MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 867 October, 1985

Exploiting Sequential Phonetic Constraints In Recognizing Spoken Words

Daniel P. Huttenlocher

Machine recognition of naturally spoken language requires developing more robust recognition algorithms. A recent study by Shipman and Zue suggests using partial descriptions of speech sounds to eliminate all but a handful of word candidates from a large lexicon. The current paper extends their work by investigating the power of partial phonetic descriptions for developing recognition algorithms. First, we demonstrate that sequences of *manner of articulation* classes are more reliable and provide more constraint than certain other classes. Alone these results are of limited utility, due to the high degree of variability in natural speech. This variability is not uniform however, as most modifications and deletions occur in unstressed syllables. Comparing the relative constraint provided by sounds in stressed versus unstressed syllables, we discover that the stressed syllables provide substantially more constraint. This indicates that recognition algorithms can be made more robust by exploiting the manner of articulation information in stressed syllables.

Acknowledgments. This report describes research done at the Artificial Intelligence Laboratory and the Speech Communications Group of the Research Laboratory of Electronics at the Massachusetts Institute of Technology. Support for the Artificial Intelligence Laboratory's research is provided in part by the Advanced Research Projects Agency under Office of Naval Research contract N00014-80-C-0505. Support for the Speech Communications Group's research is provided in part by the Advanced Research Projects Agency under Office of Naval Research contract N00014-82-K-0727.

The Speech Recognition Problem

Communication between humans and machines would be greatly facilitated by natural speech input-output capability. While speech output devices are approaching natural performance levels, speech input is far from natural. Current speech recognizers, both commercial and experimental, provide only restricted recognition capability. The vocabulary must be specified in advance, most systems must be trained to a particular talker, and sentences must generally be spoken as sequences of isolated words and phrases [1] [2] [3] [4]. Attempts to extend the current recognition technology to less restricted tasks have been relatively unsuccessful.

The sensitivity of existing recognition systems to specific tasks indicates that the current technology will not scale up to the problem of recognizing naturally spoken language. More robust recognition algorithms will be needed to handle the high degree of variability and noise in natural speech. The present paper takes a step towards the development of such algorithms by determining what information is important in recognition. Examination of the phonemic structure of a large English dictionary reveals that certain speech sounds provide much more constraint in differentiating words from one another than do other speech sounds. The same sounds are also highly perceptually and acoustically salient, indicating that these sounds carry much of the linguistic information in speech.

The Approach

Not all information is of equal importance in recognition. In order to develop robust recognition algorithms it is first necessary to determine what the important information is. There are two dimensions along which importance may vary. First, *certain information provides more constraint* in recognition than other information. Second, *certain information is more reliably present in the input* than other information. We will term something a *good recognition cue* if it is both highly constraining and relatively reliable.

By finding and exploiting good recognition cues we seek to maximize constraining power while minimizing sensitivity to noise and poor sensory input. This would also seem to be an underlying motivation for Marr's idea of identifying *natural constraints* [5]. Natural constraints are those which are imposed on a recognition problem by the physics of the objects being recognized, the physics of the recognition modality, or the biology of a natural

recognition system. Since natural constraints capture important properties of a recognition domain they serve the goal of being highly constraining while also being relatively insensitive to noise.

Our investigation is divided into two main parts. First we identify certain reliable properties of speech sounds (manner of articulation classes), and determine how much constraint these properties provide in differentiating words from one another. Then we identify certain reliable parts of words (stressed syllables) and determine their constraining power. However, some general robustness criteria for recognition algorithms are presented first.

Criteria for a Robust Recognition Algorithm

Several central requirements for a successful recognition algorithm can be identified. These issues have previously been raised by researchers in computer vision and object recognition [5] [6] [7], however they are of importance for recognition tasks in general.

- A recognition algorithm should degrade gracefully with increasing noise in the sensory input.
- A recognition algorithm should degrade gracefully with increasing complexity in the recognition task.
- A recognition algorithm should be able to handle partial or missing data.

There are two underlying motivations for these requirements. First, real-world recognition tasks occur in the presence of noise, poor sensors, and missing data. If a system is to be useful for such tasks its performance must degrade gracefully in poor environments. Second, human performance degrades relatively gracefully. People do not make the gross sorts of errors made by current recognizers — such as recognizing a poly-syllabic word where there is only one vowel nucleus. If a system is to be natural to humans, it must preserve the robustness of human perception.

Current speech recognizers are extremely fragile when evaluated according to these criteria. Even small changes in environmental noise, speaker characteristics, or the recognition vocabulary have a major impact on performance. For future generations of speech recognizers it is important to consider how to make the recognition algorithms more robust, or we have little hope of reaching the goal of natural speech input capability.

Phonetic Classes Might Make Good Recognition Cues

We are interested in the problem of recognizing words from a sequence of sounds. In English the inventory of these sounds is limited to forty or so *phonemes.* Each phoneme can have several different acoustic realizations, called *allophones.* This *allophonic variation* is caused both by local context and by the individual differences between speakers. For instance a /t/ in a retroflex context as in "truck" is extremely strong, making it more like the /č/ "chuck" than the /t/ in "tuck". This contextual variation depends greatly on the speaker — some speakers have very strong /t/'s which are often similar to /č/'s while others have very weak /t/'s which rarely resemble /č/'s.

Since allophonic variation is partly due to speaker characteristics, it is difficult to build phonemic recognizers which are not trained to a given speaker. Even for a given speaker, there is a high degree of variability in individual phonemes. Thus while sequences of phonemes are highly constraining – they uniquely specify words to within homophones – they are also highly variable. According to our criteria a good recognition cue should be both reliable and highly constraining, meaning that phonemes are not particularly good recognition cues.

A given phoneme can be characterized by both its *place* and *manner* of articulation. For example, the place of articulation for the phoneme /š/ (as in "ship") is *palatal*, because the sound is made by raising the articulators towards the roof of the mouth (the palate). On the other hand, the manner of articulation for /š/ is *frication*, because the sound is made by exhaling through a partial closure of the vocal tract, causing aperiodic (or fricative) noise. There are approximately a half dozen manner classes and a half dozen place classes which together can be used to define the space of English phonemes.

The manner of articulation of a phoneme refers to gross characteristics of the speech production process. Therefore manner of articulation differences are very pronounced. This is observable both in the acoustic signal and in studies of human perception of speech sounds. The acoustic characteristics of different manner classes are visually striking in spectrographic displays of speech [8]. The speech spectrogram of the word "snack" in Figure 1 illustrates this marked acoustic difference. The first segment is the *fricative* /s/, the second segment is the *nasal* /n/, the third segment is the *vocalic* /æ/, and the fourth segment is the *stop consonant* /k/. Each of these four manner classes has a characteristic appearance.

A set of perceptual studies examining the confusability of English phonemes further demonstrates the salience of manner of articulation classes. In

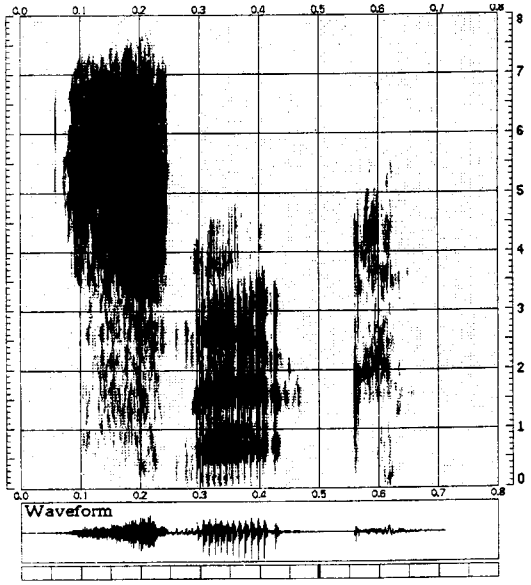

Figure 1. A spectrogram of the word "snack", illustrating the difference between the four manner of articulation classes: fricative, nasal, vocalic, and stop. A spectrogram is a three dimensional display of the speech signal where time is along the x-axis, frequency is along the y-axis, and amplitude (the z-axis) is encoded in the darkness of the display.

these studies peoples' phoneme recognition errors were examined. Almost all of the confusions were between phonemes in the same manner of articulation class [9]. There is also anecdotal evidence that manner classes are perceptually important. For example, the non-word "shpeech" is still recognizable as the word "speech", while "tpeech" is not. This may be due to the fact that /s/ and /š/ both belong to the same manner class — strong fricative, while /t/ belongs to a different class — stop consonant.

The acoustic and perceptual salience of manner of articulation phonetic classes indicates that they are reliable cues for recognition. However, it remains to be seen how much constraint these manner classes provide. Clearly by moving from a space of forty phonemes, to six classes we are losing some amount of information. The question is just how much.

Sequential Constraints Are Important

In addition to the limited inventory of speech sounds, only certain combinations of sounds may occur in a given language. For instance "vnuk" is clearly not a valid English word, because the sound sequence /vn/ is illegal. These

sequential constraints may be useful when there are unknown sounds in a sequence — a sequence beginning with /v/ cannot have an /n/ as the next sound.

Sometimes sequential constraints can be very powerful. For the partial sound sequence /kæ–t/, where /–/ denotes an unknown sound, only three of approximately forty possible sound sequences form valid English words. These are /kæst/, /kænt/ and /kæpt/, corresponding to the words "cast", "can't", and "capped". Such contextual constraints are highly local. If the same partial sound sequence /kæ–t/ is embedded in sequences of any length, /s/, /n/, and /p/ are still the only phonemes which can replace the /–/ to form valid English words. †

Even when contextual information cannot be specified in detail, sequential constraints can still be very powerful. In the above example assume further that the specific identity of the vowel is unknown. Thus we have the partial sequence /kV–t/, where /V/ denotes the presence of some unknown vowel. Despite the fact that the vowel identity is unknown, there are only five phonemes which can replace the /–/ in order to form English words. These are /r/ and /l/ as well as the three phonemes /s/, /n/ and /p/ from above.

How Powerful are Sequential Constraints?

The examples of the previous section suggest that sequential information can provide substantial constraint on the identity of unknown speech sounds. However, in these examples we assumed that the identity of at least some of the neighboring phonemes was known in detail. We have already seen that phonemes are difficult to recognize in the sensory input. Thus, even if sequential constraints are very powerful, there is little hope of reliably recognizing the necessary phonemes.

Since manner of articulation phonetic classes are acoustically relatively reliable, they make a reasonable candidate for investigating sequential phonetic constraints. Unlike a phoneme sequence, a broad phonetic sequence is already only a partial specification. Therefore, the paradigm becomes that of determining how many words in a particular lexicon match a given broad phonetic sequence. For instance, given the six manner classes — stop, vocalic, nasal, liquid or glide, strong fricative, and weak fricative — we find that the

†The two words "cactus" and "caftan" are exceptions to this. Words borrowed from other languages are often exceptions. Since the exceptions tend to be very low frequency, likelihood information could be useful in recognition.

sequence

[STOP] [VOCALIC] [STRONG-FRIC] [STOP]

matches 35 words in Webster's Pocket Dictionary of 20,000 words.

In order to perform this investigation systematically we can map each word in a large lexicon into its corresponding broad phonetic sequence, and then see how many different words map into the same sequence. That is, we can partition the lexicon into equivalence classes of words according to their broad phonetic descriptions. The smaller these equivalence classes, the more constraint is provided by the broad phonetic representation. In the limiting case, when sequences of phonemes are used rather than sequences of broad classes, the equivalence classes will almost all be of size one.†

Shipman and Zue [10] performed exactly this study using Webster's Pocket Dictionary, and mapping the words into manner of articulation sequences. In the next four sections we present their results and extend them in several ways. First, we add information about the stress pattern of a word to the representation. Second, we contrast the constraint provided by *place* of articulation classes with that provided by *manner* classes. Third, we consider using only the phonetic information in the stressed syllables of words.

Investigating Broad Phonetic Constraints

The power of broad phonetic constraints was demonstrated by a set of studies reported by Shipman and Zue [10]. These studies examined the phonemic distribution of words in the 20,000-word Merriam Webster's Pocket Dictionary. In one study the phonemes of each word were mapped into one of the six broad manner of articulation classes: vocalic, stop, nasal, liquid or glide, strong fricative, and weak fricative. For example, the word "speak", with the phoneme string /spik/, was mapped into the sequence

[STRONG-FRIC] [STOP] [VOCALIC] [STOP]

The result of this mapping is a partition of the lexicon into equivalence classes of words with the same broad phonetic class sequence.

It was found that, even at this broad phonetic level, approximately one third of the words in the 20,000-word lexicon could be uniquely specified — were in equivalence classes of size one. The average number of words in the

†The equivalence classes won't all be size one because of homophones – words which sound the same.

same equivalence class was approximately two, and the maximum was approximately 200. In other words, in the worst case this broad phonetic representation reduces the number of possible word candidates to about one percent of the 20,000-word lexicon. Shipman and Zue examined several smaller lexicons and found this to be stable for lexicons of about 2,000 or more words; for smaller lexicons the specific choice of words can make a large difference in the distribution.

The average equivalence class size measure used by Shipman and Zue is somewhat misleading because it only reflects the number of equivalence classes, and not the distribution of words across classes. A better measure is the expected value of the class size, which is the average number of words matching an arbitrarily chosen word in the lexicon. To see the difference between these two measures, consider a partitioning of 10 elements into two classes of size 5 each, versus a partitioning into two classes of sizes 1 and 9 each. In both cases the average class size is 5, whereas the expected class size is considerably smaller in the first case than in the second (5 versus 8.2).

To the extent that words cluster in a particular equivalence class, the average class size is an overly optimistic estimate of the number of words matching a given broad phonetic class sequence. Thus, we use the expected equivalence class size, denoted $E(w)$, and given by

$$\sum_{i=1}^{C} S_i \cdot L_i$$

where S_i is the size of the i-th equivalence class, L_i is the relative frequency of the i-th equivalence class - $\frac{S_i}{N}$, N is the size of the lexicon, and C is the number of equivalence classes.

The expected class size for Shipman and Zue's study is 21, approximately an order of magnitude greater than the mean class size. However, this still only represents approximately 0.1 percent of the entire lexicon. The results of this study are summarized in the first row of Table 1.

Word Frequency Effects

Partitioning the lexicon into equivalence classes implicitly gives all words equal weighting, because each word is counted once regardless of how frequently it occurs in English. However, word frequency in English is far from uniform. This means that the previous results are not particularly useful in determining how many words can be expected to match an arbitrary English word. In order to determine how word frequency affects the broad phonetic organization of

a lexicon, it is possible to weight each word in the lexicon in relation to its frequency of occurrence. Thus, the class size S_i is replaced by the frequency weighted class size F_i, given by

$$S_i \cdot \sum_{w \epsilon W_i} p(w)$$

where $p(w)$ is the probability of finding word w in a large corpus of text, and W_i is the set of words in the i-th equivalence class.

In the frequency weighted case, the expected value is the number of words which will match an arbitrarily chosen word from written English text, as opposed to an arbitrarily chosen word from the lexicon. Similarly the percent unique is the percentage of words in running text which map into equivalence classes of size one. In their original study, Shipman and Zue examined the frequency weighted lexical distribution for the words in the Pocket dictionary, using the million-word Brown Corpus of written English [11]. It was found that when word frequency is taken into account the expected equivalence class size grows somewhat. These results are summarized in the second row of Table 1.

Condition	$E(x)$	Max	% Unique
Unweighted	21	223	32%
Freq. Weighted	34	223	6%

Table 1. Equivalence class sizes obtained by mapping the words in Webster's Pocket Dictionary into manner of articulation phonetic class sequences. After Shipman and Zue.

The fact that expected class size increases when words are weighted by their frequency of occurrence means that more common English words tend to fall in slightly larger than average equivalence classes. This is contrary to expectation — common words should be more easily distinguishable from one another, not less. However word frequency is confounded with the fact that common words tend to be shorter than uncommon words, and therefore contain less phonetic information. Thus, word length may be contributing to the observed result that more common words fall in slightly larger equivalence classes. It may also be that the broad phonetic representation fails to capture certain information which is important in differentiating common words from one another. In the next section we investigate syllabic stress patterns as another potential source of constraint in lexical access.

Stress as an Additional Source of Constraint

One salient characteristic of isolated words which we have not utilized thus far is *lexical stress*, the stress pattern of the syllables in a word. Lexical stress appears to be important in distinguishing certain words from one another. For example, the words "campus" and "compose" both map into the broad phonetic sequence

[STOP] [VOCALIC] [NASAL] [STOP] [VOCALIC] [STRONG-FRIC]

However, "campus" is stressed on the first syllable whereas "compose" is stressed on the second syllable. This alone is enough to easily distinguish these words from one another. Lexical stress patterns can be successfully extracted from the speech signal. A system for identifying stress patterns in isolated words has recently been implemented [12]. The system performs 87% correct classification into three stress levels, and 97% correct classification into two stress levels.

In order to investigate the constraint imposed by lexical stress patterns, a lexicon study was run where stress information was added to the representation. Each word was encoded according to its broad phonetic classification and its syllabic stress pattern. In this scheme, a syllable is classified as being either stressed - [S], or unstressed - [U]. Thus the word "piston", with the phonetic string /pIs-tIn/, would be represented as

[STOP] [VOCALIC] [STRONG-FRIC] [STOP] [VOCALIC] [NASAL]+[S] [U]

There were two experimental conditions. In the first condition words were not weighted according to their frequency of occurrence. These results are presented in the first row of Table 2, and can be compared with those of Shipman and Zue in the first row of Table 1. The results of the two studies are quite similar, indicating that adding stress information provides some, but not much, additional constraint.

In the second condition, the frequency weighted class size was used. These results are given in the second row of Table 2, and can be compared with the second row of Table 1. In this condition, the lexical stress information provides some additional constraint. In particular, the stress pattern substantially increases the percentage of the lexicon which is uniquely specifiable — from 6 to 25 percent.

Since stress information plays a larger role when word frequency is taken into account, this indicates that stress is important in differentiating certain common words from one another. Perhaps subsequent psychophysical investigation can test whether stress is important in human perception of common words. In later sections we will return to the role of stress in recognition, when

Condition	$E(x)$	Max	% Unique
Unweighted	18	209	39%
Freq. Weighted	28	209	25%

Table 2. Equivalence class sizes obtained by adding stress information to the manner of articulation phonetic class sequences.

we examine the interaction between stress information and the variability of speech sounds.

Place of Articulation Classes

Above we saw that manner of articulation differences are highly salient, and are based on gross characteristics of the speech production process. Place of articulation differences on the other hand, are much more subtle. For example the difference between the *palatal* fricative /š/ and the *dental* fricative /s/ is a slight lowering of frequency. This difference is illustrated by the spectrograms of the words "shoe" and "sue" in Figure 2. In addition to being relatively subtle, place differences are highly variable across different speakers and phonemic contexts. One speaker's /š/ can be similar to another speaker's /s/; and the /š/ in "she" is very similar to the /s/ in "sue".

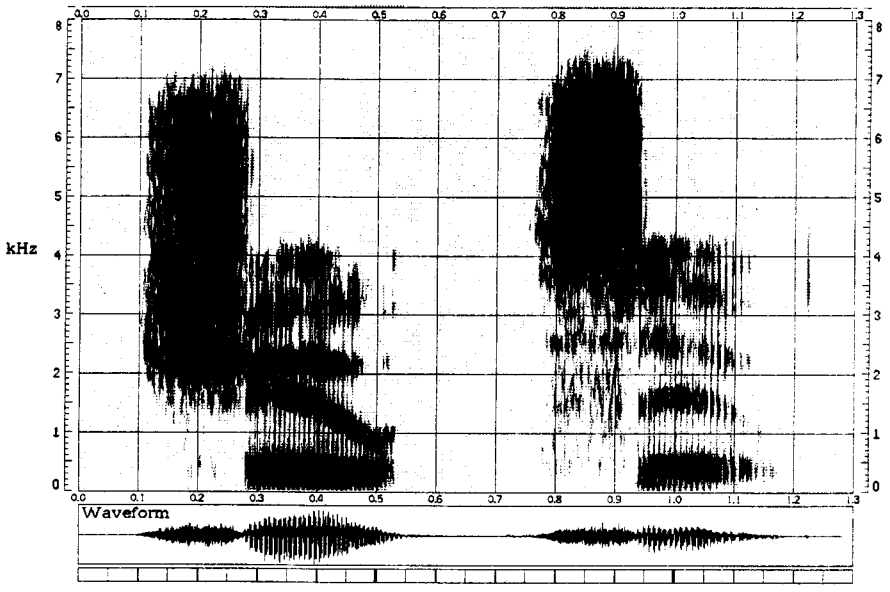

Figure 2. Spectrograms of the words "shoe" and "sue" illustrating the frequency difference between /s/ and /š/.

Since the differences between place classes are less salient than the differences between manner classes, it is reasonable to ask how much constraint is provided by place versus manner class sequences. We ran a lexicon study to investigate this, where the words in the Pocket dictionary were mapped into one of the six place classes: vocalic, palatal, labial, velar, dental, and glottal. For example, the word "speak", with the phoneme string /spik/, was represented by the pattern

[DENTAL] [LABIAL] [VOCALIC] [VELAR]

Table 3 summarizes the results of this experiment. In the first row the words are not weighted according to frequency of occurrence, and in the second row they are.

Condition	$E(x)$	Max	% Unique
Unweighted	41	336	21%
Freq. Weighted	90	336	3%

Table 3. Equivalence class sizes obtained using place of articulation phonetic class sequences.

Comparing Table 3 with Table 1 we see that place of articulation class sequences provide substantially less constraint than manner of articulation class sequences. Thus manner information is both more salient and more highly constraining than place information. This means that manner information is a better recognition cue than place information. Place of articulation does provide additional constraint — for instance place differences are all that distinguish between the sounds /p/, /t/ and /k/, it is just a less powerful cue.

Lexical Stress and Variability

The above results demonstrate that a broad phonetic classification of speech sounds can in principle be used to generate a small number of word candidates from a large lexicon. However, the acoustic realization of words and phonemes can be so variable that phonemes and syllables are deleted altogether. A multisyllabic word such as "international" can have many different realizations, some of which are illustrated in Figure 3. As can be seen from the Figure,

not only can phonemes be deleted, but some pronunciations of a word may have a different number of syllables than others.

The broad phonetic representation we have been using cannot handle the deletion or insertion of a phoneme or syllable, because such deletions affect the corresponding broad class sequence. Traditionally the problem of deletion and insertion has been solved by expanding the lexicon via phonological rules, to include various possible pronunciations of each word [13]. However this approach is an ad hoc means of accounting for the observed variability in pronunciation, rather than a general approach to modeling variability in speech. Rather than trying to explicitly model the variability, our approach is to identify the relatively invariant properties of a word.

The idea of identifying the relatively invariant portions of a word is to evaluate whether these portions of a word are also more highly constraining, and hence better recognition cues. If this were the case, it would parallel our earlier finding that the more invariant properties of phonemes also provide more constraint in recognition. In the following sections we investigate this hypothesis.

Exploiting Lexical Stress

In English, the sounds in unstressed syllables are more variable than those in stressed syllables. For instance the variations in the pronunciation of "international" shown in Figure 3 all occur in the unstressed syllables. Perceptual results have also shown that the acoustic cues for phonemes in stressed syllables are more reliable than those in unstressed syllables [14]. Since the information in stressed syllables appears more salient, a lexicon study was run comparing the importance of phonetic information in stressed versus unstressed syllables.

In this experiment there were two conditions. The first condition preserved the broad phonetic sequence in the stressed syllables, while the second preserved the broad phonetic sequence in the unstressed syllables. For example, in the first condition the word "piston", with the phoneme string /pIs-tIn/ and the stress on the first syllable, would be represented by the pattern

[STOP] [VOCALIC] [STRONG-FRIC] [*]

where [*] marks the missing unstressed syllable. In the second condition the same word would be represented by the pattern

[*] [STOP] [VOCALIC] [NASAL]

where [*] marks the missing stressed syllable.

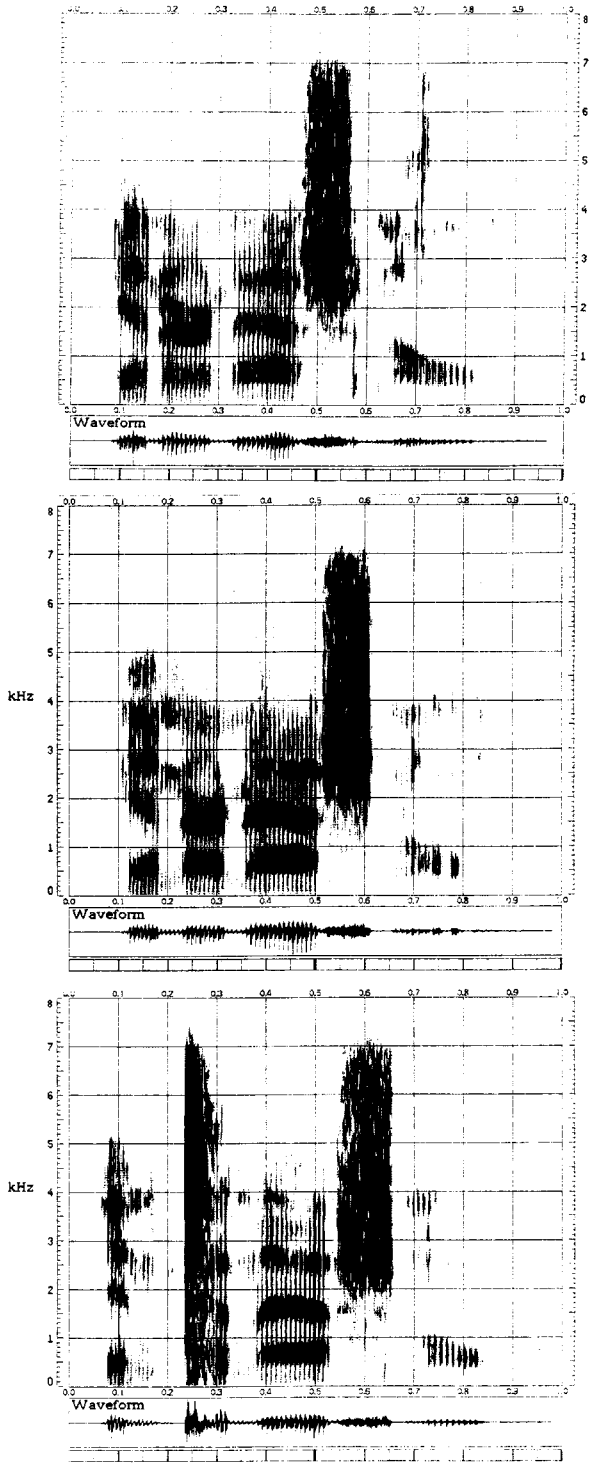

Figure 3. Different pronunciations of the word "international", illustrating that variability occurs primarily in the unstressed syllables. The top two pronunciations both have deleted /t/'s, the first with 5 syllables and the second with 4. The bottom pronunciation has an aspirated /t/ and 4 syllables.

Mapping the stressed or unstressed syllables of a word into a placeholder symbol preserves the stress pattern because the [*] corresponds to either a stressed or an unstressed syllable depending on the condition. An equivalent representation which makes this explicit combines the partial broad phonetic sequence with the syllabic stress pattern. Thus, in the first condition the word "piston" could equivalently be represented as

[STOP] [VOCALIC] [STRONG-FRIC] + [S] [U]

where [S] and [U] correspond to stressed and unstressed syllables, respectively.

The results of this experiment are given in the first three rows of Table 4. The first row is the same as the second row of Table 2, where the broad class sequence from the entire word is augmented with the stress pattern. The second and third rows show the cases where the broad class sequence is derived only from the stressed or the unstressed syllables, respectively. In all cases, the words are weighted according to their frequency in the Brown Corpus.

Condition	$E(x)$	Max
Whole Word	28	223
Stressed Only	62	260
Unstressed Only	2052	3703
Unstr. Only (Poly)	321	1725

Table 4. Equivalence class sizes obtained when the manner of articulation phonetic classes in either only the stressed or the unstressed syllables are used.

We see from the second row of the Table that the phonetic class information in the stressed syllables alone still provides substantial constraint. In sharp contrast, the third row shows that the phonetic class information in the unstressed syllables alone provides almost no constraint. However in this latter condition there was one very large equivalence class of 3703 words. This class corresponds to all the monsyllabic words in the lexicon, which simply map to the stressed syllable marker [S]. All the phonetic information for these words has been ignored because it is in a stressed syllable.

To get a better picture of the difference between the importance of stressed and unstressed syllables another experimental condition was run. In this condition only the multi-syllabic words in the Pocket dictionary were

used. Again only the phonetic class sequence in the unstressed syllables was preserved. The results are presented in the last row of Table 4. They are similar to, although somewhat less dramatic than, the results for the whole lexicon given in the third row of the Table.

Phonemes in Stressed Syllables Are Important

The results of the previous section demonstrate that the broad phonetic information in stressed syllables provides much more lexical constraint than that in unstressed syllables. This is true of phonemes as well as broad manner classes. Table 5 presents the results of a study where the actual phonemes in either the stressed or the unstressed syllables of a word were preserved, together with the lexical stress pattern. The study was run using only those words in the Pocket dictionary which contain at least one unstressed syllable. The results are weighted by word frequency.

Condition	$E(x)$	Max
Stressed Only	6	44
Unstressed Only	34	241

Table 5. Equivalence class sizes obtained when only those phonemes in the stressed or the unstressed syllables are used.

It should be noted that the difference in importance between the sounds in stressed versus unstressed syllables is not simply due to the number of phonemes which occur in stressed versus unstressed syllables. For the entire (frequency weighted) lexicon, there are approximately 1.5 times as many phonemes in stressed as in unstressed syllables, whereas the magnitude of the effect is much larger than this. In addition, when only the multi-syllabic words in the lexicon are considered, there are almost equal numbers of phonemes in stressed and unstressed syllables.

The expected class size reflects two properties of a partition — the number of equivalence classes, and the distribution of words across those classes. Best use of a given lexical partitioning (set of equivalence classes) is made when there is a uniform distribution of words across classes. In the previous experiment the distribution of words across classes was much more uniform in the stressed condition than in the unstressed condition. This means that the space of unstressed syllables is not as well utilized, strongly supporting the claim that unstressed syllables carry less information than stressed syllables.

The phonetic information in unstressed syllables is both less reliable and provides less constraint than the phonetic information in stressed syllables. However, as was the case with place of articulation information, the unstressed syllables still provide additional constraint in recognition. This suggests that the information in unstressed syllables should be used primarily in verification.

Ambisyllabic Phonemes

In many English words, it is not clear to which syllable certain phonemes belong. For example, in the word "mission" /mIš-In/, the /š/ can belong to either the first or the second syllable. Such phonemes are called *ambisyllabic*, because they can belong to either of two neighboring syllables [15]. In the studies of previous sections, phonemes were assigned to syllables using a maximal onset algorithm [16]. Ambisyllabic phonemes were not handled specially. Therefore, it is possible that ambisyllabic phonemes were assigned to stressed syllables, and this contributed to the relative importance of the stressed syllables. In order to investigate this possibility, another lexicon study was performed using a version of the Pocket lexicon where ambisyllabic phonemes had been marked.

Condition	$E(x)$	Max
Stressed Only	7	52
Unstressed Only	31	220

Table 6. Equivalence class sizes for stressed versus unstressed syllables when the ambisyllabic phonemes are moved into the unstressed syllables.

In this study ambisyllabic phonemes were moved into the unstressed syllable if either syllable was unstressed. For example in /mIš-In/ the /š/ was moved into the unstressed syllable, producing /mI-šIn/. As in the previous study the actual phonemes in the stressed or unstressed syllables were used for partitioning the lexicon. The results are summarized in Table 6. If the ambisyllabic phonemes were responsible for the relative importance of the stressed syllables, then moving them into the unstressed syllables should have produced substantially different results. However, comparing Table 6 and Table 5, we see that moving the ambisyllabic phonemes into the unstressed syllables had almost no effect. Therefore the assignment of ambisyllabic phonemes did not contribute substantially to relative importance of the stressed syllables.

Implications for Recognition Algorithms

Our computational investigation of the Pocket dictionary has demonstrated that certain speech sounds are both more reliable and more important in recognition than other speech sounds. These results suggest using a partial representation of words based on the manner of articulation information in stressed syllables for recognition. Since such a representation does not necessarily uniquely specify a given word or syllable, more detailed analysis will sometimes be needed in order to determine what word was uttered.

We consider two possible control structures for taking advantage of partial information in recognition. These are a *hypothesize and test* strategy, and a *coarse to fine* strategy. In the hypothesize and test approach a broad classification of the speech signal is used to hypothesize words from the lexicon, and then more detailed analysis is used to discriminate among the word candidates. In the coarse to fine approach a partial classification of the speech signal is performed, and then more detailed analysis is used to recognize specific phonemes, using the broad phonetic context.

Each strategy has its relative advantages and disadvantages. The next two sections consider some of the issues involved.

Hypothesize and Test

The hypothesize and test model, where partial information is used to hypothesize words from the lexicon, consists of 3 stages. First, broad phonetic sequence and stress information is extracted from the acoustic signal. Second, the partial phonetic sequence in the stressed syllables is used to access words from the lexicon. Third, if the lexical equivalence class contains more than one word, more detailed phonetic analysis is performed.

Since lexical access matches the input sound sequence against the sound sequences in the lexicon, it embodies more constraint than just sequential phonetic information. If a given input does not match, it can either be because the sequence violates the sequential constraints of the language, or because the underlying word is not in the lexicon. For instance the sequence

[NASAL] [VOCAL] [STOP] [STOP] [VOCAL] [STRONG-FRIC]+[S] [U] [S] [U]

corresponding to the word "madagascar" does not match against the Pocket dictionary, even though it does not violate the sequential phonetic constraints of English. Thus one potential disadvantage of the hypothesize and test strat-

egy is that only those phonetic sequences occurring in a given lexicon can be recognized.

One potential advantage of the hypothesize and test strategy is that the set of word hypotheses can be used in phonetic verification. For instance it can be determined which phonemes are most important in differentiating among the word candidates. In the extreme case, when all the words in a particular class have the same phoneme in the same position, there is no need to do further verification of that phoneme to distinguish among the words. In general, if some of the words in a given equivalence class share the same phoneme in the same position, an importance-based ordering can be imposed on more detailed analyses.

To determine how much constraint is provided by the word candidates, the phonetic makeup of the words in each equivalence class was examined. This experiment used manner of articulation phonetic information to form the equivalence classes — as in the original Shipman and Zue study. Recall that 32 percent of these equivalence classes contained only one word, meaning that no further discrimination is necessary. Of the remaining equivalence classes approximately 40 percent have words sharing the same phoneme in the same position. These results suggest that having a set of word candidates based on partial information can provide substantial constraint in detailed phonetic recognition.

Coarse to Fine

In the coarse to fine strategy, first a broad phonetic segmentation is done. Then the broad phonetic information is used to provide a context for more fine-grained analysis. Finally, lexical access is done using the detailed phonetic sequence. The major disadvantage of this approach is that it doesn't exploit sequential phonetic constraints until after detailed phonetic analysis is performed.

In order to use sequential phonetic constraints at the broad phonetic recognition stage, these constraints must be decoupled from the lexical representation. Below we investigate explicitly representing the sequential phonetic constraints of English in terms of allowable n-tuples of broad phonetic classes. To the extent that this representation is independent of any particular lexicon, it can be said to capture general sequential properties of English.

Sequential phonetic constraints are relatively local, extending over at most three consecutive sounds in English. For example, there are constraints

such that English has the sequences /spl/ and /spr/, but not /spt/. At a broad phonetic level this rule can be characterized as "[STRONG-FRIC]\-[STOP]\-[LIQUID] is allowable but [STRONG-FRIC]\-[STOP]\-[STOP] is not". The locality of such rules implies that a second or third order characterization of legal sound sequences should be sufficient for capturing sequential phonetic constraints.

One way of discovering the allowable sequences is to observe all the n-th order phenomena occurring in a large body of phonetic sequences. These observations can be used to construct a finite state model of broad phonetic constraints. The states of the model are n-tuples of broad phonetic classes, and the transitions are single broad classes. A transition from state $(x_1, x_2, ..., x_n)$ to state $(x_2, ..., x_n, x_t)$ occurs on input x_t, where the x_i are broad phonetic classes.

For a broad phonetic scheme such as the one we have been using, constructing these models is a relatively tractable problem because of the small number of symbols. A third order characterization of our 6 class system has only 216 possible states. For a more detailed representational scheme, with 40 or 50 labels the number of possible states rapidly becomes intractable.

A given model is formed by using the broad phonetic class sequences in a large lexicon as the initial observations. For example the lexicon consisting of the one word "cast", with the phoneme string /kæst/ and the broad phonetic sequence

[STOP] [VOCALIC] [STRONG-FRIC] [STOP]

would generate a second order model with three states and two transitions. However this model does not capture the legal sequences at the beginnings and ends of words. Therefore we make use of two additional classes [BEG] and [END] which mark before and after a word. Using these two additional classes, the model shown in Figure 4 is obtained for the one-word lexicon, "cast".

To determine the extent to which broad phonetic sequence constraints can be represented independent of a given lexicon, we compared second and third order models for two lexicons – the Pocket Dictionary and Lorge and Thorndike's 3500 most frequent English words. The same six manner of articulation classes used in the lexicon studies were used for generating the models. The number of states and transitions for each model are presented in Table 7. The second order model of the 3500 most frequent English words contains nearly all the broad phonetic sequences found in the 20,000 word Pocket dictionary. For the third order model, the 3500 word lexicon still contains most of the broad phonetic sequences found in the larger lexicon.

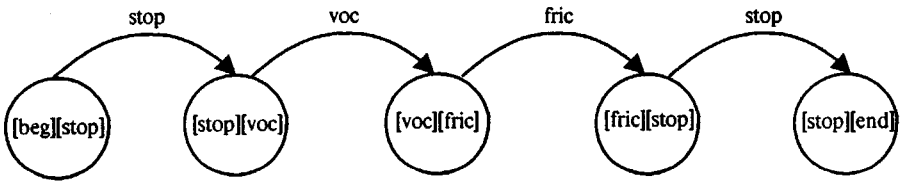

Figure 4. Second order model of a one-word lexicon.

Lexicon Size	Model Order	States	Trans
3,500	Second	51	163
20,000	Second	52	186
3,500	Third	165	528
20,000	Third	188	677

Table 7. Number of states and transitions in second and third order models of the broad phonetic sequences in the 20,000 and 3,500 word lexicons.

Another way of determining the extent to which sequential constraints can be represented independent of a particular lexicon is to use a model of one lexicon as a recognizer for another lexicon. To the extent that a model of one lexicon recognizes the sequences in other lexicons, it has presumably captured general properties of English sound sequences. When models of the 3500 word lexicon were used as recognizers for the broad phonetic sequences in the 20,000 word lexicon, the second order model recognized 99.3% and the third order model recognized 95.9% of the sequences. This strongly supports the fact that the models are independent of a given lexicon.

Explicitly representing broad phonetic constraints independent of a given lexicon allows a system to recognize most legal broad phonetic sequences in the

language, whereas hypothesize and test is limited by the fixed set of sequences in a given lexicon. This makes the coarse to fine model more attractive for recognition. On the other hand, lexical access can provide additional constraint in verification. Given the tradeoffs between the two approaches, the question becomes one of *how much* verification to do before lexical access. The more verification which is done before lexical access, the more general the algorithm in terms of the sequences it can potentially recognize, and the less it can take advantage of the specific word hypotheses.

Extending the Model: Continuous Speech

The results of the previous sections demonstrate that a partial phonetic representation can be very powerful for recognizing words. However, in continuous speech individual words are not delineated in the sensory input. In order to recognize words from continuous speech, potential word boundaries must be located so that sound sequences in the input may be matched against words in the lexicon. One straightforward approach is to hypothesize the beginning of a new word for each successive sound in the input. However the combinatorics of this approach are prohibitive, because for each sound in the input it is necessary to hypothesize words of all possible lengths starting at that point.

Certain sound sequences occur only at word boundaries. Therefore, it has been proposed that word boundaries can be identified by exploiting sequential phonetic constraints [16]. However, the useful phonetic cues for identifying word boundaries are extremely detailed — at the level of specific phonemes and allophones. Since detailed phonetic information is difficult to extract reliably from the sensory input, this approach is probably more useful for disambiguating word boundary hypotheses than for identifying potential word boundaries.

Even at a broad phonetic level, the cues to potential syllable boundaries appear to be quite strong. For instance there are only 14 possible manner of articulation sequences which correspond to syllable onsets (out of more than a hundred possibilities). This suggests that partial information may also be of high utility in recognizing continuous speech.

Summary

Not all information is of equal importance in recognition. We have demonstrated – for both phonetic and syllabic sized units – that the more reliable information is also more important in differentiating among the words in a large lexicon. Specifically, we have seen that sequences of *manner of articulation* classes are both more reliable and provide more constraint than *place of articulation* classes. Furthermore, the stressed syllables of a word are both more reliable and more highly constraining than the unstressed syllables. These results indicate that more robust recognition algorithms may be developed by exploiting the more important information in the speech signal.

Since the more reliable information in the speech signal also provides more constraint in recognition, any speech recognizer which treats the signal uniformly will suffer unnecessarily high error rates. We suggest that these results be used to develop recognizers based on partial representations of speech sounds. Another approach is to use statistical classifiers to implicitly capture differences in importance. The fact that statistically based systems, such as the one developed at IBM [1], capture the relative importance of different speech sounds probably contributes to their high performance compared with other recognizers.

While we have demonstrated that sequential constraints at the broad phonetic level are very powerful, there are also strong constraints at the acoustic and detailed phonetic levels. For instance, expert human spectrogram readers can achieve 85% phonetic labelling accuracy for syntactically and semantically anomalous sentences [17]. This performance is substantially better than that of any automatic phonetic or allophonic recognizer. Identifying these constraints may be the key to performing detailed phonetic recognition, in order to differentiate among word candidates.

Acoustic, phonetic, and lexical constraints are particularly useful because they apply early in the recognition process. In fact, such early constraints are probably *necessary* for accurate recognition. Template matching and clustering systems which do not use such constraints are very sensitive to noise, phonetic context and speaker characteristics. Higher level constraints such as syntax and semantics, while clearly important in recognition, cannot in general make up for poor phonetic level recognition performance. This was made painfully evident by the need in Hearsay-II to impose highly artificial task constraints, in order to obtain passable recognition performance given a poor phonetic front-end [18].

Acknowledgments

I would like to thank Victor Zue for his advice and guidance in supervising this work. This paper has also benefited from comments by and discussions with Phil Agre, Eric Grimson, Gary Kopec, Lori Lamel, Dick Lyon, and Jerry Roylance.

References

[1] Bahl, L.R., Cole, A.G., Jelinek, F., Mercer, R.L., Nadas, A., Nahamoo, D. and Picheny, M.A. "Recognition of Isolated Word Sentences from a 5000-Word Vocabulary Office Correspondence Task", *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, 1983.

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

[3] Doddington, G.D., and Schalk, T.B. "Speech Recognition: Turning Theory into Practice", *IEEE Spectrum*, vol. 18, no. 9, pp. 26-32, 1981.

[4] Rabiner, L.R. and Myers, C.S. "Connected Digit Recognition Using a Level-Building DTW Algorithm" *IEEE Trans. Acoust. Speech Signal Process.* vol. ASSP-29, no. 3, 1981.

[5] Marr, D. *Vision*, Freeman, 1982.

[6] Marr, D. and Nishihara, H.K. "Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes", *Proc. R. Soc. London B 200*, pp. 269-294, 1978.

[7] Grimson, W.E., Lozano-Pérez, T. "Recognition and Localization of Overlapping Parts from Sparse Data", Artificial Intelligence Laboratory, Massachusetts Institute of Technology, A.I. Memo 841, 1985.

[8] Zue, V.W. "The Use of Speech Knowledge in Automatic Speech Recognition", *Proc. IEEE Special Issue on Human-Machine Communication by Speech*, 1985.

[9] Miller, G.A. and Nicely, P.E. "An Analysis of Perceptual Confusions Among Some English Consonants", *Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338-352, 1954.

[10] Shipman, D.W. and Zue, V.W. "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems", *Proc. IEEE International Conference on Speech Acoustics and Signal Processing*, Paris, France. pp. 546-549, 1982.

[11] Kucera H., and Francis, W.N. *Computational Analysis of Present-Day American English*, Brown University Press, Providence, R.I, 1967.

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

[13] Woods, W. and Zue, V.W. "Dictionary Expansion via Phonological Rules for a Speech Understanding System", *Proc. IEEE International Conference on Speech Acoustics and Signal Processing*, Phila, Pa. pp. 561-564, 1976.

[14] Cutler, A. and Foss D.J. "On the Role of Sentence Stress in Sentence Processing", *Language and Speech, 20*, pp. 1-10, 1977.

[15] Kahn, D. "Syllable-Based Generalizations in English Phonology", Doctoral Dissertation, Massachusetts Institute of Technology. Available from Indiana University Linguistics Club, 1976.

[16] Church, K.W. "Phrase Structure Parsing: A Method for Taking Advantage of Allophonic Constraints", Doctoral Dissertation, Massachusetts Institute of Technology. Available from Indiana University Linguistics Club, 1983.

[17] Cole, R.A., Rudnickey A.I., Zue, V.W. and Reddy, D.R. "Speech as Patterns on Paper", in *Perception and Production of Fluent Speech*, R.A. Cole Ed, pp. 3-50. Erlbaum, 1980.

[18] Erman, L.D., Hayes-Roth, F., Lesser, V. and Reddy, D.R. "The Hearsay Speech Understanding System: Integrating Knowledge to Resolve Uncertainty", *Computing Surveys*, vol. 12, no. 2, pp. 213-253, 1980.

CS-TR Scanning Project
Document Control Form

Date : 11 / 9 / 95

Report # AIM-867

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
☐ Laboratory for Computer Science (LCS)

Document Type:

☐ Technical Report (TR) ☒ Technical Memo (TM)
☐ Other:_____

Document Information

Number of pages: 26(32-IMAGES)
Not to include DOD forms, printer intstructions, etc... original pages only.

Originals are:
☒ Single-sided or
☐ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☒ Laser Print
☐ InkJet Printer ☐ Unknown ☐ Other:_____

Check each if included with document:

☒ DOD Form(2) ☐ Funding Agent Form ☐ Cover Page
☐ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: _____

Page Data:

Blank Pages(by page number):_____

Photographs/Tonal Material (by page number):_____

Other (note description/page number):

	Description :	Page Number:
(A)	IMAGE MAPS	(1-26)
	UNMARKED TITLE PAGE,	1-25
	SCANCONTROL, DOD (2), TARGETS	(27-32)
(B)	Cut&Paste PICS ON PAGES	4, 10, 13, 20

Scanning Agent Signoff:
Date Received: 11/9/95 Date Scanned: 11/29/95 Date Returned: 11/30/95

Scanning Agent Signature: Michael W. Cook

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: A.I. Memo No. 867	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER: AD-A165913	
4. TITLE (and Subtitle): Exploiting Sequential Phonetic Constraints in Recognizing Spoken Words	5. TYPE OF REPORT & PERIOD COVERED: A.I. Memo
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Daniel P. Huttenlocher	8. CONTRACT OR GRANT NUMBER(s): N0014-80-C-0505
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency, 1400 Wilson Blvd., Arlington, VA 22209	12. REPORT DATE: October, 1985
	13. NUMBER OF PAGES: 25
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Office of Naval Research, Information Systems, Arlington, VA 22217	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

natural constraints, partial information, word recognition, speech recognition

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

Machine recognition of spoken language requires developing more robust recognition algorithms. A recent study by Shipman and Zue suggest using partial descriptions of speech sounds to eliminate all but a handful of word candidates from a large lexicon. The current paper extends their work by investigating the power of partial phonetic descriptions for developing recognition algorithms. First, we demonstrate that sequences of manner of articulation classes are more reliable and provide more constraint than

20. Abstract (cont'd)

certain other classes. Alone these results are of limited utility, due to the high degree of variability in natural speech. This variability is not uniform however, as most modifications and deletions occur in unstressed syllables. Comparing the relative constraint provided by sounds in stressed versus unstressed syllables, we discover that the stressed syllables provide substantially more constraint. This indicates that recognition algorithms can be made more robust by exploiting the manner of articulation information in stressed syllables.

Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**