CMU-CS-82-144

Towards Very Large Vocabulary Word Recognition

A. Waibel

30 November 1982

Carnegie-Mellon University
Computer Science Department
Speech Project

This research was sponsored in part by the National Science Foundation, Grant MCS-7825824 and in part by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory Under Contract F33615-78-C-1551.

Table of Contents

1. Introduction to the Problem

1.1 Dynamic Programming Template Matching
1.2 Harpy
1.3 Search Space Reduction Techniques

2. Creating a Database for VLVR

2.1 Four Sources for the Design of a VLV Database
2.1.1 Webster's Dictionary
2.1.2 The Brown Corpus - Form B
2.1.3 The Carterette and Freedman Corpus
2.1.4 cmut - an additional source of information
2.2 Design and Realization

3. Suprasegmental and Segmental Filters in VLVR

3.1 Methodological Comments
3.2 Some Properties of a Very Large Vocabulary
3.3 Syllable Counts
3.4 Stress patterns
3.5 Rhythm and Suprasegmental Duration Patterns
3.5.1 Syllable Durations
3.5.2 Voiced/Unvoiced Ratio
3.6 Filter Combinations - Results

4. Conclusions

Appendix I Table of Phonemes and Feature Labels

List of Figures

Figure 2-1:	Average Number of Syllables vs. Word Frequency Rank	6
Figure 2-2:	Percent of Polysyllabic Words vs. Word Frequency Rank	7
Figure 3-1:	Comparison of Measures for Various Cohort Sizes	12
Figure 3-2:	Number of Occurrences vs. Number of Syllables	15
Figure 3-3:	Histogram for Non-Word-Final Syllable Durations in Polysyllabic Words	18
Figure 3-4:	Histogram for Word-Final Syllable Durations in Polysyllabic Words	19
Figure 3-5:	Histogram for Syllables in Monosyllabic Words	20
Figure 3-6:	Histogram for Adjusted Syllable Durations in Polysyllabic Words	21
Figure 3-7:	Histogram for the Ratio of Voiced/Unvoiced Segment Durations in all Syllables	23
Figure 3-8:	Expected Cohort Sizes Using Various Filter Combinations	26
Figure 3-9:	Expected Cohort Sizes for the Frequency Weighted Vocabulary Using Various Filters	27

Abstract

In this paper, preliminary considerations and some experimental results are presented in an effort to design Very Large Vocabulary Recognition (VLVR) systems. We will first consider the applicability of current recognition techniques and argue their inadequacy for VLVR. Possible alternate strategies will be explored and their potential usefulness statistically evaluated. Our results indicate that suprasegmental cues such as syllabification, stress patterns, rhythmic patterns and the voiced - unvoiced patterns in the syllables of a word provide powerful mechanisms for search space reduction. Suprasegmental features could thus operate in a complementary fashion to segmental features.

1. Introduction to the Problem

A typical adult human being with average education can (on the average) recognize words from a vocabulary on the order of 40,000 words quite reliably. Current speech recognition technology is capable of handling vocabularies of only up to 200 words when no contextual, semantic, pragmatic or syntactic information is given to such a system[1] and vocabularies of up to 1000 words in speech understanding systems when a full sentence and a rigid recognition grammar is given[2]. Although it is no doubt true that such systems can already preform satisfactorily in a number of practical applications, they are nevertheless severely limited in generality , extensibility and robustness and do not approach human performance. As a step in the direction of unrestricted speech recognition the barriers imposed by vocabulary size must be resolved. These limitations first have to be removed in the acoustic domain before we attempt unrestricted speech recognition. As a task we propose the design of a 20,000 isolated word recognition system. A vocabulary of this size is in the order of magnitude of the command of language of human beings. It also contains the whole spectrum of word recognition problems, since various levels and kinds of confusability will certainly be encountered. The most successful current recognition strategies have in their present usage insurmountable limitations, when the vocabulary size rises to the proposed dimensions.

1.1 Dynamic Programming Template Matching

Dynamic Programming Template Matching is out of several reasons not easily extensible to very large vocabularies. First, the practicality of a system that has to be trained for very large vocabularies is questionable. Possible extensions can therefore only be obtained if subunits (e.g., syllables, demisyllables or phonemes) smaller than the word are extracted from an unknown word and matched to the pertinent templates. A second difficulty is given by the increase in recognition difficulty. Large vocabularies contain phonetically very similar sounding words (BUCK-DUCK, TWO-TO) and disambiguation requires computationally expensive detailed phonetic analysis. In contrast, there are phonetically totally non-ambiguous word pairs (ANTIDISESTABLISHMENTARIANISM - IN) and inappropriate candidates should be discarded immediately. Thus it is important to recognize word *classes* to eliminate the inappropriate candidates before identifying the recognized word. In this fashion, DP-matching methods have been successfully applied to somewhat larger vocabularies than 200 words[3].

1.2 Harpy

Efficient search of a large pronunciation network has successfully been achieved in HARPY[2] for tasks involving larger vocabularies (~1,000 words). HARPY's success is due to such virtues such as the HARPY

network which provided a constrained search space incorporating syntactic and semantic information, and the efficiency of the search that yielded the overall correct recognition result in spite of errorfull phonetic labeling. For Very Large Vocabularies, in an isolated word recognition task (no syntax/semantics), however, straight application of the HARPY approach leads to very large branching factors that make search an expensive operation. Moreover, phonetic recognition errors will more readily result in high word recognition error rates because of the size and confusability of the vocabulary. More detailed acoustic and phonetic as well as prosodic information needs to be exploited.

1.3 Search Space Reduction Techniques

For Very Large Vocabulary Recognition (henceforth VLVR) computationally inexpensive, robust and powerful mechanisms for *Search Space Reduction* are necessary ingredients for a successful system. Zue and Shipman[4] have recently demonstrated that substantial search space reduction can be achieved using 2-way (Consonant-Vowel) or 6-way featural segmental classification schemes. Nevertheless, large subvocabularies remain, particularly when increased class sizes must be presumed in the presence of error. Furthermore, some class distinctions based on linguistic notions require detailed analysis and are by no means a robust trivial first pass elimination heuristic. A consonant-vowel distinction, for example, can be exceedingly difficult in cases like liquids, glides and nasals. Cole et al.[5] have recently shown (for the alpha-digit task) how a systematic knowledge engineering approach can yield superior performance by applying featural knowledge to making fine phonetic distinctions. Yet, robust criteria to perform highly selective preclassification in all generality have not been demonstrated to date and await further study. One aspect of human speech is known to have great impact on intelligibility and naturalness of speech and yet has been largely ignored for speech recognition devices: prosody, or more generally, suprasegmentals. In this paper we demonstrate the potential impact that a set of suprasegmental features might have on Very Large Vocabulary Recognition.

In the next chapter we will introduce several Very Large Vocabulary Databases that were compiled and evaluated. The properties of Very Large Vocabularies will be discussed. The remainder of the paper will demonstrate the potential of using a combination of suprasegmental and segmental features as filters in the recognition process. Experimental results using the dictionaries described will be reported.

2. Creating a Database for VLVR

A database as a research vehicle for the VLVR task has to be designed according to two major criteria. First, it has to comprise a selection of words that both are commonly used in natural language/speech and impose the whole spectrum of recognition difficulties encountered in VLVR. Second, it has to provide various kinds of information that are needed or useful in the actual recognition process. In the following sections we describe four corpora of very large vocabularies that have been investigated.

2.1 Four Sources for the Design of a VLV Database

2.1.1 Webster's Dictionary

One of the corpora available is a machine readable form of Webster's Dictionary containing the orthographic and phonemic representation for 20,000 words. In the phonemic spelling syllable boundary markers are provided. Homographs have separate entries with pertinent separate phonemic spellings. Some of the problems encountered with this corpus are: typographical errors, archaic, inaccurate or incorrect phonemic transcriptions and the inclusion of words that are not common in present-day American English.

2.1.2 The Brown Corpus - Form B

A corpus of about 1,000,000 words selected from various American texts was collected and evaluated at Brown University[6]. This particular version of the corpus contains the orthographic spelling of the words as well as a count of number of occurrences (word frequencies) in the various source texts. The strength of this corpus is the provision of the word frequency counts and as a consequence the fact that only commonly used American English words are included. Its major drawbacks for our purposes are:

- *Homographs* are collapsed due to their identical spelling, an issue that when dealing with speech recognition tasks has to be resolved.

- The corpus is based on *written* text and thus is biased towards common occurrences in writing rather than common words in speech. For example, formulas and punctuation marks are included here. Indicative of the bias is, for example, the fact that the most frequent word in written text is the word "THE", while in spoken speech "I" occurs most often.

- There is no provision for *phonemic information* or other pronunciation related cues.

Unlike Webster's Dictionary, this corpus also includes various forms derived from a basic root word. Thus for example, "USE", "USES", "USED", "USUALLY", etc. are all listed separately. This property is actually

desirable, since it reflects a real problem in VLVR because of the high phonetic similarity among some of these words.

A further caveat is warranted here: the word frequencies introduce a very strong bias towards a set of about 100 most frequent words, which occur about 50% of the time in written English text. Building a VLVR system optimized for words as they occur most frequently would mean building a 100-200 word recognition system specialized in dealing with the highly (for VLV) atypical class of the 100 most frequent words in English text. This class largely consists of monosyllabic function words that may or may not be useful for VLVR depending on the recognition task. These properties of a large vocabulary are illustrated in figures 2-1 and 2-2. In Fig.2-1 the average number of syllables per hundred words is shown for the 20,000 words of the Webster's dictionary sorted according to the word frequencies provided by the Brown Corpus. Fig.2-2 shows analogously the percentage of polysyllabic words per hundred for the same vocabulary. It can be seen that, the distribution of syllable counts over the word frequency sorted 20,000 word vocabulary stabilizes after the first 300 most frequent words. The words in rank less frequent than the first 200 or 300 might have different properties from those very frequent words.

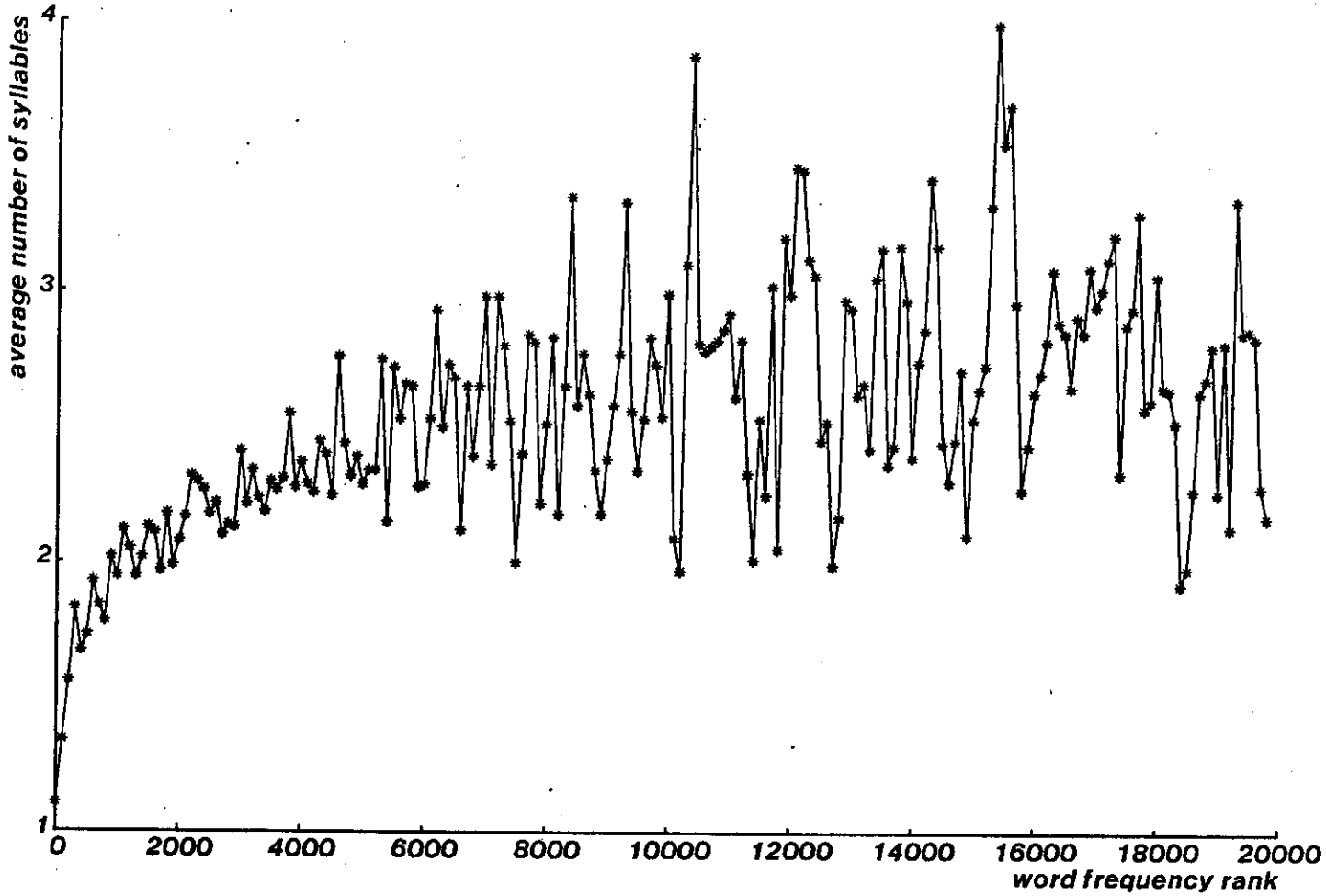

Figure 2-1: Average Number of Syllables vs. Word Frequency Rank

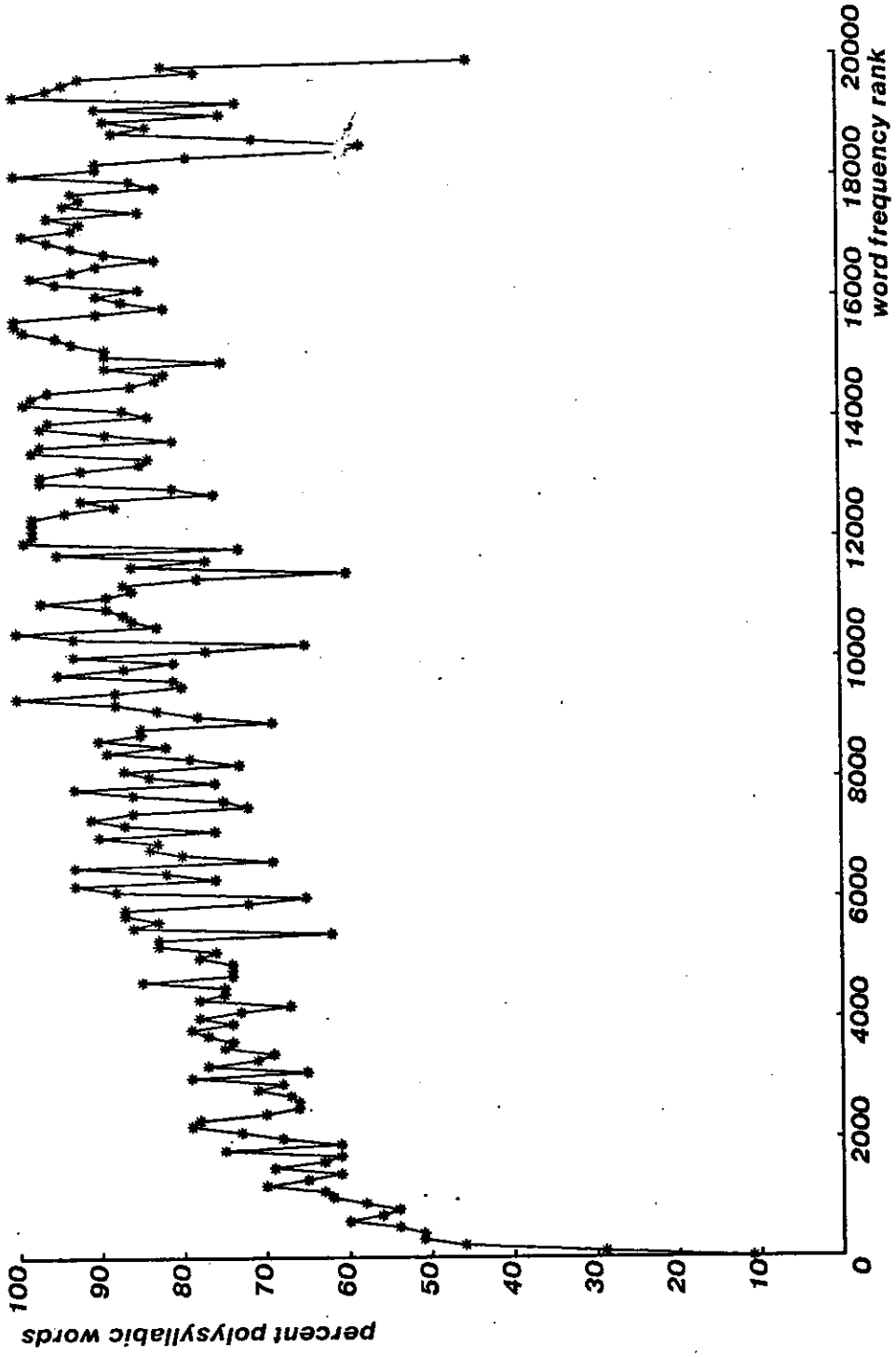

Figure 2-2: Percent of Polysyllabic Words vs. Word Frequency Rank

2.1.3 The Carterette and Freedman Corpus

The Carterette and Freedman Corpus[7] is a corpus of 16,000 spoken words. The speech was recorded without the prior knowledge of the subjects. It is conversational speech ("small talk") of students in a waiting room. Not unlike Brown Corpus - Form B, a list of words and their word frequencies have been generated and are available on line. Although there is significant overlap in the overall collection of words between the words in this corpus and those in the corpora discussed above, new aspects are gained from this particular selection of words. First, a word frequency count is obtained based on spoken language rather than written language. Second, a set of words more particular to spoken speech appear more noticeably: names of people and cities, numbers (digits, years, etc.), letters (presumably for spelling of acronyms) and finally exclamations (Oh, Uh, Nah, Um, etc.). Caveats and drawbacks similar to those for the Brown Corpus apply here, too.

2.1.4 cmut - an additional source of information

An additional source of information available in our facilities, a PASCAL version of the MITalk Text-To-Speech system[8, 9], can be run to obtain speech related information, such as:

- phoneme strings
- phoneme durations
- F0 target values
- various markers, such as Morph Boundary Markers, Syllable Boundary Markers, Stress Markers, Function or Content Word Marker, and, if so desired, Part of Speech Information

It is therefore possible to derive various kinds of speech related information from an orthographic representation, without the necessity of transcribing by hand all the words in a corpus or actual speech measurements. Of course, this information is synthetically derived and the processes that generated it are not problem-free. Thus one has to use this information cautiously as an *approximation* to real speech. Nevertheless, a corpus can be generated that contains useful additions to the corpora discussed above.

2.2 Design and Realization

In this section, we will briefly discuss two databases that we have created for two distinct research efforts leading towards VLVR.

The first, a 20,000 word database, is intended to investigate and experiment with the properties of VLV's. The objective is to consider the effects of various recognition strategies (given certain assumptions) on

performance. The prime goal is to develop methods that filter out a subvocabulary of preferably small size by means of robust detectors of features of various kinds. Since these detectors will include various aspects of speech, it is useful to have a database of words complemented by phonemic, prosodic and possibly morphemic and syntactic information. To this end, an augmented version of the Webster's Dictionary has been created. It contains in addition to the orthographic and phonemic representation, various additional aspects derived from running the individual words from this dictionary through cmut as described above and adding the word frequency count from the Brown Corpus.

The second, a speech database, is intended to provide the framework for the initial stage of the actual implementation of a VLVR system. The union of the 900 most frequent written words (Brown Corpus) and the 900 most frequent spoken words (Carterette and Freedman Corpus) was selected to provide the basis for this database. The union contains equal shares of about 450 words that are either unique to the first 900 spoken words, unique to the first 900 written words, or common to both sets. All exclamations, acronyms, titles and names were preserved. The special punctuation symbols, formulas, etc., contained in the written corpus were eliminated, as well as the somewhat arbitrary selection of numbers and isolated letters. Instead, a list of all the numbers from 1 through 20, the numbers 30, 40, ..., 90, 100 and 1000, and a list of all the letters in the alphabet were added to the corpus. Also to provide instances of long words, a set of 115 words was added that have four or more syllables. The resulting collection of words thus contains approximately 1500 tokens.

The speech database based on this selection of words is currently being collected using four native speakers of American English, several reading sessions each.

3. Suprasegmental and Segmental Filters in VLVR

3.1 Methodological Comments

In this chapter results of a theoretical exploration of possible sources of information will be presented. By "theoretical" we mean that no actual speech database was used. Rather in order to obtain insights into the properties of very large corpora (20,000) the information from the dictionaries discussed in the previous chapter was used. More specifically, the augmented Webster's dictionary discussed in section 2.4 was used. This corpus of 19955 (nominally 20,000) words not only has the orthographic and phonetic representation, but also contains entries for segmental durations, F0 targets and various markers as discussed previously. As was pointed out, this information has been obtained in part synthetically. Syllable boundaries, segmental durations, etc. have so far not been available for corpora of this size, such that synthetic data provides the best interim solution currently obtainable. It must be pointed out, however, that the results presented here must therefore be interpreted only as a first order approximations to the properties of 20,000 words in Webster's dictionary when spoken by humans. As a motivation to using this corpus, however, a few supportive comments are in order.

- The phonetic transcription obtained from synthesis has been found useful for our purpose. It was designed to produce intelligible speech and it is believed to be closer to actual speech and contain finer phonetic distinctions than the description found in the regular Webster's dictionary. This can and has provided valuable additional information that would otherwise not be available. One might understand this distinction by the underlying philosophy behind the transcriptions. The synthetic transcription does not intend to instruct proper pronunciation, but rather attempts to be a close approximation to real contemporary American speech. It is thus more useful for the specific problem we set out to solve, to recognize contemporary American speech.

- Synthesis was obtained (see chapter 2) using a version of MITalk-79[8, 9] an ambitious large scale effort aimed at unrestricted Text-to-Speech synthesis. MITalk-79 was undoubtably developed to a level that is comparable to human speech in intelligibility and naturalness. Many of the pronunciation rules as well as prosodic parameters (such as segmental durations) were obtained from measurements of spectrograms taken from one speaker. The particular sound speech quality might therefore reflect one speaker's peculiarities but resembles closely actual human speech.

- Durations have been obtained from spectrogram measurements at the segmental level. Based on measured segmental durations and a set of 11 modification rules MITalk predicts segmental durations, synthetically. The synthetic segmental durations have been found to differ from measurements on independently collected speech by a standard deviation of 17 msec[8]. These short deviations are less than the just noticeable difference (JND) of temporal variations in speech. The synthetic durations could thus be considered perceptually accurate. All statistics collected in the following involving durations will be limited to the suprasegmental structure of

the words. This eliminates to some extent the possibility of a circular reasoning between synthetic data and the desire to find regularity in the data. More specifically, the segmental durations used in MITalk have been obtained without particular attention to isochrony in English or to the concept of rhythmic beats.[1] We believe therefore that it is valid to consider syllable durations obtained in this fashion.

The word frequency count (obtained from the "Brown Corpus") was added to our database. Most statistics reported in the following, however, will not be based on frequency weighted occurrences for the reasons outlined in the previous chapter, i.e., to avoid a heavy bias towards a rather limited set of function words.

Before we turn to a statistical evaluation of possible suprasegmental filters a few methodological comments are in order. In the following sections, we attempt to define measures of the speech signal that are robust and reduce the number of remaining candidates as much as possible. In other words we seek to evaluate a measure's power to prune the vocabulary to preferably small remaining "cohorts". As a means of evaluation, average cohort size has been used in previous evaluations. This statistic has the disadvantage of not accurately reflecting the amount of pruning of a given measure if the cohort sizes are rather disperse.

We therefore propose the usage of *expected cohort size* given by

$$ECS[s] = \sum_{s_0} s_0 * p_s(s_0)$$

where $p_s(s_0)$ is the probability of any given word to fall into a cohort of size s_0. Expected cohort size would thus take into consideration the likelihood of any particular cohort size to occur. The result could be interpreted as the size of the cohort that a given unknown utterance is expected to fall into after application of one or more search space reduction mechanisms (filters).

Fig.3-1 illustrates the difference between average cohort size and expected cohort size. If for example, (after application of a given measure) two remaining cohorts have size 19,999 and size 1 respectively, then the average cohort size would be 10,000. If in turn the two cohorts both had size 10,000 their average cohort size would also be 10,000. The practical usefulness, however, of a measure giving rise to this latter distribution of words into cohorts is much greater, since the likelihood of a random word to be found in either cohort is 50% and thus the effective pruning much greater. The *expected* cohort size (ECS) in contrast can be evaluated for the first case to be 19998 which corresponds to a pruning to only 99.99% of the original vocabulary, while in the second case the ECS is 10,000 which in turn corresponds to 50% pruning.

Based on these considerations we will report in the following either *expected* cohort size or else whenever useful *maximum* cohort size (i.e., worst case assumption).

---

[1] Carlson et al., however, report partial isochrony as a result of application of the prosodic component in MITalk

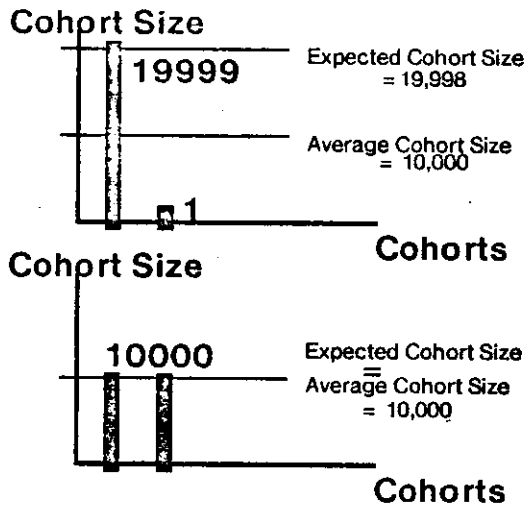

Figure 3-1: Comparison of Measures for Various Cohort Sizes

3.2 Some Properties of a Very Large Vocabulary

Various studies have already examined the statistical properties of phoneme distributions in large vocabularies. Denes[10] in 1963 reports phoneme and digram distributions for 72,000 phonemes as well as the frequency distribution of consonantal minimal pairs. He found AX (schwa), IH, T, N, S, D the most frequent phonemes, thus making consonants with alveolar place of articulation the most common. These and other results are supported by our data. Denes also reports the most common minimal pairs in such a database, i.e., the discriminating phoneme pairs in word pairs that differ only by these phonemes. Most minimal pairs are distinguished by their manner of articulation rather than their place of articulation.

We have found the number of such "similar"[2] word pairs to be surprisingly large. In a previous study using the original phonetic labeling from the Webster's dictionary it has been found that a total of 28,335 pairs of words can be found that differ by one phoneme only. A total of 6263 word pairs differ by the absence (deletion) of one phoneme in one word with respect to the other.

VLVR, however, is complicated not only by the very high number phonetically similar word pairs but also by similarities that cannot be disambiguated on the basis of phonetic identity alone in the first place.

---

[2]Note that these pairs do not in all cases have to be similar sounding from a perceptual point of view. ANIMATION - ANNOTATION may for example cause less confusion to humans than MEDITATION - MEDICATION. In some of these cases, prosodic differences or similarities might give rise to better or worse discriminability. Different stress levels might improve discriminability perceptually. Finally, different phonetic categories might differ in their discriminatory power.

- There are 376 word pairs (i.e., the recognition of 752 words is affected) that are indistinguishable by discrimination of phonemes or stress patterns or syllable boundary location (e.g., TWO - TOO, RED - READ, etc.) Discrimination can be done only on the basis of contextual cues, or likelihood of occurrence (as primed by cultural bias, experience, context or simply word frequency).

- An additional 55 word pairs are discriminable on the basis of stress only (e.g., 'Increase - incr'Ease). Note that this number is derived from words having identical phonetic spelling with the exception of their stress location. Most words pairs, however, that differ in stress patterns also differ in some aspect of their phonetic realisation (P'Erfect - Perf'Ect).[3] Stressed vowels when destressed, frequently change to reduced vowels and therefore their spectral characteristics change. The real number of pairs that are distinguishable mainly by stress is therefore probably much higher.

- 24 word pairs differ in the presence or the location of a syllable boundary only (e.g, Unreel - Unreal, Dower - Dour). Again this number was computed on the basis of words that other than the syllable boundary have identical phonetic strings. This number is presumably much greater in reality also, since some word pairs differing in syllable boundary only, will nevertheless be represented in our corpus by differing phonetic strings. This might for example be the case for vowel - vowel sequences that within one syllable would be represented by diphtongs. Discrimination between these pairs might be possible based on accurate location of the syllable boundaries or by analysis of the temporal structure of the word in question.

From the points raised above it seems clear that there is a substantial number of words that are discriminable from others on the basis of prosodic information. But could prosodic or more generally suprasegmental information be of use as preliminary filters to eliminate unlikely candidates in general ?

3.3 Syllable Counts

One possibility for a crude search space reduction mechanism would be to reject candidates that do not have the same number of syllables as the unknown utterance. It has been shown that the detection of syllable boundaries can be performed with an accuracy of better than 90% correct on continuous speech. In Fig.3-2 the number of occurrences of words with a specific number of syllables can be seen. The broken line shows the distribution over the 20,000 word vocabulary discussed. The solid line indicates the distribution of the same vocabulary, but weighted by the frequency of occurrence in the brown corpus[6]. The implicit assumptions in these two graphs are that in the first case all words occur with equal frequency and in the second case that a given word occurs as frequently as indicated by the Brown Corpus frequency. In the latter

---

[3] incidentally, most of the 118 Homographs found in our corpus differ by stress only with or without corresponding phonemic alterations.

case, for example, this means that in a recognition task the word THE is assumed to have occurred 69971 times and is counted as such while the word ABNORMALLY is counted only once. It is clear from this example and from Figures 2-1 and 2-2 that the frequency weighted distribution has a strong bias towards monosyllabic words: frequent words, in particular function words, tend to be shorter in number of syllables. In English entire paragraphs of monosyllabic words[11] are possible without any noticeable distortion in naturalness. Thus syllable counts could be considered as a means for classification in a large vocabulary with relatively limited potential for search space reduction. The ECS for the frequency weighted vocabulary is 12,628 which corresponds to an effective reduction to 63%. Assuming that all words are equally probable the ECS is 5013, i.e., an effective pruning down to 25% of the corpus.

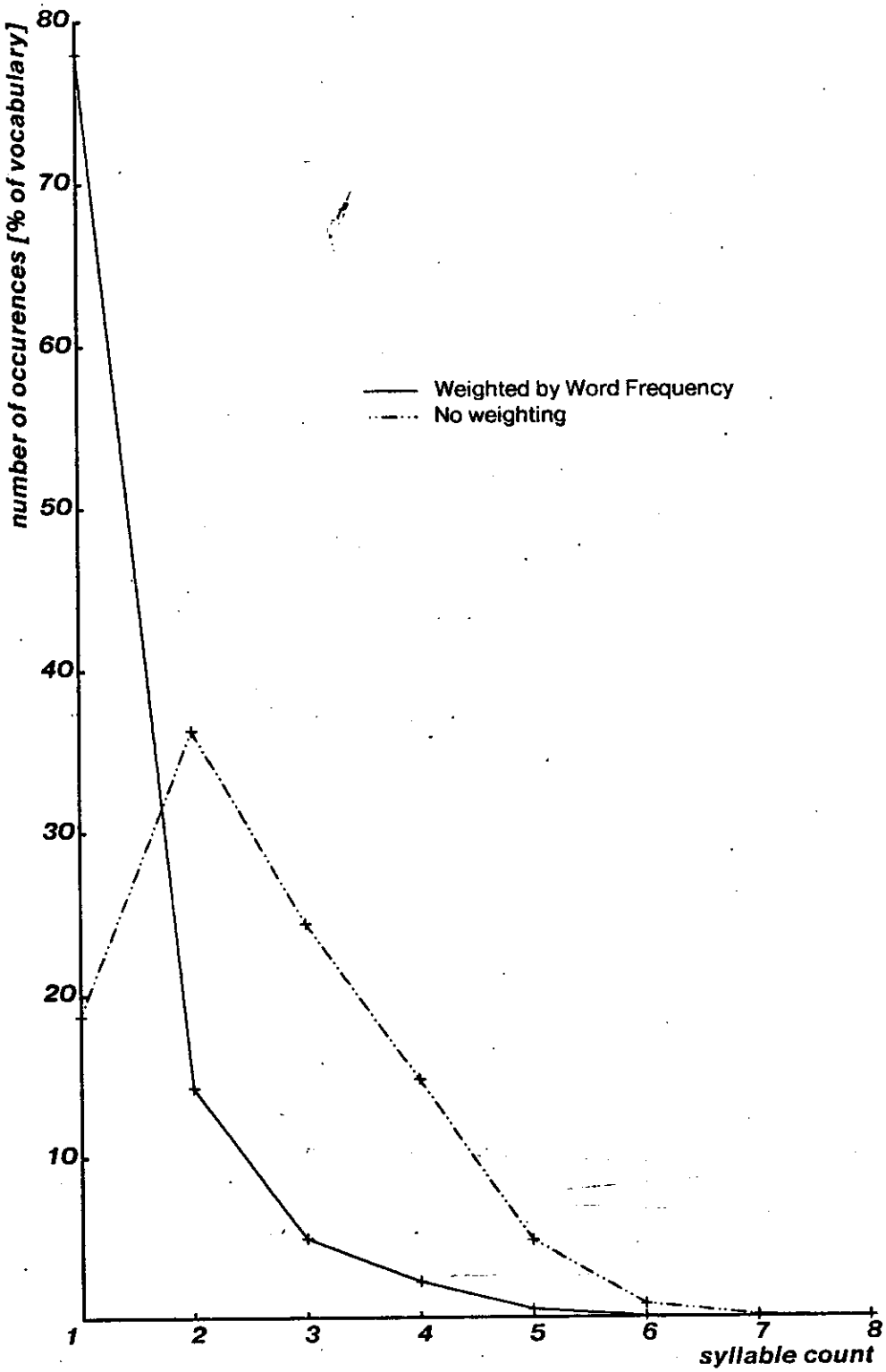

Figure 3-2: Number of Occurrences vs. Number of Syllables

3.4 Stress patterns

Stress and rhythmic patterns have been discussed by several authors[12, 13, 14, 15]. Depending on the psycholinguistic model they develop these two linguistic concepts may not easily be separable. For this experimental evaluation, however, we have chosen to explore stress and rhythm as independent entities.

From our 20,000 word corpus patterns were extracted that were either unstressed or carried primary stress. We have not considered stress levels other than primary stress, since the possibility of detecting them robustly is rather unlikely[4].

A total of only 45 unique stress patterns was identified, some of which occured very infrequently and may in fact due to incorrect or questionnable stress labels in the input. The three most frequent stress patterns are (1 = primary stress; u = unstressed) : 1-u, 1, and 1-u-u. In all three cases the stressed syllable is in word initial position. 12,252 words (more than half the dictionary) fall into either category. Following in frequency are u-1 (like in UNITE) and u-1-u, etc. If we were to use stress patterns as a filter for VLVR, the resulting ECS would be 3055, which corresponds to 15.3% of the corpus.

3.5 Rhythm and Suprasegmental Duration Patterns

Deaf speech[16] and foreign accents[17] cause considerable difficulty in intelligibility to normal native speakers of a language. One major reason for the poor intelligibility is that the temporal structure in both cases is anomalous. The English language is isochronous and stress timed. In fluent speech speakers of English place intervals of approximately equal duration between syllables carrying primary stress. If several unstressed syllables are to fill this interval they are reduced in duration, in theory[15] to 1/2 or 1/4 units. A consequence of shorter or longer syllables are the metric feet that make up the rhythmic structure of English speech[18]. Other languages (for example French) are syllable timed languages, i.e., all syllables are of equal length. These differences in rhythmic patterns give rise to some of the difficulties foreigners encounter when learning a new language and of course create perceptual problems when trying to decode foreign accents[17]. Indeed the lack of rhythm is one of the major difficulties in deaf speech[16].

If we assume that in normal English speech there is a consistent rhythmic structure, and if native English speakers seem to be making strong use of this structure in the perceptual process, then it is reasonable to examine the durational patterns for VLVR. In this section, we will examine two forms of suprasegmental temporal patterns: 1.) syllable durations and 2.) the ratio in durations of voiced and unvoiced segments in a syllable.

---

[4] In the psycholinguistic literature this has been found to be a task, that is difficult even to the human listener. Stress may in fact be partially a psychological phenomenon, that may or may not be readily available from the signal)

In order to render this measure meaningful, we need to define the extent of a syllable first. The location of the syllable boundaries given by our synthetic data was found rather inconsistent over the whole 20,000 word vocabulary. It should also be our concern here to identify a syllable boundary location that is identifiable in the speech signal. Allen[19] presents a very thorough treatment of syllable boundary location by humans. He suggests that the apparent perception of syllable boundaries is in fact the perception of rhythmic beats and measures their location in a series of click matching and tapping tasks. Syllable boundaries are found quite reliably (with little variability) at the onset of stressed syllables. Generally a syllable boundary is placed somewhat before the onset of the nuclear vowel of the syllable in question. The time interval by which the syllable boundary precedes the vowel nucleus is determined by the consonant (cluster) at the boundary. For sonorant syllable junctures the onset of the return from the maximum formant excursion towards the vowel nucleus is the boundary.

As a first order approximation, the onset of the vowel nucleus can be considered to be the syllable boundary. For our database, all syllable boundary markers were adjusted to reflect this change. One of the disadvantageous side affects of this adjustment is that segments leading a word initial syllable stand alone and are not included in any syllable. From a practical point of view, however, this is a quite useful situation, since the duration of leading segments such as voiceless stops can not easily be measured (due to the leading silence interval). If one sets out to measure real speech one does have to resort to this solution[18].

3.5.1 Syllable Durations

Syllable durations for each syllable defined in this fashion were computed by summing up the segmental durations. In Fig.3-3 and 3-4 we find histograms of syllable durations with the syllable being in non-word-final position or in word-final position for polysyllabic words and in Fig.3-5 with the syllable durations of monosyllabic words. As should be expected, the average syllable length in word-final position is longer than in non-word-final syllables. To provide a unified measure, the distributions of Fig.3-3 and 3-4 were collapsed by "shortening" all word-final syllables by 9 csec (i.e., left shifting the histogram of Fig.3-4 by 9 csec) and combining it with the histogram of 3-3. The resulting histogram is shown in Fig.3-6. Fig.3-6 shows three major excursions and we will define these peaks as the short, medium and long syllables assumed in the theory. When we place boundaries at the major dips in the histogram, i.e., at 10 and at 16 we obtain three groups that will be labeled L, M or H for Low, Medium, or High Syllable duration. In the case of monosyllabic words, only one boundary was chosen at 44 csec, resulting in only two classes, Low and High.

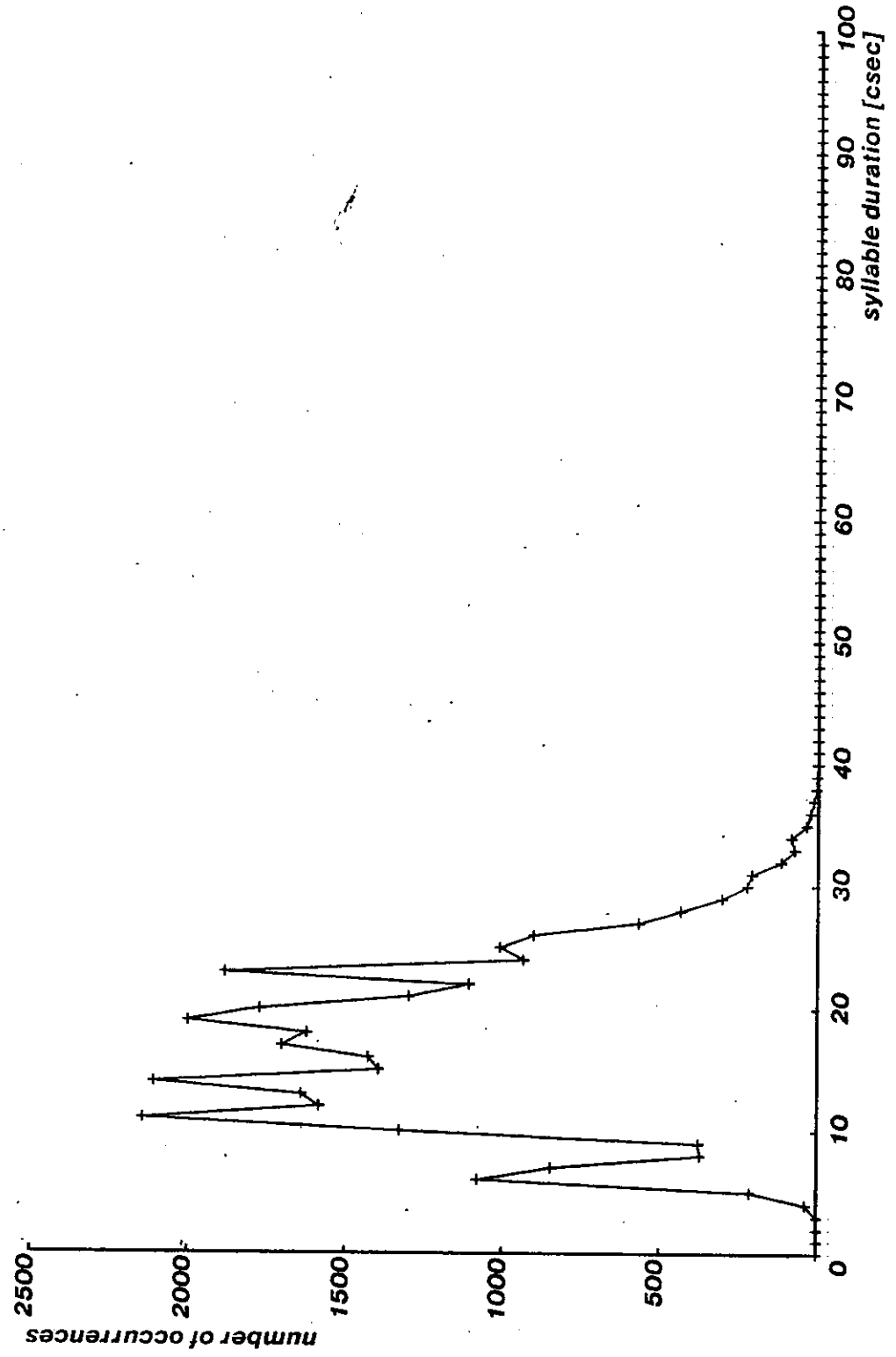

Figure 3-3: Histogram for Non-Word-Final Syllable Durations in Polysyllabic Words

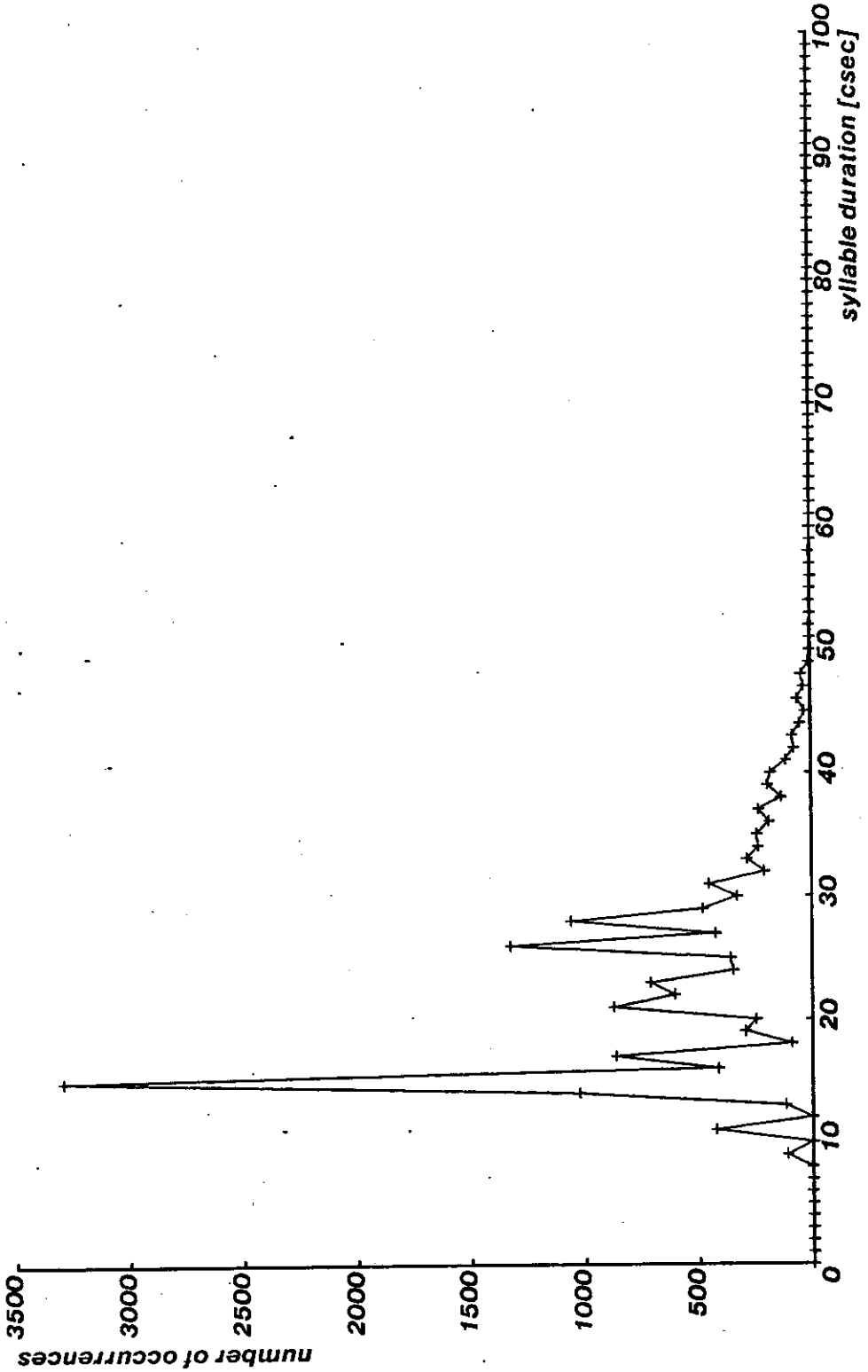

Figure 3-4: Histogram for Word-Final Syllable Durations in Polysyllabic Words

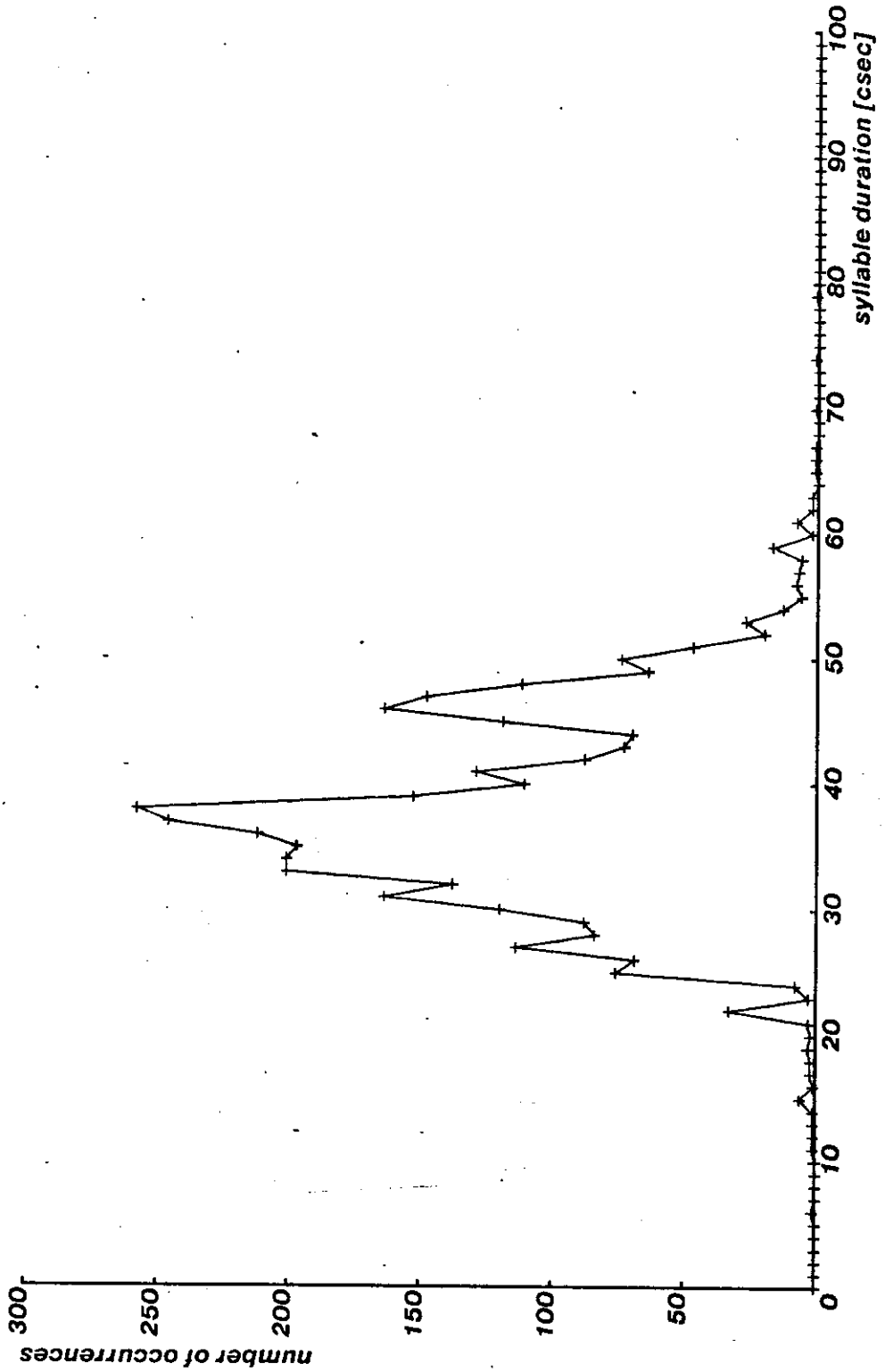

Figure 3-5: Histogram for Syllables in Monosyllabic Words

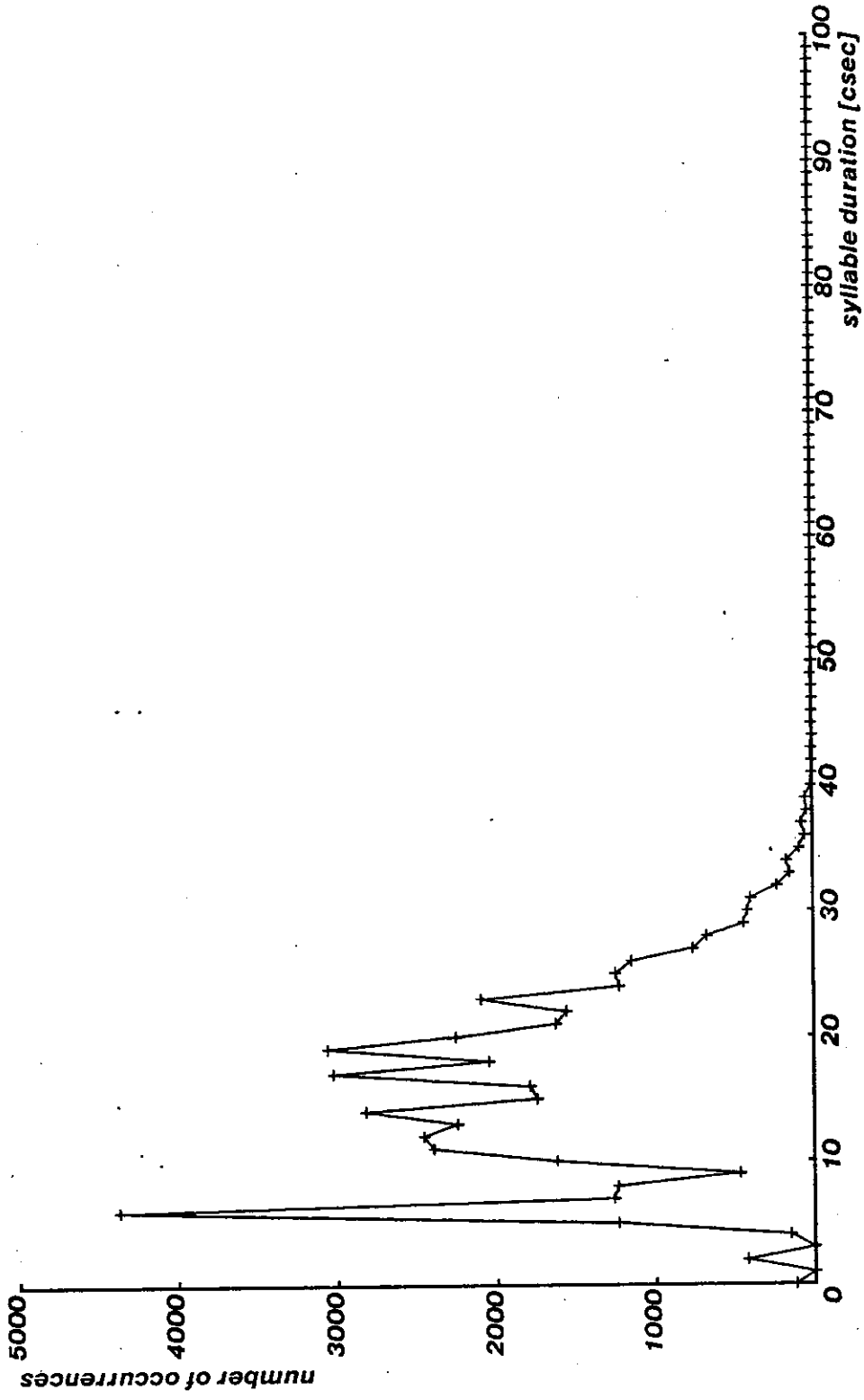

Figure 3-6: Histogram for Adjusted Syllable Durations in Polysyllabic Words

Using these labels duration and stress were first evaluated together. 75% of all syllables carrying primary stress fall under category High, 10% under category Medium and 15% under Low. 14% of the 15% Low duration primary stressed syllables, however, are monosyllabic words. Thus it could be said that in the majority of polysyllabic words stressed syllables carry a High label.

Using this classification scheme all words in our corpus were represented by n syllable duration labels (H, M or L) where n is the number of syllables in the word.[5] A total of 362 unique patterns were found. The largest group consists of 2965 words with identical patterns. The ECS is 1249, i.e., after elimination of the inappropriate prosodic patterns a subvocabulary of only 6% the size of the original vocabulary remains if we assume that any word of our original corpus is equally likely to be the unknown word.

3.5.2 Voiced/Unvoiced Ratio

An additional measure was motivated by the possibility that the relative share of voiced or unvoiced segments in a syllable could provide some overall early rejection or acceptance of a word candidate. In spectrogram reading experiments labels such as "mainly voiced", "all voiced" or "mainly unvoiced" have proven to be useful methods to early rejection of unlikely word candidates[20]. We have here evaluated this measure in an analogous fashion as the syllable durations. For each syllable the voiced to unvoiced ratio is computed. The resulting histogram is shown in Fig 3-7.[6] Again three groups can be identified and will be labeled as Low, Medium and High, where High represents the "all voiced" syllable case. Syllables that have voiced to unvoiced ratio of less than 1, i.e., that contain more unvoiced speech (frication, silence, aspiration) than voiced (e.g., SIX) are labeled L. M are all the syllables containing a smaller proportion of unvoiced than voiced segments. Finally, all uniquely voiced syllables are labeled H. In Fig.3-7 H syllables are indicated by the triangle in the upper right corner (the UV/V - ratio is infinity in this case). It should be noted here that we call here "unvoiced" or "voiced", respectively, what we believe can be detected in the signal as an aperiodic or a periodic signal. Thus, for example, we call voiced stops (B, D, G) unvoiced, since a pitchtracker will typically label the segment unvoiced in spite of the occasional presence of periodic low amplitude prevocalization pulses. Alternatively, flaps (e.g., the T in WRITING) are labeled as voiced.

Grouping all words according to their V/UV label patterns, we obtain 352 cohorts, the largest of which contains 2098 words. The ECS is 909 words or 4.5% of the original corpus.

---

[5] Note that syllable count information is implicitly used here.

[6] Separate treatment of word-initial, word-final syllables or monosyllabic words has resulted in similar results. Hence, only the collapsed distribution for all the syllables in the corpus is presented here.

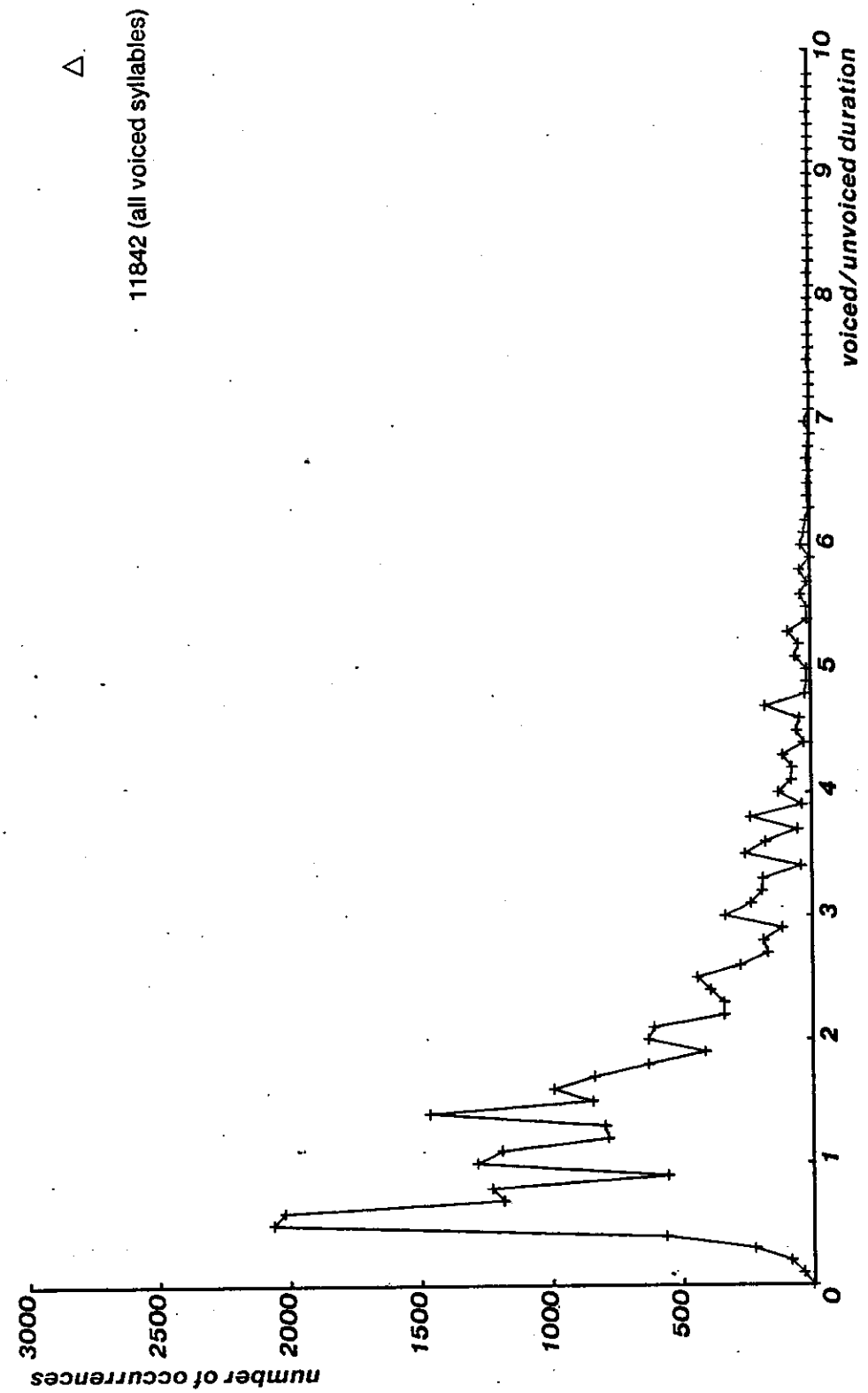

Figure 3-7: Histogram for the Ratio of Voiced/Unvoiced Segment Durations in all Syllables

3.6 Filter Combinations - Results

A number of experiments were performed to examine the effect that a combination of the labeling schemes discussed in the previous sections could have on pruning the 20,000 dictionary. Once again, what we are interested in is the application of filters that provide crude, first pass, robust classifications that eliminate all entries from consideration that do not belong to a set of near miss alternatives. The thesis of this paper is that for a task of this size, prosody might provide a powerful robust near miss mechanism that could potentially operate in parallel with segmental classification schemes. In this section we will combine some of the suprasegmental classification schemes developed above with each other as well as with carefully selected sets of segmental classifications. Above all, it has been our attempt to use classification criteria that we believe can robustly be derived from the speech signal. The criteria have been selected based on the experiences gained from a feature based recognition approach to small vocabulary word recognition[5].

The results of the various experiments are displayed in Fig.3-8. In all cases the vertical axis shows the ECS for the various filter combinations. In the first and second column the ECS is given for the case that words are classified by crude segmental feature patterns only or by duration patterns only, respectively, as discussed in the previous section. The third column shows the case where primary stress markers were added into the duration patterns. The resulting ECS of 978 corresponds to an expected pruning of the vocabulary down to 4.9%. The fourth column represents the results for the duration ratios of voiced to unvoiced segments (V/UV-ratio) in a syllable as discussed in the previous section. In the fifth column the duration labels and the V/UV ratio-labels were jointly used to classify the vocabulary into 1891 cohorts the largest of which identifies 1411 words (no stress markers were used here). The ECS is 381 (1.9% of the original vocabulary). In column 6, like in column 5, the duration labels, V/UV ratios as pattern generators are given with the addition of the primary stress markers whenever appropriate.

Columns 7 and 8 finally illustrate the usage of the suprasegmental filters as in column 6 with the addition of segmental filters. For segmental filters, two levels of detail were chosen. The first attempts to only capture very crude phonetic features, e.g., the strong fricatives S, SH, Z, ZH and the voiceless stops P, T, K, CH, J. It is believed[5] that these labels can easily be detected in most cases. The resulting ECS is only 62 words which corresponds to a reduction to 0.3% of the vocabulary. When allowing for a slightly more detailed featural analysis an even more remarkable search space reduction can be achieved. Included were (see Appendix) subsets of the closed or open vowels (major criterium in the selection was again identifiability - here the guideline was whether a particular sound can be robustly classified by a low or high F1. Ambiguous sounds were left unlabeled and hence do not appear in the patterns). Moreover voiceless stops were included as well as weak and strong fricatives and the liquids W and WH. 14,080 unique patterns were identified. The largest cohort contains 94 words. The ECS was found to be 6 words which corresponds to a reduction to 0.03% of the search space.

The drop in ECS when combining segmental and suprasegmental features is surprisingly large. This behavior might be hypothesized (at least in part) by the complementary nature of the two domains. It appears that suprasegmental information provides a powerful new perspective to analyze a given unknown utterance.

For completeness, Fig.3-9 shows the same filter combinations for the frequency weighted vocabulary. As could be predicted from the previous discussion, the high frequency of the monosyllabic function words introduces a strong bias towards the properties typical for monosyllabic words. The most prominent effect in this case, the comparably small benefit of durational patterns is easily explained by the fact that for monosyllabic function words only a two way distinction had been made. In contrast, phonetic features, however crude they may be, provide more discriminatory information in this particular case.

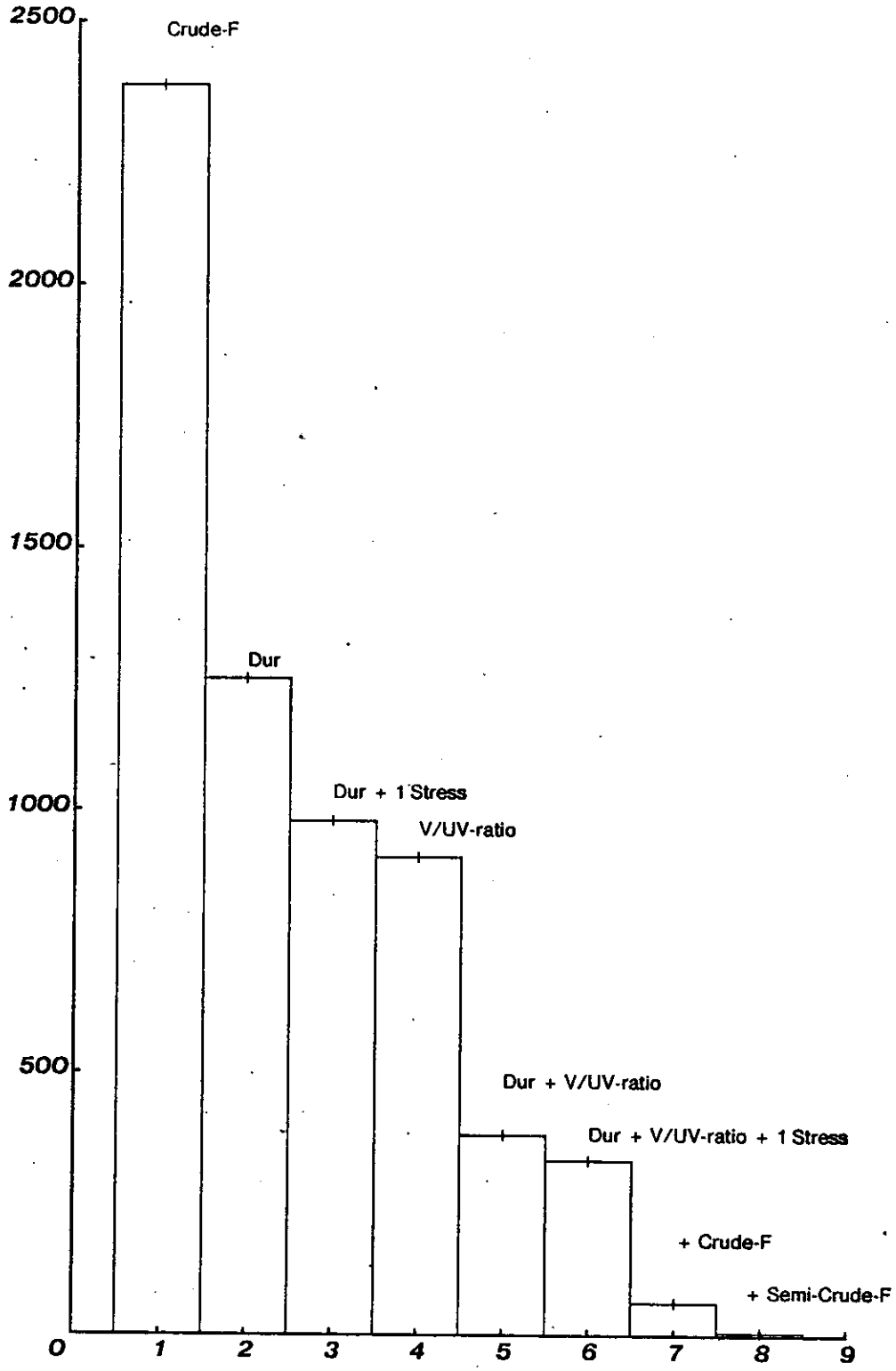

Figure 3-8: Expected Cohort Sizes Using Various Filter Combinations

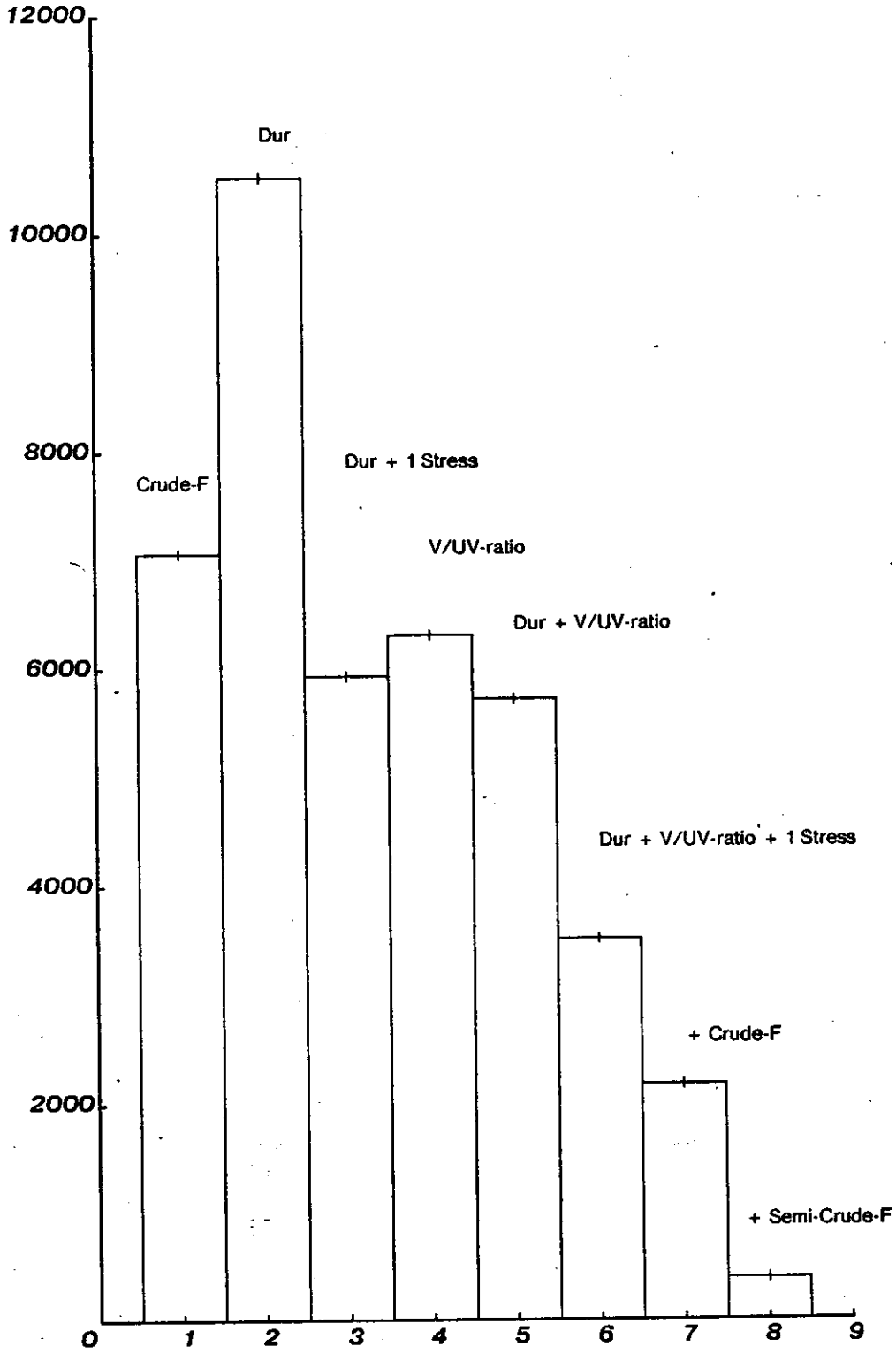

Figure 3-9: Expected Cohort Sizes for the Frequency Weighted Vocabulary Using Various Filters

4. Conclusions

In this paper we have examined the problems and some potential solutions to very large vocabulary recognition. We have presented a large vocabulary database. Based on synthetic data generated for this vocabulary, we have shown that prosodic features as well as a set of phonetic features can provide powerful cues to narrow down the large search space given by a very large vocabulary recognition task. The prosodic features of rhythm and the ratio of unvoiced to voiced segments have been found to be largely complementary to phonetic features, i.e., they are not redundant. Stress patterns yield little additional reduction of the search space, given rhythmic (durational) patterns.

Additional prosodic features, such as amplitude patterns and pitch contours have not been examined in this paper. Such additional, potentially useful prosodic features as well as the robustness of those discussed in this paper will have to be tested in a recognition system. The 1500-word speech database described above will serve as a starting point for our efforts in this direction.

The present results indicate that the study of prosody might provide substantial additional information for a very large vocabualary (~20,000 words) isolated word recognition system. Moreover, when expanding towards connected speech (e.g., a dictation machine), recognition of the prosodic information in speech, appears indispensable.

Appendix I Table of Phonemes and Feature Labels

In the following a table of features is given. The phonetic alphabet in the first column is the Arpabet as used by MITalk. The labels in the second column indicate voiced (1 or 2), unvoiced (0) segments or other markers produced by MITalk (-1). In the third column the segmental labels are given: open vowels (o), closed vowels (c), weak fricatives (v), strong fricatives (f), voiceless stops (s), W-sounds (w) and nasals (n). The labels are in some cases unorthodox depending whether they are readily extractable in speech. The fields on the right are example words from our corpus. Notice that the syllable boundaries are *not* adjusted here, i.e., differ from the representation used for the prosodic measurements.

#C	-1	*	content word	
#F	-1	*	function word	
*	-1	*	morph boundary	
-	-1	*	syllable boundary	
1	-1	*	primary stress	
2	-1	*	secondary stress	
AA	2	o	NOT	.#C.N.1.AA.T.AXP
AE	2	o	HAVE	.#C.H.1.AE.V
AH	2	o	OTHER	.#C.1.AH.DH.-.ER
AO	2	o	LONG	.#C.L.1.AO.NG
AW	2	o	OUT	.#C.1.AW.T.AXP
AX	2	*	ABOUT	.#C.AX.-.B.1.AW.T.AXP
AXP	0	*	UP	.#C.1.AH.P.AXP
AXR	1	o	PART	.#C.P.1.AXR.T.AXP
AY	2	o	NIGHT	.#C.N.1.AY.T.AXP
B	0	*	BETTER	.#C.B.1.EH.DX.-.ER
CH	0	s	CHURCH	.#C.CH.SH.1.ER.CH.SH
D	0	*	DONE	.#C.D.1.AH.N
DH	1	v	THUS	.#C.DH.1.AH.S
DX	1	*	MATTER	.#C.M.1.AE.DX.-.ER
EH	2	*	THEM	.#F.DH.EH.M
EL	1	*	FINALLY	.#C.F.1.AY.N.-.EL.*.-.L.IY
EM	1	n	ISM	.#C.1.IH.-.Z.EM
EN	1	n	PERSONAL	.#C.P.1.ER.S.-.EN.*.-.AX.LX
ER	1	*	FIRST	.#C.F.1.ER.S.T.AXP
EXR	1	*	THERE	.#C.DH.1.EXR
EY	2	*	MADE	.#C.M.1.EY.D.AXP
F	0	v	FOR	.#F.F.OXR
G	0	*	GOOD	.#C.G.1.UH.D.AXP
GP	0	*	GET	.#C.GP.1.EH.T.AXP
H	0	v	HE	.#F.H.IY
HX	0	v	PERHAPS	.#C.P.ER.-.HX.1.AE.P.S
IH	2	c	IN	.#F.IH.N
IX	2	*	MEXICAN	.#C.M.1.EH.K.S.*.-.IX.K.*.-.AX.N
IXR	1	c	HERE	.#C.H.1.IXR
IY	2	c	THESE	.#F.DH.2.IY.Z

J	0	s	JUST	.#C.J.ZH.1.AH.S.T.AXP
K	0	s	MAKE	.#C.M.1.EY.K.AXP
KP	0	s	CAME	.#C.KP.1.EY.M
L	1	*	LEFT	.#C.L.1.EH.F.T.AXP
LX	1	*	ALWAYS	.#C.1.AO.LX.-.W.EY.Z
M	1	n	MORE	.#F.M.2.OXR
N	1	n	NO	.#C.N.1.OW
NG	1	n	THINK	.#C.TH.1.IH.NG.K.AXP
OW	2	o	ONLY	.#F.2.OW.N.-.L.IY
OXR	1	o	COURSE	.#C.K.1.OXR.S
OY	2	*	POINT	.#C.P.1.OY.N.T.AXP
P	0	s	PEOPLE	.#C.P.1.IY.-.P.EL
R	1	*	VERY	.#C.V.1.EH.R.-.IY
S	0	f	SO	.#C.S.1.OW
SH	0	f	SHOULD	.#F.SH.UH.D.AXP
T	0	s	TO	.#C.T.1.UW
TH	0	v	THREE	.#C.TH.R.1.IY
TQ	1	*	WRITING	.#C.R.1.AY.TQ.*.-.IH.NG
UH	2	*	COULD	.#F.K.UH.D.AXP
UW	2	c	SCHOOL	.#C.S.K.1.UW.LX
UXR	1	*	YOUR	.#F.Y.UXR
V	1	v	EVERY	.#F.2.EH.V.-.R.IY
W	1	w	WAR	.#C.W.1.OXR
WH	1	w	WHY	.#C.WH.1.AY
Y	1	*	YOUNG	.#C.Y.1.AH.NG
YU	1	c	HUMAN	.#C.H.1.YU.-.M.AX.N
Z	1	f	PRESIDENT	.#C.P.R.1.EH.Z.-.AX.-.TQ.EN.T.AXP
ZH	1	f	DECISION	.#C.D.IH.*.-.S.1.IH.ZH.-.AX.N

References

1. J.G. Wilpon, L.R. Rabiner and A. Bergh, "Speaker-Independent Isolated Word Recognition Using a 129-Word Airline Vocabulary," *The Journal of the Acoustical Society of America*, Vol. 72, No. 2, August 1982, pp. 390-396.

2. B.T. Lowerre, *The Harpy Speech Recognition System*, PhD dissertation, Computer Science Department, Carnegie Mellon University, 1976.

3. C.S. Myers and L.R. Rabiner, "An Automated Directory Listing Retrieval System Based on Recognition of Connected Letter Strings," *The Journal of the Acoustical Society of America*, Vol. 71, No. 3, March 1982, pp. 716-727.

4. D.W. Shipman and V.W. Zue, "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," *ICASSP '82*, IEEE ASSP, 1982, pp. 546-549.

5. R.A. Cole, M.S. Phillips, S.M. Brill, P. Specker, A.P. Pilant, "Speaker Independent Recognition of English Letters," *The Journal of the Acoustical Society of America*, Vol. 72, No. Supplement 1, Fall 1982, , session R11 (abstract published)

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

7. E.C. Carterette and M.H. Jones, *Informal Speech*, Berkeley: University of California Press, 1974, 1974.

8. J. Allen, R. Carlson, B. Grandstrom, S. Hunnicutt, D. Klatt, D. Pisoni, *Conversion of Unrestricted English Text to Speech*, Massachusetts Institute of Technology, 1979.

9. J. Allen, "Synthesis of Speech from Unrestricted Text," *64*, IEEE, 1976, pp. 422-433.

10. P.B. Denes, "On the Statistics of Spoken English," *The Journal of the Acoustical Society of America*, Vol. 35, No. 6, June 1963, pp. 892-904.

11. T.H. Crystal and A.S. House, "Segmental Durations in Connected Speech Signals: Preliminary Results," *The Journal of the Acoustical Society of America*, Vol. 72, No. 3, Sept 1982, pp. 705-716.

12. I. Lehiste, *Suprasegmentals*, MIT-Press, 1970.

13. M.Y. Liberman, *The Intonational System of English*, Indiana University Linguistics Club, 1978.

14. A. Classe, *Rhythm of English Prose*, Basil Blackwell, Oxford, 1939.

15. J.G. Martin, "Rhythmic (Hierarchical) versus Serial Structure in Speech and Other Behavior," *Psychological Review*, Vol. 79, No. 6, 1972, pp. 487-509.

16. R.S. Nickerson, "Characteristics of the Speech of Deaf Persons," *The Volta Review*, Vol. 77, No. 6, 1975, pp. 342-362.

17. F.M. Chreist, *Foreign Accent*, Prentice-Hall, Inc., 1964.

18. I. Lehiste, "Rhythmic Units and Syntactic Units in Production and Perception," *The Journal of the Acoustical Society of America*, Vol. 54, No. 5, 1973, pp. 1228-1234.

19. G.D. Allen, "The Location of Rhythmic Stress Beats in English: An Experimental Study I and II," *Language and Speech*, Vol. 15 and 16, 1972, pp. 72-100 and 179-195.

20. R. Cole, "Detectable Features", Personal Communication