RECOGNITION OF WORDS FROM THEIR SPELLINGS: INTEGRATION OF MULTIPLE KNOWLEDGE SOURCES

by

NANCY ANN DALY

B.S.E.E., University of Rhode Island
(1985)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS OF THE
DEGREE OF

MASTER OF SCIENCE
IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May, 1987

©Nancy Ann Daly 1987

The author hereby grants to M.I.T. permission to reproduce and to distribute copies of this thesis document in whole or in part.

Signature of Author _____
Department of Electrical Engineering and Computer Science
May 20, 1987

Certified by _____
Victor W. Zue
Associate Professor of Electrical Engineering
Thesis Supervisor

Accepted by _____
Professor Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

JUL 08 1987 Archives

RECOGNITION OF WORDS FROM THEIR SPELLINGS: INTEGRATION OF MULTIPLE KNOWLEDGE SOURCES

by

Nancy Ann Daly

Submitted to the Department of Electrical Engineering
on May 20, 1987 in partial fulfillment of the requirements
for the degree of Master of Science in Electrical Engineering.

Abstract

Because of the acoustic similarities between some letters, automatic recognition of continuously-spoken letters is a difficult task. By constraining the problem to the recognition of spelled words, knowledge of the rules of spelling may be exploited to aid in recognition. This thesis studies the acoustic-phonetic and lexical characteristics of continuously-spelled words to determine how to combine information from these sources of knowledge.

A lexical study using a large dictionary is conducted to quantify some of the rules of spelling. Statistics dealing with the frequency of letter sequences are gathered.

Experiments are performed to determine the sufficiency of acoustic information for the recognition of spelled strings. Both auditory perception tests and spectrogram reading tests are conducted, and results are compared. An acoustic study of the spelling corpus is conducted to determine the characteristics of spelled speech that differ from ordinary speech. The study also examines specific errors made by subjects of the recognition experiments to determine their causes. Experiments in acoustic resolution of the worst substitution errors are also conducted to find acoustic parameters to distinguish between easily confused pairs of letters.

Finally, ways of integrating acoustic-phonetic and lexical knowledge are explored. A model for a spelling recognition that incorporates information from both sources is proposed and discussed.

Name and Title of Thesis Supervisor: Victor W. Zue
Associate Professor of Electrical Engineering

Acknowledgements

There are many people I would like to thank for making it possible for me to do this thesis:

First and foremost, my thesis advisor, Victor Zue, for his support, enthusiasm and guidance. With his encouragement, I have learned and accomplished more in the last two years than I dreamed was possible.

The members of the Speech Group, who have taught me about speech and provided advice and friendship as well. I would particularly like to thank my spectrogram readers, Jim Glass, Caroline Huang, Lori Lamel, John Pitrelli, Stephanie Seneff and Victor Zue for their patience and the care they took with my spectrograms.

Rob Kassel for his help with ALexiS, which made it possible for me to conduct my lexical study.

Mark Randolph for his help with SEARCH, which made it possible for me to conduct my acoustic study.

Peter Nuth for his help in putting this thesis together, and for his support and encouragement throughout my time at MIT.

And finally, my family, especially my parents, for their prayers, love, and patience, and for instilling in me a love of learning.

This research was supported by the National Science Foundation and the Defense Advanced Research Projects Agency.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Speech Recognition

The computer is one of the most important tools employed by people today, and as time goes on, its use will become more widespread and its functions more diverse. Therefore, finding ways to provide graceful communication between humans and computers is both desirable and essential. Currently, people communicate with computers primarily via text, a method which is reliable, but also slow and often awkward. Since voice is the most natural and efficient means of communication for humans, it would be advantageous to provide voice as an alternative method for communication with computers.

1.1.1 Current Speech Recognition Systems

So far, almost all speech recognition systems that have been successfully implemented are speaker-dependent, isolated-word recognizers with limited vocabulary. Such systems use a variety of techniques to recognize words, including template matching and dynamic programming techniques [18]. In this method, the input signal is compared with stored templates using dynamic time warping and a distance measure (e.g., the Itakura distance [8]) until the best match is found. This

technique yields a recognition rate of better than 95% for limited vocabulary tasks in which the system has been trained for a particular speaker. Pattern matching works fairly well for isolated word recognition, but is not readily extendible to continuous speech recognition. In continuous speech, boundaries between words are not clearly defined and coarticulation, the influence adjacent sounds or words have on each other, becomes an important factor.

IBM [9] has developed both a successful speaker-dependent isolated word recognition system and speaker-dependent continuous word recognition system. Both systems employ Hidden Markov Modeling [13], a probabilistic approach to recognizing speech. In this approach, the input speech signal is sliced into segments and statistics are used to find the best phonetic match. Using a vocabulary of 1000 words, the continuous speech recognizer has a success rate of about 91%, and with a vocabulary of 5000 words, the isolated word recognizer is correct 95% of the time.

Other systems, such as HARPY [14] and Hearsay [5], rely more heavily on higher-level speech knowledge. HARPY, which was developed in the 1970s as part of the ARPA speech understanding project, is a continuous speech recognition system that allows a limited set of grammatical constructions. Its recognition rate is over 95%. Similarly, Hearsay, another continuous speech recognizer, uses high-level knowledge of semantics and syntax, but very little low-level knowledge of the acoustic-phonetic features of the signal. With a vocabulary of 1000 words, the system is only able to correctly guess that a word is one of 50 candidates in 70% of all cases. However, Hearsay's overall recognition rate (after syntactic and semantic constraints have been applied) was as good as HARPY.

1.1.2 The Use of Speech Knowledge

Despite some successes, none of these systems represent the realization of the ultimate goal of speaker-independent unlimited vocabulary continuous speech recognition. Current technology in speech recognition possesses many limitations. For

example, most systems can only recognize isolated words; the few that recognize continuous speech can only do so in certain highly constrained circumstances, such as only allowing a small number of possible sentence structures. In addition, most of these systems require training on a single speaker and are only capable of accurately recognizing the speech of that person. Also, all of the systems mentioned above are limited vocabulary recognizers, and because of the ways they have been implemented, increasing the vocabulary size means increasing the amount of memory required, increasing the amount of necessary training, or needing additional time to perform the task. None of these requirements is desirable, so a different approach must be taken to solve the problem.

While helpful for a restricted set of applications, the current technology does not extend directly to the desired goal of continuous speech recognition. Speech is more difficult to deal with when words are spoken continuously because the acoustic properties of a word can vary depending on its context. On the other hand, as in isolated word recognizers, syntactic and semantic constraints aid in recognition. Also, the system ideally ought to be speaker-independent, and therefore needs to exploit interspeaker properties of speech signals, using acoustic features and syntactic constraints in order to recognize utterances. Present and future work on speaker-independent, unlimited vocabulary continuous speech recognizers depends not only on conventional signal processing techniques, but also on being able to apply speech knowledge, such as information about stress [1] or broad phonetic features [21,7] toward solving the problem.

A phonetically-based approach may offer the solution, but the problem is too difficult to tackle without imposing some restrictions. Solving a small portion of the problem will hopefully make the overall goal of speaker-independent unlimited vocabulary continuous speech recognition one step closer to realization.

One way to reduce the size of the problem is to restrict one of the parameters mentioned above, such as vocabulary size, when developing a phonetically-based

recognizer. This makes it easier to extract both low-level and high-level knowledge and to determine what information is relevant to the task.

One vocabulary that has been widely used in this approach is that of the digits zero through nine. Obviously, continuous digit recognition is a popular task because it can be used in a wide variety of applications. Digits form a good vocabulary to use because they are acoustically distinct. However, continuous digit recognition does present some challenges, because coarticulation greatly modifies the phonetic features of speech, and syntactic constraints are non-existent, since any digit may follow another in a given string. Several successful continuous digit recognition systems have already been developed [2,12]. Another interesting vocabulary, one that is somewhat more complicated than digits, is that of the letters of the alphabet. However, continuous letter recognition has not yet been successfully achieved.

1.2 The Spelling Task

1.2.1 Motivation

Continuous letter recognition is a meaningful task both because of its contribution toward solving the continuous speech recognition problem and because of its immediate practical applications. Like continuous digit recognition, recognition of continuously spoken letters is a small enough task to be manageable since only a limited vocabulary is used. However, letter recognition is more difficult than digit recognition. First of all, the number of words in the vocabulary has increased, from ten to twenty-six. Secondly, letters are not as acoustically distinct as numbers. People often have difficulty distinguishing the letters of the alphabet from one other, hence the common practice of giving a clarifying example (e.g.,"D as in DOG") when spelling words. However, there are a few ways in which letter recognition may be easier than digit recognition. For instance, digit strings may be affected more by coarticulation than do letter strings because, in general, speakers may say

digit strings casually. Also, syntactic constraints are non-existent in digit strings: knowledge of the ordering or structure of a string gives no useful information since digits may appear any number of times in any order. On the other hand, unless random letters are being spoken, syntactic constraints that may aid in recognition do exist for letter strings.

The development of a continuous spelling recognizer is a worthwhile task which has several applications. It can be used to distinguish between homonyms (e.g., "bear" and "bare"), or to recognize acronyms (e.g., "MIT" or "IBM" or "VLSI") which occur frequently in technical situations. In addition, a spelling recognizer would be useful in cases in which an utterance is ambiguous: a speaker could be asked to spell a word not recognized by the system. It could also be used to add words to the vocabulary of a speech recognition system. Of course, a continuous letter recognizer could be used to recognize any string, but recognizing spelled English words is a manageable and well-defined task.

1.2.2 Difficulties of Task

The twenty-six letters of the alphabet can be divided into subclasses based on their acoustic-phonetic properties. One such approach is to classify letters based on their contained vowels. This means the letters B, C, D, E, G, P, T, V and Z form a subclass (the /i^y/ set), as do A, J and K (the /e^y/ set), F, L, M, N, S and X (the /ɛ/ set), I and Y (the /a^y/ set), and Q and U (the /u/ set). O, R and W are singletons or unique elements. Another method is to group letters based on general phonetic characteristics. For example, the letters that fit the pattern [FRICATIVE][VOWEL] are C V and Z. These two classification methods can be combined to further subdivide the vocabulary. Ideally, there should be enough acoustic-phonetic cues to place each word in its own subclass, thereby facilitating recognition. However, this goal has not as of yet been reached. The obvious acoustic similarities between some letters, such as B and V, or M and N, make continuous

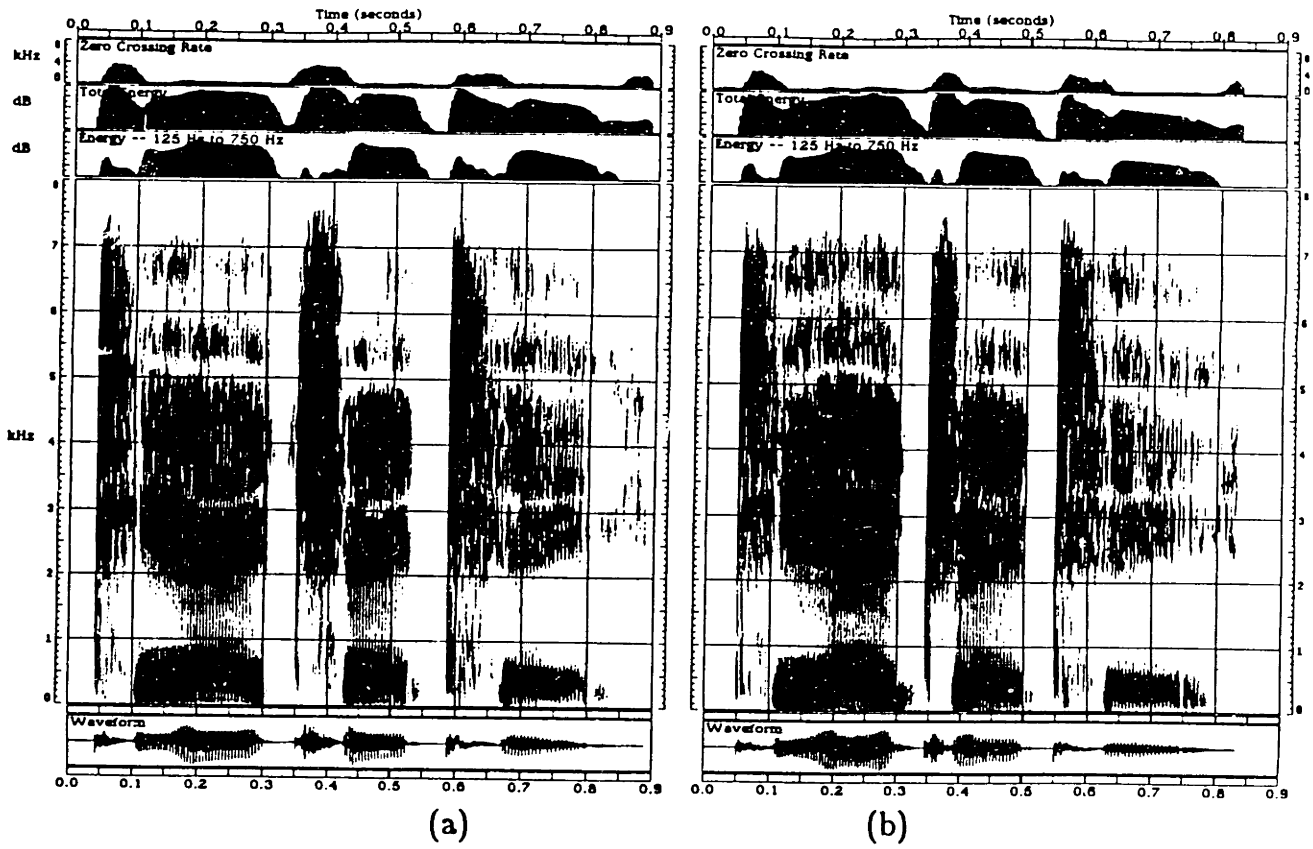


Figure 1.1: Spectrograms of (a) THAT and (b) TAJT

letter recognition a difficult task.

In order to get a better idea of the difficulties involved in continuous letter recognition, it is instructive to examine isolated letters first. A system for recognizing isolated letters and digits which uses acoustic features for discriminating among sounds has been developed by researchers at Carnegie-Mellon University [3,4]. The system, known as FEATURE, has an average accuracy rate of 89.5% when tested on 10 male and 10 female speakers. However, since FEATURE's analysis depends on the fact that the endpoints of letters are known, its extendibility to continuous letter recognition is questionable.

In general, it is difficult to apply isolated word recognition techniques to continuous speech because the signal is difficult to segment into individual words. For example, a system may be hard-pressed to determine whether an unknown utterance

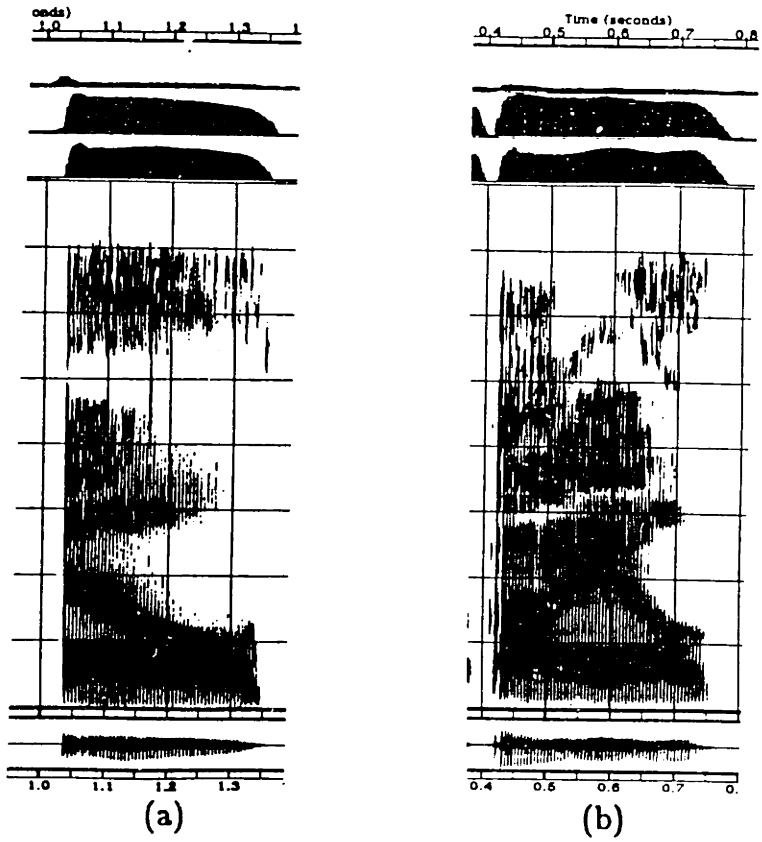


Figure 1.2: Spectrograms of (a) L and (b) IL

is AJ or HA without knowing where the boundary is. Figure 1.1 shows wideband spectrograms of the utterances THAT and TAJT spoken by the same person, and it can be seen that the two spectrograms are virtually identical. Also, coarticulation can be quite severe in spelled strings. Part (a) of Figure 1.2 shows a spectrogram of the letter L spoken in isolation, and part (b) shows a spectrogram of IL extracted from continuous speech. It can be seen that the L in part (b) of the figure is modified by its phonetic environment: the preceding I has raised the beginning of the second formant of the L.

The letters are remarkably similar acoustically (especially the /i^y/ set) and people often have difficulty distinguishing between them. In addition, segmenting the an utterance of spelled speech into individual letters may be difficult. The development of a recognition method must take into account both the characteristics of spelled speech and the difficulties associated with it.

1.2.3 Knowledge Sources

The best way to approach the spelling task is to use information from all relevant sources of knowledge. The two primary sources of knowledge that are available are acoustic and syntactic.

The acoustic knowledge source is rich in information, and listeners are usually able to extract enough from it to recognize continuous speech. However, current speech recognition systems are unable to perform as well as humans. Some recognition cues are too subtle and cannot be detected using currently available signal processing techniques. This means that acoustic information is insufficient for the realization of this task.

Since the problem cannot be solved solely by relying on acoustic features, other methods of analysis must be considered. In the general speech recognition problem, if the permissible combinations of the words are constrained, then syntax may be used to aid in recognition. Similarly, in this task, if the strings of letters to be recognized form words, then the rules of English spelling may be used to help recognize the letters.

In situations where acoustic ambiguities cannot be completely resolved, as in trying to determine if an utterance is either "CHAT" or "ZAJT," knowledge of spelling rules of English would definitely point to the first alternative as being the correct choice. So the solution to the problem of connected letter recognition can be found by combining information from the two knowledge sources.

1.3 Thesis Overview

1.3.1 Problem Statement

In order to recognize words from their spellings, both acoustic-phonetic information and lexical constraints may be used. The purpose of this thesis is to study the

acoustic-phonetic and lexical knowledge sources and to determine what information is useful to spelling recognition and how the knowledge sources might be integrated to accomplish the task.

1.3.2 Summary

A number of steps are taken to realize the goals of this thesis. First, a lexical study is undertaken in an effort to obtain information about syntactic constraints in spelled words and to try to quantify the rules of spelling.

Also, the relationship between acoustic-phonetic and lexical information is examined. We surmise that both knowledge sources are used to recognize spelled words, but the relative importance of each one to the realization of the task is not known. In order to determine the individual usefulness of the knowledge sources, experiments to determine the sufficiency of acoustic information for recognizing spelled speech are performed. Auditory perception tests are conducted to establish a benchmark recognition rate as a goal for a speech recognition system, and spectrogram reading tests are conducted because spectrogram readers use a feature-based approach to speech recognition that we could emulate in order to implement a spelling recognition system.

The results of these experiments are analyzed and errors made by listeners and readers are compared to try to determine why they occur and how they might be resolved. As part of this analysis, the acoustic characteristics of spelled speech are studied to try to determine what makes it different from ordinary continuous speech.

Finally, ways of integrating acoustic-phonetic and lexical knowledge are explored. A model for a spelling recognition system that incorporates information from both sources is proposed and discussed.

Chapter 2

Exploring Lexical Constraints

2.1 Introduction

2.1.1 Description of Task Vocabulary

The letters of the alphabet form a vocabulary with several distinctive properties. The vocabulary contains twenty-six symbols, all but one of which are monosyllabic. The letters are structurally similar to one another: most follow either the pattern [CONSONANT][VOWEL] or [VOWEL][CONSONANT]. The letters are composed of twenty-six different phonemes out of the set of forty ordinarily found in English. All the letters except W contain one vowel out of the set /ɑ, ɑ^{y}, ɛ, e^{y}, i^{y}, o^{w}, ʌ, u/ (W contains two). Consequently, many letters share the same vowel, and this results in a great deal of acoustic similarity between letters. As can be seen by the example of the spectrogram of the letters GPT shown in Figure 2.1, the parts of the letters that are different are often overwhelmed by the parts that are similar. These acoustic similarities make many letters difficult to distinguish from one another. Acoustic similarities between letters can not only cause problems in recognizing individual letters, but can also create additional difficulties when trying to recognize letters in continuously-spelled strings. For example, Figure 2.2 shows

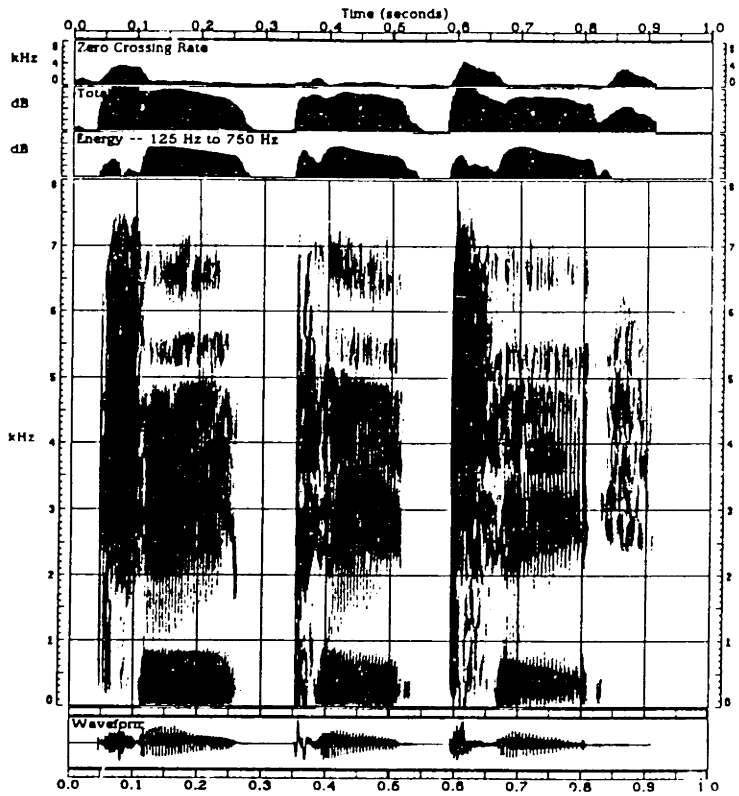


Figure 2.1: Spectrogram of the letters GPT

a spectrogram of the letters O and L, each spoken in isolation. Figure 2.3 shows O and L spoken continuously. In the former case, the letters are separated from each other and are quite distinct. However, in the latter case, it is much harder to decide how many acoustic segments there are and where the boundary between them is. As another example, if spelled quickly, the string BEET could be mistaken for BET.

Even if the signal contains all the acoustic cues necessary for identifying the letters, some of these cues are more subtle than others and are more difficult to extract. Consequently, attempts to recognize continuously-spoken letters solely based on acoustic cues are prone to errors. In order to recognize the letters reliably from the acoustic signal, other sources of information are necessary.

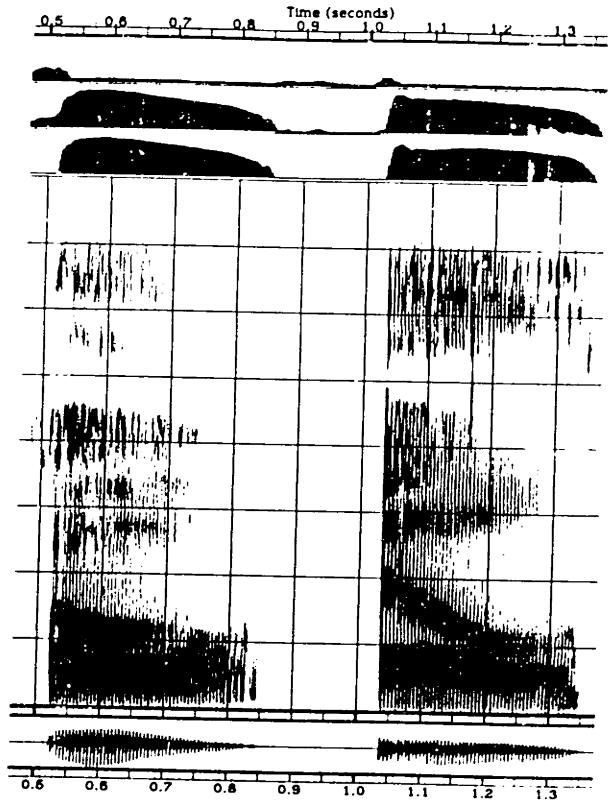


Figure 2.2: Spectrogram of the letters OL /oʷɛl/

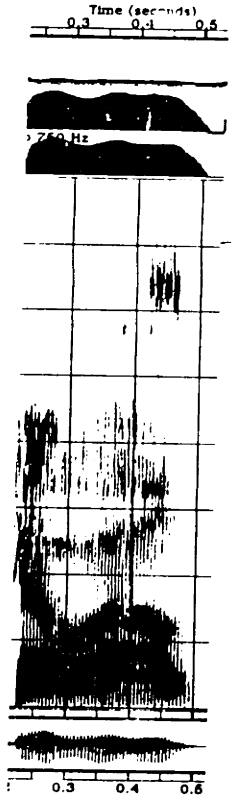


Figure 2.3: Spectrogram of the letters OL /oʷwɛl/

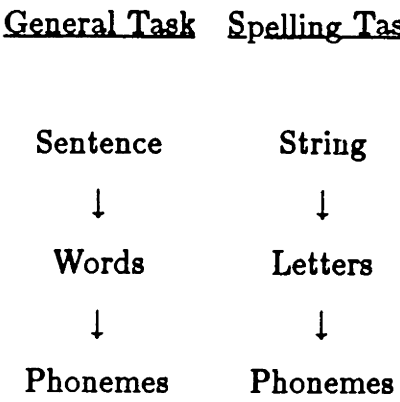


Figure 2.4: Comparison of Spelling to General Speech Recognition Task

2.1.2 Characteristics of Syntax

In the general speech recognition problem, knowledge of syntax often aids in the realization of the task. Syntax rules place constraints on the possible sequence of recognition units. As shown in Figure 2.4, if we know that a string of words to be recognized comprise a sentence, we can use the rules of English grammar to facilitate recognition. Similarly, in continuous letter recognition, if we know that there are syntactic constraints on spelled strings, we can exploit such knowledge to achieve our goal. Specifically, if the task is limited to the recognition of spelled English words, then the rules of spelling can be used to aid in recognition. In order to determine how strong lexical constraints are, and how much lexical knowledge might help in spelling recognition, an effort to determine what they are must be made. Some constraints are easier to define than others: for example, the letter Q is always followed by U. However, other rules are not as obvious; these rules of spelling must all be quantified.

2.2 Data Collection

2.2.1 Lexicon

In endeavoring to determine the rules of spelling, it is instructive to study as many words as possible, in hopes that certain lexical patterns will emerge. If they do, these patterns may be used to induce spelling rules. The largest body of words available to us for a lexical study is the twenty-thousand word Merriam Pocket Dictionary (MPD) with Brown's Corpus counts for word frequency.

A good way to find lexical patterns in a large lexicon such as this is to gather statistics about the frequency of letters and sequences of letters, both dependent on and independent of context. This is necessary in order to provide an indication of what letter sequences are more likely than others in certain situations. Also, frequency statistics such as these can also show what letter sequences are possible, if not for English in general, at least for the lexicon in question. However, one may expect that the larger the lexicon, the closer the statistical characteristics of the lexicon are to general English.

Finding letter frequencies in the lexicon by weighting the words by frequency of occurrence in English can give an idea of what word patterns are common. On the other hand, studying the lexicon in the same way, but weighting each word equally gives a clearer picture of what word patterns are possible. In this study, the MPD is analyzed in both ways.

2.2.2 Gathering Letter Frequency Statistics

The statistics gathered in this lexical study were obtained by using a lexical analysis package called ALexiS [10]. Statistics were gathered about the frequency of the following letter sequences: individual letters, pairs of letters and triplets of letters. These statistics included overall frequency of appearance of letter sequences, and also frequencies of occurrence of sequences at the beginnings and ends of words. In

Event	Most Common	Freq(%)	Top N	Comprise P %
Single Letter	E	12.4	10	75
Word Initial Letter	T	19.0	10	80
Word Final Letter	E	24.0	10	80
Pair of Letters	TH	5.4	10	25
"	"	"	125	85
"	"	"	200	95
Word Initial Pair	TH	14.1	10	41.0
Word Final Pair	HE	10.8	10	40.2
Triplet of Letters	THE	5.6	20	18.6
"	"	"	100	38.0

Table 2.1: Weighted Case

addition, forward and backward dependent probabilities of appearance were also calculated.

An examination of the results reveals some interesting facts. First of all, although the statistics for words weighted by frequency of occurrence differ from those for words weighted equally, they share some of the same characteristics. This can be seen by comparing the statistics in Tables 2.1 and 2.2.

Table 2.1 contains a summary of statistics for letter frequencies using words weighted by appearance. Each row of the table gives information about a certain aspect of the statistics. For example, the first row indicates that E is the most common single letter in the MPD; 12.4% of all letters in the lexicon are E. The first row of Table 2.1 also shows that the ten most frequent letters occur 75% of the time;

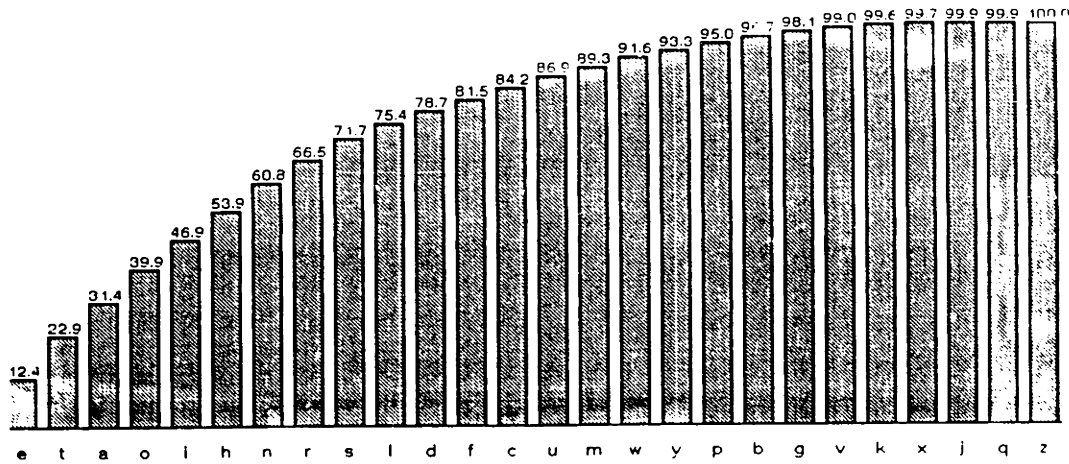


Figure 2.5: Cumulative individual letter frequencies (weighted)

that is to say, any given letter of a word has a 75% chance of being one of these ten letters. (The cumulative individual letter frequencies are shown in Figure 2.5.) The other rows of the table can be interpreted in the same way. This figure also shows the ordering of the letters of the alphabet by frequency of appearance. The constraints on letters in word-initial and word-final positions are even stronger: in both cases, the ten most frequent letters occur 80% of the time.

For pairs and triplets of letters, similar frequencies were found, and some results are shown in the table. It can be seen that the results are greatly influenced by the word THE, which is extremely common.

Table 2.2 lists similar statistics found when each word in the MPD was weighted equally. Although the frequency of appearance of specific letter sequences are different from the weighted case, it is true here, as in the weighted case, that the ten most frequent letters occur 75% of the time. Figure 2.6 shows the cumulative letter frequencies for the unweighted case, and it can be seen that the cumulative distributions are similar in the two cases.

By weighing all words equally when analyzing the MPD, knowledge of what

Event	Most Common	Freq(%)	Top N	Comprise P %
Single Letter	E	10.7	10	75
Word Initial Letter	I	16.7	10	80
Word Final Letter	E	15.1	10	80
Pair of Letters	IN	4.2	10	25
"	"	"	125	85
"	"	"	200	95
Word Initial Pair	CO	3.9	10	23.0
Word Final Pair	ON	6.3	10	35.8
Triplet of Letters	ION	1.0	20	10.6
"	"	"	100	25.2

Table 2.2: Unweighted Case

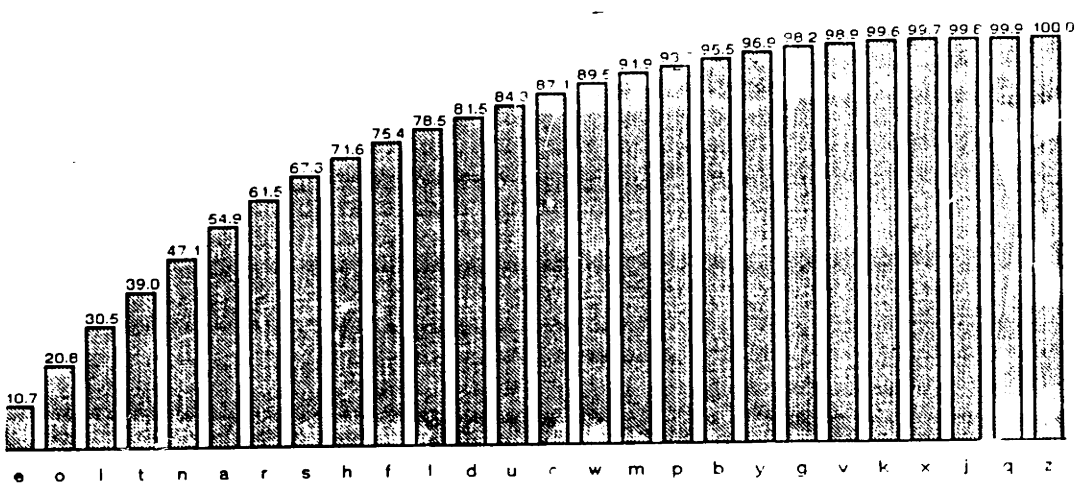


Figure 2.6: Cumulative individual letter frequencies (unweighted)

letter sequences occur can be obtained. It was found that all one-letter sequences, A to Z, can be found in the lexicon. Also, it was discovered that 82.2% of all possible two-letter sequences and only 28.3% of all possible three-letter sequences can be found in lexicon. Of the two-letter sequences, the most frequent one-third of all existing letter pairs comprise 95% of all letter pair occurrences. This means that the majority of possible letter pairs rarely occur, and that most words are composed of a combination of letter pairs drawn from a total of approximately two hundred.

The conclusion that can be drawn from these results is that the more letters known in a word, the greater the constraints that are placed on what the other letters in the word could be.

Another statistic obtained from the MPD measures the frequency of appearance of phonemes. The most common phoneme is /i^{y}/, which is not surprising. This is because /i^{y}/ is found in nine letters, including E and T. Both are among most common letters and together comprise approximately 23% of all letter occurrences (Figure 2.5). As expected, the four most common phonemes are all vowels, since every letter must contain a vowel and the set of vowels found in this vocabulary is somewhat limited. The frequencies for this statistic are shown in Figure 2.7 for the case when the words are weighted by frequency of appearance. These frequencies map directly to the letters in the MPD because each letter was substituted for its phonemic transcription in order to obtain this statistic.

The final statistic of importance deals with the lengths of words in the MPD. It was found that all the words in the lexicon are between one and sixteen letters long, and that the average number of letters per word is 7.35 when the words are weighted equally and 3.98 when the words are weighted by frequency of appearance. The graphs in Figure 2.8 show that the distributions for word lengths, particularly for the case in which words are weighted equally, look Gaussian in nature. The standard deviations for word lengths, weighted and unweighted, are 2.40 and 2.12, respectively.

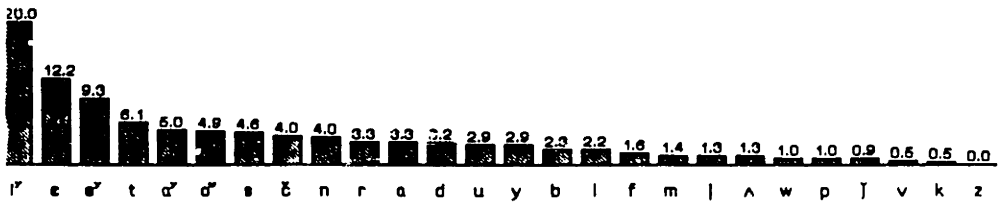
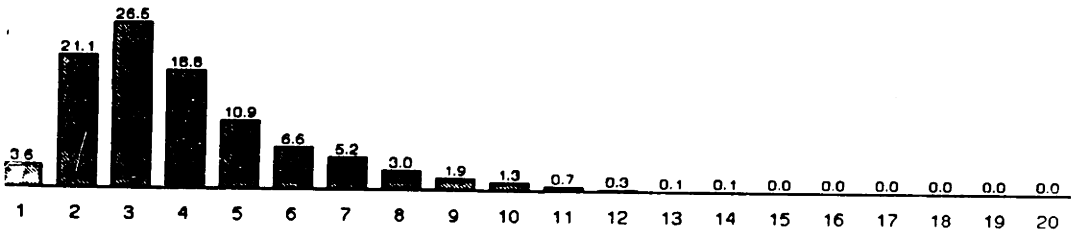


Figure 2.7: Phoneme frequencies (weighted)

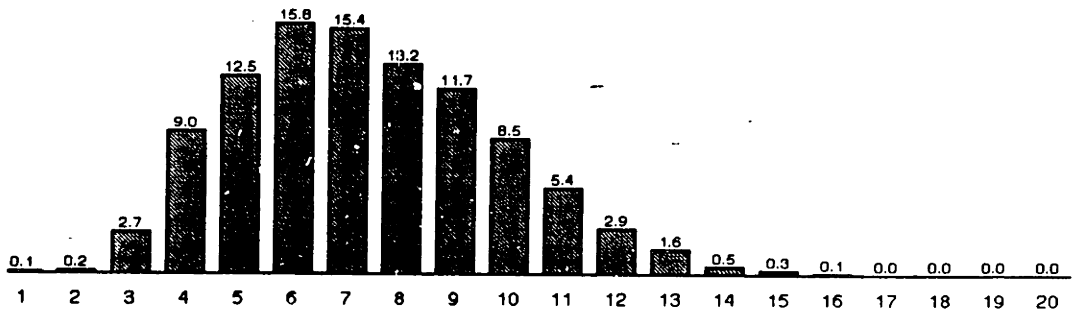
The statistics obtained for the MPD described above are valid for the lexicon, and one may argue that these statistics can be considered to describe the entire English language. However, for lexicons of smaller sizes, the statistics may not reliably reflect properties of the language.

In order to establish the robustness of the statistics, twenty lexicons of two-thousand randomly-selected words each were taken from the MPD and the means and variances of single letter and letter pair frequency statistics were obtained, weighting each of the words equally. Means of single letter frequencies for the MPD and these smaller lexicons are shown in Figure 2.9 and Figure 2.10. The ordering and actual probabilities of occurrence for the two lexicons are very similar. A closer look at the statistics show that, while the frequency means for these smaller lexicons are close to the original ones, the standard deviations are very large. This is due to the fact that the size of the sublexicons is too small.

Graphs for the letter frequency statistics obtained for the MPD (words weighted and unweighted by frequency of appearance) and the smaller lexicons (words unweighted by frequency of appearance) can be found in Appendix A, along with letter triplet frequency statistics obtained for the MPD (words weighted and unweighted).



(a)



(b)

Figure 2.8: Lengths of words in the MPD, weighted (a) and unweighted (b)

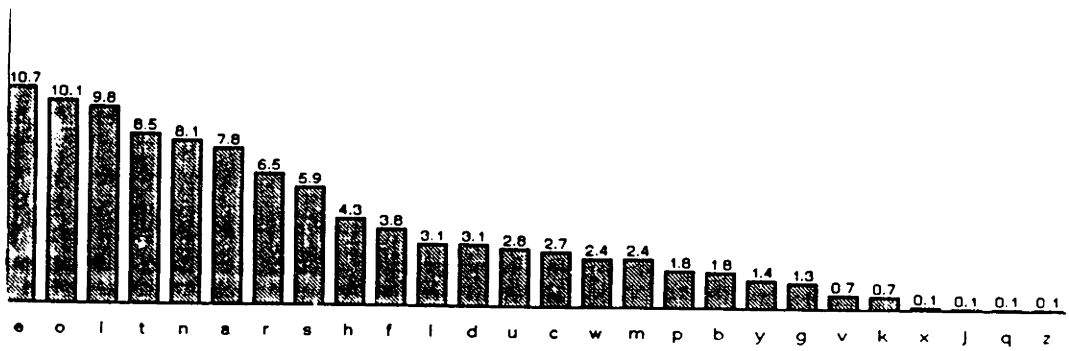


Figure 2.9: Individual letter frequencies for MPD (unweighted)

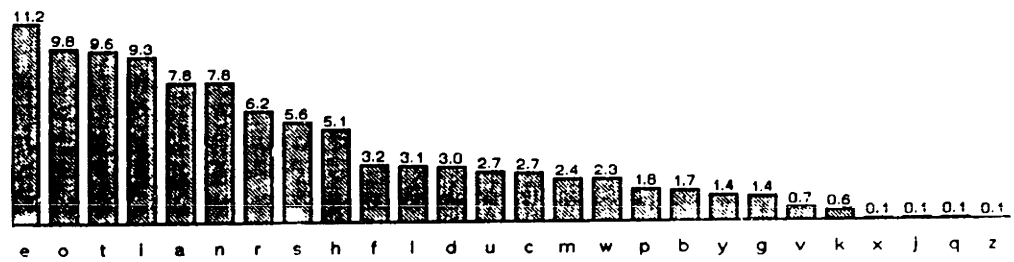


Figure 2.10: Individual letter frequencies for smaller lexicons (unweighted)

2.3 Discussion of Lexical Constraints

2.3.1 Analysis of Results

It was found in the last section that there are a large number of letter sequences that rarely or never occur, which places strong syntactic constraints on what letters may make up a particular word. This was particularly striking for three-letter sequences: less than 30% of all possible letter triplets can actually be found in words. Also, the letter sequence frequency statistics were found to be robust for the smaller lexicons. These findings allow us to hypothesize that over a small set of words, the statistics gathered will be reasonably sound, unless the word set is pathological or skewed in some way. Of course, very small lexicons cannot be expected to behave this way: the larger the lexicon, the more closely its frequency statistics will match those of the MPD. Also, the statistics can be considered valid for the English language in general. Increasing the size of a lexicon means that its frequency statistics become closer to their true values, but as the size of the lexicon increases, the marginal change in frequency statistics decreases to the point where a further increase in lexicon size produces no noticeable change in its statistical makeup. The MPD, by its robust statistics, can be considered to capture letter combinations of the English language as a whole.

The apparent strong constrains on possible sequences of letters point to redundancy in spelled strings. For example, in the case of the letter sequence QUA, the U following the Q is redundant: that is, it conveys no additional information. Measuring this redundancy is helpful in determining the predictability of letters in English words.

2.3.2 Redundancy of Letters in Words

Claude Shannon [20] attempted to measure the information content of letters in words by determining the redundancy of spelling. Redundancy measures the amount

of constraint imposed on a text in the language due to its statistical structure. He attempted to measure the entropy (H), or average number of bits per letter necessary to represent a word. Shannon studied N-gram entropies first, in which N was the number of adjacent past letters known, in order to see how much increasing amounts of knowledge about past letters fostered redundancy. To calculate N-gram entropy, Shannon used frequency of letter sequences tables used by cryptographers [16] and the following formula:

$$F_N = -\sum_{i,j} p(b_i, j) \log_2 p_{b_i}(j)$$

in which F_N is the N-gram entropy

b_i is a block of $N-1$ letters [$(N-1)$-gram]

$p(b_i, j)$ is the probability of the N-gram b_i, j

$p_{b_i}(j)$ is the conditional probability of letter j after the block b_i, and is given by $p(b_i, j)/p(b_i)$.

The entropy of letters can be obtained in the following way:

$$H = \lim_{N \to \infty} F_N$$

Table 2.3 below compares Shannon's N-gram entropies for N = 1, 2, 3 and over an entire word to those obtained for the MPD.

Shannon calculated the N-gram entropies using 26 symbol and 27 symbol (the letters of the alphabet, plus the blank symbol) character sets. He also discounted

	N-gram Entropy		
N	Shannon (26)	Shannon (27)	MPD
1	4.14	4.03	4.13
2	3.56	3.32	3.08
3	3.30	3.1	2.52
Word	2.62	2.14	2.12

Table 2.3: Comparison of N-gram Entropies

boundaries between words in text, so that many two- and three-letter sequences not found in the MPD are included in his measure. Consequently, the predictability of Shannon's letter sequences is lessened. Also, Shannon's method for obtaining F_3 is somewhat questionable: since the only three-letter sequence statistics he had available to him were for letter triplets within words, he approximates probabilities for three-letter sequences across word boundaries using a "rough formula" that gives an F_3 he admits is "less reliable" than the other entropies he calculates.

Shannon's results, as well as the results obtained for the MPD, show that past information is helpful in predicting future events: the more letters known, the greater the redundancy of information, as demonstrated by the lowering entropy rate for higher N. According to Shannon, the entropy of spelled words, F_{word}, is 2.62 bits per letter. His entropy rate is higher than it is for the MPD because Shannon used word frequency statistics for the entire language, whereas in this study, statistics were obtained using for only twenty-thousand words.

2.4 Possible Uses of This Knowledge Source

2.4.1 Exploiting the Predictability of English

The lexical study conducted using the MPD indicates that the constraints on letter sequences within words are very strong. These constraints can be used in a variety of ways, among which could be using them as rules for the synthesis of words.

A string generator that uses letter frequency statistics to compose a string would be more likely to generate real words than a random string generator, and the chance of generating actual words increases as the order of the statistics used increases. For example, a string generator is more likely to synthesize a word if it uses information about the frequency of letter pairs rather than single letters. In addition, knowing proper lengths of words is also helpful in generating words.

A string generator using information about letter pairs and triplets was developed to aid in another aspect of this thesis. It is described in Chapter 3.

2.4.2 Conclusion

The conclusion that can be drawn from this study is that lexical knowledge aids spelling recognition because it greatly constrains letter syntax. While the primary source of information is still acoustic-phonetic, syntactic constraints are important because we are not always able to extract adequate acoustic-phonetic information from the signal to recognize continuously-spoken letters.

Lexical knowledge is important, but it is difficult to quantify its importance in the spelling task: how much lexical information is necessary for a listener to recognize a spelled word? Also, how much of the lexical information available to a listener does he use to recognize the letters?

In order to determine how important lexical information is to spelling recognition, it is necessary to determine the sufficiency of acoustic information. This can be done by conducting continuous-letter recognition experiments in which the

only available knowledge source is that of acoustic-phonetic constraints. The performance of subjects in these experiments will help to determine the importance of lexical information to this task.

Chapter 3

Establishing Confusability

3.1 Introduction

Because of acoustic similarities between various letters of the alphabet, confusions are bound to occur. However, what confusions actually occur is not known, nor is the severity of these confusions.

In order to find out more about the acoustic confusability of letters in spelled strings, a set of recognition experiments was conducted. In these tests, subjects were asked to recognize letters using only acoustic-phonetic information. This was done to determine the sufficiency of acoustic information and to measure acoustic confusability. Both words and non-word strings were used in these experiments for two reasons. First of all, although we ideally would like to conduct experiments using only words since the task is spelling recognition, lexical knowledge might be used to guess some letters. Secondly, using both types of strings allows for comparisons of results.

Auditory perception tests were conducted to find out what letters were confusable to listeners, and spectrogram reading tests were conducted because the techniques employed by spectrogram readers incorporate explicit speech knowledge, and acoustic similarities between letters are easier to quantify in this acoustic feature-

based approach than in listening tests.

3.2 Preliminary Experiments

3.2.1 Isolated Letter Reading Experiment

To obtain an initial impression of what letters are often confused with each other, a pilot experiment was conducted. Four speakers spoke the letters of the alphabet in isolation and in random order, and ten trained spectrogram readers were asked to read spectrograms of the utterances and to identify the letters. Besides the spectrogram itself, the only information given about an utterance was the identity of the speaker.

It was found that in 1040 trials, the readers correctly identified the letter being spoken 923 times on the first choice and an additional 30 times on the second choice, giving first and top two choice accuracy rates of 88.8% and 91.6%, respectively. An extensive analysis of errors was then done, and a confusion matrix was formulated (Table 3.1). The confusion matrix is a plot of actual utterances versus confusions.

In analyzing the results, several interesting patterns emerge. The majority of confusions fall within letter groups that contain the same vowel (87 out of 117, or 74%), so in most cases, vowel recognition is not the problem. Most of the confusions resulted from mistaking members of the /iʸ/ set for one another: Out of 117 confusions, 81 fall in this category. Some of the confusions appear to be among consonants having the same place of articulation. For instance, B-V and V-B confusions occur presumably because they are both labial, and thus have similar formant transitions into /iʸ/. Also, there may not have been much frication noise in the /v/, causing it to be mislabeled as a /b/.

Unusually large amounts of frication noise were observed in many consonants, often causing unvoiced stops such as /t/ to be mistaken for affricates such as /ʧ/. Also, voiced and unvoiced stops were confused because in most instances, voice

Mistaken For:

Correct Letters	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A		1				1																1				
B	2				4									1								1				
C																				1						
D	2															1				6						
E				4												2				1						
F												1							1							
G										4	3					5				3						2
H																								4		
I																										
J											4					1										
K						1				3										3						
L																					3			1		
M														1		1										
N															1											
O														1												
P							3													2		6				
Q							1																			
R																										
S																										
T			1				10			1	1					4										1
U																										
V		6	1											_		1										2
W																										
X																										
Y																										
Z			7																							

Table 3.1: Confusion matrix for isolated letters

onset times (VOTs) for voiced stops were longer than usual, and there was a greater amount of turbulence noise than expected in the voiced stops. This may be due to the fact that speakers tried to enunciate the letters as clearly as possible, but instead created distortions due to overarticulation.

3.2.2 Speaker Dependent Nonsense Strings

Experiments on isolated letters are important in order to determine what features could be used to distinguish among them. However, since the task is the recognition of spelled words, experiments on continuously spoken letters should also be conducted. In continuous speech recognition, coarticulation across word boundaries makes the segmentation of utterances into recognizable words much more difficult. In our case, segmentation means the breaking up of spelled strings into their corresponding letters. However, we suspect that coarticulation may not be a severe problem here because of the nature of the task. Letters are not spoken as continuously as other sounds; speakers subconsciously tend to insert pauses or glottal stops between letters to clarify the utterance (Figure 3.1). Also, letter pairs thought to be confusable, such as UI and UY may have enough acoustic differences that they can be distinguished from each other (Figure 3.2).

In order to study the effects of coarticulation, the following steps were taken: first, a list of all pairs of letters occurring in English words was made. Then, strings of random length were generated by selecting pairs at random from the list in such a manner that each pair of consecutive letters in a list actually occurs in English. This procedure ensures that we do not examine coarticulation for situations that will not occur. Thus, the random string OXQUI would be acceptable, while the string OXQJI would not. Random strings were used to ensure that readers would not guess letters based on lexical information. Next, fifty such strings were given to a speaker, who was asked to spell each as if it were an actual word. Next, spectrograms were made of the utterances, and several expert spectrogram readers

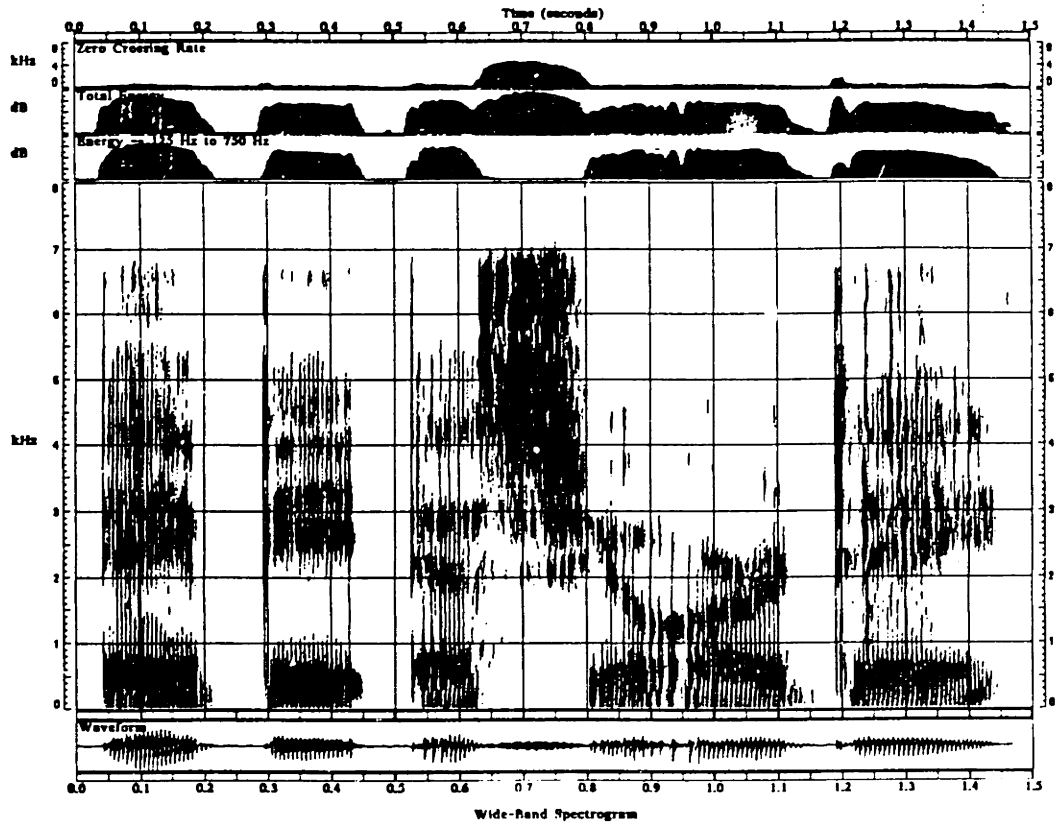
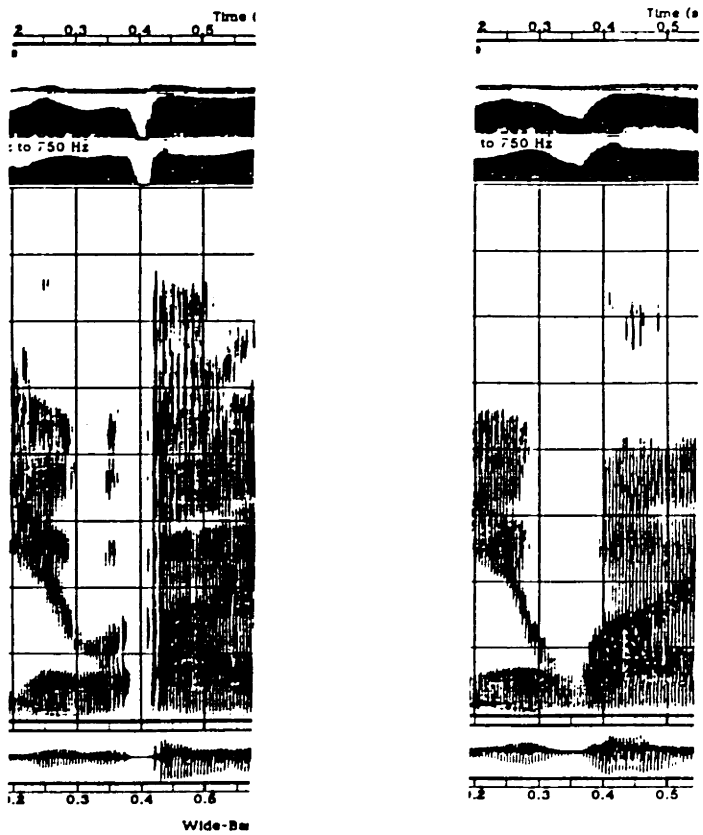


Figure 3.1: Spectrogram of ABSURD which shows pauses and glottal stops being inserted at letter boundaries



(a) (b)

Figure 3.2: Spectrograms of (a) UI and (b) UY

Type of Error	% of Total
Substitution	71.3
Insertion	14.7
Deletion	14.0

Table 3.2: Description of errors made in a continuous letter recognition experiment

were asked to read them. Once the readers had completed their task, their answers were analyzed to determine the effects of coarticulation on the spoken letters. The results are shown in Table 3.2.

3.2.3 Evaluation of Results

Results of these preliminary experiments indicate that acoustic confusability is clearly a problem in spelling recognition. Similar confusions were made in both experiments, but the overall results from the first test were slightly better than the second: readers scored 91.6% on isolated letters versus 92.3% on continuous letters.

There are a number of reasons why readers may have done better in the second experiment. First of all, some cues may be clearer in continuous letters than in isolated letters. For example, B-V confusions are less likely to be made in continuous letter recognition because the closure portion of the /b/ of B, not found in V, is discernible, whereas in isolated letters, since /b/ appears at the beginning of the utterance, the stop gap is not observable.

Also, letters embedded in a string are not prone to endpoint errors. Finally, the readers were more familiar with the task in the second experiment: the first experiment could be regarded as "training" of the readers in letter recognition.

However, statistics on these confusions cannot be obtained reliably from such a small set of data. In order to do an extensive study of confusions, a much larger amount of data collected from multiple speakers must be used.

3.3 Data Collection

3.3.1 Corpus Development

In order for strings to be considered devoid of lexical information and thus eligible to be included in the corpus, they must meet certain requirements. The strings must be "wordlike," that is, they must have some of the same characteristics as words, while not necessarily being words. For instance, within strings, each pair of letters should be one that actually exists in English words. The effect of coarticulation on two adjacent segments that are an impossible combination in an English word (e.g., QX) are not relevant to the task.

As mentioned before, we have argued that the corpus should not be entirely composed of real words because lexical information can potentially distort the results of an acoustic confusability experiment. On the other hand, the corpus should not be made up entirely of non-words for the same reason: because knowledge that a string *cannot* be a word is in itself a lexical constraint. The solution is to create a corpus containing words and non-words, and withhold information on the distribution of words and non-words from the subjects of an experiment.

The corpus is made up of a total of 1000 strings, 350 of which are words and 650 of which are non-words. All strings are between three and eight letters in length, because approximately 70% of all words are of those lengths, as shown in Figure 2.8(b). No nine- and ten-letter strings are included in the corpus, even though words with these many letters are quite common, as can be seen in the figure. This is because very long strings are harder to spell naturally. In addition, lexical information is more likely to be used to identify longer words.

Of the 350 words in the corpus, 310 were selected at random from MPD without regard to their frequency of appearance in English. In order to include enough J, Q, X and Z tokens, 10 each of strings containing these letters were added in. There are no duplicate words.

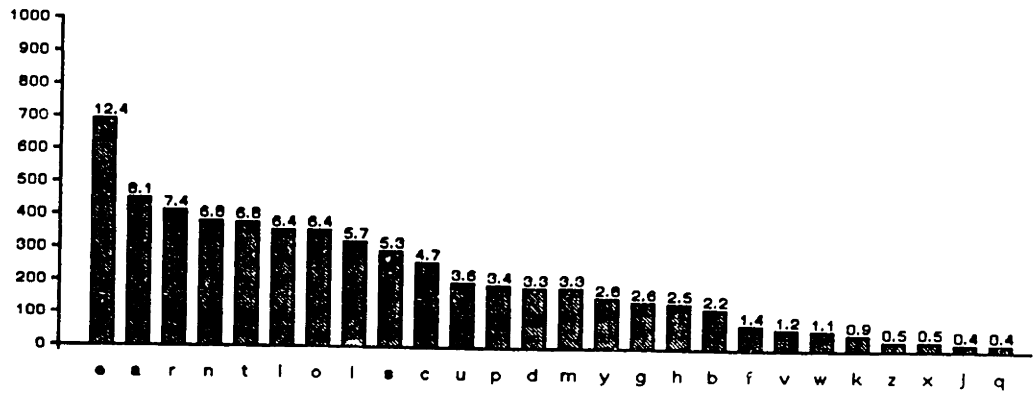


Figure 3.3: Histogram of letter occurrences for spelling corpus

650 strings were generated using the statistics obtained in the lexical study and a set of rules. The rules are as follows: strings must begin with a pair of letters that could begin a real word, and must end with a pair of letters that could end a real word. Within the word, three-letter sequences are ones that could be found in a real word. This means that strings like CAPPOST could be generated, while ones like GTAQIZ could not. Beginning and ending pairs, as well as intraword triplets, were selected at random from a list of pairs and triplets that are found in words, weighted by frequency of appearance. Of the total number of letter pairs that can be found in English, 68.7% are covered in this database. There are a total of 5607 letters in the spelling corpus.

Statistics of single letter occurrences can be found in Figure 3.3. When comparing them to Figure 2.9, it can be seen that the distributions of letters in this corpus are similar to those in the large lexicon analyzed in the lexical study. Eight of the ten most frequently-occurring letters (E, A, R, N, T, I, O and S) are common to both the MPD and the spelling corpus.

Use of these rules yield very wordlike strings: in fact, out of the 650 generated for this corpus, 56 (8.6%) were real words. Many of the non-words differed by only

one letter from a word (e.g., LYLLABLE), and most were at least "pronounceable." Also, because statistics were used to create the strings, some letter sequences were included in several strings: for example, CON was generated five times.

3.3.2 Recording

After the corpus was created, it was recorded by twenty speakers, ten male and ten female, of standard American English. Recording was done using a Sony chest microphone in a sound-treated room. Each subject spelled 50 strings, of which, on the average, 35% were words and 65% were non-words. Each string in the corpus was spelled once by only one speaker. All the letter strings were subsequently digitized and stored on a computer using the SPIRE [10] facility.

3.4 Auditory Perception Experiment

3.4.1 Purpose and Procedure

An auditory perception experiment was conducted to establish a baseline recognition performance against which spectrogram reading and recognition system performance can be measured. The corpus was divided into ten groups of one hundred words each, five words from each speaker. h of these groups constituted a listening test. Eight subjects listened to one or two tests each, for a total of fourteen tests. The tests were administered using headphones in a sound-treated room. The utterances were randomized within each test, and each utterance was said twice in succession. Subjects were told that they were listening to spelled strings, and were allowed to provide one answer per string.

	Word Length							
Error Type	3	4	5	6	7	8	Total	% of Total
Substitution	7.5	11.5	6.0	14.0	9.5	12.0	60.5	68.4
Insertion	0	0	0.5	1.0	1.5	7.5	10.5	11.9
Deletion	0	0	0	1.0	2.0	5.5	8.5	9.6
Exchange	0	0	0	0	2.5	2.5	5.0	5.6
Boundary	0.5	2.0	0.5	1.0	0	0	4.0	4.5
Total	8.0	13.5	7.0	17.0	15.5	27.5	88.5	100

Table 3.3: Distribution of listening test errors

3.4.2 Results

The overall listener accuracy rate in recognizing letters was 98.4% with a standard deviation of 0.72% (a detailed breakdown of errors made in this test can be found in Table 3.3). Also, the subjects performed with an accuracy rate of 98.4% and a standard deviation of 0.87% across speakers (Figure 3.4). Listeners made proportionately the same number of errors on words as on non-words: 41% of the strings in the corpus were words, and listeners made 42% of their errors on word strings. Errors made by listeners included substitution, insertion, deletion and gemination or boundary errors. Each error made was weighted according to how many people listened to the string in question. In one type of substitution error, a letter is incorrectly transcribed (e.g., J transcribed as G.) In another type of substitution error, a phoneme is incorrectly transcribed, resulting in two incorrect letters. For example, if part of an utterance is labeled /eɹči/, when instead it should be /eɹǰi/, the reader will transcribe those segments as HE rather than AG. In a deletion error, a letter is omitted from the transcription, and in an insertion error, a letter is added. In a gemination or boundary error, a phoneme is incorrectly divided, usually at a letter boundary. For example, SE could be transcribed as SC if the

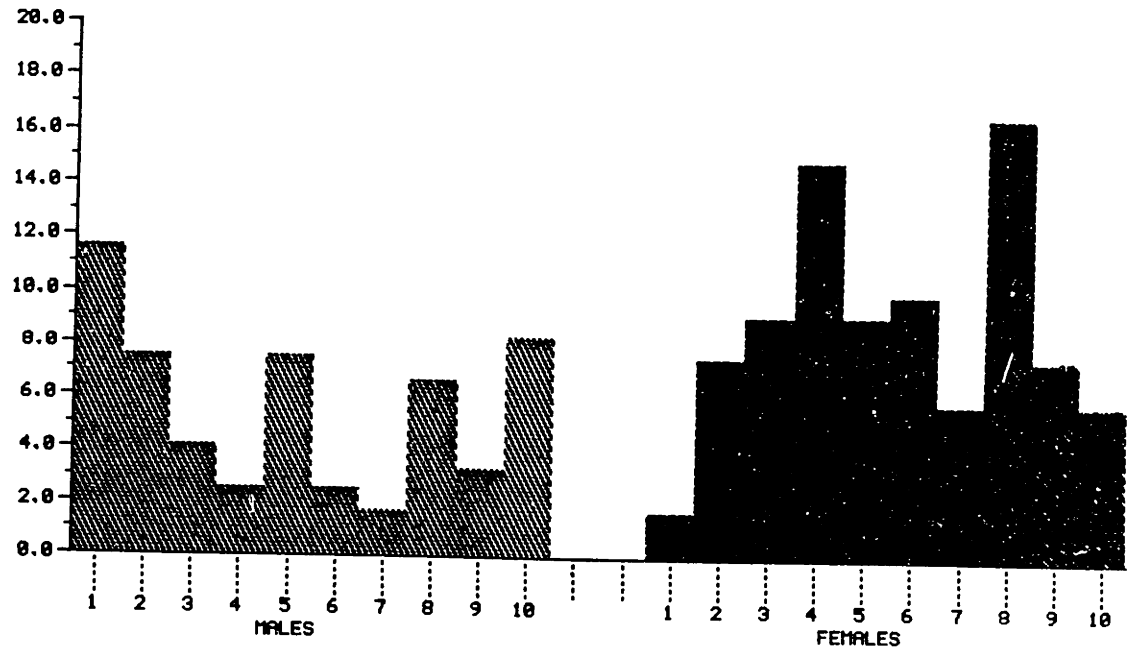


Figure 3.4: Listening test errors grouped by speaker

subject mistakenly assumes that /s/ is shared by two letters.

The most common errors made by listeners were substitution errors. Of all errors made, 68.4% were of this type. The worst confusions made by listeners were B-D, S-F, M-N, O-L and P-T (Table 3.4).

Other significant errors made include insertion or deletion errors, which account for 21.5% of all errors. These errors tended to occur in sonorant regions, and were usually due to the insertion or deletion of a vowel (e.g., BOL mistaken for BL).

Listeners also made a few exchange errors (5.6%) and gemination/boundary errors (4.5%). In exchange errors, letters are correctly identified, but are in the wrong order (e.g., TAC mistaken for CAT). This happened only on seven- or eight-letter non-word strings, and could be attributed to listeners' lack of attention or poor short-term memory. Boundary errors occurred primarily on short, quickly spelled strings, which makes letter segmentation somewhat more difficult than usual.

Listeners made very few string length errors (1.9%). Of these errors, 68.4% were made on eight-letter strings. The fact that so many of these errors were made on

Mistaken For:

Correct Letters	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A									2			0.5														
B				4	1																	2.5				
C																				0.5		2				0.5
D																										
E	1.5	1							0.5																0.5	
F																			1							
G																				1						
H																										
I																		0.5							0.5	
J																										
K																										
L															1.5											
M														3.5												
N													1.5													
O												3														
P		2		0.5	1		0.5													3						
Q							1																			
R																										
S						4																				
T				2												4										
U															0.5											
V		0.5														0.5										
W																										
X																										
Y																										
Z			1.5																							

Table 3.4: Confusion matrix for substitution errors made by listeners

long strings may again be partly due to listeners' poor recall.

3.5 Spectrogram Reading Experiment

3.5.1 Purpose and Procedure

The purpose of this experiment was to determine the sufficiency of acoustic information in recognizing letters from spectrograms. A spectrogram reading experiment is useful because, in contrast to the listening test, subjects are explicitly using acoustic-phonetic knowledge. Because of this, we can get an idea of the recognition performance that we can expect based on our current acoustic-phonetic knowledge. However, these results may provide only an upper bound, since spectrogram reading results are typically better than currently-available acoustic-phonetic front-ends.

Six trained spectrogram readers attempted to read spectrograms of some of the one-thousand utterances in order to simulate computer recognition of speech. Each of the six readers was given one hundred utterances, five from each of the twenty speakers. Approximately one-third of the one hundred spectrograms given to each reader were spectrograms of real words, and the rest were of non-words.

Readers were told that some of the spectrograms were words, but were not told the exact proportion. Other information provided included the identity of the speaker, the fact that each utterance contained between three and eight letters, and that all the strings were "wordlike," as described in the previous section. They were asked to transcribe each utterance using letters of the alphabet rather than phonetic symbols. In cases of uncertainty, readers were encouraged to write down second or third choices for segment transcriptions.

In general, spectrogram readers transcribe an utterance phonetically, and then propose an orthography for the sentence based on this transcription. There were two reasons for asking readers to transcribe the utterances with letters. First of all, it would make the conditions of the spectrogram reading test similar to those

Reader	% Correct	% Correct (Top 3 Choices)
1	94.8	96.1
2	93.3	97.2
3	91.8	94.1
4	90.7	93.2
5	88.4	94.7
6	86.8	92.6

Table 3.5: Individual recognition rates of spectrogram readers

of the auditory perception test, thereby enabling a direct comparison of results. Secondly, a reader's proposal is based not only on acoustic evidence but also on lexical access and syntactic constraints. However, in this experiment, syntactic constraints were minimal, so a reader's guess would be primarily based on acoustic information, and in cases of uncertainty, on the best available acoustic features for correctly identifying a segment. For example, if a reader phonetically transcribes a segment as /tc^{j}/, he then must decide if the segment should really be /kc^{j}/, for K, or /ti^{j}/, for T. The letter the reader chooses indicates which features he considers most important.

3.5.2 Results

As a group, spectrogram readers were asked to identify a total of 5601 tokens in 600 spectrograms. They did so with an overall accuracy rate of 91%. Individual accuracy rates ranged approximately between 86% and 95% (Table 3.5). Although accuracy rates improve somewhat when second and third choice transcriptions are included, rising from 91.0 $\pm$ 2.6% to 94.6 $\pm$ 1.6%, the higher rate is not as informative as the original one, because some readers are more conservative in guessing than others. Interspeaker variability in error rate was more striking than in the

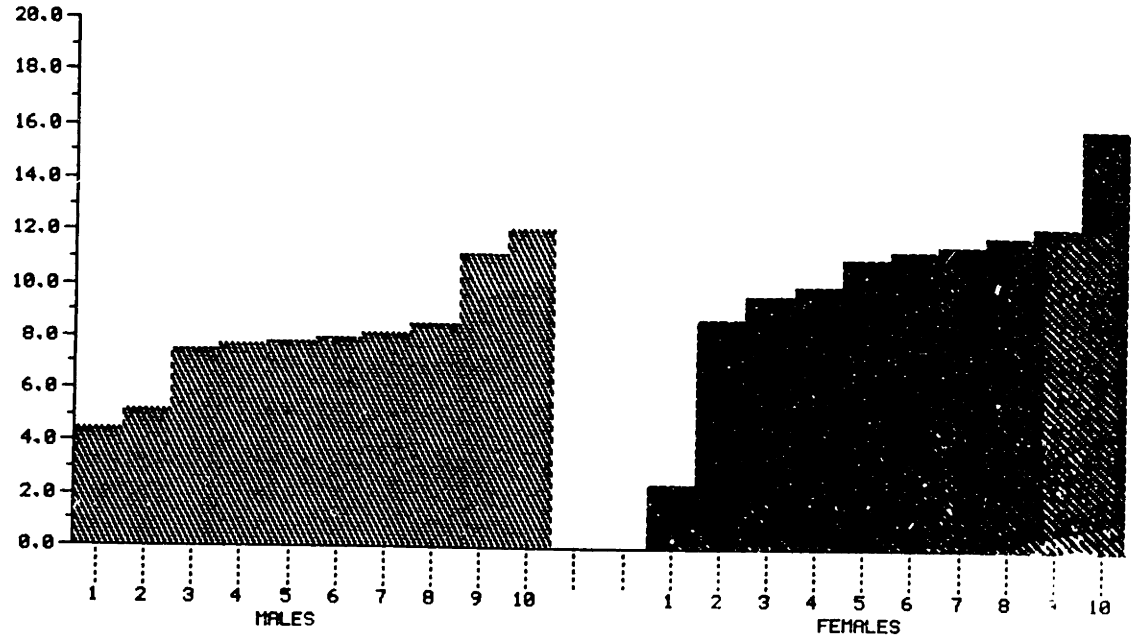


Figure 3.5: Spectrogram reading test errors grouped by speaker

listening tests: across the twenty speakers, error rates were between 2.4% and 16% (Figure 3.5). As can be seen from the figure, readers had more difficulty recognizing female speech than male speech.

Readers were more likely to make mistakes on non-words than on words. Although 41% of the utterances read by the readers were words, they only made 27% of their errors on this group. When questioned, all of the readers emphatically stated that they did not use lexical access to aid their transcriptions when uncertainties arose. However, knowing that strings were "wordlike" may have had some influence on their final transcription.

The types of mistakes made by the readers were substitution, deletion, insertion, and gemination errors. A detailed breakdown of results from this test can be found in Table 3.6. Exchange errors were not made by spectrogram readers, presumably because they need not rely on memory.

As might be expected, substitution errors accounted for the majority of the errors. 92% were substitution errors, and of these, over 80% were single-letter

Error Type	Number	% of Total
Substitution:		
By Letter	274	84.3
Across Letters	25	7.7
Insertion	15	4.6
Deletion	8	2.5
Boundary	3	0.9
Total	325	100.0

Table 3.6: Distribution of spectrogram reading test errors

substitution errors. As can be seen in Table 3.6, insertion and deletion errors were infrequent: about 7% of the errors are of either type. In fact, out of 600 strings, only 23, or 3.8%, were transcribed with the wrong number of letters. A confusion matrix for single-letter substitution errors was constructed (Table 3.7). Substitution errors made by both listeners and readers are plotted here to aid direct comparison. The confusion matrices contain a great deal of information about the types of errors made by readers and listeners. As can be seen from the plot, some errors are symmetric, that is, roughly the same number of Letter 1 to Letter 2 confusions were made as Letter 2 to Letter 1 confusions, while others were not. Some of the errors are unimportant; for example, U in the string-final position was once transcribed as F, a mistake not likely to be made often. A summary of the most common errors can be found in Table 3.8. The table is arranged so that Letter 1 to Letter 2 errors are paired with Letter 2 to Letter 1 errors so that the presence or absence of symmetry can be seen.

This summary shows that the most common errors made by spectrogram readers are symmetric, and that most of the errors can be attributed to confusions between only a few letter pairs. In fact, the four most frequent confusions, G-T, A-E, M-N

Mistaken For:

Correct Letters

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A					20																1					
B				2	3											1						4				
C							1															2				2
D							1									2				4		3				
E	16	2		1												1										
F																			7							
G			2							7							1			10		1				5
H																			1							
I	5														2			2							4	
J							1																			
K										4										1						
L									3				1		15											
M														17												
N					1								12													
O									5			11									3					
P			1				6			2							1			7						
Q										1																
R									6						2										1	
S						3																				
T			4				27			2	3					1										1
U						1									3											
V			1	4	1											3										4
W																										
X																										
Y																										
Z			2																							

Table 3.7: Confusion matrix for substitution errors made by readers

Pair	# of Errors	Pair	# of Errors	Total
T-G	27	G-T	10	37
A-E	19	E-A	16	35
M-N	16	N-M	11	27
L-O	15	O-L	11	26
F-S	7	S-F	3	10
G-J	7	J-G	1	8
P-T	7	T-P	1	8
R-I	6	I-R	2	8
O-I	5	I-O	2	7

(a)

Pair	# of Errors	Pair	# of Errors	Total
T-P	4	P-T	3	7
S-F	4	F-S	1	5
M-N	3.5	N-M	1.5	5
O-L	3	L-O	1.5	4.5
B-D	4	D-B	0	4

(b)

Table 3.8: Most common substitution errors for (a) readers and (b) listeners

and O-L, account for 47% of the single-letter confusions. Also, these four confusions occur significantly more frequently than any other; the fourth most common one, L-O, occurred 26 times, while the fifth most common, F-S, occurred only 10 times. Some of the errors were asymmetric, such as R-I confusions. R was incorrectly transcribed as I 6 times, while I was mistaken for R only twice.

The accuracy rate for this experiment was slightly lower than that in the pilot study in which readers tried to identify letters in random strings spoken by one person (91% versus 92.3%), and this may simply be due to the fact that this experiment used multiple speakers, so there was more variability in speech than in the pilot experiment.

3.6 Conclusions

3.6.1 Comparison of Experiments

The two sets of experiments performed were similar in that they used the same corpus, and each test taken by subjects contained the same number of strings, but there are many more differences between them. The subjects used in the perception test were different from the ones used in the reading test. Also, none of the speakers were subjects for either experiment. Different information about the strings were given in the tests: listeners were told they would be hearing strings of letters, while readers were told they would be seeing "wordlike" strings of letters. Also, readers were told speaker identities, and they knew each string had to be between three and eight letters long. Listeners only heard each string twice, whereas readers were allowed unlimited time, and were also allowed to collaborate with other readers. Listeners were given less information than readers because they have a slight advantage over readers to begin with: the auditory system easily and automatically processes speech.

The purpose of the auditory perception experiment and the spectrogram reading

experiment was to determine the sufficiency of acoustic information. It can be seen from the accuracy results, 98.4% and 91.0% respectively, that acoustic information is the primarily knowledge source for obtaining information to recognize spelled strings. A comparison of results of the tests suggests some interesting similarities and differences between them.

Listeners did significantly better than readers, and had less variation in results, both across subject and across speaker. Both listeners and readers guessed the correct number of letters very accurately (98.1% and 96.2%). Substitution errors predominated for both listeners and readers (68% and 92%). Also, most of the substitution errors made by readers were also made, to a lesser degree, by listeners, as shown in Tables 3.4 and 3.7. However, listeners and readers usually did not make the same specific errors: that is to say, they rarely made mistakes on the same tokens.

3.6.2 Summary of Acoustic Confusabilities

As mentioned above, most errors made in both experiments were substitution errors. Some of the errors were more likely to be made by readers than by listeners. For example, the confusion B-D, one of the worst errors made by listeners, was rarely made by readers. Some of the errors, such as G-T, A-E and M-N were symmetric, and others were asymmetric, such as P-G and I-R.

The results of these experiments raise a number of questions about the nature of spelled strings and errors that are made in trying to recognize them. An acoustic study is necessary to determine what characteristics of spelled strings make them different from ordinary speech, to answer questions about why certain errors occur and to explore ways in which these errors can be resolved.

Chapter 4

Acoustic Study of Spelling Corpus

4.1 Purpose of Acoustic Study

In order to develop a method for recognizing spelled strings, an understanding of their acoustic properties is essential. Therefore, the next step is to undertake an acoustic study in an effort to determine what differences exist between spelled strings and ordinary speech, and whether or not these differences could be exploited to aid in recognition. Alsc, this study offers the opportunity to study the spelling corpus more closely. The results of the auditory perception and spectrogram reading experiments lead to a number of questions about the types of errors made that are best answered by a study of this kind. For example, why did the mistakes made by listeners differ so much from the mistakes made by readers?

Some of the possible errors anticipated before beginning the recognition experiments rarely or never occurred. For example, the problem of insertion and deletion of segments was much less serious than expected. A study of sonorant regions (where this problem was expected to appear), concentrating on vowels in the context of a vowel followed by a vowel would help determine how two adjacent vowels can be distinguished from a single vowel.

In addition, the errors made by the subjects of the experiment were mainly

substitution errors, and an acoustic study presents the means for examining these errors, determining their causes, and exploring ways to resolve them.

This acoustic study was undertaken using SPIRE, a speech processing software package, and SEARCH, another software package which allows users to interactively explore ways to analyze acoustic data [10].

4.2 Phonological Properties of the Corpus

4.2.1 Characteristics of Vocabulary

The acoustic properties of individual letters are not discussed here in detail, because they have been documented in the literature [3,4]. For example, the success of the FEATURE system indicates that a great deal about these acoustic features is known. But continuous speech has the problem of ambiguous letter boundaries, which means the acoustic features cannot be solely relied upon for accurate recognition. However, the letter recognition task is aided by syntactic constraints on letters and the insertion of glottal stops. Unfortunately, continuously spoken letters are subject to gemination errors as well, especially at boundaries between vowels.

4.2.2 Lexical Constraints on Letters

Spelled strings differ from ordinary speech in a number of ways. First of all, they are composed of a limited set of symbols, namely, the twenty-six spoken letters of the alphabet. The letters contain only twenty-six of the forty phonemes found in English, and the possible combinations of phonemes that may occur in spelled strings is limited. For example, if a phoneme is known to be /ɛ/, it must be followed by either /f/, /l/, /m/, /n/, /s/ or /ks/ because it must be part of one of the letters F, L, M, N, S or X.

Even less specific phonetic constraints, such as broad classification by manner

FRIC	VOWEL	VOWEL	AFF	VOWEL	STOP	VOWEL
C		H		A	B	
V				E	D	
Z				I	K	
				O	P	
		A	G		T	
			J			

Figure 4.1: Letter combinations for [FRIC][V][V][AFF][V][S][V]

of articulation [21,7], greatly reduces the possible sequences of letters that could be found in a spelled string.

For example the word CHAT when spelled, can be phonetically transcribed using broad manner classes as

[FRICATIVE][VOWEL][VOWEL][AFFRICATE][VOWEL][STOP][VOWEL]

The only letters that can begin the string are C, V and Z, because they are the only ones that are composed of a fricative followed by a vowel. Similar statements can be made about the other segments in the string, and all the possible combinations of letters are shown in Figure 4.1.

Another distinctive property of spelled strings is that most syllables are stressed. This characteristic is beneficial to recognition because the acoustic-phonetic features of stressed syllables are clearer and easier to extract than those of unstressed or reduced syllables.

4.2.3 Glottal Stop Insertion

One of the most interesting characteristic of the spelling corpus is that it contains a far greater number of glottal stops that would be found in ordinary speech. The average number of glottal stops in the corpus is about 2.3 per string. A closer look at this feature may lead to an understanding of the properties of glottal stops and why they are so prevalent in spelled speech.

In Chapter 1, differences between isolated and continuously spoken letters were discussed. We surmised that for the problem of finding letter endpoints in continuous speech, letter boundary detection would not be easy, because finding word boundaries in ordinary continuous speech is a difficult task.

If this is truly the case, then it is to be expected that attempts to recognize spelled speech would be prone to a large number of insertion or deletion errors. However, in the auditory perception and spectrogram reading experiments described in the previous chapter, both listeners and readers made far more substitution errors than insertion and deletion errors combined. 68% of the listeners' errors were substitution errors, and 21.5% were either insertion or deletion errors. Results for the readers are more striking: 92% of their errors were substitutions, while only 7% were insertions or deletions. In fact, both listeners and readers chose the correct number of letters very accurately (98.1% and 96.2%). This leads to the conclusion that finding letter boundaries in spelled speech is not as difficult as anticipated.

It appears to be the case that when people spell words, they know from experience that many letters are easily confusable. As a result, they tend to enunciate clearly to make the letters easier for listeners to recognize. In sonorant regions of speech, the consequence is often the insertion of glottal stops.

Glottal stops are produced by a change in the rate at which the vocal cords vibrate by a sudden closing and opening of the glottis during voiced speech, without changing the rest of the vocal tract configuration [17, pp. 38-42]. Acoustically, this means that the speech waveform becomes irregular in the fundamental period, but

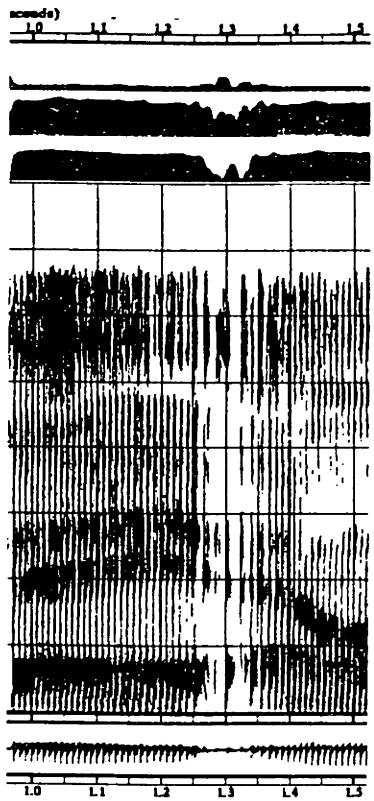


Figure 4.2: An example of a glottal stop

the formant frequencies remain the same. Figure 4.2 shows an example of a glottal stop.

Glottal stops account for 17.2% of the phonetic segments in the spelling corpus, and 99% occur between letters, forming clear letter boundaries. The other 1% of the glottal stops occur between a /ə/ or /ɪ/ and a vowel. In all cases found, the preceding letter is an M, N or H. Figure 4.3 shows an example of this type of glottal stop insertion. If the inserted /ə/ is considered to be part of the preceding letter, then all glottal stops occur at letter boundaries.

Although there are many situations in which two vowels are adjacent in the phonemic transcription of a string, these vowels are often separated by a glottal stop in the phonetic transcription. In the spelling corpus, glottal stops were inserted between 66.5% of the adjacent vowels, while an additional 22.2% were separated by a glide. This meant that in the spelling corpus, the sequence [VOWEL]/ʔ/[VOWEL]

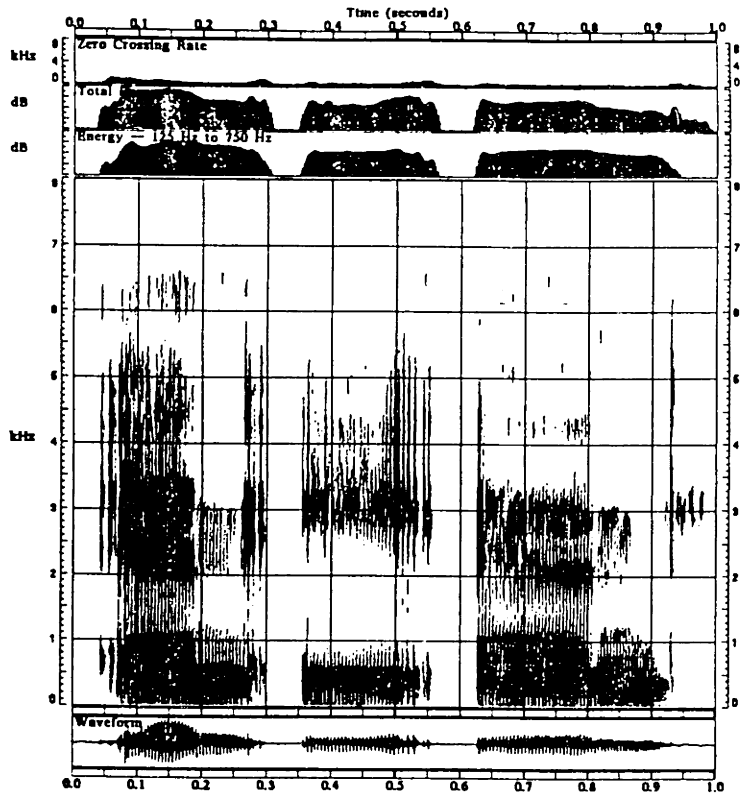


Figure 4.3: An example of an inserted /ə/ in the word NEN (/ɛnəiˠɛn/)

was six times more common than the sequence [VOWEL][VOWEL] and three times more common than [VOWEL][(inserted)GLIDE][VOWEL].

Speakers tend to deliberately insert glottal stops between vowels: 60.2% of the glottal stops in the corpus occur between vowels, and an additional 16.2% occur before a word-initial vowel. All of the remaining glottal stops occur either in the environment [VOWEL]/ʔ/[GLIDE] or [GLIDE]/ʔ/[VOWEL]. Since so many vowels are separated by glottal stops, the likelihood of insertion or deletion errors is reduced. This is confirmed by the fact that the number of insertion and deletion errors was small in both the auditory perception and spectrogram reading experiments.

A closer look at insertion and deletion errors reveals that listeners and readers respectively made about 71% and 80% of their insertion and deletion errors on vowels, and about 14% and 20% on glides. As discussed in Chapter 3, most of these errors occur in short, rapidly-spoken strings. In these cases, fewer glottal stops are inserted and vowel durations are shortened, making insertion and deletion errors

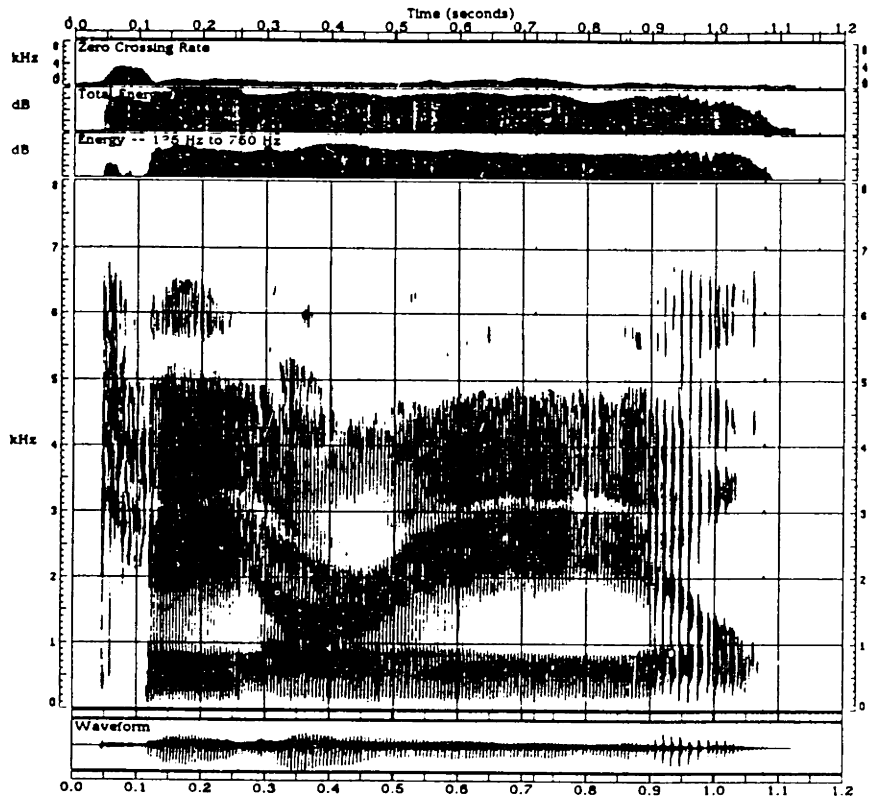


Figure 4.4: KRAAL /kɛˠɑɾɛˠɛˠɛl/

more likely.

4.2.4 Analysis of Vowel Gemination Errors

Another anticipated problem in spelling recognition is that of errors due to gemination, that is, the blending of two similar or identical into one. An example of this is recognizing the string BEET as BET by mistaking /iˠiˠ/ for /iˠ/.

When gemination occurs in ordinary continuous speech, the total duration of the two segments is usually lengthened, but the total duration is less than twice the combined durations of the individual segments in other contexts. In the spelling task, single vowels are sometimes mistaken for two consecutive vowels, and vice-versa. Figure 4.4 shows the spelled string KRAAL. In situations such as this, the number of vowel segments in the region may be determined from its duration.

A study of [VOWEL] and [VOWEL][VOWEL] regions confirms this hypothesis.

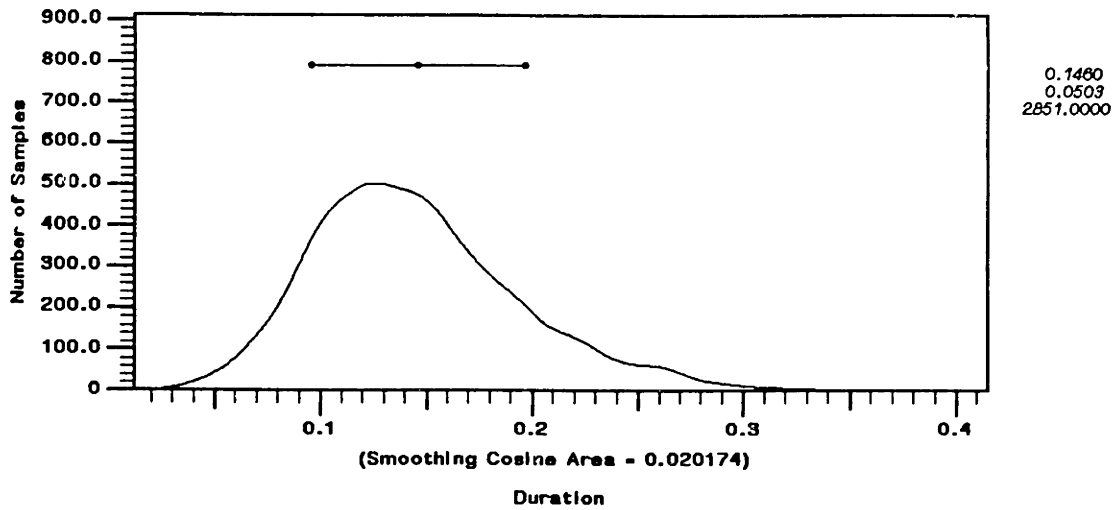
The average duration of single vowels is 142 milliseconds and average duration of vowel pairs is 286 milliseconds. Some examples of duration distributions are shown in Figures 4.5a and b. It can be seen that the duration of two consecutive vowels is almost exactly double that of a single vowel, suggesting that gemination of vowels does not greatly increase the difficulty of the task.

According to Klatt [11], the median duration of a stressed vowel is 130 milliseconds. The longer average duration of these vowels may be attributed to the fact that approximately 75% of the vowels in this corpus are tense. Figure 4.6 shows smoothed distribution for durations of tense and lax vowels. The tense vowels in this corpus, /iʸ, eʸ, ɑʸ, a, oʷ, u, ü, æ/, have an average duration of 155 milliseconds, while the lax vowels, /ɛ, ʌ, ɪ/ have a average duration of 117 milliseconds. A table of average durations of individual vowels spoken by male speakers can be found in Table 4.1.

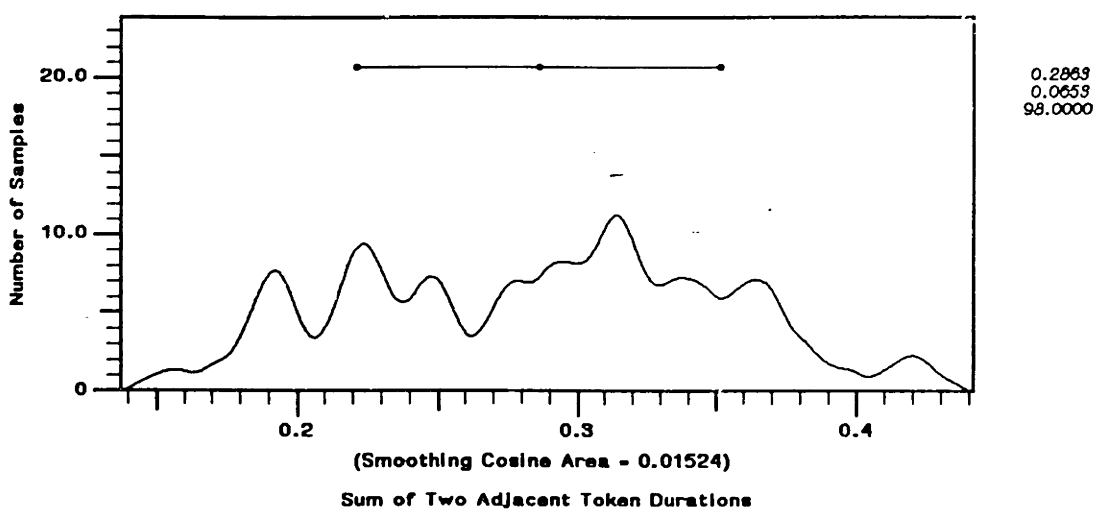
In ordinary continuous speech, pre-pausal lengthening tends to increase the duration of phrase- or sentence-final segments [15]. This trend is also found in the spelling corpus. The average durations of vowels in string-final and non-string-final positions are 201 and 139 milliseconds, respectively (Figure 4.7).

4.3 Comparison of Errors

The results of the experiments described in Chapter 3 confirm that some of the letters of the alphabet are easy to distinguish from each other acoustically, but some are very difficult. As discussed in Chapter 2, some letters are similar in their phonological structure, with the vowel portion of a letter being similar or identical. While the vowel serves to reduce the number of letter candidates, the rest of the letter, usually a relatively small part of it, must provide the acoustic information necessary to make a final decision. As an illustration, consider a letter whose structure is known to be [CONSONANT]/iʸ/. Given this information, the letter



(a)



(b)

Figure 4.5: Durations of (a) Single Vowels and (b) Vowel Pairs

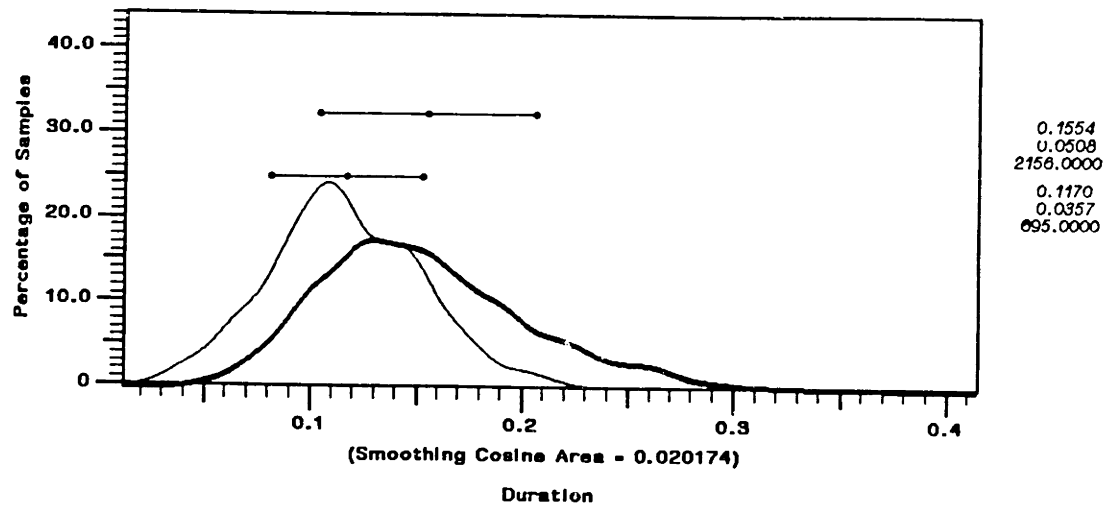


Figure 4.6: Durations of Tense and Lax Vowels

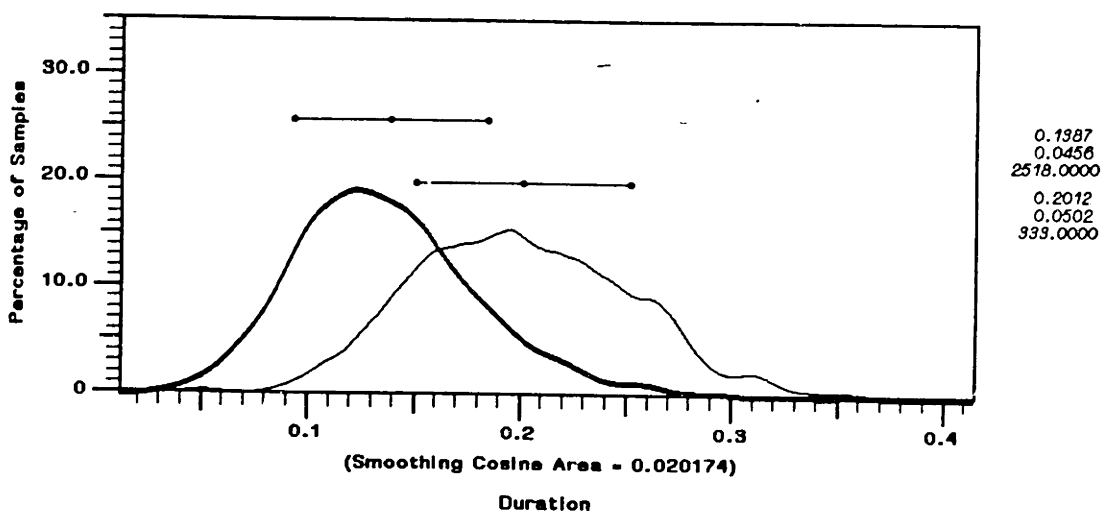


Figure 4.7: Durations of Final and Non-Final Vowels

Vowel	μ (msec)	σ (msec)	# of Tokens
i^y	146.7	47.9	1027
e^y	157.6	44.2	324
a^y	208.8	48.7	273
a	140.7	33.6	205
æ	138.2	22.9	7
o^w	157.2	42.9	188
u	138.4	54.9	81
ü	113.6	42.1	51
ɛ	121.3	33.1	639
ɪ	58.6	25.9	28
ʌ	77.2	21.8	28
OVERALL:	146.0	50.3	2851

Table 4.1: Statistics for Vowel Durations

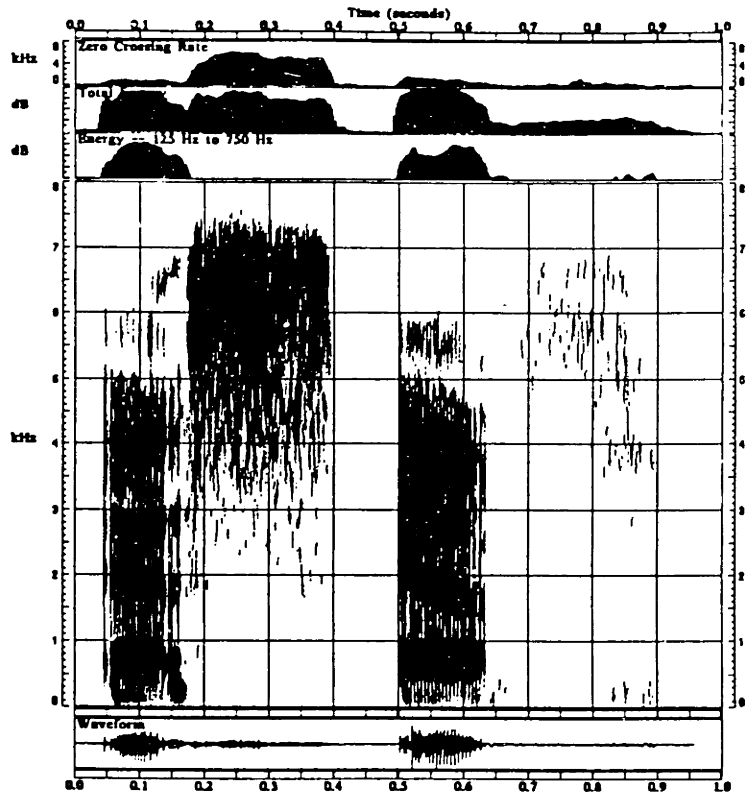


Figure 4.8: Spectrogram of S (/ɛs/) and F (/ɛf/)

could be either B, C, D, G, P, T, V or Z.

Based on existing speech recognition systems' performance, [9,14,5], we can hypothesize that they would be able to recognize acoustically dissimilar letters. However, such a system would probably have great difficulty distinguishing some letters, such as M and N. It is therefore instructive to focus on errors made by subjects of the auditory perception and spectrogram reading tests described in Chapter 3.

One of the questions that arises from analyzing the results is why the listeners made different mistakes from the spectrogram readers. Although listeners and readers sometimes made the same type of mistake (e.g., substituting B for D), one of the groups made it proportionately far more often than the other, and usually not on the same particular token.

An illustration of the difference in results is shown in Figure 4.8. The figure shows a spectrogram of the letters S and F, which is a pair of letters that the

listeners confused more often than the readers. Spectrogram readers can distinguish between the /s/ of S and the /f/ of F more easily than listeners can because they can see the difference in energy between the two phonemes in the mid-frequency range more easily than listeners can hear it. Spectrogram readers performed poorer in other instances, presumably due to the fact that they had not learned to utilize subtle acoustic cues. From this we may conclude that listeners and readers make different errors because some acoustic cues are more obvious to listeners than to readers, and vice-versa.

When examining the errors, we should focus on those made by spectrogram readers rather than listeners, because spectrogram reading makes explicit use of acoustic-phonetic knowledge that can potentially be extracted and implemented in a recognition system. Also, the emphasis should be placed on studying substitution errors, since they comprise 68% and 92% of listening and reading test errors, respectively.

Substitution errors made in these tests were described in Chapter 3. Some of the errors were symmetric; Letter 1 was mistaken for Letter 2 about as often as Letter 2 was for Letter 1. Other errors were asymmetric. Why these asymmetric errors occur and how they can be resolved are questions that may be answered by examining specific asymmetric confusions.

4.4 Analysis of Readers' Asymmetric Errors

Some of the most common asymmetric errors are listed in Table 4.2. Together, they comprise 30% of all asymmetric errors.

The letter R is more likely to be called an I than the other way around, and an examination of I-R errors helps explain why these confusions occur. Figure 4.9 shows a spectrogram of the string CRUR, which was transcribed as CIUR by a spectrogram reader. Unlike the second R, the first R of the string does not have

Letter Pair		# of Errors	
1st	2nd	1st mistake for 2nd	2nd mistaken for 1st
I	R	2	6
I	O	2	5
G	P	0	6

Table 4.2: Most Common Asymmetric Errors Made by Readers

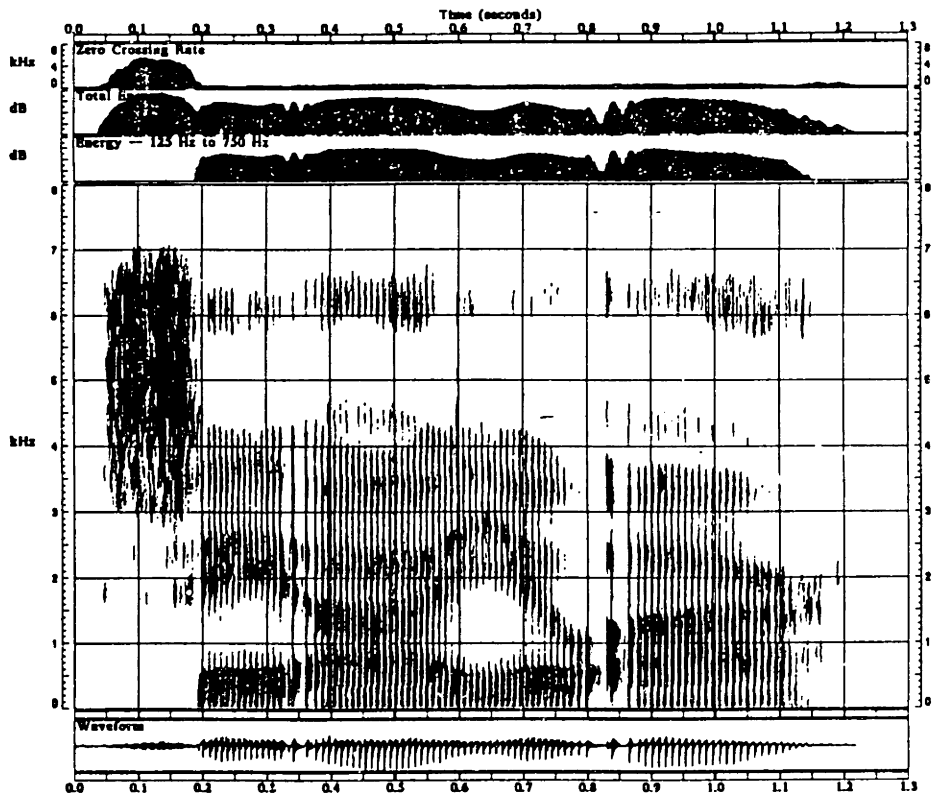


Figure 4.9: Spectrogram of CRUR (/siʔaryuar/)

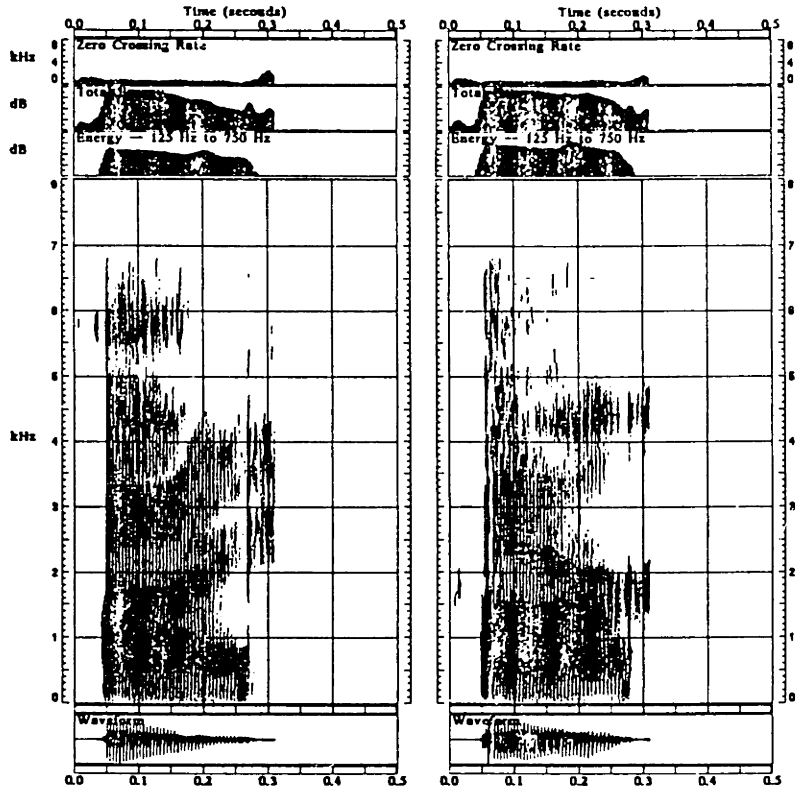
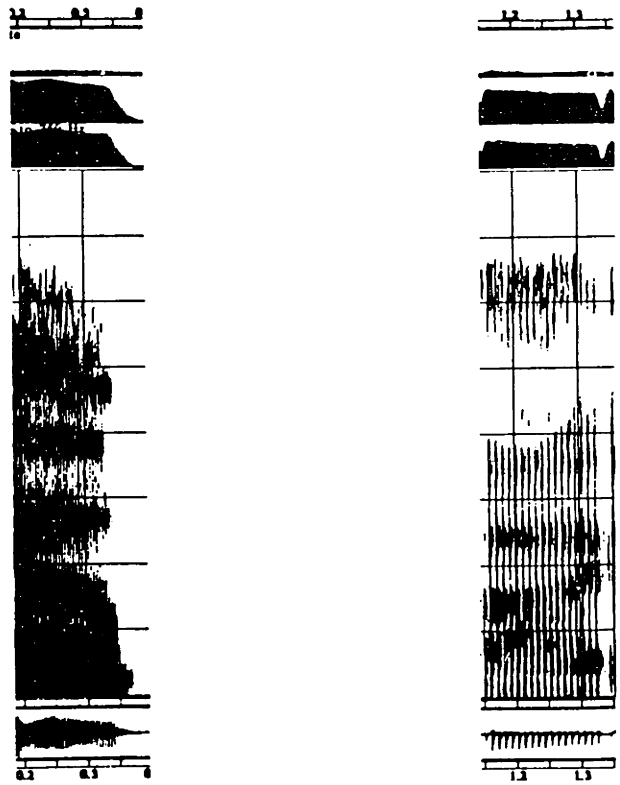


Figure 4.10: Spectrograms of (a) /ɑʸ/ and (b) /ɑr/

a low third formant characteristic of /r/. Instead, it is raised due to the influence of the following /y/, causing it to strongly resemble /ɑʸ/, shown in part (a) of Figure 4.10. Most of the R tokens that were mistaken for I were followed by /y/ or /i/. Part (b) shows a typical /ɑr/, and a comparison of the two shows that if an R is followed by a segment that raises or lowers the second and third formants, it can be confused with an I.

The asymmetric confusion between I and O has a similar explanation. O was more likely to be mistaken for I than vice-versa, and an examination of the tokens on which this error was made show why. If O was followed by U, it was sometimes called I, because, as in the I-R confusion, the third formant of the O was raised from its characteristic low position (see part (a) of Figure 4.11) to a higher frequency more typically seen in the letter I (shown in part (b) of Figure 4.11). Once again,



(a) (b)

Figure 4.11: Spectrograms of (a) /oᵛ/ and (b) /aᵞ/

as in the previously described confusion, the right context of the O can cause it to be mistaken for I.

This explains why O and R are sometimes called I, but it does not explain why the reverse is not as common. In order for I to be called an R or O, it could be followed by a segment that lowers the third and second formants, respectively. This situation did not occur in the spelling corpus. However, it was found that both I-R confusions occurred when I was at the end of a string. Segments at the ends of utterances are subject to pre-pausal lengthening, and this makes the formant transitions more gradual than is usually seen in /aᵞ/. Also, the signal near the end of an utterance can be noisy due to excess aspiration, and in both confusions, the trajectory of the third formant was hard to track. Both of these characteristics are shown in Figure 4.12 for the last two letters of the string RIANCEPI. The figure

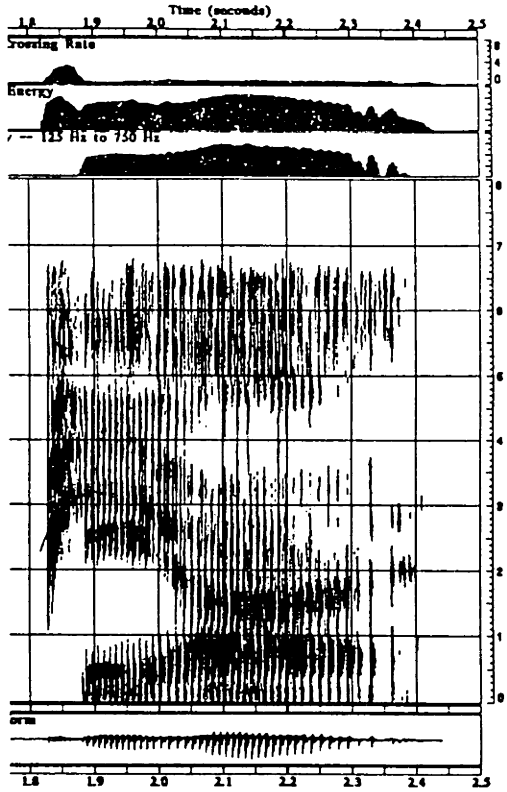


Figure 4.12: Spectrogram of PI (/pi^{y}a^{y}/)

shows the last two letters, PI, which were transcribed by a spectrogram reader as PR.

I-O confusions occurred when the right context of the I caused the second formant of /a^{y}/ to be lowered so that it resembled /o^{w}/. Figure 4.13 shows a spectrogram of IL, the last two letters of the string MISTIL, which were transcribed by the reader as OL.

The third asymmetric confusion in the table is for G versus P. The letter P was mistaken for G six times, but the opposite mistake was never made. A closer look at this confusion reveals that 5 of the 6 P-G errors were made when P occurred in a string-initial position, as shown in the spectrogram of P from the string PRIN in part (a) of Figure 4.14. String-initial /p/ is unusually strong and the release contains a great deal of aspiration noise, so that it resembles the /p̆/ shown in Figure 4.14b. An ordinary /p/ is far less likely to be mistaken for a /p̆/, since it has the pencil-

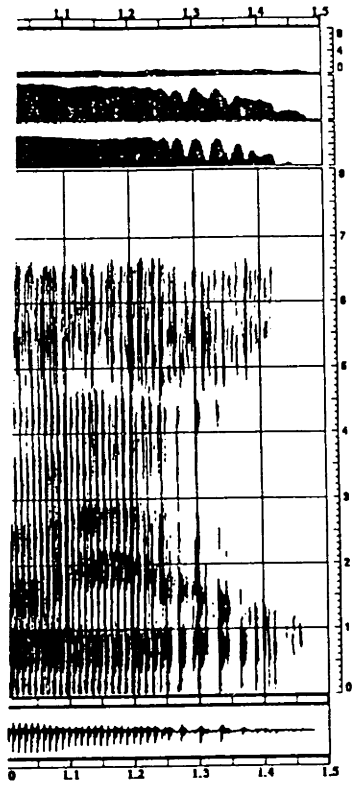


Figure 4.13: Spectrogram of IL (/aɪɛl/)

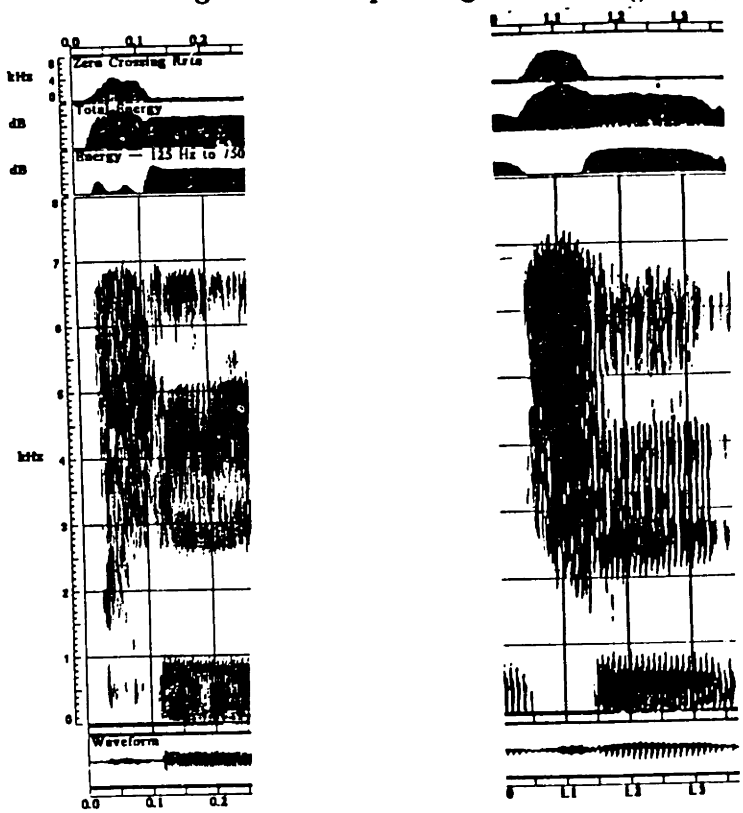


Figure 4.14: Spectrogram of (a) P (/piɪ/) and (b) G (/ʃiɪ/)

thin burst and comparatively little frication noise. Also, /ʧ/ is voiced while /p/ is unvoiced, and evidence of this distinction is usually seen by examining the voice bar and voice onset time of the segment. The voice bar is found in the closure portion of voiced stops and affricates, and is caused by tissue vibration around the neck. The voice onset time is shorter for voiced segments than unvoiced ones. However, sentence-initial segments usually do not contain prevoicing, whether or not they are voiced, so that cue for distinguishing between /p/ and /ʧ/ is not available to the reader. Therefore, he is forced to rely on the presence of aspiration noise in the burst and voice onset time, both of which are misleading for /p/. These /p/ segments are not pathological, they are merely products of overarticulation which can sometimes be a problem.

Examining specific asymmetric confusions has led to some interesting insights as to why they occur, and allows us to conclude that such confusions arise because the acoustic properties of some phonemes are modified when they occur in certain phonetic environments. These confusions may be resolved if context is taken into account when attempting to recognize the letter.

4.5 Analysis of Readers' Symmetric Errors

4.5.1 Introduction

While some confusions are asymmetric and can be explained and resolved by taking their context into account, others occur independent of phonetic environment and are more symmetric. Symmetric errors are more prevalent than asymmetric errors, and they occur presumably because subjects cannot find the right acoustic-phonetic cues for distinguishing between certain pairs of letters or phonemes. Resolution of these errors may be possible by studying the confusing pairs and finding acoustic cues for distinguishing between them.

Spectrogram readers made fifty-one different substitution errors, but the four

most frequent confusions, G-T, A-E, M-N and O-L together comprise 42.8% of the total. If acoustic cues can be found for resolving these symmetric errors, the number of confusions and the overall error rate will be drastically reduced. Therefore, we conducted a set of experiments focusing on finding acoustic features that can distinguish these letter pairs.

4.5.2 Description of the Experiments

In these experiments, acoustic features are used to determine the identities of letters. However, the conditions under which these experiments are performed differ from those of the auditory perception and spectrogram reading experiments. First of all, in this experiment, the endpoints of the segments we are trying to recognize are given: that is to say, we assume that segmentation of the signal has already been done. Also, the decision being made here is a binary one: the segment in question must be one of only two. These two combine to make the task easier than that of the listeners and spectrogram readers. Other differences between the experiments include difference in information given about speaker identity. Listeners were given no speaker information, readers were given speaker identities, and in the acoustic resolution experiment, male tokens were separated from female tokens.

Most of the acoustic resolution experiments were performed on male data only. Because of the smaller dimensions of the female vocal tract, the fundamental frequency of female speech and is higher than for male speech. The optimal window for processing male speech is too long for female speech [17, pp. 310-314], which means that the frequency resolution of female speech is greater than desired. As shown in the spectrogram of Figure 4.15, strong harmonic structures, particularly in the region around the first formant, are often present for female speakers. A trained spectrogram reader has learned to ignore these extraneous spectral peaks. However, automatic formant trackers will have a great deal of difficulty with them. For this reason, female speech is not used in most of these experiments.

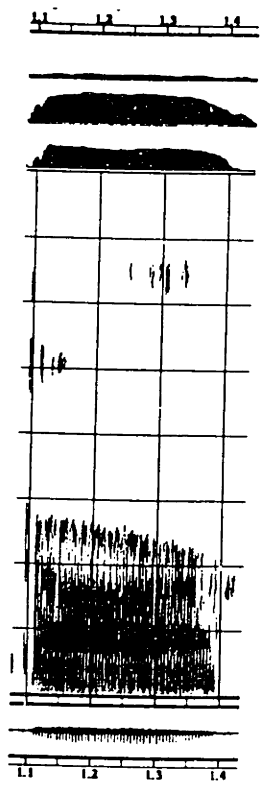
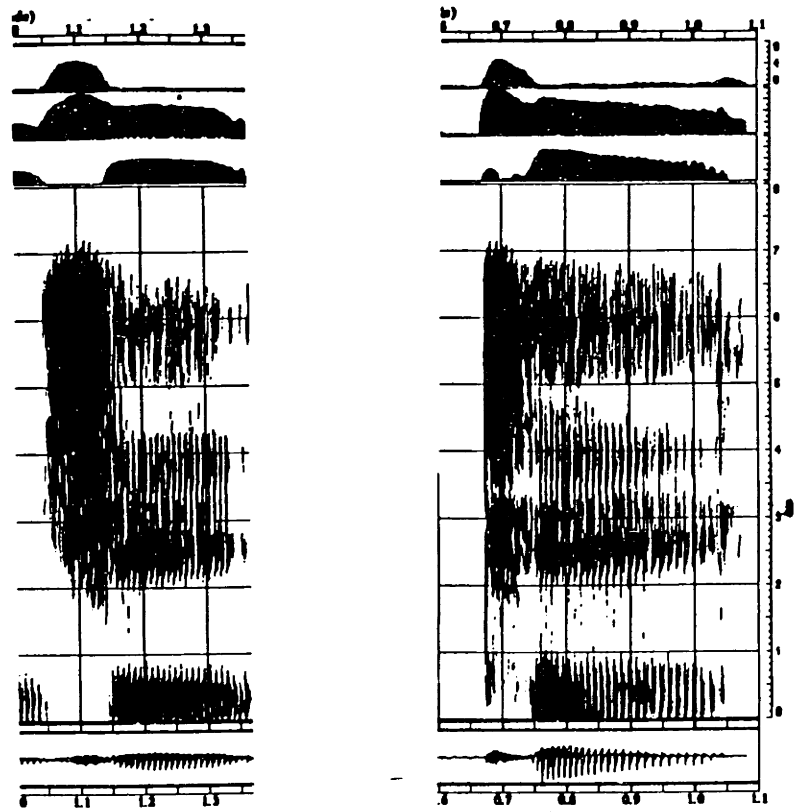


Figure 4.15: Spectrogram of R (/ɑr/) spoken by a female speaker.

Different acoustic parameters determined by examining approximately 90% of the data. Once the appropriate parameters were determined from the training data, these cues were tested on the remaining 10% of the data to determine their effectiveness.

4.5.3 G-T Confusions

The most common substitution error made by spectrogram readers was mistaking G for T, and vice-versa. Spectrograms of the two letters are shown in Figure 4.16. The confusion is between the /t/ in T, which is often unusually strong in spelled speech due to overarticulation, and the /ʧ/. Two features were used to resolve this confusion. The first is the presence or absence of voicing in the closure portion of the consonant, before the burst. Since /ʧ/ is voiced and /t/ is not, we would expect to see some prevoicing during the closure for /ʧ/ but not for /t/. This is



(a) (b)

Figure 4.16: Spectrograms of (a) G (/ʄiʸ/) and (b) T (/tiʸ/)

a good feature except for string-initial G and T tokens, because prevoicing does not ordinarily occur at the beginning of an utterance. The second feature is the characteristics of the noise following the burst. Since /ʧ/ is an affricate, it contains frication noise, and since /t/ is a stop, it contains aspiration noise. Frication noise tends to have a flat spectrum, while the aspiration noise contains peaks in energy around the higher formant frequencies of the following sonorant. For /t/ this means that the second and third formant are visible in the noise, as can be seen in Figure 4.16. This difference in noise type is expressed quantitatively by the amount of energy found in the region 3100-3600 Hz for males. For /t/, this represents the region between the emerging third formant and higher-frequency frication noise. Even though the appropriate frequency band varies from speaker to speaker, such variability is greatly reduced since all the /t/ tokens are followed by /iʸ/.

The results of this experiment are shown in the first row of Figure 4.17 with the training and testing accuracy rates combined, along with the results from the auditory perception and spectrogram reading experiments. The results from this experiment are shown for male speakers only, whereas the results from the other experiments are for both male and female speakers. These results are shown in the form of confusion matrices that indicate how well each individual confusions are resolved. Average error rates are shown in Figure 4.18 for easier comparison of overall results. It can be seen from both figures that while listeners have the best performance record for distinguishing G from T (99.5% correct), the acoustic resolution test using only one or two acoustic features has a higher accuracy rate than spectrogram readers (96.8% versus 89.9%).

4.5.4 A-E Confusions

The second largest group of substitution errors were A-E confusions. Spectrograms of these two letters are shown in Figure 4.19. The formant trajectories of the vowels /eʸ/ and /iʸ/ are sometimes modified by phonetic context in such a way as to cause

Guessed Letters (%)

Correct Letters (%)

Listeners

	G	T
G	100	0
T	0.9	99.1

	E	A
E	99.1	0.9
A	0	100

	O	L
O	98.4	1.6
L	0	100

	M	N
M	98.5	1.5
N	0.8	99.2

Readers

	G	T
G	89.3	10.7
T	11.5	88.5

	E	A
E	97.3	2.7
A	3.3	96.7

	O	L
O	98.4	1.6
L	10.3	89.7

	M	N
M	89.4	10.6
N	7.8	92.2

Acoustic Experiment

	G	T
G	94.6	5.4
T	0.9	99.1

	E	A
E	97.8	2.2
A	1.1	98.9

	O	L
O	94.4	5.6
L	4.5	95.5

	M	N
M	80.3	19.7
N	6.2	93.8

Figure 4.17: Analysis of Worst Substitution Errors

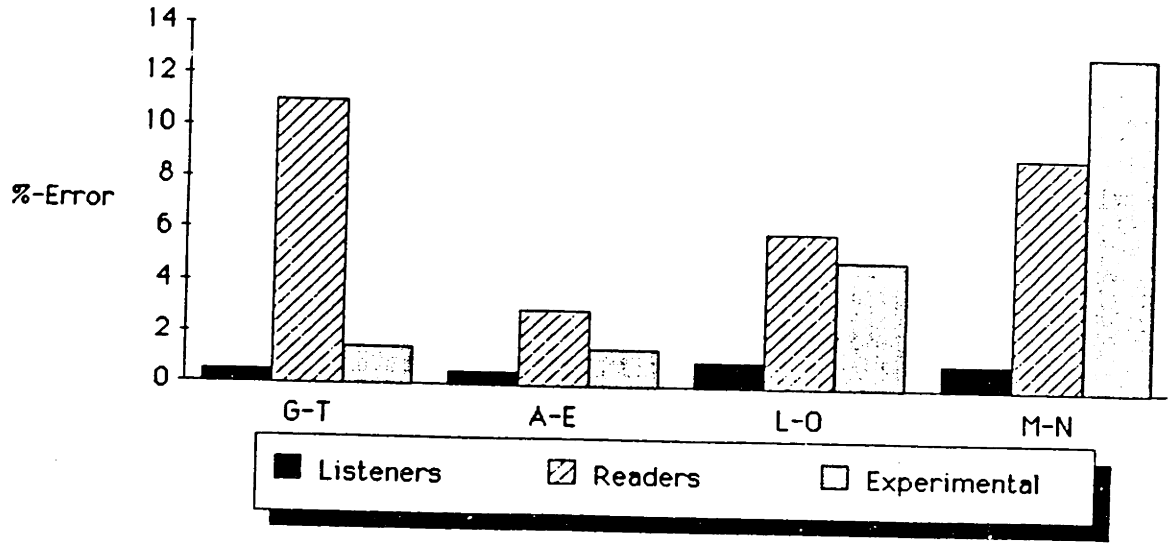


Figure 4.18: Symmetric Errors

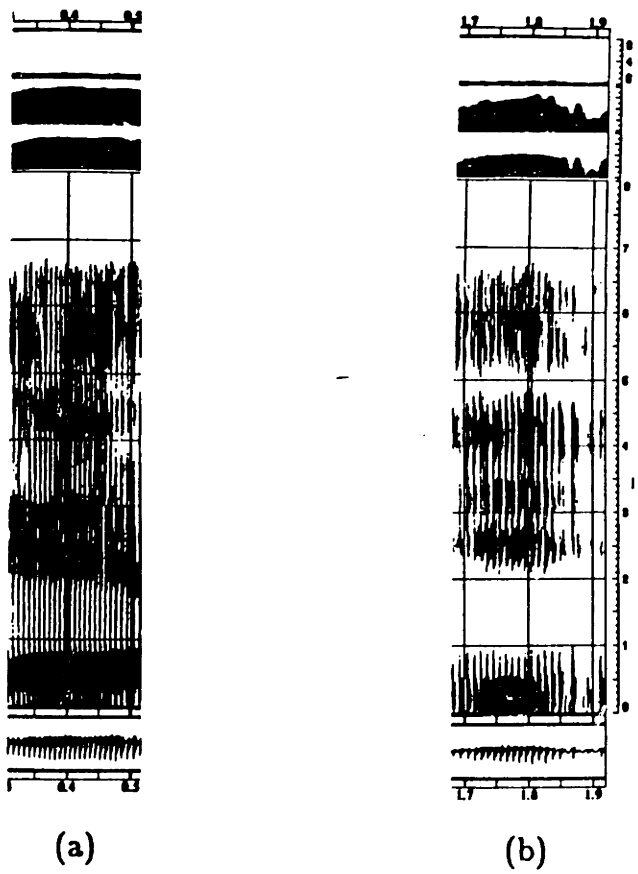


Figure 4.19: Spectrograms of (a) A (/e^{y}/) and (b) E (/i^{y}/)

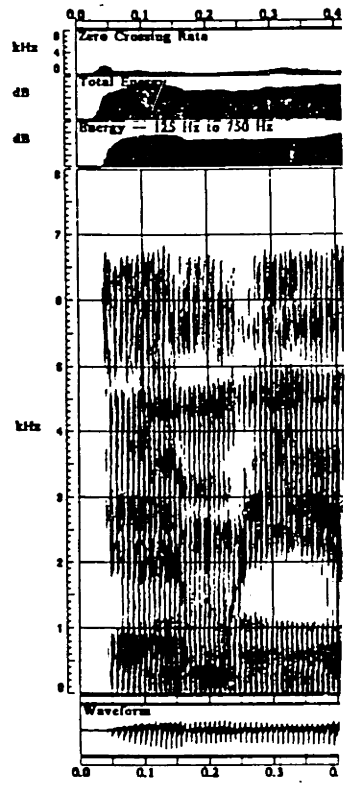
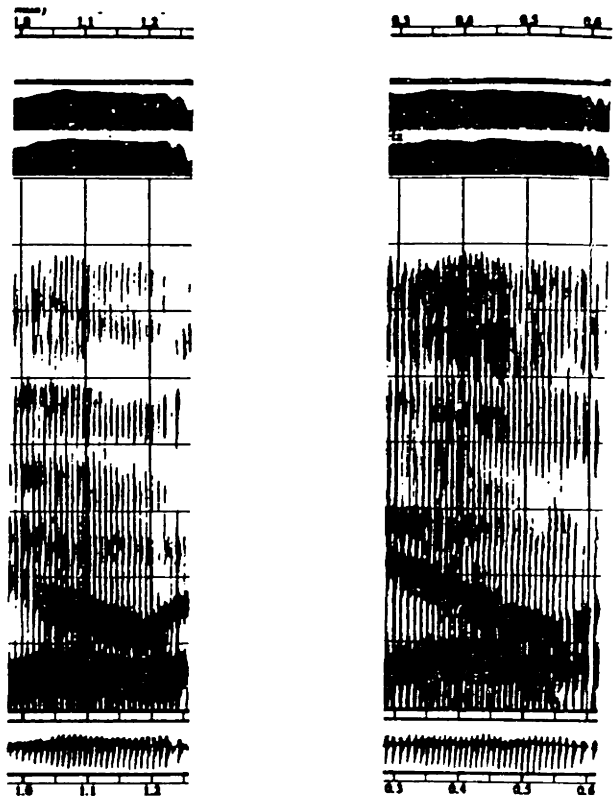


Figure 4.20: Spectrograms of ME (/ɛmiˀ/)

them to be mistaken for each other. As shown in Figure 4.20, for example, if the letter E is preceded by the letter M, the /m/ of the M can lower the second formant of the following /iˀ/ so that it resembles /eˀ/.

A number of acoustic features were tested on the A and E tokens, and it was found that the best separation results were obtained when the tokens were separated according to left phonetic context. Tokens preceded by phonemes such as /l/, /w/ or /m/ were partitioned from the rest, and then the same features were used to resolve tokens in both groups. The two features used were the average value of the first and second formants across each token, which are generally lower for /iˀ/ or /eˀ/ preceded by /l/, /w/ or /m/.

A-E confusion matrices and overall error rates for the three recognition experiments can be found in Figures 4.17 and 4.18, respectively. A comparison of results for this experiment to those of the previous recognition experiments show that, as



(a) (b)

Figure 4.21: Spectrograms of (a) O (/oʷ/) and (b) L (/ɛl/)

in the case of G-T, the listening test yielded the highest accuracy rate (99.5%), followed by this acoustic resolution experiment (98.3%) and the spectrogram reading experiment (97.0%). Once again, a careful acoustic analysis gives better results than those obtained by spectrogram readers, and this not only be because formants were more accurately measured, but also because the identity of the left phonetic context was known.

4.5.5 O-L Confusions

Spectrogram readers also had difficulty distinguishing O from L. At first glance, this confusion is a surprising one, since the acoustic differences between these letters are evident to a listener. However, the letters are actually very similar acoustically, so much so that even listeners occasionally mistook one for the other. Figure 4.21

shows spectrograms of the two letters which demonstrate the resemblance between the letters; each is composed of a vowel followed by a semivowel. The semivowels, /w/ and /l/ are one of the most difficult pairs of English phonemes to resolve. Efforts made to use the semivowel part of each letter to help distinguish them from one another proved fruitless, so attention was instead directed towards the vowel portion.

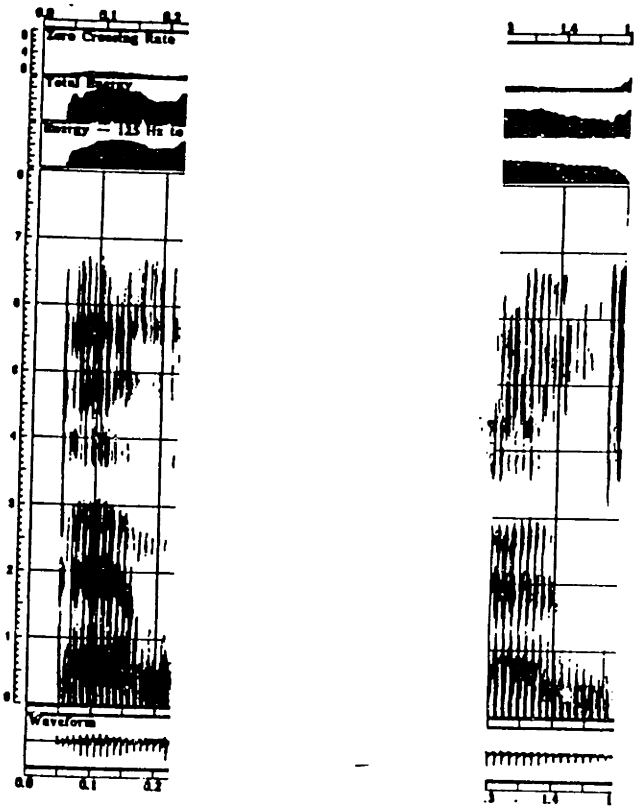
The vowel of O, /o^{w}/, is a back vowel, while the vowel of L, /ɛ/ is a front vowel, so the average value of the second formant is a good feature for distinguishing between them. However, using the average value of the formant over the entire vowel yields poor results because the following semivowel lowers the last part of the second formant, resulting in average second formant frequencies for /o^{w}/ and /ɛ/ that are virtually the same. Using the average second formant calculated over the first seventy-five milliseconds of vowel gives better separation results.

As in the A-E resolution experiment, the vowel formants are modified by the phonetic environment, so the data is partitioned by context and the same features is used to distinguish O tokens from L tokens within each group. Here, tokens that are preceded by phonemes that tend to raise the second formant, such as /i^{y}/, /y/ and /č/, are separated from the rest.

Besides the average value of the beginning of the second formant, duration of the vowel segment is also helpful for resolving O-L confusions. The vowel /ɛ/ is a lax vowel, while /o^{w}/ is not, and therefore typically has a shorter duration than /o^{w}/.

Using these two features, we can acoustically resolve O and L tokens with an overall accuracy rate of 95.0%. This is a higher accuracy rate than that obtained by spectrogram readers (94.0%), but, once again, the listeners' performance was significantly better (99.2%).

4.5.6 M-N Confusions



(a) (b)

Figure 4.22: Spectrograms of (a) M (/ɛm/) and (b) N (/ɛn/)

The letters M and N are the final pair of symmetric confusions to be examined. This confusion was often made not only by spectrogram readers, but by listeners as well. Spectrograms of M and N are shown in Figure 4.22. Both M and N consist of the vowel /ɛ/ followed by a nasal, /m/ or /n/.

Acoustically, the letters M and N are almost identical. The primary difference between them is in the place of articulation The place of articulation, labial for /m/ and alveolar for /n/, influences the trajectory of the preceding vowel. Figure 4.22 shows that in M, the labial /m/ causes the formants of the preceding /ɛ/ to fall sharply at the end of the vowel. An examination of the formant frequencies of N in the same figure show no such rapid changes in /ɛ/.

The second formant of /ɛ/ in M is affected by the following labial more than the other formants, while the second formant of /ɛ/ in N is more stable than other formants. The locus for the second formant of an alveolar sound is approximately 1800 Hz for male speakers, so we would expect that the second formant of /ɛ/ would be at that frequency immediately before the /n/, and that it would be fairly level. A good way to express this difference quantitatively is as a measure the slope of the second formant during the last ten milliseconds of the vowel.

Using this feature, an attempt was made to separate M tokens from N tokens. As in the previous experiments, comparisons of the results of the auditory perception, spectrogram reading and acoustic resolution experiments are shown in Figures 4.17 and 4.18. This time, both the overall accuracy rates of the auditory perception and spectrogram reading experiments (98.9% and 90.8%) were better than those obtained in the acoustic resolution experiment (87.1%).

The fact that this acoustic resolution experiment did not yield better separation results than those obtained in the spectrogram reading experiment is due to two factors. First, unlike the other acoustic resolution experiments, only one feature was used to try to accurately partition the data. All three of the other experiments used two features, and higher accuracy rates than those in the spectrogram reading

experiment were obtained. Obviously, the feature used did not adequately capture the acoustic differences between M and N. Second, as was mentioned before, the techniques used in these experiments to find formant frequencies do not work well for female speech, and sometimes perform poorly on male speech. Formant information is imperative for resolving many confusions. However, formant tracking is error-prone, which partially explains the difficulty in acoustically resolving M and N. Since only the last ten milliseconds of the vowel were used, this meant an error in formant tracking could not be smoothed out very well.

There are two paths that may be taken to better resolve this confusion. First of all, the acoustic resolution experiment can continue as before, and other features can be tested to see how well they separate the tokens. For example, the nasal murmur itself has not yet been used to try to distinguish N from M. According to Glass [6], there are some spectral differences between /m/ and /n/, but they are usually diminished in a large data-set such as this because the differences are speaker- and context-dependent. However, in this experiment, speakers are separated by sex and the left phonetic context is the same for all tokens. Features of the nasal, along with better measurements of formant movement at the end of the vowel may lead to better separation of M tokens and N tokens.

Secondly, a different approach developed by Seneff [19], in which the spectrum of the vowel portion of a letter is characterized without specifically tracking formants, may be the answer. This method, which incorporates a non-linear auditory model into the analysis of vowel spectra, yields spectrographic representations of these vowels that consist of a series of lines, called "line-formants." Once obtained, these line formants contain enough information about formant frequencies and trajectories to be used to discriminate between vowels.

Line formants for /ɛ/ followed by /m/ and /n/ are represented in a two-dimensional probability distribution of the frequency and slope of the lines, and are shown in Figure 4.23. It can be seen that acoustic differences between the two sets of /ɛ/

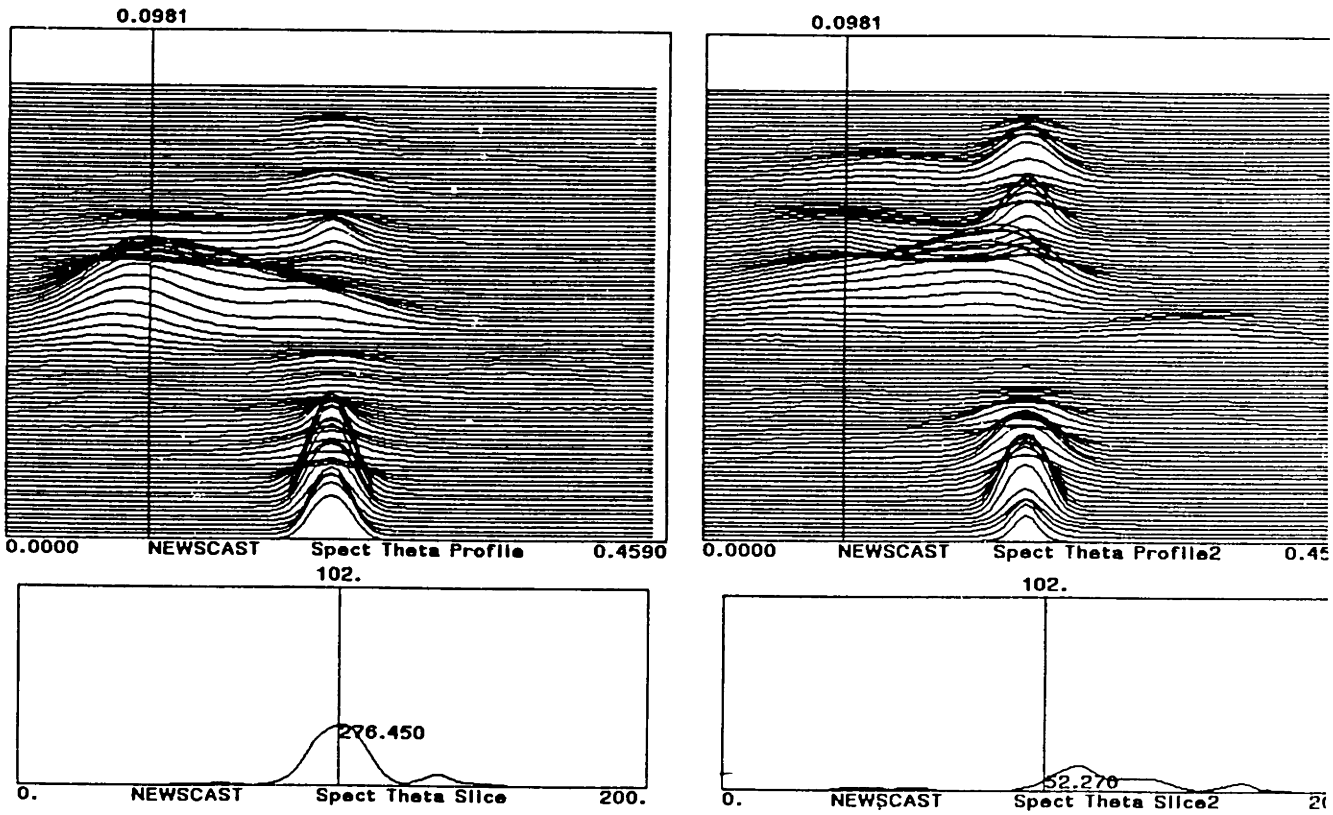


Figure 4.23: Line formants for /ɛ/ followed by /m/ and /n/

Proposed Letters (%)

Correct Letters (%)

	Training			Testing	
	M	N		M	N
M	88.6	11.4	N	88.2	11.8
M	15.4	84.6	N	24.2	75.8

Figure 4.24: Resolution of M vs. N using Line Formants

tokens are accentuated by the application of the auditory model. In a preliminary experiment using the majority of /ɛ/ tokens for training and the remainder for testing, the auditory model was used to attempt to discriminate between M and N tokens. The results are shown in Figure 4.24.

The overall recognition rate for this test, which was performed on both male and female data combined, was 88.6% for training data and 82.0% for test data. Although this is lower than the rates obtained in the other recognition experiments, the data does include both male and female speakers. This approach seems to be promising and may eventually lead to improved resolution of M-N and other confusions.

4.6 Conclusions

Spelled strings differs from ordinary continuous speech in three major ways: spelled strings are composed of a smaller phoneme set and a limited number of permissible phonetic sequences within letters, they are primarily made up of stressed syllables, and they contain a far greater number of glottal stops. All of these features can be used to facilitate continuous letter recognition.

Errors made in trying to recognize spelled strings are primarily susbstitution

ones. Other errors, which result from not being able to find letter endpoints within a string, do not often occur because natural boundaries formed by, among other things, glottal stops, help to segment strings into letters. Some substitution errors were asymmetric, and occur because the effects of coarticulation cause one letter to resemble another, while the opposite problem does not occur. These errors may be resolved by taking the phonetic environment of a letter into account when trying to determine its identity.

Other errors were symmetric, and tended to occur independent of context. By measuring certain acoustic differences between the letters, three of the four worst symmetric confusions were resolved with a higher accuracy rate than that obtained by spectrogram readers, who used a similar approach.

These results lead to two conclusions. First of all, since better overall performance than spectrogram readers was achieved in these acoustic resolution experiments, using only one or two simple and rather crude acoustic measurements, we expect that accuracy results would be even better if a greater number of more sophisticated acoustic features were used. Second, since the accuracy rate for these experiments is so high for these difficult confusions, we expect even higher accuracy rates for other, less acoustically similar confusions. Therefore, we hypothesize that if spectrogram readers can achieve an accuracy rate of approximately 91% using only acoustic-phonetic information, a spelling recognition system using only acoustic measurements similar to those described in the above acoustic resolution experiments may be able to achieve an even better performance rate.

Chapter 5

Conclusion

5.1 Summary of Results

Although acoustic-phonetic information is important for recognition, it is not sufficient; continuously-spoken letters are difficult to recognize due to acoustic similarities between some of them. Information from other knowledge sources may aid in spelling recognition.

In the general continuous speech recognition problem, syntactic constraints may be exploited to facilitate recognition. In continuous letter recognition, if the task is restricted to recognizing spelled words, then knowledge of the rules of spelling can improve accuracy.

Knowledge of spelling rules aids in continuous letter recognition because lexical constraints on words are strong. A lexical study conducted using the MPD showed that not only were some letters and sequences of letters much more likely to occur than others, but also that there were a limited set of letter combinations that were permissible. The predictability of English was demonstrated; the more letters known in a word, the greater the constraints on what the other letters could be and the greater the redundancy of information contained in the word.

Both acoustic-phonetic and lexical information are used to achieve recognition of

ordinary spelled words. However, it is difficult to determine how much information is derived from each of the knowledge sources. Although acoustic-phonetic information alone is not adequate for perfect spelling recognition, its actual performance rate is not known. Determining the sufficiency of acoustic information shows the relative importance of each of the available knowledge sources.

Spelling recognition experiments were conducted using a corpus composed of words and "wordlike" non-words to determine the adequacy of acoustic-phonetic knowledge alone. In an auditory perception experiment, listeners achieved an accuracy rate of 98.4% and in a spectrogram reading experiment, spectrogram readers achieved an accuracy rate of 90.7%. These results show that listeners may rely almost exclusively on acoustic-phonetic information to recognize continuously-spoken letters. Also, spectrogram readers, who use similar recognition techniques as would be used by a spelling recognition system, perform fairly well using only acoustic-phonetic information. Adding lexical information and doing a more sophisticated acoustic analysis should further increase the accuracy rate of the acoustic-phonetic feature-based approach used by spectrogram readers. The next step is to explore possible ways of integrating information from the acoustic-phonetic and lexical knowledge sources.

5.2 Integration of Knowledge Sources

Based on the results of the spectrogram reading experiment, the assumption that we can develop a fairly accurate spelling recognizer using just acoustic-phonetic information and the techniques used by spectrogram readers is a valid one. Spelled speech possesses certain acoustic characteristics which are not found in ordinarily continuous speech. These include a limited vocabulary and phoneme set, a large number of glottal stops and a predominance of stressed syllables. These features may be exploited to aid in recognition.

s i e č e t i

z ǰ

Figure 5.1: Phonetic transcription lattice for the word CHAT.

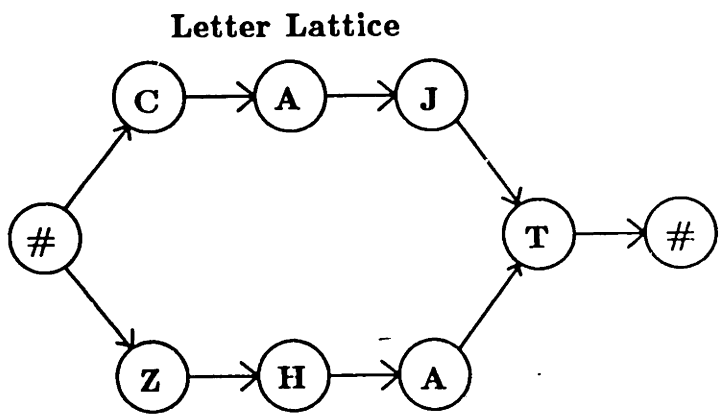


Figure 5.2: Letter lattice for the word CHAT.

	0th Order	1st Order	2nd Order
CHAT	947.8	1.0921	.00381
ZHAT	20.97	0	0
ZAJT	0.589	0	0
CAJT	26.698	0	0

Table 5.1: Path probabilities ($\times 10^{-8}$) using Markov Models

Acoustic-phonetic information alone can reduce the number of possible letter transcriptions of a spelled string. As an example, suppose we are asked to recognize the spelled string CHAT. Using only acoustic information, a phonetic transcription lattice, shown in Figure 5.1, may be obtained. Using knowledge about the phonetic characteristics of letters, the phonetic transcription lattice can be translated into a letter lattice, which is shown in Figure 5.2.

Any one of the paths shown in the letter lattice of Figure 5.2 is acoustically valid. However, only one at most is actually correct. The next step is to determine the best way to decide which path to follow.

One approach is to simply follow the best acoustic path. When creating the phonetic transcription lattice, the signal is segmented and one or more phonetic transcriptions is proposed for each segment. Ordinarily, the transcriptions are ranked according to probability of correctness, and this ranking could be taken into account when determining the final transcription of the word.

The fact that there is more than one reasonable path proves the insufficiency of acoustic information. However, if the letter string must form a word, then knowledge of the syntax of English words as expressed by the rules of spelling can be used. Lexical information can be applied toward finding the best path through the lattice to come up with the most likely word candidate.

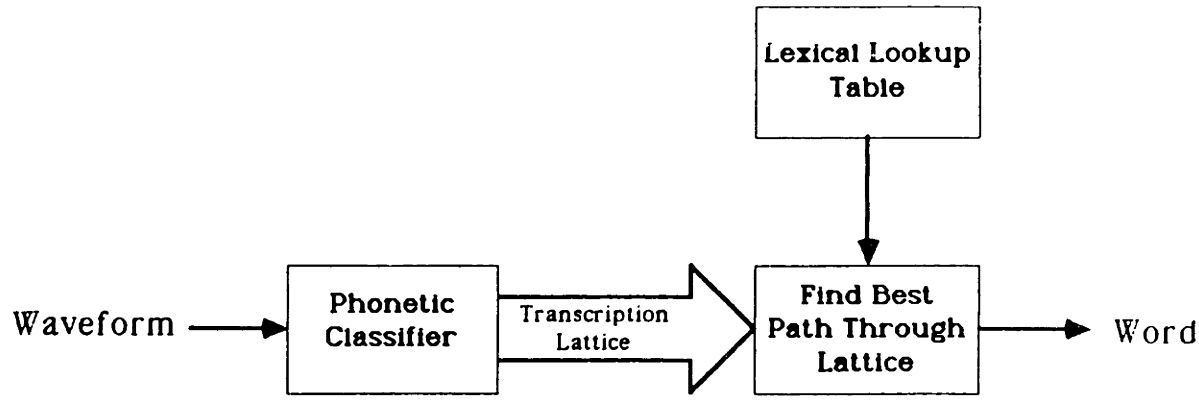


Figure 5.3: Proposed spelling recognition system.

The application of lexical knowledge can be demonstrated using CHAT once more as an example. Information about the frequencies of letters and letter sequences gathered in the previously-conducted lexical study can be used to find the best path through the letter lattice. Table 5.1 lists path probabilities using zeroth-, first- and second-order Markov Models. The order of the Markov model describes how many past states are used to determine the probability of the proposed next state. For example, in a second-order model, given that a two letter sequence is CH, what is the probability that the next letter of the sequence is A? From the table, it can be seen that no matter which Markov model is used, the best path is always the one for CHAT, which also happens to be the only word among all the candidate strings.

The example described above shows a methodology for recognizing words from their spellings that could be incorporated into a model for a spelling recognition system. Figure 5.3 shows a block diagram for a proposed recognition system. The system takes the input spelled speech waveform and performs as fine an acoustic analysis on it as possible. This acoustic analysis yields a phonetic transcription

lattice, which is in turn transformed into a letter lattice using the phonetic characteristics of letters as a guide. Lexical knowledge is then applied to the letter lattice to find the best path through it, and the result is an orthographic transcription that hopefully corresponds to the input spelled word.

5.3 Suggestions for Future Work

There are many ways in which this work may be extended. First of all, the acoustic study of the spelling corpus can be continued in an effort to find out more about acoustic-phonetic features particular to spelled speech. Also, ways to better resolve spelled speech acoustically can be explored.

The system described in the previous section assumes that the acoustic analysis of the waveform will result in detailed segmental classification in order to obtain a sparse letter transcription lattice. However, as was demonstrated in Chapter 4, even broad classification reduces the number of possible letter sequences due to the structural characteristics of letters. Although broad classification generally leads to a more dense letter transcription lattice than detailed classification, lexical knowledge may still be able to find the correct path through the lattice. Experimentation would indicate how detailed the segmental classification should be in order to obtain accurate orthographic transcriptions.

Work can also continue in the area of fine acoustic resolution. As discussed in Chapter 4, although the most difficult confusions could be resolved with a few acoustic parameters better than by spectrogram readers, the accuracy rates obtained were still not in the same range as those realized by the listeners. It may be possible to further improve scores by using more sophisticated features. Also, using alternate means of representing the signal, such as in the form of line formants, may provide another way of improving recognition scores.

The lexical study should also be extended because more information about lex-

ical constraints are needed. Although the statistics obtained about the frequency and existence of letter sequences are powerful and very useful to this task, they do not fully capture the rules of spelling. The inherent structure of words has not been exploited; for example, the rule that all words must contain at least one vowel letter (i.e., A, E, I, O, U or Y) has not been used.

Although substitution errors are by far the most common error made in spelling recognition, other errors, such as deletion and insertion errors, do occur. Ways for resolving these and other types of errors should also be studied.

In order to implement a spelling recognition system, information from the lexical and acoustic-phonetic knowledge sources must be combined. The optimal integration of information from these two sources may be obtained through experimentation. From the results of the recognition experiments described in Chapter 3, it can be seen that the primary source of information is acoustic-phonetic, but the proper weighting of information from the two sources is not yet known.

In addition, the relative importance of knowledge from each of the sources may vary. In some cases, a fine acoustic resolution of a spelled string is not necessary, since lexical knowledge can compensate for acoustic uncertainty. For example, in Chapter 4, attempts were made to disambiguate the four most common substitution errors made by spectrogram readers. However, not being able to distinguish between these letters may not matter if the distinction can be made using lexical information.

An analysis of the MPD was conducted to explore this hypothesis. Specifically, we looked for all minimal pairs of words that differ only in one of the four minimal pairs of confusable letters that we investigated. For example, the word BAT could be confused with the word BAG if T were confused with G. Table 5.2 shows the percent of words containing at least one of the confusable letters that would be subject to such a confusion. The table shows that an inability to resolve one of these confusions matters for only a small percentage of words containing one of the confusable letters. Therefore, we conclude that perfect acoustic resolution may not

Confusable Pair	% Confusable Words
G-T	2.3
A-E	2.9
O-L	0.6
M-N	1.5

Table 5.2: Percent of words that are confusable due to containing one of a confusable letter pair

be necessary to obtain the correct solution.

Appendix A

Summary of Letter Frequency Statistics

This appendix contains information about letter frequencies to supplement what is shown in the text of this thesis.

A.1 Equally-Weighted Words

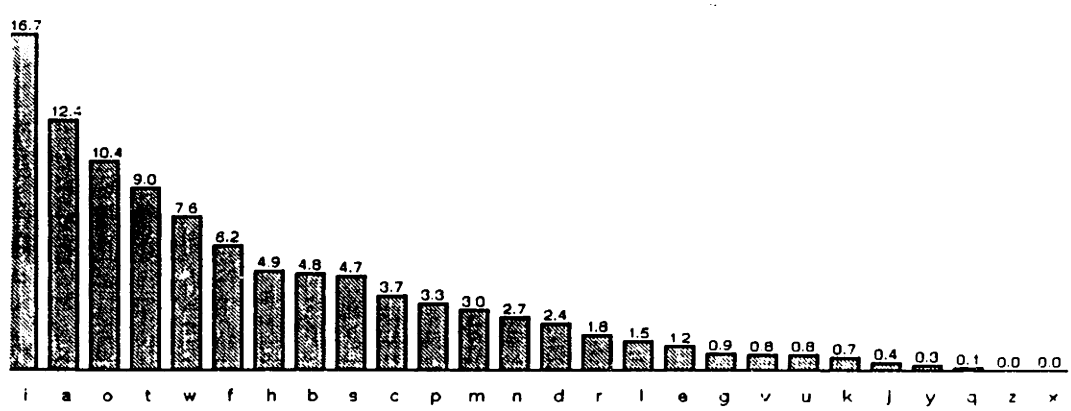


Figure A.1: Histogram of Beginning Letter Occurrences

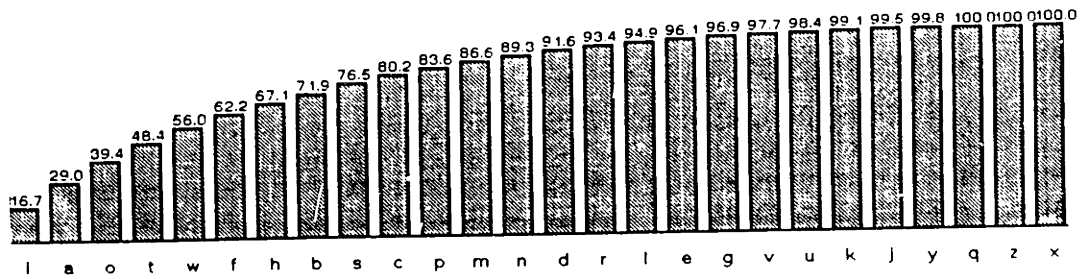


Figure A.2: Histogram of Cumulative Beginning Letter Occurrences

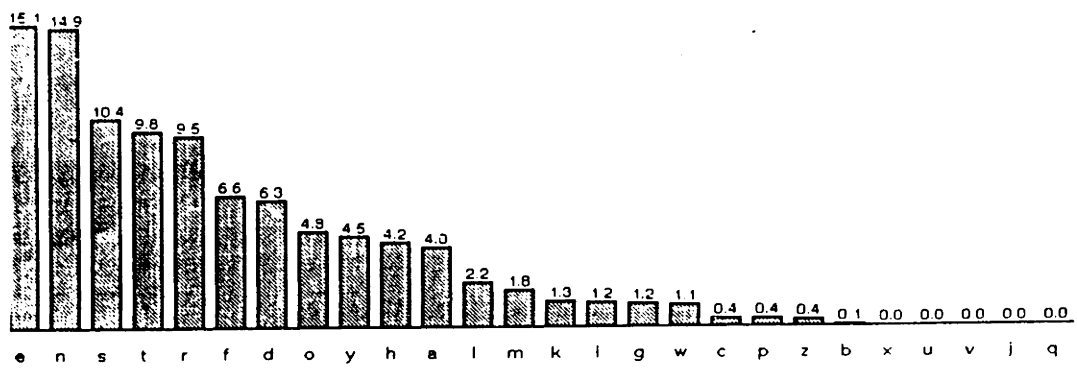


Figure A.3: Histogram of Ending Letter Occurrences

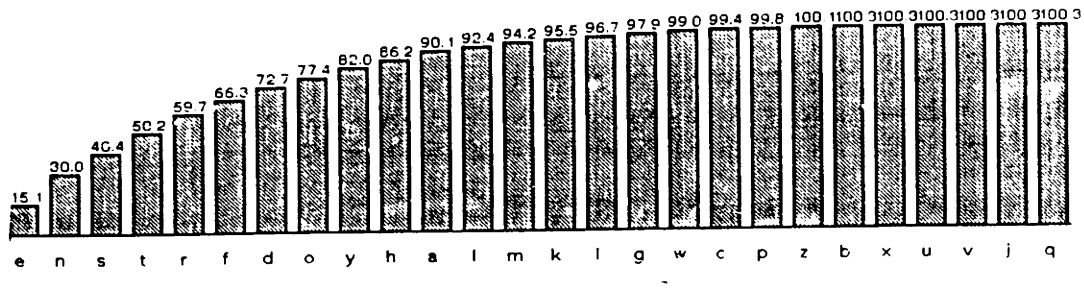


Figure A.4: Histogram of Cumulative Ending Letter Occurrences

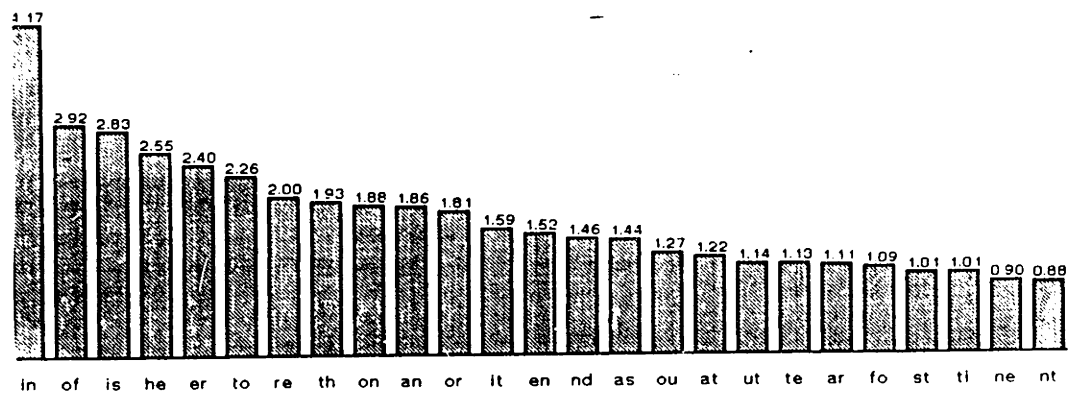


Figure A.5: Histogram of Joint Letter Occurrences

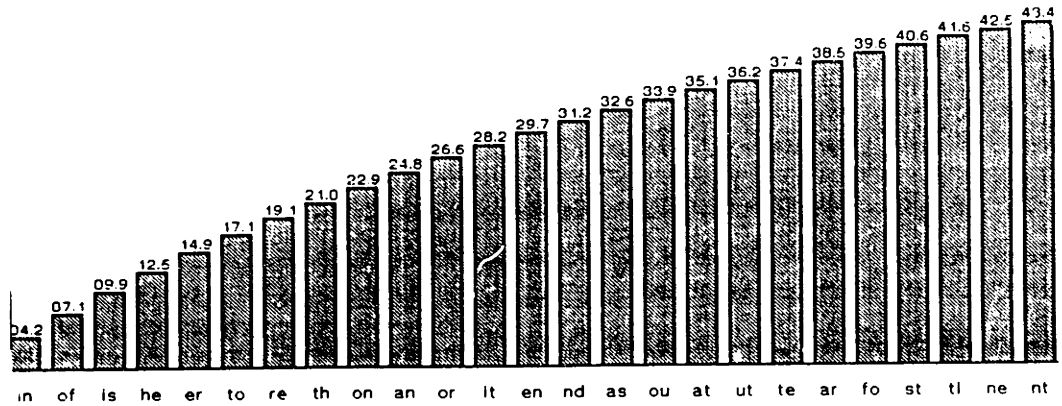


Figure A.6: Histogram of Cumulative Joint Letter Occurrences

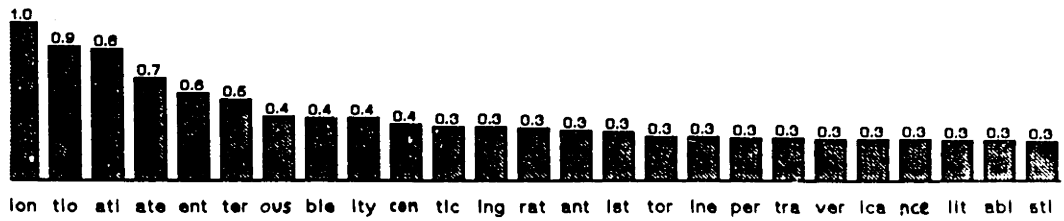


Figure A.7: Histogram of Beginning Letter Triplets Occurrences

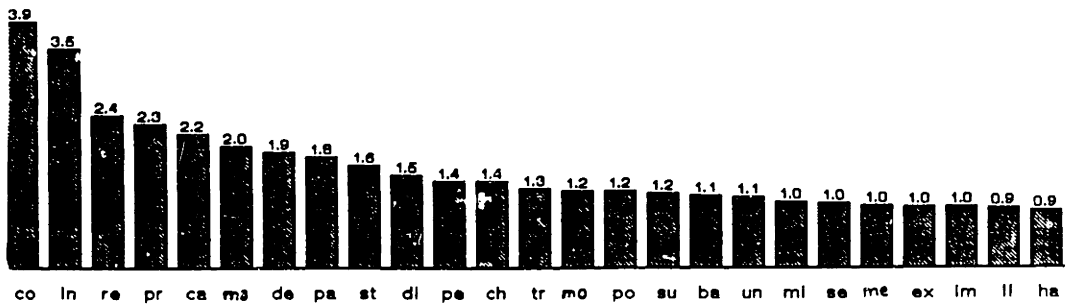


Figure A.8: Histogram of Ending Letter Triplets Occurrences

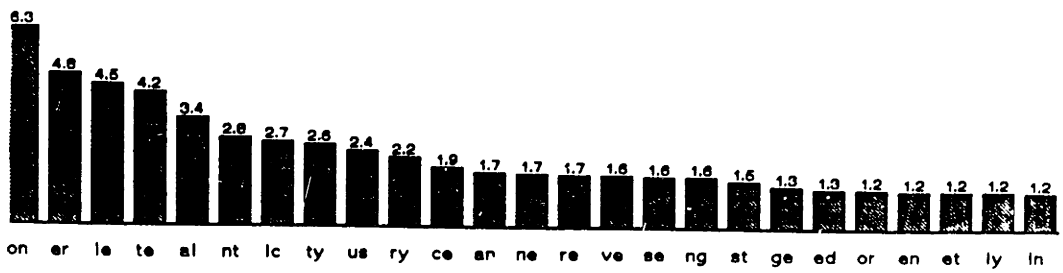


Figure A.9: Histogram of Joint Letter Triplets Occurrences

A.2 Words Weighted by Frequency of Appearance

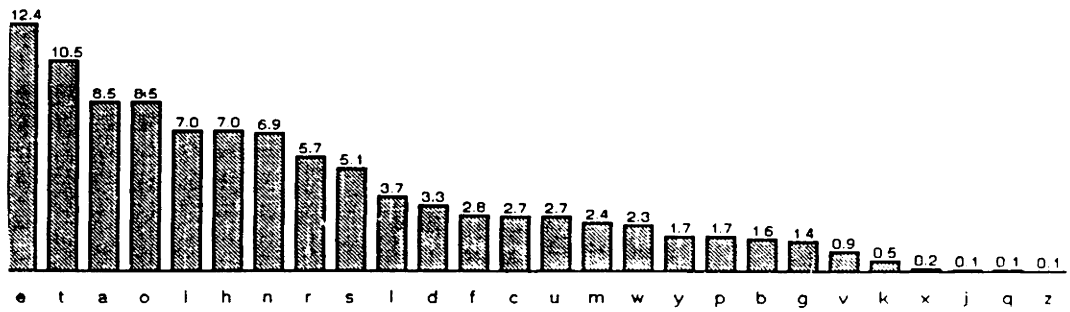


Figure A.10: Histogram of Single Letter Occurrences

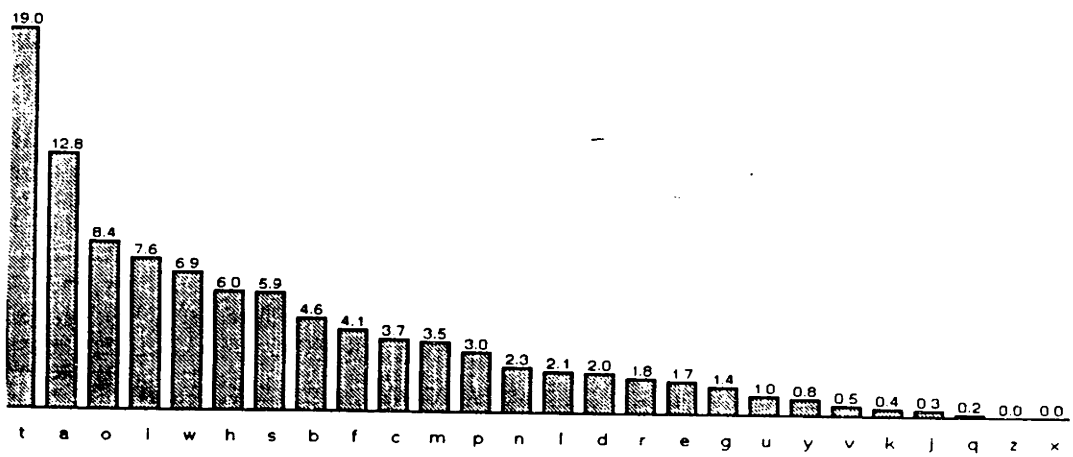


Figure A.11: Histogram of Beginning Letter Occurrences

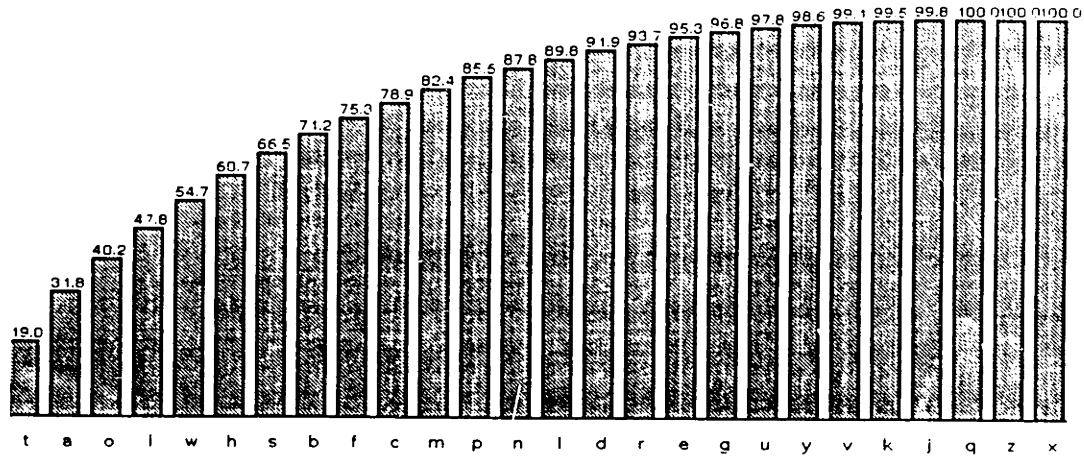


Figure A.12: Histogram of Cumulative Beginning Letter Occurrences

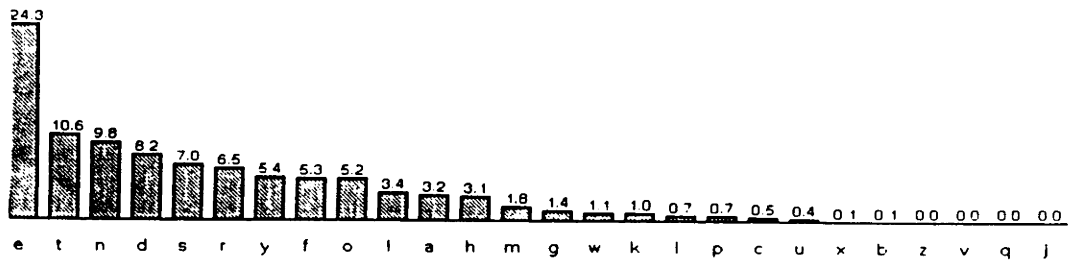


Figure A.13: Histogram of Ending Letter Occurrences

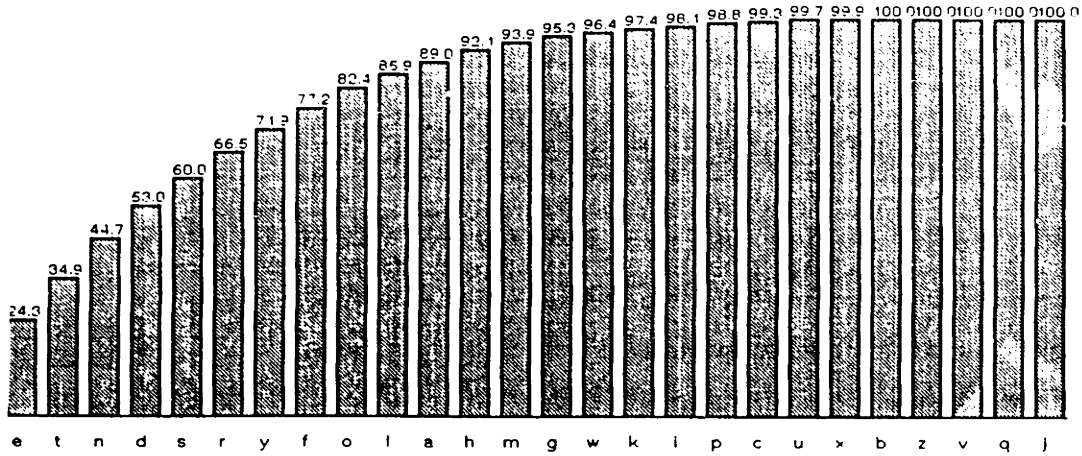


Figure A.14: Histogram of Cumulative Ending Letter Occurrences

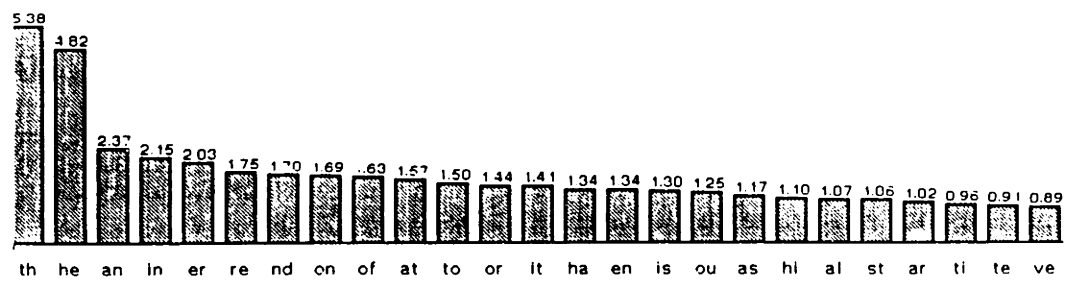


Figure A.15: Histogram of Joint Letter Occurrences

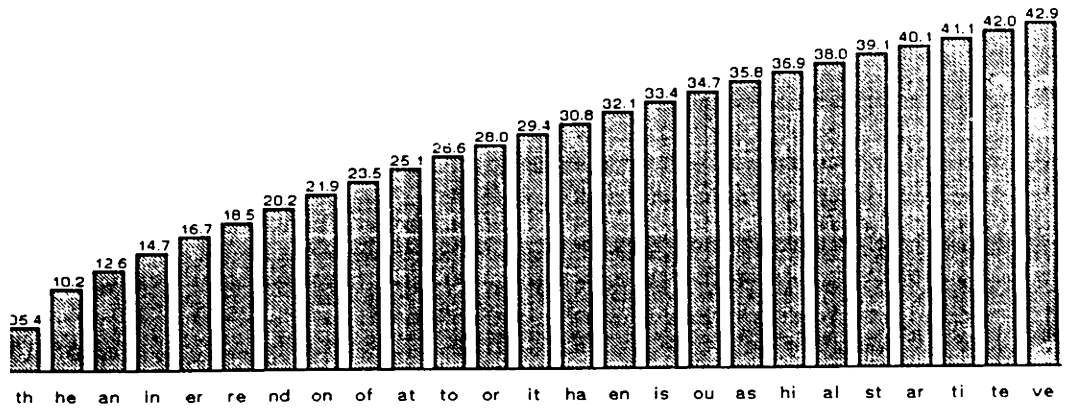


Figure A.16: Histogram of Cumulative Joint Letter Occurrences

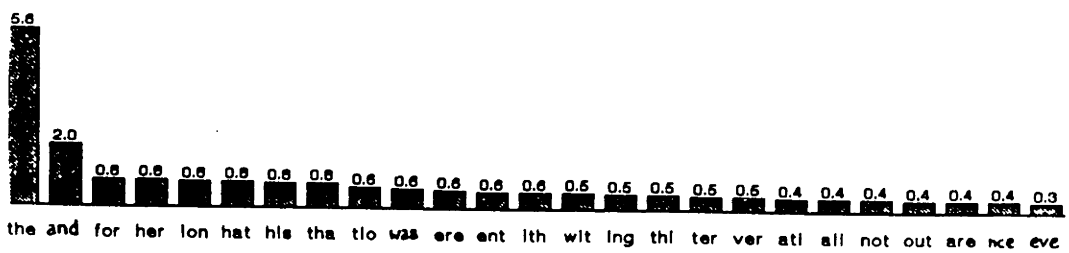


Figure A.17: Histogram of Beginning Letter Triplets Occurrences

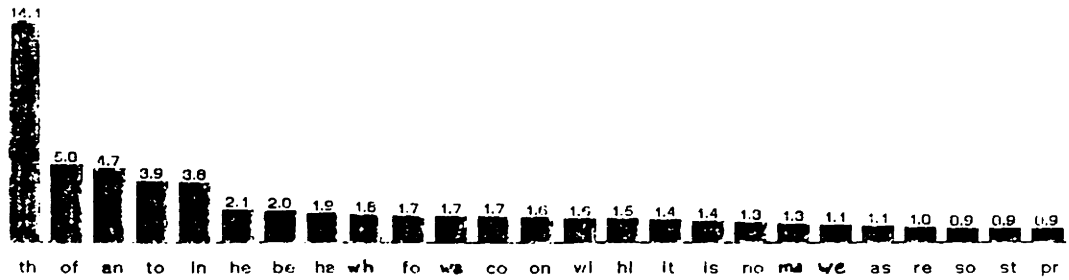


Figure A.18: Histogram of Ending Letter Triplets Occurrences

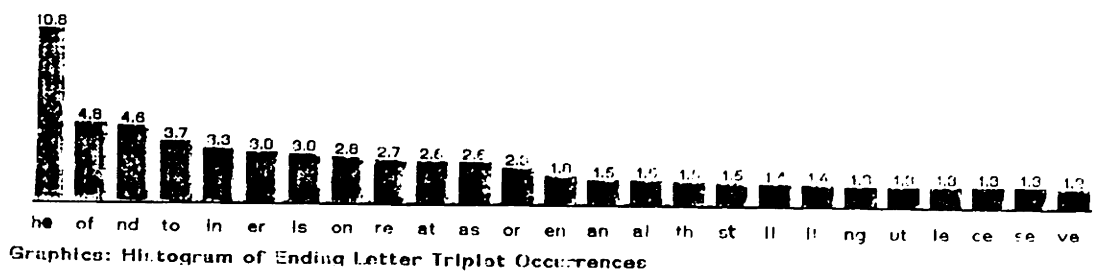


Figure A.19: Histogram of Joint Letter Triplets Occurrences

A.3 Statistics for Unweighted Words from Twenty Lexicons

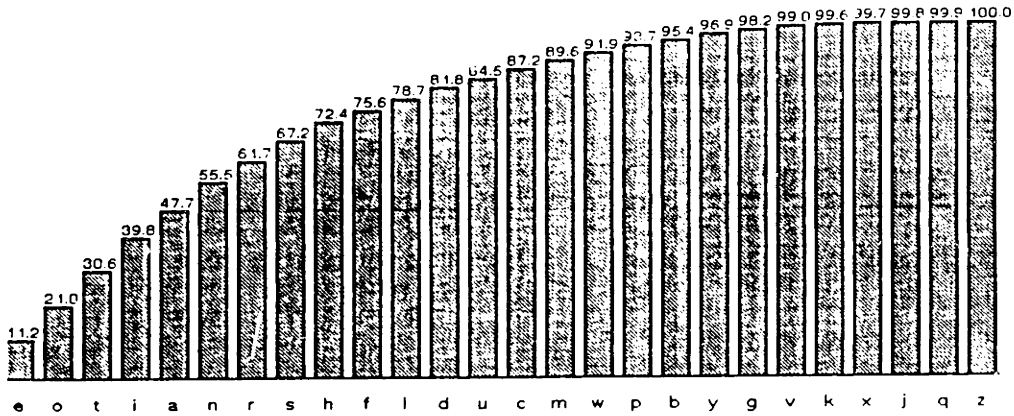


Figure A.20: Histogram of Cumulative Single Letter Occurrences

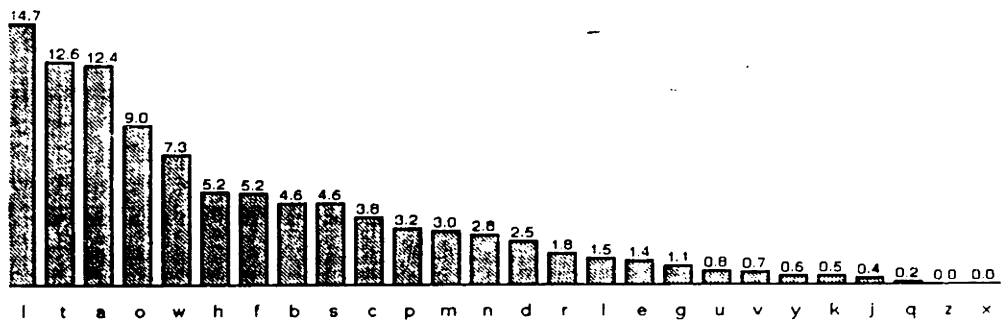


Figure A.21: Histogram of Beginning Letter Occurrences

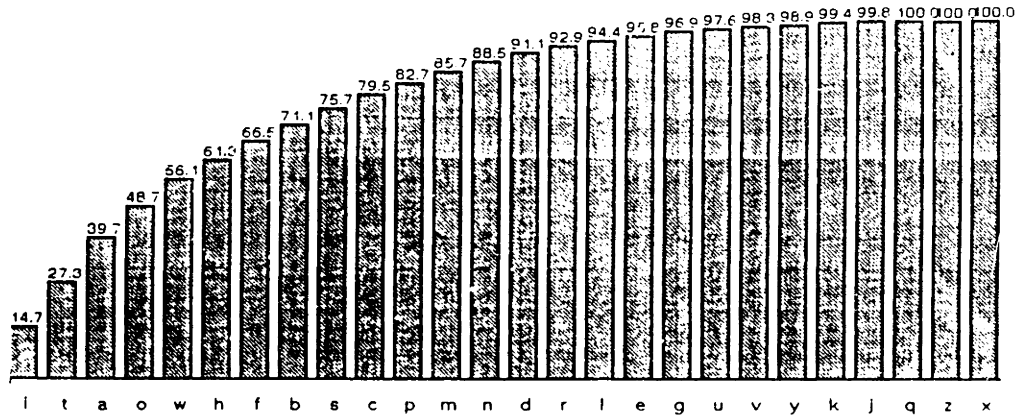


Figure A.22: Histogram of Cumulative Beginning Letter Occurrences

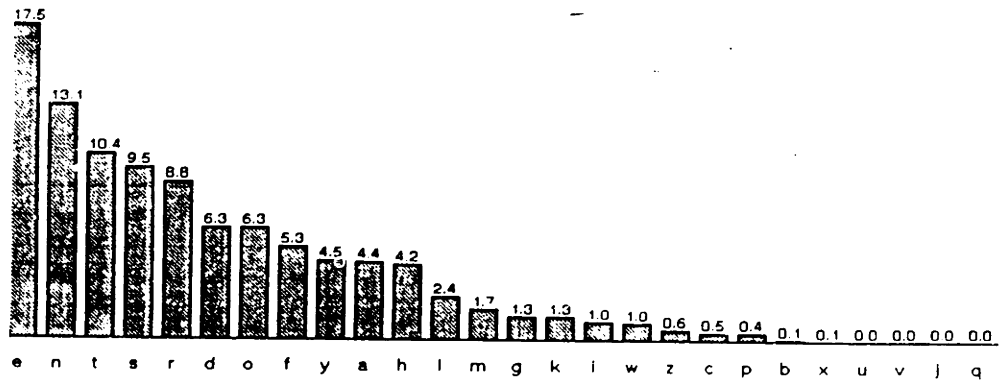


Figure A.23: Histogram of Ending Letter Occurrences

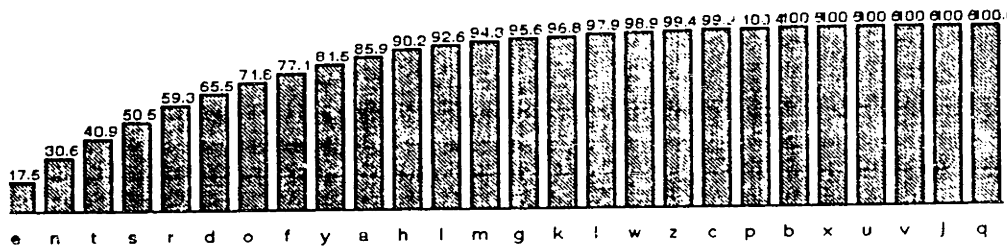


Figure A.24: Histogram of Cumulative Ending Letter Occurrences

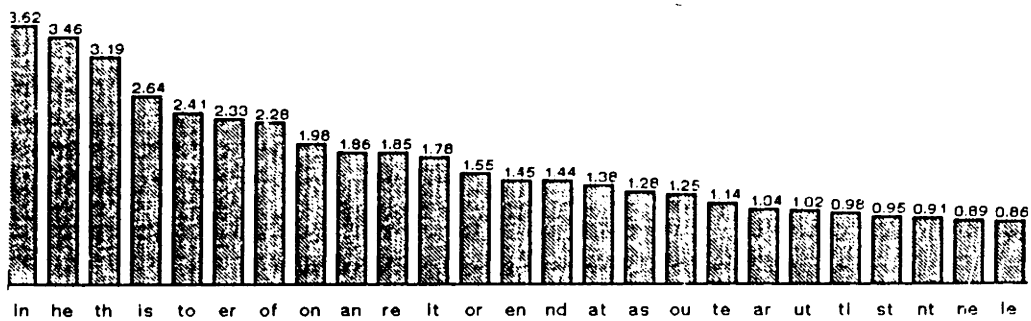


Figure A.25: Histogram of Joint Letter Occurrences

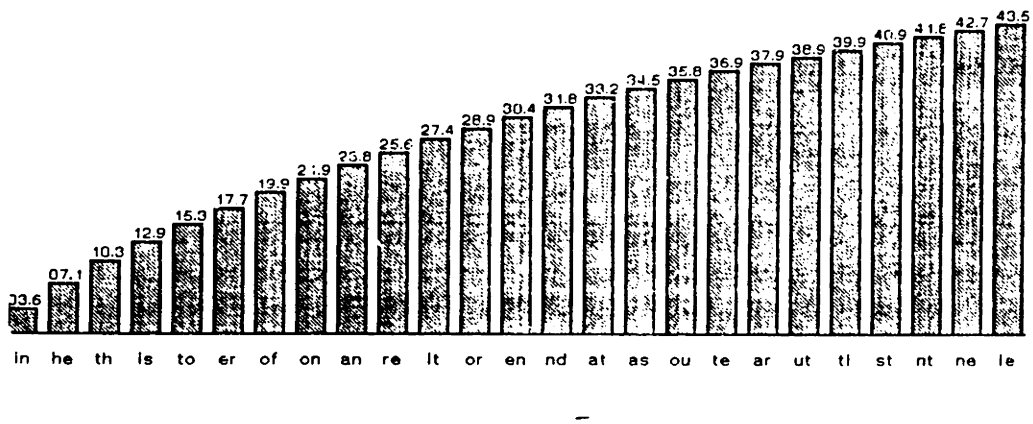


Figure A.26: Histogram of Cumulative Joint Letter Occurrences

Bibliography

[1] A. M. Aull, "Lexical Stress and Its Application in Large Vocabulary Speech Recognition," S.M. Thesis, Massachusetts Institute of Technology, 1984.

[2] F. R. Chen, "Acoustic-Phonetic Constraints in Continuous Speech Recognition: A Case Study Using the Digit Vocabulary," Ph.D. Thesis, Massachusetts Institute of Technology, 1985.

[3] R. A. Cole, R. M. Stern and M. J. Lasry, "Performing Fine Phonetic Distinctions: Templates versus Features," *Variability and Invariance in Speech Processes,* J. S. Perkell and D. H. Klatt, Eds. Hillsdale: Lawrence Erlbaum Assoc., 1985.

[4] R. Cole et al., "FEATURE: Feature-based, speaker-independent, isolated letter recognition," Technical Report, Department of Computer Science, Carnegie-Mellon University, August 1982.

[5] L. D. Erman and V. R. Lesser, "The Hearsay II Speech Understanding System: A Tutorial," *Trends in Speech Recognition,* edited by W. A. Lea, Englewood Cliffs: Prentic-Hall, Inc., 1980.

[6] J. R. Glass, "Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment," S.M. Thesis, Massachusetts Institute of Technology, 1985.

[7] D.P. Huttenlocher "Acoustic-Phonetic and Lexical Constraints in Word Recognition: Lexical Access Using Partial Phonetic Information," S.M. Thesis, Massachusetts Institute of Technology, 1984.

[8] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, pp. 67-72, 1975.

[9] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of IEEE,* vol. 64, pp. 532-556, 1976.

[10] V. W. Zue et al, "The Development of the MIT Lisp-Machine Based Speech Research Work Station, *Proc. ICASSP-86,* pp. 329-333, 1986.

[11] D. H. Klatt, "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence," *Journal of the Acoustical Society of America,* vol. 59, no. 5, pp. 1208-1221, May 1976.

[12] G. E. Kopec and M. A. Bush, "Network-Based Isolated Digit Recognition Using Vector Quantization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 4, pp. 850-867.

[13] S. E. Levinson, L. R. Rabiner, M. M. Sondhi, "An Introduction to the Application of the Theory of Probabalistic Functions of a Markov Process in Automatic Speech Recognition," *Bell Systems Technical Journal,* vol. 62, pp. 1035-1074, 1983.

[14] B. Lowerre and D. R. Reddy, "The Harpy Speech Understanding System," *Trends in Speech Recognition,* edited by W. A. Lea, Englewood Cliffs: Prentice-Hall, Inc., 1980.

[15] G. E. Peterson and I. Lehiste, "Duration of Syllabic Nuclei in English," *Journal of the Acoustical Society of America,* vol. 32, no. 6, pp. 693-703, June 1960

[16] F. Pratt, *Secret and Urgent*, Blue Ribbon Books, 1949.

[17] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Englewood Cliffs: Prentice-Hall, Inc., 1978.

[18] D. R. Reddy and V. Zue, "Recognizing continuous speech remains an elusive goal,", from "Tomorrow's Computers: The Challenges," *IEEE Spectrum,* vol. 20, no. 11, pp. 84-87, 1983.

[19] Stephanie Seneff, "Vowel Recognition Based on 'Line-Formants' Derived from an Auditory-Based Spectral Representation," to be presented at the International Congress of Phonetic Sciences, Tallinn, Estonia, USSR, August, 1987.

[20] C. E. Shannon, "Predictability and Entropy of Printed English," *Bell System Technical Journal,* vol. 30, pp. 50-64, January 1951.

[21] D. W. Shipman and V. W. Zue, "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," *Proc. ICASSP-82,* pp. 546-549, 1982.