complex that contains non-contradictory information. A feature contained in such a stimulus may be readily discriminated and therefore discriminated by the child before other stimuli containing that feature. Combinations where there exists conflicting cues about a feature's identity may be expected to be acquired last by the child. Nor can it be expected that a specific feature will be perceived in all phonetic contexts instantaneously. The acoustic attributes that correspond more or less directly with certain features may be expected to be perceived earlier than those that have not been shown to have any type of invariant relationship. These attributes will be perceived easiest in phonetic contexts where there exists no other acoustic cues that contradict their specification of the feature. For example, the presence or absence of a stop consonant, cued by a sharp increase in spectral energy after a period of very low energy, should be an early feature acquisition. In both Shvachkin's and Garnica's studies detection of the presence or absence of a stop consonant preceded most other feature distinctions.

Barton's (1975a) legitimate criticism of the criterion used by Garnica to determine success, renders Garnica's results uninterpretable. However again the same issues can be raised as for Shvachkin's work.

Barton's later study (1975b) produced three interesting results. First, the correct responses were more consistent for words that were learned than those that were taught. Second, a consistent pattern of results existed across all subjects. The individual differences were a matter of degree, not of being able to perceive a contrast or not. Third, features were not discriminated equally well when combined with different combinations of acoustic information for specific contrasts.
That is, the feature 'voice' was more easily detected in the /k-g/ pair than in the /f-v/ pair.

Discrimination of learned words involved contrasts that the child was able to articulate. Presumably the articulatory distinction aided in perceiving the phonetic distinction. The fact that the subjects could be taught the words and then perform the task indicates again that the ability to perceive distinctions precedes the ability to produce them. By the time that the child could produce the distinction it might be expected that the ability to perceive it was simply more established. However once the child can articulate the contrast, the phonetic "meaning" of the contrast has been firmly established and the phonetic percept has been achieved. Barton's study reveals the difference between the ability to make a phonetic discrimination based solely on the results of auditory analysis and based on acoustic-articulatory correspondences.

The fact that the subjects could all perform the discriminations but that individual performances differed by their consistency, supports the notion that all normal children have the same innate ability to process stimuli in their auditory system, but that each individual learns to use these abilities at his own rate and in his own order, i.e. by using individual strategies.

The third consideration reveals that phonetic processing involves learning to recognize the existence of features in a complex, rather than in isolation. The auditory system is able to identify and discriminate isolated acoustic cues that resemble those found in speech, but the recognition of the total stimulus complex evokes use of the left hemisphere when the complex is recognized as a component of the language
system that needs decoding. Such recognition is achieved when the auditory-articulatory correspondence is discovered by the child. The correspondence may be discovered before the child has acquired the motor skills required for articulation. But when the articulatory skills are developed the link is established that will permit the acoustic patterns to be traced to their articulatory source and hence, consistently phonetically perceived.

Kornfeld's (1971) hypothesis that the child produces what he perceives and does so according to a system that is more abstract than the adult's is worth of careful consideration. The evidence for this theory comes from spectrographic analysis of children's productions that reveals that the children mark distinctions that are not heard by the adult. The adult would assign the production to a different phonetic class but the acoustic data shows that the child has marked distinctions that are not found in either phoneme. It is difficult to understand how a child whose auditory system has been shown to perform the same acoustic analysis and make the same discriminations as the adult, could hear both the adult production and his own production as anything other than that heard by the adult. In this case the child, using his abstract system would perceive the adult form and his form as being identical. His production matches his perception. Several difficulties arise from such a theory. First of all, if the child hears both forms as being identical, what promotes the development of his system steadily towards that of the adult? Related to this question, Kornfeld maintained that a child rejects the adult's imitations of his productions because the adult does not mark the essential differences that the child perceives. But numerous reports on the child acquiring language report that the child
is aware that his productions are not in adult form. If this is indeed the case then the child can perceive the difference between his form and the adult form and he rejects the adult imitation because he can detect the difference and expects the adult form from the adult. This behaviour is repeatedly reported in Smith (1973).

Kornfeld (1976) elaborated upon the reasons for the child's use of a more abstract system. She supposed that certain combinations of acoustic information can be ambiguously interpreted during analysis and therefore result in a "misperception". In other words, the child's perceptual abilities are developing as well as his productive abilities. His more abstract perceptual system is explainable in terms of the interaction of acoustic attributes in a stimulus complex.

Studying words with initial consonant clusters, Kornfeld cited examples of phonetic contexts where there exists conflicts of acoustic cues. She argued that lacking the adult's knowledge of morpheme structure constraints and phonological redundancies, the child is more reliant upon phonetic information during speech recognition. The misperceptions result not from imperfect "tuning" at the level of auditory analysis but from conflicts that arise at the level of phonetic processing.

Kornfeld outlined the morpheme constraint knowledge that the adult has available for his use when decoding $C_1C_2V$ and $C_1C_2C_3V$ clusters. The morpheme structure for English immediately predicts that for a three-segment cluster the first segment must be /s/, the second a voiceless plosive, and the third either /r/ or /l/. For two segment clusters not containing an initial /s/, the first plosive may be voiced or voiceless, the second segment, an /r/ or /l/. In this paper, Kornfeld
discussed the perception of the voicing contrast because there is much information available on its acoustic-phonetic relationship.

In addition to the determination of absolute VOT, Kornfeld listed five other acoustic patterns that have been shown to serve to judge voicelessness in a segment. These include cues such as the absence of a rapid spectral change at the onset of voicing (see Stevens & Klatt, 1974, above), higher onset frequency of F₁, the presence of aspiration, greater intensity and duration of the burst phase and a high pitch on adjacent vowels. In different contexts these dimensions vary in their effectiveness to cue the voiceless distinction. Those cues that can override other conflicting cues would be used primarily to make the phonetic decision. In CV contexts where the consonant is a stop, it is primarily the VOT and detection of a rapid spectral change at voicing onset that are used to make the voiced-voiceless contrast (Stevens & Klatt, 1974).

In #stop - /r/ clusters, the presence of the /r/ causes the VOT for the voiceless stop to be increased by 30-50 msec. This change also causes the F₁-transition to be lengthened, which lowers the F₁ onset frequency and therefore produces cues for a voiced plosive. Kornfeld argued then that if the child is using F₁-transition data as a primary cue, he will perceive the stop as voiced. She supplied data from Ingram (1974) and Smith (1973) that reveal the child substitutes a voiced stop singleton consonant for such clusters.

Stevens and Klatt (1974) explained that the lengthening of the VOT in such clusters corresponds with an increased duration of the /r/ segment. The sonorant is lengthened so that the voicing onset is not
concurrent with the rapid formant motions of the sonorant vowel transition and does not falsely cue a voiced segment. Hence conflicting cues for voicelessness, i.e. increased VOT and a rapid spectrum change at the onset of voicing, are resolved by an articulatory adjustment, lengthening of the sonorant. In the child who has not achieved the articulatory skills necessary to produce a cluster, he can not use such knowledge to unambiguously perceive the cluster produced by an external source, the adult, let alone modify his own production based on auditory feedback so that he perceives his productions as correct. While Stevens and Klatt provided evidence that the $F_1$-transition cues and the VOT cues are unambiguous in nature, Kornfeld provided evidence that the $F_1$ onset frequency is a conflicting cue. It can only be assumed that the child will use whatever cues he has previously found to be successful for making phonetic decisions. Until the child can relate the stimulus complex to the articulatory adjustments required to produce it, his analysis may rest upon ambiguous cues.

Kornfeld reported that children often analyze and produce an adult cluster as an affricate or a single initial consonant marked by heavy aspiration. Haggard (1973) and Klatt (1973) showed that in /tr/ clusters, the delayed onset of voicing may be accompanied by turbulence prior to the first glottal pulse on the following vowel. Kornfeld proposed that the child may turn to the turbulence and loudness as cues when VOT fails. In adult utterances such a cluster may become affricated due to an open glottis that allows greater airflow and greater acoustic intensity. Again, not having the articulatory correspondence to rely upon for interpretation of the acoustic information contained in the cluster, it is
likely that the child may interpret the /tr/ as its homorganic affricate /tʃ/. As a third example, Kornfeld presented for discussion the cluster /gr/, when the VOT value for /g/ extends into the normally voiceless category. If the child were depending upon canonical VOT value to mark the distinction, he would perceive the stop as being voiceless and aspirated. Kornfeld presents examples from the literature that support the occurrence of such substitutions in children's speech.

Stevens and Klatt (1974) offered an articulatory explanation for the increased VOT values that exist for voiceless aspirated stops in the velar position. As the articulation for /g/ is performed also with the slow-moving tongue body it might be expected that VOT is somewhat lengthened although not enough to ensure that the transition is complete before voicing commences, as for /k/. Hence if the child lacks knowledge of the articulatory manipulations required to co-ordinate a complex of non-conflicting cues, he can not rely upon them to process external auditory stimuli and hence may use a more basic cue such as the absolute VOT value. In this case, not knowing that the VOT value for a /k/ in a /kr/ cluster is extended to the length that it is, the VOT for the /g/ in a /gr/ cluster may be judged as voiceless.

Finally, the case of voiceless unaspirated stop consonants contained in clusters beginning with /s/ was cited. Here both the VOT value and the $F_1$-transition cues for voicing indicate that the consonant is voiced. The cues of burst intensity and duration also go in the wrong direction. Davidsen-Neilsen (1969) reported that when the /s/ segment is removed, 92% of the subjects' responses identified the re-
sulting stimuli as voiced. Kornfeld cited the high frequency of such forms produced by children as substantiating that they misperceive the voicing of the plosive in such clusters.

In summary, two important conclusions can be drawn from this study. First of all, children may use certain acoustic cues to make the voicing distinction, that will lead the child to a misperception. With the acquisition of the articulatory knowledge required to produce the segment under analysis, the child learns to identify the acoustic cues with their source and hence learns to "dis-ambiguate" any conflicting cues. (Evidence has been provided to show that during articulation the cues are "dis-ambiguated".) In order to articulate a sequence the child must have as input, a sequence of phonemes, hence he must have conceptualized the existence of the entities in the utterance. To do this, the child must trace the acoustic attributes to their source through manipulation of his articulatory and auditory abilities. Second, once the possible phoneme combinations are familiar to the child, then he has knowledge of morpheme constraints in his language, and this knowledge can be utilized to aid in decoding the speech signal especially in cases where the acoustic information is contradictory, such as in some clusters.

One objection can be raised to Kornfeld's hypothesis. If the child's production does reveal his perception of the adult form, then the child does not discriminate any difference between the two productions. In this case, the external model and the child's auditory feedback are equivalent and he could assume that his articulation was correct. Without some kind of more sensitive self-monitoring it is difficult to see how the child's forms advance towards the adult's. In the
example cited earlier from Smith (1973) the child was aware that his production was not equivalent to the adult's. Smith assumed that the child perceives in the same manner as the adult. Kornfeld, on the other hand has presented evidence that the child's perceptual abilities are developing, as well as his production abilities. Taking both of these viewpoints into consideration it is most likely that the child's perception abilities are more advanced (yet not in adult-form) than his production abilities. The child can perceive the difference between his form and the adult form, yet until he is able to develop the motor skills required to equate the forms, he is not able to perceive what constitutes the difference. This is accomplished by being able to trace the auditory stimulus back to the articulatory configuration that produced it. Hence the child develops his productive repertoire by matching his auditory feedback to the external model.

The fact that Smith's son, A, could discriminate the objects by name, before he could speak at all indicates that he was able to do so by referring to the results of auditory analysis and forming an association between objects and their names. He was not required to identify what the difference was between the lexical items and therefore did not need any articulatory knowledge. The task resembled an auditory discrimination task.

3.1.2 Literature Examining the Child's Perception of Speech

Tikofsky and McInish (1968a) tested the ability of seven-year olds to discriminate stimuli that varied in the initial segment. The subjects performed a forced-choice same-different discrimination task.
The stimuli were 120-item lists composed of word-word pairs, word-nonsense syllable pairs or nonsense syllable-nonsense syllable pairs. The initial segment in each pair differed in one to five distinctive features.

The subjects were able to discriminate well by seven years of age and only 2% errors occurred. Most errors occurred when the consonants differed by only one distinctive feature. Some features were more discriminable than others and within a distinctive feature category some pairs were more discriminable than others. This last finding indicates that the stimulus complex as a whole undergoes phonetic analysis and that the interaction of acoustic information may either facilitate or impair the recognition of specific features. In cases where the acoustic information produces conflicting cues the child may depend upon misleading cues and further impair his ability to make the relevant discrimination.

Most of the errors were based on the features place or voicing. The highest error rate occurred for the /f-θ/ and the /v-ɹ/ pairs. These pairs have also been found to be poorly discriminated by adults (Tikofsky & McInish, 1968b; Miller & Nicely, 1955; Strevens, 1960). The errors that occurred for voicing occurred in the pairs /v-f/, /z-s/ and /f-θ/.

Abbs and Minifie (1969) examined the acoustic cues that were found to be important for perceptual discrimination of fricatives by preschool children. Each of /f/, /v/, /θ/, /ɹ/, /s/ and /z/ were paired with each of /a/, /i/ and /ai/ in VC and CV combinations. The syllables were paired so that they differed only in the consonant. Seventeen children ranging in age from 3;0 to 5;1 were subjects. As each syllable in a pair was presented, one of two pictures in front of the subject was
illuminated. The subject was then asked, "Who said ____?" After responding the subject was either reinforced or corrected, allowed to change his response, and then reinforced.

The results showed that there were no differences in errors due to the changing of vowel environment, yet overall there were significantly fewer errors when the consonant followed the vowel than when it preceded the vowel. However the pairs /v-z/, /f-s/ and /v-d/ produced more errors as VC than as CV, contrary to the overall trend. The most difficult pairs to discriminate were /f-x/ and /v-d/, while the easiest were /θ-z/ and /s-z/.

The stimuli were then analyzed to determine what acoustic cues were utilized by the subjects in performing the discriminations. Duration, intensity and spectral measures were analyzed.

Fricatives in final position were found to be longer than those in initial position and unvoiced fricatives were longer than the voiced fricatives in either position. /s/ was formed to be significantly longer than any of the other fricatives, while /v/ and /d/ were significantly shorter. The analysis of the ratio of fricative duration to vowel duration showed that the ratios for unvoiced fricatives were longer than those for the voiced fricatives. The position effect showed that in CV syllables the vowel is much longer in relation to the preceding fricative, whereas in VC syllables the vowel was only slightly longer or shorter than the following consonant. The differences in the C:V ratios reflect changes in both vowel and consonant duration. An important cue for the perception of voicing may be the ratio of these durations.
/s/ and /z/ were found to be 10-15dB more intense than the other fricatives. There were no differences in intensity that corresponded to the position of the fricative in the syllable or voicing differences.

Spectrographic analysis revealed that /s/ and /z/ have resonances at the high end of the spectrum while /f/, /v/, /θ/ and /ð/ are significantly lower. There were no differences in center frequency associated with position in the syllable. It was also determined that the half-power bandwidths for /s/ and /z/ were considerably shorter than those for the rest of the fricatives. Therefore /s/ and /z/ are more intense and have major resonances at a higher frequency than the rest of the fricatives.

Overall, fewer errors occurred when one of the consonants was voiced and the other was unvoiced than when both were voiced or unvoiced. Fewer errors occurred when the fricative was in final position rather than initial position. When one of /s/ or /z/ were compared with one of /f/, /v/, /θ/, /ð/ fewer errors resulted than when two members of the latter group were compared with each other.

The voiced-unvoiced distinction for fricatives appears to be facilitated especially in VC syllables by the ratio of the consonant to vowel duration. The C:V ratios are significantly larger for VC syllables containing unvoiced fricatives than for those containing a voiced fricative. Although similar changes occur in CV syllables they are not large enough to be statistically significant. The other cues that may be used when the fricative is in initial position are the presence of a low frequency component (the voice bar) and the well defined formant structure of voiced fricatives. In the Miller et al. (1976) experiment it was de-
termined that subjects judged noise-buzz sequences to contain voice if the noise-lead was 16 msec or greater. If the noise lead was smaller than 16 msec the stimuli were identified as "no-noise". Although Miller et al. used nonspeech stimuli there may exist a parallel with the detection of voicing in fricatives. For an unvoiced fricative in initial position, the noise produced at the point of constriction extends for a period far greater than 16 msec and then voicing begins with the onset of the following vowel. Voiced fricatives in initial position have two components, the friction produced at the point of articulation and a low frequency voicing component from the glottal source. The voicing distinction may be determined by detecting the noise-lead time in a manner similar to that in the Miller et al. experiment. In addition, it was found in the study by Abbs and Minifie (1969) that voiceless fricatives have a longer duration than voiced fricatives, possibly to ensure a noise-lead. The voiced distinction may also be facilitated by the presence of strong formant patterns that are detected in a manner similar to those in stop-consonant CVs (Stevens & Klatt, 1974).

/s/ and /z/ are principally distinguished from other fricatives by their high resonance frequency and their relatively intense, short spectrum. They also have relatively high peak amplitudes.

The highest error rate occurred when two members of /f/, /v/, /θ/ and /ʒ/ were contrasted and there were no major voicing, frequency, bandwidth or intensity differences. But when two of these dimensions were contrasted in a pair the error rate dropped considerably.

Essentially the acoustic cues used by children to discriminate the fricatives were of a temporal nature or due to significant frequency
or intensity contrasts. These acoustic cues were used to distinguish the features + stridency and + coronal. When neither feature were present to be discriminated, the pairs were poorly discriminated. The type of analysis required to perform these discriminations resembles the auditory analysis that has been shown to take place for the categorical perception of stop consonants. Conceivably with further research, the type of auditory analysis necessary to perform fricative discrimination could also be shown to take place in a categorical manner. Cutting and Rosner (1974) have shown that variable acoustic dimension, rise-time, can result in the categorical perception of speech (to cue the affricate/fricative distinction) and of nonspeech ("plucks" and "bows").

Graham and House (1971) investigated the adequacy of current linguistic descriptions to describe the cues used by a child to deduce the adult's phonological system. Thirty girls from three to four and one-half years of age performed a same-different discrimination task on pairs of "words". The "words" consisted of the frame /haCad/ containing one of the sixteen commonly occurring English consonants. The words were paired either with themselves or another word in the group. Analysis of the errors revealed that the children made the same type of errors as adults, only more. Individual differences were a matter of degree, not type of error. The error rate was the highest when the consonants differed by only one distinctive feature. When the stimulus pair differed by two or more distinctive features, the error rate dropped and then levelled off.

Graham and House also examined the "equivalence" of distinctive features for purposes of discriminability. When one-feature contrast
pairs were examined, some features contributed more to discrimination than others. However these features were not "additive", i.e. when combined so that a pair differed by two highly "discriminable" features, the result was not always a highly discriminable pair.

The authors used multidimensional scaling techniques to determine if they could find the indicated four dimensions that would account for ninety-percent of their data's "fit", in terms of several distinctive feature systems. None of the feature systems would combine into feature combinations that describe those relevant for children's perceptual judgments. The authors concluded that while linguistic distinctive features may adequately describe the articulatory gestures used to produce speech they are not adequate to describe the perceptual parameters used by a listener in categorizing speech. The pattern of errors made by the children did support their coding acoustic information in terms of articulatory data such as manner, place and voicing. Therefore the children may use articulatory data when decoding the speech signal.

Considering the fact that Graham and House determined that the discriminability of features was not additive and that the interaction of acoustic cues often obscure the specification of the features present, it is not surprising that the authors failed to find an adequate system to describe the perceptual parameters used in decoding the speech signal. Unless more invariant acoustic attributes are found to correspond with feature distinctions such an analysis would not be possible. The results of this experiment again indicate that the processing of acoustic information about a feature is performed with respect to the total stimulus complex. The information pertaining to a feature is variable in nature
according to the rest of the acoustic information in the stimulus complex and interactions among the acoustic cues may yield conflicting information about the features that are present. It should also be noted in the Graham and House study that the pair /f-θ/ produced the most errors. Here is another example where the lack of a well-defined acoustic basis for a phonetic decision results in poor discrimination.

Moskowitz (1975) stated that a differentiation must be made between the child's acquisition of the phonetic system, i.e. learning to associate the articulatory gestures with specific acoustic effects, the variations imposed upon a sound by its phonetic context and the timing required to produce articulatory gestures with respect to each other, and the acquisition of phonology, i.e. learning the systematic aspects of the sound system including functional oppositions and the system in which they are functional. If this differentiation is kept in view then many apparent discrepancies in the child's system can be resolved. It is possible that the child's system may contain a phonological distinction yet he cannot phonetically realize the items or the reverse, that the child may be able to produce the items forming a distinction yet not use the items to form a phonological contrast. Both systems interact and either one may lead or lag behind the other, not revealing the true status of the other system. In cases where the child indicates that his perception is more advanced than his production, it is usually supposed that the child's phonetic system is not as advanced as his phonological system.

Moskowitz reviewed data from eight subjects to determine the acquisition of the eight English fricatives in both systems. She used
a substitution analysis framework. Several relevant observations were made. First the most common substitution for /θ/ is /f/, precisely the phoneme that is so poorly discriminated from /θ/ by children as well as adults. Second, the usual substitute for /ð/ was /d/. However on elicited imitation tasks the phoneme /v/ was frequently substituted for /ð/.

Moskowitz established the following criterion to establish that a phoneme has been phonologically acquired:

"A phoneme x, can be said to have been acquired when the pattern of phonetic realization of X is consistently distinct from the pattern of phonetic realization of any other phoneme, Y [p. 146]."

Using this criterion the voiceless fricatives /f,s,ʃ,θ/ were acquired before the voiced fricatives /v,z,ʒ,ɹ/. In addition all fricatives showed evidence of phonological acquisition before they were phonetically stable. When a phoneme has been phonologically learned, it has been established as an entity in the system of language. It has been distinguished from other phonemes as being acoustically different and being able to differentiate lexical items. The phoneme is recognized as a speech-sound contained in the system and hence becomes eligible for phonetic processing. Phonetic learning can now begin to determine what distinguishes the phoneme from other phonemes in articulatory terms.

The three unvoiced fricatives /f,s,ʃ/ occurred phonetically before the four voiced fricatives. Phonetically, /θ/ was acquired last, as the children persisted in substituting /f/ for /θ/ for long periods of time. The four voiced fricatives exhibited four different patterns of phonetic acquisition and did not duplicate the acquisition pattern of the unvoiced fricatives.
Moskowitz explained the alternative substitutions for /d/ as resulting from conflicts between the two sources of phonetic information available to the child, the perceptual mode and the articulatory mode. The child who has not mastered /d/ and who does not have /θ/ will substitute /d/ in his spontaneous speech. /d/ is his closest possible articulatory substitution and during production he attempts to modify the articulation of /d/ to match /d/. In the imitation task though, he does not need to maximize articulatory practice, but needs to maximize acoustic effect. To do this, he substitutes /v/, a closer match acoustically than /d/.

Moskowitz presented the suppression process as an example of a process that limits the child's phonetic capacities. Briefly, for each allophone of a phoneme, a speaker has recognized that there exists a nucleus of possible intended productions that result in acceptable instances of that phone. Outside this nucleus is a second layer of phonetic possibilities that are close enough that they may be considered acceptable during acquisition. A third layer exists beyond the second layer and it is composed of all the phones that could not serve as an acceptable realization of the phone in question. During acquisition the child learns the boundary between the "acceptable" and the "second layer" productions. In the process, he suppresses the phonetic realizations that exist in the second layer. If however two phones are being acquired and their "second layers" overlaps then this area will be strongly suppressed. If ranges of values are acceptable for one phone yet being suppressed for another phone, the acceptable ranges may shift to avoid a conflict of demands.
/θ/ falls into the suppressed layers of the phonetic realizations of /t/, /s/ and /f/. Phonologically /θ/ emerges but phonetically it battles with the phonetic realizations of /t/, /s/ and /f/ and hence, is triply suppressed. The child may substitute /f/ for /θ/ during this stage as /f/ offers the closest match acoustically. /f/ occurs as a substitute for /θ/ when he cannot approach its production in an articulatory manner. /d/ occurs as a substitute for /ð/ while attempting to modify his articulation to approach /ð/. /v/ occurs as a substitute for /ð/ in an imitation task because it is an expedient acoustic substitute.

Moskowitz then, outlined the information that a child has available for acquiring phonetic realizations. He uses his articulatory knowledge and his perceptual knowledge to achieve the realization in the most expedient manner for the task at hand. This account suggests that the child monitors his productions and attempts to modify his kinesthetic feedback until, on successive attempts, a match is achieved between his auditory feedback and his perception of the external model. Until he has achieved this goal his productions will be composed of segments that are either a close articulatory match or a close acoustic match.

Menyuk (1973) surveyed data collected on the mastery of speech sounds by American and Japanese children to determine if there existed a specific order for the acquisition of feature distinctions. This data was compared with data on sound substitutions made by children with articulatory problems and data on perceptual confusions made by adults. Menyuk hoped to analyze these collections of data and derive some information on the cues used in perception and production of consonants by children during the developmental period.
The features examined were gravity, diffuseness, stridency, continuancy, nasality and voicing. The data on the mastery of features was determined by calculating the percentage of sounds containing a feature that were produced with this feature at various stages during the children's development. Menyuk did not state whether both members of a feature opposition were required before a feature was considered learned. As no mention was made concerning use of the opposition in a meaningful way, it is assumed that Menyuk was examining phonetic capacities not phonological abilities. The specific phonemes in which features were contained were not given so no information can be derived from this study, concerning the specific physiological and acoustic parameters that were used to make these distinctions.

The data showed that Japanese and American children mastered the features in a very similar order. The order derived did not correspond well with data collected on the frequency of occurrence of features in the adult language. Therefore Menyuk gave little support to the theory that children first produce the features that they hear most frequently.

A comparison was made between the substitutions made by the American children and the group with articulation problems, with the perceptual errors made by adults. The normal and the deviant groups performed in a similar manner, but the deviant group generally maintained features less well. Stridency was best maintained by adults in their recall of CV syllables, while it was the feature least maintained by the children in the "articulation problem" group. Voicing and nasality ranked high in all groups. Substitutions made by the normal children correlated very well with the adult perceptual errors, indicating that there
may exist some underlying acoustic cues in those features that are not readily perceived, recalled and produced.

The features + nasal and + voiced ranked high in the maintenance of features by both groups of children and the adults' perceptual errors. Therefore they may be the easiest features to perceive, recall and produce. The feature + strident was best maintained in the recall of consonants by adults. It was among the last of the features to be acquired by the American and Japanese children, yet it was best maintained after the feature + voice and + nasal in the substitutions of normal children. This feature ranks last though, in the features maintained by the children with articulation problems.

Menyuk proposed that some features are more easily mastered due to their on-off characteristics. That is, for the feature + voice the folds either vibrate or do not, and the feature is either present or absent. Features such as + continuant, though, cannot be as clearly delineated; the acoustic signal lasts for sometime longer than a burst, but cues are not clearly specified in absolute terms.

The features that were best maintained by the children (+ voice and + nasal) were found to be discriminated by children with varying efficiency depending on the phonemes that contained the contrast (see Graham & House, 1971). For example, the feature voice was discriminated well in the pair /t-d/, yet poorly in the /s-z/ pair. Features may be detected through basic auditory processing in some stimulus complexes and not in others. The specific acoustic information that needed to be decoded in the above results cannot be examined because specific contexts of occurrence were not provided and features are manifested in
different manners in different phonemes.

Menyuk (1971) sought evidence that phonological processing takes place in parallel with semantic-syntactic processing. Nonsense syllable sequences and three- to five-word sentences were presented to a normal and a language delayed group of children for repetition. Menyuk argued that the language delayed children should not perform as well as the normal children if semantic-syntactic information was utilized during phonetic processing.

The results indicated that language delayed children made more consonant errors when repeating both types of stimuli. The largest number of errors was made by both groups when repeating nonsense syllables. The language delayed group could not analyze phonological sequences in terms of segmental or syllabic phonological features. When the substitutions made by both groups were analyzed a similar feature maintenance pattern resulted. Although the scores deteriorated when the stimuli were meaningless, the pattern of recall was the same for both nonsense syllables and words.

Menyuk concluded that phonological processing into lexical representations must include consideration of semantic information and that "phonological sequences are identified first in terms of lexical form (semantic and phonological features) and then in terms of their speech sound sequences - syllabic or segmental [p. 188]". It is interesting also to note that the same pattern and type of errors occurred for both sets of stimuli. This infers that while semantic information may improve positive identification of a speech sound sequence, all auditory stimuli undergo the same basic auditory analysis (including non-
speech). Furthermore a signal that has been judged to be a potential component of the language system will be processed phonetically and information from one level of analysis is available at another level. This is a requirement of the analysis by synthesis model.

Menyuk and Anderson (1969) examined the production and identification of the three liquids /w/, /r/ and /l/ by preschool age children. Adults and thirteen-year olds served as control groups. The authors wished to determine whether the children could distinguish a sharp phonetic boundary for these liquids, as has been found for stop consonant stimuli varying in several acoustic dimensions. The values produced by the children were also recorded.

Results showed that the children did not establish sharp phonetic boundaries for the liquid sounds, either in their production or during the identification task. The children maintained a distinct boundary better in the perception task than the production task. The authors reasoned that when children are developing their phonetic system, they first identify differences between members in a set and then learn to produce them. Menyuk and Anderson were also concerned as to whether children match an articulatory gesture to a previously identified sound, or "mimic" sounds and base their perceptual categories on the resulting articulatory gestures. The adults in this study were found to observe the boundaries more consistently when both producing and identifying the three sounds. However, the thirteen-year-olds performed the best of all three groups. This result was interpreted as possible support for the use of the analysis by synthesis model of speech perception. At thirteen years of age, the time when linguistic knowledge is considered
to be fully developed (and the left hemisphere also), children will have the most reliable acoustic-articulatory data available for their use that they will ever have. Up until this point in their lives, they have been experimenting with articulation and its acoustic effects. Having the raw data that results from auditory analysis available, they then go about matching an articulatory configuration to a desired sound and hence build up a bank of articulatory-acoustic data. Furthermore this store can be drawn upon for phonetic processing. At this age then, the child's system is probably "tuned" the finest it will ever be. The adult loses the acuity of his auditory system and his discrimination abilities and production may deteriorate. The acoustic-articulatory image becomes less vivid. Menyuk and Anderson were unable to place much significance on this finding due to the small number of thirteen-year old subjects (N=4). It would be of interest if researchers further investigated any processing differences that do occur at this age.

3.2 Conclusions

The above research on children's acquisition of the sound system is consistent with what is known about adult processing of speech stimuli. The results can be integrated to show that the analysis by synthesis model of speech perception is used by children during the time that they acquire language.

From birth, the human auditory system is equipped to perform complex auditory analysis on speech or nonspeech stimuli. The nature of this processing involves the extraction of data from certain acoustic dimensions and the resulting division of stimuli into its natural cate-
gories. The categories perceived in speech processing characterize the "names" of sounds and it may be more than coincidental that the acoustic dimensions that differentiate many features correspond to distinctions that the auditory system determines readily. Accordingly, infants have been shown to be able to perceive speech stimuli in a categorical manner similar to the adult. In these studies, the stimuli vary in only one acoustic dimension. In the case of natural speech, each sound is composed of a complex of the acoustic data representing each feature. In these complexes the acoustic data interacts and can produce poorly defined, if not conflicting, data for a specific feature's identification. Also in different stimulus complexes the information for a specific feature is realized in a different manner. Hence features are not simply detected as an isomorphic mapping between acoustic information and feature. The child can be expected to first learn a feature in some contexts before others. For example, Barton (1975b) found that two and one-half year olds could readily discriminate the feature 'voice' for the stimulus pair /k-g/, yet Tikofsky and McInish (1968a) found that seven-year olds still had difficulty with that feature in the pairs /f-s/, /s-z/ and /θ-ʃ/.

There is a great difference in detecting one feature from one acoustic dimension and processing the entire stimulus complex to yield a set of features. Herein lies the difference between auditory and phonetic processing. What causes a signal to be recognized as belonging to the system of language and undergo left hemisphere processing? For speech or nonspeech, if the acoustic characteristics of a stimuli define it as an element of a system (therefore it may be classified) and the
listener has developed a concept about the existence of that element in the
system then special perceptual processing occurs. In the case of speech
the sounds are recognized as originating in the vocal tract and during
phonetic processing are traced back to their place of origin. Decoding
of the total stimulus complex is achieved through seeking the acoustic-
articulatory relationships.

The child may begin to process sounds phonetically when he has
recognized that the sound originated in the vocal tract, but does not
need to have mastered the articulatory realizations. The studies by
Shvachkin (1973), Garnica (1973) and Barton (1975b) required that the
child detect a contrast that produced a difference in meaning between
lexical items. To do this the child has formed a concept about the exist­
ence of the phonemes being discriminated. He may detect a difference
between two sounds yet until he has mastered their articulation he can­
not identify the difference. The phonemes become real for the child
when he can articulate them. Before then he uses what articulatory
knowledge that he has to approximate a phoneme during production. It
is postulated then, that the child begins to process stimuli phonetically
when he has formed a concept concerning the segment's place in the
speech system and begins to seek the articulatory configuration that pro­
duces the segment. Hence the articulatory pattern is not necessary for
a segment's phonetic interpretation, but the realization that the seg­
ment can be traced to its source and the active pursuit of the source,
is necessary. Once this information is available to the child, he no
longer depends so much on auditory analysis because he has a reliable
means of perceiving all tokens of the same stimuli.
Moskowitz provided evidence that the child has formed some relationships between acoustic and articulatory images during phonetic acquisition of a phoneme. In substituting /v/ for /ð/ the children were able to match acoustic effects when necessary, although in their spontaneous speech they provided an articulatory match, /d/. Hence the child has formed ideas about best matches in both articulatory and acoustic terms. In the development of articulation he modifies the articulatory gesture until it matches acoustically, the model.

Smith's son A was able to understand a recording of his speech provided he still was at the same stage. If several weeks lapsed before he listened to the recording, he could not understand his speech. The child must be able to use his articulatory data to perceive his speech in a manner similar to that proposed by the analysis by synthesis model. When listening to the adult though, if the adult used an "adult form" that also corresponded to a different lexical item in the child's speech, the child would interpret the adult meaning. In this case the child could not have used his articulatory data in a direct manner. However the child also indicated on numerous occasions that he was aware of his "incorrect" pronunciation. Provided that this is so, the child knew there was a difference yet could not perceive the nature of the difference in articulatory terms. He had a concept of the phoneme's entity and he could perceive the phoneme without being able to produce it.

In the analysis by synthesis model each time a sequence is processed the articulatory representation is activated and compared with the acoustic data. Hence the acoustic-articulatory link is reinforced. In a case where the articulation is not 'correct' this reinforcement...
serves to re-establish the difference between the existing production and the acoustic goal. Hence processing in this manner also explains the source of the child's awareness that his articulation is deficient. The comparator provides this information. During analysis of adult speech, the child uses his articulatory knowledge and can attend to the differences between his form and the adult form but identification of the sequence is more dependent upon auditory analysis during this stage. Therefore it can be explained why the more acoustically identifiable features are perceived first.

Perception then develops along with production, but precedes production by the interval extending from when the phoneme is recognized as a part of the system to when it is correctly articulated. During this interval the link between the perceived acoustic characteristics and the produced articulatory representation is established and the two forms are manipulated until they coincide. Once articulation is mastered, the child can convert a sequence of symbols (the "names" of the speech sounds) into articulatory representations and he begins to acquire knowledge of the morphological system. As morpheme constraints are learned, this knowledge further aids in perceiving sequences of sounds.

Although all human auditory systems are capable of the same auditory analysis from birth, all human languages do not contain the same combinations of features in their phonemes. Therefore it is unlikely that more than general trends of acquisition could occur due to the different facility with which features contained in different combinations are perceived.
Between the time of birth and adulthood, the human's perception and production of a feature takes on the specific values determined in his native language. Eimas et al. (1971) showed that infants in an English-speaking environment perceived the phonetic boundary for voicing at a location very near that found in adult English. Chinchillas did the same (Kuhl & Miller, 1975) so it is unlikely that infants in a Spanish-speaking environment would do otherwise. Lisker and Abramson (1970) found that Spanish speakers produce and identify a different phonetic boundary for the feature 'voice'. During the period of phonetic acquisition the child's auditory system becomes "tuned" to the specific values found in his language environment. There is some modification of the characteristics of innate auditory processing to specific values perceived in a language, and production takes on these values also. It would be interesting to further investigate at what age these language specific values for phoneme boundaries would be revealed in categorical perception experiments and whether production takes on the specific value from its onset or is modified at a later date.

The results of discrimination testing showed that children perceived in a similar manner to the adult, only poorer. Stimuli that were poorly discriminated by the adult were also poorly discriminated by the child. In these cases, it was shown that acoustic information did not clearly define features. In a discrimination experiment where the adult cannot resort to using contextual information for identification of a stimulus, performance is restricted to pure auditory-phonetic processing and a condition more comparable to the child's mode of processing results.
Hence it is not surprising that both populations made similar errors; they have similar auditory systems. Nor is it surprising that children are poorer at discriminating in many cases, because they are more dependent upon auditory analysis which is not as reliable as drawing upon acoustic-articulatory correlations.

In certain cases it was also determined that phonemes acquired last are among those that are poorly discriminated. The pair /f-θ/ were poorly discriminated by children and adults and furthermore, /f/ persists as a substitute for /θ/ for a long period in many children's speech. In this case the children are able to draw upon an acoustic substitute when an articulatory substitute is not available.

Menyuk (1973) proposed that some features are more easily perceived and produced due to their on-off characteristics. The detection of an acoustic attribute can be associated with the presence or absence of a well-defined articulatory motion. But Menyuk's study failed to examine the acquisition of features in specific contexts and features are not learned in all contexts simultaneously. Again the example arises that in Barton's study (1975b) the /g-k/ voicing contrast was easily discriminated but other voicing contrasts were not. The same result was found by Graham and House (1971). Hence again, it appears that the clear interpretation of acoustic data interacting in a stimulus complex, determines which features in their respective contexts are discriminated first. This discrimination is a prerequisite for the analysis of what differences exist between this stimulus and other stimuli and hence identification of the segment in articulatory terms.
The development of phonetic perception that occurs in this manner explains the development of the acoustic-articulatory relations that are contained in the comparator unit of the analysis by synthesis model. Once these relationships have been determined, the child is fully equipped to use this mode of analysis. But prior to this time how does the child use such a model? The child can perceive two forms of speech, his own and the adult form. The child can understand his speech during the brief time that he remains at that level. In order to do this he uses the analysis by synthesis method and processing employs his articulatory representations. In perceiving the adult though, he lacks the articulatory specifications for some segments. For those segments that are phonologically learned, he has discriminated the segment from others in the language and he has formed a concept of the segment's role in the language system. The child is also aware that his articulatory knowledge is not sufficient to realize the segment in production. But this very fact, that the child knows his articulation and the acoustic model do not correspond, explains how the child uses the analysis by synthesis model for processing and the value of its use. When a segment is analyzed that is beyond the child's articulatory abilities, a hypothesis concerning its identity is made by the control unit. The articulatory representation used by the child is generated and the comparator analyzes the difference between this representation and the results of preliminary analysis (auditory analysis). The results from the comparator serves to produce the child's awareness of the difference between forms, as the acoustic-articulatory differences are exactly what are determined. Information about the difference can then be utilized in the child's next production. Hence
perception actively modifies the child's productions and articulation ultimately aids in perception.

Stevens and House suggest that in speech processing a match of acoustic information with articulatory representations does not always need to be performed by the adult because often a reliable hypothesis can be derived from the results of preliminary analysis, however it is performed as a check. The child is aware that he cannot rely upon this match and relies more on earlier processing and his knowledge of his articulatory discrepancy from the model. If the child was not aware of his limited production capacities, he would fail to understand the adult's speech. This awareness originates in the comparator during perception of the adult forms but also occurs during the child's production.

In a theory that presumes perception and production are components of the same system it is not surprising to find that both processes interact in their development. The comparator with its vast store of acoustic-articulatory relations provides the "goal" information for production and attempts to determine how articulation is deficient, based on previous analysis.

When the child produces a phoneme not yet phonetically learned, he has a goal for this production, based on his acoustic and articulatory knowledge of that phoneme and the discrepancies determined by the comparator unit. During production the child receives kinesthetic feedback from the vocal tract and auditory feedback that is processed in the same manner as external stimuli. The hypothesis is precisely the goal that contains correctional articulatory information. The comparator analyzes this hypothesis with the actual articulation and computes the difference.
The articulators can then be informed how far off-target they were and this correction factor can be used in the next production. In this case a comparison is made between the actual articulatory representation and the hypothesized articulatory representation supplied by the comparator. The comparator can anticipate that the articulatory representation used in its next analysis of that segment will be in this form.

It is postulated that the child utilizes the analysis by synthesis model from the time he begins to phonetically perceive speech. While the articulatory representation may not be available for his use, processing utilizes the results of auditory analysis and the comparator's knowledge of the discrepancies to be expected between the child and adult forms. While the comparator recognizes that these forms are not identical, the child is aware of his immature production. During perception of the adult, the comparison of the acoustic model with the child's articulatory representation provides information that may be used later to modify production. During production the comparison is made between the actual articulatory representation used and the articulatory adjustments proposed by the "goal". This information is used to keep the comparator informed of the child's production compared with the goal so that the comparator can anticipate this relationship for its next perceptual comparison. Until the child has developed his articulatory competence and can trace stimuli to their origin in the vocal tract, it is to be expected that he will have difficulty perceiving sequences where clearly defined acoustic-phonetic information is at a premium.
Chapter 4

DISCUSSION

4.1 The Importance of Babbling

The introduction raised some issues that may be better explained with some insight into the perceptual process used by children. The child can perceive both his own speech and that of the adult. He does this using the same method of processing. When perceiving himself he is able to use the analysis by synthesis model in the same way that the adult uses it. To perceive the adult, he uses this model again but also uses the knowledge that the comparator contains about the relationship between the forms. This information is also used to modify articulation towards the external adult model. It should be noted that the comparator does not necessarily know the exact articulatory adjustments required to correct an articulation, but its knowledge is probably in general acoustic-articulatory terms and has been derived from previous experimentation with articulation.

The child's perception is not in adult form from a very young age. Although his auditory system is capable of analyzing stimuli in an adultlike manner, phonetic-perception develops along with, but probably slightly ahead of, phonetic production. These two aspects of the same system depend on each other for their development. Although the results of auditory analysis may contribute largely to phonetic processing it is the final resolution with an articulatory representation
that completes this learning.

One consideration of theories of phonological development is whether babbling contributes to acquisition of the sound system. During this period the child produces a complex variety of sounds, some of which are not contained in his language. He may not babble some sounds contained in his language. Therefore, babbling is not language specific. After this period, the vocalizations trail off somewhat and then assume qualities of sound sequences found in language. Sometimes after the child's first birthday, "words" appear; they are assumed to have some meaning for the child.

Deaf children babble as normal children do, but do not progress beyond this stage. It may be supposed that during the babbling period the infant is experimenting with his articulatory abilities, but not in a linguistically relevant manner. Babbling and its auditory feedback are necessary for the next stage of development where the acoustic-articulatory correspondences are established. The infant is born with innate auditory capabilities but although he may hear sounds he has not yet learned what to listen for. Similarly the infant is equipped with the oral musculature necessary for speech but he has not yet discovered the product that results from manipulating the peripheral oral structures. During babbling the child discovers the sensations associated with speaking and hearing, without perceiving their role in the system of language.

Studdert-Kennedy (1976) outlined a study by Marler (1975) in which he studied the abilities of several species of birds to develop normal bird song dependent on hearing their own productions and hearing an external model. The dove or chicken needs to hear neither an external model or its own voice for song to develop normally. The song sparrow
sparrow needs to hear its own voice but does not need an external model. The white-crowned sparrow needs both the external model and its own feedback in order to develop normal song. If it is deafened early in life a highly abnormal song develops. If reared in isolation an abnormal song with some natural characteristics will develop.

Marler suggested that this rudimentary song reflects the existence of an auditory template that consists of information for auditory processing about the structure of vocal sounds. Matched with an external model the template tunes the development of rudimentary speech to characteristics of the model. The bird gradually discovers the motor controls needed to match its output with the modified template. Without a model the template establishes some basic features of normal song in the rudimentary form.

Marler proposed that the child learns language in a similar manner. Early in life sensory mechanisms analyze the sounds of others and then turn to analyzing their own productions. On the motor side vocal development is dependent on auditory feedback and there has developed "neural circuitry necessary to modify patterns of motor outflow so that sounds generated can be matched to pre-established auditory templates [p. 33]."

Similar patterns of speech develop in the child who is born deaf or raised in isolation. The congenitally deaf child produces little if any, speech and voice with abnormal "deaf" qualities. If raised in isolation the child produces a highly abnormal pattern of speech (see Fromkin, Krashen, Curtiss, Rigler, & Rigler, 1974). Therefore there is support for some type of speech-related auditory sensorimotor mechanism that may modify patterns of motor output to match sounds generated by
the vocal apparatus against some standard. This sensorimotor interaction would provide the mechanism for discovering auditory-articulatory correspondences.

4.2 Implications For Research

The results of studies on children's perceptual status can be explained in terms of the child's use of the analysis by synthesis model. If the results are not sufficient to wholly support the existence of this model in children's processing systems, it is due simply to the lack of research in this area. While there are inherent difficulties associated with experimentation in this area, several theoretical issues arise that could yield information about the acquisition process.

Beginning from a very young age, it would be of interest to determine at what stage in the acquisition process and for what type of tasks, phonetic processing occurs. Gilbert and Climan (1974) found in dichotic listening experiment, that an REA was present in children as young as two and one-half years. Using recordings of average evoked potentials, such as that used in Wood (1975), some idea of what is necessary to evoke phonetic processing could be obtained.

More integrated research into both production and perception at successive stages in a child's acquisition of his sound system could yield more specific information on the relationship and interaction between these forms. Most of the present studies involve one of these processes and infer about the other. Analysis of the acoustic information contained in both forms may indicate what information is processed relatively directly from the signal and where discrepancies in acoustic cues arise, as well as indicating whether the perceived form is in fact
produced.

Another issue upon which investigators disagree is whether the child can perceive the difference between his form and the adult form. This is difficult to assess if the child can perceive both forms when produced by the appropriate speaker. To test this, both forms would need to be produced by the same speaker. Therefore spectrographic analysis of the child's productions would be necessary to ensure that the speaker replicates exactly the child's form.

During phonological acquisition, between the time when distinctions are taught and named do other differences exist in processing besides consistency? Processing time could be examined as well as any difference in the acoustic information that is being extracted as a basis for the phonetic decision.

As mentioned earlier, the result observed by Menyuk and Anderson (1969) could be further researched to determine whether thirteen-year-olds really do more closely observe phonetic boundaries in perception and production than either younger children or adults. This could lead to indirect evidence for a close tie between perception and production and the theory that this tie is developed in the pre-puberty stage of life.

Also mentioned earlier was the investigation into when specific language values for features such as voicing occur, and whether these boundaries are both perceived and produced at this value from the beginning, or whether production is later modified to these specific values.
SELECTED BIBLIOGRAPHY


Barton, D. The discrimination of minimally different pairs of real words by children aged 2;3 to 2;11. Paper presented at the Third Child Language Symposium, London, 1975. (b)


Cooper, W.E. Contingent feature analysis in speech perception. Perception and Psychophysics, 1974, 16, 201-204. (b)

Crowder, R.G. The sound of vowels and consonants in immediate memory. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 587-659. (a)


Menyuk, P. Direction of language processing. *Q.P.R., Research Laboratory of Electronics, M.I.T.*, 1971, **101**, 182-188.


Pisoni, D.B. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception and Psychophysics, 1973, 12, 253-260. (a)


Sachs, R.M. Vowel identification and discrimination in isolation vs. word context. Q.P.R., Research Laboratory of Electronics, M.I.T., 1969, 93, 220-229.

Saslow, M.G. Reaction time to consonant-vowel syllables in ensembles of various sizes. Q.P.R., Research Laboratory of Electronics, M.I.T., 1958, 50, 143-144.


Tikofsky, R.S., and McInish, J.R. Effects of intrastimulus delay on speech and accuracy in a forced choice speech discrimination task. Report Number 20, October 1, 1968, University of Michigan, Grant Number 1 P01 HD 01368-04, National Institute for Child Health and Human Development. (b)


