RADC-TR-89-208
Final Technical Report
October 1989

TOWARD A NATURAL SPEECH UNDERSTANDING SYSTEM

University of Colorado

Gary Bradshaw, Terry Halwes, Alan Bell

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, NY 13441-5700

This report has been reviewed by the RADC Public Affairs Division (PA) and is releasable to the National Technical Information Services (NTIS) At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-89-208 has been reviewed and is approved for publication.

APPROVED: David B. Stockton

DAVID B. STOCKTON, CAPT, USAF
Project Engineer

APPROVED:

WALTER J. SENUS
Technical Director
Directorate of Intelligence & Reconnaissance

FOR THE COMMANDER James W. Hyde III

JAMES W. HYDE III
Directorate of Plans & Programs

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (IRAA) Griffiss AFB NY 13441-5700. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS N/A
2a. SECURITY CLASSIFICATION AUTHORITY N/A	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) N/A	5. MONITORING ORGANIZATION REPORT NUMBER(S) RADC-TR-89-208
6a. NAME OF PERFORMING ORGANIZATION University of Colorado 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Rome Air Development Center (IRAA)
6c. ADDRESS (City, State, and ZIP Code) Institute of Cognitive Science Campus Box 345 Boulder CO 80309-0345	7b. ADDRESS (City, State, and ZIP Code) Griffiss AFB NY 13441-5700
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F30602-81-C-0193
8c. ADDRESS (City, State, and ZIP Code) Training Systems Division Brooks AFB TX 78235	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 62205F; PROJECT NO. 9567; TASK NO 02; WORK UNIT ACCESSION NO. P2
11. TITLE (Include Security Classification) TOWARD A NATURAL SPEECH UNDERSTANDING SYSTEM	
12. PERSONAL AUTHOR(S) Gary Bradshaw, Terry Halwes, Alan Bell	
13a. TYPE OF REPORT Final	13b. TIME COVERED FROM Jan 87 TO Oct 87
14. DATE OF REPORT (Year, Month, Day) October 1989	15. PAGE COUNT 216
16. SUPPLEMENTARY NOTATION N/A	

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	07	
23	02	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Speech
Linguistics
Man-machine Interface

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Template matching speech recognition algorithms have proven effective in isolated-word speaker-dependent applications. Improvements may permit these algorithms to be used for a wider range of applications. This report describes several experiments on a template-matching system that are prerequisites for extending the system to multi-speaker applications. One set of experiments was performed on a two-speaker database, evaluating the learning algorithm and recognition performance with these two speakers. A second group of experiments evaluated system performance as a function of the amount of template-matching that is permitted. Next, several tests were conducted to evaluate system performance on a new database of eight speakers. In addition to the recognition tests, we identified three different acoustic representations. The three representations are successively closer approximations to selective characteristics of the human ear. Finally, two new recognition vocabularies, based on important linguistic considerations, are described. Because of their linguistic generality, these vocabularies can be used to provide (Continued on Reverse)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL David B. Stockton, Capt, USAF	22b. TELEPHONE (Include Area Code) (315) 330-4024 22c. OFFICE SYMBOL RADC (IRAA)

DD Form 1473, JUN 86 *Previous editions are obsolete.*

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

Block 17 Continued:

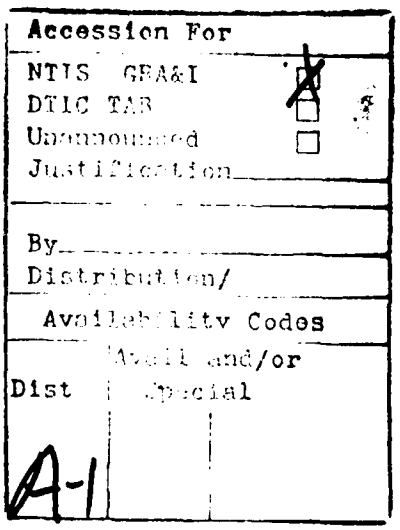
12 09

Block 19 Continued:

a systematic and representative evaluation of recognition system performance. The design criteria used to develop these vocabularies is described, so that other investigators may evaluate existing vocabularies or develop new ones to suit special applications.

Acknowledgements

The authors thank Rebecca Burns and Gary Tajchman for their assistance on the project and for their many contributions to this final report. This work was sponsored by Rome Air Development Center and by the Air Force Human Resources Development Laboratory under contract F30602-81-C-0193 from Rome Air Development Center. The support of Captain David Stockton, Lt. Colonel Hugh Burns, and Brian Dallman during the contract period is also gratefully acknowledged.



Summary of Research Progress

The research contract extended from January 1, 1987 through October 31, 1987. During this time, we completed the following research tasks:

1. Tests were performed on the NEXUS speech recognition system to investigate the potential of the system to function correctly in a multi-speaker recognition task. (see Section II)

2. Tests were performed on the NEXUS recognition system to reveal the optimal setting of matching parameters that will lead to fast recognition without sacrificing accuracy. (see Section III)

3. Multiple-speaker recognition tests were performed on a new speech database consisting of utterances from eight speakers. Performance tests evaluated the performance of a recognition system when trained on all eight speakers, or trained on a subset of four of the speakers and tested on all eight speakers. (see Section IV)

4. Differences between the current acoustic representation and that developed by the human ear were studied. Three new acoustic representations, each more closely approximating the characteristics of the human ear, were identified. (see Section VI)

5. A set of criteria to evaluate the adequacy of vocabularies for speech recognition tasks were designed. Using the criteria, two new vocabularies were developed. One vocabulary is specialized to test the ability of a recognition system to make fine phonetic distinctions, while the other vocabulary is intended as a continuous speech recognition vocabulary. (see Section VII)

6. A survey of relevant articles was performed, and an annotated bibliography of papers on speech perception and speech recognition was compiled (Appendix D)

The following project had to be abandoned, due to unanticipated technical problems:

1. A speech workbench, combining the ability to graphically manipulate, synthesize, analyze, and display speech, was proposed. Limitations on disk size and system software errors prevented meaningful work on this project. (see Section V)

Table of Contents

List of Tables

List of Figures

I. Introduction

Template-matching speech recognition architectures have enjoyed widespread popularity since their introduction by Itakura [1]. Major reasons for this popularity include their language independence, low error rates for distinctive words produced in isolation by a single speaker, and their simple programming requirements. Template-matching systems rank among the best available recognition architectures, but they are far from perfect. Unsolved problems include 1) recognition of connected speech; 2) recognition of speech produced by a wide range of talkers; and 3) recognition of distorted words uttered by speakers under conditions of mental or physical stress.

Exploiting the unique sequences of phonemes that appear in many vocabularies, template-matching systems avoid the difficult problem of making fine phonetic distinctions. As long as all words are acoustically distinct, this strategy works very well. When similar words are included in the vocabulary, recognition performance can drop drastically (Bradshaw, Cole, and Li, [2]). Assume that a particular vocabulary included the words "all", "ball", "call", "fall", "gall", "hall", "mall", "Paul", "stall", "tall", and "wall". Each of these words begins with a consonant and ends with the phonetic sequence /ɔl/. Current recognition systems would be likely to confuse many of these words with one another. Errors can be traced to a simple fact about human speech: word articulation is never precise. Articulatory variation will first be manifest in the reference patterns; the vowel-consonant (VC) cluster /ɔl/ will vary slightly among the eleven stored words. Because the VC cluster is much longer than the initial consonant, minor variations in cluster quality can outweigh important differences that occur at the beginning of these words. Failing to recognize the sharing of phonetic segments between words, the matcher is misled by spurious differences and makes recognition errors.

Limited recognition performance is only one problem created when systems ignore the phonetic structure of words in a language. Storing and matching processes are also adversely affected. In principle, only a small number of phonetic patterns (between 40 and 100) are sufficient to match all of the sounds produced in a language. Whole-word systems store unique acoustic

patterns for each word, leading to redundancy in storage and corresponding inefficiencies in matching as the vocabulary size increases. Even in our simple "ball" vocabulary shown above, the matcher must compare the same sequence, /ɔl/, eleven times against an input. This situation is manageable as long as the vocabulary size remains small, approximately a thousand words or less. Although some practical applications can be developed with limited vocabularies, most users will quickly find that larger vocabularies are desirable; some applications may require vocabulary sizes in the ten- or hundred-thousand word range. The inefficiencies of whole-word systems become so overwhelming that the approach must be abandoned. Large vocabulary systems must exploit the phonetic structure of a language in order to remain affordable and responsive.

Human listeners recognize that words are created from a common set of phonetic segments, and focus their attention on the distinctive parts of utterances. Successful speech recognition depends on finding mechanisms that permit systems to recognize the similarity of words and to selectively attend to distinctive sections. A set of discovery mechanisms was included in the NEXUS speech recognition architecture (Bradshaw, [3], attached as Appendix A) for just this purpose. Although NEXUS is a template-matching architecture, with all of the advantages mentioned in the first paragraph, it is not a whole-word matching system. A broad segmenter in NEXUS breaks words down into sub-word units, corresponding roughly to the phonetic segments of a language. Learning heuristics in NEXUS can recognize that two words, such as "ball" and "Paul", share identical phonetic segments. When common sequences of phonetic segments are identified, NEXUS eliminates the redundant information, and constructs word models that reflect the sharing of phonetic segments between different words.

This capability is not sufficient to permit perfect recognition performance. Variability of phonetic segments can be so large that no single template adequately represents the phonetic class. Through an evaluation of recognition errors, NEXUS recognizes that a particular acoustic segment has variant forms. Each variant is encoded into an **equivalence set** of patterns that have the same phonetic identity. NEXUS does not bother to encode every

minor variant of a phonetic segment, but only stores distinctive variants necessary to represent segment variability necessary for correct word recognition.

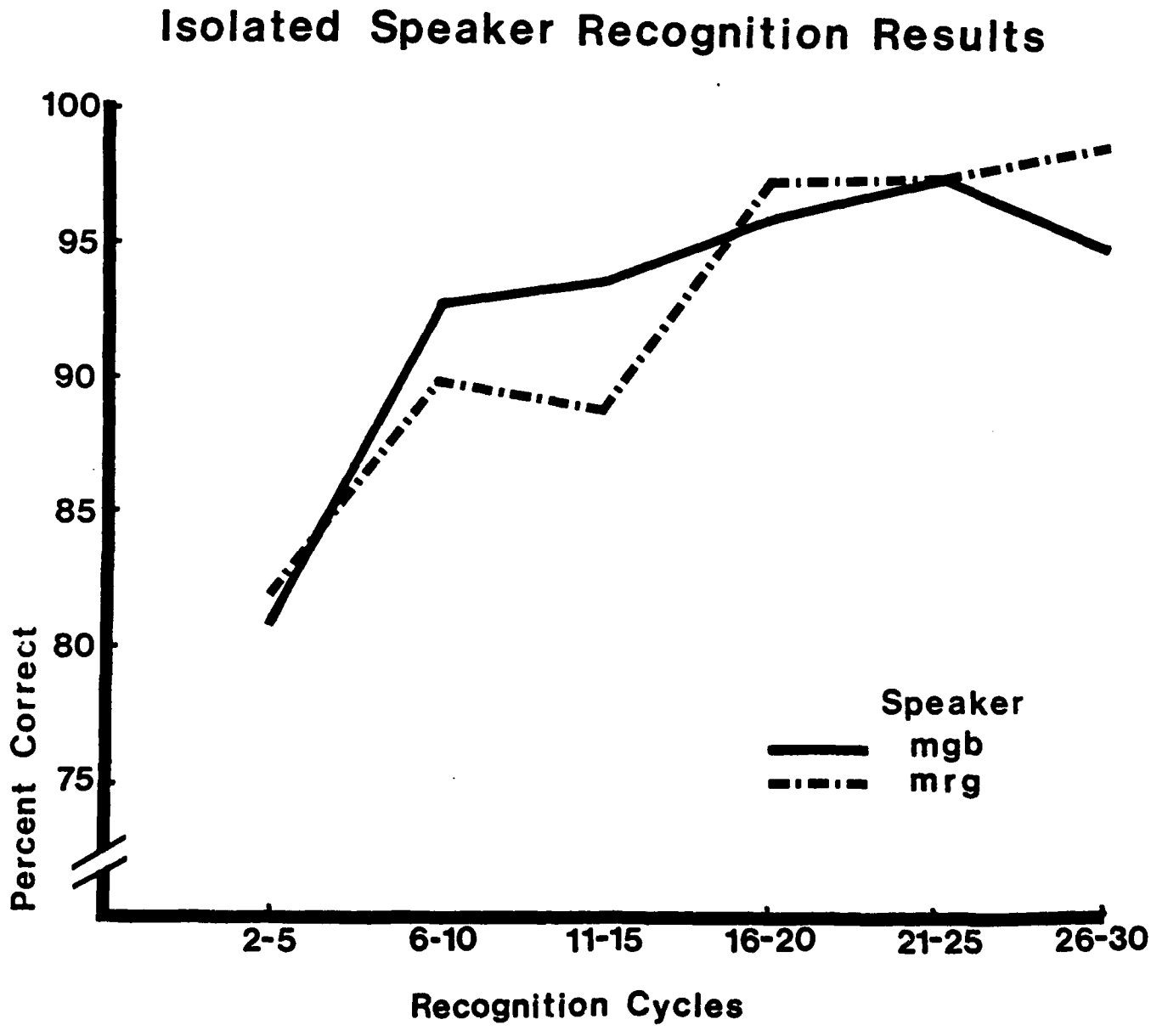
Learning heuristics in NEXUS, based on machine learning principles, perform these operations. The heuristics analyze word similarity and identify the factors responsible for a recognition error. On the basis of this information, NEXUS modifies its representations of words and phonetic segments. The learning heuristics operate continuously, so that NEXUS does not need to make an arbitrary distinction between training and performance modes. As a result, NEXUS has a great deal more flexibility than systems that rigidly enforce this distinction. NEXUS can add new words into its vocabulary at any time; or modify its inventory of phonetic segments to reflect new variability that is leading to recognition errors.

Bradshaw [3] reports comparisons between NEXUS and CICADA, a state-of-the-art template-matching system, on a common database. Comparative tests, where two or more systems classify the same utterances, are invaluable in identifying the relative strengths of different approaches. To make the test challenging, the "alphabet" vocabulary was used. This vocabulary is difficult because similar words are common. The CICADA system correctly identified 80% of the 1560 utterances. At the time of the original tests, NEXUS correctly recognized 89.5% of the same utterances. Since those results were published, two errors in the NEXUS code were identified and corrected, and the tests re-done, improving NEXUS' overall performance to 93%.

Figure 1 shows the current recognition performance for NEXUS. The database used for these experiments contains thirty repetitions of each letter of the alphabet produced by two different speakers, for a total of sixty repetitions. Single-speaker experiments are organized into recognition cycles. During each cycle, the twenty-six letters of the alphabet are presented in sequence to NEXUS. A full single-speaker experiment consists of thirty recognition cycles. Recognition cycles for each speaker are averaged into five-cycle blocks in Figure 1. The first point, labeled 2-5, does not include cycle 1 performance, and so represents an average of only four

cycles. On the first recognition cycle, NEXUS is trying to classify words that it has never heard before. Not surprisingly, every utterance on cycle 1 is misclassified.

Figure 1



It is evident from the improvement that occurs over recognition cycles in Figure 1 that the learning heuristics are quite effective. Performance during cycles 2-5 averages only 81%, while performance on the final block,

26-30, is up to 97%. Because of this learning, reporting the average performance of NEXUS on cycles 2-30 as 93% is a misleading indicator of the true power of the system. NEXUS must have experience with words to permit the learning heuristics to operate. The 93% figure averages together cycles where NEXUS is still acquiring this experience with cycles where performance has greatly improved because of successful learning. Most of the improvement occurs during the first fifteen recognition cycles. A better indicator of asymptotic performance is to determine recognition rates during the last half of the experiments. During this period, NEXUS' performance is up to 97.3%, representing an error rate only one-seventh as large as the CICADA system.

These results demonstrate that some of the limitations of template-matching technology are not inherent, but are merely restrictions imposed by current implementations. Extending template-matching systems to natural speech input from a wide variety of talkers will require solutions to two stubborn problems: connected speech and multi-speaker recognition. The results we have described do not address these issues, but do offer hope that these problems may be solved. The ICS Speech Laboratory at the University of Colorado has an ongoing effort to build a natural speech recognition system. Given the advantages of template-matching technology, our strategy is to explore extensions to this technology that may permit recognition of connected speech produced by a wide range of talkers. In the long run, it may be impossible or impractical to completely solve these problems from the template-matching framework. Only by pushing the technology to the limits will these boundaries be identified.

This report describes some efforts we have made to extend template-matching technology to a multi-speaker situation. Some of our efforts are intended to improve performance of the existing system, while other experiments actually begin to address the problem of recognizing speech produced by multiple talkers. All of these results are still in the exploratory phase. The report concludes with a discussion of projected changes necessary to approach the difficult problem of connected speech.

II. Two-speaker tests of NEXUS

Template-matching speech recognition systems operate by comparing acoustic patterns of speech. The representation of acoustic information can take many forms, but every representation contains information about three dimensions of sound: the **intensity** of sound that occurs at a number of **spectral frequencies** over a period of **time.** By their very nature, such representations are sensitive to the characteristics of a particular speaker's voice. Spectral contours are strongly affected by the shape of the speaker's vocal tract, while temporal contours are affected by the speaker's dialect. If different speakers are asked to utter the same word, the resulting acoustic patterns will be different. As a result, errors become commonplace when reference templates developed from one speaker are used to recognize words produced by a different speaker.

Because of these differences, template-matching systems have often been restricted to single-speaker applications. Although this limitation may not be important in some applications of voice technology, most applications will require the system to recognize speech produced by several different talkers. Given this requirement, it is helpful to identify extensions that will permit a template-matching system to be used in a multi-speaker situation.

A trivial multi-speaker adaptation for template-matching systems is to store copies of each word template produced by every speaker. Recognition proceeds by comparing the unknown input against all of the stored templates in the multi-speaker database. Few investigators have explored this alternative because of the unacceptable computational demands placed on the recognition system. The size of the reference template set would increase linearly with the number of speakers, and the time required by the matching operation is likely to approach a near-linear function of the number of speakers (the use of pruning algorithms to terminate poor matches early would mitigate the increase in matching time to a certain extent). Data storage and template matching are the most expensive hardware and software components of recognition systems. Given these constraints, it may be possible to adapt a single-speaker recognition system to a small number of users, but increases of a hundred- or thousand-fold would not be practical. A second restriction

on this approach is that each speaker must train the recognition system on his or her voice. This may be practical as long as the recognition vocabulary is a few hundred words or less, but will become impractical when the recognition vocabulary reaches a few thousand words.

If similarities between different speakers can be recognized and exploited, redundant templates could be eliminated. Although there may be many different types of speakers, the number of different types must be less than the number of people who speak a language. Rabiner et al. [4] proposed the use of statistical clustering techniques on word templates to identify the reference template set. These techniques were applied to a database of words produced by 100 different speakers. The clustering process requires input data on the similarity of words. The process begins by matching each instance of a particular word with all other instances of the same word, to produce an index of **within-word** similarity. Also, each instance of every word must be compared with all instances of other words, producing an index of **between-word** similarity. Clusters are identified where the within-word distances were small compared with the between-word distances. Rabiner et al. found that some words exhibited more variability than others, so a maximum of twelve clusters was identified for each word in the vocabulary. A representative token from each cluster was used to create the reference set. Through this process, Rabiner and his colleagues were able to limit the number of reference templates for each word to no more than 12, with commensurate savings in storage and matching.

Several difficulties have prevented the widespread utilization of clustering techniques. First, the clustering operation is resource intensive. A large database must be collected, then every possible comparison made between word tokens before clusters can be identified. Second, the process is not incremental. The clustering procedure must be repeated each time a new speaker or a new word is added into the recognition set. Both of these factors preclude a dynamic, adaptive clustering procedure from being developed, restricting the technique to situations where the group of speakers and the target vocabulary do not often change. Finally, the procedure is not completely automatic, but requires human intervention to

determine the number of clusters that optimally fit the input data set. At each site, system operators would have to be trained on clustering procedures to assist the recognition system in the development of the reference set.

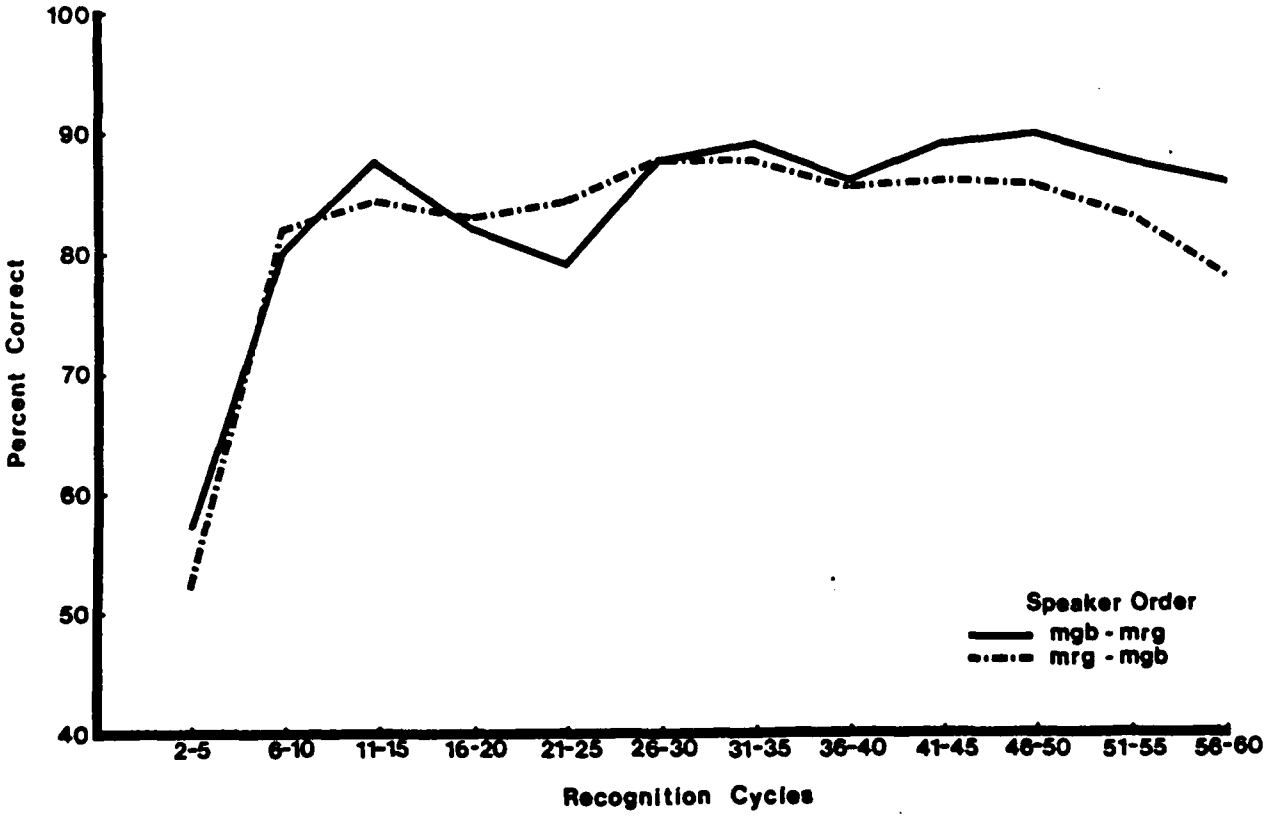
The machine learning techniques included in NEXUS do not suffer from these limitations, and appear to be a viable alternative to traditional statistical procedures. Variability in utterances can be accommodated in NEXUS by encoding a set of acoustic patterns to represent each phonetic segment. When NEXUS is trained on a single speaker, many distinctive phonetic segments are represented by only a single node; the phonetic equivalence set contains only one template. A few segments have greater variability, so NEXUS includes two or more patterns in the equivalence set. NEXUS adds instances to equivalence sets only when the current set is unrepresentative of phonetic segment class variation. This strategy can be applied to the multi-speaker situation as well. If the current set of patterns can be applied to a new speaker, NEXUS will do so without extending any equivalence sets. If the new speaker produces phonetic segments in a manner distinct from all previous speakers, leading to recognition errors, NEXUS will encode new variants of the segments, adding them to the appropriate equivalence sets.

In effect, NEXUS has always been capable of acting as a multi-speaker recognition system. No modifications were necessary to test NEXUS on multiple speakers. To be an effective multi-speaker system, NEXUS must generalize patterns between speakers. If no generalization occurs, NEXUS effectively degenerates to the trivial solution of storing independent templates for all speakers. The best index of generalization is to evaluate the total number of phonetic patterns that are encoded in the recognition network. If the number of patterns in multi-speaker mode equals the sum of the patterns in single-speaker mode, no between-speaker generalization has occurred. Perfect generalization would occur if the number of phonetic segments was equal to the largest number that occurred in individual speaker mode. Naturally, it would be unrealistic to expect NEXUS to generalize to that extent.

Several experiments were performed to evaluate NEXUS as a multi-speaker system. These experiments were limited by the fact that the current database only contains utterances produced by two speakers, both of whom are male. The entire database consists of thirty repetitions of each letter produced by each speaker, for a total of 1,560 utterances. Multi-speaker tests were performed by having NEXUS classify one complete cycle of 26 tokens produced by the first speaker, then switch to the other speaker for one cycle, returning to the first speaker to classify the second cycle, and so on until all thirty sets of letters were classified for both speakers. Figure 2 presents recognition performance, summarized in blocks of five recognition

Figure 2

Multi-Speaker Recognition Results



cycles. Compared against the average of the corresponding isolated-speaker tests, the overall recognition rate dropped slightly, from 93% to 84%.

During the first recognition cycle, NEXUS does not have sufficient experience with words for the learning heuristics to operate. NEXUS resorts to memorizing each input as a new utterance. As a result, the first cycle has a major impact on the development of the recognition network. It is important to establish that NEXUS is relatively insensitive to the order that speakers are experienced, because in practice there can be no guarantee that training examples occur in the order that maximizes long-term performance. Figure 2 shows multi-speaker experiments in both orders. The solid line indicates the experiment where speaker mgb came first; the dashed line indicates performance where speaker mrg was first. As can be seen from the graph, there is little difference in recognition performance as a function of speaker order.

The overall pattern is remarkably similar to isolated-speaker results shown in Figure 1. Most of the change occurs in cycles 1 to 10, with gradual improvement throughout the remaining 55 cycles. The mrg-mgb speaker order experiences a mild downturn in the last fifteen cycles. This decline does not appear to be significant.

The results plotted in Figure 2 are smoothed in averaging across blocks of recognition cycles. Inspection of individual recognition cycles reveals variability in recognition performance, particularly in the final 30 cycles. Performance can drop as much as 23% from 96% on one trial to 73% on the next, then rebound to 92% on the third cycle in the sequence. These pronounced shifts in the number of errors are not observed in isolated-speaker performance. One possible explanation of this problem derives from the pattern of interleaving utterances from the two speakers. NEXUS classifies one cycle from speaker A, then one from speaker B, then returns to A, and so on. Observed performance dips never persist for more than one cycle. The dip might derive from interleaving a difficult cycle from one speaker between a pair of easy cycles for the other speaker. To evaluate this hypothesis, the error pattern in speaker-dependent mode was compared against the

performance for corresponding data sets in the two-speaker tests. This comparison did not show the expected correlation in cycle performance, eliminating this simple first explanation. A more careful analysis is currently underway, along with some additional experiments, to help isolate and remove the source of this variability.

III. Experiments on matching parameters

A part of the funded research effort was to evaluate certain parameters of NEXUS, to help improve system performance, and prepare for development of a connected-speech version of NEXUS. Specifically, we manipulated two parameters of the matching operation to evaluate their impact on recognition performance and learning processes.

Matching in NEXUS occurs by comparing all of the reference templates in parallel against successive frames of the input. The Euclidian distance between the short-term spectra of reference patterns and the input are computed. Following a dynamic programming algorithm, the current match is fed into a matrix where minimum-warping-path distances are accumulated. If the reference pattern is very different from the input, such as comparing the steady-state vowel /e/ to the fricative /s/ in "c", the cumulative warping distance will be very large. Since the matching operation is a computationally-intensive procedure that requires more than 80% of all CPU time in NEXUS, recognition time may be reduced eliminating such unnecessary comparisons. Lowerre (Lowerre, [5]; Lowerre and Reddy, [6]) developed the concept of a **beam search** to terminate implausible candidates early. The best match at any particular frame can be identified by selecting the minimum cumulative distance in the warping matrix for the frame. A **pruning threshold** is computed by adding a constant to the minimum distance. All current matches that exceed the current pruning threshold represent distances significantly larger than the best currently achieved, and are eliminated from further consideration. Matches with scores below the pruning threshold are retained for further consideration.

Two important parameters are associated with the beam searching process. The first parameter is the pruning threshold, which can be varied by changing the constant added to the best score. As the constant increases, the threshold is raised, permitting additional patterns to remain in the candidate set. Reducing the constant has the opposite effect, permitting fewer matches to complete. Previous experience had shown that the time to complete an entire recognition experiment was a linear function of the

pruning threshold, which suggests that a minimization of the threshold consistent with accurate performance should be selected.

The second parameter is derived from the frame lag used to obtain the minimum distance score. Matching in NEXUS is based on processing frames of the input in sequence. Appropriate reference frames are matched against the first frame of the input, then the second frame of the input, and so on, until all input frames have been matched. Consider how pruning might occur in the first input frame. The first frame of the first reference might be quite different from the first input frame, and receive a poor matching score of 1000. Without any experience on this input, it is not clear whether this reference pattern should be eliminated from experience. The first frame of the second reference pattern might be a better match, with a score of 200. However, we will not know what the best match is until all matches for the current frame have been computed. At this point, the threshold can be calculated. The system must go back over all matches, eliminating all of those that exceed the threshold.

This threshold computation method requires the system to pass over all matches twice. The first time to compute and update each score, the second time to check each match to see if it should be eliminated. The inefficiency is not great, but since it occurs during the most-frequently-repeated process in the system, the overall penalty is high. An alternative is to compute the threshold by adding the pruning constant to the best score from the previous frame. Every score on frame two can be compared against the pruning threshold obtained from frame one, while frame three matches can be compared against the threshold from frame two, and so on. With this simple modification, we no longer need to pass over each matching score twice. Instead, we can compute the value of the score and immediately compare it against the threshold to determine whether it should be continued or eliminated. We refer to this as the **frame lag** pruning algorithm, because the threshold for frame N is computed from frame N-1.

One difficulty in using a frame lag threshold is that no previous score exists for the first frame. In this case, the assumption is made that the

previous matching score was zero, so the pruning threshold for the first frame is identical to the pruning constant (0 + pruning constant = pruning constant).

A second factor that must be considered is that the frame lag and the pruning constant are not independent. The first method we considered, which we will refer to as the **lag 0** method, computes the best match on the current frame, then adds the pruning constant to that value to generate the pruning threshold. If the pruning constant were zero, only the best match in each frame would be retained. Instead, the pruning constant is usually set to a larger value to permit close matches to be retained, while more distant matches are eliminated. The pruning constant reflects a "slop factor" needed to permit close matches to continue. The situation differs at a frame lag of one, because the best match at frame N-1 will be smaller than the best match at frame N. If the pruning constant were set to zero, all matches would be terminated. The pruning constant must be increased to include an expected-distance for a single frame along with the proximity factor needed to permit close matches to continue.

Of course, using the distance obtained in the previous frame is not the only possibility. We might choose a lag of two frames, so that the threshold for frame three was based on frame one, the threshold for frame four was based on frame two, and so on. As the frame lag increases, the pruning constant must also be increased to maintain a constant amount of pruning. This results from the cumulative nature of the distance metric. If we are computing a score for frame N of the input, and our lag is $\underline{m}$ frames, the pruning threshold is based on frame N-$\underline{m}$. The score at frame N represents the sum of the distances from frame 1 to frame N-$\underline{m}$ and from frame N-$\underline{m}$+1 to N. As $\underline{m}$ gets larger, the cumulative distance over $\underline{m}$-1 frames will increase for all patterns. Thus, the pruning constant must be increased to reflect the average expected distance in $\underline{m}$ frames.

Varying the frame lag has a subtle but important consequence for the pruning process. Consider what would happen if a short section of the input was quite different from the corresponding sections of all reference

patterns. A pruning threshold obtained at the last frame of the normal section N - m would have a constant added to it, representing the average cumulative distance that would be expected in the average lag period. At time N, every cumulative score may exceed the threshold, so all matching would be terminated. This situation does not arise when the frame lag is zero: the pruning threshold is always computed from the best match at the current frame, so that matching is guaranteed to continue.

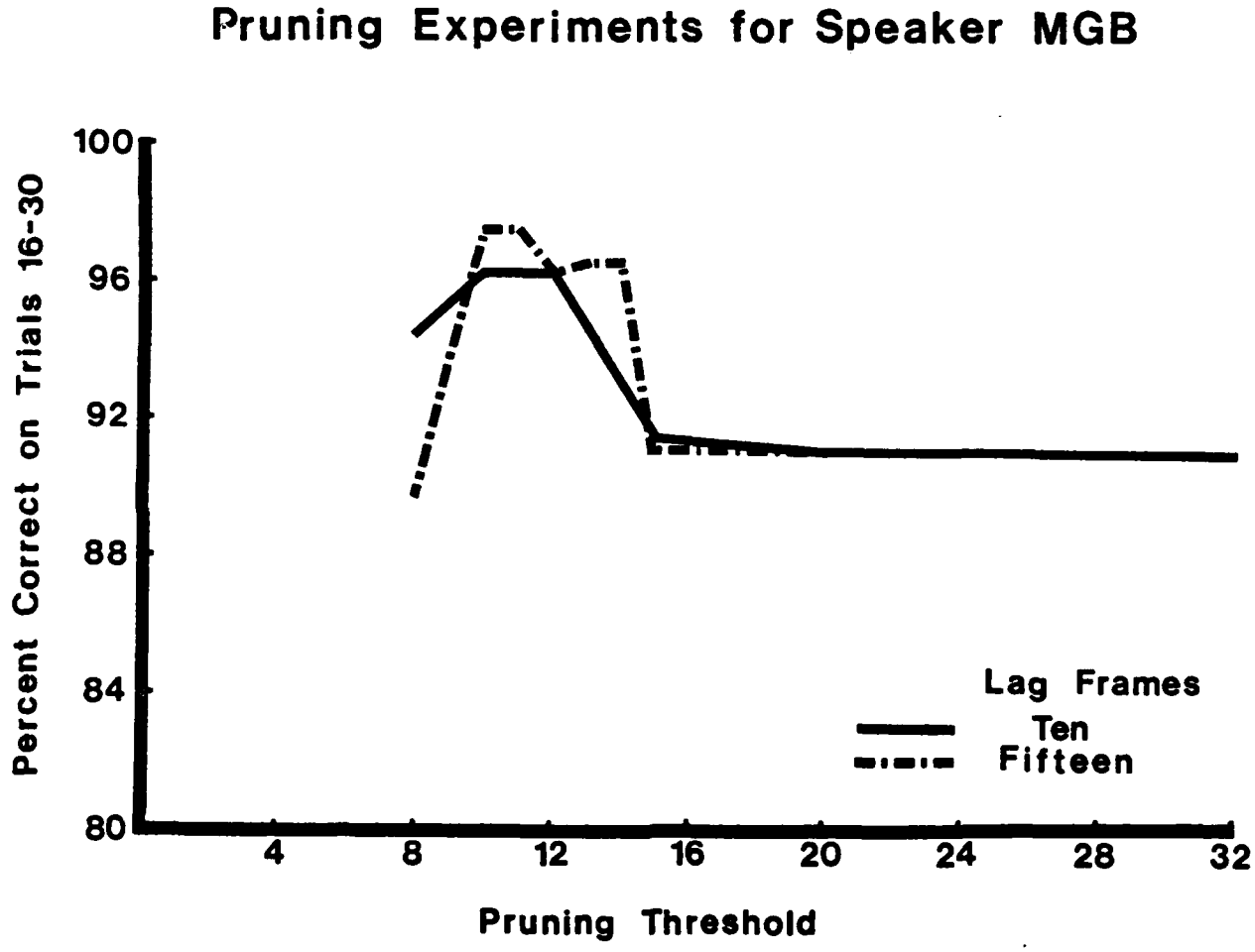
In effect, using the frame lag method to compute the pruning threshold balances both actual matches and standardized matches against one another. In the lag 0 method, the pruning constant consisted only of the proximity factor. The lag 1 method requires the pruning constant to reflect both the proximity factor and the average distance for a single frame. As the lag increases, the cumulative average distance will increase, but the proximity factor will remain constant. The proximity factor will accordingly have a proportionally smaller weight in the size of the pruning constant. The cumulative average distance is a "standardized" amount, independent of the actual score at a particular point. If the match between a reference and the input is poor during the lag region, the cumulative distance will exceed the threshold, terminating the match. As we saw in our example earlier, lag matches can be terminated even if they are the best match at a particular frame.

There appears to be no way to determine the optimal frame lag and value of the pruning constant on theoretical grounds. Instead, several experiments on NEXUS were performed to identify optimal settings of these parameters. Figure 3 summarizes our experiments for speaker MGB, while Figure 4 summarizes similar experiments for speaker MRG.

Both of these figures present average performance on recognition cycles 16-30 as a function of the frame lag and pruning threshold. One curious result can be seen in Figure 3. As the pruning threshold is increased, NEXUS makes more errors. This effect is probably due to a peculiarity of the alphabet. Most minimal pairs in the alphabet are CV syllables such as "B", "D", "P", and "T", where the initial constant serves to discriminate between

items. Large differences occur at the beginning of these words, so a tight pruning threshold eliminates related words. A looser pruning threshold will preserve these matches, then minor differences in the final vowel quality will outweigh the larger initial differences, causing a recognition error.

Figure 3

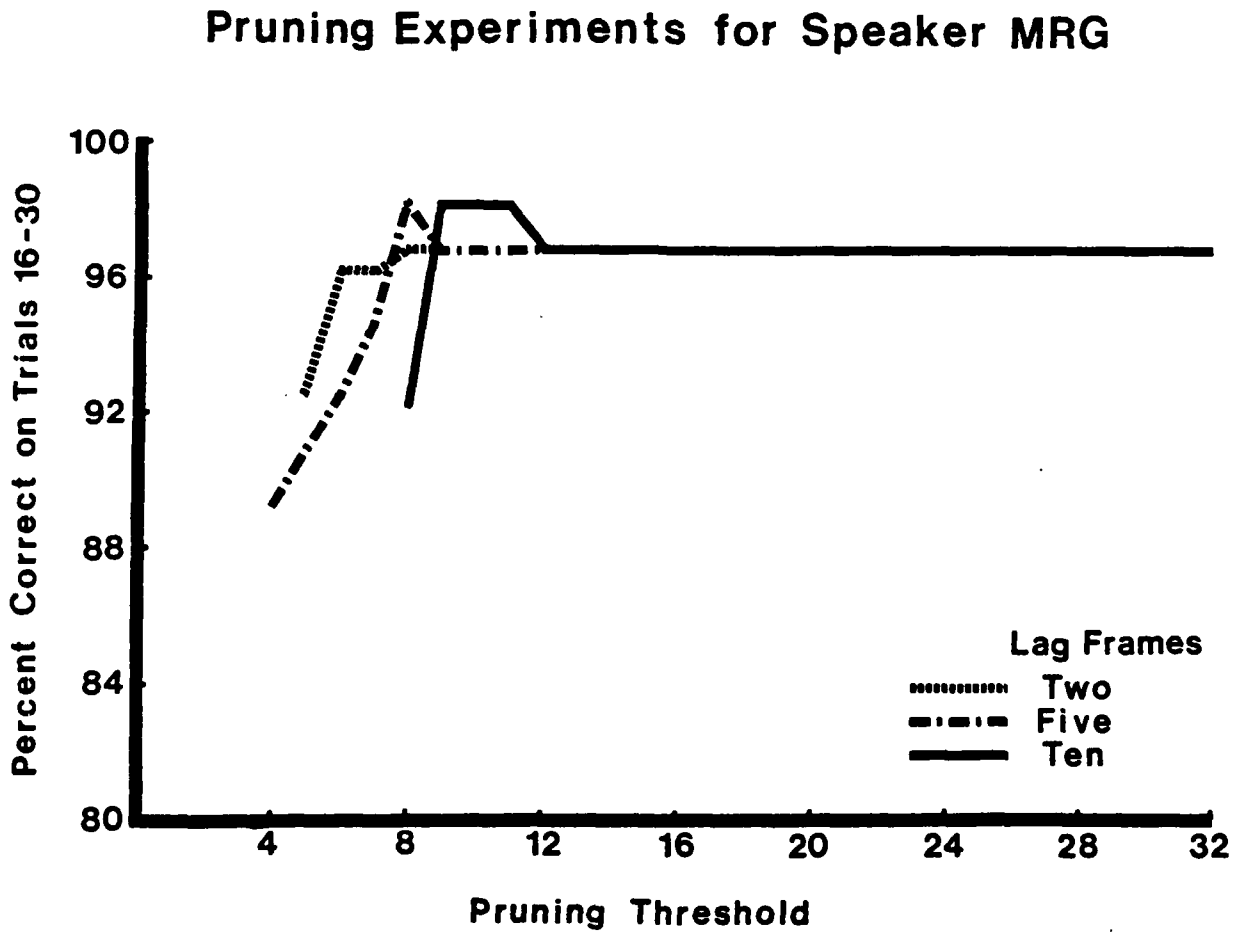


These results should not be taken to support a low threshold value, since the effect appears to be vocabulary-dependent. A different vocabulary is unlikely to preserve the initial-segment-difference bias, whereupon a low threshold would prove inferior to a higher value.

Based on these experiments, we have standardized the frame lag parameter to be 10 frames and the pruning constant to be 20. Both selections represent conservative settings that permit all reasonable matches to be pursued, yet

minimize total experiment time under this constraint. Lower values may increase apparent performance, but these results would be unlikely to generalize to new vocabularies. A second reason for choosing conservative values is that the pruning mechanism was never intended as a mechanism to enable NEXUS to make fine phonetic distinctions. Underlying the time-normalization matching strategy is the philosophy that very local information

Figure 4



may be unreliable; only the global match is important. If the pruning threshold is set too tightly, reasonable candidates will be eliminated from consideration on the basis of one or two spurious frame mismatches. Since pruning is an irrevocable decision, it should only occur when the system has firm reason to eliminate the candidate template. Pruning is best thought of as a way to speed matching by eliminating broad phonetic mismatches. The goal of NEXUS' learning mechanisms is to make fine phonetic distinctions.

These mechanisms require information about word similarity in order to make refinements in the network. If similar matches are terminated by the pruning mechanism, NEXUS will not generate the information needed by the learning mechanisms. The pattern of errors as a function of pruning threshold suggests that the current version of NEXUS may not be exploiting all available information on word differences, and that improvements to the learning mechanisms are still possible.

IV. Multi-speaker tests of NEXUS

A new database was collected to explore the potential of NEXUS to simultaneously recognize utterances produced by a broader range of speakers. Eight speakers, four males and four females, each produced 30 tokens of the augmented alphabet vocabulary. All tokens were collected on our laboratory MicroVax computers, using a new a/d system that differs slightly from the system used to collect the previous two-speaker database. As in our previous system, the a/d converter quantizes each sample to 16-bit accuracy and uses a 16 KHz input rate. The major change is in the analog hardware. The current system has a much higher ambient noise level than the previous system, reducing the effective signal-to-noise ratio. During the recording sessions, each token was immediately checked for clipping and for adequate volume. If an utterance was found to be outside of the acceptable amplitude range, the user was asked to repeat the item.

The augmented alphabet contains the words and pseudowords "INK", "-ING", "HIGH", "SHY", "GAY", "-ETH", and "THEE" in addition to 25 letters from the alphabet[1] . The items added to the augmented alphabet extend its coverage of English phonemes; the vocabulary thus represents a more thorough test of recognition systems than does the standard alphabet task. By including a broader range of phonetic contrasts, researchers can ensure that their recognition systems will be readily extensible to other vocabularies. The augmented alphabet vocabulary is also more challenging for recognition systems, as several new minimal pairs are introduced. Table 1 shows the major subsets of the augmented alphabet vocabulary.

Although 30 tokens of each word were recorded, we restricted all testing to the first eight tokens produced by each speaker. Testing all tokens from every speaker in multi-speaker mode would require several months of computer

---

1. "W", the only polysyllabic word in the alphabet, was omitted from the current database. It is seldom confused with the other monosyllabic items. Inclusion of this item only serves to artificially inflate recognition performance.

time per experiment. The selection procedure reduced the database to more manageable proportions. No further use was made of the remaining 22 tokens

Table 1: Subsets of the Augmented Alphabet Vocabulary

Set	Items
E	B C D E G P T V Z THEE
EH	F -ETH S X L M N
A	A H J GAY K
I	I HIGH SHY Y
IH	INK -ING
U	U Q
Other	O R

of each item. The new multiple-speaker database contains 8 tokens of 32 words produced by each of 8 speakers, for a total of 2048 unique utterances. Each token in the database was transformed from the time domain into the frequency domain using the normal begin/end detection and signal processing routines.

Isolated-Speaker Baseline Tests

Each speaker was tested independently (in single-speaker mode) to establish baseline performance levels. Table 2 summarizes the results for the eight speakers. The percentage of items correctly recognized in each cycle was determined, and average performance figures for cycles 2-4 and 5-8 was computed. Table 2 also shows the average performance across the entire test from cycles 2-8. (Cycle 1 performance is always zero, and so is omitted from all analyses.) To permit comparisons with later tests, performance on cycle 2 as well as the average performance on cycles 3 and 4 is shown separately.

As can be seen in Table 2, performance on the augmented alphabet is lower than asymptotic performance on the previous database. This reduction in performance probably derives from at least three differences between the databases: 1) the augmented alphabet is a more challenging vocabulary than the standard alphabet; 2) the background noise level was higher using the

current a/d system; and 3) the number of training sets was reduced from 30 to 8 for each speaker.

Table 2: Isolated Speaker Performance for Augmented Alphabet Database

	Cycles		Average	Cycles	
Females	2-4	5-8	2-8	2	3-4
fcb	75%	77%	76%	78%	73%
flh	66%	77%	72%	50%	73%
frr	89%	91%	90%	84%	91%
ftv	77%	84%	81%	65%	83%
Average	77%	82%	80%	70%	80%
Males					
mel	76%	93%	86%	72%	78%
mjm	84%	88%	87%	84%	84%
mmp	81%	86%	84%	78%	83%
msw	84%	92%	89%	78%	88%
Average	82%	90%	86%	78%	83%
All Speakers					
Average	79%	86%	83%	74%	82%

Further inspection of Table 2 suggests another factor that may be influencing overall performance. NEXUS is sensitive to the sex of a particular speaker. The average performance for male speakers, 86%, is 6% better than performance for female speakers. Differences of this sort are commonly observed in most speech recognition systems. We have traced some of the difficulty to segmentation errors, which are more common in female speakers than in male speakers. A hand segmentation of our database, currently in progress, can be substituted for automatic segmentation to establish the relation between mis-segmentation and recognition errors.

With only eight recognition cycles per speaker in the tests, NEXUS has much less experience to develop a speech network than when tested for 30 cycles per speaker using the previous database. Nevertheless, the learning trend shown in Table 2 broadly resembles our previous results. Here we find a 7% improvement from the first block, cycles 2-4, to the final block, cycles 5-8.

Having established the similarity in performance between the old and new databases in single speaker mode, we next performed more extensive multiple-

speaker tests of NEXUS, exploiting the larger number of speakers and the two sexes present in our new database. The next sections will describe two series of multi-speaker tests. The initial series explored the ability of NEXUS to recognize words produced by a group of speakers all of the same sex, while the second series explored the ability of NEXUS to recognize words produced by both males and females.

Single-sex tests of NEXUS

Our first experiments with multiple speakers restricted training to either the male speakers or the female speakers. Training again followed the interleaved speaker pattern described in Section II. One set of utterances from the first speaker was given to the system, followed by one set of utterances from the second speaker, the third speaker, and the fourth speaker. After one instance of every word from each speaker had been given to NEXUS, the second set from the first through fourth speakers was given, and so on. The large number of speakers precluded testing the full database of utterances, so only the first four sets of words were tested. Table 3 shows NEXUS' recognition performance (in percent of tokens correctly recognized) for each recognition set. As performance is strongly affected by

Table 3: Single-sex Multiple-speaker Performance for the Augmented Alphabet Database

	Recognition Sets			Average
	1	2	3-4	2-4
Females				
fcb	0%	78%	69%	72%
flh	19%	47%	59%	55%
frr	34%	91%	89%	90%
ftv	59%	75%	77%	76%
Average	38%	72%	73%	73%
Males				
mel	0%	72%	81%	78%
mjm	0%	81%	86%	84%
mmp	21%	72%	81%	78%
msw	12%	81%	84%	83%
Average	11%	77%	83%	81%
All Speakers				
	24%	75%	78%	77%

the speaker, Table 3 presents recognition rates separately for each speaker, as well as group averages.

These two tests provide information about three important characteristics of NEXUS: 1) the ability of NEXUS to generalize to novel speakers; 2) learning trends in the data; and 3) performance changes when moving from single-speaker to multiple-speaker tests. A major factor that cuts across these characteristics is differences in performance between male and female speakers. Differences in recognition performance observed in single-speaker tests are also evident in the single-sex runs. Average performance on sets 2-4 for the female speakers was 73%, while the recognition rate for male speakers was 81%. Our description of results will separately consider the two groups, where other important differences will become apparent.

The single-sex tests of NEXUS permit a limited evaluation of the ability of NEXUS to generalize to new speakers. In single speaker runs, NEXUS misclassifies all items on the first cycle, which are novel utterances. In multiple-speaker runs, NEXUS encounters a new speaker on the second recognition cycle. All vocabulary items have been encoded into the speech network from the first speaker, so NEXUS attempts to recognize items from the second speaker using a network created with the first speaker. Recognition performance on these sets (flh and mel) reflects the ability of NEXUS to generalize from one speaker to another. This is not a pure test of generalization, as NEXUS continually refines the network based on early errors in the second speaker's recognition set. In the same way, performance on the third speakers (frr and mjm) represent a rough estimate of the ability of NEXUS to generalize to a new speaker given experience with two previous speakers, and so on. Thus, recognition performance shown in the first column of Table 3 can be taken as a ballpark estimate of generalization. A more precise test of generalization following extensive training will be reported later.

The generalization data reveal another major difference between males and females: Significant generalization occurs for female speakers, while there is little evidence of generalization for male speakers. Apparently the

acoustic differences between male speakers were sufficiently large that NEXUS was unable to profit from limited experience with previous speakers. In all likelihood, our limited database does not reflect the full range of variation of speakers with different dialects, heights, weights, and so on. Under these circumstances, we should be conservative in our interpretation of the results. The minimal generalization observed between male speakers should be accepted as the best estimate, while the higher level of generalization observed between female speakers discounted as a fortunate accident. Yet it should also be recognized that these results reflect only a limited experience to the initial speakers. NEXUS may be able to generalize more accurately to novel speakers after greater training with the initial set of speakers.

Evaluating the learning trends in the data, we again observe important differences between males and females. Comparing performance on recognition set 2 with the average of recognition sets 3 and 4, little change is evident for the female speakers (72% to 73%), while male speakers improved by 6% (77% to 83%). Only two female speakers showed a net improvement between recognition set two and the final block (sets 3 and 4), while all four males improved in this period. This result is somewhat surprising as both groups showed an improvement over the same period in single-speaker tests (average improvement for females: 70% to 80%; average improvement for males: 74% to 82%. See Table 2.) No clear interpretation can be given to this difference without additional experiments.

Comparing the performance of single-speaker and multiple-speaker tests, we find a surprising result. Traditionally, template-based recognition systems make more errors when they must classify utterances produced by a group of speakers than in single-speaker mode. This phenomena was observed in our two-speaker tests reported in Section II. The current single-sex experiments do not show this result. Indeed, the performance on male speakers actually improves in the multiple-speaker case. Average isolated-speaker performance for males on cycles 2-4, 79% (Table 2), is 2% lower than recognition performance on the same data sets in multi-speaker mode. For the male speakers, NEXUS is apparently able to use information from the group to

create templates that are more effective than templates based on a single speaker. Isolated-speaker performance for females on cycles 2-4, 77% (Table 2), is slightly higher than performance on the same recognition sets in multiple-speaker mode (73%), but the drop is much smaller than commonly observed (9% in the two-speaker tests reported in Section II). Although these results are preliminary, they suggest that NEXUS may not suffer from some of the limitations of existing recognition architectures in working with multiple speakers. To further evaluate the potential of NEXUS in multiple-speaker situations, we next performed a series of experiments where NEXUS was trained and tested on both male and female speakers.

Recognition Tests with Male and Female Speakers

Four tests were conducted where NEXUS received training and testing on both male and female speakers. These experiments were designed to test: 1) The ability of NEXUS to recognize utterances from both male and female speakers; 2) The sensitivity of NEXUS to specific training orders; 3) The sensitivity of NEXUS to grouping versus interleaving speakers of the same sex; and 4) The ability of NEXUS to generalize to novel speakers when the learning mechanisms have been disabled.

The first two multi-sex tests of NEXUS used a blocked training order, where the first recognition set of all male speakers was presented to the system, followed by the first recognition set of all female speakers. Next, the second recognition sets were presented in the same blocked order, and so on until all four recognition sets had been presented. The speaker order in experiment MS-1 was MEL, MJM, MMP, MSW, FCB, FLH, FRR, FTV. Experiment MS-2 reversed the order of speakers within sexes: MSW, MMP, MJM, MEL, FTV, FRR, FLH, FCB. The results of these experiments are shown in Tables 4 and 5.

Several important trends are evident in these tables. The most obvious is that recognition performance failed to reach the level obtained in the single-sex experiments. Using the average performance score for trials 2-4 as an index of performance, recognition performance dropped from 77% on the single-sex experiments to 42% in MS-1 and 43% in MS-2. Expressed as apercentage error score, the error rate more than doubled from the single-sex

to the multiple-sex experiments. Obviously NEXUS is making many confusions between reference patterns speakers of opposite sex. With the limited training we were able to give the system, NEXUS was not able to develop networks general enough to simultaneously recognize both sexes of speakers.

Table 4: Percent Correct Recognition Rates for Augmented Alphabet in Experiment MS-1

	Recognition Sets 1	2	3-4	Average 2-4
Females				
fcb	6%	47%	48%	48%
flh	19%	22%	31%	28%
frr	22%	44%	52%	49%
ftv	59%	62%	64%	64%
Average	27%	44%	49%	47%
Males				
mel	0%	25%	30%	28%
mjm	0%	34%	31%	32%
mmp	22%	41%	50%	47%
msw	12%	34%	42%	40%
Average	9%	34%	38%	37%
All Speakers	18%	39%	44%	42%

Table 5: Percent Correct Recognition Rates for Augmented Alphabet in Experiment MS-2

	Recognition Sets 1	2	3-4	Average 2-4
Females				
fcb	66%	50%	55%	53%
flh	12%	22%	28%	26%
frr	28%	53%	56%	55%
ftv	6%	47%	60%	56%
Average	28%	43%	50%	48%
Males				
mel	16%	22%	23%	23%
mjm	22%	31%	31%	31%
mmp	22%	41%	58%	52%
msw	0%	34%	48%	44%
Average	15%	32%	40%	38%
All Speakers	21%	38%	45%	43%

In spite of this difference, there are many similarities between these results and earlier findings. Considering the learning trend shown across recognition sets, NEXUS improves throughout the experiments. In both experiments MS-1 and MS-2 we find a large improvement between set 1 and set 2, with a smaller improvement between recognition set 2 and the average of 3 and 4.

Contrasting experiments MS-1 and MS-2, we can evaluate the sensitivity of NEXUS to different training orders. NEXUS uses incremental learning techniques that are influenced heavily by initial items. Under these circumstances, the similarity between the performance results for the two different speaker orders is astonishing. On the first recognition sets, only a three percent difference in recognition rate is observed. After that point, the recognition scores are within one percent of each other. These results were not anticipated, as the networks created by the learning mechanisms are strongly influenced by the first recognition cycle. Different speakers were used as the nucleus of the recognition network in the two experiments, which quickly grew to resemble one another. This evidence strongly suggests that <u>in practice</u> NEXUS is insensitive to training histories.

Experiment MS-3 represents an additional test of the sensitivity of NEXUS to training history, and explores an interleaving of male and female speakers in training. The order of presentation of speakers was MEL, FCB, MJM, FLH, MMP, FRR, MSW, FTV. Although a male speaker was first, speakers were not grouped by sex as in earlier experiments. Table 6 presents the performance observed in this experiment.

The average performance observed in Experiment MS-3 closely resembles the performance observed in the two previous experiments. This is true for overall performance, which is within one percent of earlier experiments, as well as for the acquisition curves. Experiment MS-3 differs by no more than 3 percent from Experiment MS-1 on any of the utterance sets. Comparing all three experiments, it appears that massive changes in training history have

little effect on the performance of NEXUS. The learning algorithm seems insensitive to training order.

Table 6: Percent Correct Recognition Rates for Augmented Alphabet in Experiment MS-3

	Recognition Sets			Average
	1	2	3-4	2-4
Females				
fcb	0%	56%	55%	55%
flh	22%	25%	39%	34%
frr	31%	44%	52%	49%
ftv	59%	66%	56%	59%
Average	28%	48%	50%	49%
Males				
mel	0%	25%	25%	25%
mjm	0%	34%	36%	35%
mmp	22%	47%	47%	47%
msw	12%	38%	39%	39%
Average	9%	36%	37%	36%
All Speakers				
	18%	42%	44%	43%

One curious difference between these three experiments and their single-sex counterparts is the shift in relative accuracy between the male and female groups. Males were better-recognized in the single-sex tests, averaging 8% higher than females. In all three of the multiple-sex tests, performance on males was inferior to females. This result is particularly puzzling since NEXUS depends heavily on early training examples to build the networks. In all three experiments, the first recognition set is male, and in the first two experiments, the first four recognition sets are male. Thus the training history would appear to bias the network to perform better on males than on females. Additional tests where training orders are completely reversed have been planned to help explore this change in greater detail.

The final experiment in the multiple-sex test differs substantially from the first three. Training was limited to two male and two female speakers: MJM, MMP, FLH, FRR. After the first four recognition sets had been given to NEXUS, the learning mechanism was turned off. This prohibited further changes to the network. Under these circumstances, NEXUS more closely

resembles traditional speech recognition architectures, which rigidly separate training and performance modes.

After the training phase, NEXUS was tested on the 5th and 6th recognition sets for all speakers, including the two male and two female speakers that had never been used in training. This experiment represents a good test of the ability of NEXUS to generalize a recognition network to new speakers. As learning is no longer permitted, NEXUS cannot quickly begin adapting to a novel speaker. This test avoids the limitations of previous "novel speaker" tests. The results of experiment MS-4 are shown in table 7.

Table 7: Percent Correct Recognition Rates for Augmented Alphabet in Experiment MS-4

	Recognition Sets			
	1	2	3-4	5-6
Females				
fcb	--	--	--	67%
flh	0%	56%	55%	59%
frr	12%	90%	92%	88%
ftv	--	--	--	67%
Average	6%	73%	73%	70%
Males				
mel	--	--	--	30%
mjm	0%	81%	78%	75%
mmp	22%	69%	78%	72%
msw	--	--	--	56%
Average	11%	75%	78%	58%
All Speakers	9%	74%	76%	64%

Curiously, NEXUS performs better in this experiment than in previous multiple-sex experiments. Performance in the training phase is roughly comparable to performance in single-speaker tests, even though one of the four speakers (flh) consistently received the lowest recognition rates throughout all experiments. In the test phase, after learning was no longer permitted, NEXUS correctly recognized 73% of utterances from familiar speakers, and 55% of utterances from unfamiliar speakers. It is puzzling that performance on novel speakers in this experiment should exceed the performance when the training set includes those speakers. Unfortunately,

the experimental design was slightly flawed. The recognition sets tested in performance mode were not identical to those tested in previous experiments. Of course, this is impossible for familiar speakers, as these sets had been used to train the system. Yet the first four recognition sets for the unfamiliar speakers could have been given to the system, which would permit more specific comparisons with previous experiments.

Collectively the four experiments provide an interesting overview of the ability of NEXUS to simultaneously recognize speakers of both sexes. The differences observed between the single-sex and multiple-sex tests suggest that NEXUS is making confusions between male and female templates much more often than templates are confused within members of the same sex. Knowing the differences in vocal tract length and corresponding differences in formant locations between males and females provides a possible explanation of these confusions.

Reducing these confusions will doubtless require changes in the acoustic representation used in NEXUS. The learning mechanisms depend on reliable differences in sounds that do not appear to be present in the current representations, or NEXUS would not asymptote at such a low recognition level. Many investigators have suggested the use of **frequency-ratio** relationships, where the ratios of formant frequencies are extracted. These ratios are more invariant across differences in vocal tract length than are absolute formant locations (Peterson and Barney, [7]). Results of the single-sex tests suggest another possible avenue of exploration. Instead of attempting to create a single representation for all speakers, it might be possible to add information in to the representation that made it clear whether the speaker was a male or a female. In this case, the learning mechanisms might build parallel networks for males and females. Although this approach seems wasteful in storage and processing resources, human listeners automatically detect the sex of a person who they are listening to, even when they cannot understand the language the person is speaking (Cherry, [8]). The best means of exploring this issue might be to conduct perceptual studies on human speech perception to determine the importance of speaker-sex in human speech understanding.

V. Tools for speech analysis

A part of the contracted effort was to have been the development of special purpose software that permitted users to work with speech in several different representations. The workbench was to have been graphically oriented, displaying speech in several different representations (pulse-code modulation, fft, lpc, etc), coupled with functions that allowed direct manipulation of speech: cutting out sections of waveforms, splicing between waveforms, amplifying or attenuating waveforms, and so on.

Groundwork for this effort had already been done. A speechbench package, developed by Gary Bradshaw, permitted graphic manipulation of pcm waveforms, and the laboratory had purchased the ILS package from Signal Technology, Inc. The development of the workbench required the integration of these two software packages, and was intended to be performed on the ICS Speech Lab MicroVax I computers.

Several factors precluded significant progress in the development of an integrated speech workbench. The ILS package was developed for Berkley 4.2 Unix. This operating system was not available on the MicroVax computers, but a related operating system, ULTRIX 32-m, was purchased. Incompatibilities between the different operating systems led to problems in compiling and loading the ILS package. These incompatibilities were corrected. Next, the 31 MByte hard disk on our MicroVax I could not simultaneously store the source code, the object code, and the executable code for the ILS package. Modules had to be restored from the floppy drive, compiled, and the source code deleted. This piecemeal process was exceedingly slow, requiring several days to build the ILS system. The speech lab applied for University funding to purchase a new disk drive to supplement the limited-capacity drive available on the MicroVax. Unfortunately, we did not receive these funds until January, after the end of the research contract period. We now have a high-capacity, high-speed disk drive for one of the MicroVaxes.

The version of ULTRIX supplied with the MicroVaxes had several bugs in the system terminal library file. Many of these functions were used in the Speechbench package. Since we did not have source code, it proved to be

difficult to correct the system-library problems. Negotiations began with DEC to get source-code tapes for ULTRIX. The University of Colorado agreed to purchase a site license for ULTRIX sources, although DEC has not yet supplied the source code tapes.

Given all of these problems, we decided to set aside our efforts to create an integrated workbench on the MicroVaxes. Instead, we turned our attention to the NEXUS system, resulting in the progress discussed in the previous three sections.

VI. Towards an improved acoustic representation

Since human speech recognition capabilities far exceed the best current recognition systems, it may be useful to model recognition systems on human characteristics. Lea [9] expressed the matter this way: "...While a machine need not operate internally in the same manner as the human, the human speech processing abilities can serve as a successful 'prototype system' for guiding the development of machine algorithms for speech recognition." (p. 42). This has been our strategy in the development of a new frequency scale for the acoustic representation of sound.

Although we believe it is important to use our knowledge of human processes to develop an acoustic representation for speech systems, slavish duplication of low-level details of human auditory processing seems inadvisable. Biological systems emerge as a balance among conflicting evolutionary pressures, and existing structures are adapted to serve new ends. Instead of capturing all of the minutia of human auditory processing, our philosophy is to use human data as guidance and inspiration for the design of acoustic processes. Nowhere is this problem more evident than in the information we have about various levels of human auditory processes. Sound is transduced by the cochlea into neural impulses, then carried by the auditory nerve up to the auditory cortex of the brain. Each of the structures in the chain serves more than just to passively relay raw acoustic information to the brain, but instead actively transforms the signal it transmits. An exacting duplication of these processes may be possible, but is certain to be difficult. Instead, we have focussed on developing a representation similar to that observed in the auditory cortex, which is the encoding most likely to be utilized by other areas of the brain in speech recognition. Thus we seek to mimic the final result of auditory processing but will not attempt to duplicate the processes the auditory system utilizes to achieve that end.

Although detailed electrophysiological data exists on various levels, psychophysical data on human perception is probably more relevant to our needs. Electrophysiological studies often focus on a single level, and consider the response of isolated cells. The brain has information from many

of these cells available at all times, and can perform complex operations on cell groups. Studies at the auditory nerve (e.g., Kiang, Sachs and Peake, [10]; Evans, [11]) reveal a particular complex of filter functions that change markedly in bandwidth and in filter shape as a function of frequency. Those results have been used to guide the development of acoustic representations for speech recognition systems (Searle, et al., [12]). Studies of single-unit response in the auditory cortex, where phonetic decisions are probably made, reveal quite different response patterns (Goldstein and Abeles, [13]). Physiological information remains too sketchy and incomplete to provide an integrated account of the auditory system. At this time, psychological studies seem to provide a better understanding about how the intact auditory system functions as a unit.

Stevens and Davis [14] provide one of the best treatments available of auditory psychophysics. Their work and more recent advances are summarized in texts by Durrant and Lovrinic [15] and by Moore [16]. These works form the basis of the following comparison between the human auditory representation and sound representation in NEXUS.

Representation of speech in the human auditory system and in NEXUS

The acoustic representation processes currently included in NEXUS were derived from a standard template-matching representation system, CICADA, developed at Carnegie-Mellon University. Several differences exist between this representation and what is known about human auditory processing. This section will discuss all pertinent differences, while the next section outlines changes we intend to make to reduce important differences.

Frequency range: A digitized representation can only encode frequencies at slightly less than half the sampling rate (the so-called nyquist frequency). NEXUS samples data at 16 KHz, effectively limiting frequency response to 8 KHz. Adult males hear typically have a hearing range from 20 Hz to 16-20 KHz. Adult females and juveniles may have a higher limit. Sampling the microphone output at a 40 KHz rate would extend the frequency range up to 20 KHz. Most acoustic phoneticians believe that 8 KHz is the functional limit for phonetic information, so it is doubtful that increasing

the sampling rate would improve recognition performance. The additional information would slow down the system operation, because the higher sampling rate increases acoustic storage by 2.5 times; matching times would increase by a commensurate amount. Changes to the sampling rate in order to improve the frequency range do not appear necessary or advisable at the current time.

Loudness response: The perceived loudness of sounds differs markedly as a function of the sound frequency in human perception. Human listeners will judge a constant-intensity tone to change in loudness as it is swept along the frequency scale. The perceived loudness of a constant intensity tone will be greatest when the tone frequency is in the range between 500 Hz and 5 KHz, and will appear to fall off as the tone frequency increases above this range or decreases below this range. Equal-loudness functions specify the intensity of tones across frequencies necessary to produce a constant perceived loudness (Durrant and Lovrinic, [15]). Different equal-loudness functions are generated depending on the intensity of sound at center frequencies.

NEXUS does not include any loudness compensation. Intensity values determined from the FFT routine have a uniform effect across the entire frequency range, although a standard pre-emphasis filter boosts the loudness of high frequencies slightly. The inclusion of a loudness judgement function might have a considerable effect on processing speech signals, particularly in the range below 500 Hz. We consider the equal-loudness function to be worthy of exploration. However, the function is a complex interaction between sound frequency and sound intensity, and will be difficult to duplicate. We intend to investigate this factor in a later phase of the project.

Spectral representation: The human ear extracts information about the energy present at a number of frequencies simultaneously. Microphones, used to input data into computer systems, generate a simpler coding system that encodes the waveform intensity across time. NEXUS uses a Fourier transform to translate the data from the time domain into the frequency domain. The output of the Fourier analysis crudely approximates the representation

derived by the auditory system, encoding the energy present into a small number of channels that have fixed center frequencies and bandwidths. Several important differences exist between the current spectral representation and related auditory processes. These discrepancies are undoubtedly important in limiting the performance of NEXUS, and we intend to extensively revise the processes involved in generating the spectral representation.

The major differences we plan to address in the immediate future include the sensitivity function of each channel to various frequencies, the logarithmic placement of channel center frequencies, and the change in channel bandwidth across frequency range.

The current spectral representation is derived from a 128-point FFT. Each FFT point encodes the intensity of acoustic information present in a 62.5 Hz region. To reduce the number of channels that NEXUS must consider, five adjacent FFT points are added together. This reduces the number of channels to 26, each of which has a bandwidth of 312 Hz.

NEXUS cannot distinguish sounds that occur within a single channel. If a sine-wave tone were varied in the range from 374 to 686 Hz, NEXUS would be unable to distinguish any variation in pitch, because this range is covered by a single channel. Human listeners can distinguish between 20 and 40 distinct pitches in the same interval. At higher frequencies, the human ability to resolve pitch changes is reduced, yet the mapping in NEXUS remains constant, so the difference between human performance and the current spectral representation is attenuated.

Summing FFT channels in this simple fashion has a second important consequence. Imagine that NEXUS were presented with a sine-wave tone that swept between the center frequency of two adjacent channels. As long as the frequency change fell within a single channel, no difference would be noticed. Then, as the tone passed over the boundary between the two channels, the energy would rapidly shift between the two channels. It would appear that an abrupt step change in frequency had occurred. The

representation is insensitive to large changes that occur within a single channel, then indicates a substantial change as the tone makes the small transition between channels. This representation does a poor job of tracking a relatively simple acoustic event.

If the channel shape is changed, so that the responsiveness of each channel varies as a function of the distance between the signal and the characteristic frequency of the channel, this problem is eliminated. The human auditory system appears to work in this way. Channels have a broad sensitivity gradient that declines as the distance between the channel and the center frequency increases. Channels overlap substantially in their sensitivity gradients, so that several channels will be activated by a single tone. Tone sweeps are not indicated by channels switching on or off, but instead by shifts in the relative activation between channels.

The second type of spectral difference concerns the frequency scale. The human auditory system does not treat frequencies as if they derived from a linear scale, but instead seems to map frequencies on a logarithmic scale. Notes on a musical instrument, such as a piano, do not differ by a fixed number of Hz across the musical scale. Instead, the change from one note to the next must increase for higher notes. It seems that the channels used to extract pitch information are not equally placed along the linear frequency range, but instead are equally placed on a logarithmic frequency scale. NEXUS uses a linear scale, where changes occur in equal increments across the spectrum.

Details of the human frequency scale have not yet been resolved. Different experimental procedures give different results, ranging from the purely logarithmic musical scale (Attneave and Olson, [17]) to the more complex functions of the mel scale (Stevens and Davis, [14], p. 81) and of the threshold function for frequency discrimination (Wier et al., [18]). All of these scales share the common factor of decreasing frequency resolution as the frequency increases.

Finally, the auditory system exhibits a bandwidth change that parallels changes in frequency resolution. Channels responsive to high-frequency energy have a wide sensitivity function, while lower channels have a narrow sensitivity function. In contrast, the frequency channels in the current spectral representation are of constant bandwidth across the frequency range.

In our judgement, these differences in frequency resolution, frequency scale, and in the channel bandwidth function between the human auditory system and the current spectral representation are important. We anticipate that the proposed changes (described in the next section) will lead to substantial improvement in the overall recognition performance of NEXUS.

Temporal resolution: NEXUS uses 20 msec sections of the waveform to generate each short-term spectra. The waveform is multiplied by a Hamming window, effectively restricting the waveform to 12 msec. The window is shifted 3 msec to generate successive short-term spectra. The ability of NEXUS to resolve sudden temporal events is mostly determined by the effective duration of the Hamming-windowed speech, 12 msec. Any events that occur within this interval will be averaged together.

Evidence indicates that the temporal resolution of the auditory system is not constant across frequencies. Resolution is poor at low frequencies, and is much higher at high frequencies. NEXUS incorporates an average value, representative of only center frequencies. Some phonetic distinctions depend on small temporal changes in variables such as attack time (e.g., 'sha' vs. 'cha'). This again seems to be an important discrepancy that we propose to address.

Parameterization: All matching in NEXUS is based on the full spectrum, with no attempt to emphasize certain aspects of the signal. No attempt was made to extract specific parameters such as formant values, fundamental pitch, or similar types of information. Parameterization seems to be pervasive in the human auditory system, especially at the higher (cortical) levels. Auditory functions such as lateral inhibition, which emphasizes spectral peaks and zeros, and temporal inhibition, which emphasizes

relatively fast changes in any aspect of the spectrum, may make major contributions to the human ability to recognize speech. Some of these functions are relatively simple to compute, but require a better input than the simple 128-point FFTs currently in use. Other information, such as formant tracks, has proven extremely difficult to derive automatically. Even when methods are identified to reliably extract these parameters, another problem arises: how should spectral and parametric information be combined to yield a matching score. These issues are doubtless important problems worthy of study, but we do not feel prepared to begin investigating them before other problems in the representation have been addressed.

Improvements to the current auditory representation

As we have seen, the current representation differs from human auditory processes in many significant ways. We have identified several changes to the acoustic representation that seem practical and also seem likely to lead to substantial improvements in recognition performance. Instead of trying to completely revise the current representation all at once, we intend to make incremental improvements to the representation. The effect of each change will be assessed independently. This strategy helps direct future efforts to changes likely to be effective, and will prevent extensive investment in modifications that turn out to have an insignificant impact on recognition performance.

The changes we propose could be accomplished more cleanly if the FFT routine was changed from 128 to 512 points. The current 128-point analysis provides a coefficient to represent spectral intensity every 62.5 Hz along the frequency scale. This limits the resolution of mapping FFT points into the channels of the representation. A 512-point FFT provides a more detailed analysis of spectral energy, with coefficients spaced every 15.6 Hz over the frequency range. Given the heavy storage and computation costs associated with this modification, we feel that an initial assessment of more limited-scale changes is warranted. Experience with these changes will provide a good indication of whether the shift to a 512-point FFT is worth the potential benefits.

Three proposed changes to the current speech representation are discussed immediately below. The changes are incremental; each modification builds upon previous improvements. The changes to the auditory representation form a series of representations where each step brings the representation closer to the important characteristics of human auditory representation described in the current section.

REPRESENTATION 1: Mel Frequency Scale

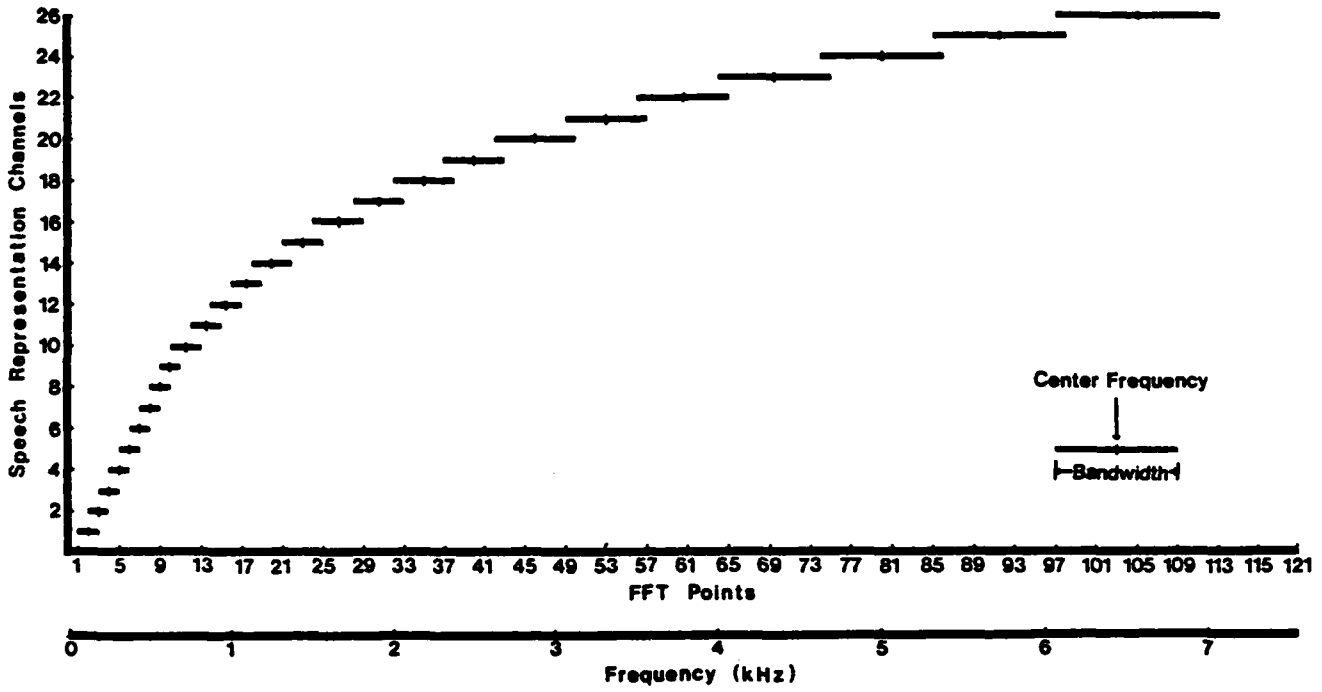
The choice of an auditorially appropriate frequency scale was difficult, because of the conflicting evidence in the literature, discussed above. Our preference was to simulate the frequency discrimination function, but limitations in the resolution of the FFT prevented an accurate simulation of this function in the low frequencies. A more veridical simulation will be tested later when the FFT is changed to 512 points. The same limitations also forbade immediate exploration of the musical scale. The one auditorially plausible scale of frequency that could be simulated with reasonable accuracy from the 128-point FFT was the mel scale.

The mel scale does not conform to a simple mathematical function. In the region below 500 Hz, the mel scale is nearly linear; as it rises to higher frequencies, it comes to approximate a logarithmic scale. Figure 5 shows the frequency-mapping function that we developed. This function is relatively accurate in the higher frequencies, based on the logarithmic approximation to the mel scale. However, the function deviates somewhat from the mel scale in the region between 500 and 600 Hz. In this region, the mel scale changes rapidly from a linear scale to a log one. Due to the low number of channels in this region provided by the 128-point FFT, we had to preserve a linear mapping up to 600 Hz. Thus, the linear portion of the frequency scale and bandwidth function used in the representation extends roughly 100 Hz higher than the corresponding linear region of the Mel scale.

Fifth-octave channel bandwidths were selected to equally subdivide the frequency scale into 26 channels, the number used in the current NEXUS representation.

Figure 5

Speech Channel Parameters based on Critical Band Function



Implementation of the mel scale representation will be accomplished by a matrix multiplication -- for each of the 26 channels the vector of 128 FFT values is multiplied by a vector of weights (0 = no amplitude outside the channel band, 1 = full amplitude inside the channel band, .5 = half amplitude at the channel boundaries). Since the number of FFT points per channel increases along with frequency, the actual values in the weighting vector are normalized (each value divided by the sum of the values in its vector) to give all channels equal weight. We have coded the matrix multiplication routine, and have created a table of channel weights. A new version of the database will be created and tested using the new frequency mapping.

REPRESENTATION 2: Channel Band Shape and Overlap

The frequency channels of the current NEXUS representation only have a minimal overlap. Each channel shares with the adjacent channels half of the

energy of the FFT point at the boundary between the two channels. As described earlier, this method does not reflect smooth frequency changes in a graceful way. Channels in the human auditory representation overlap considerably, permitting a smoother reaction to frequency change.

The second change we intend to make in the representation is to use a smooth channel sensitivity gradient similar to human auditory processing. Channels will receive input from a large number of FFT points, although the impact of an FFT point on a particular channel will diminish as a function of the distance between the center frequency of the FFT point and the channel center. Channel centers will be similar to those shown in Figure 5, with minor local displacements necessitated by the constraint that each channel had to center on one of the FFT points. This constraint was required to equate the maximum peak amplitudes of the various channels. Channel overlap was arbitrarily selected to match fifth-octave band-pass filters with slopes of -60 dB per octave. Given the roughly fifth-octave spacing of the channel centers, fifth-octave channel bandwidths produce channels that overlap at their half-power (3 dB down) points.

Like the initial change in representation, this change will be implemented by a matrix multiplication between the short-term spectra of FFT points and a channel-mapping matrix to implement the channel sensitivity function. The weights used in the mapping matrix range from 0 dB, at the channel centers, to -60 dB, at the channel boundaries. The weights used in the matrix are the voltage-ratio equivalents of the normalized attenuation values.

REPRESENTATION 3: Temporal Resolution

As described above, the human auditory system does not maintain a uniform temporal and spectral resolution across the frequency range. Instead, temporal resolution increases at higher frequencies, with a corresponding decrease of spectral resolution. While the modifications proposed in the first two representations mimic the decrease in spectral resolution with increasing frequency, they do not provide for a corresponding improvement in temporal resolution. The temporal resolution is determined by the width of

the Hamming window used in the FFT process. This window is currently set to a relatively wide (20 msec) value.

The final proposed change is to improve the temporal resolution of high-frequency channels. The optimal method would be to compute separate FFT's for each channel, or use digital filtering methods to accomplish the same end. This modification would be computationally very costly, so we propose to test the potential impact of such a modification using a simple split-spectrum method. Two FFTs will be computed, each using a different Hamming windows. A 12 msec window will be used for the higher frequencies, improving temporal resolution; in the lower frequencies a 20 msec window will be used. These FFTs will be combined before the channel mapping operation is applied to the data.

The choice of the cutting point, the frequency at which the spectral representation switches to the data from the 12 msec to the 20 msec window, is somewhat arbitrary. The selection of the crossover point was influenced by two considerations. One plan was to split the quasi-mel frequency scale, proposed in the earlier sections, roughly in half. The other goal was to include in the region of higher temporal resolution the frequency regions that contain most of the relatively fast acoustic changes that are phonetically important. It turned out that there was little conflict between these two strategies, since most of the energy in stop bursts, for example, falls in relatively high-frequency regions. Accordingly, the crossover frequency for the split spectrum was set at 1156 Hz, at the boundary between channels 13 and 14.

The 3 msec step size will be equated for the two halves of the spectrum. After the split spectrum is recomposed, the 26 channels of the speech representation are created by multiplying the resultant values by the channel mapping matrix.

Evaluating the Modified Speech Representations

A series of experiments has been designed to determine whether any of the three proposed alternative representations lead to significant improvement in the NEXUS recognition system. For each proposed change in representation, the experimental procedure will be as follows: first, the original time-domain waveforms will be processed by the revised acoustic procedures, creating an alternative database of utterances. Then four runs of NEXUS will be made with pruning parameters varied between runs to create a representative sample of how the new representations interact with pruning parameters. The three revised representations will be compared with current system performance on the same database of utterances, to provide systematic data informing us about how much each change contributes to improvements in recognition rate.

VII. Development of new vocabularies to test recognition systems

The selection of a vocabulary for a speech recognition system can have a major effect on performance measurements. Vocabularies that include many similar words force a recognition system to discriminate on the basis of fine phonetic distinctions, while vocabularies with unique items provide a much less demanding test. Itakura [1] tested his seminal template-matching system on a 200-word vocabulary of Japanese city names, and the 36-word alpha-digit vocabulary. Recognition errors occurred on 2.7% of the trials with the city name vocabulary, but increased by a factor of four-fold to 11.4% for the much smaller alpha-digit vocabulary. Dixon and Silverman [19] report that the only significant factors affecting recognition system performance are the experience of a speaker in using speech-recognition devices and the recognition vocabulary.

Investigators in the area seldom utilize the same vocabulary, making comparisons of different recognition systems problematic. Apparent success or failure of a recognition system may lie in the choice of vocabulary, not in the correctness of the algorithms and procedures used. This makes it difficult to separate effective techniques from ineffective ones on the basis of reported results. Harder still is an identification of the relative strengths and weaknesses of effective techniques. The situation can be improved by outlining vocabulary design criteria and using these criteria to develop general-purpose databases. Design criteria provide a formal program to evaluate existing vocabularies, revealing their strengths and weaknesses. The criteria can also be used to help develop new vocabularies that are more comprehensive than those currently in use. This process requires substantial effort and linguistic skill, so it is also helpful to identify several general-purpose vocabularies that are applicable in a broad range of recognition tasks. Researchers may decide to utilize a carefully designed vocabulary instead of utilizing an ad hoc database.

Developing, testing, and refining a speech recognition system is an arduous process. An effective vocabulary must be compact, minimizing the number of different lexical items. This helps to ensure that performance tests can be run quickly, speeding the development cycle. Compactness,

though, must not be obtained at the expense of representativeness. Investigators often adjust their system to maximize performance on the database. If important dimensions of speech variability are omitted, the system may become so specialized that transitions to real-world applications are hindered. Finally, our vocabularies should present a true challenge to speech recognition systems. A trivial vocabulary can be so easy that all systems score 99% correct or better, eliminating performance results as a discriminative tool. Difficult vocabularies help to organize techniques on the basis of their effectiveness, and also yield a conservative estimate of recognition system performance. Nearly every recognition system can achieve high recognition rates under ideal conditions. "Worst-case" estimates of system performance are of greater significance both to researchers and to end users, because they specify how robust a recognition system is.

No single vocabulary can simultaneously satisfy the criteria of compactness, representativeness, and difficulty for all of the many proposed speech recognition applications. If a small number of distinct vocabularies can be identified and used as standards, comparisons of different systems will be greatly simplified.

Knowledge and insight into the structure and diversity of natural language must guide vocabulary design in order to permit development of a database that tests recognition systems along appropriate dimensions of speech variability. The database must provide a suitable representation of the wide range of speech sounds and contextual variations encountered in natural spoken language as well as of the fine discriminations made in natural speech recognition. Such representation is acquired only through careful consideration of linguistic phenomena. Inadequate representation of speech variability in a database has disastrous consequences for the application of a speech recognition system.

Linguistic considerations center around language structures (phonemes, syllables, words, etc.) and language processes (deletion, addition, or changes to sounds). It is perhaps easier to evaluate the language structures in a database than to understand what processes are likely to occur. Even an

evaluation of the language structures can be a difficult task. A simple broad transcription of vocabulary items will reveal how many phonemes are present, and revealing which phonemes are missing. Yet phonemes do not occur with equal frequency in languages, and it would be desirable for the vocabulary to duplicate the relative distribution of phonemes in the language. Next, phonetic sequences must be considered, again both in terms of their existence and relative distribution. The same factors must again be considered at the syllable level. A connected speech vocabulary creates opportunities for sequences that do not occur within words, so this factor must also be studied. An evaluation of language processes (cliticization, reduction, etc.) cannot be performed simply by looking at the vocabulary and grammar of a database; knowledge of the language as typically used by its speakers is needed. We cannot present all of the details involved in evaluating a vocabulary here. What follows is a general outline of vocabulary design criteria and a discussion of the factors represented in our two vocabularies.

Speech variability is affected by the speaker, the speech situation, and the vocabulary items. To simplify the discourse analysis, these functions are usually treated as if they did not interact, but they are not in fact independent. Speakers display a wide range of individual differences in their reactions to speech situations and manifestations of speech processes in producing vocabulary items. Differences in the production of words and phrases range from random variation (no two utterances are identical, no speech situation is perfectly replicable, and no two people perceive a word identically) to individual systematic variation (idiolect) and group systematic variation (dialect). Specifics of speech variation as a function of the speech situation will be treated in subsequent reports when data recording takes place.

Speaker "independent" vocabulary design criteria include an analysis at the phonemic level (existence and distribution of phonemes), the phonetic sequence level, and the syllable level. Since there are only a small number of English phonemes, relatively small vocabularies can include all phonemes with approximately the correct distribution. However, the number of phonetic

sequences and syllables is so great that most practical vocabularies only provide a small sample of the possible units. In this situation, a more appropriate analysis is to investigate the representation of linguistic units by class or type. At the phonemic level, sounds are considered by groupings of manner of articulation (stops, fricatives, nasals, etc.) or place of articulation (bilabials, dentals, palatals, etc.) which is useful because members of sound classes behave very similarly in the same phonetic contexts. Syllables are open (V, CV) or closed (VC, CVC) and combine to represent the canonical forms of words. Phrases, clauses, and sentences are described by phrase structure types. At each level of vocabulary structure, and at this superficial level of analysis, representativeness consists of accounting for numbers and classes of units known to exist in the language. Many vocabulary design descriptions go no further, which, in our view, is a serious shortcoming.

Much more important than tallies or frequency counts is the distribution of linguistic units across phonetic contexts. Many phonetic processes which underlie variation in speech are conditioned by phonetic environments. Distributional representativeness is a consideration of the combinatorics of linguistic units and important linguistic patterns. Prominent linguistic patterns are unit initial position, unit medial position, unit final position, and unit adjacencies; and basic linguistic units are the phoneme, syllable, word, phrase, sentence, discourse, and intonation contour. The occurrence and result of phonetic processes in the data base can be predicted on the basis of what is known about speech combinatorics.

Phonetic processes are particularly complex in connected speech. When the sentence "Didn't you go eat yet?" is spoken fluently, it may be pronounced as [dɻčəgəwi?ǰɛ?] rather than [dɪ dn?t ju go it jît][2] . Several speech processes can be observed: when [dɪdn?t ju] becomes [dɻčə], clearly several segments have been deleted, in addition to the contraction process already indicated in standard orthography. The vowel [ɪ] assimilates the nasal feature of [n]; the stop [t] and glide [j] assimilate to form the affricate [č], the vowel

---

2. Ladefoged [20] describes the phonetic representation system we are using.

[u] is reduced to the unstressed [ə], and the negative and pronominal elements are pronounced as part of the verb (cliticization). When [go it jīt] becomes [gəwi?jī?], [o] is reduced to [ə] while its rounded quality remains in the [w] sound, and the alveolar stop [t] is approximated by glottal stop [?]. Vocabularies for connected speech recognition must be carefully chosen to maximize the occurrences and varieties of speech processes they can evoke.

Two vocabularies are presented in Appendix B. The first vocabulary is intended as a test of the ability of a system to make fine phonetic distinctions. It is an augmented form of the alphabet vocabulary that has been used in many laboratories. The augmented alphabet vocabulary is quite compact, reasonably representative, and very tough. It consists of 32 monosyllabic items. 55% of syllables are CV in structure (consonant followed by vowel), 35% are (C)VC, and 10% are V. All the consonant phonemes of English are represented except [ž] which occurs rarely in English. Glottal stop is automatically present in vowel-initial words and is not indicated in the transcription. Over half the vowels and diphthongs of English are represented, including front, back, high, mid, and low vowels.

All the items in the vocabulary make a minimal pair with at least one other item--six minimal pairs sets in all. Minimal pairs are easily confusable items which force the recognition system to match items on the basis of different phonetic qualities rather than or unique item shapes, like "double u." The vocabulary may be made even more compact to accommodate the computational load of a multi-speaker recognition test by sampling items across minimal pair sets.

Consonants are more heavily distributed in initial position; no stop consonants occur in word final position except in two consonant clusters with [k]. Final position consonants consist of fricatives, nasals, and glides. While these distribution patterns make the vocabulary less generally representative of English, they maximize potential confusability of items which provide a rigorous test of an important aspect of word recognition.

The second vocabulary is intended as a connected speech vocabulary, which includes many opportunities for complex interactions between words. The vocabulary of short phrases is flexible in size, is fairly representative, and is very difficult. It has a lexicon of 38 words which can be combined to produce many varieties of three common sentence types: imperative sentences, declarative sentences, and interrogative sentences. Most of the speech sounds of English are represented. With the inclusion of two nouns (thing, book) and three verbs (show, give, save), all sounds are represented except the infrequent sounds of [ž], [ɔɪ], and [au]. In spite of its compactness, the vocabulary contains many of the important distributional patterns of English. All the major consonant classes (stops, fricatives, nasals, laterals, and glides) appear in word initial and word final position. All four syllable types are represented in distributions appropriate for English, and the same syllables appear in a variety of contexts providing a full range of stressed, full, and reduced syllables. Also, both rising and falling intonation patterns are represented, and the distribution of either can be generated within the data base by manipulating the context in which the utterances are spoken. (File the text with the letter. File the text with the letter.)

The particular challenge to recognition provided by this vocabulary is the pronunciation of cliticized words. Pronouns, auxiliary verbs, articles, and prepositions all share the property of undergoing variable reduction, which may be drastic in naturally spoken English. In large part, this phenomena occurs because these words are often pronounced as part of the preceding word. Clitics will be extremely difficult to recognize because clitics closely resemble one another, and there is so little phonetic character that remains to disambiguate these items. The vocabulary maximizes the reductions of vowels, simplification of consonant clusters, assimilations of proximal sounds, and deletions of consonants and vowels in connected speech, providing a conservative recognition test.

Available recognition strategies in NEXUS will be utilized to investigate these processes. Most of the words are distinct from one another in many ways, providing the matcher with several cues. Minimal pairs are present in the vocabulary as well, forcing the recognizer to handle both gross and fine

distinctions. The sentences are all short enough to be spoken in one breath group eliminating any pauses which could be otherwise used as cues. The vocabulary provides a wide variety of contextual variants, since words can appear in a variety of positions and in combination with numerous other words.

New strategies, probably including some knowledge of cliticization processes, will have to be devised for NEXUS to recognize the processes operating in this vocabulary. Differences in length and acoustic trajectories make it improbable that full words can be matched against cliticized pronunciations. Simple variants on the current matching system, such as weighting prosodically prominent syllables, are inadequate because the correct identification of clitics is often crucial to understanding the meaning of a sentence, yet clitics are usually unstressed and very short in duration. The connected speech vocabulary is appropriate to test recognition strategies which make use of rhythmic structures for nonlinear matching rules.

VIII. Planned enhancements of the NEXUS system

The long-term goal of our laboratory is to develop speech recognition technology to the point where natural speech recognition systems become practical. These systems must exhibit near-human performance in their ability to accurately identify connected speech produced by a wide range of talkers. The most critical shortcoming of existing systems is their limited ability to correctly recognize and exploit acoustic information. Accordingly, our laboratory has concentrated on the processes required to perform the mapping between an acoustic waveform and phonetic sound classes.

The work described in this report has been limited to isolated-word situations. Since many of the ideas behind NEXUS were novel and untested, it seemed necessary to gather practical experience with these concepts in a simple isolated-word situation. Without some demonstration of the effectiveness of the techniques, an attempt to solve the general speech recognition problem could not be justified. We feel that the current results support the learning approach of NEXUS, so we are turning our attention to the larger problem of connected speech.

Connected speech differs radically from words produced in isolation (Oshika, et al., [21]). Probably the most important difference results from coarticulatory effects between words in a sequence. The section on vocabulary development mentioned other problems: cliticization and reduction. All of these phenomena cause severe distortions to the acoustic patterns associated with words. We should not expect recognition systems that ignore these phenomena to be successful; published reports of attempts to generalize isolated speech recognition systems to connected speech confirm this expectation.

The philosophy behind the NEXUS project has always been that machines ought to learn about speech phenomena. The current version of NEXUS is not capable of acquiring all of the types of information needed to interpret connected speech correctly. Our first steps toward connected speech will involve the development of new learning mechanisms. We do not yet understand

whether learning will be appropriate to solve all of the new problems of connected speech recognition. Some connected-speech phenomena may be more naturally solved by building certain capacities directly into the recognition system, instead of forcing NEXUS to acquire these capabilities on its own. Our approach is best interpreted as a bias towards using learning mechanisms, yet this bias does not blind us from the possibility that it may be more appropriate to directly program some "innate" capabilities into machines.

The most important limitation of the current version of NEXUS is that word models ignore their acoustic context. NEXUS has no way of knowing that the articulation of words will vary as a function of their phonetic surroundings. NEXUS must be equipped with the capacity to detect context changes in speech, and understand the implication of these changes in recognizing words. The current learning mechanisms must be modified to provide these capabilities to NEXUS.

The first change will be to provide NEXUS with a symbolic description of words. NEXUS currently encodes all word descriptions as a sequence of nodes in the recognition network. Alternative pronunciations of words are represented as separate node sequences. The current version might encode the word "B" in the following way:

```
      node 37  --  node 67
OR    node 33  --  node 86
OR    node 37  --  node 95
```

Although it seems that node 33 and node 37 might be equivalent, NEXUS does not make this assumption. The segmenter, responsible for the identification of word node structure, sometimes detects the onset burst and sometimes detects an offset breath release. Node 37 might correspond to /b/ and node 67 to /i/, while node 33 might correspond to /bi/ and node 86 to /h/. The current representation does not highlight specific features of nodes, such as their length or contents. Without this information, it becomes difficult to reason about node sequences that combine to form alternative word structures.

NEXUS needs a more abstract description to highlight this information. An example of this symbolic description for the same word "B" might be:

```
(B consists-of  short-turbulence-segment 37
                followed-by long-vocalic-segment 67
            OR  followed-by long-vocalic-segment 95
 OR consists-of long-vocalic-segment 33
                followed-by short-turbulence-segment 86)
```

Symbolic structures of this sort will facilitate the development of sophisticated reasoning mechanisms. NEXUS should be able to compare these abstract descriptions of different words to uncover their potential similarities and differences. Hypotheses about these differences can be created, and confidence in hypotheses increased as a function of later experience. We anticipate that the next version of NEXUS will learn more slowly, but that it will be able to create more sophisticated descriptions of words.

A related development will be to explicitly identify equivalence sets of acoustic patterns. In our description of "B" above, the patterns 67 and 95 represent two different articulations of the phonetic segment /i/. NEXUS does not encode these as equivalent, but simply treats them that way for the word "B". If NEXUS discovers that pattern 67 can also be used in describing the word "D", it does not automatically assume that pattern 95 may be used in "D" also. This strategy prevents overgeneralization, yet forces NEXUS to discover the equivalence of certain patterns in many different contexts. When we shift to symbolic word models, NEXUS will be changed to include an explicit representation of equivalence sets. Thus, a better description of the new word models might be:

```
(B consists-of  short-turbulence-segment class-1
                followed-by long-vocalic-segment class-2
 OR consists-of long-vocalic-segment class-3
                followed-by short-turbulence-segment class-4)
```

(class-1 consists-of 37)
(class-2 consists-of 67 OR 95)
(class-3 consists-of 33)
(class-4 consists-of 86)

Each class of sounds will initially be separate for different words. However, the learning mechanisms may see that two classes share many items, conjecture that the classes are identical, collect evidence relevant to this hypothesis, and decide to collapse the two distinct classes into one. It will be possible for NEXUS to use each acoustic pattern in many different classes. This creates a context-dependent equivalence. In effect, NEXUS will be determining that in a certain context, two sounds can be used interchangeably. However, NEXUS will not be forced into over-generalizations where two sounds are universally considered as interchangeable. This modification will help in the development of contextual dependencies in words.

Finally, modifications to the segmenter will be made so that transition segments are identified. When words are produced in sequence, word beginnings and endings are often sharply changed by inter-word coarticulation, while the center segments of words remain more stable. If the segmenter identified word-initial and word-final transition regions as segments distinct from other sections of a word, NEXUS could discover that the variability of these segments is greater than more central segments. In this way, NEXUS could learn to attend to stable center segments as the best cues to word identity, and would accordingly focus its attention on these regions.

References

1. Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-23, 67-72.

2. Bradshaw, G. L., Cole, R. A., and Li, Z. (1982). A comparison of learning techniques in speech recognition. Proceedings of the 1982 International Conference on Acoustics, Speech, and Signal Processing, Paris, France, 570-573.

3. Bradshaw, G. L. (1985). Learning to recognize speech sounds: A theory and model. CMU Technical Report CMU-CS-85-136.

4. Rabiner, L. R., Levinson, S. E., Rosenberg, A. E., and Wilpon, J. G. (1979). Speaker-independent recognition of isolated words using clustering techniques. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-27, 336-349.

5. Lowerre, B. T. (1976). The Harpy Speech Recognition system. Ph. D. Dissertation, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pa.

6. Lowerre, B. T., and Reddy, D. R. (1979). The Harpy speech understanding system. In W. A. Lea (Ed.), Trends in Speech Recognition. Englewood Cliffs, New Jersey: Prentice-Hall.

7. Peterson, G. E., and Barney, H. L. (1952). Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 24, 175-184.

8. Cherry, E. C. (1953). Some experiments on the recognition of speech with one and with two ears. Journal of the acoustical Society of America, 25, 975-979.

9. Lea, W. A. (1980). Speech recognition: Past, present, and future. In W. A. Lea (Ed.), Trends in Speech Recognition. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

10. Kiang, N. Y.-s., Sachs, M. B., and Peake, W. T. (1967). Shapes of tuning curves for single auditory nerve fibers. Journal of the Acoustical Society of America, 42 1341-1342.

11. Evans, E. F. (1975). Cochlear nerve and cochlear nucleus. In W. D. Keidel and W. D. Neff (Eds.) Auditory System: Physiology, Behavioral Studies, Psychoacoustics. New York, Springer-Verlag.

12. Searle, C. L., Jacobson, J. Z., and Kimberly, B. P. (1980). In R. A. Cole (Ed.), Perception and Production of Fluent Speech. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

13. Goldstein, M. H., Jr., and Abeles, M. (1975). Single unit activity of the auditory cortex. In W. D. Keidel and W. D. Neff (Eds.) Auditory

System: Physiology, Behavioral Studies, Psychoacoustics. New York: Springer-Verlag.

14. Stevens, S. S., and Davis, H. (1938). Hearing: Its Psychology and Physiology. New York: American Institute of Physics.

15. Durrant, J. D., and Lovrinic, J. H. (1984). Bases of Hearing Science. Baltimore, Md.: Williams and Wilkins.

16. Moore, B. C. J. (1982). An Introduction to the Psychology of Hearing. New York: Academic Press.

17. Attneave, F., and Olson, R. K. (1971). Pitch as a medium: A new approach to psychophysical scaling. American Journal of Psychology, 2,147-166.

18. Wier, C. C., Jestaedt, W., and Green, D. M. (1977). Frequency discrimination as a function of frequency and sensation level. Journal of the Acoustical Society of America, 61, 407-416.

19. Dixon, N. R., and Silverman, H. F. (1981). What are the significant variables in dynamic programming for discrete utterance recognition? Proceedings of the 1981 International Conference on Acoustics, Speech, and Signal Processing, Atlanta, Georgia, 728-731.

20. Ladefoged, P. (1982). A Course in Phonetics, Second Edition. New York: Harcourt Brace Jovanovich, Inc.

21. Oshika, B. T., Zue, V. W., Weeks, R. V., Neu, H., and Aurbach, J. (1975). The role of phonological rules in speech understanding research. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-23, 104-112.

APPENDIX A

Learning to Understand Speech Sounds: A theory and model

by
Gary L. Bradshaw

Carnegie-Mellon University
Pittsburgh, Pennsylvania

Table of Contents

List of Figures

List of Tables

Acknowledgements

The work described in this document was a protracted labor of love. While I was involved with this project, I incurred many debts to friends and colleagues, and I would like to take this opportunity to acknowledge these debts.

First and foremost, I would like to thank my dissertation committee, Herbert A. Simon, John R. Anderson, D. Raj Reddy, and Brian MacWhinney, for allowing me the freedom to pursue my own vision. Particular thanks go to Professor Simon for chairing a dissertation outside of his many domains of expertise, and to Professor Reddy for providing facilities and support necessary for this dissertation.

I owe a very special debt to the speech group at CMU. Ron Cole introduced me to the field of speech research, and provided me with background information. Alex Rudnicky, with an encyclopedic knowledge of the literature, helped to provide some of the empirical buttress for my learning theory, and was a valuable listener to my ill-formed ideas. Fil Alleva helped on countless programming problems, and provided a matcher that formed the nucleus of my final version. In addition, Fil's elegant style of programming helped me discover both artistry and style in computer science. Fil, Alex Waibel, and Philippe Specker provided support routines, and information on technical issues. Richard Green patiently recorded the second data base to help evaluate the performance of the NEXUS system. Kai-Fu Li helped organize my descriptions of various pieces of NEXUS by listening to technical descriptions of code. These and the rest of the members of the speech group provided a necessary community for speech research. My debt to this group is unrepayable.

I would like to thank Pat Langley both for introducing me to the mysteries of computational models of learning, and for many enjoyable and valuable discussions of learning issues. Thanks also go to David Nicholas and Jeff Shrager for listening to my ideas on the project.

Janet McDonald's extensive and cogent comments on a primitive draft of this dissertation helped provide necessary organization for this document. Alex Rudnicky, Ron Cole, and David Nicholas suffered through early drafts while Jeff Shrager, Terry Halwes, and Ilse Gayl read later versions. All of these people made invaluable comments and suggestions. Of course, this group cannot be held accountable for remaining errors; such errors are solely due to stubbornness on the part of the author.

Finally, I would like to thank my friends for helping me get through: Matthew Lewis, Ann Beattie, Sandy Milberg, Jeff Shrager, Peter Pirolli, Janet McDonald, and David Nicholas all had wet shoulders at some points.

To all of the above people, I can only say, "Thank you for everything."

This research required substantial resources, both computational and human. In addition to the human resources cited above, this project required a great deal of time and computational resources. I gratefully acknowledge the support provided by NSF grant DCR-8205539 for this research.

1. Introduction

Psycholinguists have long studied human speech communication. One area of particular interest is the process that decodes a spoken utterance into component words. Although a great deal of research has been directed towards an understanding of this process, current models of word recognition are inadequate and incomplete. The present chapter serves to introduce speech recognition, describing key concepts and empirical findings. The following chapters will review limitations of current models, and a new model will be proposed that avoids these limitations. A key assumption of the new model is that basic elements of speech recognition are not innate, but are acquired through experience with language. This assumption is shown to be consistent with a broad range of empirical findings of human speech perception. A computer implementation of the model, NEXUS, is described along with performance results of learning and recognition trials.

1.1 Word recognition

Word recognition is the process of translating acoustic sounds into a sequence of words. It can be understood as a kind of pattern classification activity, beginning with an encoding of the basic stimulus event, and ending with the identification of particular words known to the listener.

Word recognition seems to be a simple and trivial process. Clark and Clark (1977) describe the naive model of word recognition as an analogy to reading printed text. In this model, each phonetic segment is independently produced in sequence, as are letters in a word. Word boundaries would be marked by silences. Still longer silences would mark the boundaries between phrases and sentences. According to this view, listeners only need to be able to discriminate between the basic phonetic segments in their language in order to identify words. Empirical data prove this model to be wrong in several important ways. Phonetic segments, which can be thought of as the "alphabet" of speech, are not isolated independent entities. Tape-splicing experiments show that many phonemes cannot be understood out of context. For example, it is impossible to isolate a segment of speech that listeners can identify as the phoneme /p/ (Harris, 1953; Peterson, Wang, and Silvertsen, 1958). Printed letters have a consistent shape in all words, but phonetic segments do not have invariant shape or properties. The acoustic realization of each segment is a complex function of its context (Cooper, et al., 1952; Harris, 1958; Denes, 1955), the speaker, stress and speaking rate (Lisker and Abramson, 1970; Summerfield, 1975), and general speaking conditions. Also, events in speech are not in a one-to-one correspondence with phonetic segments. A continuous section of an utterance may simultaneously carry information about multiple phonetic segments (Liberman, et al., 1954; Liberman, et al., 1967). Finally, words in a text are marked by blank spaces, but visual displays of speech do not show pauses or silences marking word or phrase boundaries. Without these pauses, a single string of phonetic segments may not have a unique interpretation as a sequence of words. Instead, ambiguous parsings exist that will map a single phonetic sequence into different strings of words (Cole & Jakimik, 1980).

1.2 Speech perception

Currently, research on word recognition is divided into two distinct areas: **speech perception** and **word identification**. Speech perception emphasizes the perceptual processes that occur during word recognition, including the encoding of the acoustic input and its phonetic analysis. Word identification research emphasizes the interdependence of word recognition and communication.

Issues in word identification include the interaction of perceived sounds with lexical, syntactic, and semantic knowledge in the meaningful interpretation of an utterance. Although lexical access is a necessary component of the word recognition process, the recognition system described in this paper is directed towards an understanding of speech perception, and so includes only a simple method of lexical access. Our discussion of the word recognition process will be restricted to issues of speech perception.

In the analysis of speech perception processes, investigators have uncovered two empirical phenomena that have come to form core issues in speech perception. The first phenomenon relates to the nature of speech as it is produced: the sounds of speech are remarkably variable. The second phenomenon concerns our perception of speech: in apparent defiance of acoustic reality, perceptual experience is remarkably constant across a broad range of acoustically different stimuli. These phenomena are referred to as **variability in production** and **perceptual constancy**, respectively. Perceptual constancy can be further analyzed into two important phenomena. The first perceptual phenomenon is that phonetic segments can be said to have **multiple acoustic realizations**. Very different acoustic events are often perceived as identical phonetic segments (Liberman and Studdert-Kennedy, 1977). The second perceptual phenomenon is that identical acoustic events can have different phonetic labels, depending upon the contexts in which events occur (Liberman, Delattre, and Cooper, 1952). Therefore, a single event can have **multiple phonetic interpretations**. The mapping between acoustic events and phonetic percept cannot be one-to-one or even many-to-one; the relation between stimulus and percept must be many-to-many. Understanding how a human listener performs this complex mapping operation is a central problem in speech perception, which we will refer to as the **noninvariance** problem. The following sections discuss these two phenomena, and the resulting noninvariance problem, in greater detail.

1.2.1 Variability in speech

Perhaps the most important problem in speech perception is the lack of invariance of phonetic segments. There are three sources of phonetic variability: contextual variability, talker variability, and random variability. Contextual variability occurs whenever the expression of a phoneme depends upon its environment. A well-studied example of contextual variability is the formant transitions important in identifying the place of articulation of stop consonants (Liberman, et al., 1967). This variability results from physical limitations on movements of the human articulatory structures. Contextual variability is not an accidental and unnecessary property of speech; synthetic speech created without such contextual influences cannot be understood by listeners (Harris, 1953).

Talker variability can be broken down into three major categories: differences in vocal tract length and shape, differences in talker accents, and differences in characteristic articulatory gestures used to make a particular sound. Very little is known about the relative importance of each of these factors in overall differences between talkers. A number of studies have simply investigated gross between-talker differences. Peterson and Barney (1952) measured vowel formant frequencies across a number of speakers, including adults and children, from a number of regions around the country. They found enormous variability for the formant frequencies across talkers. This variability was somewhat reduced when they considered the ratio of f_1 to f_2, but even this measurement exhibited a great deal of scatter across talkers. The properties of fricatives have also been found to differ widely across talkers. In one study Hughes and Halle (1956) found differences to be so great that if the fricatives were excised from one speaker's utterance and spliced into another, a phoneme different from the original would have been understood.

Random variability includes all variability that cannot be attributed to phonetic context or talker identity. Very little research has been done to systematically investigate the amount of random variability normally present in speech. Presumably such variability is a result of random perturbations in articulatory movements occurring even when a single talker tries to repeat the same word in exactly the same manner. Random variability occurs in nearly all coordinated muscular movements, such as serving a tennis ball or threading a needle. The dearth of investigations of random variability precludes formal statements about whether or not such variability is perceptible to listeners, or how much random variability is present. Experience with machine recognition systems has revealed something about the extent of random variability. The simplest word recognition task is to identify isolated words drawn from a small vocabulary of tokens produced by a single talker. Confusion rates for traditional pattern recognition systems may be as high as 50% for even this apparently trivial task (Bradshaw, Cole, and Li, 1982). Pattern recognition systems classify utterances on the basis of their global similarity to an existing set of templates. Recognition errors occur because the distinctive information useful in classifying a word is less salient than the random similarity between different words. Further investigation is needed to help understand the amount and nature of random variability.

1.2.2 Perceptual constancy and categorical perception

In marked contrast to the variability of speech is the constancy of our perceptual experience. As mentioned earlier, most listeners do not even recognize that the sounds they hear are profoundly changed according to their immediate context. **Perceptual constancy** occurs when two different stimuli are treated as functionally equivalent. We have already mentioned that talker characteristics can affect the acoustic properties of an utterance. Listeners readily compensate for such variability as they perceive and recognize speech. Perceptual constancy may occur because the human perceptual system is insensitive to differences in the acoustic signal, or by a more active compensation process in which detected differences are ignored. **Categorical perception** is a special case of perceptual constancy, where two different stimuli can only be discriminated from one another if they fall into different categories. Investigations of categorical perception have been important in speech perception because they provide support for models of speech perception based on perceptual mechanisms, and argue against models that involve higher-level processes to compensate for differences in utterances.

Categorical perception in speech recognition was first demonstrated by Liberman, Harris, Hoffman, and Griffith (1957). These experimenters created a series of synthetic stimuli that differed along the continuum of second-formant transition, an important cue for place-of-articulation. Fourteen stimuli were created with onset frequencies ranging from 1320 to 2880 cps in 120 cps intervals. When subjects were asked to classify the stimuli as /b/, /d/, or /g/, very sharp decision boundaries were found. Next, subjects were tested on their ability to discriminate between items. Pronounced peaks were observed in the discrimination function for adjacent items; pairs drawn from different classes could be much more reliably discriminated than pairs drawn from the same class, even though the acoustic differences were of equal size. Indeed, within-class discrimination performance for these pairs was found to be very close to chance. Tests of pairs of items that differed by two or three steps along the continuum revealed the same basic pattern: Pairs could be more reliably discriminated when items were drawn from different categories than when both items were drawn from the same category. However, discrimination performance on within-category pairs improved as the distance along the continuum between items increased. Thus, perception cannot be said to be truly categorical: empirical tests have instead shown a heightened sensitivity for stimuli that cross categorical boundaries.

Categorical perception has also been shown for voice-onset time (VOT) (Abramson and Lisker, 1970; Liberman, et al., 1961; Lisker and Abramson, 1970), /l/ versus /r/ discrimination (Miyawaki et al. 1975), stop versus continuant discrimination (Cooper et al., 1976; Diehl, 1976), and fricative-affricate distinctions (Cole and Cooper, 1975). Not all speech sounds are categorically perceived, however. Long vowel sounds, for example, can be discriminated in a continuous fashion (Fry, et al., 1962).

Although categorical perception of speech sounds is a robust phenomenon, we must be cautious in our interpretation of such effects. Many researchers in the field have claimed that these experiments demonstrate a perceptual insensitivity to differences between various tokens of a particular phoneme. An important implication of such a claim is that the mechanisms that compensate for such variability occur in the perceptual system, rather than at higher levels. Such a conclusion biases theories of speech perception towards perceptual explanations. However, the generality of categorical perception effects has not been well tested. Existing demonstrations rely on synthetic utterances varied along a single dimension. Little research has been done to evaluate the discriminability of natural tokens. Categorical perception experiments demonstrate a partial perceptual insensitivity to a single acoustic cue and, as such, these experiments are interesting evidence relevant to theories of speech perception, but until more general tests are performed, such demonstrations cannot exclude all other levels as potential explanations for perceptual constancy.

1.3 The noninvariance problem

This chapter briefly introduced important concepts in speech perception. From a production viewpoint, speech was shown to exhibit a great deal of variability due to the influence of systematic and random factors. In contrast, perception remains remarkably unaffected by variability in production. Accounting for the paradox of perceptual constancy under production variability is normally referred to as the noninvariance problem. The noninvariance problem is a central focus in speech perception research. The next chapter will describe several theories of speech perception, and discuss how each theory accounts for the noninvariance problem.

2. Models of speech perception

Given the centrality of speech perception in the word recognition process, it is not surprising that several different theories have been formulated. Each theory must specify the basic units of recognition, their origin, and the processes which operate on them. Most models of speech perception include several stages of processing. A prototypical multiple-stage model utilizes a **phonetic perception** stage to overcome the noninvariance problem, a **phonological stage** to handle allophonic variants, and a **word stage** for final word-recognition (Liberman, et al., 1967; Studdert-Kennedy, 1974, 1976; Pisoni and Sawusch, 1975). From a speech perception perspective, the phonetic perception stage is the most important, because it tries to account for the invariance problem. The next section will review various theoretical accounts of the mechanisms of the phonetic perception stage.

2.1 Models of phonetic perception

The phonetic stage is responsible for the transformation of a continuous acoustic signal into a sequence of phonetic segments. Current accounts of this stage include motor theory (Liberman, et al., 1967; Repp, 1982), analysis-by-synthesis (Halle and Stevens, 1972; Stevens and House, 1972), Pisoni and Sawusch's stage theory (1975), Oden and Massarro's (1978) feature integration theory, and feature detection theories (Eimas & Corbit, 1973; Cooper, 1975).

2.1.1 Motor theory

The motor theory is one of the oldest and most widely known theories of phonetic perception. The basic assumption of this theory is that "speech is perceived by processes that are also involved in its production" (Liberman, et al., 1967). The acoustic signal is said to activate the production system at the level of neural commands to the articulatory system. Through an inversion of the production process, these commands are mapped onto phonetic segments. A one-to-one mapping is assumed between neural commands and phonetic segments. Invariance is thus obtained at the level of phonetic segments. One fundamental difficulty with this approach is the lack of invariance of motor commands for phonemes produced in different contexts (MacNeilage, 1970). Lenneberg (1962) also documented a case of a child who learned to understand speech without being able to speak. Finally, motor theory is not an explanatory theory in the sense of specifying the nature of invariant motor commands, and does not provide a process model of how auditory input is associated with motor commands. Ultimately, this theory casts little light on actual mechanisms of phonetic perception.

2.1.2 Analysis by synthesis theory

Analysis-by-synthesis theory (Stevens, 1960; Stevens and Halle, 1967; Halle and Stevens, 1972) provides an alternative to motor theory. The basic auditory stimulus is submitted to a preliminary analysis. The results of this analysis lead to the formation of a hypothesis about the abstract representation of the utterance. Internal phonological rules, used both for perception and production of speech, generate an internal auditory pattern that is compared with the input. If the resulting match is sufficiently close, recognition occurs; otherwise a new hypothesis is generated, and the process is repeated.

The first difficulty with analysis-by-synthesis is that many rules needed to derive internal auditory patterns are not necessary or useful for the production of speech. Next, the nature of the comparison process is somewhat of a mystery, because a listener has to recognize sounds produced by talkers with quite different vocal tract shapes. It is not clear how an analysis-by-synthesis model can allow a listener to recognize patterns which the listener cannot produce. Like motor theory, analysis-by-synthesis is so poorly specified that it cannot be taken as a process model of phonetic perception.

2.1.3 Feature detector theory

Feature detector theories were developed as a simple alternative to complex production-based decoding schemes, and have become predominant in current speech perception literature. Detector theories drew their inspiration from two sources: linguistic feature theories (Jacobsen and Halle, 1956; Jacobsen, Fant, and Halle, 1963; Chomsky and Halle, 1968) which consider phonemes to be bundles of concurrent distinctive features, and the identification of complex biological feature detectors including visual and auditory detectors in frogs (Lettvin, Maturana, McColloch, and Pitts, 1959; Frishkopf and Goldstein, 1963).

Eimas and Corbit (1973) provided the first explicit feature detector model to account for judgements of voicing contrasts in syllable-initial stops (i.e., /b/ vs. /p/, /d/ vs. /t/, and so on). The model includes two voice-onset-time (VOT) detectors which are maximally sensitive to the modal value of onset times for voiced and voiceless stops, respectively. Each detector has a broad sensitivity gradient proportional to the distance from the detector center. Because the sensitivity gradients overlap substantially, both detectors are likely to respond to a given stimulus. The decision boundary is located at the point where the sensitivity gradients overlap. A further assumption of the model is that only the output of the detector with greater activation is available for further processing. The initial model was later generalized to other dimensions of speech, including the place dimension (Ades, 1974; Cooper, 1974), the stop-continuant dimension (Cooper, Ebert, and Cole, 1976; Diehl, 1976), the fricative-affricate dimension (Cole and Cooper, 1975), and the vowel-height dimension (Morse, Kass, and Turkienicz, 1976).

The basic feature detection model does not immediately explain categorical perception or solve the noninvariance problem. Eimas and Corbit (1973) postulated that higher levels of processing were insensitive to the magnitude of feature detector output, responding only in a binary fashion according to the identity of the detector with the greatest level of activation. Categorical perception was not considered to be a function of the feature detectors themselves, but instead a general property of the perceptual architecture.

Two quite different accounts of the noninvariance problem have been advanced. Cooper (1974) proposed that feature detectors might be sensitive to multiple sets of acoustic cues. This explanation seems somewhat *ad hoc*: no evidence other than adaptation results suggests this architecture, and it is unclear that evolutionary pressures would select for detectors conveniently attuned to the appropriate set of cues. Stevens (1971, 1972, 1975) provided a more elegant account of the noninvariance problem. He argued that even though the acoustic signal was quite variable, certain relations between acoustic events were invariant. The implication is that there is no noninvariance problem if these invariant relations can be detected. In the words of Blumstein and Stevens (1981): "invariant acoustic properties can be derived directly from the acoustic signal." (p. 27). This proposal, along with the assumptions of Eimas and Corbit, allowed feature detector theory to account for categorical perception and noninvariance.

The feature detector model gained further credibility with demonstrations of categorical perception for speech stimuli in human infants. Eimas, Siqueland, Jusczyk, and Vigorito (1971), utilizing a sucking paradigm with one- and four-month-old infants, were able to demonstrate categorical boundaries for VOT that closely resembled the boundaries of adult English speakers. These results have been extended to place of articulation (Eimas, 1974; Miller and Morse, 1976) and to fricative contrasts (Eilers and Minifie, 1975). The implication of this evidence, particularly given the young age of subjects in these experiments, was that feature detectors were innate mechanisms in the human perceptual system, probably specialized for human speech.

Independent support for feature detector theory was obtained through studies of selective adaptation. Eimas and Corbit (1973) reasoned that a feature detector might be fatigued by repeated presentation of an appropriate stimulus. This fatigue would temporarily reduce the responsiveness of the detector to test stimuli. Decision boundaries, which occur at the point of equal detector outputs, would be expected to shift towards the adaptor stimulus. Eimas and Corbit found such adaptation along the VOT continuum. Adaptation occurred not only when the adaptor was a member of the test continuum (i.e., testing /ba/ – /pa/ after adaptation with /ba/ or /pa/), but also when other voiced or voiceless stimuli were used as adaptors (i.e., testing /ba/ – /pa/ after adaptation with /da/ or /ta/). Boundary shifts have also been shown for place of articulation (Cooper, 1974), the fricative-affricate dimension (Cole and Cooper, 1975), and the stop-continuant dimension (Cooper, Ebert, and Cole, 1976; Diehl, 1976) using the adaptation paradigm.

2.1.4 Limitations of Feature Detector Theory

With this broad base of empirical support, feature detector theory quickly superseded motor theory and analysis-by-synthesis theory, becoming the most prominent theory of speech perception. Yet nearly from the time of its first proposal, important empirical evidence has been available which questions the validity of detector theory. Empirical evidence suggests (1) that selective adaptation occurs to acoustic properties of the signal, not to phonetic categories; (2) that phonetic decisions are not simple detections made on the basis of auditory patterns, but instead are complex interpretations of all information available to the listener; and (3) that the ability to make phonetic decisions is not innate, but instead a learned skill. We will review each of these criticisms in turn.

Although the name "feature detector theory" is strongly reminiscent of the physiological analyses of frog vision and audition, the term is misleading. Every speech feature detector theory shares the assumption that detectors are sensitive to abstract linguistic properties of speech. The phonetic feature voiced/voiceless depends not only on the relationship between burst release and voicing onset, but also on the duration of voiced transitions, the intensity of aspiration, first-formant onset frequency, and so on (Diehl, 1981). A feature detector must be responsive to all of the appropriate factors to make a voiced/voiceless determination.

To provide a contrast for feature detectors, we will consider a property detector theory. Property detector theory postulates the existence of specialized detectors in the human auditory system that are sensitive to specific acoustic events. Property detectors are analogous to visual line or edge detectors, and might be responsive to observable events in speech, such as rising or falling formant transition, voice onset time, etc. The distinction between the two types of detectors lies along the dimensions of complexity and specialization. This distinction is not arbitrary: feature detectors cannot be simple property detectors, because it is assumed that feature detectors respond to different speech patterns that do not share simple acoustic properties. Conversely, a theory of speech understanding based on property detectors must include an additional mechanism to perform the

many-to-many mapping between stimulus and percept; the machinery of a property detector is insufficiently powerful to generate this invariance on its own.

Adaptation Studies

Adaptation studies form an important buttress in the empirical support of feature detector theory, because these studies provide the most direct evidence available to discriminate between simple auditory theories and feature detector theory. Early studies claimed adaptation occurred along featural dimensions, not on a simple acoustic basis. The adaptation effect did not generalize perfectly, however. Adaptation was greatly attenuated when adapting stimuli were not drawn from the test continuum. This section will describe several experiments designed to clarify the nature of adaptation.

The most direct contrast is to evaluate detector properties by comparing acoustic and phonetic adaptation. Vowel transitions, which cue place of articulation, are strongly conditioned on the associated vowel. It is possible, therefore, to create two CV syllables which share the same initial phonemes, but which have acoustically distinct patterns, such as /bæ/ and /bi/. Adaptation effects for /bæ/ – /dæ/ judgements were compared by Cooper (1974) after adapting with either /bæ/ (second formant transition between 900 Hz and 1.2 KHz) or /bi/ (second formant transition between 2 and 2.8 KHz). A property detector theory would predict no adaptation effects for /bi/ on /bæ/ – /dæ/ judgements due to the differences in vowel transitions for /bi/ and /bæ/, while a feature detector would predict equal adaptation effects. Boundary shifts were significantly different from zero for both /bæ/ and /bi/ adaptors. However, the magnitude of the adaptation effect for /bi/ was less than half of the adaptation for /bæ/, which was a statistically significant difference. Although this experiment provides partial support for both phonetic and feature detector theories, a final interpretation of the results is problematic, because the /bi/ adaptor was a token of a real utterance, and so may have inadvertently shared auditory properties with the synthetic /bæ/ stimulus.

Ades (1974) provided a replication and extension of adaptation effects, again using place of articulation as the relevant dimension. In one experiment, two /be/ and three /de/ adapting stimuli were prepared, and boundary shifts were measured along the /bæ/ – /dæ/ continuum. /de/ stimuli were varied in their acoustic similarity to /dæ/. A feature detector theory would predict equal adapting effects for /dæ/ and each of the three /de/ stimuli, while auditory theory predicts that two of the /de/ stimuli, acoustically similar to /bæ/, will have an opposite boundary shift. Again this experiment failed to clearly support either theory. Although all of the /de/ stimuli had an adapting effect similar to /dæ/, the magnitude of the effect was only about 40% of the size of the /dæ/ adaptation, and varied according to the acoustic similarity of each /de/ stimulus to /dæ/.

In another experiment, Ades created two continua, from /bæ/ to /dæ/ and from /æb/ to /æd/. Each token along the /æb/ to /æd/ continuum is the "mirror image" of a token along the /bæ/ to /dæ/ continuum. If phonetic feature detectors are present, we would expect /dæ/ and /æd/ adapting stimuli to have equal effects on /bæ/ – /dæ/ and /æb/ – /æd/ judgements. Because of the mirror image nature of the stimuli, rising formant transitions in the /bæ/ stimuli are realized as falling transitions in /æb/, so an auditory property detector theory would predict no adaptation effects. Within-continuum adaptation had the expected effects: mean differential boundary shifts were 3.67 for /bæ/ – /dæ/ and 3.14 for /æb/ – /æd/. Between continuum stimuli had no adaptation effects: mean differential shifts were 0.08 for /bæ/ – /dæ/ and – 0.01 for /æb/ – /æd/.

These experiments present problematic results for both feature detection theories and property detector theories. Feature detection theories cannot explain why the adaptation was not identical across stimuli having identical phonetic interpretations but distinct acoustic properties, and property detector theories cannot readily explain why any adaptation occurred at all for the acoustically distinct stimuli. Without a well-specified theory of acoustic properties, however, it is difficult to rule out a purely auditory explanation of these experiments. Pisoni (1980) has suggested that complex auditory interactions might be present in the stimuli, that lead to the observed adaptation results.

In order to avoid the auditory-interaction problem, Sawusch and Jusczyk (1981) noted that a labial stop consonant which follows an initial /s/ is interpreted by English listeners as a voiceless stop. For example, [sbit] is interpreted by most listeners as /spit/, even though the same syllable with the initial /s/ spliced out is heard as /bit/ (Lotz, et al., 1960). Adapting stimuli of [ba], [sba], [p^ha] and [sp^ha] were presented to listeners. The respective boundary shifts for the four syllables are: −0.35, −0.34, +0.14, and +0.34, where a positive value implies a shift towards the /p/ end of the continuum. Each of these four shifts was found to be significant. Of major importance is the boundary shift for [sba]. Because this stimulus was interpreted by listeners as /spa/, feature detector theory would predict a positive shift. The negative shift which occurred for this stimulus is instead in the direction of its acoustic properties.

Diehl (1976) reasoned that if adaptation was a result of auditory, as opposed to phonetic, properties of the stimuli, it might be possible to use non-speech adaptors to shift the boundary for speech stimuli. A phonetic theory would appear to limit adaptation to only those cases where phonemes are detected. Accordingly, "pluck" and "bow" adapting stimuli were presented to subjects who then classified stimuli from the /ba/ – /wa/ continuum. The pluck stimulus, which shares an initial abrupt amplitude change with /ba/, was found to be an effective adaptor, although the bow stimulus had no adaptation effect. Samuel and Newport (1979) extended this finding to a second continuum, /sha/ – /cha/, and showed adaptation only occurs for stimuli which share acoustic properties with the test stimuli. Further experimentation revealed that adaptation could occur even when the non-speech adaptors and speech test stimuli did not share the same spectral characteristics. They concluded adaptation was a result of shared abstract auditory properties, not necessarily a function of simple spectral properties.

Factors affecting phonetic judgements

Although adaptation experiments were long felt to support feature detector theory, the above experiments appear to be strong evidence against this theory. A second major criticism concerns the implicit structure of the feature detectors themselves. Feature detectors are interesting theoretical constructs because they provide a simple mechanism capable of performing the complex identification of phonetic segments. If the processes underlying feature detection can be shown to be highly complex, the simple, presumably neurological basis of feature detection would be suspect. Empirical studies reveal that human listeners are sensitive to a large number of factors in making phonetic judgements, implying feature detectors with extraordinary neurological sensitivities. Diehl (1981) compiled a list of local acoustic factors affecting voicing judgeme' :

> Consider, for example, that judgements about the voicing category of initial stops depend on all of the following acoustic variables: VOT, duration of voiced formant transitions, first-formant onset frequency, onset frequencies and directions of second- and third-formant transitions, spectral characteristics of the following vowel, duration of the following vowel, duration of aspiration, intensity of aspiration, and direction of fundamental

frequency change at voicing onset. Lisker has provided an even longer list of acoustic variables that affect voicing judgements for medial stops.

Compiling such lists of variables cannot directly be taken as evidence against feature detector theory. This evidence simply underscores the complexity of human phonetic judgements, and implies feature detectors must be sensitive to a number of factors during the detection process. Some recent experiments reveal that phonetic decisions are not simply made on the basis of acoustic information, but depend on non-acoustic factors as well. Ganong (1980), for example, identified VOT boundaries with word-nonword pairs. He found a boundary shift for VOT contingent upon the lexical status of particular tokens. In one instance, the boundary between "dash" and "tash" was found to be closer to "tash" than the boundary between "dask" and "task", even though the relevant contrast between /t/ and /d/ was identical for the two pairs. Subjects had a tendency to "correct" the stimulus boundary so that they heard ambiguous stimuli as real words. In this case, feature detectors are presumably sensitive to word-nonword distinctions available in the subjects' lexicon.

McGurk and MacDonald (1976) found another important factor which influenced phonetic judgements: visual stimuli. In this experiment, video tapes were created which included conflicting visual and auditory stimuli. In one instance, an utterance /baba/ was dubbed onto lip movements for the utterance /gaga/. 98% of all adult subjects reported hearing the sequence /dada/. MacDonald and McGurk (1978) replicated and extended their original study for six plosive stops and the nasals /m/ and /n/. They found that whenever the place of articulation conflicted between the auditory and visual stimuli, an auditory illusion occurred. Stimuli which share the same place of articulation, but which had a different manner, were veridically perceived. These results were interpreted according to a **manner-place** hypothesis, where information about the manner of articulation is obtained from the auditory stimulus, and information about the place of articulation is obtained from the visual stimulus. Adaptation experiments using such stimuli again found boundary shifts consistent with the acoustic properties of the signal, not the phonetic percept (Roberts and Summerfield, 1981).

These studies reveal that phonetic decisions are not simple detections made on the basis of auditory patterns, but instead are complex interpretations of many acoustic, visual, and lexical properties of the signal. Postulating feature detectors that are sensitive to a broad range of acoustic properties seems unpalatable at best. Any theory which would allow a feature detector to evaluate non-acoustic information, such as lexical or visual information, seems untenable.

Developmental studies of speech perception

Finally, demonstrations of categorical perception in infants as young as one month led to the hypothesis that feature detectors are innate mechanisms of the human perceptual system (Eimas, et al., 1971; Eimas, 1974, 1975). Learning was excluded by the theory, and so, by corollary, were developmental changes in phonetic understanding and differences in adult performance. Fortunately, this theoretical stance did not completely prevent the collection of the relevant data. This section will review studies investigating developmental differences.

Curiously, the earliest demonstrations of a strong developmental component are not acquisitional studies, but instead studies that investigate the outcome of language development, as realized in adult performance. In a classic study, Lisker and Abramson (1970) observed differences in perceptual classification of synthetic stimuli varying along a VOT continuum. Thai makes a distinction between pre-voiced, voiced, and voiceless stops. Native speakers of Thai separated labial pre-voiced from voiced stimuli at 40 msec, while the boundary between voiced and unvoiced occurred at − 20

msec. Spanish and English speakers typically divide the continuum into only voiced and voiceless stops, but the boundaries are quite different between the two languages. Spanish speaking adults divide the continuum at 17 msec, while native English speakers have a boundary at 25 msec.

Williams (1977, 1979) explored the voiced/voiceless distinction in bilingual speakers of Spanish and English. In both languages, stops with VOT of 25 msec or greater are labeled unvoiced, and stops with VOT's of less than − 4 msec are voiced. Of particular interest is how bilingual subjects respond to stops with intermediate VOT's between − 4 and 25 msec. In the first study (Williams 1977), five out of eight bilingual subjects were found to have an English-like discrimination pattern. The remaining three subjects exhibited a more Spanish-like discrimination pattern, although they had a broader discrimination function for intermediate exemplars. Since all subjects were known to be highly fluent in both languages, Williams concluded that these subjects had learned to pay more attention to secondary cues in making voiced/voiceless distinctions. The impoverished synthetic stimuli did not preserve such additional information, forcing subjects to make judgements on the basis of VOT alone. In a second study, Williams (1979) found developmental evidence for the boundary shift in Puerto Rican children who had moved to the Boston area and were learning English as a second language. For all subjects, the VOT boundary was typically sharp. Increasing exposure to English resulted in a shift of the boundary toward the monolingual English 25 msec value.

Miyawaki et al. (1975) investigated the /ra/ − /la/ continuum with English and Japanese speakers. Typical categorical performance was observed for English subjects, while discrimination of items for Japanese was little better than chance across the entire continuum. Eimas (1975) found infants exhibited categorical performance on this same continuum, implicating a habituation or loss of sensitivity in Japanese adults, rather than a learned sensitivity by English speakers.

Other studies, investigating the process of phonetic development, have also found clear evidence of learning. Simon and Fourchin (1978) tested French and English children from age 2 to 14 with synthetic stimuli that varied along the VOT continuum from coat to goat (English) or from Toto to dodo (French). Correlated with the changes in VOT were changes in F1 onset transitions, providing a double-cue for voiced-unvoiced distinctions. Both groups of subjects showed three distinct developmental stages of recognition for these stimuli. In the first stage, children from both countries responded correctly to the endpoint stimuli, but exhibited random classification performance on intermediate tokens. In the second stage, subjects' responses paralleled the linear nature of the continuum: the probability of giving a particular word response to a token varied inversely with the distance to the endpoint stimuli. In the third stage, subjects showed adult-like categorical performance for items along the continuum.

Although English and French children showed similar developmental trends, the ages for each stage varied markedly across the two languages. English children showed endpoint response patterns at age two, linear responses at age three, and categorical performance from age four on. French subjects maintained the endpoint pattern from two until age five, showed progressive labeling from five to age nine, and showed the adult pattern after age nine. Simon and Fourchin attributed this difference in development to differences in cue salience in the two languages. In English, F1 onset is an important cue for making voiced-unvoiced distinctions, while this cue is not normally present in spoken French. In addition, pre-voicing is common for French voiced consonants, so that the range of stimuli tested was closer to natural English pronunciation than French. These factors probably increased the level of difficulty for the French children, and so these children appear to take longer to develop adult levels of competence.

Templin (1957) found a similar developmental trend in 480 three- to eight- year-old children. Although no phonological breakdown was provided for age groups, overall test scores on a speech-sound discrimination test improved steadily as a function of age. Even eight-year-old subjects were not perfect; they averaged 94% correct on a task where normal adults are errorless.

Although infant studies tended to emphasize the innate nature of feature detectors, the developmental studies we have reviewed suggest a quite different picture. Although it is possible to observe a sharp discrimination function when infants are presented with synthetic stimuli, we cannot conclude that the same infants possess all of the subtle abilities of adults to discriminate between natural stimuli. The studies mentioned here instead have found that children do not possess full adult phonological competencies. Furthermore, adults do not all have equivalent competencies; instead they may lose sensitivities to contrasts that do not signal phonological differences in their language. Bilingual subjects learn to rely on a multiplicity of cues in making phonetic judgements for different languages. Collectively, this evidence contradicts a notion of innate feature detectors, instead implicating learning mechanisms in making phonetic judgements.

I have summarized three areas of research that appear to contradict feature detection theory: adaptation studies, which demonstrate adaptation effects to be primarily acoustic in nature; perceptual studies demonstrating the complexity of information utilized in making phonetic judgements; and developmental studies showing differences in performance for adults and evidence of learning in children. This evidence is particularly compelling because it utilizes the same experimental paradigms originally used to support feature detector theory. As a whole, the evidence strongly suggests that phonetic decisions are not made by innate feature detectors. The next chapter will outline an alternative model which appears to be more consistent with the existing literature.

3. Property integration theory

A viable alternative to a feature-detector theory would appear to be a property detection/central integration theory. The **property integration** theory postulates that peripheral property detectors, attuned to selective acoustic events, provide information about an utterance to a central decision unit. Instead of a multi-stage model, where decisions are made in a hierarchical fashion, the property integration theory postulates the existence of only a single decision unit. This unit, incorporating syntactic, semantic, and pragmatic information as well as acoustic information, makes the actual phonetic judgement. The overall approach of property integration theory owes a larger debt to the Hearsay speech recognition system (Erman, et al., 1980) than to existing theories of speech perception.

In support of the property integration theory, we have already reviewed evidence suggesting the auditory system is sensitive not to phonological features, but to simple acoustic properties. However, it is not intuitively obvious that the property integration theory can account for an important empirical paradox: if speech properties can be detected so readily by infants, why does it require eight or more years to learn to make phonetic judgements? It would seem that a simple and trivial learning process could identify the relationship between acoustic properties and phonological features. Three- year-olds are quite competent at understanding speech, so it is unclear why additional improvement should occur, or even be necessary.

To reconcile this discrepancy, we must distinguish between *detecting properties* and *making phonetic judgements*. Property detection is an analytical function that can be performed solely on the basis of acoustic information. Phonetic judgements are decision functions, requiring the listener to integrate several distinctive features, and make an appropriate classification of the sound. Such judgements depend not only on the acoustic properties of the signal, but also depend on the set of pre-defined categories used to classify the signal. Pisoni et al. (1982) provided an elegant demonstration of this phenomenon for the VOT continuum. They first asked subjects to make /ba/ – /pa/ judgements, and obtained a standard dichotomous decision function. Subjects were then given two repetitions of a synthetic pre-voiced /ba/ ([^{m}ba]), and asked to classify the same items into /^{m}ba/, /ba/, or /pa/. Despite the absence of highly pre-voiced stops in initial position in English, all four subjects were able to consistently discriminate examples of the three syllables from one another. A further difference between analytic and decision functions is that the information necessary to make a phonetic decision is not always simultaneously available in the waveform, but requires an extraction and integration process that extends over a fairly large segment of the input, probably at least a syllable, whereas detection functions are presumed to be very localized processes with little temporal integration.

Curiously, nearly all of the theoretical accounts of speech recognition have made the distinction between property detection and phonetic judgements, yet the same accounts have not recognized that innate detection does not imply an innate ability to make phonetic decisions. In part, this confusion occurred because researchers have failed to discriminate between *sufficient* and *necessary* cues for phonetic identification. The ability to create synthetic speech stimuli resulted in a flurry of studies demonstrating the efficacy of manipulating a single cue, such as VOT, in determining phonetic decisions. There is reason to believe that human listeners attend to multiple cues in making phonetic decisions, the implication being that judgements represent a synthesis of many cues, all of which are attended to by listeners. In cases where information relevant to a particular cue is unavailable, such as trying to determine VOT in whispered speech, human listeners may be capable of attending to other information to make phonetic judgements. In many situations, multiple cues may

be present and provide conflicting or reinforcing evidence about the identity of a sound. Fitch et al. (1080) exploited the compound nature of perception to create a set of stimuli containing two independent cues for stop-consonant manner: silence duration and formant transition. These cues were found to engage in a **trading relation** where one cue could, within limits, substitute for another.

3.1 Stevens' cue detection model

Stevens (1972) suggested an elegant model which can resolve the inconsistency between innate detectors and learning phenomena in phoneme identification. Consonants were assumed to have **primary** and **secondary** attributes, or cues. Primary cues are detected by a few general property detectors in the auditory system, while secondary cues are more subtle attributes of a sound. Children presumably learn to discriminate between consonants based on their primary attributes in stressed initial positions. The co-occurrence of secondary cues allows an association between primary and secondary cues. This association, in turn, enables the child to learn to discriminate between unstressed consonants, where the primary cue may be absent or weak. In the original paper, Stevens (1972) did not specifically address the question as to whether or not primary cues are always present for stressed consonants. Later accounts (Stevens & Blumstein, 1978, 1981) specifically assumed that primary cues might be absent or disguised by noise; in such cases, adult listeners could rely on secondary cues to make phonetic judgements.

Unfortunately, the Stevens model has not been well-tested. The theory seems to have been largely ignored in other theoretical and empirical work. A simple test of this model might be to investigate whether or not the trading relation found by Fitch et al. has any developmental component. Another approach might be to determine whether the reliance on a cue or cues can be affected by the noise background of a particular stimulus. If cues can be selectively masked by appropriate information, developmental trends could be readily explored. Hopefully, experiments such as these will be performed in the near future, to help clarify the role of learning in speech perception.

3.2 A learning model of word recognition

We have already described property integration theory as an alternative to feature-based theories. This section will introduce a specific learning model based loosely within the property integration framework. The **learning theory of word recognition** concentrates on learning processes that might occur as a human acquires a new language. The model does not incorporate specific perceptual processes to derive an acoustic representation from an utterance; our primitive understanding of the relevant processes makes any such approach premature. Instead, the significance of the model is that it tries to demonstrate how a phonetic structure might be discovered within a particular language. This discovery is a non-trivial problem because phonetic structure is never taught directly, but must be inferred through experience.

The fundamental premise of the learning model is that perceptual units are acquired through experience and are used in the interpretation of speech. In addition, the theory makes an important distinction between the detection of acoustic events and their interpretation. This allows the learning theory to model, in principle, non-acoustic influences on phonetic decisions. I will next discuss the perceptual units used in the theory. Following sections will describe the acquisition of these units and how they are interpreted.

3.2.1 The units of perception

The theories we reviewed have incorporated phonetic segments as intermediate units between the original sensory experience and the identification of words. There are important theoretical reasons for including some type of intermediate units: an overall theory of word recognition is simpler if it permits independent decisions to be made. Word identification processes, for example, do not have to correct for the variability of speech or recognize coarticulatory phenomena. In rejecting the perceptual detection of phonetic segments, we raise the question as to what, if any, intermediate levels are present in the word identification process. Of course, we should not preclude phonetic segments as intermediate units; instead we simply require a non-perceptual explanation of their identification. Unfortunately, there is little direct evidence about the nature of units detected by the perceptual system, or about any other possible intermediate units. One reason to reject phonetic segments as basic units is the extraordinary discrepancy between the apparent rate of comprehension of phonetic segments and the rate at which ordinary sounds can be ordered: speech can be understood at a rate of approximately 50 phonemes per second (Foulke and Sticht, 1969). Warren, Obusek, Farmer, and Warren (1969) found that a sequence of noises could not be ordered if items were presented faster than 1.5 elements per second. If phonemes were independently processed, we would expect this to be the upper limit on the rate of speech, because order is important in discriminating between words. The discrepancy may be resolved by assuming that a perceptual unit of speech simultaneously encodes several phonetic segments. Liberman, et al. (1967) had something like this in mind when they suggested speech is a type of code that represents phonetic segments in a compact form.

Another important reason for considering other perceptual units is simply the nature of the speech waveform itself: phonetic segments do not occur as independent isolated events in a waveform. Instead, a continuous section of the waveform will simultaneously convey information about a number of phonetic segments. These facts suggest that a larger unit of speech, such as a syllable, may be the fundamental units of recognition. Martin and Bunnell (1982) provided evidence for larger units when they showed human listeners are sensitive to anticipatory coarticulation phenomena in making phonetic judgements. Physical constraints on human articulatory gestures encourage speakers to begin rounding their lips two or three segments before it is actually necessary. Listeners can detect the anticipatory rounding, and use this information to speed recognition decisions. In short, not all instances of each phonetic segment are treated alike. Listeners can discriminate between different instances of the same segment, and utilize these differences to speed up the recognition process.

Collectively, these data suggest that perceptual units might be larger than phonemes. A few investigators have proposed that syllables are potential units. A major objection to syllable-based recognition is the large number of units needed. In English, for example, approximately 40,000 triphones can be identified (Wickelgren, 1969). Also, syllables are not free from coarticulatory phenomena; they still retain some of the variability present for smaller units. The same variability occurs for other intermediate units, such as diphones and demisyllables, that have also been suggested as possible units of recognition.

In the absence of direct psychological evidence about basic units, I feel it is unwise to make strong commitments to a particular type of unit. For the moment, I will assume that basic elements are smaller than words, and that they can be isolated by the perceptual system. Because of the heavy computational demands of speech understanding, both for human listeners and for machine systems, it is desirable to develop a small, efficient set of descriptors in which contextual interactions in interpretation are minimized. We know, from analyses of human speech production mechanisms, that

vowels and turbulence segments (fricatives, affricates, etc.) are produced in different resonant cavities of the vocal tract. Accordingly, each of these segments exhibits a great deal of independence in production. However, within a single vowel or turbulence segment, important coarticulatory effects are present. I have defined a new unit of speech, the **continuous acoustic segment**, to be a segment that has continuous spectral properties. There are three basic types of continuous segments: periodic, aperiodic, and silence segments. Segments of this type were identified by an expert spectrogram reader, VZ, who was correctly able to identify 97% of all phonetic segments from a spectrogram (Cole, Rudnicky, Zue, and Reddy, 1980)[1]. Further details about identification of continuous acoustic segments are provided in the chapter describing NEXUS. The learning theory simply assumes continuous acoustic segments to be the basic perceptual units of speech.

3.2.2 Developing a descriptor set

A recognition task requires unknown stimuli to be matched against a previously encoded set of descriptors. A basic premise of the learning theory is that the set of descriptors is acquired through experience with language. This acquisition process would be trivial if we were given direct practice on basic perceptual units. Infants do not receive much practice on minimal phonetic pairs, and do not appear to need this practice. An assumption underlying the learning theory is that language is taught through experience with words, not through repetitions of subword units such as syllables or phones. A correlate of this assumption is that all language learners must infer an appropriate set of descriptors. This section describes a set of learning mechanisms that can identify perceptual units through experiences with whole words.

The inference process to construct a set of descriptors is based on an analysis of the similarities and differences between words. The learner must develop an understanding of both meaningful and irrelevant variation in words. A prerequisite for these processes is the ability to recognize words, so that appropriate analyses may be performed. This leads to a "chicken and egg" problem: If the system can recognize words, why bother to learn about sub-word units? Conversely, how does a system recognize words without having an understanding of their component parts? Apparently humans learn about the fundamental units in a language without needing to do so.

One resolution of this problem is to propose that learning processes operate not only to acquire speech units but also to refine them. The model suggests that words are initially encoded and recognized in a relatively unanalyzed form. As the vocabulary increases in size, it becomes more parsimonious for a listener to infer internal structure. Another consequence of vocabulary expansion is that greater opportunities exist for minimal pairs to occur. The similarity between such items encourages appropriate discrimination. Finally, contextual dependencies are discovered as words appear in varied contexts. This would result in the identification of coarticulation effects that are independent from individual words. It is assumed that all compensatory learning of this sort, operating on sub-word units, would automatically generalize to all words that contained the appropriate units.

Let us consider briefly some of the specifics of the inference process. The general goal is to acquire a set of descriptors that will allow a listener to recognize all utterances in a language. Any set acquisition task can be analyzed into **set modification actions** and the conditions under which

---

[1] VZ further discriminated two types of periodic segments: vowel segments and nasal segments. Due to the difficulty of isolating nasal segments from vowel segments, no such distinction was attempted in NEXUS.

each action is applied. There are only four primitive set modification actions: adding new descriptors, modifying them, merging two or more descriptors, and eliminating items from the set. The flexibility and elegance in learning systems of this sort derive not from the actions taken to the description set, but instead lie in the conditions under which each action is taken. A statistical clustering program, for example, might identify the basic descriptors from an analysis of a very large sample of speech. Computationally, the requirement that the total sample be available during the clustering analysis would appear to put all such techniques beyond the capabilities of human listeners. A second class of heuristic techniques exhibits incremental learning of descriptors, which seems a more feasible learning pattern. The NEXUS learning model of speech perception uses heuristics to continuously guide an incremental learning process.

One fundamental assumption underlies all heuristics used in the NEXUS system: no two words can share the same acoustic sequence. Homophones form a small class of words that violate this assumption, but for the majority of words it is correct. We will also assume that when a recognition error occurs the learner is told that an error occurred and given feedback about the identity of the correct word[2]. This information allows the learner to do a **similarity analysis** between the acoustic sequences of the input, the incorrectly recognized word, and an instance of the correct word. From the similarity analysis, the listener can determine an appropriate course of action. The following chapter elaborates on the similarity analysis process, and provides details on the actions NEXUS takes in each case.

3.2.3 The interpretation of events

Due to the large number of factors that influence phonetic judgements, the learning model makes a distinction between the detection of acoustic events and their interpretation. This separation appears to be necessary for non-acoustic information to influence perceptual judgements. Such a separation also permits, but does not require, coarticulatory influences between sequential acoustic events. In order to achieve this separation the learning system must not only develop a set of acoustic descriptors, but must also learn an interpretation function which can assign meaning to detected events. One important implication of this model is that an acoustic descriptor does not have a direct phonetic interpretation. Only through the application of the interpretation function can a phonetic label be assigned to a description.

The interpretation function is also assumed to be acquired through simple generalization and discrimination processes. The learner presumably has information about the context in which a sound occurred. This context is preserved as a meaningful datum in the interpretation of the sound. If the same sound is identified under somewhat different contexts, the interpretation function for that sound is generalized. Sounds that are encountered only in a small number of contexts never achieve this generalization.

3.3 Summary

This chapter introduced property integration theory as an alternative to existing theories of speech perception. Property integration differs from other theories in assuming that learning is

---

[2] This assumption is unrealistic to a certain extent. Human listeners do not always receive feedback about errors. Sensitivity to syntactic, semantic, and pragmatic information may allow a listener to independently detect and correct errors even without feedback from the talker. The utilization of such factors is far beyond the scope of the present project.

necessary for a system to recognize the fundamental sound patterns that make up a language. The learning theory of speech perception was introduced as a model of how learning might occur. The following chapter elaborates upon a computer implementation of the learning theory, NEXUS, while remaining chapters analyze the performance of the model.

4. A description of the NEXUS speech recognition system

nex' us[1], 1 neks' us; 2 neks' us, n. [NEX' US, pl.] 1. A bond or tie between the several members of a group.

> Nobody ever has discovered, in the external universe, merely by observation through the senses, the *nexus* which so binds two events together, that the production of one of them must be followed by the occurrence of the other.
>
> F. Bowen, *Modern Philos.* p. 280 [s. 1877]

[L., < nexus, pp. of necto, tie.]

Funk and Wagnalls New Standard Dictionary

In order to evaluate some of the claims of the learning theory of speech recognition, a computer model was developed. The NEXUS program, a complete word recognition system, is not intended as a literal model of a human listener understanding words. One important difference between a human listener and NEXUS is that human listeners utilize all available knowledge, including higher level knowledge, in the word identification process, while NEXUS is not provided with any semantic or syntactic knowledge of language; word identification is made solely on the basis of acoustic information. Furthermore, NEXUS is an incomplete realization of the learning theory of speech perception outlined in Chapter Three. In order to construct an efficient working system, the acoustic representation and matching functions have been simplified. Discussion of important restrictions will be included in following sections. Although NEXUS is an incomplete and slightly compromised model, the system acts as a sufficiency proof of the fundamental learning assumption: perceptual units can be identified through experience with speech, then utilized to recognize and identify later utterances. These units can be interpreted in a context-dependent manner similar to that of human performance.

NEXUS can be thought of as three semi-autonomous processing modules that share a common data structure called the **speech network**. The **encoding module** performs signal processing functions on the input, the **recognition module** compares the input against the speech network to match and identify speech, while the **learning module** implements set acquisition processes to create and modify the speech network. Although encoding, recognition, and learning functions occur in separate modules, these modes of activity are tightly interleaved. NEXUS operates in an encode-recognize-learn cycle, so that on any given trial the system encodes and recognizes an utterance, then may modify the speech network on the basis of its experience by adding new words to its vocabulary, adding new perceptual units to help discriminate between confusable items, or simplifying the network to reduce processing time. Intertwining recognition and learning functions produces an incremental pattern of learning in NEXUS, an important similarity to human behavior. In contrast, many other speech learning systems (Itakura, 1975; Rabiner, 1978; Rabiner et al., 1979) incorporate training processes whose data requirements prohibit incremental learning, forcing the training phase to be completely separated from later recognition performance. The next section will describe the basic speech network data structure. Following sections provide greater detail on encoding, recognition and learning modules.

4.1 The NEXUS speech network

The speech network is a hierarchical structure that embodies all of NEXUS' knowledge about speech. The top level of the network consists of a list of word descriptions ($W_{0,0}$ $W_{0,1}$ $W_{1,0}$ $W_{2,0}$ ... $W_{m,n}$) representing the current vocabulary of the system. Each $W_{i,j}$ represents the j^{th} instance of word i. By allowing each word to have multiple descriptions, NEXUS can quickly and simply encode different pronunciations of words.

The second level in the network is the set of word descriptions. Each description is stored as a structure containing a word label $WL_{i,j}$ and an associated list of pointers (p_0 p_1 p_2 ... p_n) where each p_i points to a continuous acoustic segment. The list is interpreted from left-to-right as a sequence of segments concatenated together. Figure 4-1 shows a hypothetical section of a speech network where the spoken letter "P" is represented by four different word descriptions: $[p]_1$ [i], $[p]_1$ [h] [i], $[p]_1$ [i] [h], and $[p]_2$ [i]. In this figure $[p]_1$ and $[p]_2$ are presumed to represent distinct articulations of the phoneme /p/.

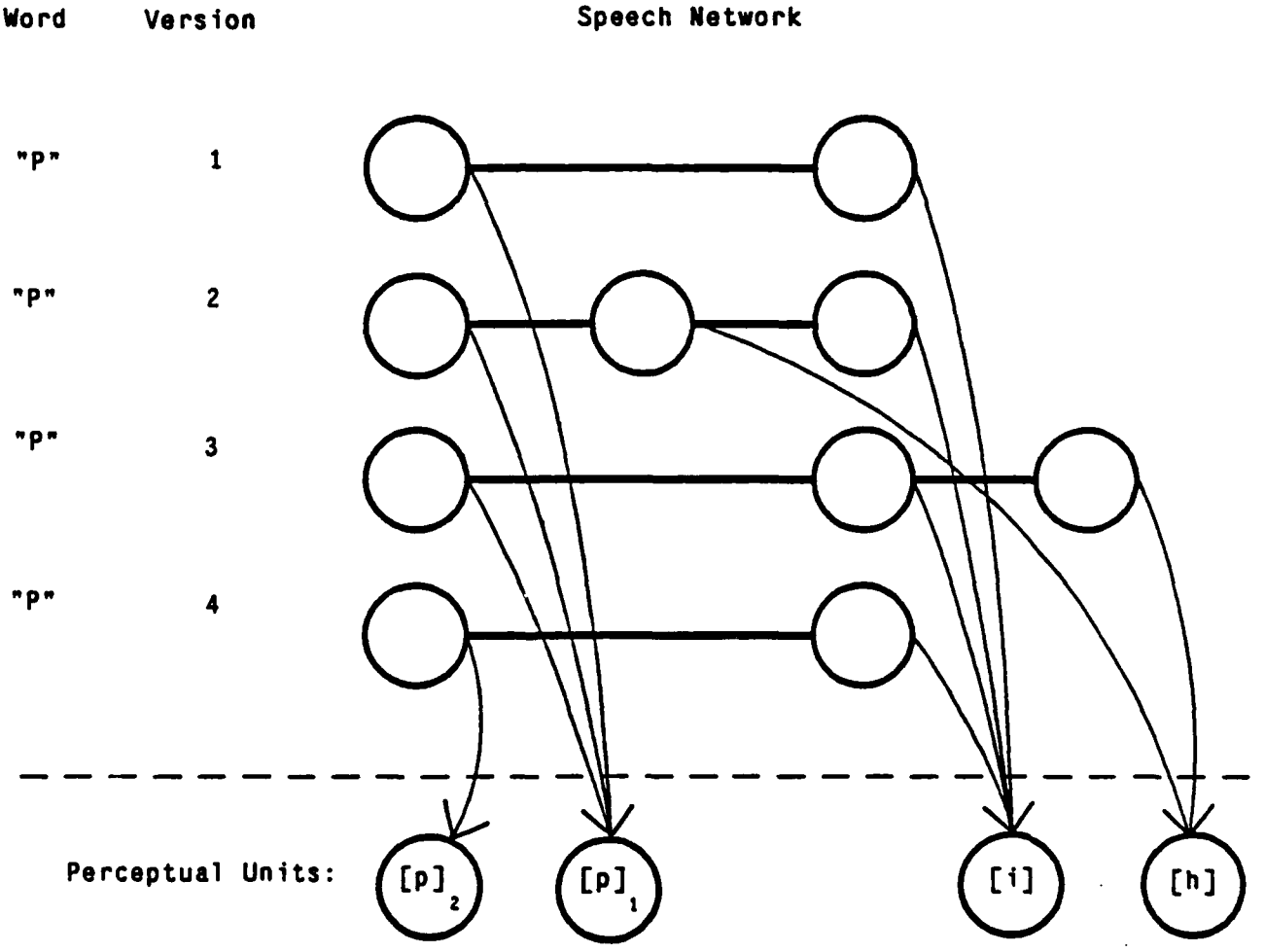


Figure 4-1: Hypothetical NEXUS Speech Network

Continuous acoustic segments form the final level of speech representation. Each segment forms a pattern, or template, of a primitive acoustic event. In the current representation, this pattern consists of a series of temporally ordered short-term spectra. The pattern is designed to simulate the

frequency coding present in a human "neural spectrogram". The signal processing section describes continuous acoustic patterns in more detail.

The separation of word descriptions and continuous acoustic segments allows the network to represent speech in a very compact, efficient form. Several different word descriptions may point to the same continuous acoustic sequence, which need be defined only once. In contrast, many current speech recognition systems store independent acoustic patterns for each instance of every word in the system's vocabulary (Itakura, 1975; Rabiner et al., 1978; Sakoe, 1979). The NEXUS representation also permits a simple, powerful generalization mechanism. Under certain circumstances, NEXUS decides that a particular acoustic segment should be modified. Making the appropriate changes simultaneously updates all words referencing that segment.

The next section describes how acoustic events are encoded and initially represented in NEXUS. Following sections will show how the speech network is used to recognize incoming speech, and how the network is constructed and modified by NEXUS learning mechanisms.

4.2 Encoding processes

The first step in every NEXUS processing cycle is to digitize and encode a real acoustic input. NEXUS performs two major encoding functions: **signal processing** and **segmentation**. Translation into the frequency domain and intensity adjustment are signal processing functions, while identification of segment boundaries and begin-end detection are the primary segmentation functions.

4.2.1 Signal processing

Speech is initially encoded in NEXUS by digitizing the output of a microphone. Each input sample represents the amplitude of the microphone output at a given time. This time/amplitude representation is quite different from the representation derived by the human ear, which simultaneously analyzes the intensity of sounds at a number of different frequencies. The signal processing module is responsible for translating speech into the **frequency domain**. To understand the translation process, it is important to note that a single time-domain sample does not convey information about frequency. This information can only be obtained by looking at a sequence of samples. Computationally, the transformation is accomplished by means of a fast Fourier analysis (fft). The fft processes a short interval of the waveform to produce a vector, or **short-term spectrum**, representing the intensity of a number of frequencies during the interval. One important property of the fft algorithm is that changes in frequency within the sample interval are ignored. By choosing an appropriately small input interval, such changes are minimized. Specifically, the sample interval in NEXUS is 20 msec. This interval is multiplied by a hamming window, leaving an effective interval length of 12 msec. The output of the process is a short-term spectral description of 128 frequency coefficients. Each coefficient represents the intensity in a 62.5 Hz range of frequencies, so the entire vector covers the range of frequencies from 0 to 8000 Hz.

The initial 128 coefficients are linearly mapped onto a 26 point scale in order to reduce the complexity of the representation. The amplitude value of each coefficient is mapped onto a log-dB scale to simulate amplitude-coding of the human ear. Global amplitude variation, caused by changes in speaking volume, microphone distance, and so on, is compensated for by calculating the first derivative of the spectral contour. Each of the resulting 25 coefficients is compressed into an 8-bit

representation, and so can assume any whole value between 0 and 256. Collectively, these coefficients represent the short-term spectra, and are referred to as a *frame* of data. One frame is calculated from the original waveform every 3 ms. The output of the signal processing stage is an array of short-term spectra from the beginning of the utterance until its end[3].

One important limitation of the current representation is that it does not include any properties in the description of speech events. Although this is not psychologically accurate, robust detection of acoustic properties is a difficult task which has not yet been solved. Indeed, little is known about the nature of properties detected by the human listener, other than information about general timing of events such as onset-differences useful in analyzing VOT. Therefore, a simple representation was chosen which resembles, in part, the acoustic information detected by the cochlea of the ear. There are three important differences between a "neural spectrogram" and the spectrogram represented in NEXUS. NEXUS represents frequencies on a linear frequency scale, implying a uniform ability to discriminate between adjacent frequencies across the scale. The ability of the human ear to discriminate between frequencies is a near-logarithmic function of the frequency, and has been represented by the **bark scale** (Zwicker, 1961). Next, the ability of NEXUS to resolve events in time is constant for all frequencies. This limitation is necessary to take advantage of the efficiencies of the fast Fourier transform algorithm. However, human listeners can resolve temporal events more accurately at high frequencies than at lower frequencies. This ability helps to localize onset bursts with greater accuracy than NEXUS. Finally, the human ear has better absolute frequency discrimination than the 25 point spectrum included in NEXUS. Although these differences may significantly affect performance, they were felt to be necessary compromises for an efficient computer implementation of NEXUS.

4.2.2 Segmentation

NEXUS recognizes words by matching sequences of perceptual units against the input. Because NEXUS does not incorporate any innate perceptual units, it must identify these units through an analysis of words. The first step in this process, described in the current section, is to identify segment boundaries corresponding to the appropriate sequence of continuous acoustic segments. The learning mechanisms, described in a later section, can then add continuous acoustic segments into the network as new perceptual units.

Boundary detection

The segmenter is designed to isolate four types of segments: silence segments, turbulence segments (including fricatives, affricates, etc.), voiced segments, and unknown segments. Silence segments are regions with little or no acoustic energy present. Turbulence segments are produced by a constriction in the vocal tract leading to a turbulent air flow. Spectrally, these segments are characterized by a predominance of quasi-random high-frequency energy. Voiced segments, identified on a spectrogram by a low frequency formant structure and by the presence of glottal-pulse striations, occur whenever the vocal cavity is being excited by the vocal folds. Unknown segments usually represent transition segments that occur between clear acoustic events. Most unknown segments represent a discontinuous vowel offset; a few are due to weak turbulence segments.

The segmentation routine does not work on a spectral representation of the data, but instead

---

[3]This array may be thought of as a two-dimensional matrix, where frequency is represented on the vertical dimension and time on the horizontal axis. In this form, the matrix is similar to a "digital spectrogram".

generates a set of parameters from the original time-domain waveform. The parameters consist of zero crossing counts of the original waveform and of a copy of the waveform low-pass filtered at 1000 Hz. Zero crossing counts are more useful than amplitude information or spectral shapes because many fricatives contain very little acoustic energy, yet still have high zero crossing rates. Zero crossings therefore provide a more reliable means of discriminating turbulence segments from silences. Low-pass zero crossing counts are similarly useful for detecting vowel decays, when the overall amplitude has returned nearly to baseline.

In order to obtain zero crossing counts, every local maximum and minimum is identified for both the original waveform and the low-pass filtered waveform. The distances between successive maxima and minima are calculated. Counts of distances that exceed high and moderate thresholds are collected for 12 msec windows, with a window offset of 3 msec. The values of 12 msec and 3 msec were chosen to correspond to each frame of data obtained in the signal processing phase. These four counts (High frequency--Medium amplitude, High frequency--High amplitude, Low frequency--Medium amplitude, and Low frequency--High amplitude) form the basic data upon which segmentation decisions are made.

The segmenter assigns a label to each sample of the waveform through a series of identification processes. In the first pass, reliable segment **islands** are identified by utilizing stringent requirements for the three major segment types SILENCE, VOICE, AND TURBULENCE. Although the actual detection requirements differ for each segment type, the process requires a characteristic pattern to persist over a number of frames before a segment is detected. The second phase of segmentation extends each segment island using a more moderate set of thresholds to encompass unlabeled samples. At the end of this process, many samples are still unlabeled. The segmenter utilizes heuristic rules to identify as many of these samples as possible. For example, if there are only a few unlabeled samples in an otherwise continuous segment, they are assigned to that segment. Unlabeled samples at the end of a voicing segment are likely to have resulted from an unstable vowel offset, and so are associated with the voiced segment. All remaining samples are labeled as unknown segments and usually represent the transitional region between one segment type and another. The final operation of the segmenter is to identify and report out all segment types and boundaries. Figure 4-2 shows the segments identified in the letter "G".

Begin-End detection

The digitization process used to record words or sentences leaves a silence segment at the beginning and at the end of the utterance. The purpose of begin-end detection is to identify these silences, and restrict matching operations to the embedded utterance. Begin-end detection is accomplished simply by looking for initial and final silence segments, and adjusting the utterance begin time to the ending of the initial silence segment (if any) and end time to the beginning of the final silence segment (if any). Effectively, all information before the detected utterance beginning and after the detected ending is eliminated from further processing.

Begin-end detection is not actually necessary for NEXUS to operate. NEXUS could learn about silence segments, and include them as "words" with no meaningful value. Deletion of these segments eliminates a great deal of unnecessary matching, which is by far the most expensive operation in the system. In order to maintain efficiency, matching and searching are restricted to parts of the waveform assumed to be meaningful.

Figure 4-2: Segment Boundaries found in "G" (speaker mgb)

4.3 Recognition

During the recognition process, NEXUS performs two tasks: the input is matched against known perceptual units in the speech network, and the optimal word sequence is identified. The recognition process was designed to be as general as possible: NEXUS does not make any assumptions about the length of the utterance, and is capable of identifying an utterance consisting of several words. The **matching process** works by means of a forward-looking dynamic programming algorithm, and in principle can be used to interpret a continuous, unbounded stream of input. In the **classification stage**, the sequence of perceptual units associated with the input is interpreted as a sequence of words through a simple parsing process.

4.3.1 The matching process

The matching process represents the interface between NEXUS and the external world. By matching known perceptual units against the unknown utterance, events in the input may be recognized, and the utterance can be interpreted as a sequence of known units. The basis of the matching process is spectral comparison. The input and perceptual units each consist of a sequence or vector of frames of short-term spectra. $I_{i,k}$ denotes the kth coefficient in frame i of the input, while $P(x)_{j,l}$ denotes the lth coefficient in frame j of perceptual unit x. Note that $0 \leq i \leq m$, $0 \leq j \leq n$, and $0 \leq (k \text{ and } l) \leq 25$, i.e., there are m frames in the input, n frames in the perceptual unit x, and there are 25 coefficients in each frame. We can compare a frame of the input with a frame from a perceptual unit by finding the Euclidian distance between the two intensity vectors. The **spectral distance** $SD_{I_i,P(x)_j}$ between I_i and $P(x)_j$ is computed to be:

$$SD_{I_i,P(x)_j} = \sqrt{\sum_{k=0}^{25} [I_{i,k} - P(x)_{j,k}]^2}$$

and can be thought of as a rough measure of the similarity of shape between the two frames.

In order to compare an entire perceptual unit, consisting of a number of frames of data, against a section of the input, we could simply calculate the sum of the spectral distances between each frame of the perceptual unit and a corresponding frame of the input. This simple Euclidian distance is not a very good measure of the similarity between the input and the perceptual unit, because events in speech do not always take the same amount of time. To overcome this problem, NEXUS utilizes a warping procedure (Itakura, 1975; Sakoe and Chiba, 1978) to stretch and fold the time axis of perceptual units to optimally align them against segments of the input.

Consider the **warping matrix** generated by calculating the spectral distance between each input frame and each perceptual-unit frame. If there are m frames in the input pattern, and n frames in the perceptual unit, the warping matrix is an m x n cross-product of spectral distances between these patterns, and the matrix entry $M_{i,j}$ would be the spectral distance $SD_{I_i,P(x)_j}$ calculated above. The warping matrix defines an alignment space. Any possible alignment between the input and perceptual units may be derived by choosing entries from the alignment space. For example, the linear alignment mentioned above can be derived by choosing the diagonal elements of the matrix. The global similarity between two patterns is usually defined to be the average of the selected alignment entries.

Alignment, or **warping** algorithms, provide a systematic way of choosing the correspondence between perceptual unit frames and input frames. Conceptually, warping is a means of providing a

mapping between the time indices *i* and *j* such that an optimal time alignment between the input and perceptual unit occurs. The mapping *w* between j and i is denoted by $j = w(i)$. Most warping processes generate a functional relationship between frames in the input and frames in the perceptual unit: $w(i)$ is defined for each value of $0 \leq i < m$. This functional relationship is easily seen in the warping matrix, where each row in the warping matrix is associated with one and only one column. However, the inverse function may not exist (i.e., some columns may not have an associated row, or may have multiple rows associated with them). A few investigators (Sakoe & Chiba, 1978) permit skipping and duplication of both rows and columns in the matrix. In this case, the number input frames matched is not constant for different perceptual units, causing a normalization problem in comparing the matching scores.

To preserve the general temporal characteristics of both patterns, all warping functions constrain the path that the alignment function may take. Conceptually the constraint is that one or both patterns may be stretched or folded, but not cut or duplicated, during alignment. A typical continuity constraint (Rabiner et. al, 1978), used in the NEXUS system, imposes the following conditions on the warping function:

$$w(i + 1) - w(i) = 0, 1, 2 \; (w(i) \leq w(i - 1))$$
$$w(i + 1) - w(i) = 1, 2 \; (w(i) = w(i - 1))$$

These two equations require that $w(i)$ be monotonically increasing, with a maximum slope of two, and a minimum slope of one-half. An important corollary of the slope restriction is that the perceptual unit may not be folded to less than one-half, or stretched to more than twice, its original length.

Endpoint constraints are also present for warping functions in NEXUS. These constraints require the comparison to align the first input and perceptual unit frames, and terminate the match by aligning the final input frame with a final frame in a perceptual unit:

$$0 = w(0)$$
$$n = w(m)$$

A warping path is selected to minimize the global spectral distance between the input I and the perceptual unit P(x). This minimum is calculated using dynamic programming, a recursive procedure that calculates the minimum aggregate distance D_A to any grid point in the warping matrix $M(i, j)$ as:

$$D_A(i,j) = M_{i,j} + \min_{j-2 \leq q \leq j} D_A(i - 1, q)$$

The final distance $D_{I,P(x)}$ between I and P(x) is $D_A(m,n) / m$, representing the average spectral distance between input frames and perceptual unit frames associated by the warping path. A second way of expressing this distance is:

$$D_{I,P(x)} = \sum_{i=0}^{m} SD_{I_i,P(x)_{w(i)}} / m.$$

NEXUS assumes the input consists of a sequence of one or more perceptual units. Every sequence is subject to the following constraints: each frame of the input must be identified with a perceptual unit and perceptual units are not allowed to overlap (i.e., each input frame must be associated with one and only one known perceptual unit). For a given sequence, NEXUS searches for an optimal alignment of units through a two-level dynamic programming algorithm (Sakoe, 1979).

Figure 4-3 shows a 3-node word branch being compared to the input. The two-level algorithm compares the first perceptual unit against the initial frames of the input. If m is the length of this perceptual unit, then by maximally folding this unit, a match can be completed at the $m/2^{th}$ frame of the input. By maximally stretching the unit, a match can be completed at the $2*m^{th}$ frame. Other complete matches are possible at every frame between these points. These completions determine a **transition region**, where the match of the first node ends and the match of the second begins. The second unit can be thought of as having a series of possible beginning times, corresponding to the possible end times of the first unit in the transition region. The matching process continues, and a second transition region occurs between the end of the second and the beginning of the third nodes. Matching is continued in this fashion down the sequence of input frames and down the sequence of perceptual units. A global path constraint requires the first input frame to be matched against the first frame of the first perceptual unit in the sequence, and the last frame of the input to be matched against the last frame of the last unit in the sequence.

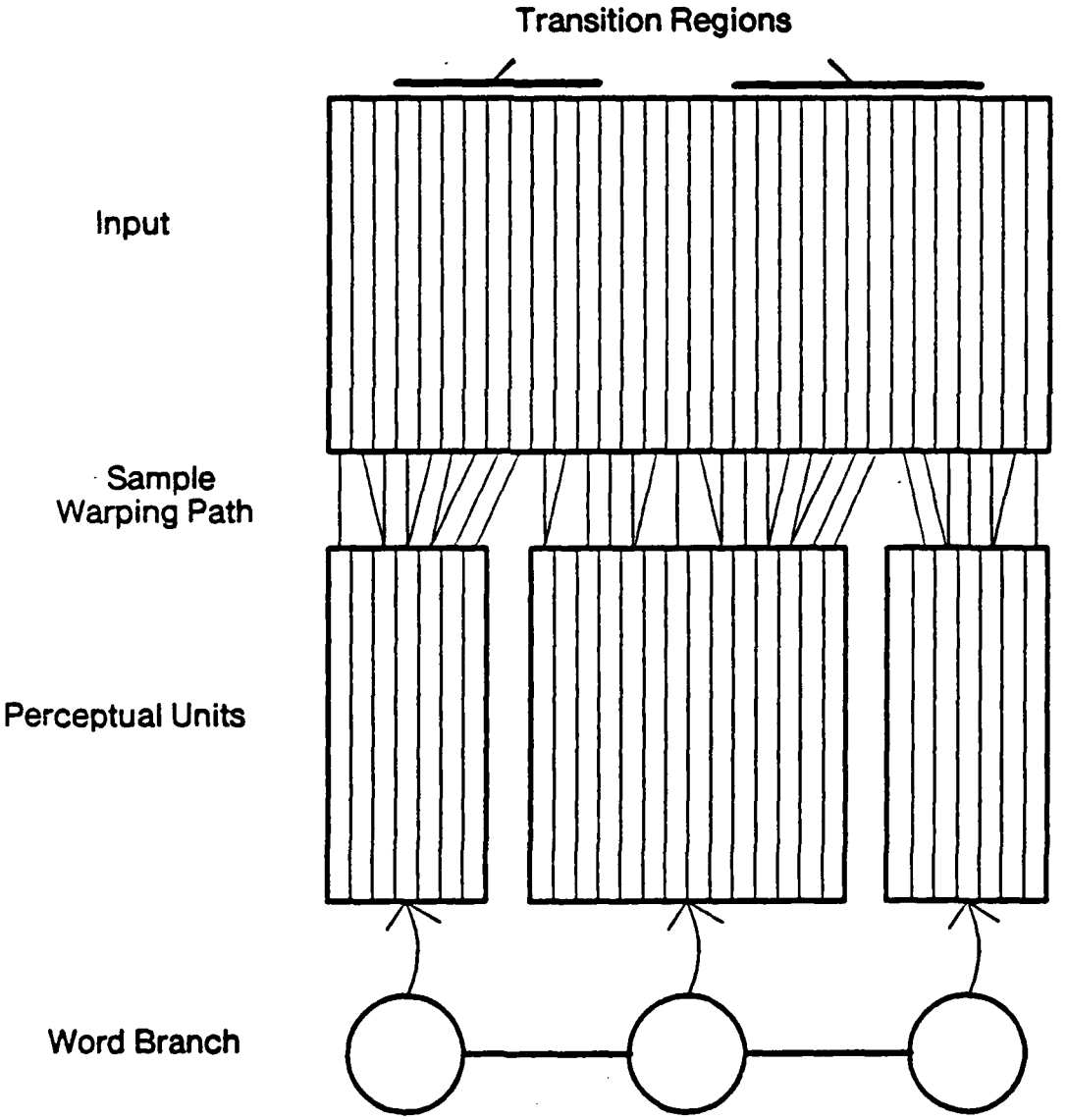


Figure 4-3: Comparison process

NEXUS must try to match all known words against the input. One method to perform all of this matching might be to sequentially match each word against the input and choose the best match. This method is a poor one for a number of reasons. First and foremost, the matching process is computationally very expensive. Halfway through a match, it might be possible to decide that a particular word is an unlikely candidate to match input. It is highly advantageous to prune all unlikely matches as early as possible. The word-sequential match has another important flaw, however. Many words share identical sequences of perceptual units. Under a sequential matching algorithm, network redundancy causes the same perceptual unit to be matched against the same input frames multiple times, again leading to inefficiencies in the matching process.

An alternative method would be to organize the comparison process not by words in the vocabulary but by perceptual units. Each perceptual unit could be matched against the beginning of the input utterance. Where each perceptual unit ends, a new complete set of matches is begun, until all frames in the input have been matched. There are two difficulties with this approach. First, most of the possible sequences of perceptual units violate the phonological constraints of a particular language. Also, there is no assurance that the matched sequence of perceptual units could be interpreted as a word or sequence of words.

The solution chosen in NEXUS is to only match sequences of perceptual units that form words in its vocabulary. These sequences are matched in parallel against the input frames, and a beam searching mechanism (Lowerre and Reddy, 1979; Bisiani and Waibel, 1982) terminates the search of poorly-matching sequences, reducing the matching load. The parallel matching permits NEXUS to match each perceptual unit against a section of the input only once, and share the matching scores with all words containing the appropriate sequence of perceptual units. Nevertheless, the matching process alone accounts for 85% of all CPU time utilized by NEXUS.

Although we have been considering single-word sequences, NEXUS searches for multiple-word interpretations of the utterance as well. This search is a straightforward generalization of the single-word matching process. Whenever a word sequence can be completed, a word-transition region can be defined. NEXUS does not impose any constraints on word sequences, so it begins matching each word against the input at every word-transition frame. NEXUS then chooses the best matching word sequence for the input.

NEXUS cannot determine whether or not a particular word will finish completion at the end of the input. Accordingly, the system does not rule out full-word paths followed by partially completed sequences. Even during recognition of isolated words, NEXUS frequently identifies a word fragment at the end of the input string. This word fragment usually represents an alternative ending for a particular utterance. Currently, the identification of a word fragment is not counted as an error during the recognition process. Unless a serious endpoint detection error occurs, the perceptual units matched in the word fragment are short, low amplitude units.

4.3.2 Classification

Following the matching process, an utterance must be identified. This is a relatively simple process beginning with the identification of the best-matching warping path. After the sequence of perceptual units is identified, NEXUS must parse this sequence as a series of words. The parsing is done using a simple recursive procedure. All word descriptions are compared against the perceptual sequence. The branch exactly matching the longest consecutive sequence of initial units is selected, and remaining units are recursively identified in the same fashion until all units have been associated

with words. This simple parsing technique would be inadequate for full language comprehension. A phonetic sequence such as /kargo/ would be interpreted as "cargo" even in the sentence "Where did the car go?". A more general parsing algorithm is needed in the system to deal with cases like this.

4.4 Learning processes

NEXUS is designed as a data-driven discrimination learning system. The main learning goal is to build a compact network capable of discriminating between all words currently in the recognition vocabulary. NEXUS incorporates general learning heuristics to modify the network as it performs its recognition task. These heuristics can either add new perceptual units and branches into the network, or prune and simplify the network. During the learning process, the system discovers the simnilarities and differences between words, and so can create perceptual units common to many words, along with multiple representations of units needed to discriminate between those words. NEXUS is designed as a **continuous learning system**. It does not make an arbitrary distinction between a training phase and a testing phase. For that reason, new words can be added into the NEXUS vocabulary at any time, and the system will readily incorporate these words into its recognition network, using existing perceptual units whenever possible.

There are two different types of learning mechanisms in NEXUS. **Instance-based** learning heuristics require information derived by comparing a particular input against the current speech network, while **network maintenance** heuristics do not need such input. Instance-based mechanisms create new perceptual units, recognize shared structure between different words, merge perceptual units, and develop prototypical units. The simpler network maintenance heuristics can only prune away branches that have proven unnecessary or have led to a large number of recognition errors.

4.4.1 Instance-Based learning

During the recognize-learn cycle, NEXUS can learn both from incorrect and from correct identifications. Incorrect instances cause the addition and/or deletion of perceptual units from the network, while NEXUS uses correct instances to refine its set of perceptual units. During the **confusion recovery** process, invoked whenever a recognition error occurs, NEXUS performs a **similarity analysis** to generate hypotheses about why the error occurred, and what actions can help prevent the error in the future. A similarity analysis is not necessary during the **positive exemplar learning** process: the input and recognized word branch are sufficient data for the process.

NEXUS make two different types of recognition errors: **new-word errors** ane **confusion errors**. Human listeners can recognize novel words correctly even on their first presentation. Due to the constraints on matching, NEXUS must interpret an utterance in terms of words it already knows, and will consequently make a recognition error whenever it encounters a new word. By inspection of the network, NEXUS can identify new-word cases from confusion errors. I will first discuss the error recovery process. The heuristics for learning in a new-word case are quite different, and will be discussed in a subsequent section.

4.4.1.1 Confusion Recovery

Confusion errors occur because the global match between the incorrectly identified word and the input is better than the global match for any of the same-identity words in the network. Usually errors of this sort occur because the information useful in discriminating between two words comprises only a small percentage of the total word durations. For example, the spoken letters "B" and "D" have different onset characteristics, but share the same vowel /i/. Because the vowel is of much longer duration than onset bursts, relatively minor vowel differences may outweigh the initial burst differences, causing a recognition error (Bradshaw, Cole, and Li, 1982). If the vowel redundancy can be recognized, the representation may be simplified by forcing both word patterns to share the same vowel segment. When new utterances are classified, only the unshared segments will lead to matching differences, and so will determine utterance classification.

The confusion recovery process, invoked whenever a word classification error is made, attempts to identify the discriminating information necessary to distinguish between the two words, and to separate that information from the information common to both words. If possible, the network may be simplified by joining together common sequences. The first step is to analyze the cause of the error through a word similarity analysis. Next, a node-correspondence analysis is performed, in order to allow for modifications to the existing speech network. If the analysis is successful in identifying the discriminating sections of the two words, NEXUS will take corrective action to the speech network, and may add new branches to the network as well.

The error recovery algorithm is diagrammed in detail in Figure 4-4. A requirement for all steps of the recovery process is to be able to compare the test item with tokens in the network that have the same word-identity (called **same-tokens**) and with tokens of the word mistakenly recognized (**error-tokens**). NEXUS begins the recovery process by performing an **error-recovery match**, shown at the top of Figure 4-4.

The error-recovery match is different from a standard **classification match** in two important ways. The error-recovery match only includes same-tokens and noise-tokens; no other words are included in the match. This restriction simply speeds up the error-recovery match. The pruning mechanism, responsible for eliminating unlikely candidates from the search, is also deactivated. This deactivation ensures that the best same-token match will be obtained. It may not be possible to match any of the same-token words against the input. Figure 4-4 shows that in this case, NEXUS defaults to adding the input into the network as a new branch, and terminates the error recovery process.

If a best same-token can be identified, Figure 4-4 shows NEXUS will compute a similarity analysis. The function of this procedure is to try to identify shared acoustic structure between words. A series of comparisons between the input, all same-tokens, and all error-tokens currently in the network is performed. Through these comparisons, **framewise spectral distance profiles** are computed. These profiles represent the frame-by-frame distance between words. Two average spectral-distance profiles are created: the same-word profile and the error-word profile. These two profiles are then subtracted, to create a **similarity profile** of the two words. At every point where the similarity profile has a positive value, the between-word match is larger (less similar) than the within-word match. Negative values occur when the between-word similarity is greater than the within-word similarity. The similarity profile is one of the most fundamental concepts in NEXUS. It represents the system's analysis of where two words are different, and where they share similar acoustic structure. This profile will be used to determine what network modifications are possible to simplify the network and prevent a similar error from occurring in the future.

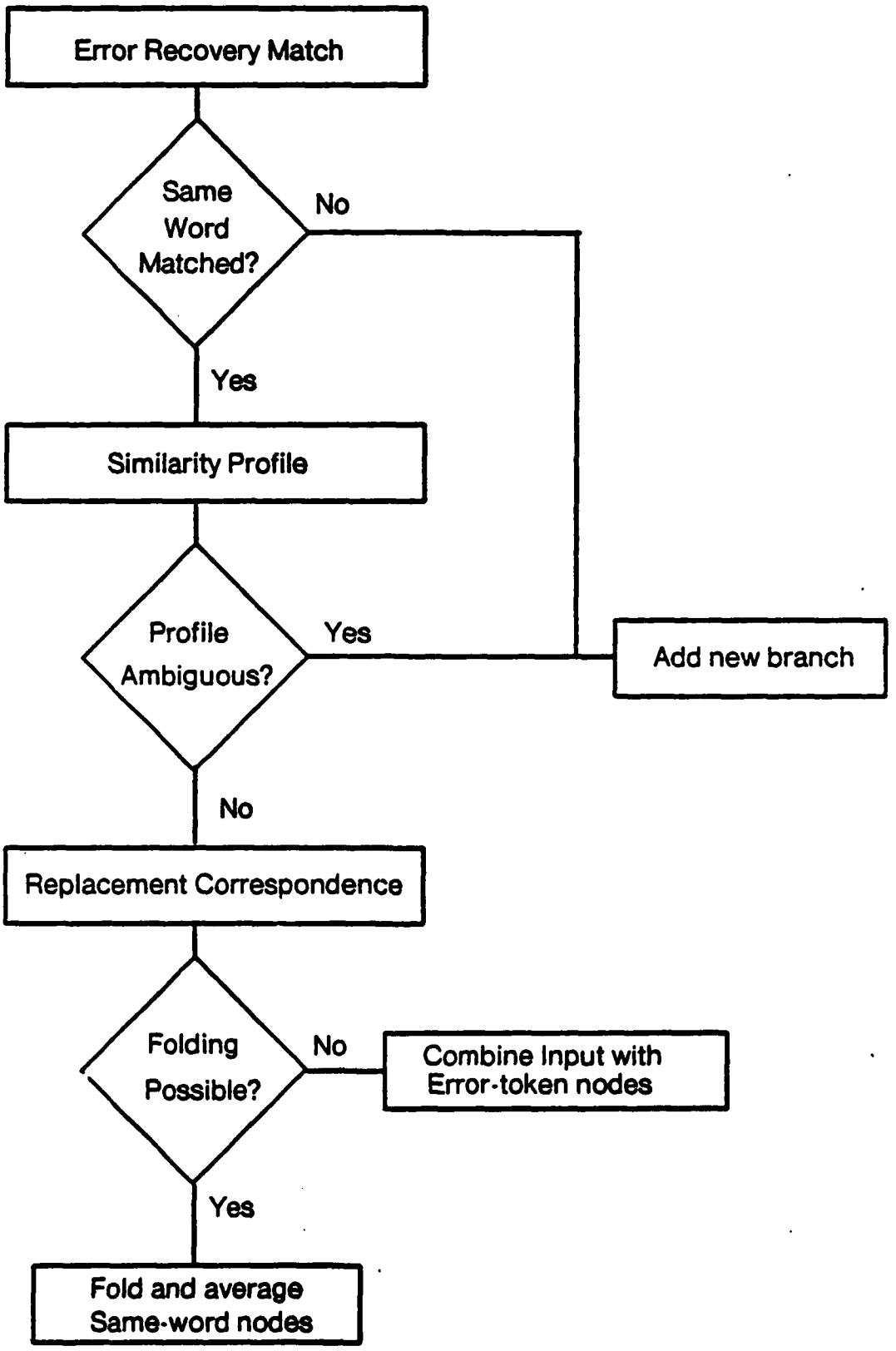


Figure 4-4: Processes in error recovery

A preliminary algorithm to calculate the similarity profile incorporated distance-profiles calculated only by matching the test-utterance against network tokens. However, since the test utterance was misclassified and therefore dissimilar to the same-word tokens, computed distance profiles were often uncertain. Occasionally these profiles indicated the wrong part of an utterance served as the discriminating segment. In such cases, NEXUS might determine that the /i/ vowel discriminated between "B" and "D", for example, then collapse the initial segments of these words together. A second more stable method of calculating the profile incorporates two base patterns: the test utterance and the best same-word token. Because the two patterns are likely to be a different length, NEXUS standardizes all profiles on the number of frames in the test utterance. A correspondence function is calculated to warp the same-word token onto the test utterance. This correspondence functions is used to standardize the lengths of the profiles.

To compute a set of profile scores, NEXUS will compare words from the network against the base patterns. Each same-token and each error-token in the network serves as a comparison pattern. A comparison pattern is warped against each base pattern, and the frame-by-frame distances along the warping path are obtained as outlined in the matching section. Each comparison pattern will generate two vectors of frame-by-frame distances, representing the distance between the pattern and the test utterance, and the distance between the pattern and the same-token, respectively. The vectors of distances are the distance profiles mentioned above. If the distance profile was calculated by matching a same-token with the base patterns, the distances are averaged into the same-word profile; error-token profiles are averaged into the error-word profile. Because these profiles are of the same length, a simple vector subtraction suffices to calculate the overall **similarity profile** of the two words.

Figures 4-5 to 4-7 are designed to provide a conceptual overview of the error recovery process. Figure 4-5 shows a simple similarity analysis for the words "B" and "D". In this figure, the speech network consists of a "B" branch $/b/_1 - /i/_1$ and a "D" branch $/d/_2 - /i/_2$[4]. The input was a second instance of the word "B", consisting of perceptual units $/b/_3 - /i/_3$. The similarity analysis finds the following frame-to-frame distances: input B – network B, input B – network D, and network B – network D distances. Frame-by-frame differences are calculated, and shown as a hypothetical similarity profile with an intuitively plausible result: the distinctive information discriminating B from D occurs at the beginning of these words.

The next step in the similarity analysis is to compare the continuous similarity profile against the discrete set of acoustic segments which were identified during the segmentation process. If a correspondence can be found, NEXUS will be able to use existing network nodes in new branches it creates. Occasionally the outcome of the similarity analysis will not be conclusive, either because the similarity profile does not reveal clear differences between words, or because the differences do not correspond to input units. In this case, NEXUS defaults to adding in all input descriptors identified during segmentation as new perceptual units, and creates a variant same-word branch containing these units.

If the similarity analysis is successful, it conveys information both about where two words are different, and about where they are similar. This discovered similarity may allow NEXUS to replace redundant perceptual units that were created in earlier network learning processes. The replacement is effectively done by folding together network units. A necessary condition for this folding process is to find the correspondence between the units in the same-word utterance and the incorrectly

[4]Subscripts are used here as arbitrary means of distinguishing between nodes, and have no other meaning

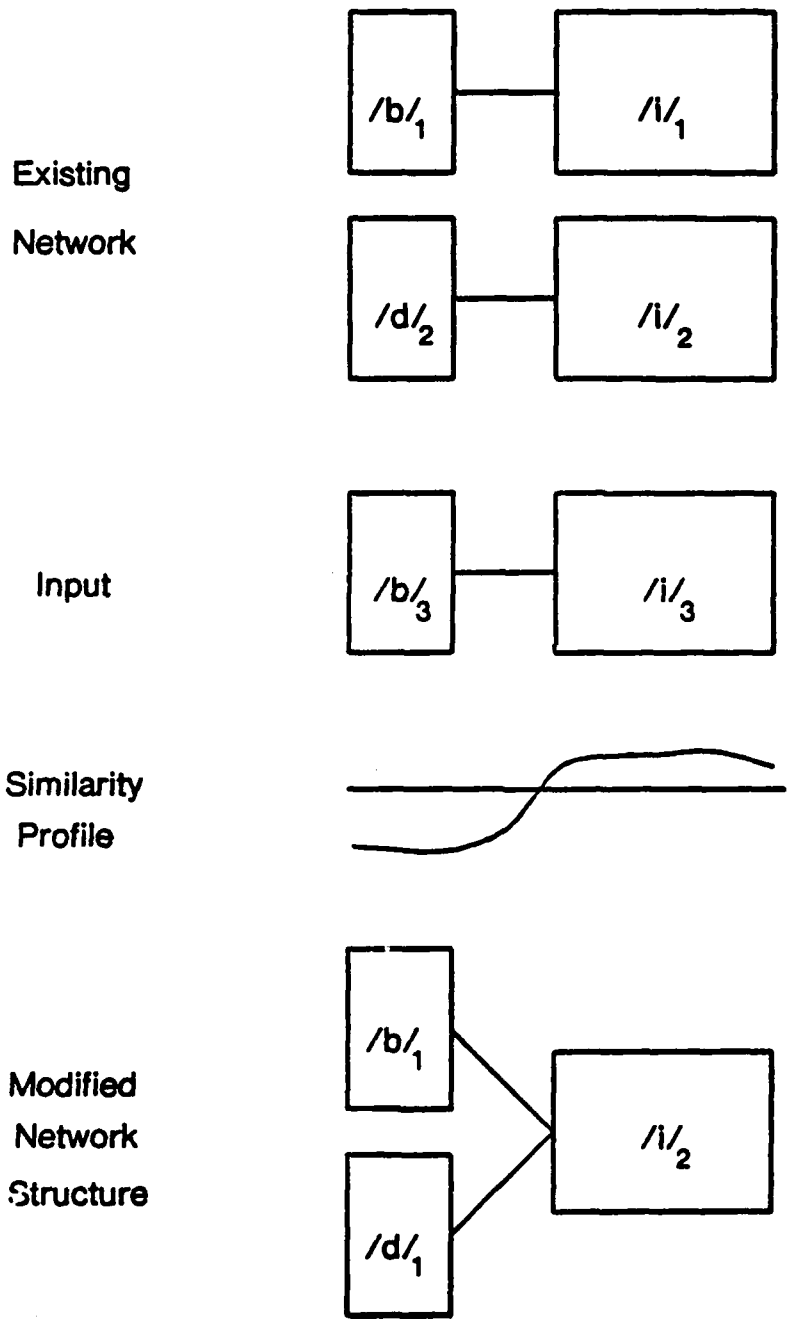


Figure 4-5: Successful network folding operation

recognized word units. The correspondence is discovered through a simple process. First, the incorrectly identified word is warped onto the same word. Using the resulting frame correspondence, each segment of the same-word utterance is compared with the segments of the incorrectly identified word. Segments are said to correspond if they have more than 60% of frames in common. The correspondence analysis does not necessarily derive a one-to-one mapping. Every same-word node must be matched against at least one segment from the incorrectly identified word, but it may match against multiple nodes as well.

If a replacement correspondence is obtained, NEXUS will eliminate redundant perceptual units from the network. In Figure 4-5, the correspondence analysis would map /b/$_1$ onto /d/$_2$ and /i/$_1$ onto /i/$_2$. The similarity analysis implies segments /i/$_1$ and /i/$_2$ are two instances of the same node. In this case, the perceptual units associated with these nodes are averaged together, and the network is folded as shown in the modified network structure at the bottom of the figure. NEXUS can determine whether or not the newly-created structure would have correctly discriminated between the two utterances. If so, the error recovery process terminates.

Figure 4-6 shows one possible cause for a node correspondence failure. In this figure, only one segment was observed for the network B. The similarity profile reveals a distinctive difference between words, but the network cannot be folded, as there is no means to arbitrarily split the existing /bi/$_1$ pattern.

Figure 4-4 illustrates a final action that may be taken by NEXUS. If the newly created network structure would not have correctly discriminated the test utterance from the error-token, or if the replacement correspondence is uncertain, a network addition is indicated. NEXUS will attempt to incorporate nodes from the recognized-token in the new branch, along with nodes from the input. A second correspondence analysis, between the recognized-token and the test utterance, is performed. Figure 4-6 illustrates a case where the second analysis is successful, and a new branch is created incorporating a new perceptual unit from the input, and the vowel unit from the existing D branch. If this analysis fails, NEXUS must revert to the default strategy of adding in an entirely new branch. Figure 4-7 shows the default action, adding a new branch due to the failure of the similarity analysis.

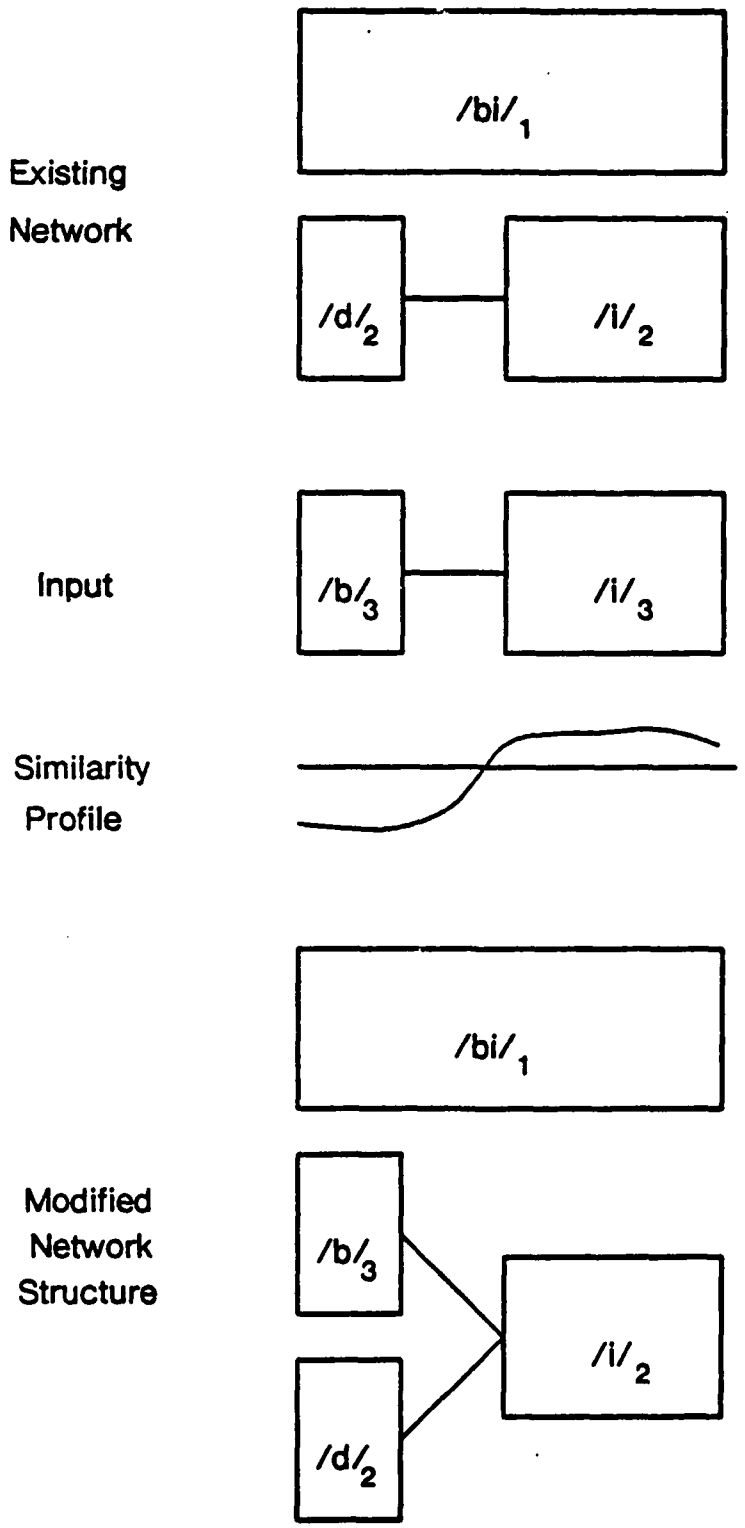


Figure 4-6: Network modifications when correspondence is unsuccessful

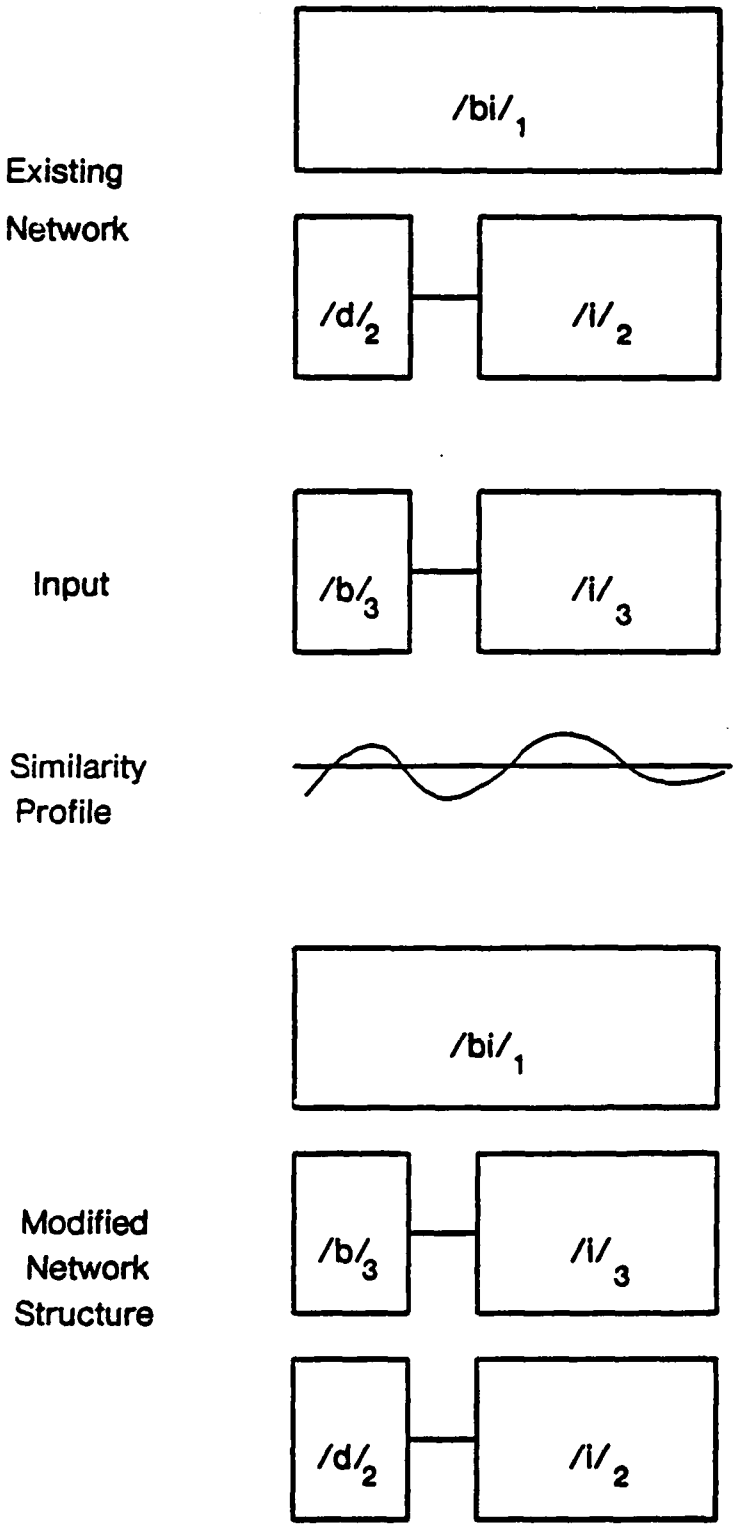


Figure 4-7: Network modifications when similarity analysis is unsuccessful

4.4.1.2 New-word learning

NEXUS will always interpret an utterance as a sequence of previously-encoded words. Such an interpretation is wrong whenever the sequence includes a word as-yet unknown to NEXUS. After the error analysis is begun, NEXUS will discover that no existing versions of the new word exist in the network, and will enter into a special new-word analysis process. The new-word learning heuristic is designed to allow NEXUS to use existing perceptual units to encode new words. The first step in new-word learning is to exhaustively match every possible sequence of existing perceptual units against the input. The best matching sequence is then associated with frames of the input. The quality of match for each input segment is calculated by finding the average frame distance. Wherever NEXUS can identify existing units that are very good matches to the segments of the input, those units will be incorporated into the new-word description. If no such units can be identified, NEXUS adds the input segments into the set of phonetic units, and creates a new network branch to represent the new word.

4.4.1.3 Positive-exemplar learning

Positive exemplar learning may occur whenever an utterance is correctly identified. In such cases, perceptual units may be refined by means of an averaging process. Perceptual units are initially derived from a specific instance of speech. The pattern associated with the perceptual unit reflects not only the characteristic properties of the perceptual unit, but also random preturbations as well. By averaging together several instances of a unit, such fluctuations will hopefully be eliminated, and a prototypical pattern can be obtained. Averaging is performed on a unit-by-unit basis for all perceptual units that have been averaged with fewer than 10 other inputs. The weighted nature of the average limits the utility of averaging beyond 10 inputs.

To average together a perceptual unit and a section of the input, we must associate an input frame with each frame of the perceptual unit. This information is not available from the initial warping, which instead associates a perceptual unit frame with each frame of the input. A second warp is performed to obtain the correct correspondence function. This error-recovery warp associates an input frame with each frame in the perceptual unit. A weighted average is calculated for each coefficient in every frame by multiplying the perceptual unit coefficient by the experience of the unit, adding the coefficient from the input, and dividing by the new experience value (e.g., old experience plus one). This averaging process helps NEXUS to derive prototypical perceptual units by removing random variability from each unit.

4.4.2 Network maintenance heuristics

In addition to the instance-based learning heuristics, which depend on data derived during the recognition process, a second class of heuristics was identified. These network maintenance heuristics act to simplify the network by eliminating unnecessary or harmful branches. Network maintenance heuristics accept as input the speech network itself, along with a set of statistics NEXUS keeps on each branch in the network. These statistics include frequency-of-usage information, the number of correct and false identifications for each branch, and the last time each branch was used correctly in a recognition decision. Although these heuristics could be applied after every recognition process, just as the instance-based heuristics are, the underlying statistics do not change much as a function of a single recognition decision. Accordingly, these heuristics were invoked after every recognition cycle, where each word in the vocabulary is tested. In more natural interactions with a speech recognition system, items in the vocabulary are not produced in strict sequence. For these conditions, network maintenance heuristics could be applied after a constant number of recognitions, whenever a large number of errors was observed, or whenever the size of the network grew too large.

The first network maintenance heuristic, designed to eliminate unused branches, was added to the system because it was observed that many branches were being added to the network that did not match against new instances and so presumably had only limited generality. Also, other initially useful branches were superceded by later versions of the same word. At the end of every recognition cycle, the system checks each branch to see how recently it has been used. After five successive cycles without being used correctly, a branch is considered for deletion. Deletion is not automatic, because it was felt that branches which had been used successfully on many occasions might still prove to be useful, so the heuristic requires four additional unused trials for each correct recognition attributed to a branch.

The second network maintenance heuristic serves to eliminate branches that are responsible for a large number of recognition errors. This heuristic calculates the ratio of incorrect identifications to the total identifications attributed to each branch. If at the end of any recognition trial this ratio is found to be larger than 40%, the branch is eliminated from the network.

4.5 summary

This chapter provided a description of the NEXUS speech recognition system. Features of the system include a sophisticated matching algorithm capable of continuously matching perceptual units against a waveform, a multiple-word classification process, and a set of learning mechanisms to acquire and utilize knowledge of speech. Rather than making an artificial separation of training and recognition phases, common to nearly all other speech recognition systems (e.g., Harpy: Lowerre, 1977; Lowerre and Reddy, 1979; Hearsay: Lesser, et al., 1975; HWIM: Wolf and Woods, 1977; IBM: Dixon and Silverman, 1977; Nadas, 1984; and template systems: Bradshaw, et al., 1982; Itakura, 1975; Rabiner, 1978; Rabiner, et al., 1979; Myers and Levinson, 1982), NEXUS incorporates continuous learning algorithms. The system is always prepared to add new words into its vocabulary, to analyze confusion errors and make corrective changes, or to eliminate unnecessary information from its database. The following chapter will provide information on several experimental tests, designed to explore the performance of NEXUS.

5. NEXUS performance results

In order to evaluate NEXUS a database of spoken utterances was recorded and maintained on a system disk. The database contains utterances from two talkers, each of whom recorded 30 versions of the spoken alphabet (letters "A" to "Z"). The recordings were made in a series of sessions; no more than 6 sets of the alphabet were recorded at any session. The recording room is a laboratory in the Computer Science department at Carnegie-Mellon University. Room acoustics are comparable to those found in normal office environments. Speaker mgb recorded all tokens with a close-talking Sure microphone during quiet hours in the laboratory. Speaker mrg used a hand-held microphone, and recorded tokens during normal laboratory hours. The background noise level is considerably higher for these tokens, including noise from people moving around, personal computers operating, and so on.

Digitization was performed by a Digital Sound Corporation a/d converter, which includes a microphone pre-amplifier and amplifier. The microphone input was filtered at 6.4 kHz and sampled at a rate of 16 thousand samples per second. Each sample is recorded with an accuracy of 16 bits. A few tokens were re-recorded due to incorrect amplitude levels or because an utterance was incompletely recorded. Otherwise, tokens were not corrected even though additional events (lip-smacks, offset breath releases, etc.) or other abnormalities might be present.

The alphabet was chosen to be the test vocabulary because it is a compact vocabulary, yet presents a great challenge to speech recognition systems (Bradshaw, Cole, and Li, 1982). This vocabulary is difficult because many of the words have a highly similar acoustic structure, such as the /i/-set ("B" "C" "D" "E" "G" "P" "T" "V" and "Z") and the /eh/-set ("F" "L" "M" "N" "S" "X"). Speech recognition systems cannot rely on global acoustic patterns to disambiguate between these words; distinctions must be made on the basis of fine phonetic structure. Thus the alphabet was thought to represent an interesting challenge for the learning theory of speech recognition.

5.1 NEXUS recognition performance

There are several different standards that might be used to evaluate the performance of NEXUS. The most stringent standard would be to compare the system with human performance on the same database. Although no formal tests of human classification performance on the tokens were performed, the high quality of the data and the simplicity of the recognition task would undoubtedly lead to error rates far less than 1%. A second, less stringent standard is to compare NEXUS with other machine speech recognition systems. Fortunately, a state-of-the-art speech recognition system, Cicada, was available to classify the utterances. Cicada, a recognition system developed at Carnegie-Mellon University, is comparable to the best currently available recognition systems. Cicada results were obtained for each speaker by allowing each of the 30 sets of utterances to serve as the reference set, and using each reference set to classify all other tokens from the corresponding speaker. For the alphabet task, Cicada had a 21% classification error rate for speaker mgb, and a 19% error rate for speaker mrg.

Because NEXUS is a continuously-learning system, a more detailed analysis of performance than average classification rate is necessary. A measure reflecting the dynamic adaptation of the system is to evaluate performance for each recognition cycle, defined to be a classification of each vocabulary item from one data set. For the current database, each cycle consists of one recognition trial for each of the spoken letters "A" to "Z". Because NEXUS begins without any knowledge of

speech whatsoever, performance on the first recognition cycle is always 0%. All discussions of NEXUS performance are based on the remaining 29 recognition cycles.

5.1.1 Performance for speaker mgb

Figure 5-1 shows the cycle-by-cycle performance of the NEXUS system for the first speaker, mgb. Across all recognition cycles, NEXUS correctly classified 698 utterances, and incorrectly classified 51. Five recognition trials are missing. Missing trials occur when a timeout occurred from the host computer's remote file server. The average recognition rate for the remaining 749 trials is 93%. Average performance for the first five cycles is 80%, and the average for the final five cycles is 97%.

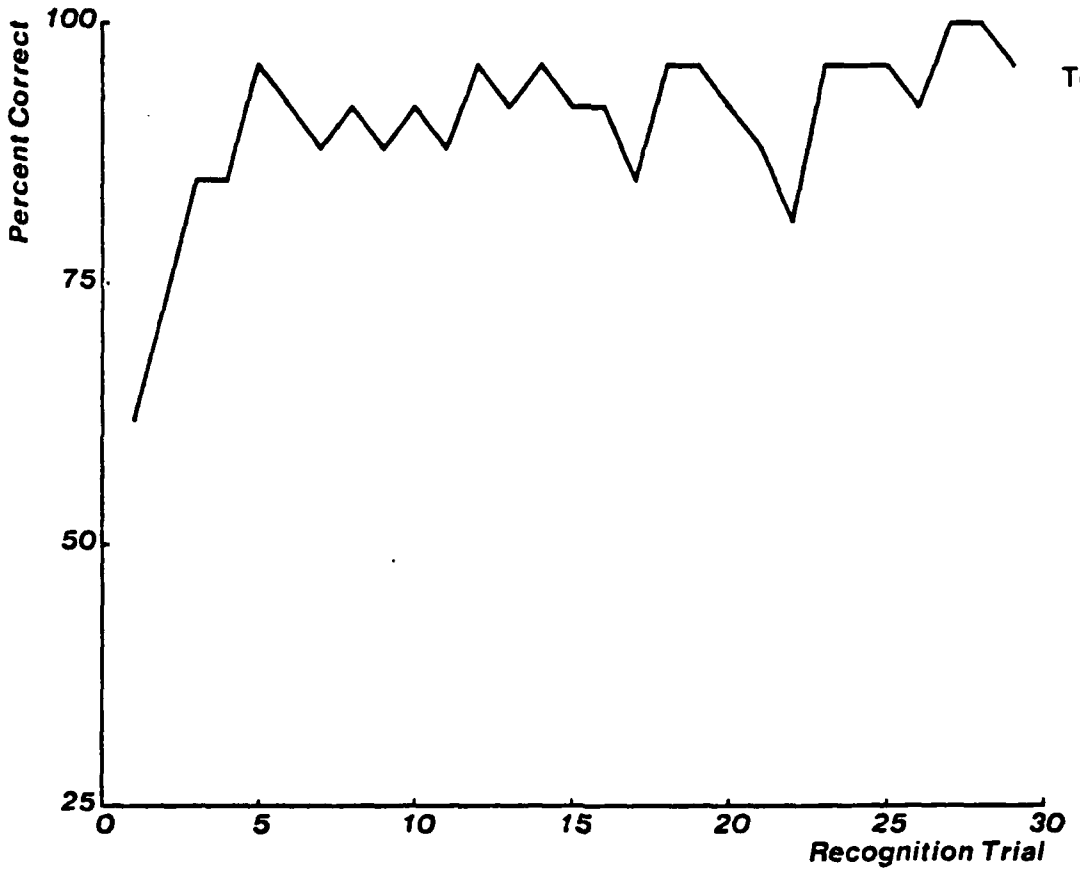


Figure 5-1: NEXUS Alphabet performance for speaker mgb

An important learning trend is readily apparent in Figure 5-1 over the first five cycles. Performance increases from 62% on cycle one to 92% on cycle five. Performance remains relatively constant after cycle five, although there appears to be a further slight improvement in the remaining twenty five cycles.

Confusion matrices, summarized by five-cycle blocks, are presented in Appendix A. Even from the very earliest cycles most confusions are not random, but appear very similar to those produced when human listeners classify nonsense syllables (Miller and Nicely, 1955; Singh and Black, 1966; Wang and Bilger, 1973). A summary of errors occurring after cycle one appears in Table 5-1. Performance is analyzed by subsets, according to the primary vowel of the spoken word. A within-set error

occurs whenever a word is classified as one of its set-cohorts. These errors are usually psychologically plausible. Between-set errors occur when the word identified belongs to a different subset than the word spoken. These errors tend to violate acoustic and phonological similarities, and so represent important departures from human performance. For speaker mgb, eight out of the 51 errors made by NEXUS are psychologically implausible errors of this sort, representing 15.7% of all mis-classifications. These eight errors are shown in Table 5-2.

Set	Vocabulary	Percentage Correct
/i/	{"B","C","D","E","G","P","T","V","Z"}	91%
/eh/	{"F","L","M","N","S","X"}	91%
/a/	{"A","H","J","K"}	96%
other	{"I","O","Q","R","U","W","Y"}	97%

Table 5-1: Speaker mgb confusion errors by sub-vocabulary

Because these errors are atypical of human performance, an analysis was performed to evaluate why these errors occurred. Table 5-2 includes the matching scores representing the observed degree of similarity between the test word and the recognized-token. As described earlier, NEXUS automatically performs an error-recovery match as part of the recovery process. The matching score for the best same-word item is shown in the final column in Table 5-2. Note that in every case the same-word match is better than the classification match.

This type of error can occur in only one circumstance. If the same-word item had been completely matched against the test utterance, the lower score obtained would have resulted in a different recognition. The pruning mechanism must have eliminated from further search an utterance whose final score is superior to the recognate. Although this does not mean that each of these words would have been correctly identified if the system did not prune, it does demonstrate that the obtained match is a poor one, and a more plausible error (or correct recognition) might have occurred if no pruning had been done. This hypothesis is further supported by an analysis of the actual scores for words. The distribution of matching scores was calculated for two groups of items. The first group contains all correct matches and all plausible errors (within set errors) shown in Table 5-2. The second group contains all implausible errors. The mean matching score for the plausible group is 264.7, while the mean for the implausible group is 570.0. Standard deviations for the two groups are 71.4 and 119.2, respectively. Whenever an implausible match occurs, the matching score is likely to be very high, averaging 4 s.d. larger than the mean of plausible matches. All but one of the anomalous errors would be caught if a rejection threshold was set at two standard deviations above the mean. These errors do not appear to be due to an unexpected similarity between words that humans judge to be very different. Rather, they appear to be due to a localized difference between instances of the same word. These localized differences may be due to extraneous noises present in some of the instances of speech, or due to different articulations of a word. Improvements in the pruning algorithm might reduce or eliminate these errors.

Figure 5-2 shows the number of branches at every cycle for the alphabet task. The number of

Error	Cycle	Score	Best Same Word Match
A =>X	30	675	235
H =>B	18	369	222
O =>W	12	690	541
O =>W	18	558	446
O =>R	22	668	424
O =>R	30	635	449
R =>Y	17	436	281
Y =>R	22	529	368

Table 5-2: Analysis of implausible errors for speaker mgb

branches increases during the first seven cycles, then asymptotes at approximately 45 branches. The final network representation includes 14 words with only a single branch, 10 words with two branches, two words with three branches, and one word ("v") with five branches.

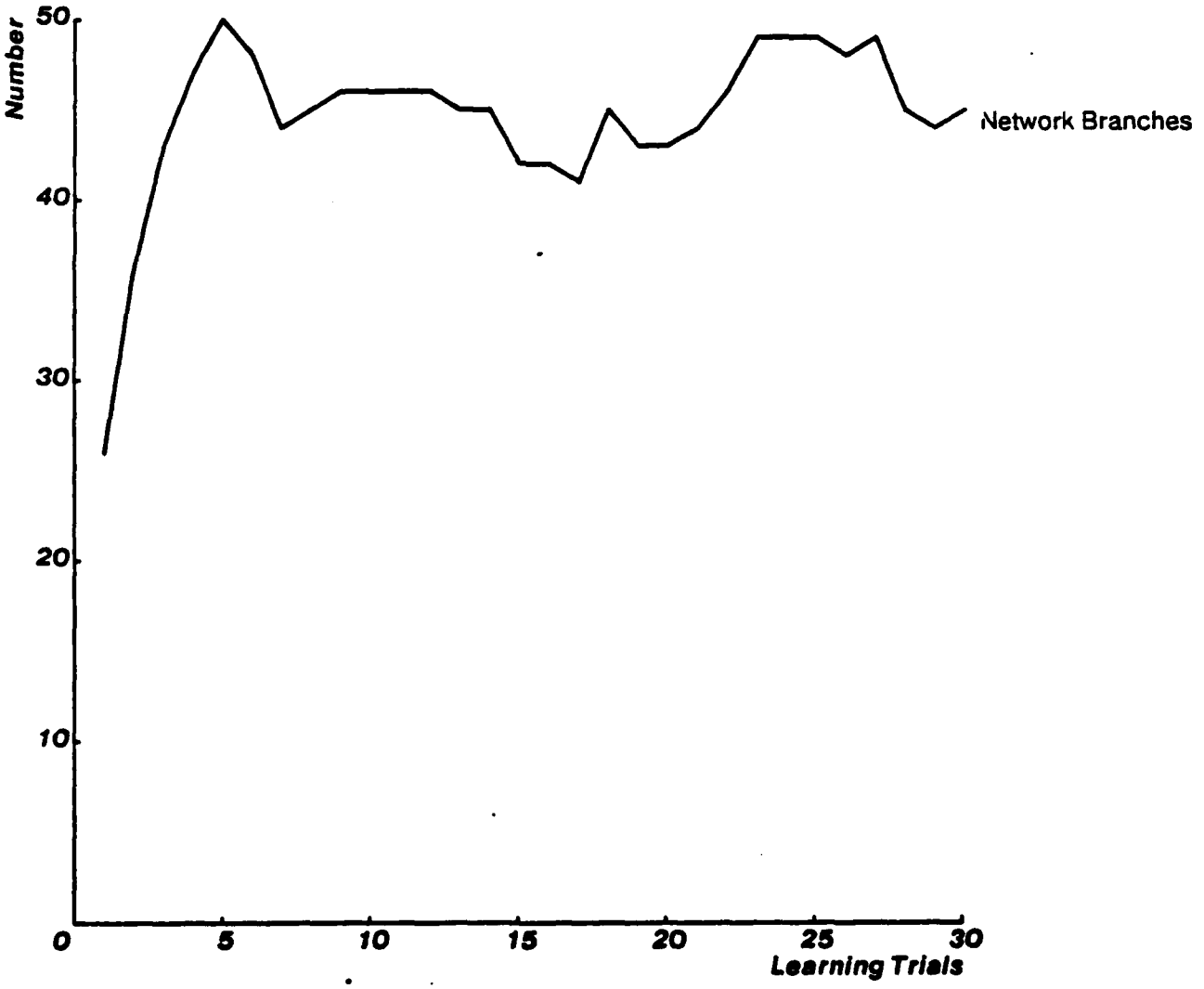


Figure 5-2: Complexity of alphabet network for speaker mgb

5.1.2 Performance for speaker mrg

Cycle by cycle error rates for the second speaker are shown in Figure 5-3. Unfortunately, 27 trials are missing, due to file server timeouts[5]. Out of the remaining 727 trials, 627 (86%) were correctly classified, while 100 (14%) were misclassified. This error rate represents a 6% improvement over Cicada's 80% classification rate. On the first five trials, the average percentage correct identification rate was 78%, while the average for the last five trials was 86%.

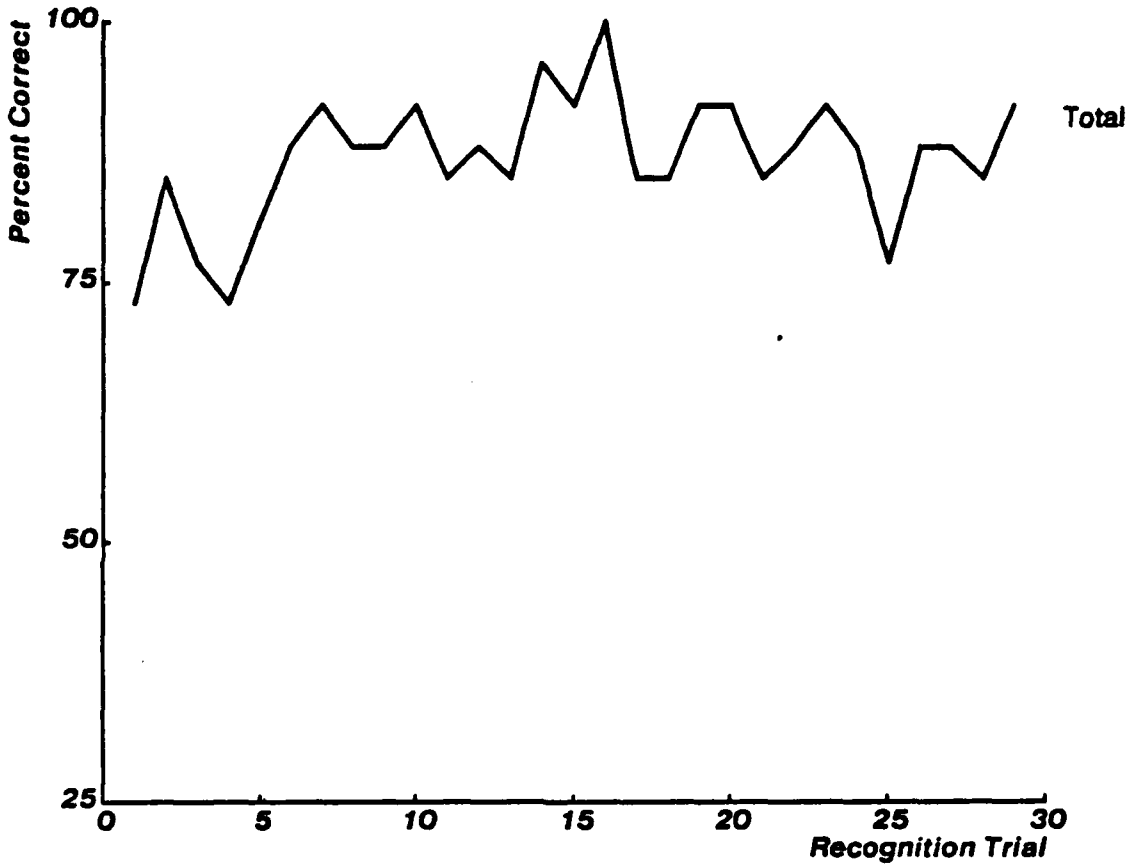


Figure 5-3: NEXUS Alphabet performance for speaker mrg

Confusion matrices, summarized by five-trial blocks, are presented in Appendix Two. The error pattern for sub-vocabularies follows a slightly different pattern than for the first speaker. Table 5-3 shows performance on the four subsets. /eh/-set performance is far below the performance of other sets, accounting for more than half the errors NEXUS makes on all recognition trials, even though the set only represents 23% of the alphabet. Errors within this subset occur for two pairs: "F" – "S" and "M" – "N". In the second recognition cycle, a programming error occurred in NEXUS. Although the replacement correspondence analysis failed to find a correspondence between the same-token and the recognized-token, the replacement procedure was invoked. This caused an anomalous path for "S" to be created, leading to many of the later errors. The identical error occurred in cycle 12 when

[5] 10 out of the 27 missing trials were "E" tokens. The spoken letter "E" is one of the more difficult to classify, so the performance figures quoted may be slightly overstated. If we assume a worst-case hypothesis that all of these E tokens were misclassified, the overall performance figure would be 627/ 737 or 85%. All error analyses are corrected for the missing trials by eliminating these trials from the candidate pool.

"M" was mistaken with "N". Unfortunately, the error in the code was not detected in time to repeat the recognition test.

Set	Vocabulary	Percentage Correct
/i/	{"B","C","D","E","G","P","T","V","Z"}	85%
/eh/	{"F","L","M","N","S","X"}	68%
/a/	{"A","H","J","K"}	94%
other	{"I","O","Q","R","U","W","Y"}	99%

Table 5-3: Speaker mrg confusion errors by sub-vocabulary

Only three of the 101 recognition errors were between-class errors. Two of the three errors followed the pruning error pattern characteristic of between-class errors for the first speaker. The final error occurred when an "M" was recognized as an "A". The error-recovery match failed, indicating that the existing "M" branch could not be matched against the input. In all three cases, scores were well outside of the normal range of matches, indicating a rejection threshold would have disqualified the obtained match.

The development of the network shows a similar pattern to that of the first speaker. Figure 5-4 presents the cycle-by-cycle size of the network. The asymptotic number of branches would appear to be about 50, although there is some indication that the network may not have reached an asymptotic value by the end of 30 trials. An inspection of network size reveals that part of the instability is due to the creation of bad paths. By the end of the experiment, both of the faulty paths have been eliminated by the error-ratio heuristic. The final network contains 11 words represented by only a single branch, nine words represented by two branches, three words represented by three branches, two words represented by four branches ("D" and "P") and one word represented by six branches ("F").

5.2 Evaluation of learning heuristics

NEXUS incorporates a complex set of learning mechanisms. During the learning process, NEXUS performs an error analysis to attempt to isolate the similarities and differences between words. This section will discuss various components of the error analysis, to evaluate their effectiveness. Our analysis of learning mechanisms will follow the learning procedure flowchart presented in the last chapter (Figure 4-4).

The first step in learning is to find the best same-word token. If no existing tokens can be found, the system cannot perform the similarity analysis, and must default to adding the input to the network. Fortunately, NEXUS is able to identify a same-word token in nearly every instance. In only two cases for speaker mgb, and four cases for speaker mrg, was it impossible to identify a matching same-word token.

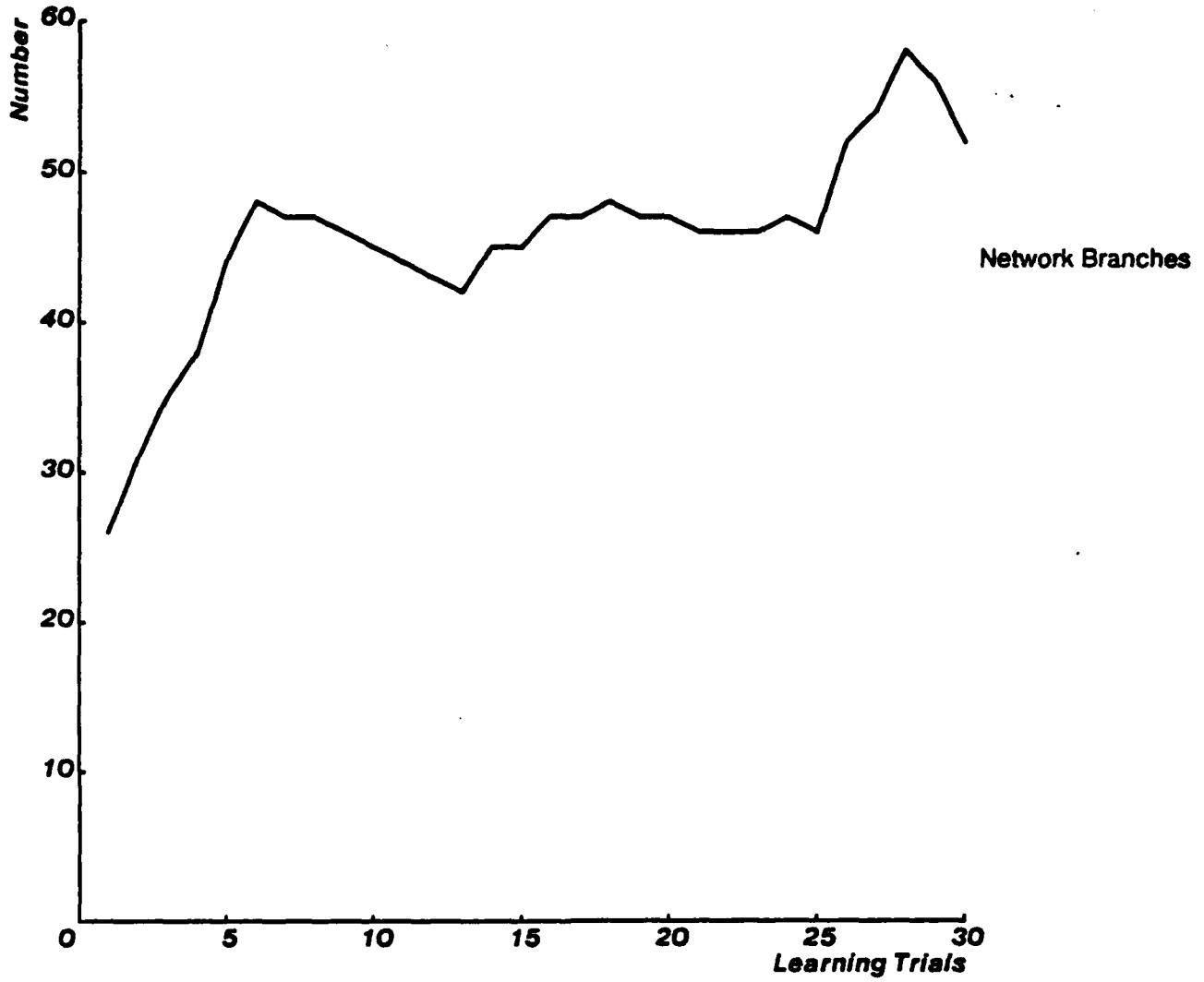


Figure 5-4: Complexity of alphabet network for speaker mrg

5.2.1 Similarity profile

An analysis of the similarity profile extraction process appears in Table 5-4. There are several different outcomes that may result, presented in the various columns of the table. A similarity profile is not calculated when NEXUS determines the error was caused by an extraneous noise in the utterance. In that case, and whenever the same-word token cannot be found, NEXUS will omit the similarity profile. This occurs in roughly 10% of all errors for both speakers. The profile is completely successful in 56% of the cases for speaker mgb and 39% of the cases for speaker mrg. A successful profile occurs whenever NEXUS correctly identifies both the discriminating and similar sections of the two words. In this case, NEXUS will attempt to fold network nodes together and/or fold the input together with existing network nodes. In roughly 20% of the cases, Table 5-4 shows the profile was positive. Positive profiles represent somewhat of an anomaly, implying that all nodes in the test utterance are, on average, closer to the same-word than to the recognized-word. Under this circumstance, an error would not be expected. However, the similarity profile is based on the average distances, instead of best cases. A more sophisticated decision rule than NEXUS' nearest-neighbor,

such as the *k-nearest-neighbor rule*, might avoid these recognition errors. Wherever the profile is positive, NEXUS tries to fold the input together with same-word nodes, to create a new branch. If no foldings would have prevented the error from occurring, NEXUS adds the input into the network as a new branch.

	Total Errors	Profile OK	Profile Positive	Profile Negative	Input Singular	Profile Indeterm	Profile Omitted
MGB	51	29 (56.7)	9 (17.6)	2 (3.9)	3 (5.9)	1 (2.0)	7 (13.7)
MRG	101	39 (38.6)	22 (21.8)	12 (11.9)	16 (15.8)	1 (1.0)	11 (10.9)

Table 5-4: Similarity Profile Analysis

Three of the columns in Table 5-4 represent failures of the similarity profile analysis. Negative profiles occur when the analysis finds the between-word distances to be always smaller than within-word distances. Indeterminate profiles occur because NEXUS incorporates a threshold for positive values. If no nodes exceed the positive criterion, the similarity profile is uncertain. Finally, if the input word is comprised of only one node, NEXUS cannot identify similar and distinct nodes of the word. A similarity profile is impossible in this case, and NEXUS is forced to add the input in as a new branch. Collectively, these failures account for 12% and 29% of the cases for speakers mgb and mrg respectively. Profile reversal errors, where the incorrect node is identified as discriminating between two words, causes great difficulties for NEXUS. Profile reversal errors never occurred in any of the 152 errors analyzed by the system. The difference in performance between speaker mgb and mrg probably reflects the error in the program that caused incorrect branches to be included for speaker mrg.

5.2.2 Correspondence Analysis

In order to fold the network, NEXUS must establish a correspondence between the node structures of the same-token and the error-token. Table 5-5 summarizes the correspondence analysis for these two tokens. Failures to generate a correspondence can be caused by different segmentations of the two words, or the fact that either of the words consists of only a single node. These failures occur in 24% and 33% of the cases for speakers mgb and mrg, respectively.

	Total Errors	Correspond Found	No Correspond	Other Singular	Same Singular	Correspond Omitted
MGB	51	32 (62.7)	7 (13.7)	5 (9.8)	0 (0.0)	7 (13.7)
MRG	101	56 (55.4)	19 (18.8)	11 (10.9)	3 (3.0)	12 (11.9)

Table 5-5: Replacement Correspondence Analysis

Finally, a correspondence between the input and the error-token is used to permit a folding of the input and error-tokens, to create new branches. The results of this analysis are presented in Table 5-6.

	Total Errors	Correspond Found	No Correspond	Other Singular	Same Singular	Correspond Omitted
MGB	51	31 (60.7)	8 (15.7)	5 (9.8)	0 (0.0)	7 (13.7)
MRG	101	32 (31.7)	13 (12.9)	5 (5.0)	1 (1.0)	50 (49.5)

Table 5-6: Input Correspondence Analysis

5.2.3 Final Actions

NEXUS has many alternative action responses available, depending on the analyses described above. A goal of the learning heuristics is to add as few new nodes and branches as possible, and to fold the network whenever possible to eliminate redundancy. Table 5-7 provides an analysis of the actions taken by NEXUS for each error trial. The default action, to add the input to the network, can occur in many different circumstances, as shown in Figure 4-4. For both speakers, this occurs in approximately 60% - 65% of all error trials. A correct fold is the most desirable outcome, because no new nodes are added into the network. This occurs in only a very small number of cases. Of all the folding conditions, it is most common for NEXUS to fold together existing branches and add a new branch containing some input nodes. This occurs in 29% of the cases for speaker mgb, and 19% for speaker mrg.

	Total Errors	Null Branch	New Fold	Corr Fold	Both Fold	Same Fold	Add Input
MGB	51	2 (3.9)	2 (3.9)	0 (0.0)	15 (29.4)	2 (3.9)	30 (58.8)
MRG	101	6 (5.9)	3 (3.0)	2 (2.0)	19 (18.9)	4 (4.0)	67 (66.3)

Table 5-7: Final Action Analysis

5.3 Summary

This chapter described performance tests on NEXUS, and presented an evaluation of various system components. NEXUS was shown to be an effective learning system, with heuristics capable of analyzing words for their component elements. These elements, when incorporated into a word network, could be used to recognize new instances of known words with high accuracy.

NEXUS correctly recognized 93% and 86% of all utterances for speaker mgb and mrg, respectively, while a state-of-the-art recognition system scored 79% and 80% on the same utterances. A useful measure of the relative performance of these systems is to calculate the ratio of errors made by the two systems. Cicada made 3 times as many errors as NEXUS for speaker mgb, and 1.43 times the errors for speaker mrg.

NEXUS learning heuristics were also shown to be effective in identifying and eliminating network redundancies. Although the heuristics could not universally identify network redundancies, NEXUS was able to recognize overlap in many cases, and fold the network extensively, as shown in Appendix II.

To complete the discussion of NEXUS, the final chapter will go beyond objective, statistical evaluation methods in an attempt to communicate remaining difficulties with the project, and provide the reader with information on important improvements to NEXUS.

6. Discussion

NEXUS is a complex program charged with a difficult task that has eluded psychologists for many years: to identify the important acoustic characteristics of speech. Although the system performed well when compared to another speech recognition system, it is important to recognize the limitations of the program. NEXUS is not able to perfectly recognize even a small vocabulary after thirty trials. The error recovery process was successful in some cases, but not as successful as might have been desired. Although improvements to NEXUS could certainly be made, the limited success obtained from a very sophisticated system underscores the difficulty of speech recognition.

NEXUS was never intended as a "no-holds-barred" engineering approach to speech recognition. Instead, the NEXUS project made a specific point: human-like learning heuristics can be designed to analyze speech, identify a set of perceptual units, and use these units to recognize speech. The importance of this demonstration is two-fold. First, it serves as an existence proof of a speech learning system. Although a computer model of this sort may not convince all psychologists of the necessity of human perceptual learning, the success of the system revives the possibility of human learning. Next, the system represents a radical departure from current speech recognition systems. Most computer systems incorporate statistically-based pattern learning and recognition techniques. NEXUS illustrates that heuristic learning techniques can be used as well. The advantages of heuristic learning techniques have already been described in the previous chapter. The disadvantage of a heuristic-based system is that heuristics do not come with guarantees, and construction of useful heuristics remains more of an art than a science. Hopefully, NEXUS will illustrate that heuristically-guided learning systems hold promise even in areas outside of their more traditional AI applications.

The previous chapters have presented, described, and analyzed NEXUS. At this point, I would like to supplement the formal descriptions and quantitative analyses with subjective comments on various components of the system. These comments are intended to further the readers understanding of what has been accomplished and of what remains to be done.

6.1 Segmentation and sensitivity to noise

NEXUS is based on a model of speech as a sequence of discrete acoustic events. A new segmentation system was built to identify speech segments. As is well-known in the field of computer speech recognition, it is very difficult to build a robust segmentation system. Unfortunately, no quantitative analyses could be performed on the segmenter. Subjectively, the segmentation algorithm performed very well. If unstable segmentations are performed, NEXUS would never be able to find the correspondence between two different paths. Tables 5-5 and 5-6 show a high number of successful correspondences obtained. Very few errors could be traced to an incorrect segmentation. However, the segmenter is slightly conservative, and will occasionally miss a true segment boundary.

NEXUS' ability to cope with additional noises present in speech, however, is still somewhat rudimentary. The current implementation tries to identify extraneous noises and recognize them as non-significant acoustic events when they occur in later instances of speech. This general approach seems to be a useful one, but there are an amazing number of unique clicks, pops, and hisses that can creep into a database. The real world, with its ringing telephones, doors opening, multiple talkers, etc. presents a much more complex set of non-speech events. No satisfactory solution to this problem appears to exist at the current time.

6.2 The bounds of continuity

One of the goals of NEXUS was to build a program that could accept a continuous, unbounded input, and continually recognize speech. This goal is in marked contrast to the design of most speech recognition systems, which listen for a very limited time and then begin a series of processes predicated on specific "begin" and "end" times for an utterance. NEXUS is implemented on a serial processor, and so performs in a cyclic fashion. However, with one exception, NEXUS could continuously process an unbounded waveform if a series of processors were arranged in a pipeline fashion. In the worst case, all of the signal processing, segmentation, and matching functions need look at only a local section of the waveform. Unfortunately, the final parsing algorithm requires a backtrace procedure. In order to determine the words of an utterance, some sort of break is needed. Designing a forward-looking parse is a difficult process. Consider an imaginary window on an utterance. At several points in this window, different words can be completed, and new matches are begun. Although the best matching word may be chosen at any given point, there is no guarantee that another word which has not yet completed its match will not provide a better overall score. Human word identification data (Cole & Jakamik, 1980) appear to indicate the recognition process is so accurate that only a very small number of word candidates are viable. Higher level processes undoubtedly serve to uniquely determine a single candidate in most instances. NEXUS, along with other speech recognition systems, is not very close to this level of performance.

6.3 The similarity profile

NEXUS stands or falls on its ability to perform a similarity analysis of utterances. A robust analysis procedure proved far more difficult to implement than had been expected. Although an algorithm was found that worked to a certain extent, the similarity analysis is still somewhat fragile. There are several solutions to this problem. The first solution is to permit errors to occur in the analysis, and permit backtracing to clean up the error at a later point. Unfortunately, this apparently trivial solution is nearly impossible to implement. NEXUS may not discover the error until further alterations have been made to the network, predicated on the existing errorful branches. NEXUS would have to maintain a complex analysis of the relationship between the network and the hypotheses that generated it. The only means to restore the network after the error had been detected would be to excise all branches incorporating the errorful sequences, and to replace those branches with the prior branches in the network. Storage requirements would be tremendous. Another one of the hidden constraints on NEXUS, that the system not be permitted an infinite, perfect memory, would be violated as well.

A second alternative is to allow NEXUS to query a tutor about its hypotheses. The tutor would eliminate the need for a similarity analysis altogether because the system could be told where to make network alterations. This method also seems counter to the discovery spirit of the NEXUS project, although a system with more limited or occasional feedback could be designed.

A third alternative is to use a more conservative set of parameters during the learning process, slowing down the rate of learning. Due to time and storage limitations, NEXUS is designed to learn quickly. A more cautious implementation might identify a confusion error and collect a number of instances upon which to base its confusion analysis. Again this requires a perfect "echoic memory" to store several literal instances of speech. An alternative would be to generate hypotheses on the basis of a specific instance or two, but require that confidence in the hypothesis be built up over time. Although the learning processes would undoubtedly be slowed down, their speed might appear less inhuman, as well.

6.4 Implications for feature theories

NEXUS was designed, in part, as a reaction against feature-detector based accounts of speech understanding. Although an exhausting account of the limitations of this theory has been given, I would like to briefly discuss the error classification performance of NEXUS with respect to feature analysis.

A point that is apparently not well appreciated by many psychologists is the confounding between acoustic similarity and featural similarity. Error clustering in human confusion matrices has long been used to support a featural analysis (Miller and Nicely, 1955; Shepard, 1972). It is quite striking that NEXUS generates nearly human confusion patterns even though the representation consists of only a general acoustic pattern, and the matching algorithm treats all patterns uniformly. It is unnecessary to resort to featural accounts to describe the error patterns of NEXUS. Until more is known about human adaptation to new speakers and about how to generalize NEXUS to multiple speakers, caution in generalizing this conclusion to human performance is advised. NEXUS is primarily a speaker-dependent system, and would generate unnatural confusions if a network built through experience with one speaker were used to classify another's utterances. Nevertheless, the confusion pattern is so strikingly human that I believe the suggestion should not be ignored.

6.5 Final Summary

Briefly, the contributions of the NEXUS project include: 1) an outline of a new theory of speech perception, in which supra-phonemic units form the basis of speech recognition, and the characteristics of these units are discovered through experience; 2) the construction of a working speech recognition system, based on the principles of the learning theory of speech perception; 3) identification of heuristics capable of guiding learning in a real-world perceptual domain; and 4) demonstration of a strong relationship between the acoustic characteristics of speech and human confusion patterns. I believe these results have important implications for researchers in the fields of speech perception, speech recognition, and artificial intelligence.

Projects of the scale and complexity of NEXUS have a humbling effect on their creators: one's understanding of the scope of the problem increases faster than solutions can be identified. Thereaction that "further work must be done" has accordingly achieved the exalted status of a tautological cliche. That cliche needs no amplification in the present document.

It may, however, be instructive to identify what type of work is needed most urgently. In order to build useful speech recognition systems, a better understanding of the variability of speech is needed. This variability should be assessed in the broadest possible context: differences in speakers, settings, and repetitions all need to be evaluated. Changes in the settings should include not only manipulations in phonemic context, but changes in lexical, emotional, and discourse contexts as well. Next, a better understanding of how human listeners adapt to such variability would be of tremendous value. Finally, a description of the neural mechanisms underlying speech perception would assist in the development of computer models of speech recognition. Until advances in these areas have been made, attempts to build robust recognition systems are likely to advance slowly through a space of infinite possibilities but few solutions.

Bibliography

Abramson, A. S., and Lisker, L. Discriminability along the voicing continuum: Cross-language tests. In ***Proceedings of the 6th International Congress of Phonetic Science, Prague, 1967.*** Prague: Czechoslovakia: Academia, 1970.

Ades, A. E. How selective is phonetic adaptation? Experiments on syllable position and vowel environment. ***Perception and Psychophysics***, 1974, 16, 61-66.

Bisiani, R., and Waibel, A. Performance trade-offs in search techniques for isolated word speech recognition. ***Proceedings of the 1982 International Conference on Acoustics, Speech, and Signal Processing***, Paris, France, pp. 570-573, May 1982.

Bradshaw, G. L., Cole, R. A., and Li, Z. A comparison of learning techniques in speech recognition. ***Proceedings of the 1982 International Conference on Acoustics, Speech, and Signal Processing***, Paris, France, pp. 554-557, May 1982.

Chomsky, N., and Halle, M. ***The sound pattern of English.*** New York: Harper and Row, 1968.

Clark, H. H., and Clark, E. V. ***Psychology and Language. An introduction to Psycholinguistics.*** New York: Harcourt Brace Jovanovich, Inc., 1977.

Cole, R. A., and Cooper, W. E. Perception of voicing in English affricates and fricatives. ***Journal of the Acoustical Society of America***, 1975, 58, 1280-1287.

Cole, R. A., and Jakimik, J. A model of speech perception. In R. A. Cole (Ed.), ***Perception and Production of Fluent Speech.*** Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Cole, R. A., Rudnicky, A. I., Zue, V. W., and Reddy, D. R. Speech as patterns on paper. In R. A. Cole (Ed.), ***Perception and Production of Fluent Speech.*** Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. Some experiments on the perception of synthetic speech sounds. ***Journal of the Acoustical Society of America***, 1952, 24, 597-606.

Cooper, W. E. Adaptation of phonetic feature analyzers for place of articulation. ***Journal of the Acoustical Society of America***, 1974, 56, 617-627.

Cooper, W. E. Selective adaptation to speech. In F. Restle, R. Shiffrin, N. Castellan, H. Lindman, and D. B. Pisoni (eds.), ***Cognitive Theory. Volume 1.*** Hillsdale, N. J.: Lawrence Erlbaum Associates, 1975.

Cooper, W. E., Ebert, R. R., and Cole, R. A. Perceptual analysis of stop consonants and glides. ***Journal of Experimental Psychology: Human Perception and Performance***, 1976, 2, 92-104.

Denes, P. B. Effects of duration on the perception of voicing. ***Journal of the Acoustical Society of America***, 1955, 27, 761-764.

Diehl, R. L. Feature analyzers for the phonetic dimension *stop vs. continuant*. ***Perception and Psychophysics***, 1976, 19, 267-272.

Diehl, R. L. Feature detectors for speech: A critical reappraisal. ***Psychological Bulletin***, 1981, 89, 1-18.

Dixon, N. R., and Silverman, H. F. The 1976 Modular Acoustic Processor (MAP). ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, 1977, ASSP-25, 367-379.

Eilers, R., and Minifie, F. Fricative discrimination in early infancy. ***Journal of Speech and Hearing Research***, 1975, 18, 158-167.

Eimas, P. D. Auditory and linguistic processing of cues for place of articulation by infants. ***Perception and Psychophy...*** 1974, 16, 513-521.

Eimas, P. D. Auditory and Phonetic coding of the cues for speech: Discrimination of ther-l distinction by young infants. ***Perception and Psychophysics***, 1975, 18, 341-347.

Eimas, P. D., Cooper, W. E., and Corbit, J. D. Some properties of linguistic feature detectors. ***Perception and Psychophysics***, 1973, 13, 247-252.

Eimas, P. D., and Corbit, J. D. Selective adaptation of linguistic feature detectors. ***Cognitive Psychology***, 1973, 4, 99-109.

Eimas, P. D. Siqueland, E. R., Jusczyk, P. and Vigorito, J. Speech perception in infants. ***Science***, 1971, 171, 303-306.

Erman, L., Hayes-Roth, F., Lesser, V., and Reddy, D. R. The Hearsay-II speech understanding system: Integrating knowledge to resolve uncertainty. ***Computing Surveys***, 1980, 12, 213-253.

Fitch, H. L., Halwes, T., Erickson, D. M., and Liberman, A. M. Perceptual equivalence of two acoustic cues for stop-consonant manner. ***Perception and Psychophysics***, 1980, 27, 343-350.

Foulke, E., and Sticht, T. Review of research on the intelligibility and comprehension of accelerated speech. ***Psychological Bulletin***, 1969, 72, 50-62.

Frishkopf, L. S., and Goldstein, M. H. Responses to acoustic stimuli from single units in the eighth nerve of the bullfrog. ***Journal of the Acoustical Society of America***, 1963, 65, 1219-1228.

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. The identification and discrimination of synthetic vowels. ***Language and Speech***, 1962, 5, 171-189.

Funk and Wagnalls New Standard Dictionary of the English Language. New York: Funk and Wagnalls Company, 1957.

Ganong, W. F. Phonetic categorization in auditory word perception. ***Journal of Experimental Psychology: Human Perception and Performance***, 1980, 6, 110-125.

Halle, M., and Stevens, K. N. Speech recognition: A model and a program for research. In J. A. Fodor and J. J. Katz (Eds.), ***The structure of language***. Englewood Cliffs, N.J.: Prentice Hall, 1972.

Harris, C. M. A study of the building blocks in speech. ***Journal of the Acoustical society of America***, 1953, 25, 962-969.

Harris, K. S. Cues for the discrimination of American English fricatives in spoken syllables. ***Language and Speech***, 1958, 1, 1-7.

Hughes, G. W., and Halle, M. Spectral properties of fricative consonants. ***Journal of the Acoustical Society of America***, 1956, 28, 303-310.

Itakura, F., Minimum prediction residual principle applied to speech recognition. ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, 1975, ASSP-26, 67-72.

Jacobsen, R., and Halle, M. ***Fundamentals of language***. The Hague, Netherlands: Mouton, 1956.

Jacobson, R., Fant, G., and Halle, M. ***Preliminaries to speech analysis: The distinctive features and their correlates***. Cambridge, Mass.: MIT Press, 1963.

Lenneberg, E. H. Understanding language without ability to speak: A case report. ***Journal of Abnormal and Social Psychology***, 1962, 65, 419-425.

Lesser, V. R., Fennell, R. D., Erman, L. D., and Reddy, D. R. Organization of the Hearsay II speech understanding system. ***IEEE Transactions on Acoustics, Speech, and signal Processing***, 1975, ASSP-23, 11-24.

Lettvin, J. Y., Maturana, H. R., McColloch, W. S., and Pitts, W. H. What the frog's eye tells the frog's brain. ***Proceedings of the Institute of Radio Engineers***, 1959, 47, 1940-1951.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. Perception of the speech code. ***Psychological Review***, 1967, 74, 431-461.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. The role of selected stimulus variables in theperception of the unvoiced stop consonants. ***American Journal of Psychology***, 1952, 65, 497-516.

Liberman, A. M., Delattre, P. C., Cooper, F. S. and Gerstman, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. ***Psychological Monographs***, 1954, 68, 1-13.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. The discrimination of speech sounds within and across phoneme boundaries. ***Journal of Experimental Psychology***, 1957, 54, 358-368.

Liberman, A. M., Harris, K. S., Kinney;, J. A., and Lane, H. The discrimination of relative onset time of the components of certain speech and nonspeech patterns. ***Journal of Experimental Psychology***, 1961, 61, 379-388.

Liberman, A. M., and Studdert-Kennedy, M. Phonetic perception. In R. Held, H. Leibowitz, and H. L. Teuber (Eds.), ***Handbook of sensory physiology, Vol. VIII.*** Heidelberg: Springer-Verlag, 1977.

Lisker, L., and Abramson, A. S. The voicing dimension: Some experiments in comparative phonetics. In ***Proceedings of the 6th International Congress of Phonetic Science, Prague, 1967.*** Prague: Czechoslovakia: Academia, 1970.

Lotz, J., Abramson, A. S., Gerstman, L. J., Ingeman, F., and Nemser, W. J. The perception of English stops by speakers of English, Spanish, Hungarian, and Thai: A tape cutting experiment. ***Language and Speech***, 1960, 3, 71-77.

Lowerre, B. T. Dynamic speaker adaptation in the Harpy speech recognition system. ***Proceedings of the 1977 International Conference on Acoustics, Speech, and Signal Processing***, pp. 788-790.

Lowerre, B. T., and Reddy, D. R. The Harpy speech understanding system. In Wayne A. Lea (Ed.), ***Trends in Speech Recognition.*** Englewood Cliffs, New Jersey: Prentice-Hall, 1979.

MacDonald, J., and McGurk, H. Visual influences on speech perception processing. ***Perception and Psychophysics***, 1978, 24, 253-257.

MacNeilage, P. F. Motor control of serial ordering of speech. ***Psychological Review***, 1970, 77, 182-196.

Martin, J. G., and Bunnell, H. T. Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. ***Journal of Experimental Psychology: Human Perception and Performance***, 1982, 8, 473-488.

McGurk, H., and MacDonald, J. Hearing lips and seeing voices. ***Nature***. 1976, 264, 746-748.

Miller, C., and Morse, P. A. The "heart" of categorical speech discrimination in young infants. ***Journal of Speech and Hearing Research***, 1976, 19, 578-589.

Miller, G. A., and Nicely, P. An analysis of perceptual confusions among some English consonants. ***Journal of the Acoustical Society of America***, 1955, 27, 338-352.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., and Fujimura, O. An effect of linguistic experiences. The discrimination of [r] and [l] by native speakers of Japanese and English. ***Perception and Psychophysics***, 1975, 18, 331-340.

Morse, P. A., Kass, J. E., and Turkienicz, R. Selective adaptation of vowels. ***Perception and Psychophysics***, 1976, 19, 137-143.

Myers, C. S., and Levinson, S. E. Speaker independent connected word recognition using a syntax-directed dynamic programming procedure. ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, 1982, ASSP-30, 561-565.

Nadas, A. Estimation of Probabilities in the language model of the IBM speech recognition system. ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, 1984, ASSP-32, 859-861.

Oden, G. C., and Massaro, D. W. Integration of featural information in speech perception. ***Psychological Review***, 1978, 85, 172-191.

Peterson, G. E., and Barney, H. L. Control methods used in a study of the vowels. ***Journal of the Acoustical Society of America***, 1952, 24, 175-184.

Peterson, G. E., Wang, W. S., and Silvertsen, E. Segmentation techniques in speech synthesis. ***Journal of the Acoustical Society of America***, 1958, 30, 739-742.

Pisoni, D. B. Some remarks on the perception of speech and nonspeech signals. In E. Fisher-Jorgensen and T. Thorsen (Eds.), ***Proceedings of the Ninth International Congress of Phonetic Sciences (Vol. 3).*** Copenhagen: Institute of Phonetics, 1980.

Pisoni, D. B., Aslin, R. N., Perey, A. J., and Hennessy, B. L. Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. ***Journal of Experimental Psychology: Human Perception and Performance***, 1982, 8, 297-314.

Pisoni, D. B., and Sawusch, J. R. Some stages of processing in speech perception. In A. Cohen and S. G. Nooteboom (Eds.), ***Structure and Process in Speech Perception.*** Heidelberg, W. Germany: Springer-Verlag, 1975.

Rabiner, L. R. On creating reference templates for speaker independent recognition of isolated words. ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, 1978, ASSP-26, 34-42.

Rabiner, L. R., Levinson, S. E., Rosenberg, A. E., and Wilpon, J. G. Speaker-independent recognition of isolated words using clustering techniques. ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, 1979, ASSP-27, 336-349.

Rabiner, L. R., Rosenberg, A. E., and Levinson, S. E. Considerations in dynamic time warping algorithms for discrete word recognition. ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, 1978, ASSp-26, 575-582.

Repp, B. H. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. ***Psychological Bulletin***, 1982, 92, 81-110.

Roberts, M., and Summerfield, Q. Auidiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. ***Perception and Psychophysics***, 1981, 30, 309-314.

Sakoe, H. Two-Level DP-matching A dynamic programming-based pattern matching algorithm for connected word recognition. ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, 1979, ASSP-27, 588-595.

Sakoe, H. and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, 1978, ASSP-26, 43-49.

Samuel, A. G., and Newport, E. L. Adaptation of speech by nonspeech: Evidence for complex acoustic cue detectors. ***Journal of Experimental Psychology: Human Perception and Performance***, 1979, 5, 563-578.

Sawusch, J. R., and Jusczyk, P. W. Adaptation and contrast in the perception of voicing. ***Journal of Experimental Psychology: Human perception and performance***, 1981, 7, 408-421.

Shepard, R. N. Psychological representation of speech sounds. In E. E. David, Jr., and P. B. Denes (Eds.), ***Human Communication: A unified view*** New York: McGraw Hill, 1372.

Simon, C., and Fourchin, A. J. Cross-language study of speech-pattern learning. ***Journal of the Acoustical Society of America***, 1978, 63, 925-935.

Singh, S. Cross language study of perceptual confusion of plosive phonemes in two conditions of distortion. ***Journal of the Acoustical Society of America***, 1966, 40, 635-656.

Singh, S., and Black, J. W. Study of twenty-six intervocalic consonants as spoken and recognized by four language groups. ***Journal of the Acoustical Society of America***, 1966, 39, 372-387.

Snyder, R. T., and Pope, P. New norms for and an item analysis of the Wepman test at the 1st grade 6-year level. ***Perceptual and Motor Skills***, 1970, 31, 1007-1010.

Stevens, K. N. Toward a model for speech recognition. ***Journal of the Acoustical Society of America***, 1960, 32, 47-55.

Stevens, K. N. Perception of phonetic segments: Evidence from phonology, acoustics, and psychoacoustics. In D. L. Horton and J. J. Jenkins (Eds.), ***The perception of language***. Columbus, Ohio: Charles E. Merrill, 1971.

Stevens, K. N. The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David and P. B. Denes (Eds.), ***Human communication: A unified view***. New York: McGraw Hill, 1972.

Stevens, K. N. Potential role of property detectors in the perception of consonants. In G. Fant and M. A. A. Tathan (Eds.), ***Auditory Analysis and Perception of Speech***. New York: Academic Press, 1975.

Stevens, K. N., and Blumstein, S. E. Invariant cues for place of articulation in stop consonants. ***Journal of the Acoustical Society of America***, 1978, 64, 1358-1368.

Stevens, K. N., and Blumstein, S. E. The search for invariant acoustic correlates of phonetic features. In P. D. Eimas and J. L. Miller (Eds.), ***Perspectives on the study of speech***. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., 1981.

Stevens, K. N., and Halle, M. Remarks on analysis by synthesis and distinctive features. In W. Wather-Dunn (Ed.), ***Models for the Perception of Speech and Visual Form***. Cambridge, Mass.: MIT Press, 1967.

Stevens, K. N. and House, A. S. Speech perception. In J. V. Tobias (Ed.), ***Foundations of Auditory Theory, Vol. II***. New York: Academic Press, 1972.

Studdert-Kennedy, M. The perception of speech. In T. A. Sebeok (Ed.), ***Current Trends in Linguistics (Volume 12)***. The Hague, Netherlands: Mouton, 1974.

Studdert-Kennedy, M. Speech perception. In N. J. Lass (Ed.), ***Contemporary Issues in Experimental Phonetics***. Springfield, Illinois: G. D. Thomas, 1976.

Summerfield, A. Q. How a full account of segmental perception depends on prosody and vice versa. In A. Cohen and S. G. Nooteboom, ***Structure and Process in Speech Perception.*** New York: Springer-Verlag, 1975.

Templen, M. C. ***Certain Language Skills in Children.*** Minneapolis, Minnesota: University of Minnesota Press, 1957.

Wang, M. D., and Bilger, R. C. Consonant confusions in noise: A study of perceptual features. ***Journal of the Acoustical Society of America***, 1973, 54, 1248-1266.

Warren, R. M., Obusek, C. J., Farmer, R. M., and Warren, R. P. Auditory sequence: Confusion of patterns other than speech and music. ***Science***, 1969, 164, 586-587.

Wickelgren, W. A. Context-sensitive coding, associative memory, and serial order in (speech) behavior. ***Psychological Review***, 1969, 76, 1-15.

Williams, L. The perception of stop consonant voicing by Spanish-English bilinguals. ***Perception and Psychophysics***, 1977, 21, 289-297.

Williams, L. The modification of speech perception and production in second-language learning. ***Perception and Psychophysics***, 1979, 26, 95-104.

Wolf, J. J., and Woods, W. A. The HWIM speech understanding system. ***Proceedings of the 1977 International Conference on Acoustics, Speech, and Signal Processing***, pp 784-787.

Zwicker, E. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). ***Journal of the Acoustical Society of America***, 1961, 33, 248.

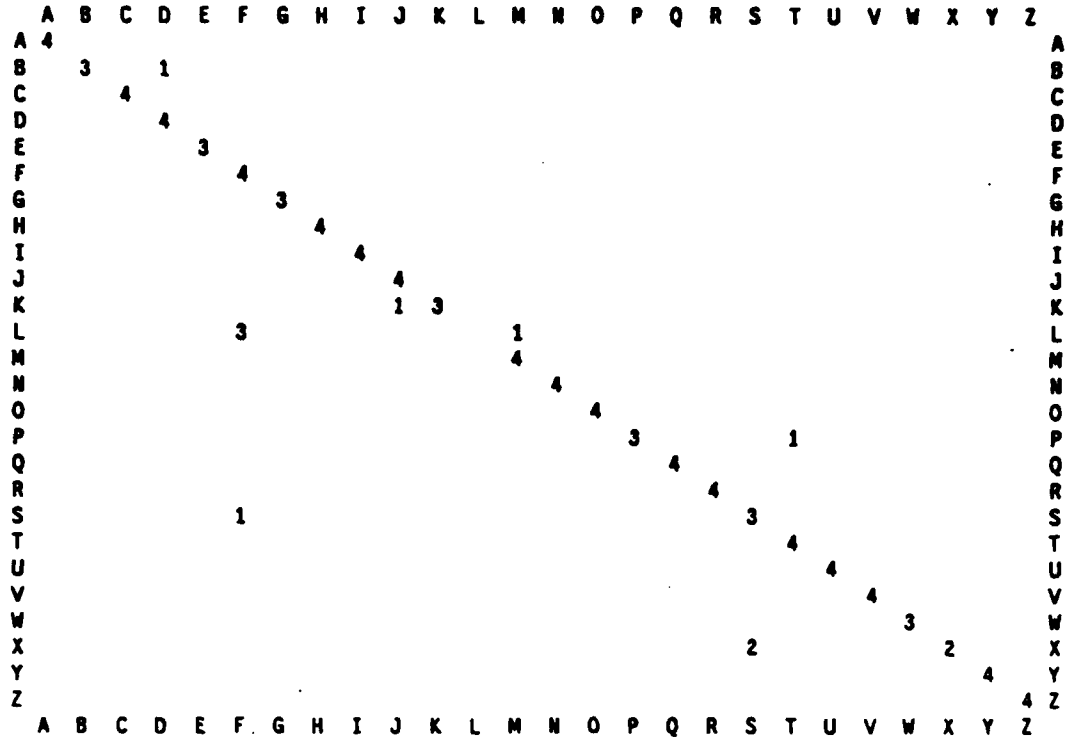
I. Alphabet Confusion Matrices

Confusion matrices for each speaker are summarized in five-trial blocks. Within each matrix, actual stimuli are presented as rows in the matrix, while recognized stimuli form the columns of the matrix.

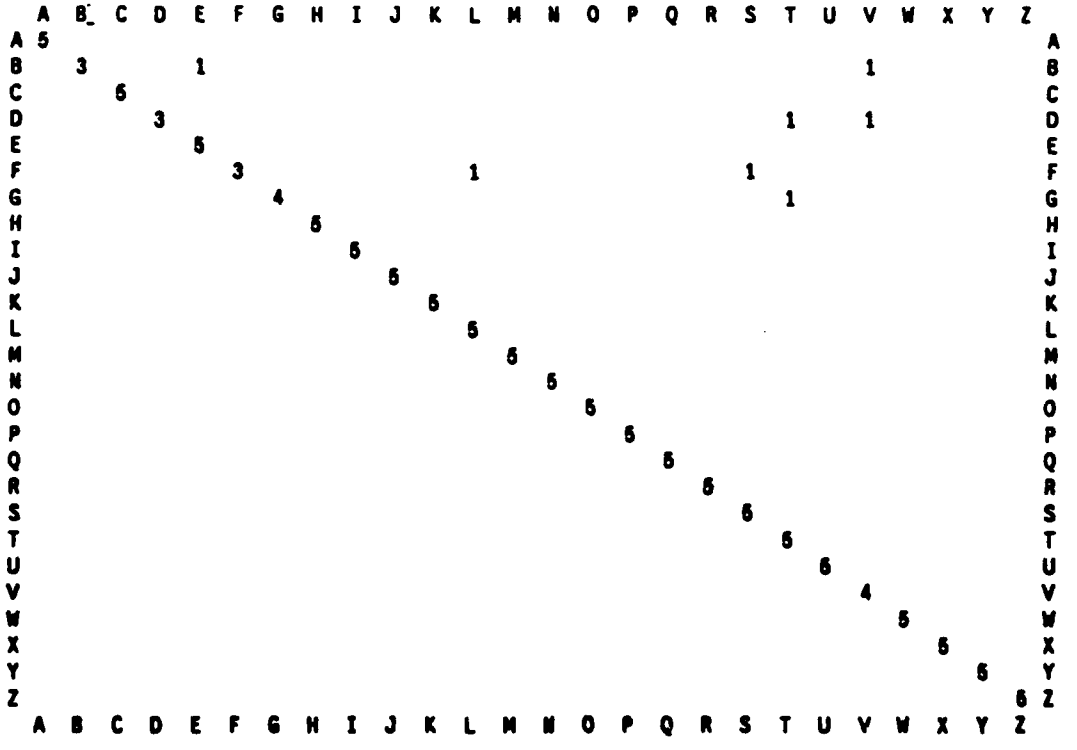
The first matrix for each speaker presents only four trials, trials 2 – 5, while remaining matrices present full five-trial blocks. Trial 1 data is omitted because "confusions" made by the system reflect a complete lack of experience with the full vocabulary.

Row totals may not always add up to the number of trials in a matrix. This occurs because the remote file server failed to obtain a speech utterance from the database kept on a different computer.

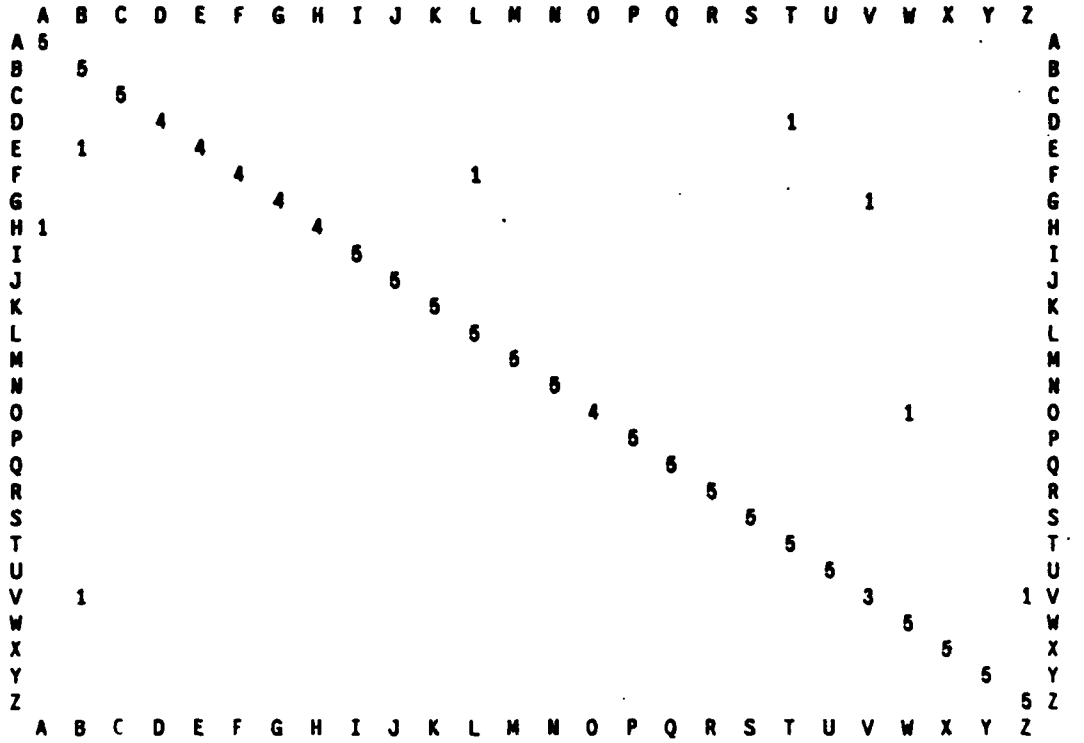
Confusion Matrix
Speaker MGB
Trials 2 - 5



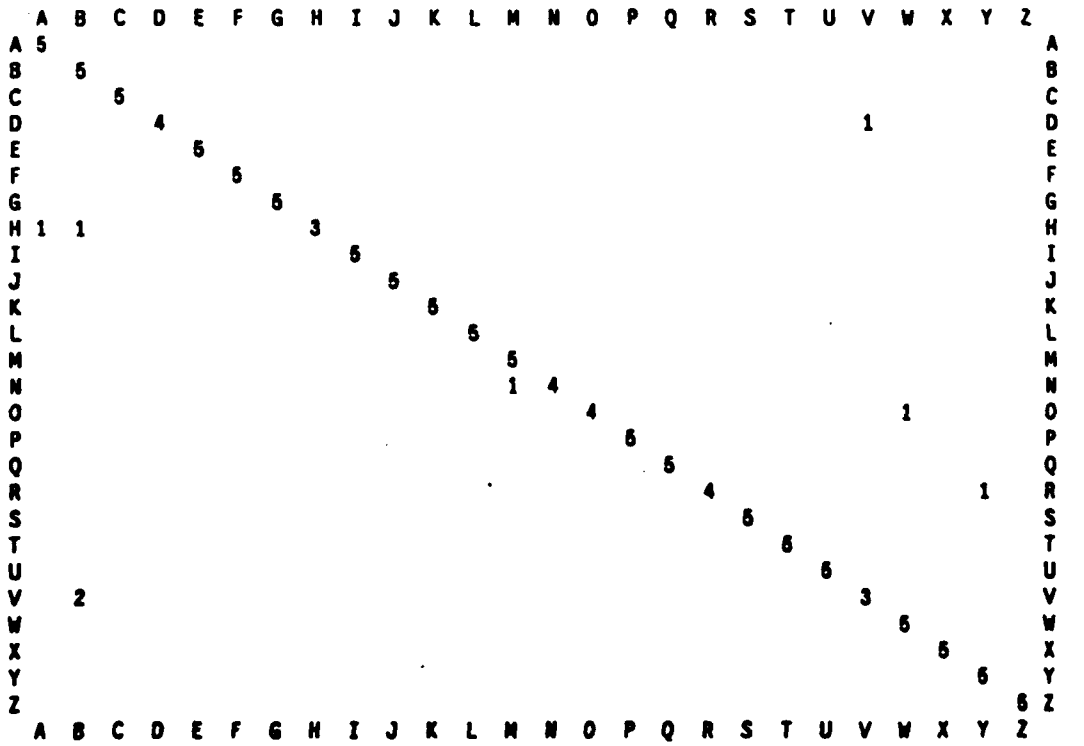
Confusion Matrix
Speaker MGB
Trials 6 - 10



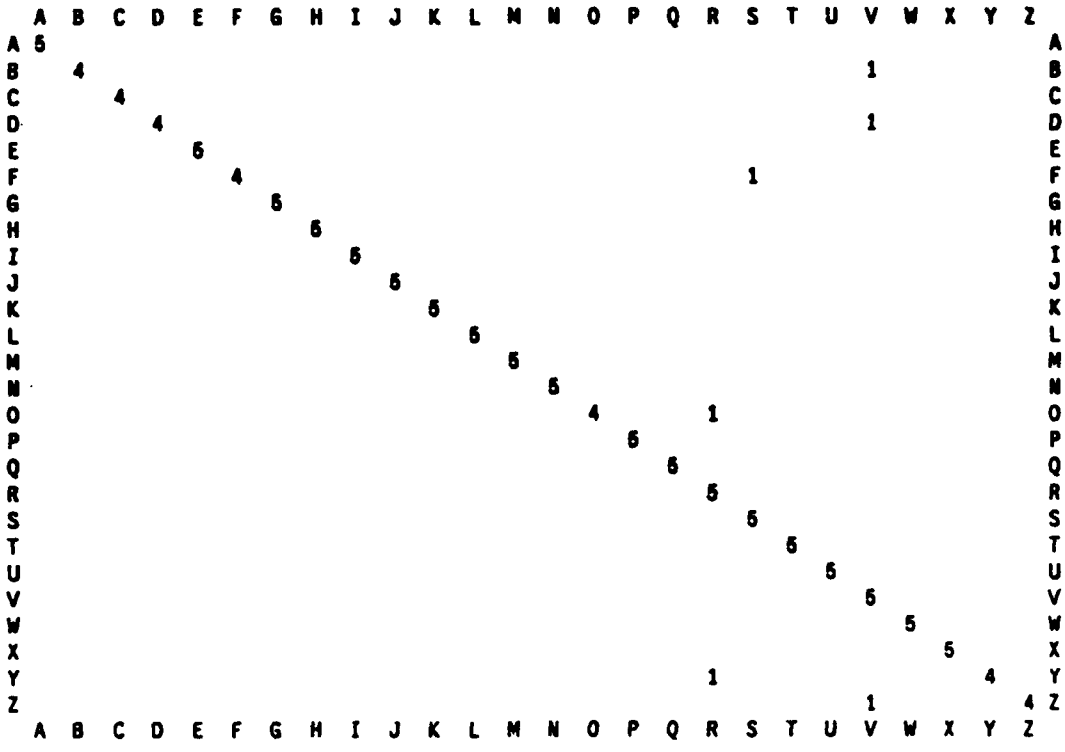
Confusion Matrix
Speaker MGB
Trials 11-15



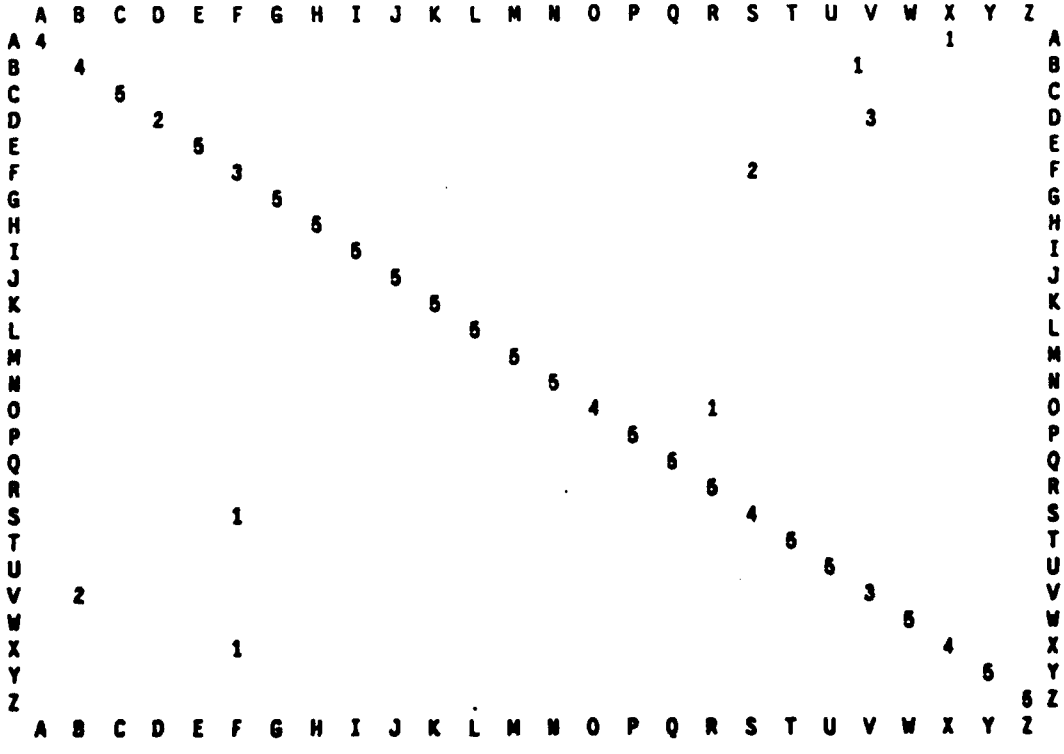
Confusion Matrix
Speaker MGB
Trials 16-20



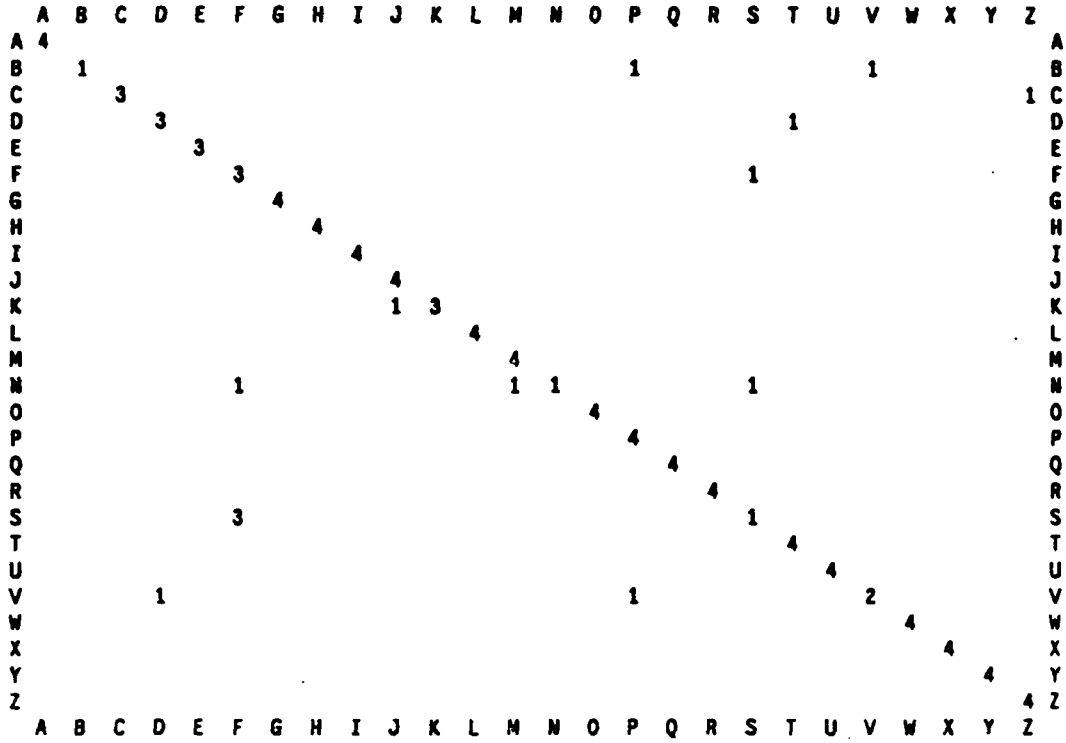
Confusion Matrix
Speaker MGB
Trials 21-25



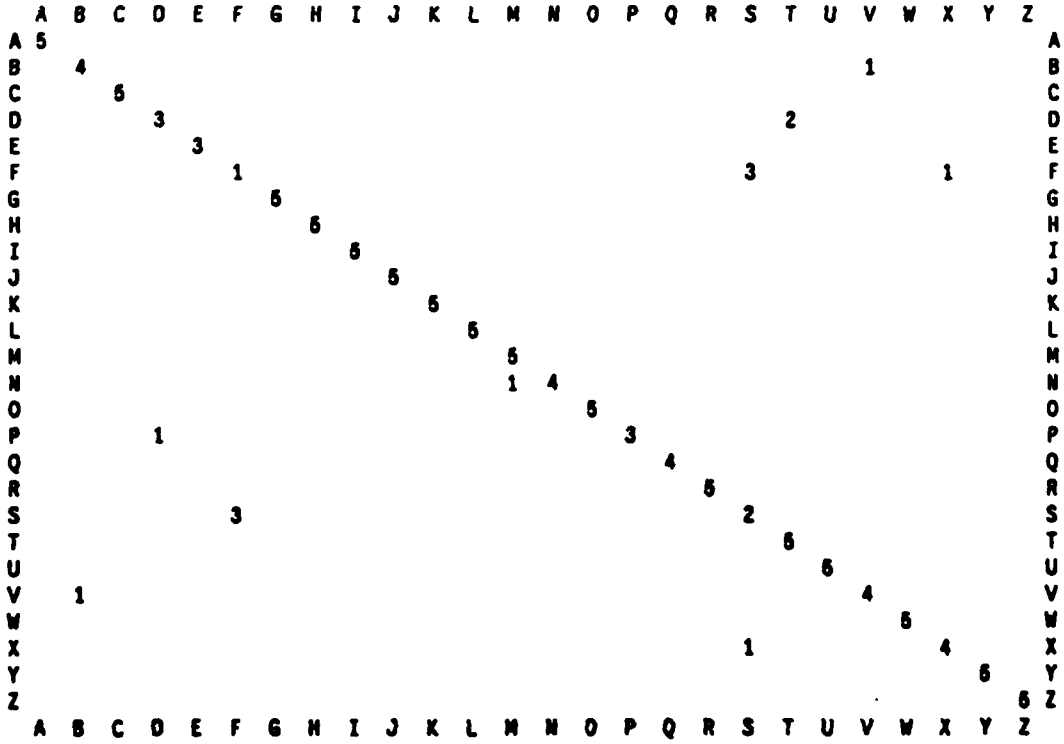
Confusion Matrix
Speaker MGB
Trials 26-30



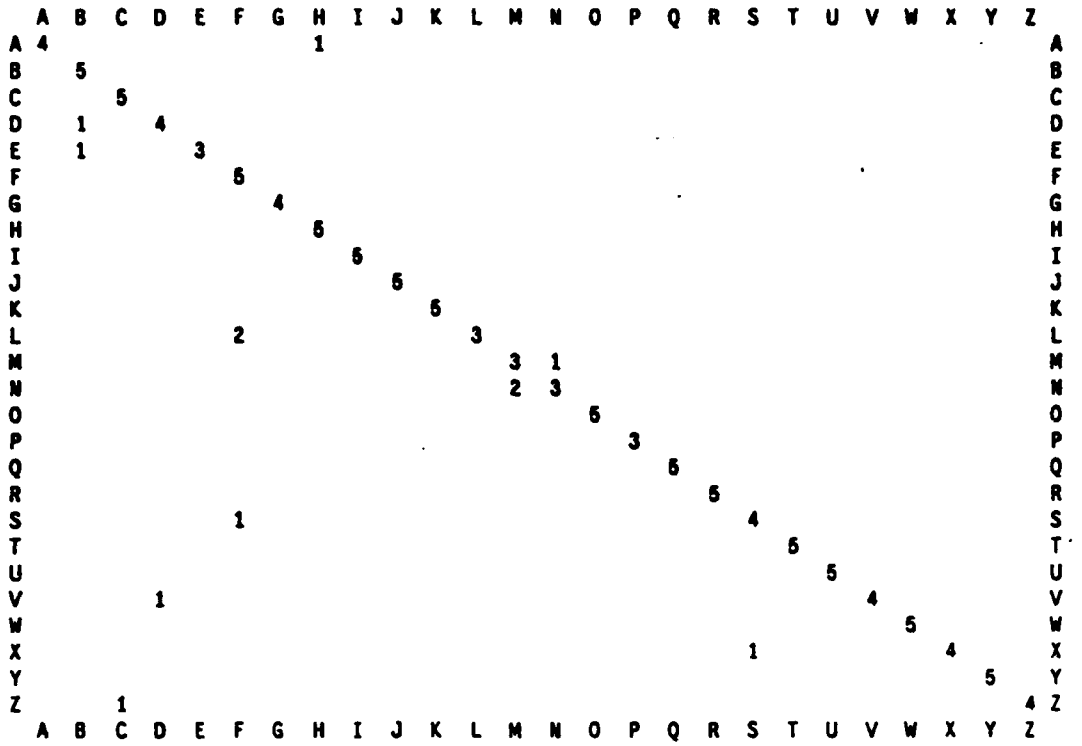
Confusion Matrix
Speaker MRG
Trials 2-5



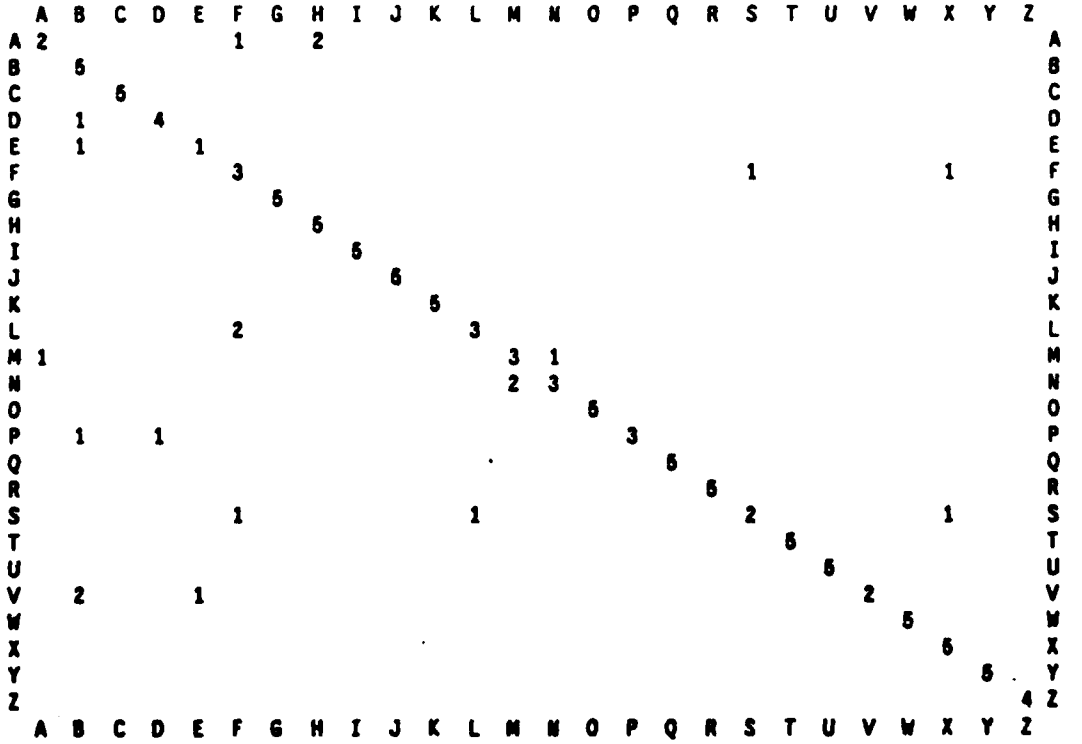
Confusion Matrix
Speaker MRG
Trials 6-10



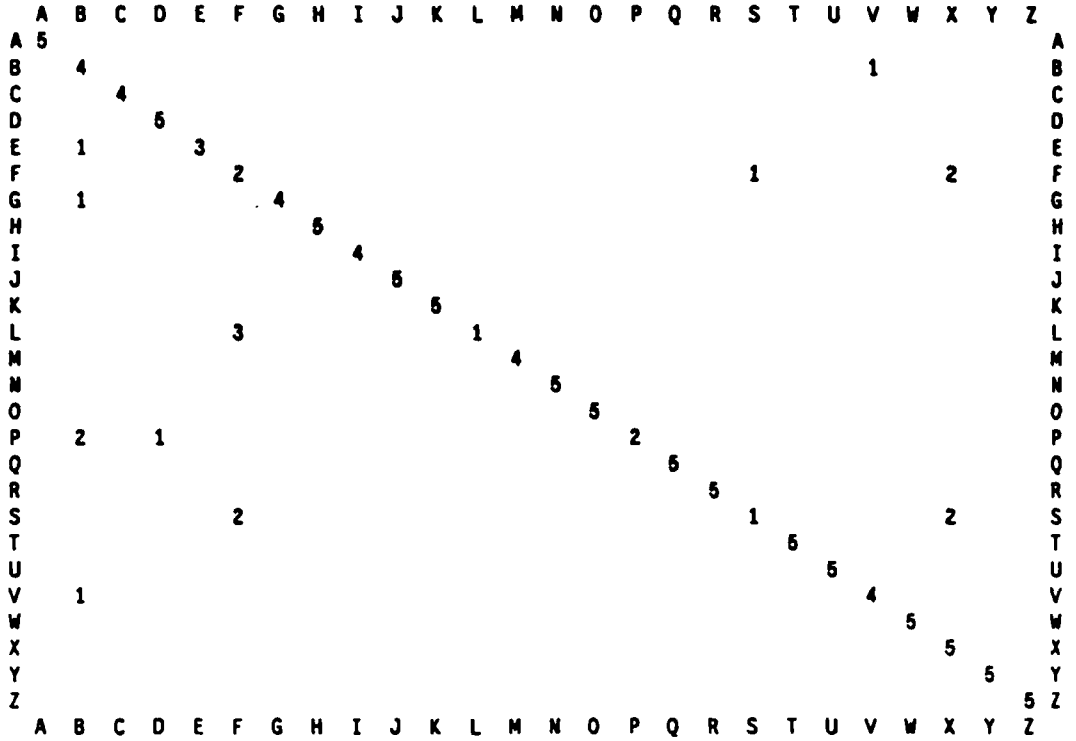
Confusion Matrix
Speaker MRG
Trials 11-15



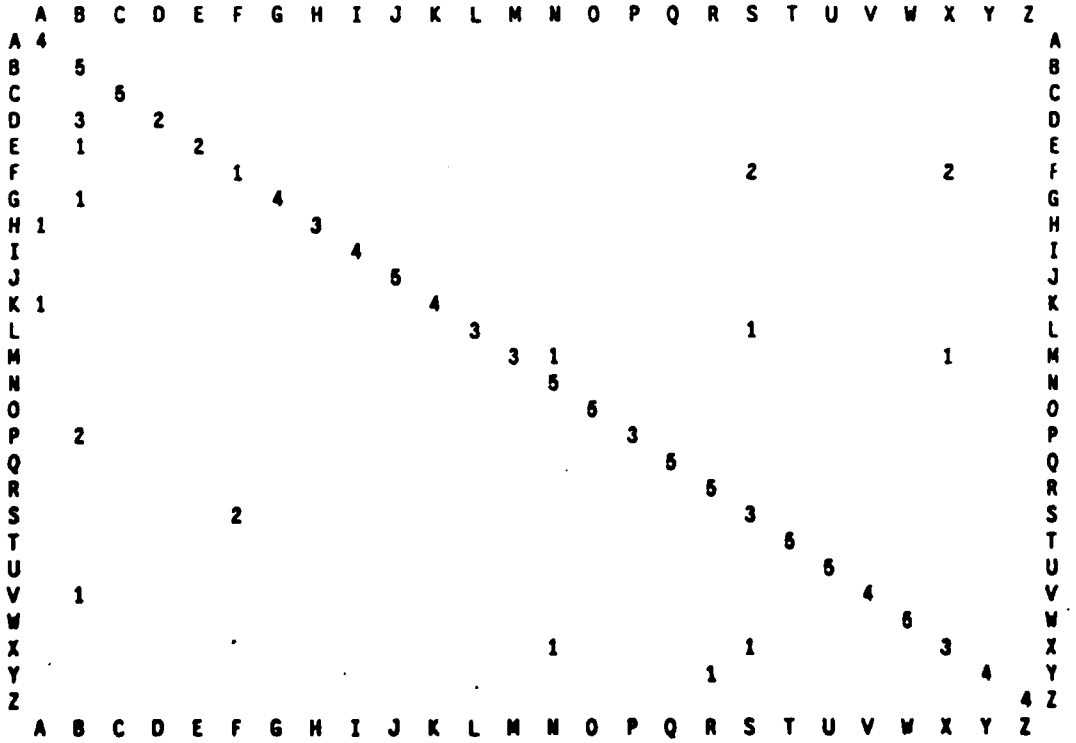
Confusion Matrix
Speaker MRG
Trials 16-20



Confusion Matrix
Speaker MRG
Trials 21-25



Confusion Matrix
Speaker MRG
Trials 26-30



II. E-set network development

The figures included in this appendix represent the development of the speech network when Nexus was presented with the "E-set" vocabulary for one speaker (mgb). Each figure represents the state of the network after the completion of one recognition cycle. Occasionally the network did not change during a recognition cycle, so a single diagram will be presented for two recognition cycles.

As NEXUS learns, independent paths are allowed to share the same nodes. The mechanism incorporated to implement sharing is not a folding process, where words share the same structure, but a pointing process, so that two paths may point to the same node. The graphic representation of this process is somewhat misleading. In Figure II-2, for example, a /b/ - /i/ path has been "folded" with a /d/ - /i/ - /h/ path. NEXUS does not assume that /b/ - /i/ - /h/ is a legitimate path for the word "B", nor does it assume that /d/ - /i/ is a legitimate path for the word "D". Instead, the system uses the same /i/ node in both existing paths. In Figure II-3, NEXUS has discovered a /d/ - /i/ path, and so generalizes the representation of the word "D" only. Each of these figures should be interpreted as simple illustrations of the type of sharing discovered by NEXUS, and not representative of the fine details of the created network structure.

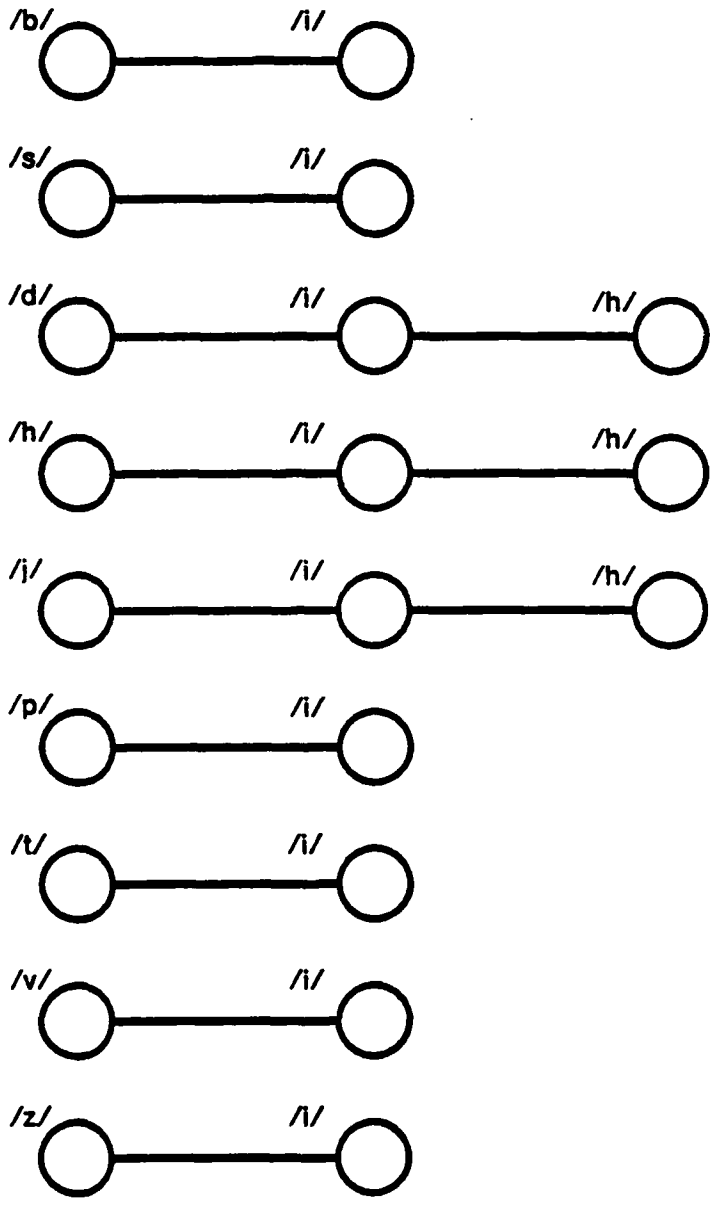


Figure II-1: E-set Network: Cycle 1

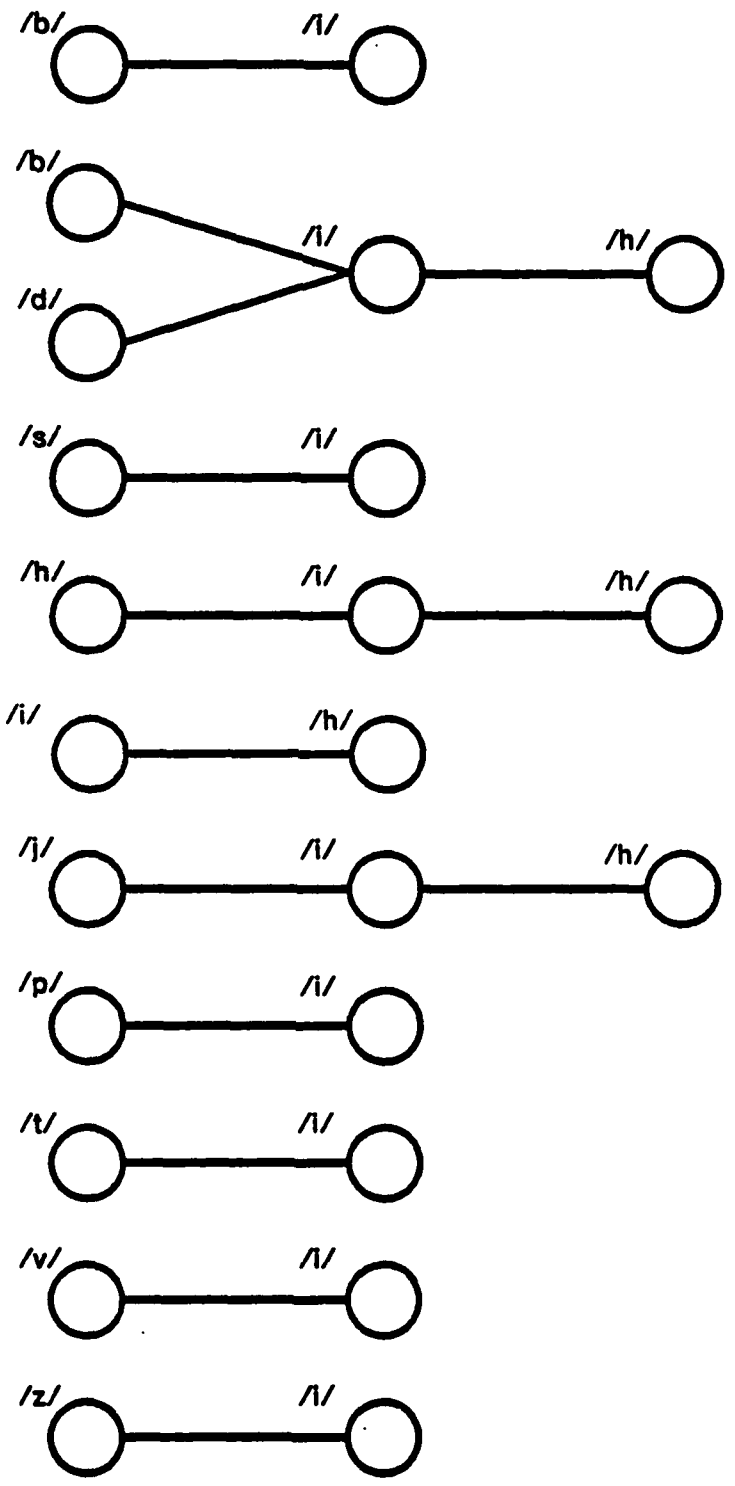


Figure II-2: E-set Network: Cycles 2 and 3

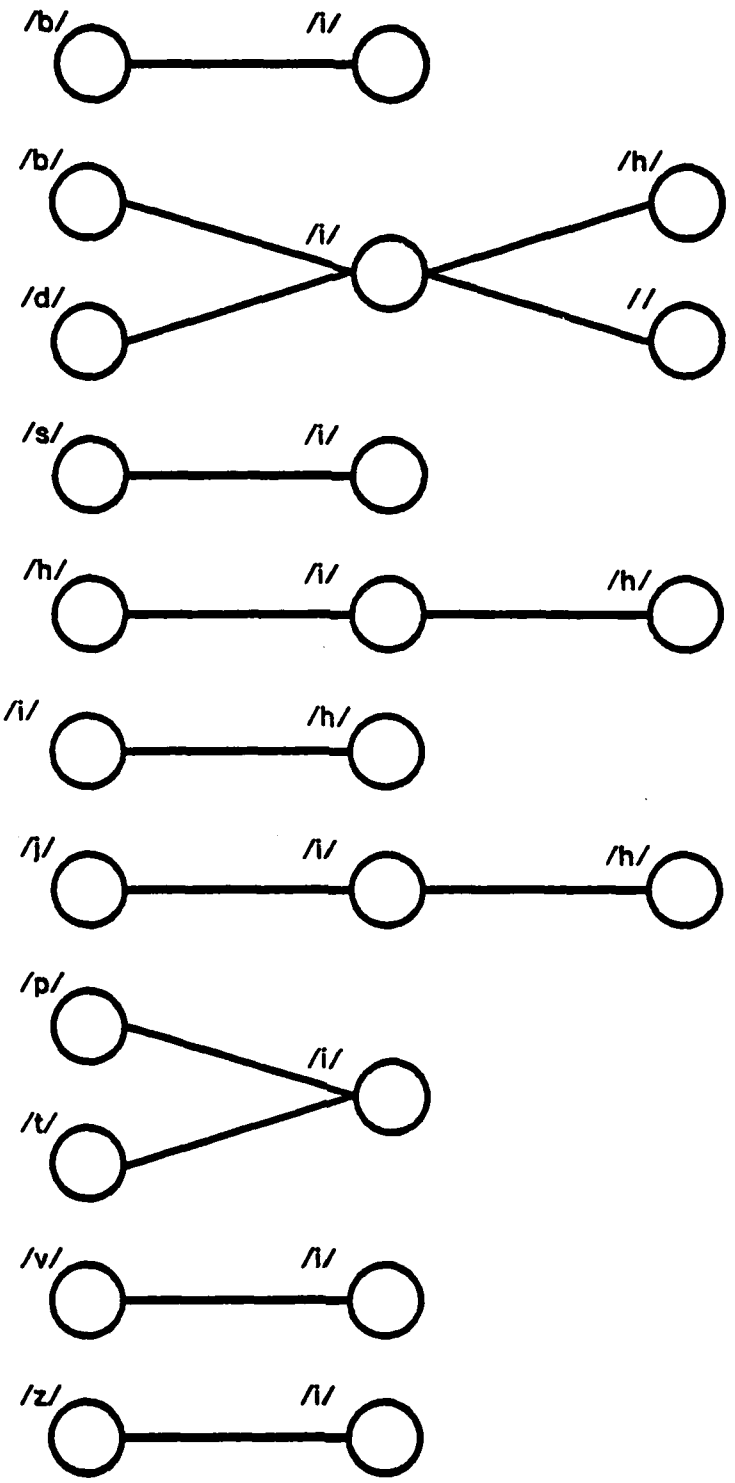


Figure II-3: E-set Network: Cycle 4

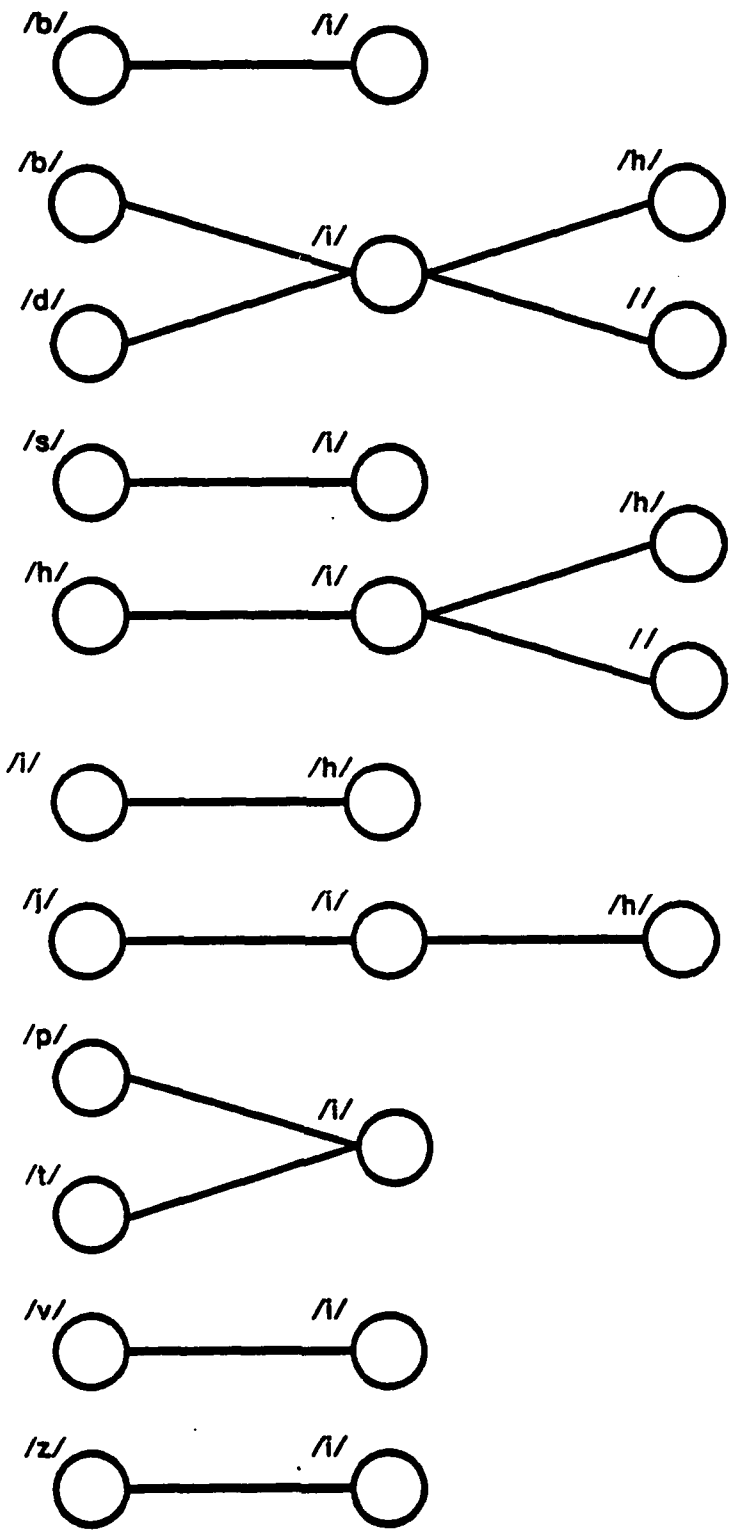


Figure II-4: E-set Network: Cycle 5

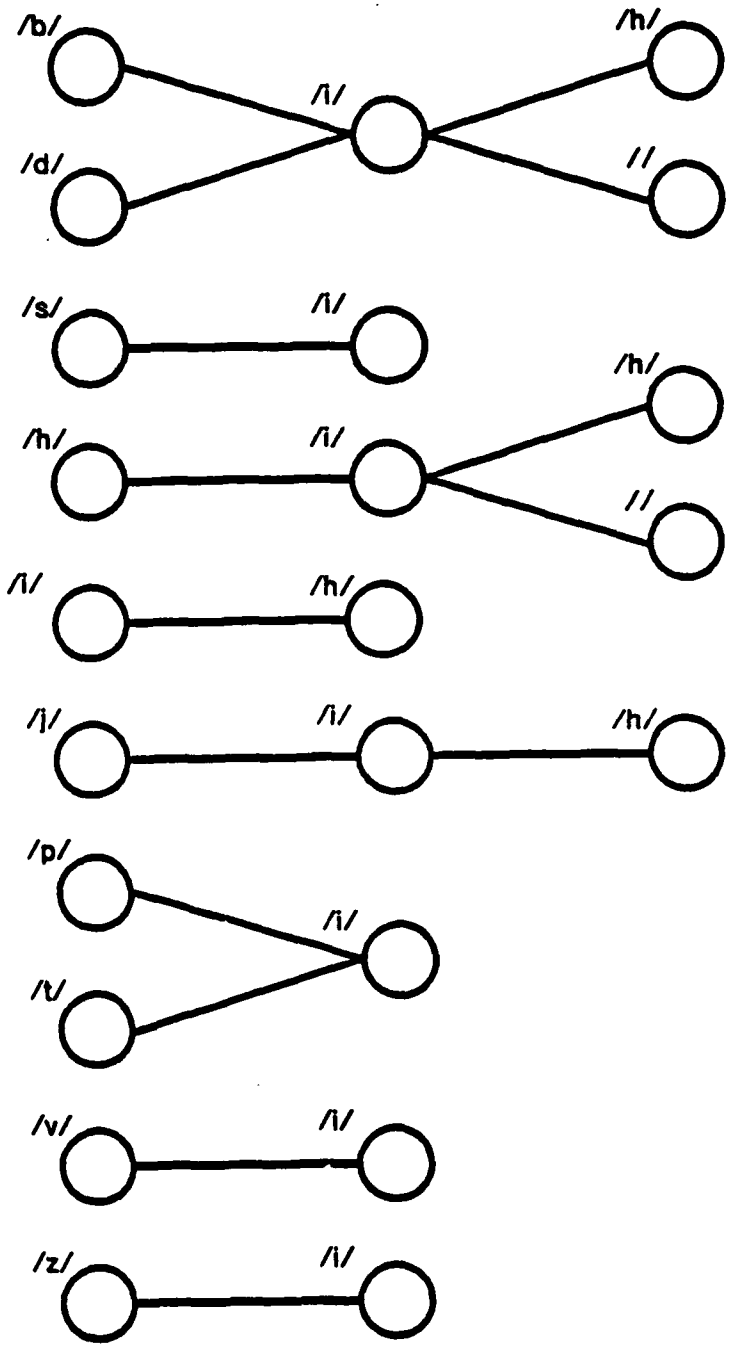


Figure II-5: E-set Network: Cycle 6

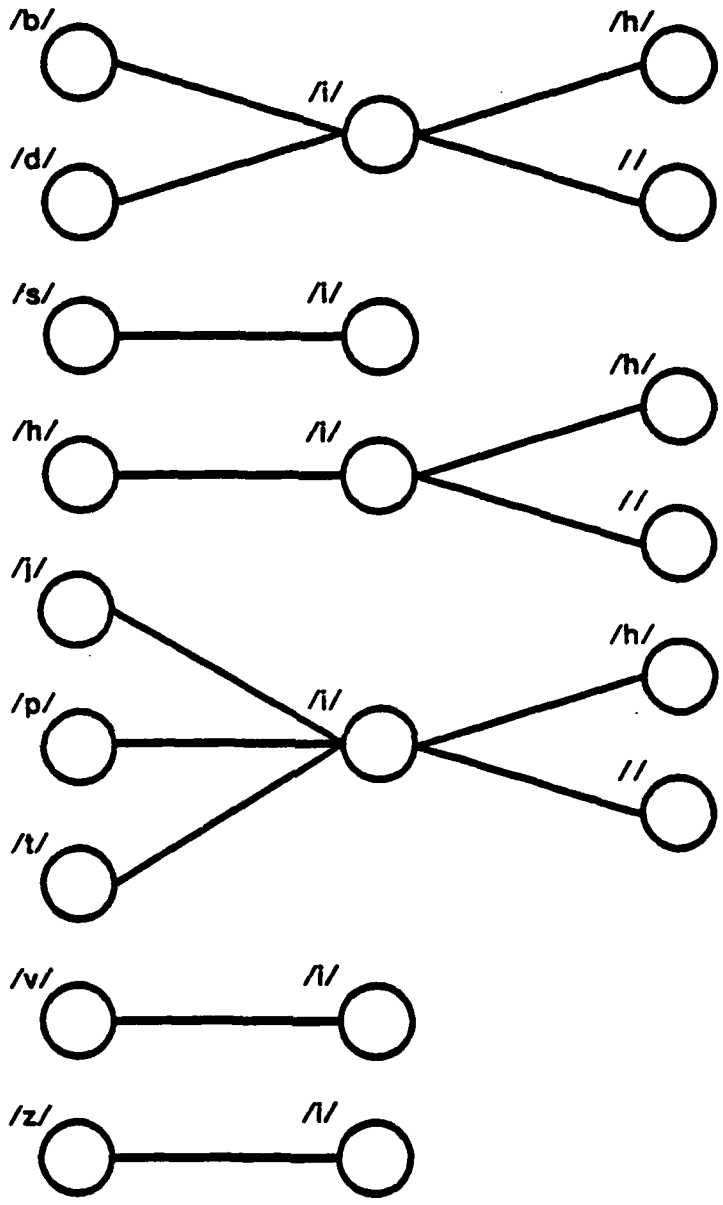


Figure II-6: E-set Network: Cycle 7

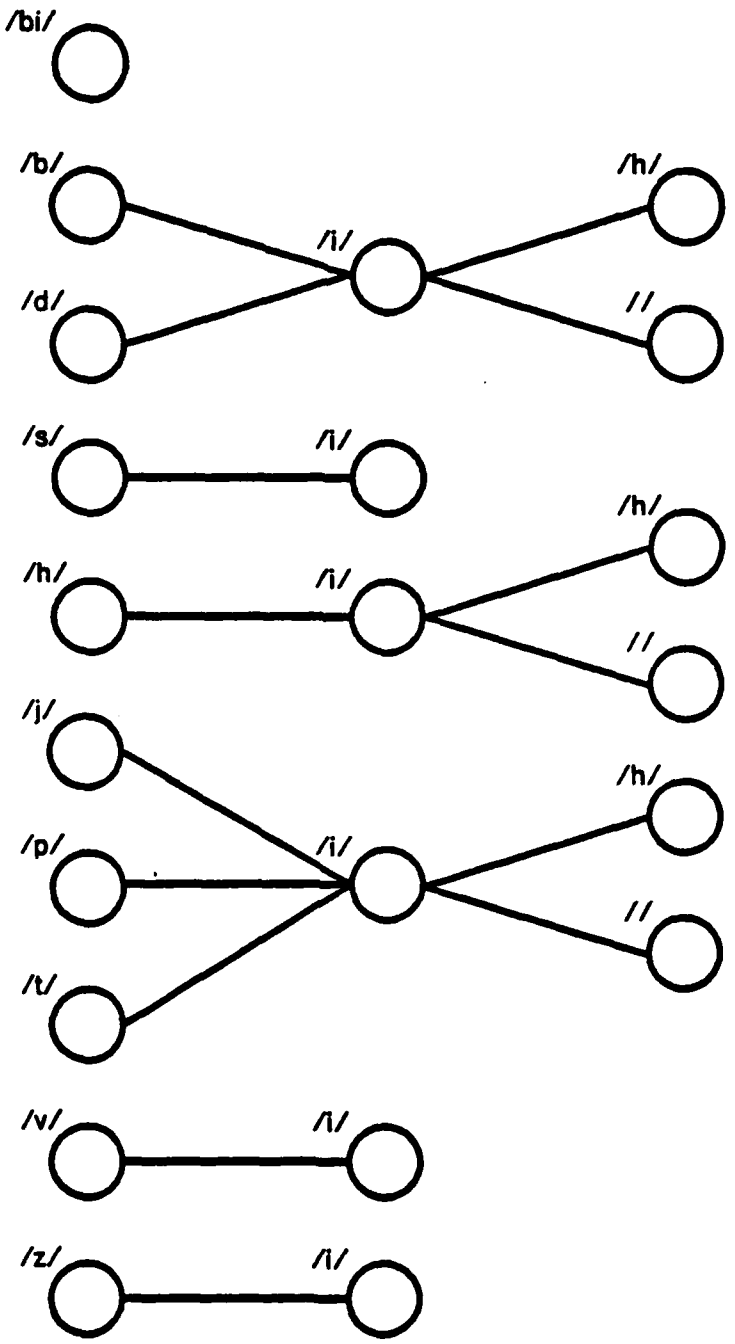


Figure II-7: E-set Network: Cycle 8

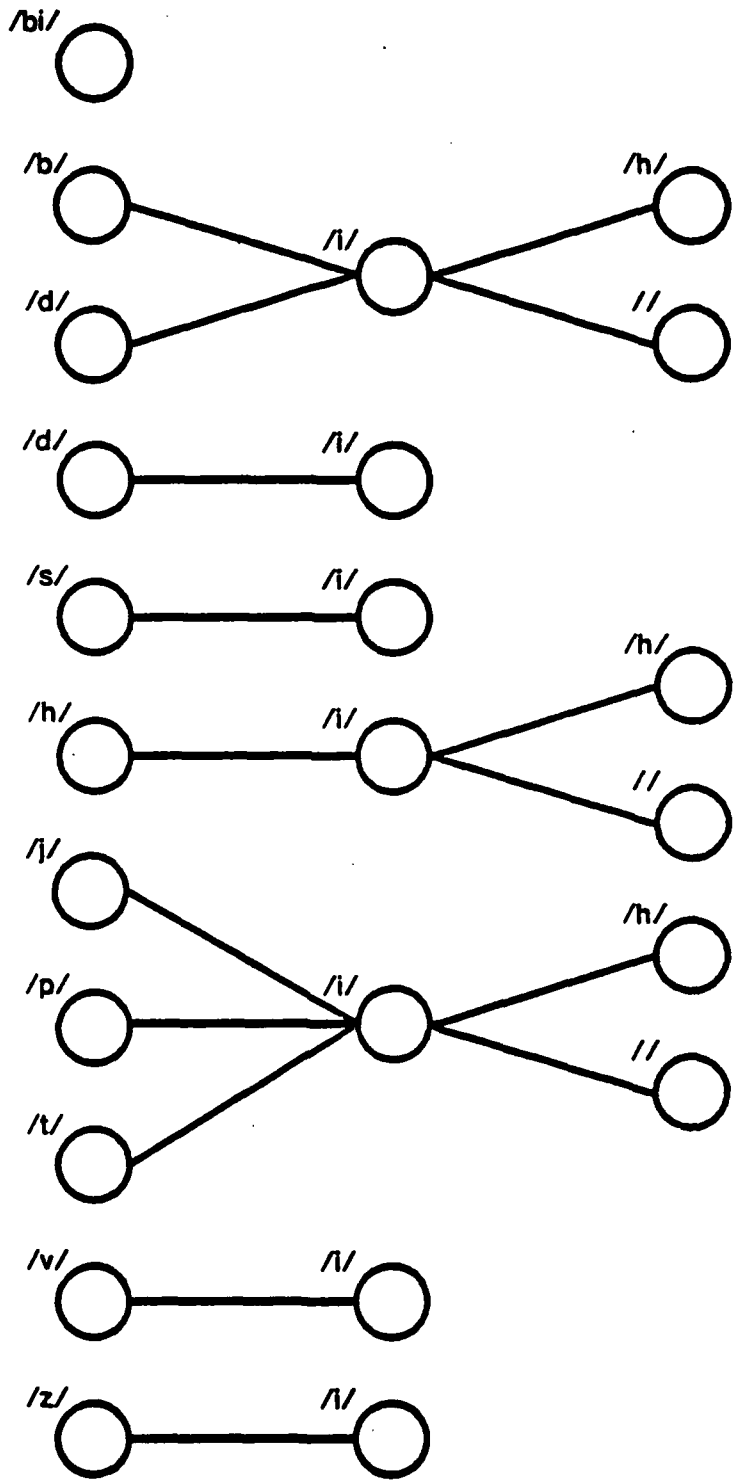


Figure II-8: E-set Network: Cycle 9

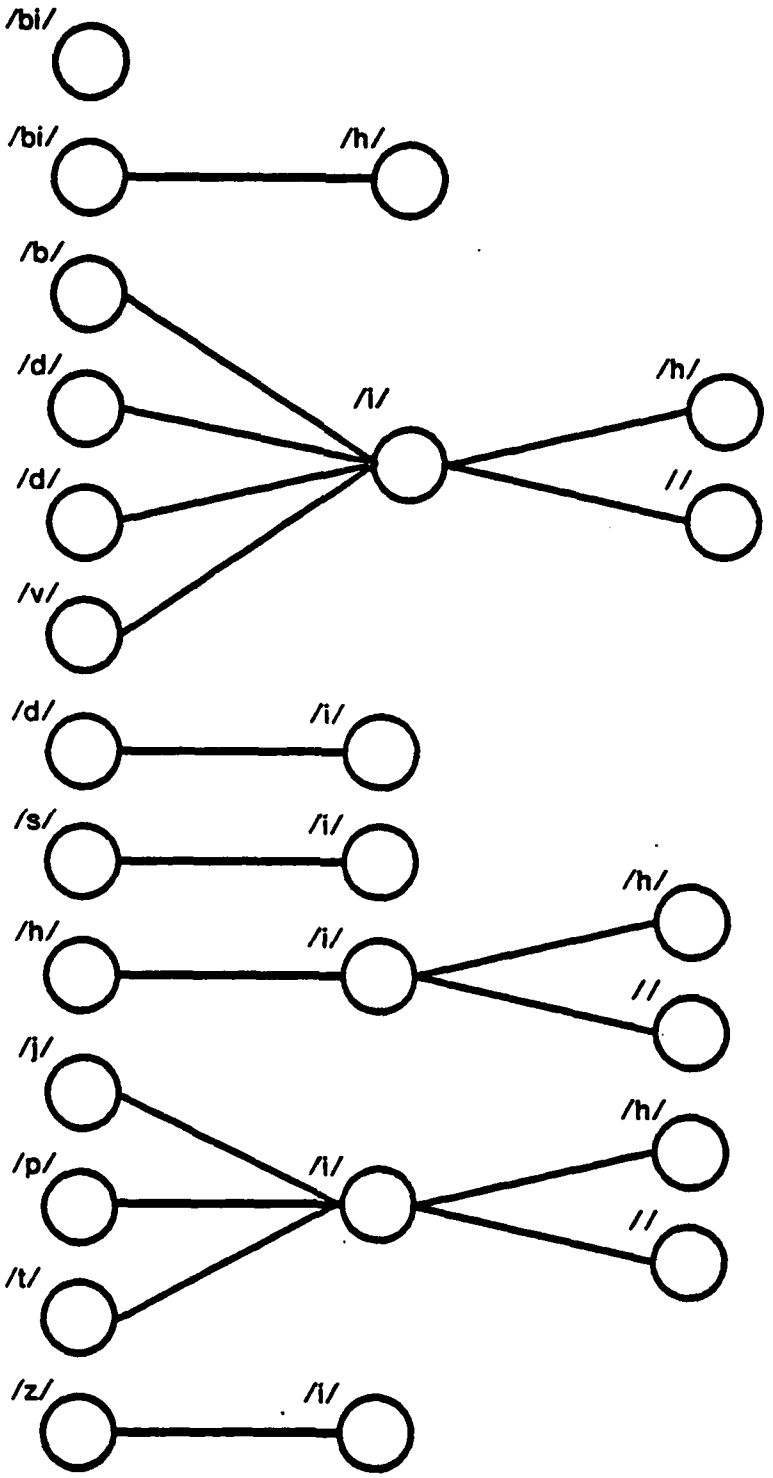


Figure II-9: E-set Network: Cycle 10

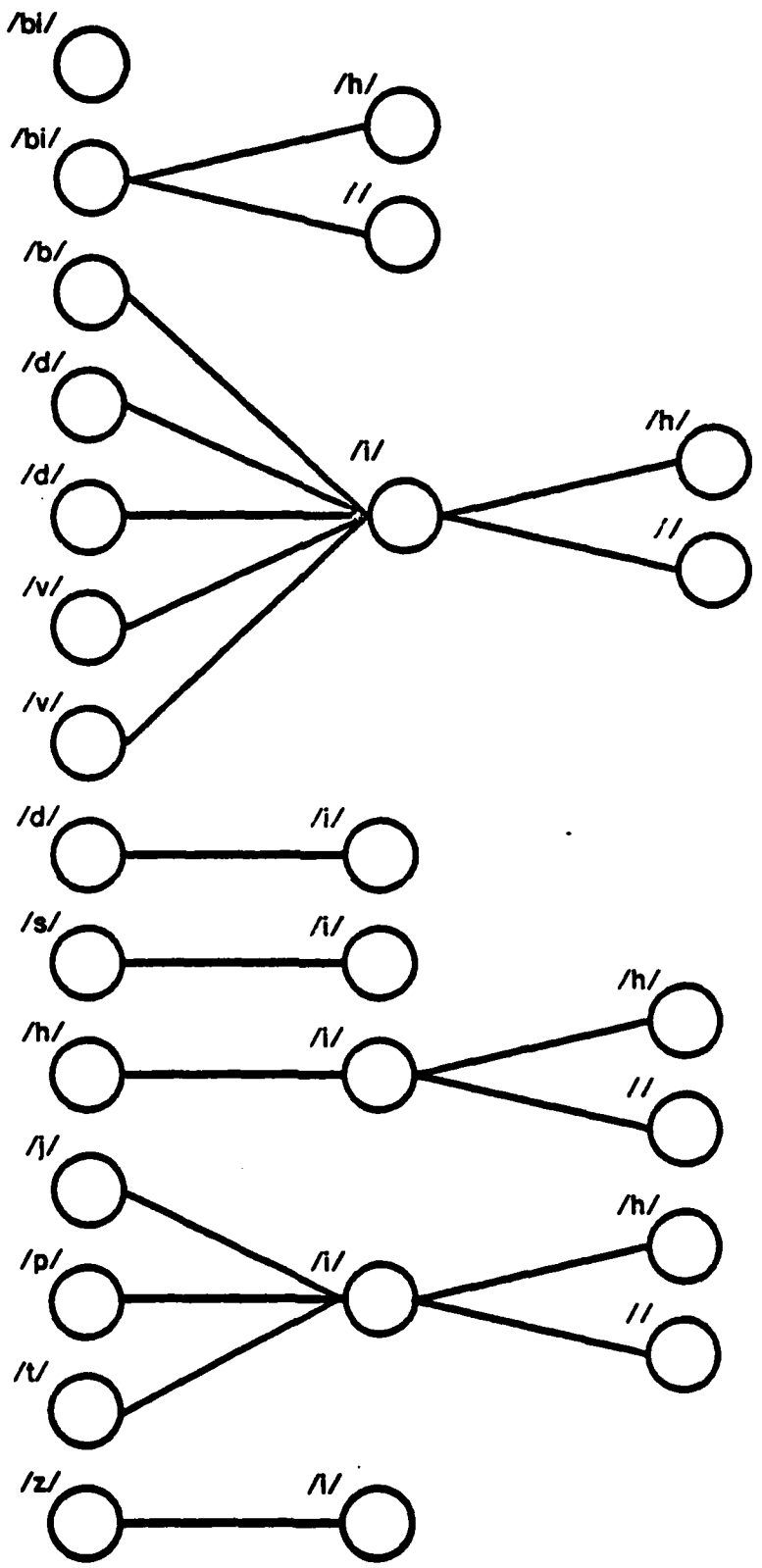


Figure II-10: E-set Network: Cycles 11 and 12

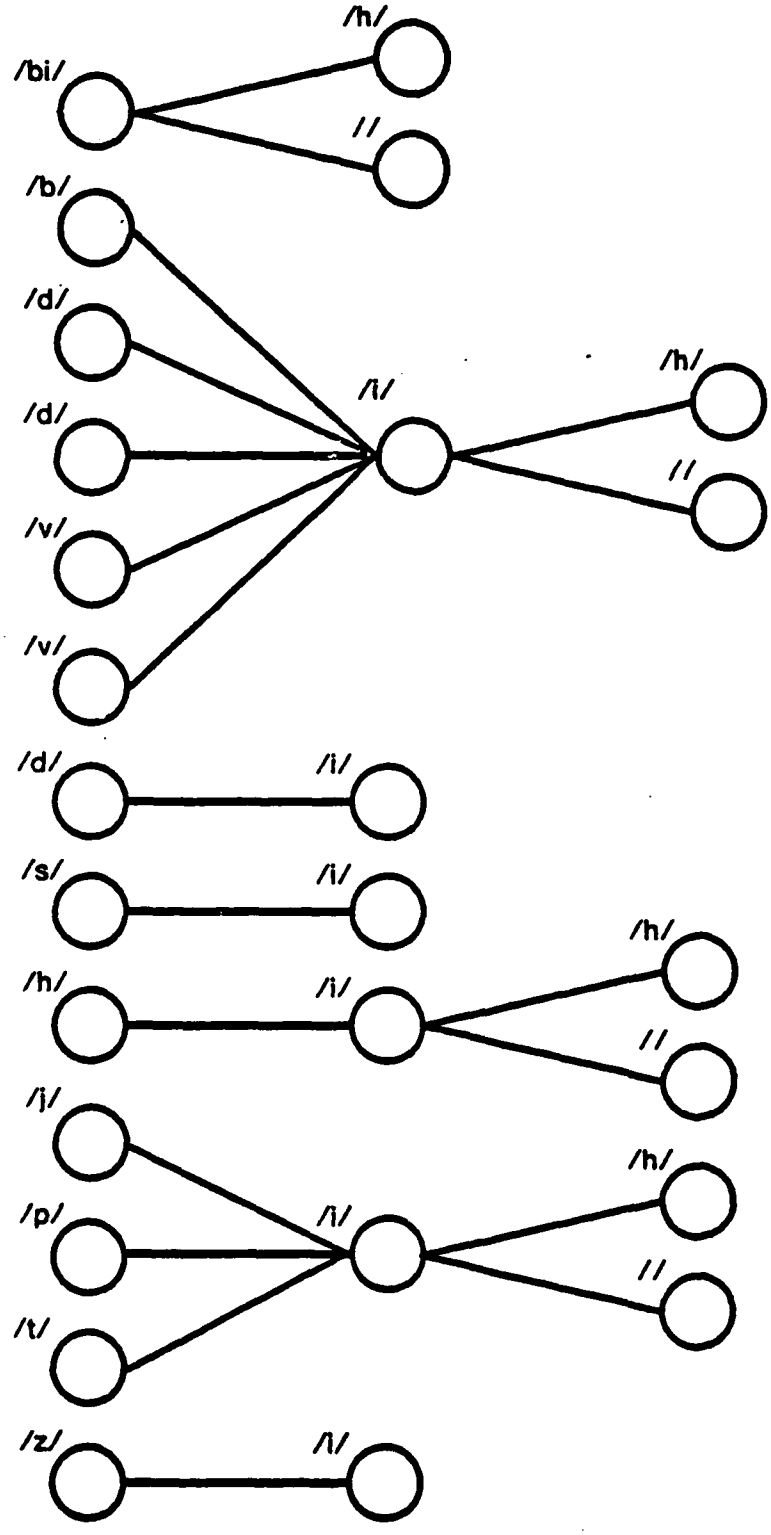


Figure II-11: E-set Network: Cycle 13

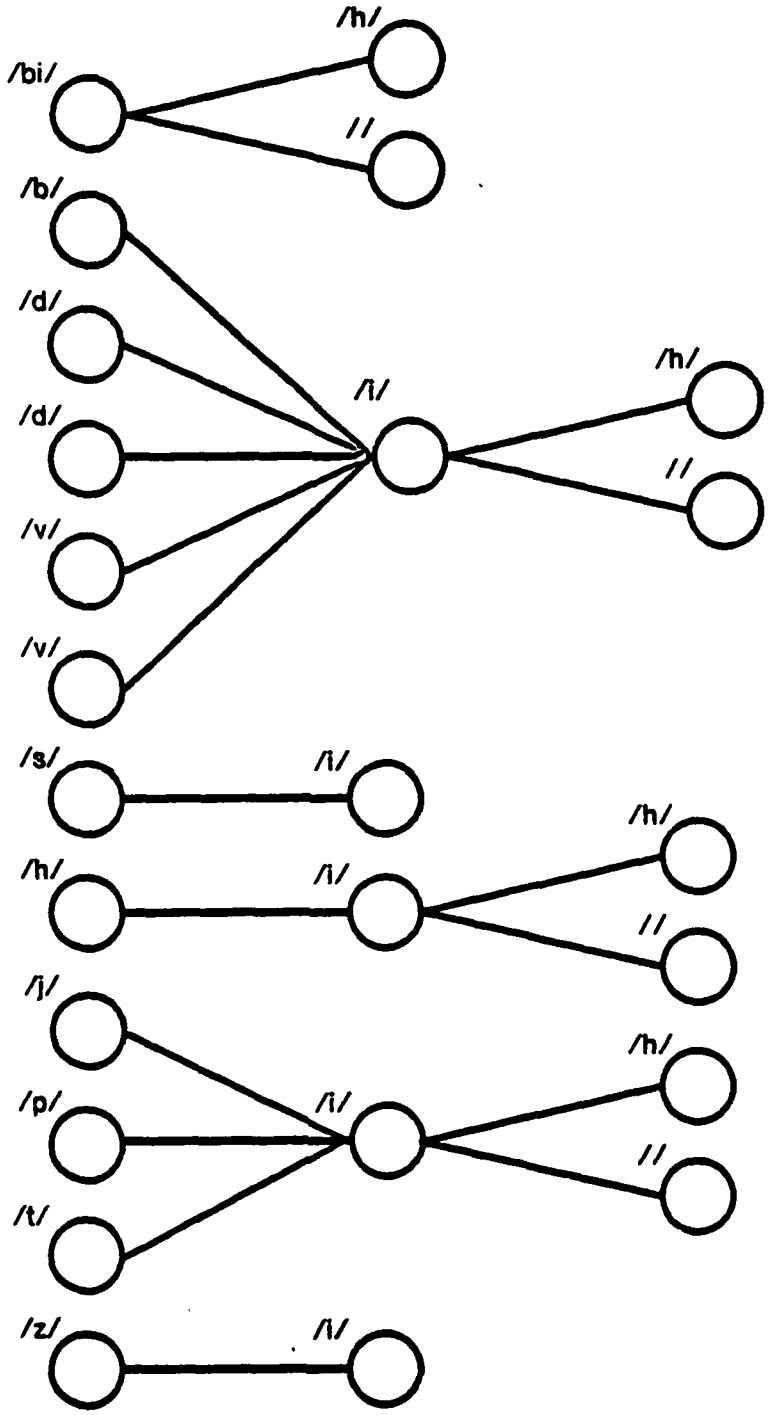


Figure II-12: E-set Network: Cycle 14

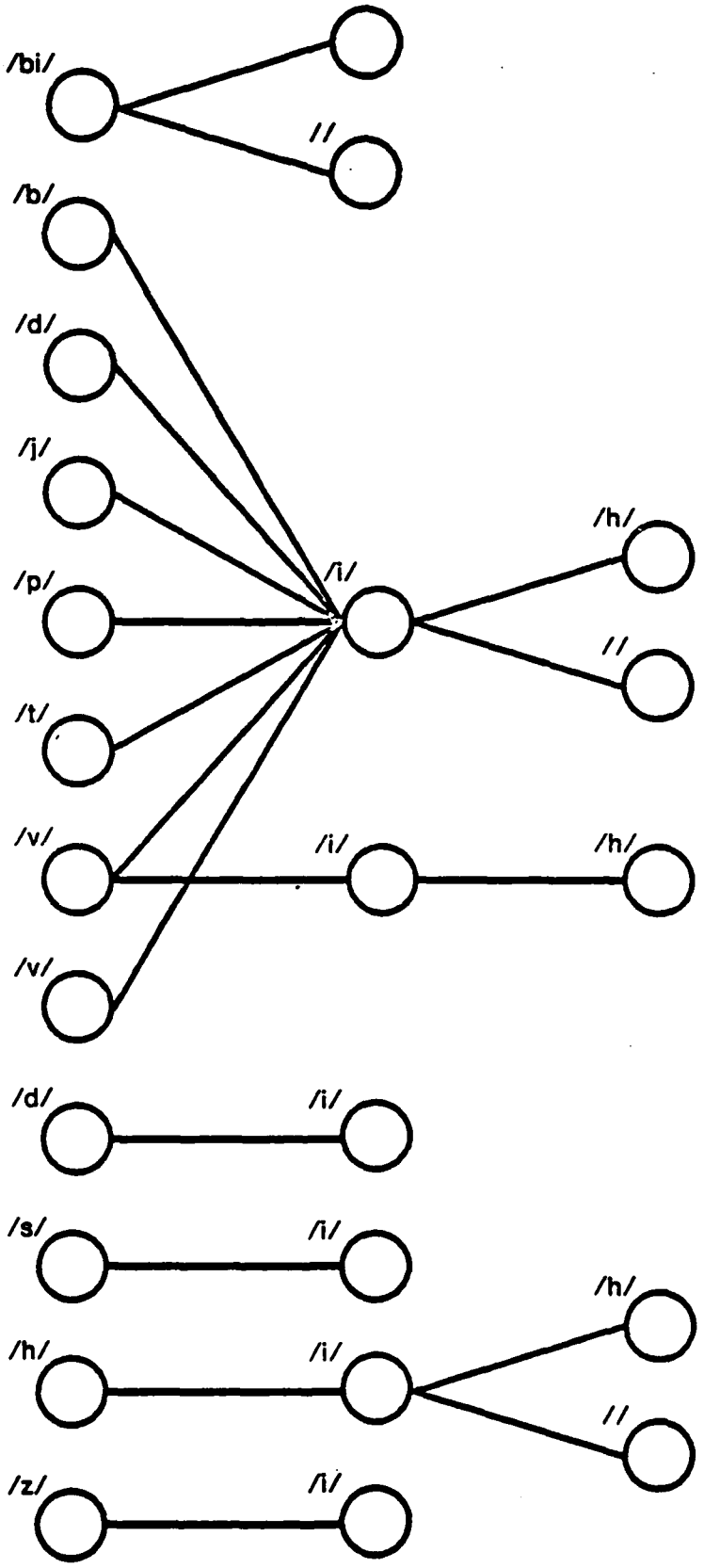


Figure II-13: E-set Network: Cycle 15

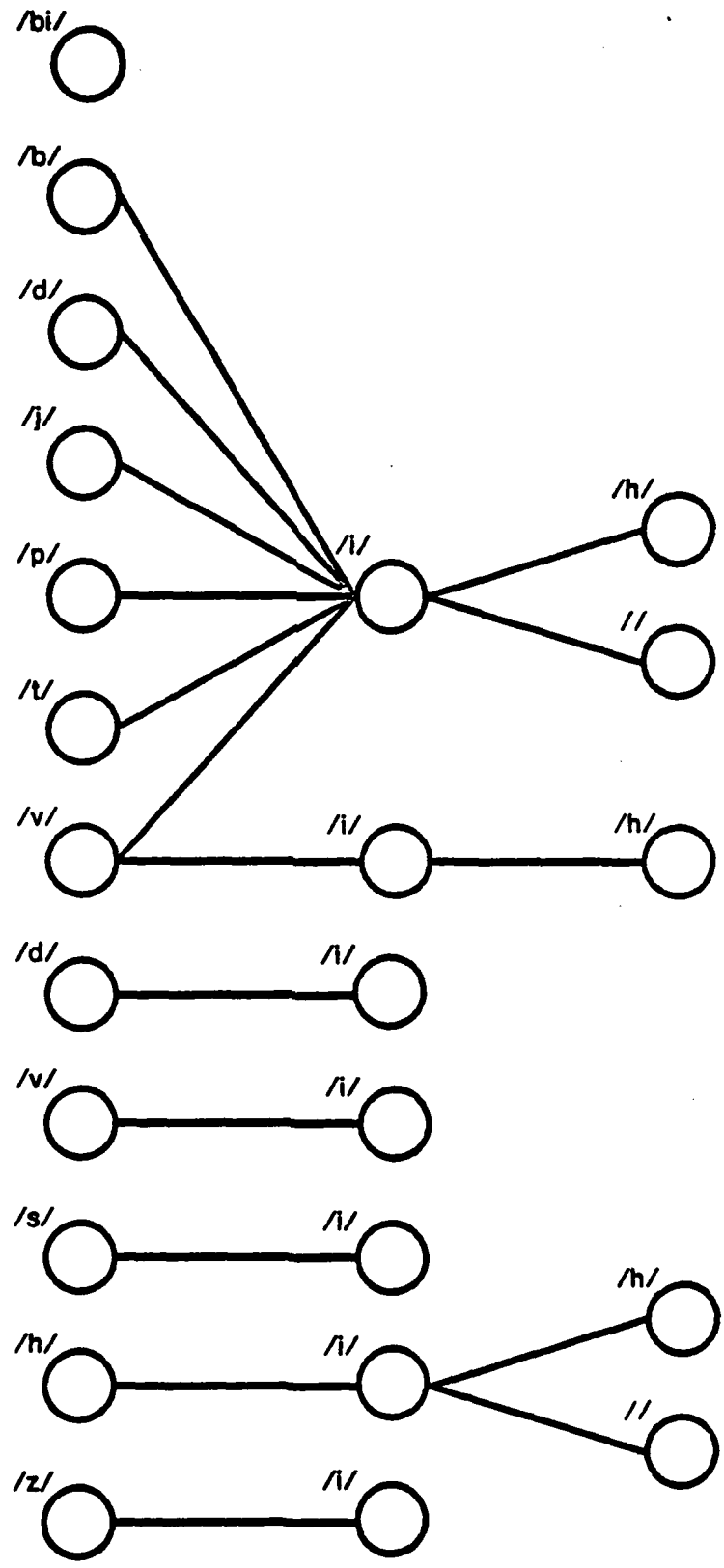


Figure II-14: E-set Network: Cycles 16 and 17

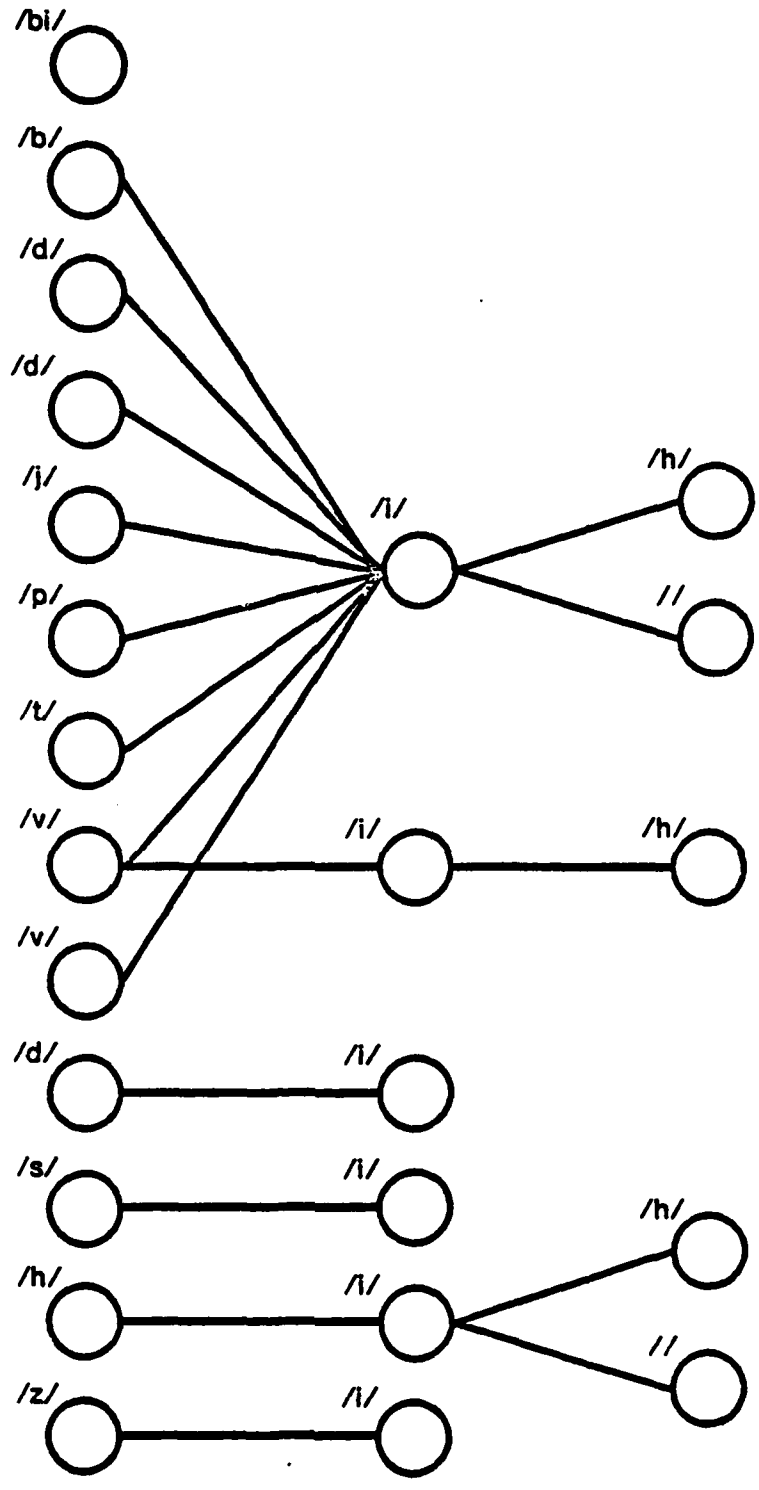


Figure II-15: E-set Network: Cycles 18 and 19

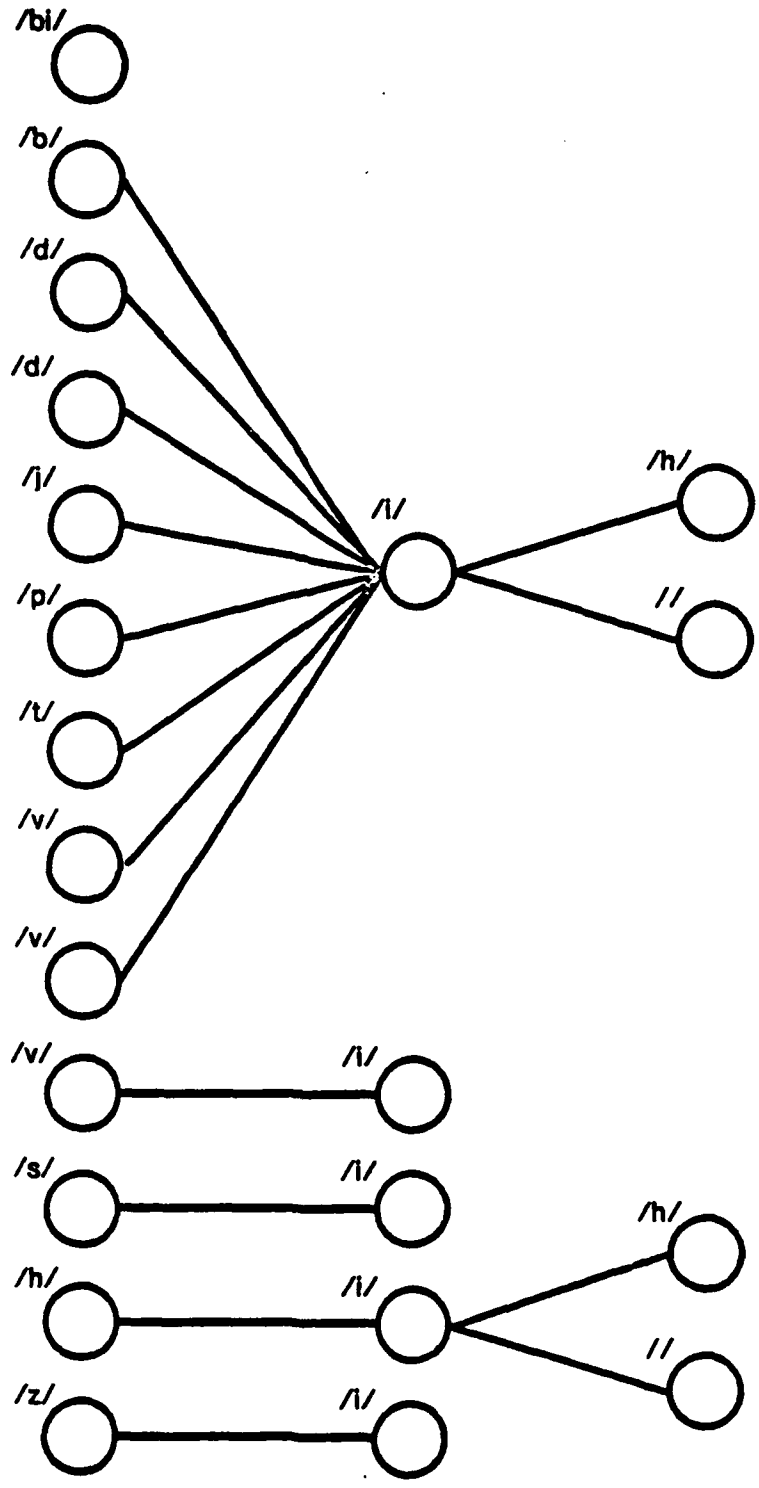


Figure II-16: E-set Network: Cycle 20

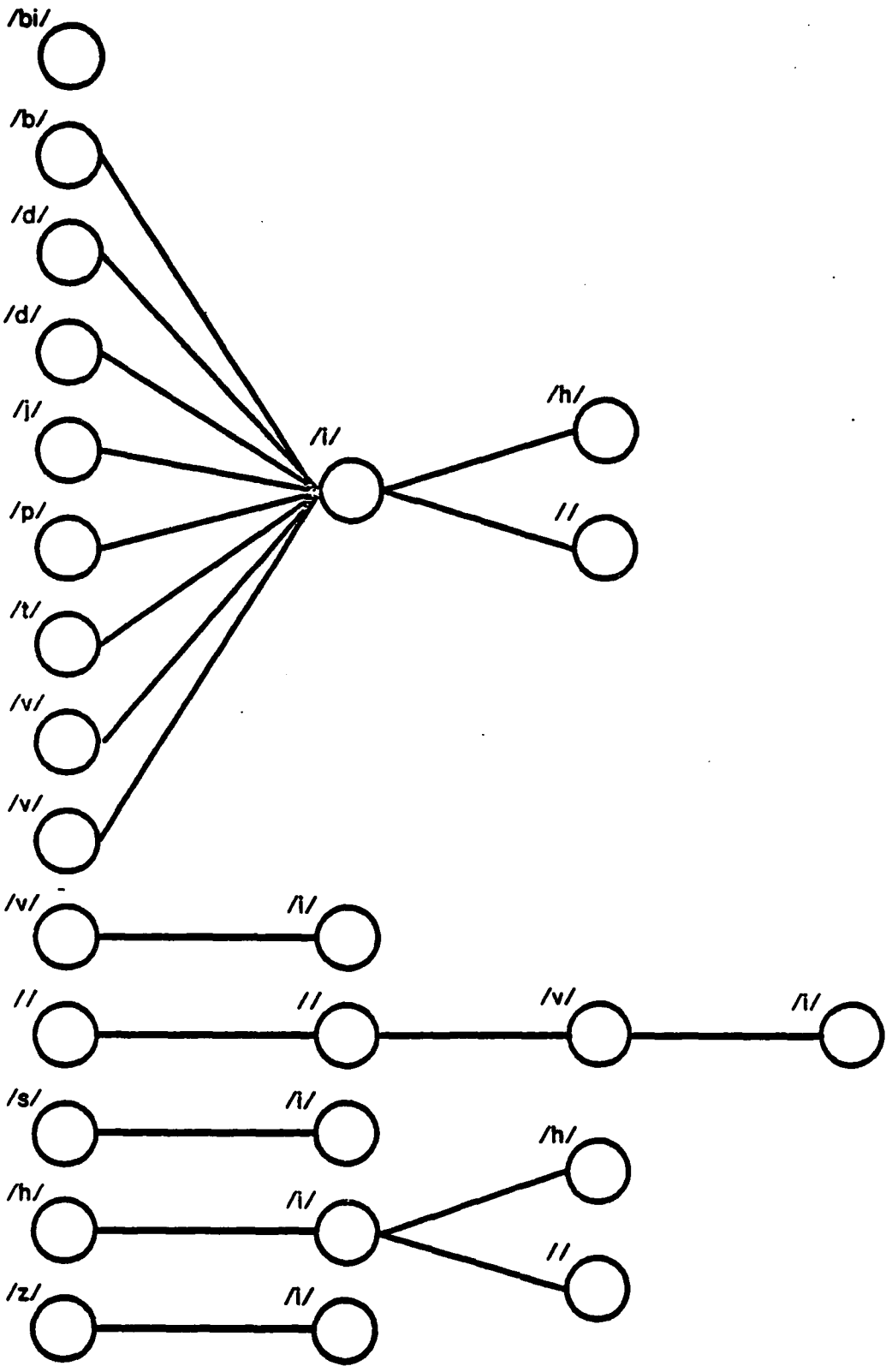


Figure II-17: E-set Network: Cycles 21 and 22

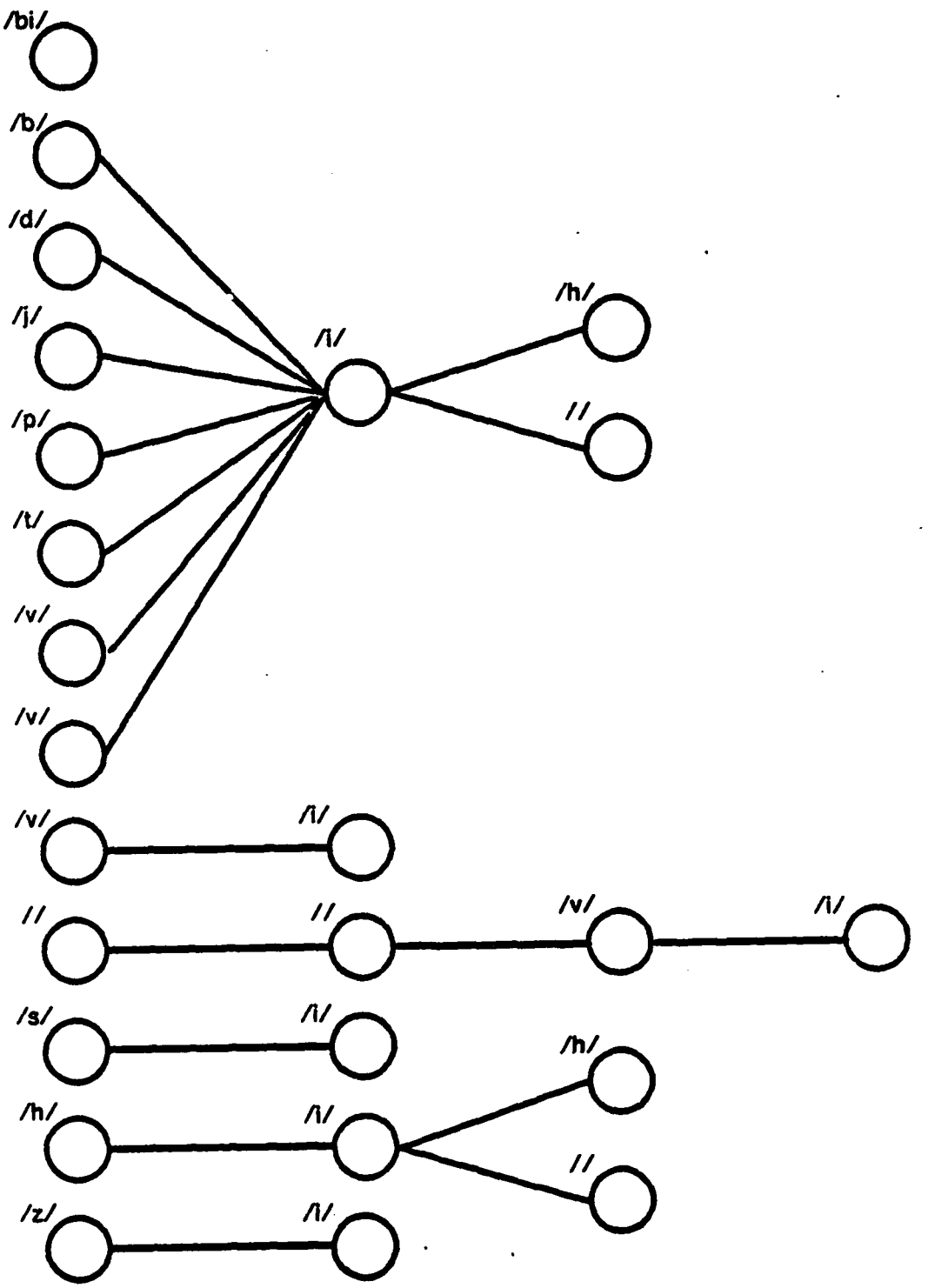


Figure II-18: E-set Network: Cycles 23 and 24

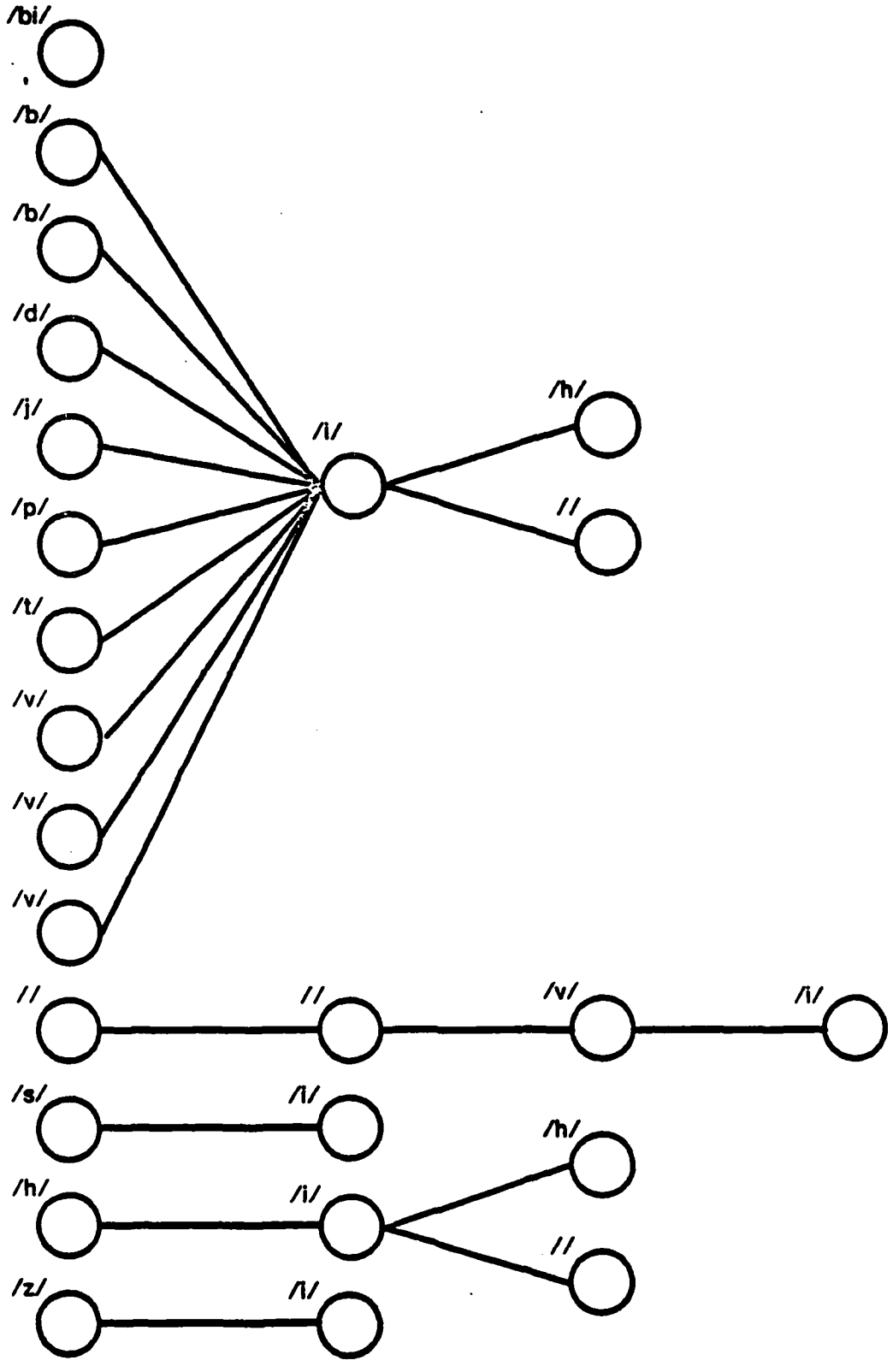


Figure II-19: E-set Network: Cycle 25

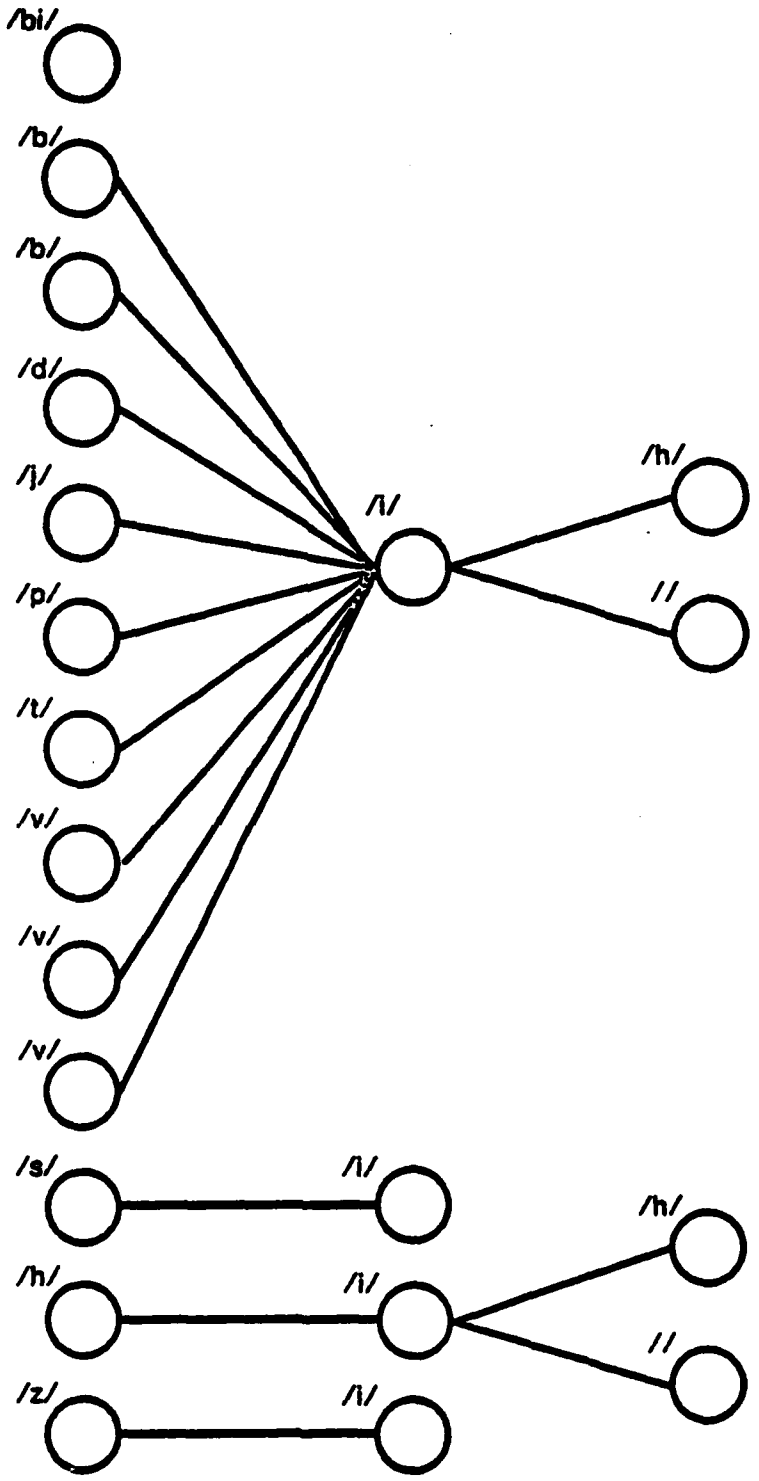


Figure II-20: E-set Network: Cycle 26

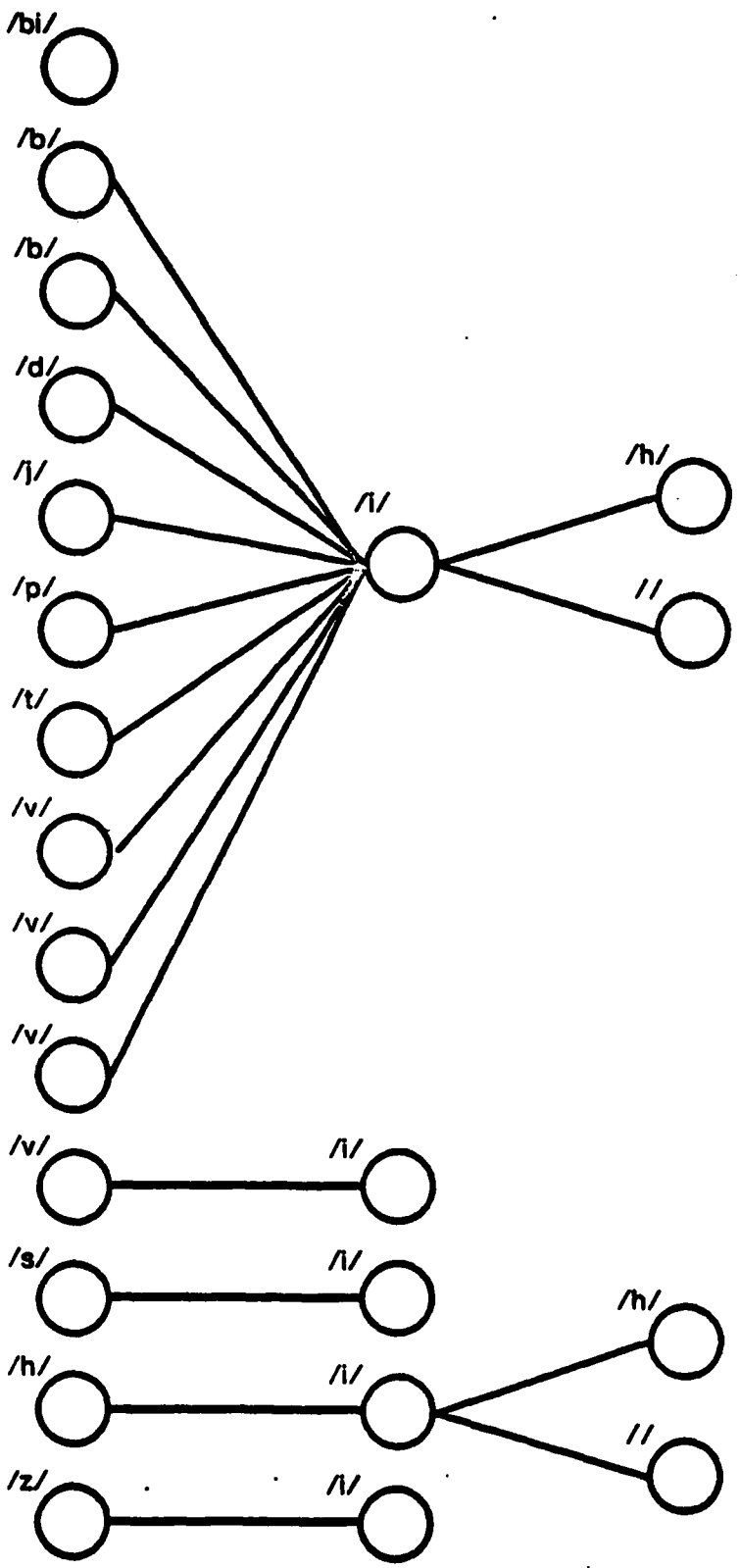


Figure II-21: E-set Network: Cycles 27 and 28

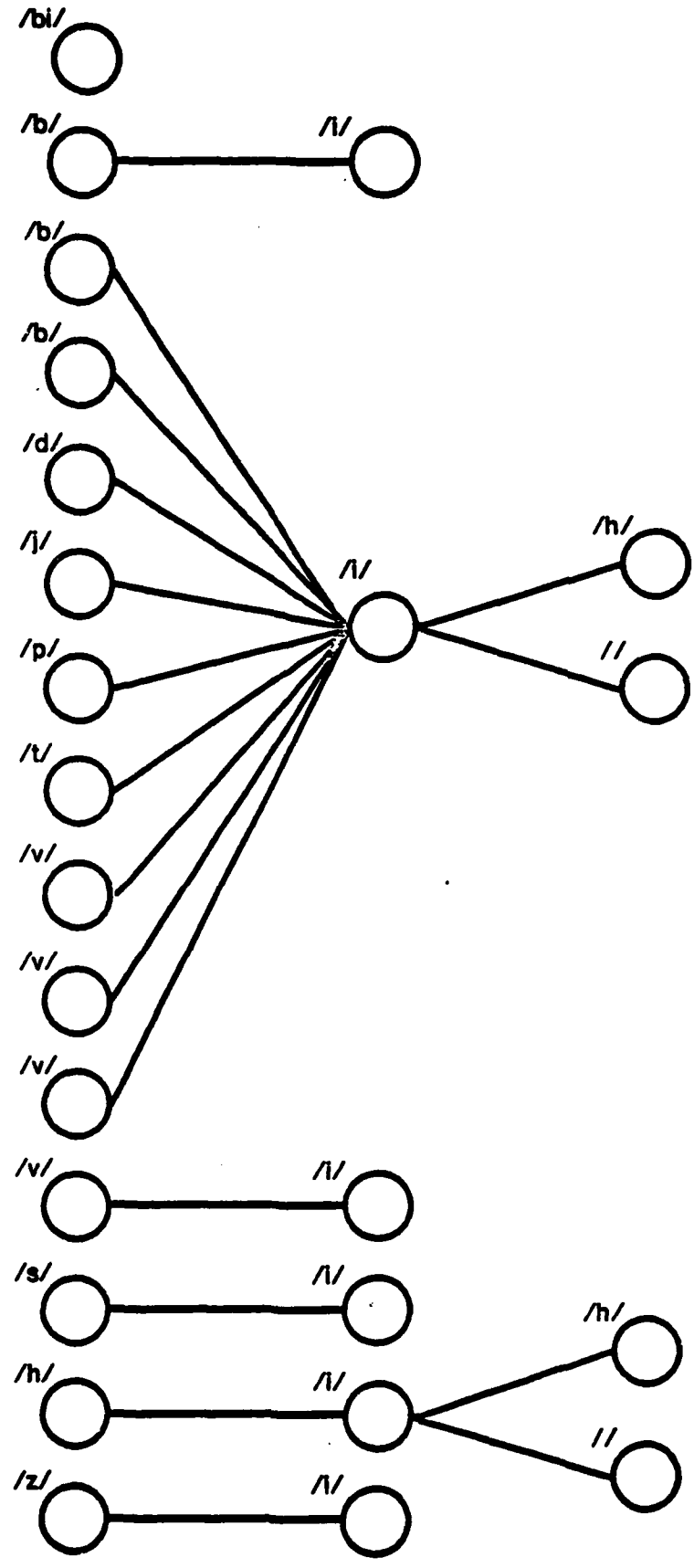


Figure II-22: E-set Network: Cycle 29

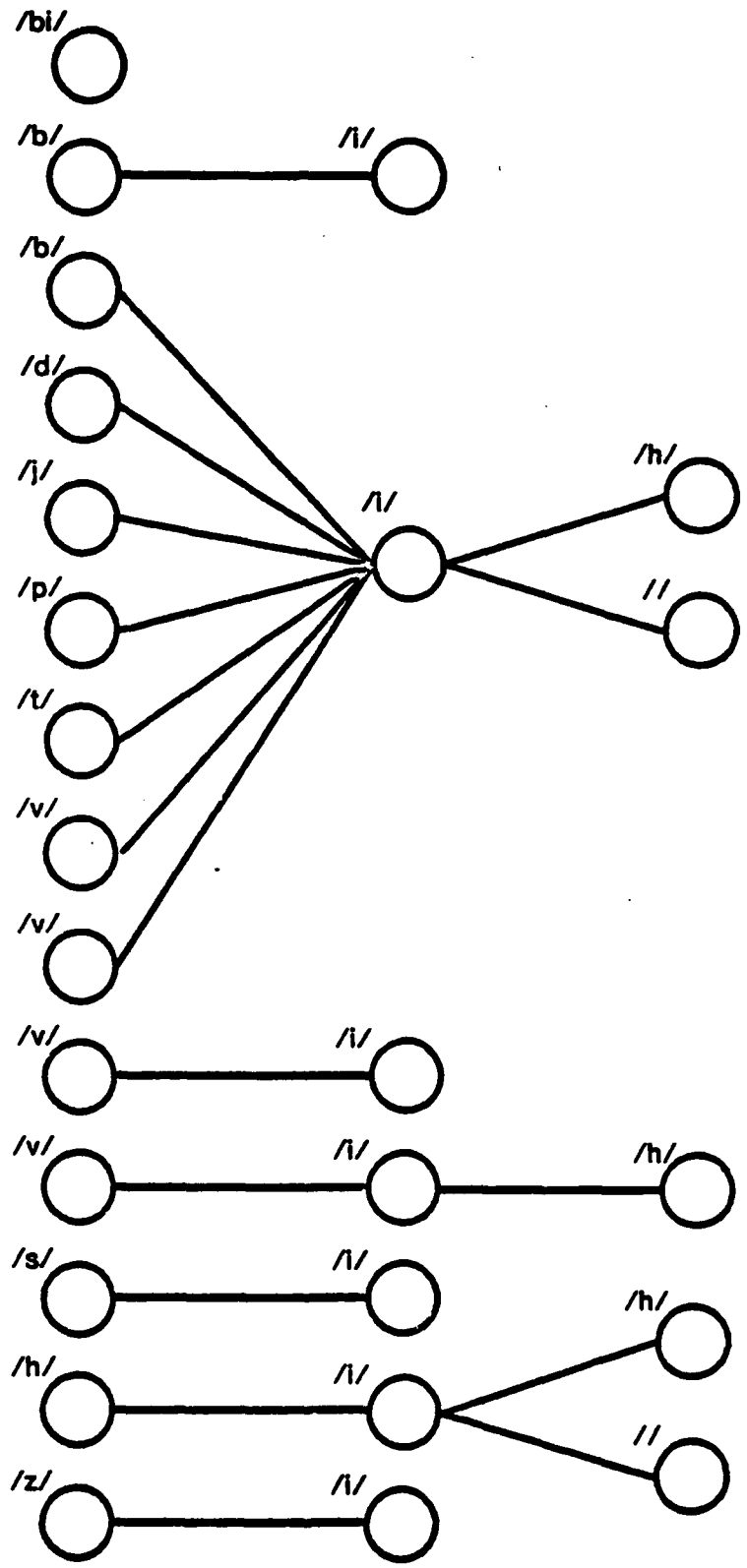


Figure II-23: E-set Network: Cycle 30

APPENDIX B

AUGMENTED ALPHABET DATA BASE FOR NEXUS

The data base consists of 32 monosyllabic items. 55% of the syllables are CV in structure, 35% are (C)VC, 10% are V. All the consonant phonemes of English are represented except [ž] which occurs rarely in English. Glottal stop is automatically present in vowel-initial words and is not indicated in the transcription. Over half the vowels and diphthongs of English are represented, including front, back, high, mid, and low vowels.

All the items in the data base make a minimal pair with at least one other item--six minimal pair sets in all. Minimal pairs are easily confusable items which force the recognition system to match items on the basis of different phonetic qualities rather than on unique word shapes (like 'double u').

Consonants are more heavily distributed in initial position; no stop consonants occur in word final position except in two consonant clusters with [k]. Final position consonants consist of fricatives, nasals, and glides. While these distribution patterns make the data base less generally representative of English, they maximize potential confusability of items and provide a rigorous test of word recognition.

Broad Transcription of Items in the Data Base

a	[e]	r	[ar]
b	[bi]	s	[ɛs]
c	[si]	t	[ti]
d	[di]	u	[yu]
e	[i]	v	[vi]
f	[ɛf]	w	---
g	[ǰi]	x	[ɛks]
h	[eǩ]	y	[waɪ]
i	[aɪ]	z	[zi]
j	[ǰe]	gay	[ge]
k	[ke]	-ing	[ɪn]
l	[ɛl]	ink	[ɪŋk]
m	[ɛm]	thee	[ϑi]
n	[ɛn]	shy	[ǩaɪ]
o	[o], [ov]	high	[haɪ]
p	[pi]	-eth	[ɛθ]
q	[kyu]		

Phonemes Represented in the Data Base

Consonants

	lab.	den.	alv.	pal.	vel.	glt.
stops	p,b		t,d		k,g	?
fric.	f,v	θ,ϑ	s,z	ǩ,ǰ ǩ,ǰ		h
nasals	m	n			ŋ	
liquids/ glides	w		l,r	y		

Vowels and Diphthongs

	front	central	back	diphthongs
high	i		u	aɪ
	ɪ		(ʊ)	eɪ
mid	e	(ə, ʌ)	o	oʊ
	ɛ		(ɔ)	(aʊ)
low	(æ)		a	(ɔɪ)

Circled segments are absent from the data base.

Distribution of consonants in word initial and word final positions.

	Initial consonants (CV)		Final Consonants (VC)	
	-voice	+voice	-voice	+voice
stops	[p]	[b]		
	[t]	[d]		
	[k]	[g]		
fric.		[v]	[f]	
		[ð]	[θ]	
	[s]	[z]	[s]	
	[š]			
		[ǰ]	[č]	
	[h]			
nasals				[m]
				[n]
				[ŋ]
glides/		[w]		[l]
liquids		[y]		[r]

Consonant Clusters: [ky] CCV, [ks] VCC, [ŋk] VCC

Minimal Pairs

[(C)i]		[ɛC]		[(C)aɪ]	
e	[i]	f	[ɛf]	i	[aɪ]
b	[bi]	l	[ɛl]	y	[waɪ]
c	[si]	m	[ɛm]	shy	[saɪ]
d	[di]	n	[ɛn]	high	[haɪ]
p	[pi]	s	[ɛs]		
t	[ti]	x	[ɛks]		
v	[vi]	-th	[ɛθ]		
z	[zi]				
thee	[ðɨi]				

[(C)e(C)]		[Cu]		[ɪC]	
j	[ǰe]	q	[kyu]	-ing	[ɪŋ]
k	[ke]	u	[yu]	ink	[ɪŋk]
a	[e]				
h	[eč]				
gay	[ge]				

APPENDIX C

PILOT DATA BASE OF SHORT PHRASES

This appendix describes a simple connected-speech data base, serving as an example of the application of the techniques developed in Section VI. Included are a copy of the lexicon, the grammar, and a discussion of several speech phenomena that should arise when talkers produce speakers from this vocabulary. Although the lexicon is small and the grammar simple, the communication system should prove challenging to most speech recognition systems, as will be seen in later sections describing the characteristics of the system.

Subject Nouns N_1
I
you
he
she
it
we
they

Articles Art.
the
a

Prepositions Prep.
to
for
in

Uninflected Verb V
connect
translate
copy
send
file
call

Object Nouns N_2
me
you
him
her
it
us
them
file
text
letter
document
call
copy

Inflected Verbs V_{ed}
transferred
connected
translated
copied
filed
called

Auxiliary Verbs Aux.
did
have
has
will

II. Grammar

A. Sentence Types and Examples

1. V NP_2 (Prep) (NP_2) Imperative
   a. Copy it.
   b. Give him the file
   c. Send it to her.
   d. Copy the letter to the text.

2. N_1 (Aux) $V_{(ed)}$ NP_2 (Prep NP_2) Declarative
   a. They sent the letter.
   b. They will send the letter.
   c. They will send it to us.

3. Aux N_1 V NP_2 (Prep NP_2) Interrogative
   a. Have you called him?
   b. Did you give it to them?
   c. Will you transfer the letter to the file?

B. Rewrite Rules

1. NP(Art (V_{ed})) N
2. Parentheses refer to optional elements.

C. Grammaticality

1. The sentences chosen for actual use will make use of all lexical items and all sentence types but not all the possible combinations of each.

2. Native speaker intuitions will assure that the sentences chosen will be grammatical English sentences.

III. Broad Phonetic Transcriptions of the Words in the Lexical Set as if Pronounced in Isolation

I	[aɪ]	filed	[faɪld]
you	[yu]	called	[kɔld]
he	[hi]	transfer	['trænsfəd]
she	[ši]	transferred	['trænsfəd]
it	[ɪt]	cop	[kapi]
we	[wi]	copied	[kapid]
they	[ðe]	send	[sɛnd]
me	[mi]	sent	[sɛnt]
him	[hɪm]	connect	[kə'nɛkt]
her	[hə˞]	connected	[kə'nɛkt'd]
us	[ʌs]	translate	['trænzlet]
them	[ðɛm]	translated	['trænzleɪəd]
a	[a]	did	[dɪd]
the	[ðə]	have	[hæv]
file	[faɪl]	will	[wɪl]
text	[tɛkst]	has	[hæz]

letter	[lɛɾə]	to	[tu]
document	['dakyumənt]	for	[fɔr]
call	[kɔl]	in	[ɪn]

IV. Sources of Variation
A. Speaker dependent variation
1. age
2. sex
3. dialect
a. vowels
1. /a/ ~ /ɔ/ call [kal] ~ [kɔl]
2. /o/ ~ /ɔ/ for [for] ~ [fɔr]
3. /ə/ ~ /ɪ/ translated [trænzleɾəd]~ [trænzleɾɪd]
4. /ɛ/ ~ /ɪ/ in [ɛn] ~ [ɪn]
them [ðɛm] ~ [ðɪn]
send [sɛnd] ~ [sɪnd]
sent [sɛnt] ~ [sɪnt]

b. lengthening and diphthongizing
1. he [hi] ~ [hiɪ]
2. copy [kapi] ~ [kapɪ]
3. file [faɪl] ~ [faɪᵊl]
4. them [ðɛm] ~ [ðɛᵊm]
5. him [hɪm] ~ [hɪᵊm]
6. will [wɪl] ~ [wɨl] ~ [wɪᵊl]
7. transfer [træns] ~ [træᵊns] ~ [trɛᵊns]
8. I [aɪ] ~ [aᴵ] ~ [a]

c. nasalization
1. sent [sɛnt] ~ [sɛ̃nt] ~ [sɛ̃t] ~ [sɛʔ]
2. transfer, translate [træns] ~ [træ̃ns] ~ [træ̃s] ~ [træ̃ʔs]

d. reduction
1. document ['dakyumɨnt] ~ ['dakyumnt] ~ ['dakyəmnt]
2. connect [kʰə'nɛkt] ~ [kʰ'nɛkt]

e. clarification of reduced vowels
1. document ['dakyumːɹt] ~ ['dakyumɪnt]

f. palatalization
1. to [tu] ~ [tyu]

g. consonants
1. call [kal] ~ [kaw]
2. you [ju] ~ [iu]

h. syllabic organization

1. document ['dakyumɨnt] ~ ['dak·yumɨnt]

B. Context Dependent Variation

1. reductions

a. The vowel in pronouns, articles, and prepositions (open or closed syllables) may reduce or even delete

1. send us the file [sɛnd(ɨ)sˠ(ɨ)faɪl]
2. to me [t(ɨ)mi]
3. letter in the file [lɛɾə(ɨ)nˠ(ᵊ)faɪl]

b. Word final consonant clusters may be simplified

1. send it, sent it [sɛnɪt]
2. connect the call [kə'nɛkˠəkal]

2. assimilations

a. Word final and word initial consonants will ususally be spoken in a single articulation if they are the same or nearly the same in place of articulation

1. send it to me [sɛndɪt:umi]
2. copied document [kapid:akyumɪnt]

b. Alveolars may palatalize when followed by a palatal glide. Glide then deletes, producing an affricate.

1. Did you send it? [dɪʤusɛndɪt]
2. I will connect you. [aɪlkə'nɛkʧu]

c. Voicing may assimilate the the voicing of the environment.

1. connect him [kə'nɛkdɪm]

3. deletions

a. initial fricative of pronouns 'him', 'her', and 'them' may delete following any other segment.

1. connect him [kə'nɛktɪm]
2. transfer her ['trænsfəə]
3. send them [sɛndəm]

b. Remaining vowels may reduce and delete (B.1.a.)

c. Initial consonant and vowel may be deleted in auxiliaries 'will' and 'have' (contraction).

1. I will [aɪl]
2. I have [aɪv]

4. 0 insertions

V. Other Features of the Data Base

A. Homophones
1. him-them [əm]
2. send it-sent it [sɛ̃ɾɪt]

B. Minimal pairs
1. complete inventory: 27 total pairs

a.	I	[aɪ]	V
	I'll	[aɪl]	VC
b.	it	[ɪt]	VC
	in	[ɪn]	VC
c.	him	[ɪm]	VC
	them	[ɛm]	VC
d.	a	[ə]	V
	the	[ðə]	CV
e.	the	[ðə]	CV
	they	[ðe]	CV
f.	they	[ðe]	CV
	a	[e]	V
g.	to	[tu]	CV
	you	[yu]	CV
h.	he	[hi]	CV
	she	[ši]	CV
	we	[wi]	CV
	me	[mi]	CV
i.	you've	[yuv]	CVC
	you'll	[yul]	CVC
	(also with 'we', 'they', I)		
j.	he's	[hiz]	CVC
	he'll	[hil]	CVC
	(also with 'she', 'it')		
k.	send	[sɛnd]	CVCC
	sent	[sɛnt]	CVCC
l.	file	[faɪl]	CVC
	filed	[faɪld]	CVCC
	(also with 'call', 'transfer', connect', 'translate', and 'copy')		

2. distribution of syllable types
a. per cent of minimal pairs in open syllables (V and CV) is 14:46 or 30%
b. per cent of minimal pairs in closed syllables (VC and CVC) is 32:46 or 70%

C Inventory of sounds (Circled segments are absent from the data base)

1. vowels

	front	central	back
high	i	ɨ	u
	ɪ		(ʊ)
mid	e	ʌ ə	(o)
	ɛ		ɔ
low	æ		a

2. diphthongs

[aɪ] [(eɪ)] [iɪ]
[(ɔɪ)] [(oʊ)] [(aʊ)]

3. consonants

	labial	interdental	alveolar	alveolar	palatal	velar	glottal
stops	p,(b)		t,d			k,(g)	?
fricatives	f,v	(θ) ð	s,z	š,(ž)	č,ǰ		h
nasals	m		n			ŋ	
laterals			l	r			
glides	w				y		

4. sound patterns
   a. consonants
      1. initial consonants: t,tr,d,k,f,ɬ,s,š,h,m,l,w,y
      2. final consonants and possible sequences with initial consonants across word boundaries:

knst,nt,kt,t # t,f,ɬ,m,h,y
rd,ld,nd,d # w,š,ɬ,m,n,y
v # t,k,s,ɬ,y
z # t,s,š,k,h
m # t,f
n # ɬ
l # š,ɬ,m,y,h,w
r # f,t,ɬ,h

      3. medial clusters: nsl, ky, nsf, kt

   b. vowels
      1. initial vowels: i, e, ə, u, aɪ
      2. final vowels: ɪ, ə
      3. vowel sequences across word boundaries:

i # ɪ, ə
u # ɪ,ə,ə̯

C. syllable patterns
   1. total different syllables: 50

   2. structural forms
      a. V (2)
      b. CV (15)
      c. C^2VC^3 (29)
      d. VC (4)

D. Intonation Patterns
   1. Interrogatives are marked by rising intonation usually beginning on the last syllable, but sometimes on an earlier stressed syllable.

   2. Declaratives are marked by falling intonation on the last syllable and a general progression from higher to lower tones.

   3. Imperatives generally progress from high to low intonation.

4. Examples

3 2 1 3
a. Did you call him?

3 2 4 3 3 1
b. I have filed the letter.

3 1 3 3 1 2 2 1
c. Call him. Copy the letter for me.

VI. Speech "Networks"

Transfer me to him.

[træ̃(n)(z)sfə̃mitu(h)(ɪ)m]
(ɾ ə)

Send the letter to the file.

[sɛ̃(n)(d)ɝ(ə)lɛɾətuɝ(ə)faɪl]
(ə)

You connected him to them.

[yukənɛktɪd(h)(ɪ)mtu(ɝ)(ɛ)m]
(ɨ) (d) (ə)
(ə)

Did you send the document to her?

[dɪd(dy)usɛ̃(n)(d)ɝ(ə)dakyum(ɪ)nt:u(h)ə]
(ǰ) (ə)(ɨ)
(ə)

VII. Discussion

A. In what ways is this data base representative of a larger scope of speech?

The data base includes most of the speech sounds of English. All but three sounds (ž, ɔɪ, aʊ) will be represented with the inclusion of the following five words into the data base:

N_2: thing, book V: show, give, save

The data base does not necessarily represent norms of sound frequency and distribution patterns found in larger samples of English, but it does contain many of the important distributional patterns. All the major classes of sounds (stops, fricatives, nasals, laterals, and glides) appear in word initial and word final position.

All four syllable types are represented and have the expected distribution patterns for English in general. Furthermore, the same syllables appear in a variety of contexts providing a full range of stressed, full, and reduced syllables.

Both rising and falling intonation patterns are represented, and the distribution of either can be generated within the data base by manipulating the context in which the utterances are spoken. (File the text with the letter. File the text with the letter.)

B. What issues does this data base present for NEXUS?

The recognition strategies already available in Nexus will be utilized. Most of the words are distinct from one another in many ways, providing the matcher with several cues. Several of the words are minimal pairs (each verb with its past inflected form and several pronouns), which are structures Nexus has been successful with. However, the data base does provide a good test of word recognition in connected speech. The sentences

are all short enough to be spoken in one breath group eliminating any pauses which could be otherwise used for cues. Furthermore, the data base offers a wide variety of contextual variants since words can appear in a variety of positions and in combination with numerous other words.

A particular challenge to recognition provided by this data base is the pronunciation of cliticized words. Pronouns, auxiliary verbs, articles, and prepositions all share the property of undergoing variable reduction, which may be drastic in degree, in naturally spoken English. This is in large part because prosodically they are pronounced as part of the preceding word. Clitics will be extremely difficult to recognize because there is so little phonetic substance remaining and because clitics closely resemble one another.

New strategies, probably including some knowledge of cliticization processes, will have to be devised for Nexus to match full words to cliticized pronunciations. The adaptation must be more sophisticated than weighting strategies which favor syllables that are prosodically prominent because clitics have exactly the features which would be unweighted and because their correct identification is often crucial to the meaning of the sentence. This is an appropriate data base for recognition strategies which make use of rhythmic structures for nonlinear matching rules.

Appendix D
Annotated Speech Bibliography
Gary Tajchman
Rebecca Burns

Aaltomen, O. (1985). The effect of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels. Journal of Phonetics, 13, 1.

-The effect of the relative amplitude levels of F2 and F3 turned out to be systematic regardless of the location of F3 in the frequency scale. Reduction of F2 amplitude increased /i/ idents, F3 correspondingly increased /y/ idents. Ident. data deal only with final product of a process and very little with the information processing.

Aaltomen, O., and Syonpää,J. (1983). Computerized two-dimensional model for Finnish vowel identifications. Audiology, 22, 410-415.

-resulting basic vowel-identification chart demonstrates areas where stimuli were identified as a certain vowel and offers the practical means for presentation and follow-up of individual articulatory and auditory capacities.

Ainsworth, W. A. (1981). Duration as a factor in the recognition of synthetic vowels. Journal of Phonetics, 9, 333-342.

-results suggest that differences in duration of 250 msec have about the same effect on vowel recognition as differences in formant frequency of about 100 Hz or 1 Bark.

Ainsworth, W. A. & Paliwal, K. K. (1984). Correlation between the production and perception of the English glides /w,r,l,j/. Journal of Phonetics, 12, 237-243.

-results tend to reject the hypothesis that a listener refers to his own articulation when perceiving speech.

Alfonso, P. J., & Baer, T. (1982). Dynamics of vowel articulation. Language and Speech, 25, 151-173.

-results provide a complete physiological description of the S's vowel space and show good agreement among the different levels of description.

Allik, J., Meelis M., & Ross, J. Comment on measurement of pitch in speech: An implementation of Goldstein's theory of Pitch Perception. Journal of the Acoustical Society of America, 13, 1.

-pitch detection algorithm proposed by Dvifhvis, Willems and Sivyter can be made more than 20x faster by replacing the harmonic sieve procedure by the approximate common denominator procedure, results differing only slightly.

Amerman, J. D. & Parnell, M. M. (1984). Variable perceptual potency of the initial 20 ms time window. Journal of Phonetics, 12(1), 1-7.

-examines the contribution of perceptual potency of the initial 20 msec "time window" associated with voiceless stop consonant release.

-found perceptual potency of 20ms window to vary significantly.

Amerman, J. D., & Parnell, M. M. (1981). Influence of context and rate of speech on stop-consonant recognition. Journal of Phonetics, 9, 323-332.

-findings suggest that the cue strengths of phonetic segments associated with consonant perception change as a function of context-rate interaction.

Atal, B. S. (1983). Speech coding: Recognizing what we do not hear in speech. Annals of the New York Academy of Sciences, 405, 18-32.

-discusses factors that influence the design of efficient speech coders as well as new speech-coding methods that attempt to maximize the perceptual similarity between the original speech signal and its coded replica.

Atlas, L. E. et al. (1983). Results of stimulus and speech-coding schemes applied to multichannel electrodes. Annals of the New York Academy of Sciences, 405, 377-386.

-finding suggest that different clues relevant to speech comprehension are optimized by each of the different modes of processing speech.

Baddeley, A., Eldridge, M., & Lewis, V. (1981). The role of subvocalization in reading. Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 33A, 439-454.

-concludes that subvocalization allows the creation of a supplementary articulatory code that is produced and utilized in parallel with other aspects of reading.

Barry, W. J. (1984). Segment or Syllable? A reaction-time investigation of phonetic processing. Language and Speech, 27(1), 1-15.

-studies examined whether the segment or the syllable is the point in the phonetic processing of the acoustic input at which the 1st perceptual decision is made in the speech-perception hierarchy.

Batstone, S., & Tuomi, S. K. (1981). Perceptual characteristics of female voices. Language and Speech, 24, 111-123.

-attempted to define perceptually salient characteristics in the voices of 30 female undergraduates as perceived by female and male listeners.

Beckman, M., & Atsuko, S. (1984). Spectral and perceptual evidence for CV coarticulation in devoiced /si/ and /syu/ in Japanese. Phonetica, 41, 61-71.

-Results suggest that a supposedly lower level coarticulation between the fricative and the vowel can occur before a higher-level process deletes the vowel, contradicting the order implied by traditional accounts of speech as a translation of discrete phonological units.

Berkovits, R. (1981). Are spoken surface structure ambiguities perceptually unambiguous? Journal of Psycholinguistics Research, 10, 41-56.

-results indicate that perception of intonation is affected by the interpretive bias of an ambiguous sentence, and that the strength of the preferred reading is attenuated to a greater degree by an opposing context than by opposing intonation.

Bernstein, L. E. (1983). Perceptual development for labeling words varying in voice onset time and fundamental frequency. Journal of Phonetics, 11, 383.

-child subjects, in contrast with adults, did not use F0 as a factor in judging the voicing of the prevocalic stops /g/ and /k/.

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. Infant Behavior and Development, 4, 247-260.

-results indicate that the syllable-like stimuli discriminated better than the non-syllable-like stimuli even though the physical change from the habituation to the dishabituation stimuli was always the same.

Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception and Psychophysics, 29, 191-211.

-trading relations between speech cues, and the perceptual equivalence that underlies them, appear to derive specifically from perception of phonetic information.

Bisiani, R. (1983). Techniques for computer recognition of speech. Annals of the New York Academy of Sciences, 405, 39-37.

-reviews state-of-the-art techniques for the computer recognition of speech and suggests that although systems have been demonstrated and some are in use, computer speech recognition still has very limited capabilities when compared with human performance.

Blank, M. A., Pisoni, D. B., & McClaskey, C. L. (1981). Effects of target monitoring on understanding fluent speech. Perception and Psychophysics, 29, 383-388.

-conducted a methodological study to determine whether the allocation of processing resources for conscious analysis of the sound structure of a speech signal affects ongoing comprehension or the ultimate level of understanding of a linguistic message.

Blumstein, S. E., & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. Cognition, 10, 25-32.

-discusses evidence for and implications of a theory of acoustic invariance, which holds that invariant acoustic properties can be derived directly from the acoustic signal and that ultimately form the inventory of speech sounds used in natural language.

Blumstein, S. E., & Stevens K. N. (1985). On some issues in the pursuit of acoustic invariance in speech: A reply to Lisker. Journal of the Acoustical Society of America, 77, 1203.

-Three points are considered--the minimal unit for acoustic invariance, the level of linguistic representation over which this unit operates, and the role that acoustic invariance plays in speech and language. Our position emphasizes the role of phonetic features in a theory of acoustic invariance, and we propose a series of working hypotheses to guide research in this area.

Bock, J. K., & Mazzella, J. R. (1983). Intonational marking of given and new information: Some consequences for comprehension. Memory and Cognition, 11, 64-76.

-results showed that comprehension time was faster in sentences in which the new information presented was accented.

Boekaerts, M. (1982). Speech production: No left-to-right serial process. Communication and Cognition, 15, 41-51.

-discusses the cognitive processes involved in speech production, focusing on the conscious and unconscious decisions made by speakers in attempting to communicate a message.

Bond, R. N., & Feldstein, S. (1982). Acoustical correlates of the perception of speech rate: An experimental investigation. Journal of Psycholinguistic Research, 11, 539-557.

-investigates the influence of vocal frequency and vocal intensity on the perception of speech rate at 3 levels of actual speech rate.

Bond, Z. S. (1982). Experiments with synthetic diphthongs. Journal of Phonetics, 10, 259-264.

-it is concluded that the perceptual requirements for identifying a synthetic token as a diphthong may be variable.

Borchgrevink, H. M. (1982). Cerebral mechanisms of complex sound perception: Consequences for the functional evaluation of hearing. Scandinavian Audiology, (Suppl. 16), 135-139.

-claims normal speech contains superfluous information, but if this redundancy is reduced below a certain level, the message communicated will not be adequately comprehended.

Bosshardt, H., & Horman, H. (1982). The influence of suprasegmental information on speech perception of 4 to 6 year old children. Archiv Fur Psychologie, 134, 81-104.

-hypothesized that children of this stage are not able to use suprasegmental information to integrate a sentence into a single unit.

Brokx, J. P., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. Journal of Phonetics, 10, 23-26.

-it was shown that intelligibility of the target message could be manipulated by introducing an artificial constant difference in pitch

between target speech and interfering speech. Intelligibility increased with increasing difference in pitch.

Brown, R. (1981). An experimental study of the relative importance of acoustic parameters for auditory speaker recognition. Language and Speech, 24, 295-310.

-results indicate that (1) fund.-freq mean, formant mean, and formant bandwidth are the most important parameters, of those investigated, for speaker recognition; and (2) although listeners differ in average score recorded, they may be treated as reacting identically to changes in the factors.

Carden, G., Levitt, A., Jusczyk, P. W., & Walley, A. (1981). Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. Perception and Psychophysics, 29, 26-36.

-results are interpreted as evidence that the identification of place of articulation involves phonetic processing and could not be purely auditory.

Carre, R. (1982). Presentation of a European speech research group: The Speech Communication Laboratory at E.N.S.E.R.G. Grenoble (France). Speech Communication, 1, 75-79.

-outlines research done on speech production, analysis and synthesis; psycho-acoustic tests and peripheral auditory system modeling in speech perception; multispeaker and continuous speech recognition systems.

Carrell, T. D., Smith, L. B., & Pisoni, D. B. (1981). Some perceptual dependencies in speeded classification of vowel color and pitch. Perception and Psychophysics, 29 (1), 1-10.

-results are consistent with both P. K. Kuhl's (1975, 1976) and J. L. Miller's (1978) findings but refine the understanding of interaction between dimensions by showing that vowel identification is also dependent on the processing of pitch information.

Christovich, L. A. (1985). Central auditory processing peripheral vowel spectra. Journal of the Acoustical Society of America, 77, 789.

-This paper presents a review of recent experiments in vowel perception done at the Pavlov Institute of Physiology in Leningrad. The data concern three topics: experimental procedures appropriate for the study of phonetic quality perception, processing of the auditory spectral shape of a vowel, and processing of the auditory dynamic spectrum of a vowel.

Chistovich, A., Malinnikova, T. G., & and Stoliarova, E. J. (1982). Perception of one-formant synthetic vowels with quasi-random variations of fundamental period and amplitudes of glottal pulses. Fiziologicheskii Zhurnal SSSR, 68, 1330-1336.

-presents data showing that the quasi-random variations of either the fundamental period or the amplitude of glottal pulses results in roughness of the vowel.

Chistovich, L. A., & Ogorodnikova, E. A. (1982). Temporal processing of spectral data in vowel perception. Speech Communication, 1, 45-54.

-data suggest that the running identification of the stimulus is integrated temporally.

Chodorow, M. S., & Manning, S. K. (1983). Syllable similarity: The effects of differences in vowels, consonants, and order. Quarterly Journal of Experimental Psychology: Human Experimental Psychology. 35A, 139-154.

-in 5 experiments with synthetic and natural speech syllables, a rating task was used to study the effects of differences in vowels, consonants, and segment order on judged syllable similarity.

Clark, J. E. (1981). A low-level speech synthesis by rule system. Journal of Phonetics, 9, 451-476.

-system is designed to be used both as a means of stimulus generation for perceptual research and in conjunction with a high level phonological rule component.

Clark, J. E. (1983). Intelligibility comparisons for two synthetic and one natural speech source. Journal of Phonetics, 11, 37-49.

-results suggest that terminal analog speech synthesizers perform least well when generating approximations to the complex spectra found in fricatives and the release of phase stops.

Coberly, M. S., & Healy, A. (1984). Accessibility of place and manner features and the place/manner dissimilation principles in a learning task. Language and Speech, 27,

-Place/manner dissimilation principle was accessed toward end of testing session, suggesting subjects encountered it as an articulatory constraint as a result of pronouncing stimuli throughout the experiment. Past tense and plural suffixing may be processed in terms of articulatory constraints.

Cole, R. A. (1981). Perception of fluent speech by children and adults. Annals of the New York Academy of Sciences, 379, 92-109.

-discusses the processes by which children and adults understand natural continuous speech - concludes that children use the phonetic information available to them in much the same way as adults do.

Cole, R. A., & Rudnicky, A. I. (1983). What's new in speech perception? The research and ideas of William Chandler Bagley, 1874-1946. Psychological Review, 90, 94-101.

-some of the main results of Bagley's research are compared to those obtained in more recent experiments. It is concluded that many of the most important insights about spoken-word recognition were first offered by Bagley.

Cooper, W. E. (1983). The perception of fluent speech. The Annals of the New York Academy of Sciences, 405, 48-63.

-points out that the perception of fluent speech involves both acoustic-and knowledge-driven sources of information -indicates suprasegmentals provide cues to linguistic factors such as word stress, syntactic structure, and semantic interpretation.

Cosmides, L. (1983). Invariances in the acoustic expression of emotion during speech. Journal of Experimental Psychology: Human Perception and Performance, 9(6), 864.

-results supported the hypothesis that different individuals produce standard acoustic configurations to express emotions. Acoustic properties reflecting contrastive stress consistently varied with emotional context over syntactically and semantically identical utterances.

Crowder, R. G. and Repp, B. H. (1984). Single formant contrast in vowel identification. Perception and Psychophysics, 35, 372-378.

-results suggest a sensory, rather than a judgmental basis for the vowel contrast effects obtained.

Cummings, G., & McCorriston, M. (1981). Evaluation of computer speech for use with CAI for young children. Journal of Computer-Based Instruction, 8, 22-27.

-compared normal speech, speech reproduced by supertalker, and Codec speech for intelligibility. Supertalker did poorly while Codec speech gave identification accuracy similar to that of normal speech.

Cutler, A., & Darwin, C. J. (1981). Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. Perception and Psychophysics, 29, 217-224.

-it is argued that sentence processing involves a flexible use of prosodic information

Dabbs, J. M., & Evans, M. S. (1982). Electronic AVTA: Signal processing for automatic vocal transaction analysis. Behavioral Research Methods and Instrumentation, 14, 461-462.

-developed an integrated circuit device that simplifies computer acquisition of the sound/silence patterns of vocalization in dyadic conversations.

Dannenbring, G. L. (1980). Perceptual discrimination of whispered phoneme pairs. Perceptual and Motor Skills, 51, 979-985.

-conducted 2 experiments to investigate the ability of 12 undergraduates to discriminate between whispered consonants that are differentiated in normal speech on the basis of voicing.

Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset time. Quarterly Journal of Experimental Psychology, 33, 185-207.

-studied the use made of a common fundamental frequency or a common starting time in grouping formants to form phonetic categories.

Darwin, C. J. (1984). Perceiving vowels in presence of another sound: Constraints on formant perception. Journal of the Acoustical Society of America, 76, 1636-1647.

-Two experiments show that formants may only be estimated after properties of the sound wave have been grouped into different apparent sound sources. The first result illustrates a general auditory mechanism for performing perceptual grouping; second result illustrates a mechanism that may use a more specific constraint on vocal tract transfer functions.

Darwin, C. J. & Pearson, M. (1982). What tells us when voicing has started? Speech Communication, 1, 29-44.

-results provide perceptual support to algorithms for detecting voiced excitation that use overall intensity as a decision parameter.

de Haan, H. J. (1982). The relationship of estimated comprehensibility to the rate of connected speech. Perception and Psychophysics, 32, 27-31.

-investigated the relationship between subjective estimates of the comprehensibility of connected, free-running speech and rate of speech for each of 2 types of time-compressed speech: pitch varying speeded speech and pitch-normalized compressed speech.

de Haan, H. J. (1978). A speech-rate intelligibility/comprehensibility threshold for speeded and time-compressed connected speech. US Army Research Institute for the Behavioral and Social Sciences, TP 297, p. 20.

-results interpreted to mean that the speech rapidity threshold (SRT) reflects on intermediate level of information processing involving the perception of the potential for interpretation or comprehension, rather that the complete act of comprehension per se.

Dirks, D. D., Morgan, D. E., & Dubno, J. R. (1982). A procedure for quantifying the effects of noise on speech recognition. Journal of Speech and Hearing Disorders, 47, 114-123.

-results suggest that the proposed adaptive strategy may provide a practical method by which the relative effects of competition on speech recognition may be quantified in an individual listener.

D'Odorico, L., Franco, F., & Vidotto, G. (1985). Temporal characteristics in infant cry and non-cry vocalization. Language and Speech, 28,

-Duration of vocalization is influenced by communicative value and non-segmental features, but in most cases not independently one factor from the other.

Donnenwerth-Nolan, S., Tanenhaus, M. K., & Seidenberg, M. S. (1981). Multiple code activation in word recognition: Evidence from rhyme monitoring. Journal of Experimental Psychology: Human Learning and Memory, 7, 170-180.

-results suggest that multiple codes are automatically accessed in word recognition.

Dubno, J. R., Dirks, D. D., & Langhofer, L. R. (1982). Evaluation of hearing-impaired listeners using a Non-sense Syllable Test: II. Syllable recognition and consonant confusion patterns. Journals of Speech and Hearing Research, 25, 141-148.

-consonant confusion analysis revealed place of articulation errors to be the most frequent, regardless of S's audiometric configuration.

Eimas, P. D. (1981). Infants, speech, and language: A look at some connections. Cognition, 10, 79-84.

-suggests that infant speech research will continue to provide descriptions of the initial state.

Eimas, P. D., & Miller, J. L. (1981). Organization in the perception of segmentation and suprasegmental information by infants. Infant Behavior and Development, 4, 395-399.

-results indicate that the S's speech processing system is sensitive to the organization in the basic elements of human language.

Elliot, L. L., Longinotte, C., Clifton, L., & Meyer, D. (1981). Detection and identification thresholds for consonant-vowel syllables. Perception and Psychophysics, 30, 411-416.

-compared with 10 year olds and adults, 6 year old listeners required a greater increases in stimulus intensity above detection threshold to identify these stimuli (5-formant synthesized CV) at a high performance level.

Ferrero, F. E., Pelamatti, G. M., & Vagges, K. (1982). Continuous and categorical perception of a fricative-affricate continuum. Journal of Phonetics, 10, 231-244.

-results indicate a categorical perception suggesting that the availability of acoustic information in auditory memory is immaterial for producing continuous perception in a discrimination task.

Fledge, J. E., & Brown, W. S. (1982). The voicing contrast between English /p/ and /b/ as a function of stress and position-in-utterance. Journal of Phonetics, 10, 335-345.

-indicates the stop voicing pair /p-b/ were clearly distinguished by voicing in utterance-medial positions, but less so at the margins of utterances.

Foltner, K. A., Beasley, D. S., & White, S. (1979). Time-compressed spondaic words as a measure of speech reception threshold. Journal of Auditory Research, 19, 255-258.

-collected speech reception thresholds from 1 ear of each of 60 19-24 yr old normal-hearing adults. No differences between 2 time-compressed lists of spondees were found.

Foss, D. J. & Gernsbacher, M. A. (1983). Toward a unitary model of phoneme identification. Journal of Verbal Learning and Verbal Behavior, 22, 609-632.

-explores nature of the speech code and role of sentence processing.
-results show predictability effects when words occurred in isolation.

Fourcin, A. J., et al. (1983). Speech perception with promontory stimulation. Annals of the New York Academy of Sciences, 405, 280-294.

-describes simple and noninvasive methods of promontory stimulation that are capable of bringing a measurable improvement in both receptive and productive abilities.

Fowler, C. A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. Journal of Experimental Psychology, 112, 386-412.

-proposes that perceived syllable timing corresponds to the timed sequencing of the vowels as produced and not to the timing either of vowel onsets as conventionally measured or of syllable-initial consonants.

Fox, R. A. (1984). Effect of lexical status on phonetic categorization. Journal of Experimental Psychology: Human Perception and Performance, 10, 525-540.

-data supports a perceptual model wherein phonetic categorization can operate separately from higher levels of analysis.

Fox, R. A. (1983). Perceptual structure of monophthongs and diphthongs in English. Language and Speech, 26, 21-60.

-investigated, by means of multidimensional scaling dimensions underlying the perception of diphthongs in American English and whether such dimensions are radically different from those found in studies utilizing monophthongal vowels.

Fox, R. A. (1985). Multidimensional scaling and perceptual features: Evidence of stimulus processing or memory prototypes? Journal of Phonetics, 13, p. 205.

-Multivariant analysis of variance indicated that subjects' perceptual distance judgments were sensitive to the subphonemic as well as phonemic distinctions among the stimuli. Results are argued to demonstrate covert, subphonemic categories during the paired-comparison task.

Fox, R. A. (1981). Influence of personal characteristics of the speaker on phonetic quality in perception of vowels. Perceptual and Motor Skills, 53, 515-519.

-results suggest that judgments of similarity were made on the basis of linguistically categorized images.

Franco, F. (1984). Differences in manner of phonation of infant cries: Relationship to communicate context. Language and Speech, 27,

-Data showed a relation between duration of vocalization and manner of phonation. Hypothesis- differences in functional meaning of cries are marked by non-segmental regularities.

Fukuda, K. (1983). Click monitoring and the perceptual segmentation of speech sequences. Psychologia: An International Journal of Psychology in the Orient, 26, 214-222.

-it is concluded that speech sequences are perceptually segmented by using both acoustic and syntactic cues.

Gass, S. (1984). Development of speech perception and speech production abilities in adult second language learners. Applied Psycholinguistics, 5, 51-74.

-examined the acquisition of production and perception by adult learners of English, focusing on the voice onset time of initial /b/'s and /p/'s.

Gervasio, A. H. (1984). Computer-assisted analysis of conversation. Behavior Research Methods, Instruments, and Computers, 16, 158-161. 61 computer assisted analysis system

-describes CALAS- parses language of 2 speakers in a conversation into words, phrases and clauses - places verbs and nouns into categories derived from theories of case grammar.

Gilhooly, K. J., & Logie, R. H. (1981). Word age-of-acquisition, reading latencies and auditory recognition. Current Psychological Research, 1, 251-262.

-results support the view that the age-of-acquisition variable mainly affects word production and has little affect on word recognition processes.

Ginzel, A., Pedersen, C. B., Spliid, P. E., & Anderson, E. (1982). The role of temporal factors in auditory perception of consonants and vowels: A study of different age groups. Scandinavian Audiology, 11, 93-100.

-categorical perception of different phonemes was demonstrated in 2 experiments in all age groups tested.

Godfrey, J. J., Syrdal-Lasky, A. K., Millay, K. K., & Knox, C. M. (1981). Performance of dyslexic children on speech perception tests. Journal of Experimental Child Psychology, 23, 401-424.

-results suggest an inconsistency in the dyslexics' phonetic classification of auditory cues. A significant relationship was found between reading level and speech discrimination.

Green, T. R., Payne, S. J., Morrison, D. L. & Shaw, A. (1983). Friendly interfacing to simple speech recognizers. Behavior and Information Technology, 2, 23-28.

-describes improvements to the recognition performance of a simple commercial speech recognizer.

Guzy, J. J. (1982). The acquisition of linguistics knowledge from visible speech spectrograms of ordinary speech: A proposal. International Journal of Man-Machine Studies, 16, 327-332.

-suggests that the computer analysis of speech be based on the direct examination of the visible speech spectrograph.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. Scandinavian Audiology, 11, 79-87.

- a list of 10 spoken Swedish sentences was computer edited to obtain new lists with exactly the same content of sound, but with new sentences. The equality in intelligibility of selected lists was investigated.

Hagerman, B. (1982). Measurement of speech reception threshold: A comparison between two methods. Scandinavian Audiology, 11, 191-193.

-tested a 2 - x 4 - word speech reception threshold method simulated in a computer with the common 10 word method.

Haggard, M., Summerfield, Q., & Roberts, M. (1981). Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading Fo cues in the voiced-voiceless distinction. Journal of Phonetics, 9, 49-62.

-showed that differences in fundamental freq. (Fo) at the onset of periodicity that accompany the voiced-voiceless distinction in initial stop consonants are not substantially diminished in the immediately subsequent Fo contour.

Hansen, J. C., Dickstein, P. W., Berka, C., & Hillyard, S. A. (1983). Event-related potential during selective attention to speech sounds. Biological Psychology, 16, 211-224.

-Results suggest that the attention mechanisms brought into play when sele ing complex phonetic stimuli for further analysis are similar to those engaged when selecting between tones of different frequencies.

Harris, L. B., & Pastore, R. E. (1983). Recognition thresholds for a speech continuum following selective adaptation. Perception and Psychophysics, 34, 268-272.

-authors contend results can't be explained by assuming action of only a simple factor such as fatigue, stimulus contrast, or response contrast.

Harris, M. O., Umeda, N., & Bourne, J. (1981). Boundary perception in fluent speech. Journal of Phonetics, 9(1), 1-18.

-describes a perceptual technique for locating boundaries in continuous reading, preliminary to an eventual definition of "boundary".

Hillenbrand, J. (1983). Perceptual organization of speech sounds by infants. Journal of Speech and Hearing Research, 26, 268-282.

-findings suggest that prelinguistic infants can perceptually organize speech sounds on the basis of auditory properties related to feature similarity.

Hojo, H. (1982). Similarity of Japanese consonant phonemes: An analysis by INDSCAL and Hayashi's Quantification Theory I. Japanese Journal of Psychology, 53, 72-79.

-identified perception dimensions of 13 Japanese consonant phonemes. Similarity responses were analyzed by individual differences multidimensional scaling (INDSCAL) and quantification theory.

Horsman, L. (1983). Disabled individuals can talk to their computers. Rehabilitation Literature, 44, 71-74.

-discusses a system that provides for data entry by voice - specifically the shadow /VET (voice-entry-terminal), has acceptable recognition reliability and is affordable.

Howell, P., Powell, D., & Khan, J. (1983). Amplitude contour of the delayed signal and interference in delayed auditory feedback tasks. Journal of Experimental Psychology: Human Perception and Performance, 9(5), 772.

-it is proposed in this article that the disturbance arises from the disruptive effects caused by the rhythm of the delayed signal and actual identity of the delayed speech does not matter.

Ishida, H. (1982). Speech compression and microcomputer. Japanese Journal of Special Education, 20(3), 1-8.

-applications of speech compression is discussed in relation to the information-processing ability of the auditory system.

Jaeger, J. J. (1984). Assessing the psychological status of the vowel shift rule. Journal of Psycholinguistic Research, 13, 13-36.

-demonstrates certain entities can be psychologically real either because they have been brought to the speaker's conscious attention as part of their education or because they have been instituted from the orthographic system of their language.

Jenkins, J. J., & Franklin, L. D. (1982). Recall of passages of synthetic speech. Bulletin of the Psychonomic Society, 20, 203-206.

-results indicate that with a little practice there is basically no difference in recall performance as a function of the intonational pattern used.

Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in "vowelless syllables". Perception and Psychophysics, 34, 441-450.

-findings challenge traditional accounts of vowel perception and point toward important sources of dynamic information.

Jusczyk, P. W., Smith, L. B., & Murphy, C. (1981). The perceptual classification of speech. Perception and Psychophysics, 30, 10-23.

-results suggest that there is information available in the full-formant chirps, but not in the 2-formant chirps, which allows Ss to group the sounds into classes corresponding to the identity of the initial consonant sounds.

Kahn, D., Rabiner, L. R., & Rosenberg, A. E. On duration and smoothing rules in a demisyllable-based isolated word recognition system. Journal of the Acoustical Society of America, 75, p. 590

-A simple rule that reduces length of rhyme demisyllables in non word-final stressed syllables to approx. half their isolated-syllable duration provides recognition accuracy as high as more complex algorithms.

Katsuki, J., Speaks, C. E., Penner, S., & Bilger, R. C. (1984). Application of theory of signal detection to dichotic listening. Journal of Speech and Hearing Research, 27, 444-448.

-suggests that findings don't support the hypothesized relation between direction of ear advantage and hemisphere asymme tries for speech and language processing.

Keating, P. & Blumenstein, S. (1978). Effects of transition length on perception of stop consonants. Journal of the Acoustic Society of America, 64, 57-64.

-results suggest slope and duration of formant transitions seem to contribute minimally to the perception of place of article in stop consonants.

Kewley-Port, D., & Luce, P. A. (1984). Time-varying features of initial stop consonants in auditory running spectra: a first report. Perception and Psychophysics, 35, 353-360.

-use of visual displays of linear prediction smoothed spectra identify place of articulation of initial voiced stops from time varying features.

Kohler, K. J. (1983). Prosodic boundary signals in German. Phonetica, 40, 89-134.

-Article discusses interplay of rhythmic and semantic structuring of German speech signalled by prosodic features, especially Fo and duration marginally also by speech pauses.

Kohler, K. J. (1985). Fo in the perception of lenis and fortis plosives. Journal of the Acoustical Society of America, 78, 21.

-The theory that we are here dealing with a general auditory phenomenon instead of a production-perception relationship is refuted on the basis of the production and perception data available for German and for English.

Kong, K., Tsoi, T.S., & Chu, S.M. (1983). A syllable recognition system based on peak and latency measures. Acta Psychologia Taiwanica, 49-60.

-tested new concept of feature extraction in a syllable recognition system of an analog speech terminal. Cantonese phoneme vocabulary of 19 used.

Krause, S. E. (1982). Developmental use of vowel duration as a cue to postvocalic stop consonant voicing. Journal of Speech and Hearing Research, 25, 388-393.

-qualitativ comparisons between the production and perception data revealed parallel refinement in the use of vowel duration as a function of age.

Krulee, G. K., Tondo, D. K., & Wightman, F. L. (1983). Speech perception as a multilevel processing system. Journal of Psycholinguistic Research, 12, 531-554.

-suggests speech signal contains features processed by 3 subsystems: a prosodic system and a 2 level concurrently operating system in which voice analyzing system makes decisions in order to inform word-processing system.

Kuhl, P. K., & Padden, D. M. (1982). Enhanced determinability at the phonetic boundaries for the voicing feature in macaques. Perceptions and Psychophysics, 32, 542-550.

-results demonstrate that discrimination performance was always best for between-category pairs of stimuli, thus replicating the "phoneme boundary effects" seen in adult listeners and in human infants.

Kurowski, K., & Blumstein, S. Perceptual integration of the murmur and format transitions for place of articulation in nasal consonants. Journal of the Acoustical Society of America, 76, p. 383.

-Results showed murmer provided as much information for perception of place of artic. as did transitions. Highest scores for place of artic. were obtained with both kinds of info. Thus, the combination form an integrated property for the perception of place of artic.

Kuwabara, H. (1982). Perception of CV-syllables isolated from Japanese connected speech. Language and Speech, 25, 175-183.

-results indicate that at least 2 syllables, 1 preceding and 1 following are necessary to provide an acoustic environment for the correct identification of a CV syllable.

Lacerda, F. P. (1982). Acoustic perceptual study of the Portuguese voiceless fricatives. Journal of Phonetics, 10, 11-22.

-findings demonstrated an unexpected ability of the perceptual system to compensate for static filtering effects.

Ladd, R., & Silverman, K. (1984). Vowel intrinsic pitch in connected speech. Phonetic, 41, 31-40.

-By comparing test vowels in comparable segmental and prosodic environments (German), it was shown that IP (intrinsic pitch) effect does occur in connected speech.

Lawson, E. A. & Gaillard, A. W. (1981). Evoked potentials to consonant-vowel syllables. Acta Psychologica, 49, 17-25.

-results show that consonants of long duration were perceived later than plosives but well before the onset of the vowel.

Lawson, E. A., & Gaillard, A. W. (1981). Mismatch negativity in a phonetic discrimination task. Biological Psychology, 13, 281,288.

-tested the hypothesis that in a discrimination task, the number of phonetic features available determines the latencies and amplitudes of the components of the evoked potentials.

Leather, J. (1983). Speaker normalization in perception of lexical tone. Journal of Phonetics, 11, 373-382.

-suggests that in the perceptual processing of F0 , phonetic decisions are referenced to an inferred scaling of the source voice range.

Liberman, A. M. (1982). On finding that speech is special. American Psychologist, 37, 148-167.

-it was found that specialized processes of phonetic perception had been made to conform to the acoustic consequences of the way articulatory movements are regulated.

Liberman, A. M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. Perception and Psychophysics, 30, 133-143.

-results indicate that the effectiveness of the silence cue was a result of distinctively phonetic (as against generally auditory) processes.

Lindgren, R., & Lindbolm, B. (1983) Speech perception processing. Scandinavian Audiology, (Suppl. 18), 57-80.

-discusses theories of speech perception processing, taking into consideration the interaction of acoustical, physiological, neurological, psychological, phonetic, and linguistic factors.

Linell, P. (1982). The concept of phonological form and the activities of speech production and speech perception. Journal of Phonetics, 10, 37-72.

-proposes that the phonological form of a given linguistic expression should be construed as a phonetic plan, i.e., a plan for performing a phonetic-behavioral act that can be perceived and comprehended by language users as the linguistic expression in question.

Lisker, L. (1985). The pursuit of invariance in speech signals. Journal of the Acoustical Society of America, 77, 1199.

-The view that this is not the phoneme, but rather the phonetic feature, to which an acoustic invariant might be attributed, raises two questions: (a) Since segments sharing a feature are rarely judged to constitute a single sound, the search for a feature-specific invariant, whose function is to explain perceptual constancy, is deprived of its essential motivation, and (2) there is no more reason to expect the acoustic cues to a feature to be context-independent than is the case with the phoneme. What seems more likely is to find that some phonemes, and some features, are more invariantly marked in the speech signal than others.

Lisker, L., & Baer, T. (1984). Laryngeal management at utterance-internal word boundary in American English. Language and Speech, 27,

-Acoustic and physiological data obtained from one American English speaker who produced utterances containing /b/ and /p/ in a variety of contexts showed at least 5 patterns of lip-larynx coordination.

Luce, P. A., Feutstel, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural speech. Human Factors, 25, 17-32.

-results suggest that difficulties observed in the perception and comprehension of synthetic speech are due, in part, to increased processing demands in short-term memory.

Lushchikhina, J. M., & Solovova, L. M. (1982). Speech and auditory functions of listeners in communication systems. Psikoligicheskii Zhural, 3, 111-119.

-speech functions tended to be related to certain mental processes and personalities of S's, a finding that suggests the need for their incorporation into views on the general structure of intelligence.

MacKain, K. S. (1982). Assessing the role of experience on infants speech discrimination. Journal of Child Language, 9, 527-542.

-author argues that phonetic input cannot be specified and "experience" cannot be defined in this context without knowing how infants perceptually structure speech input.

MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. Applied Psycholinguistics, 2, 369-390.

-results show classic categorical perception by American English speaking controls.

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. Science, 219, 1347-1349.

-findings suggest that the infants capacity to reproduce speech sounds in prelinguistic babbling may rest on a predisposition of the left hemisphere to recognize the sensorimotor connections between the auditory structure of speech and its articulatory source.

Mann, V. A., & Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. Cognition, 14. 211-235.

-suggests that differences in the speech and nonspeech perception for the experiment reflect the different properties of auditory and phonetic modes of perception.

Marcus, S.M. (1981). Associative coding and word boundary detection. IPO Annual Progress Report, Report 16, 49-56.

-author's (1979, 1981) relatively time-flexible approach to speech perception through context-sensitive coding and the use of a direct acoustic-lexical mapping is used as a basis of 3-phase framework of the work recognition process.

Marcus, S.M. (1983). From "past history" to "interactive activation" in speech recognition. IPO Annual Progress Report, Report 18, 26-31.

-proposes an alternative to the strictly sequential approach to the speech signal and its processing and recognition.

Marcus, S. M. (1981). ERIS-context sensitive coding in speech perception. Journal of Phonetics, 9, 197-220.

-discusses how ERIS, a computer speech recognizer based on a set of independent context-sensitive codes, demonstrates the validity and power of representing speech by a nonsequential associative approach.

Marcus, S. M. (1981). Acoustic determinants of perceptual center (P-center) location. Perception and Psychophysics, 30, 247-256.

-hypothesizes that the P-center location for a given stimulus is independent of the nature of adjacent stimuli in a sequence of speech sounds.

Marslen-Wilson, W. D., & Tyler, L. K. (1981). Central processes in speech understanding. Philosophical Transactions of the Royal Society of London, 295, 317-332.

-study suggests that psycholinguistic processes are organized to allow optimally effective use of the information carried by the speech signal as it becomes available over time.

Martin, J. G., & Bunnell, H. T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. Journal of Experimental Psychology: Human Perception and Performance, 8, 473-488.

-cross-spliced words to confuse coarticulatory effects in CVCV words - reaction time was slowed and also attracted more false alarms.

Massaro, D. W., & Cohen, M. M. (1983). Consonant/vowel ratio: An improbable cue in speech. Perception and Psychophysics, 35, 501-505.

-presents data in support of the idea that consonant duration relative to vowel duration is a cue to the voicing of postvocalic consonants.

Massaro, D.W., & Cohen, M.M. (1983). Evaluation and integration of visual and auditory information in speech perception. Journal of Experimental Psychology: Human Perception and Performance, 9(5), 753.

-3 experiments were carried out to investigate the evaluation and integration of visual and auditory information in speech perception. Results provide strong evidence for a fuzzy logical model of perceptual recognition.

Massaro, D. W., & Cohen M. M. (1983). Phonological context in speech perception. Perception and Psychophysics, 34, 338-348.

-speech perception can be viewed in terms of the listener's integration of 2 sources of info.: acoustic features transduced by the auditory receptor system and the context of the linguistic method.

Massaro, D. W., & Hary, J. M. (1984). Categorical results, categorical perception and hindsight. Perception and Psychophysics, 35, 586-588.

-areas discussed include categorical perception criteria, number of physical continua present in an ordered set of stimuli, and earlier theories of cat. perception.

Massaro, D.W., Thompson, L., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. Journal of Experimental Child Psychology, 41, 93-113.

-results showed children to be poorer lip readers than adults. Lip reading ability correlated positively to size of visual contribution to bimodal speech perception. Results contradict categorical perception of speech events and any non-independence in the evaluation of auditory and visual information in speech perception.

May, J. G. (1981). Acoustic factors that may contribute to categorical perception. Language and Speech, 24, 273-284.

-results suggest that voicing alone, or in combination with acoustic information about the lower formants, may be a necessary condition for continuous perception.

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. Journal of Verbal Learning and Verbal Behavior, 20, 298-305.

-an interpretation of the results is advanced in which the syllable is considered a processing unit in speech perception.

Miceli, G. (1982). The processing of speech sounds in a patient with cortical auditory disorder. Neuropsychologia, 20, 5-20.

-tasks exploring phonological analysis showed an impairment limited to the stop consonants within this framework, processing of feature place was selectively disturbed, while voicing was normally analyzed.

Miller, J. L., & Eimas, P. D. (1983). Studies on the categorization of speech by infants. Cognition, 13, 135-165.

-findings indicate that the infant possesses finely tuned, linguistically relevant perceptual abilities, which facilitate and shape the task of language acquisition.

Miller, J., Aibel, I. L., Green, K. (1984). On the nature of rate-dependent processing during phonetic perception. Perception and Psychophysics, 35, 5-15.

-findings indicated that during phonetic perception, listeners accommodate for changes in the physical rate of speech, not for changes in its subjective rate.

Miller, J., & Grosjean, F. (1981). How the components of speaking rate influence perception of phonetic segments. Journal of Experimental Psychology: Human Perception and Performance, 7, 208-215.

-two studies investigated the way in which the components of speaking rate, articulation rate, and pause rate combine to influence processing of the silence-duration cue for the voicing distinction in medial stop consonants.

Miller, J., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: a reanalysis and some implications. Phonetica, 41, 215-225.

-There was indeed substantial variation in articulation rate even within single utterance of single talker. Results contrast with research which suggests rate variability is largely due to changes in pausing and less to actual articulation rate.

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. Journal of Phonetics, 9, 283-303.

-results indicate that the position of /r/ and /l/ in a word had an effect on identification and production by the Japanese Ss.

Mohanty, A. K. (1984). On psycholinguistics: A retrospect. Psycho-lingua.

-traces dev of psycholinguistics addresses present options in integrating linguistics and psychology.

Morton, J., Marcus, S. M., & Ottley, P. (1981). The acoustic correlates of 'speech like': A use of the suffix effect. Journal of Experimental Psychology: General, 110, 568-593.

-results reinforce the view of a distinction between central, subjectively controllable factors and a strong precategorical effect.

Niederjohn, R. J., Mliner, D. G. (1982). The effects of high-pass and low-pass filtering upon the intelligibility of speech in white noise. Journal of Auditory Research, 22, 189-199.

-results provide information on the contribution to intelligibility of various frequency ranges of the speech spectrum.

Niemeyer, W., & Starlinger, J. (1981). Do the blind hear better? Investigations of auditory processing in congenital or early acquired blindness: II. Central functions. Audiology, 20, 510-515.

-hypothesis of a better utilization of auditory information after the loss of the visual information channel was tentatively confirmed, and may be ascribed to the plasticity of the CNS.

Nooteboom, S. G. (1981). Lexical retrieval from fragments of spoken words: Beginnings verse endings. Journal of Phonetics, 9, 407-424.

-it is concluded that the decision process in word retrieval is monitored in a more complex way than can be accounted for by a threshold-type model.

Nooteboom, S. G. (1983). The temporal organization of speech and the process of spoken-word recognition. IPO Annual Progress Report, (Report 18), 13-26.

-Suggest that the temporal organization of synthetic speech of less than optimum quality should be adapted to the temporal course of speech perception.

Nooteboom, S. G., & Doodeman, G. J. (1982). Speech quality and word recognition from fragments of spoken words. IPO Annual Progress Report, (Report 17), 46-50.

-examines effect of differences in speech quality, caused by differences in the degree of data reduction in vocoder speech, on the relative number of speech sounds required for the correct recognition of polysyllabic words.

Nusbaum, H. C., Schwab, E. & Sawusch, J. R. (1983). The role of "chirp" identification in duplex perception. Perception and Psychophysics, 33, 323-332.

-results suggest that listeners do not need to perceptually integrate F2 transitions or F2 and F3 transition pairs with the base in duplex perception.

Ohde, R. N. (1982). Adaptation of voicing: Effects of ear presentation and acoustic energy variables. Journal of Phonetics, 10, 265-278.

-determined the effects of ear presentation, intensity, and number of repetitions of adaptors varying in VOT on changes in stimulus rating of boundary and nonboundary stimuli.

Ohde, R. Fundamental frequency as an acoustic correlate of stop consonants voicing. Journal of the Acoustical Society of America, 75, p. 224.

-F0 contours were nearly identical for voiceless unaspirated stops and voiceless aspirated stops, both had significantly higher F0 values than voiced stops. Data do not support simple rise-fall dichotomy in F0 at VOT as invariance correlate. Consistent with absolute value of F0 influenced by position of hyoid bone and height larynx.

Osbourne, D. K. (1979). Interconsonantal difference as a predictor of the perception of consonants. Journal of Auditory Research, 19, 247-254.

-demonstrated that a psychophysically determined index of sound similarity-interconsonantal difference - can predict errors made in the aural perception of speech sounds.

Oshrin, S. E., & Siders, J. A. (1984). Discrimination of computer-synthesized speech. Perceptual and Motor Skills, 59, 619-622.

-examined effects of learning on discrimination of computer-synthesized speech by presenting 100 computer-produced monosyllabic words to 2 groups of 15 adults.

Ostreicher, H.J., & Sharf, D.J.S. (1976). Effects of coarticulation on the identification of deleted consonant and vowel sounds. Journal of Phonetics, 4, 285-301.

-experimental findings support (1) coarticulation effects are perceived and used to identify adjacent sounds in conversational speech; (2) adjacent phoneme perception involves parallel processing of features; and (3) directional influence (vs. cohesiveness of CV unit) is the major factor in the perception of coarticulation.

Paliwal, K. K., Lindsay, D. & Ainsworth, W. A. (1983). Correlation between production and perception on English vowels. Journal of Phonetics, 11, 77-83.

-results supported the rejection of the hypothesis that a listeners refers to his/her own articulation for perceiving speech.

Paliwal, K. K. (1984). Effectiveness of different vowel sounds in automatic speaker identification. Journal of Phonetics, 12, 17-21.

-examined effectiveness of 11 English vowels in automatic speaker identification task.

Paliwal, K. K., & Ainsworth, W. A. (1985). Dynamic frequency warping for speaker adaptation in automatic speech recognition. Journal of Phonetics, 13, p. 123.

-A nonlinear spectral normalization procedure for speaker adaptation is proposed. Procedure employs dynamic programming algorithm for warping the log power spectrum along frequency axis and does not require prior knowledge about input speaker. Does not improve performance of speaker independent speech recog. systems.

Petersen, H. (1982). A dichotomy in the English sound system. Journal of Phonetics, 10, 439-451.

-shows that diphthongization of the high vowels /i:/, /u:/ vs nondiphthongization of these vowels gives rise to a dichotomy in the English sound system with profound effects throughout that system.

Pickett, J. M. (1983). Theoretical considerations in testing speech perception through electroauditory stimulation. Annals of the New York Academy of Science, 405, 424-434.

-characterizes some of the basic phonetic features of speech that are necessary for its communication and relates those features to some problems in the implant testing.

Pilon, R. (1981). Segmentation of speech in a foreign language. Journal of Psycholinguistic Research, 10, 113-122.

-results challenge the credibility of traditional associationist accounts of language acquisition and speech perception.

Pisoni, D. B. (1981). Some current theoretical issues in speech perception. Cognition, 10, 249-259.

-discusses 9 areas that will probably receive greater attention in the next few years.

Pisoni, D. B., Aslin, R. N., Perey, A. J. & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. Journal of Experimental Psychology: Human Perception and Performance, 8, 297-314.

-results demonstrate that the perceptual mechanisms used by adults in categorizing stop consonants can be easily and quickly modified with laboratory techniques.

Pisoni, D. B., Carrell, T. D., & Guns, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. Perception and Psychophysics, 34, 314-322.

-results demonstrates context effects are not peculiar to the perception of speech signals or to normalization of speaking rate.

Polka, L., & Strange, W. (1985). Perceptual equivalence of acoustic cues that differentiate /r/ and /l/. Journal of the Acoustical Society of America, 78, 1187.

-A spectral cue and a temporal cue were varied in synthetic versions of "rock" and "lock". Phonetic identification and oddity discrimination tasks showed a trading relation between the two cues. Discrimination accuracy was ordered: Facilitating cues > one-cue > conflicting cues, indicating that perceivers discriminated on the basis of an integrated phonetic percept.

Pols, L. C. W., & Schouten, M. E. H. (1985). Plosive consonant identification in ambiguous sentences. Journal of the Acoustical Society of America, 78, 33.

-The hypothesis tested in this paper is that the acoustic-phonetic information at the sentence level (irrespective of the meaning of the sentence) makes it easier to interpret those cues to plosive consonant identify which are contained in vocalic transitions. Dutch sentences in experiments involving the perception of isolated VCV or CVC utterances, whether spoken in isolation or excised from longer utterances, the contribution of the CV transition to plosive consonant perception may be underestimated.

Port, R. F., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. Perception and Psychophysics, 32, 141-152.

-results suggest that the consonant-vowel ratio serves as a primary acoustic cue for English voicing in syllable-final position and imply that this ratio possibly is directly extracted from the speech signal.

Rakerd, B. (1984). Vowels in consonantal context are perceived more linguistically than are isolated vowels: Evidence from an individual differences scaling study. Perception and Psychophysics, 35, 123-136.

-examines whether the presence of neighboring consonants can exert a contextual influence on vowel perception.

Rakerd, B., & Verbrugge, R. R. (1985). Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study vowels. Journal of the Acoustical Society of America, 77, 296.

-Perceptual dimensions corresponding to the advancement, height, and tenseness vowel features were recovered. Given the determinancy of individual differences scaling, this finding is taken to provide strong evidence for the perceptual significance of those features. The perceptual dimensions are considered in relation to various acoustic parameters of the stimuli employed in this study. They are also considered in relation to perceptual dimensions that have been observed in other vowel scaling studies.

Rakerd, B., Dechovitz, D. R., & Verbrugge, R. R. (1982). An effect of sentence finality on the phonetic significance of silence. Language and Speech, 25, 267-282.

-results are consistent with the principle that silence can have phonetic significance for a listener only when it is perceived to have occurred in a stretch of speech that was articulated continuously.

Recasens, D. (1984). Vowel to vowel coarticulation in Catalan VCV sequences. Journal of the Acoustic Society of America, 76, 1624-1635.

-results show that degree of V to V coarticulation in linguopalatal fronting and F2 frequency varies monotonically and inversely with the degree of tongue-dorsum contact, carryover effects larger than anticipatory affects. V to V coarticulation in VCV sequences is dependent on mechanical constraints of tongue dorsum to achieve closure during consonant production.

Reineke, T. (1981). Simultaneous processing of music and speech. Psychomusicology, 1, 58-77.

-combination of considerable dichotic interference of melodies with melodies and digits (spoken) with digits and the relative lack of interference between the 2 classes of stimuli suggest that separate information processing systems may be used for music and speech.

Remez, R. E., & Rubin, P. E. The stream of speech. Scandinavian Journal of Psychology, 24, 63-66.

-use of sinusoidal replicas of speech signals reveals that listeners can perceive speech solely from temporally coherent spectral variation of nonspeech acoustic elements.

Remez, R. E., Rubin, P. E., & Pisoni, D. B. (1983). Coding of the speech spectrum in three time-varying sinusoids. Annals of the New York Academy of Sciences, 405, 485-489.

-phonetic perception may depend on properties of coherent spectrum variation, a 2nd-order property of the acoustic signal.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. Science, 212, 947-950.

-time-varying properties of 3-tone sinusoidal replicas of natural utterances are apparently sufficient to support perception of the linguistics message in the absence of traditional acoustic cues for phonetic segments.

Repp, B. H. (1981). Two strategies in fricative discrimination. Perception and Psychophysics, 30, 217-227.

-results support hypothesis that influences of vocalic context on fricative identification are tied to a phonetic mode of perception.

Repp, B. H. (1981). Perceptual equivalence of two kinds of ambiguous speech stimuli. Bulletin of the Psychonomic Society, 18, 12-14.

-stimuli from 2 synthetic /da/ - /ga/ continua, one generated by parameter interpolation, the other by adding wave-forms of the endpoint stimuli in varying proportions, were presented to Ss - results suggest that the 2 procedures yield equally ambiguous stimuli.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.

-findings provide strong empirical evidence for the existence of a speech-specific mode of perception.

Repp, B. H. (1983). Bidirectional contrast effects in the perception of VC-CV sequences. Perception and Psychophysics, 33, 147-155.

-it is hypothesized that the presumed contrast effects do not result from any direct interaction of spectral cues across the closure interval but are due to perceptual information conveyed by the closure itself.

Repp, B. (1984). Effects of temporal stimulus properties on perception of the [sl] - [spl] distinction. Phonetica, 41, 117.

-Findings are interpreted as reflecting a perceptual compensation for coarticulatory shortening of [s] before stop consonants, in conjunction with contrastive interactions between the perceived durations of adjacent acoustic segments. Results suggest local temporal signal properties, as distinct from global perceived speaking rate, are an important factor in phonetic perception.

Repp, B. (1984). Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. Language and Speech, 27, 3.

-Bursts contributed to perception of stop manner by reducing amount of silence required to perceive a stop. Burst amplitude was a cue for both manner and place of stop articulation. Silent closure interval is the primary cue. All these effects probably reflect listener's tacit knowledge of systematic acoustic differences in natural speech.

Repp, B. Role of release burst in perception of [s]-stop clusters. Journal of the Acoustic Society of America, 75, 1219.

-Series of experiments found several trading relation sets in perception of stop consonants. All experiments revealed listeners are remarkably sensitive to presence to even very weak release bursts.

Repp, B. H., & Williams, D. R. (1985). Influence of following context on perception of the voiced-voiceless distinction in syllable-final stop consonants. Journal of the Acoustical Society of America, 78, 445.

-The focus is on temporal cues to the distinction, with vowel duration and silent closure duration as the primary and secondary dimensions, respectively. Adding a second syllable to a monosyllable increases the number of voiced stop consonant responses, as does shortening of the closure duration in disyllables.

Rietveld, A. C. M., & Gussenhoven, C. (1985). On the relation between excursion and prominence. Journal of Phonetics, 13, 229.

-A difference of 1.5 semitone is sufficient to cause a difference in the perception of prominence, and prominence judgments of different excursion sizes follow a Hertz scale more closely than a semitone scale.

Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. Perception and Psychophysics, 30, 309-314.

-results strongly suggest that auditory rather than phonetic levels of processing are influenced in selective adaptation.

Rosen, S. M., & Howell, P. (1981). Plucks and bows are not categorically perceived. Perception and Psychophysics, 30, 156-168.

-authors were unable to replicate the finding by Cuttin and Rosner (see PA, Vol 53:10801) that discrimination measured in a ABX task was best around 40 msec, the category boundary.

Rosenhamer, H. (1983). Some mathematical tools for speech preprocessing and speech perception modeling. Scandinavian Audiology, (Suppl. 18), 71-79.

-discusses (1) methods of preprocessing speech
(2) how speech signal is received in the ear
(3) principles of speech codification in cochlear
(4) simple mathematical models for estimation and identification of speech elements.

Rosner, B. Perception of VOT continua: A signal detection analysis. Journal of the Acoustical Society of America, 75, 1231.

-Relationship between discrimination and identification d's varies with response constraints, number of noticeable differences in stimulus array and stability of judgments in tasks. Argues against dual coding model; in favor of continuous model.

Samuel, A. G. (1981). The role of bottom-up confirmation in the phonemic restoration illusion. Journal of Experimental Psychology: Human Perception and Performance, 7, 1124-1131.

-data indicate that phonemic restoration depends on the interplay between the listener's expectations and the acoustic signal.

Samuel, A. G. (1982). Phonetic prototypes. Perception and Psychophysics, 31, 307-314.

-results suggest a prototype based representation for phonetic categorization. Several process models using such a representation are considered.

Samuel, A. G., Kat, D., & Tarter, V. (1984). Which syllable does an intervocalic stop belong to? A selective adaptation story. Journal of the Acoustical Society of America, 76, 1653.

-Pattern of adaptation effects and noneffects indicate that intervocalic stop consonants are perceptually more like syllable initial than syllable final ones. Consonant in VCV is apparently treated as different from VC or CV consonants.

Sawusch, J. R., & Jusczyk, P. (1981). Adaptation and contrast in the perception of voicing. Journal of Experimental Psychology: Human Perception and Performance, 7, 408-421.

-results are interpreted as supporting the position that selective adaptation effects arise at an early, auditory level of processing that is responsive to the spectral overlap between adaptor and test items.

Sawusch, J. R., & Nusbaum, H. C. (1983). Auditory and phonetic processes in place perception for stops. Perception and Psychophysics, 34, 560-568.

-on basis of results a process model of speech perception is described.

Schäfer-Vincent, K. (1983). Pitch period detection and changing: method and evaluation. Phonetica, 40, 177-202.

-An algorithm is presented which detects quasi-periodic parts in a given speech signal, and provides the fundamental frequency for these parts. The algorithm analyses the structure of time-amplitude representation of the speech signal. It determines potential 'period twins' and then determines if they are apart of a 'period chain'.

Scheffers, M. T. (1982). The role of pitch in the perceptual separation of simultaneous vowels: II. IPO Annual Progress Report, (No. 17), 41-45.

-results support a theory of a profile analysis of the spectrum of the combination, aided by separation of low formants as belonging to different vowels on the basis of the pitch or pitches perceived in the complex sound.

Schwab, E. C., Sawusch, J. R., & Nusbaum, H. C. (1981). The role of second formant transition in the stop-semivowel distinction. Perception and Psychophysics, 29, 121-128.

-results are interpreted as arguing against models that incorporate transition rate as a cue to phonetic distinctions. It is shown that the phonetic interpretation of the obtained adaptation results is not justified.

Scott, D. R., & Cutler, A. (1984). Segmental phonology and the perception of syntactic structure. Journal of Verbal Learning and Verbal Behavior, 23, 450-466.

-research in speech production has shown syntactic structure reflected in segmental phonology - whether such segmental effects can be used as cues to syntactic structure in speech perception is examined.

Scott, D., Isard, S. D., & de Boysson-Bardies, B. (1985). Journal of Phonetics, 13, 155.

-Two experiments show that the perception of regular occurring stress events is not specific to "stress-timed: lang.s and is not even specific to speech. Concludes that the phenomenon of regularization cannot be used as evidence for an underlying isochionous rhythm in English.

Searle, C. L. (1982). Speech perception from an auditory and visual viewpoint. Canadian Journal of Psychology, 36, 402-419.

-suggests that the auditory system in humans utilizes a heavily overlapped set of filters, broader bandwidths at high frequencies, which results in good spectral resolution at low frequencies and good temporal resolutions at high frequencies.

Segui, J., Frauenfelder, U., & Mehler, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. British Journal of Psychology, 72, 471-477.

-significant differences were found between phoneme and syllable detection times, suggesting that phoneme detection is highly dependent on syllable identification. In addition, a strong correlation between the phoneme and syllable RTs for each word/nonword item was obtained.

Seldon, H. L. (1981). Structure of human auditory cortex: II. Axon distributions and morphological correlates of speech perception. Brain Research, 229, 295-310.

-possible morphological correlates are discussed for neurolinguistic phenomena such as parallel processing of phonemes and words and unilateral, categorical perception of consonants compared with bilateral recognition of vowels and continuous transitions between them.

Sharf, D. J., & Ohde, R. N. (1981). Recovery from adaption to stimuli varying in voice onset time. Journal of Phonetics, 9, 79-87.

-findings suggest the possibility of a cognitive component in adaptation recovery and the importance of including the recovery process in models of adaption.

Sharf, D. J., & Ohde, R. N. (1982). Recovery from categorical boundary shifts for the vowel duration cue to consonants voicing. Journal of Phonetics, 10, 105-111.

-results reflect 2 stages of recovery -a rapid change in the 1st min and a more gradual change during the following half hour.

Sharf, D. J., & Ohde, R. N. (1984). Effect of formant frequency onset variation on the differentiation of synthesized /w/ and /r/ sounds. Journal of Speech and Hearing Research, 27, 475-479.

-results show a systematic relationship between the differentiation of /r/ and /w/ sounds and difference between the frequency of 2nd and 3rd formants at onset for both adult and child stimuli.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. Applied Psycholinguistics, 3, 243-261.

-for some of the S's, production of the contrast was more accurate than their perception of it.

Sherak, R. (1982). A real-time software voice key and an application. Behavior Research Methods and Instrumentation, 14, 124-127.

-describes a real-time software voice key that can be used to "locate" speech onset and offset within a 4-msec "window" and to distinguish among several classes of speech sounds (e.g. sibilance, nonsibilance, aspiration and silence).

Sherrard, C. A. (1982). "Auditory feedback" or "side tone"? The effects on speech production and intelligibility of auditory stimulation from the larynx. Language and Speech, 25, 283-292.

-argues that the contingency of the variation in the true feedbacks that make them disruptive, while the noncontingency of the variation in the sidetones increases the overall level of auditory interference and enhances speakers' intelligibility as they try to overcome it.

Shinn, P., & Blumstein, S. On the role of the amplitude envelope for the perception of [b] and [w]. Journal of the Acoustical Society of America, 75, 1243.

-Amplitude cues were able to override transition slope, duration and formant frequency cues in perception of stop-glide contrast. Presence of glide amplitude envelope changed stimuli labeled as stops to fricatives-not glides. Amplitude envelope in vicinity of consonantal release is critical for continuant/noncont. contrast.

Simpson, C. A., & Marichionda-Frost, K. (1984). Synthesized speech rate and pitch effects on intelligibility of warning messages for pilots. Human Factors, 26, 509-517.

-study examined whether speech rate and voice pitch of "phoneme" synthesized speech would effect accuracy.

Singer, J. (1981). Some effects of intermodulation distortion on speech intelligibility. Journal of Auditory Research, 2, 201-206.

-experiments in which S's listened to difficult versions of Northwestern University Auditory Test #6.
-results indicate intermodulation distortion version significantly degraded speech intelligibility.

Skinner, M. W., Karstaedt, M. M., & Miller, J. D. (1982). Amplification bandwidth and speech intelligibility for two listeners with sensorineural hearing loss. Audiology, 21, 251-268.

-it was found that the wider the bandwidth, the higher the score; and the greater the amount of speech energy above threshold, the higher the score.

Sneppe, R. (1979). Intelligibility of speech against background noise. Journal de Psychologie Normale et Pathologique, 76, 403-418.

-studied vocal audiometry with a variable spectrum of 4-16 khz using synthetic phrases as stimulus materials.

Sorin, C. (1981). Functions, roles and treatments of intensity in speech. Journal of Phonetics, 9, 359-373.

-results of M. Rossi (1978) on the perception of intensity glides on vowels are compared with present findings on the discrimination of intensity glides at the end of sentences to determine which auditory treatment was used in both cases. It is shown that a "simple" psychoacoustic model could explain the observed results.

Spitzer, J. B., & Osborne, D. K. The effect of open-versus closed-set procedures in the perception of compressed speech. Journal of Auditory Research,

-evaluated open- and closed-set paradigms as procedural considerations in compressed- speech testing.

Stathopoulos, E., & Weismer, G. An aerodynamic study of stress in children and adults.

-Adults produced greater peak oral airflow Vo than children, but only in some contexts. Initial stops were produced with higher Vo than medial, and medial than final, but this trend interacted with position of stress. A context sensitive boundary cue hypothesis, a mechanical constraint hypothesis and a floor-effect hypothesis are introduced to account for the data.

Stemberger, J. R. The nature of /r/ and /l/ in English: Evidence from speech errors.

-All variants of /r/ and /l/ are members of a single consonant phonemes thus do not apparently contain a feature for syllabicness; that must be imposed on the segment from an external source such as syllable structure.

Stevens, K. N. (1983). Acoustic properties used for the identification of speech sounds. Annals of the New York Academy of Sciences, 405, 2-17.

-describes some of the acoustic properties that must be identified by a listener to distinguish between words.

Stevens, K. N. (1982). Toward a feature-based model of speech perception. IPO Annual Progress Report, (No. 17), 36-37.

-suggests data from acoustic and perceptual studies indicate that the acoustic property corresponding to a given feature is independent of the context in which the feature occurs.

Stewart, J. M., & Barach, C. (1981). Preferential scaling of certain selected distinctive features. Journal of Psycholinguistic Research, 10, 167-178.

-results reveal that memory is a crucial factor in saliency with coronal vs. strident, while feature contrasts are more important in judgments of voice vs. coronal and voice vs. strident.

Studdert-Kennedy, M. (1981). The emergence of phonetic structure. Cognition, 10, 301-306.

-outlines a new approach to speech perception which borrows from ecological perspectives, natural phonology, and phonetic theory.

Studdert-Kennedy, M. (1983). Limits on alternative auditory representations of speech. Annals of the New York Academy of Sciences, 405, 33-38.

-suggest that the limits on alternative representatives of speech appear to be that they preserve some minimal, perhaps grossly transposed, spectral structure that is sufficient to specify instantaneous articular placement.

Sugai, K. (1981). Basic sound perception to formulate speech perception. Japanese Journal of Speech Education, 19, 28-36.

-findings show that S's learned to discriminate phonemes using the auditory articulatory circuit they learned to perceive beats corresponding to their body movements.

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

-results suggest that timing should, in the main, be regarded as intrinsic to the acoustical specifications of phonetic events, a view that is compatible with recent reformulations of the problem of timing control in speech production.

Summerfield, Q., Haggard, M., Foster, J., & Gray, S. (1984). Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. Perception and Psychophysics, 35, 203-213.

-investigates whether segments of sound with uniform spectra, devoid of peaks and valleys, can be identified reliably as vowels under certain circumstances.

Suomi, K. (1985). Vowel dep. of gross spectral cues to place of articulation of stop consonants in CV syllables. Journal of Phonetics, 13, 267.

-Present results indicate the vowel exerts a reliable influence on stop-vowel waveform from very beginning interms of gross spectral properties. Consonant influence is strong throughout the 45 ms interval from beginning of CV syllable. Argues specifically against acoustic invariance.

t'Hart, J. (1981). Temporal quantization of speech. IPO Annual Progress Report, (No. 16), 44-48.

-temporal quantization is introduced as a means of assigning duration more easily to syllable-sized segments in speech synthesis.

Tallal, P. (1983). Acoustic coding of speech and normal limits on transfer of information: Discussion paper. Annals of the New York Academy of Science, 405, 64-65.

-computer alteration of the rate of change of specific temporal characteristics of some speech resulted in significant enhancement of speech perception in a group on children with language disorders.

Tanenhaus, M. K., Flanigan, H. P., & Seidenberg, M. S. (1980). Orthographic and phonological activation in auditory and visual word recognition. Memory and Cognition, 8, 513-520.

-results suggest that word recognition entails activation of multiple codes and priming of orthographically and phonologically similar words.

Tansley, B. W., Regan, D., & Suffield, J. B. (1982). Measurements of the sensitivities of information processing channels for frequency change and for amplitude change by a titration method. Canadian Journal of Psychology, 36, 723-730.

-suggest the human auditory system contains functional subunits ("channels") selectively sensitive to FM and AM.

Tarte, R. D. (1982). The relationship between monosyllables and pure tones: An investigation of phonetic symbolism. Journal of Verbal Learning and Verbal Behavior, 21, 352-360.

-it is postulated that frequency may be an important meaning-bearing component and formant frequencies of vowels may be critical in the phonetic symbolism phenomenon.

Tarte, R. D., & O'Boyle, M. N. (1982). Semantic judgments of compressed monosyllables: Evidence for phonetic symbolism. Journal of Psycholinguistic, 11, 183-196.

-analysis showed that S's did judge the monosyllables as different on the basis of both speed and frequency. The implications for phonetic symbolism research are discussed.

Tartter, V. C. (1981). A comparison of the identification and discrimination of synthetic vowel and stop consonant stimuli with various acoustic properties. Journal of Phonetics, 9, 477-486.

-the results suggest that the steady states may mask the transition cues for consonants and dominate vowels perceptually.

Tartter, V. C. (1982). Vowel and consonant manipulations and the dual-coding model of auditory storage: A re-evaluation. Journal of Phonetics, 10, 217-223.

-suggests that the model be modified to assume an auditory memory basis for identification and between-category discrimination.

Tatham, M. A. A. (1985). An integrated knowledge base for speech synthesis and automatic speech recognition. Journal of Phonetics, 13, 175.

-Developments in the cognitive theory of phonetics suggest that production and perception are mutually dependent-modalities of the same system. Speech synthesis and auto. sp. recog. are brought together by sharing a common knowledge base. An appropriate type of representation is outlined.

Treiman, R. (1983). The structure of spoken syllables: evidence from vowel word games. Cognition, 15, 49-74.

-Results provide strong support for validity of onset and rime. Findings do not support division of rime into peak and coda. At least one level of structure must exist between syllable and phoneme.

Tuller, B., Kelso, J. S., & Harris, K. S. (1983). Converging evidence for the role of relative timing in speech. Journal of Experimental Psychology and Human Perception and Performance, 9, 829-833.

-some converging lines of evidence for a functionally significant vowel-to-vowel period in speech and how this may relate to the role of temporal invariance in motor skills in general are discussed.

Umeda, N., & Quinn, A. S. (1981). Word duration as an acoustic measure of boundary perception. Journal of Phonetics, 9, 19-28.

-results showed that word duration increased monotonically with the number of boundary responses made by listeners. Results also suggest that the perception of boundary was strongly influenced by the occurrence frequency of words.

van den Berg, R. J. H., & Slis, I. H. (1985). Perception of assimilation of voice as a function of segmental duration and linguistic context. Phonetica, 42, 25-38.

-Longer durations of clusters at word boundaries gave rise to more "progressive assimilation" at the expense of "regressive assimilation" or "no assimilation". Non words behaved differently from sentences and words.

Ventsov, A. V. (1983). What is the reference that sound durations are compared with in speech perception? Phonetica, 40, 135-144.

-To devise a system of quantitative rules for the automatic phonetic interpretation of natural speech it is necessary to know with what reference the speech sound duration is to be compared. An experiment with a specially arranged set of stimuli has shown that durations of adjacent sounds in an utterance-vowels-are used as the reference for word stress perception in Russian.

Verbrugge, R.R., Shankweiler, D.P., & Edman, T.R. (1976). What information enables a listener to map a talker's vowel space? Journal of the Acoustic Society of America, 60, 198-212.

-prior experience with a talker's speech contributes little to success in vowel identification. Precursors mainly influenced listeners' response biases rather than facilitating vowel identification. Results do not support hypothesis that point vowels provide unique information for normalizing vowel space. Sentence context aids vowel identification by allowing adjustment to a talker's tempo, rather than to the talker's vocal tract.

Walsh, T. Modelling temporal relations within English syllables. Journal of Phonetics, 12, 29.

-Final consonant cluster abbreviation is due in part to tendency to produce CVC syllables with a fixed ratio between CV and final C portions of the syllable. The temporal relations within a syllable can be modelled by a set of ordered rules - based on 2 experiments.

Walsh, T., & Parker, F. (1982). Consonant cluster abbreviation: An abstract analysis. Journal of Phonetics, 10, 423-437.

-proposes that at some point in the derivation of a syllable a transformation rule applies that converts a CV/C structure into C/CV.

Wardrip-Frvin, C., & Peach, S. (1984). Developmental aspects of the perception of acoustic cues in determining the voicing feature of final stop consonants. Language and Speech, 27, pt.4.

-3 yr. olds relied more on temporal cues, 6-yr. olds relied more on spectral cues, adults used both in judging the voicing feature of final stop consonants.

Werker, J. F., & Tees, R. C. Phonemic and phonetic factors in adult cross-lang. speech perception. Journal of the Acoustical Society of America, 75, 1866.

-Results suggest that ontogenetic modification in the perception of non-native phonetic contrasts involves a change in processing strategies rather than sensorineural loss.

Whalen, D. H. (1983). Vowel information in postvocalic fricative noises. Languages and Speech, 26, 91-100.

-results illustrate that fricative noises contain considerable information about preceding high vowels.

Wilson, R. H., Arcos, J. T., & Jones, H. C. (1984). Word recognition with segmented-alternated CVC words: A preliminary report on listeners with normal hearing. Journal of Speech and Hearing Research, 27, 378-386.

-segmented CVC monosyllabic words at the phoneme boundaries and presented them to 60 adults with normal hearing.

Wingfield, A., Lombardi, L., & Sokol, S. (1984). Prosodic features and the intelligibility of accelerated speech: Syntactic verse periodic segmentation. Journal of Speech and Hearing Research, 27, 128-134.

-interactions among variables suggest ways in which prosody ordinarily facilitates determination of syntactic structure in connected speech.

Young, L. L., & Wilson, K. A. (1982). Effects of acetylsalicylic acid on speech discrimination. Audiology, 21, 342-349.

-results indicate that for some persons, aspirin produced a substantial decrease in speech understanding on noise conditions, even though there may not be a decrease in pure-tone sensitivity or speech discrimination in quiet ones.

Zatorre, R. (1983). Category-boundary effects and speeded sorting with a harmonic musical-interval continuum: Evidence for dual processing. Journal of Experimental Psychology: Human Perception and Performance, 9(5), 739.

-results demonstrate that there are changes in discriminability associated with learned categories and suggest that there may be two hierarchically organized stages. A dual-processing model is discussed.

MISSION
of
Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control, Communications and Intelligence (C^3I) activities. Technical and engineering support within areas of competence is provided to ESD Program Offices (POs) and other ESD elements to perform effective acquisition of C^3I systems. The areas of technical competence include communications, command and control, battle management information processing, surveillance sensors, intelligence data collection and handling, solid state sciences, electromagnetics, and propagation, and electronic reliability/maintainability and compatibility.