SYLLABLE STRUCTURE OF ENGLISH WORDS: IMPLICATIONS FOR LEXICAL ACCESS

Michiko Kosaka and Hisashi Wakita

Speech Technology Laboratory, 3888 State St., Santa Barbara, CA 93105.

We parsed a large corpus of English words into syllables and into their constituents to determine the difference between high and low frequency words with respect to these structural properties. There are obvious applications of the results to the lexical access problem in large-vocabulary isolated-word speech recognition systems.

INTRODUCTION

One of the problems in the theories of word recognition involves the relationship between the frequency of usage of words and the structural properties of them. This question is interesting because (1) the differences in word frequency effects might be due to factors other than the frequency of usage, and (2) we might be able to clarify the nature of lexical access, i.e. whether words are accessed on the basis of an acoustic, phonetic or phonological representation. This question is also interesting for isolated-word large-vocabulary machine recognition systems because (3) certain structural constraints in lexical access have been shown to be very powerful in reducing the search space for candidate words. The precise form of the lexical entries is very important for continuous speech recognition systems.

MATERIAL AND METHODS

Brown Corpus words were used as the data. Following Pisoni, et.al., we defined high frequency words as those equal to or greater than 1000 words per 1 million (e.g. *the, of, many*), and low frequency words as those between 10 and 30 words per 1 million inclusively (e.g. *acceleration, bronchial, conjugate*). In addition, we defined mid frequency words to be 30 to 1000 words per 1 million exclusively (e.g. *able, measurement, strike*). These words were matched against the phonetic transcriptions of the SCRL dictionary, which resulted in a data base of a total of 7443 words. There were 91 high frequency words, 3072 mid frequency words and 4280 low frequency words.

Brown Corpus words might not be an ideal sample because the material is approximately 20 years old and because it is based on printed texts as opposed to a transcription of the spoken language. Nevertheless, because of a lack of other computer-readable data bases, we took the the Brown Corpus words to be our sample. It might be argued that word information from the spoken language is not an appropriate alternative, since we do not expect people to speak to the machines in the same way that they would speak to other people.

The phonetic transcriptions (ARPAbet) of these words were parsed by a syllable parser developed at STL. The syllabication of the parser is based on the maximum onset principle. Stress resyllabication was not included in this parser, since stress information was not available in a convenient form. Therefore, the onset count should be slightly over-represented for syllable-initial consonant clusters and slightly under-represented for syllable-final coda consonant count. The quantitative effect of this ommission is not clear, but we do not expect it to be significant.

This study focuses on the frequency of usage vs. syllable length and sub-syllabic constituents. A motivation for this is that previous studies on the phonological structural properties of words dealt exclusively with the identity of phonemes and their length in terms of phonemes [1, 2, 5, 6, 7].

WORD FREQUENCY AND LENGTH

Table 1 below shows the relationship between the word length (in syllables) and the frequency ranges of high, mid and low. Table 2 shows the relative frequency of occurrence within each frequency class. The results indicate that the high frequency words are different from mid and low frequency words and that they are from two separate populations. The Pearson correlation of mid and low frequency was 0.9. Thus the mid and low frequency words can be considered to be from the same population. That the two populations are independent can be seen from the proportion of one-syllable words. They are 0.88 0.35 and 0.23 for high, mid and low frequency words, respectively. The mean length for each group was 1.12, 2.01 and 2.33 for high, mid and low frequency words, respectively. One syllable is the median of high frequency words; whereas the median of mid and low frequency words are two syllables.

Table 1: Word Frequency and Length (Syllable)

length	high	mid	low	total
1	80	1073	978	2131
2	11	1199	1681	2891
3	0	541	1033	1576
4	0	211	429	640
5+	0	48	159	207
total	91	3072	4280	7443

Table 2: Word Frequency and Length (%)

length	high	mid	low	total
1	87.91	34.93	22.85	28.63
2	12.09	39.03	39.28	38.84
3	-	17.61	24.14	21.17
4	-	6.87	10.02	8.60
5+	-	1.56	3.71	2.78

WORD FREQUENCY AND SYLLABLE CONSTITUENTS

Difficulties in intelligibility of certain words have often been, in part, attributed to the lexical distance based on the frequency [1] and to the particular phonemes, or phoneme/grapheme ratios [2]. We investigated two factors that might account for such difficulties.

Word Frequency and Onset

The onsets were classified as nil (no consonant at the beginning of a syllale), cluster (two or more consonants at the beginning of a syllale) or simple (exactly one consonant at the beginning of a syllable). These three classes cover all the possible onsets. We hypothesized that high frequency words are simpler in the sense that it is low in consonant clusters and that simple and null onsets prevail. Table 3 summarizes the ratio of these occurrences.

These results show that the characteristics of high frequency words vs. mid or low frequency words is not in the composition of simple onsets. Simple onsets are by far the greatest proportion of all words in all frequencies. High frequency words are characterized by a relatively large proportion of null onsets and a very low proportion of consonant clusters with respect to low frequency words.

The results might be interpreted as the following. Null and simple onsets are simpler in that they are perceived and produced much more easily than the clusters. Clusters are complex components. They are more difficult to perceive and to produce. Another interpretation is to say that high frequency words are much more constrained phonotactically. In other words, fewer grammar rules are necessary to process high frequency words.

Table 3 also shows that within a population, the cluster onset decreases as the length increases, and in general, the nil onset increases (with the exception of mid frequency words). An instance of simplification seems to occur as the complexity, in terms of length, increases.

Table 3: Word Frequency and Onset: Composition Ratio within Frequency Class and Length (%)

length	type	high	mid	low	total
1	nil	23.75	4.85	3.68	5.02
	cluster	1.25	22.09	30.16	25.01
	simple	75.00	73.07	66.16	69.97
2	nil	40.91	12.43	9.67	10.93
	cluster	0	13.22	16.21	14.91
	simple	59.09	74.35	74.12	74.16
3	nil	-	15.53	13.62	14.27
	cluster	-	10.41	13.39	12.37
	simple	-	74.06	72.99	73.36
4	nil	-	13.39	13.73	13.62
	cluster	-	9.00	11.80	10.88
	simple	-	77.61	74.47	75.51
5+	nil	-	13.11	14.74	14.37
	cluster	-	6.15	7.74	7.37
	simple	-	80.74	77.52	78.26
all	nil	27.45	12.08	11.42	11.77
	cluster	0.98	13.17	15.25	14.37
	simple	71.57	74.75	73.33	73.86

Table 4: Word Frequency end Coda: Composition Ratio within Frequency Class and Length (%)

length	type	high	mid	low	total
1	nil	28.75	5.96	5.62	6.66
	cluster	6.25	10.16	49.80	28.20
	simple	65.00	83.88	44.58	65.13
2	nil	72.73	46.91	43.71	45.15
	cluster	0.00	4.34	12.73	9.20
	simple	27.27	48.75	43.56	45.67
3	nil	-	55.08	54.34	54.60
	cluster	-	4.68	8.23	7.01
	simple	-	40.23	37.43	38.39
4	nil	-	68.96	68.41	68.55
	cluster	-	2.25	3.79	3.24
	simple	-	28.79	27.80	28.05
5+	nil	-	76.23	76.90	76.75
	cluster	-	0.00	1.97	1.51
	simple	-	23.77	21.13	21.74
all	nil	38.24	46.12	50.24	48.53
	cluster	4.90	4.98	12.55	9.59
	simple	56.86	48.90	37.21	41.87

Word Frequency and Coda

The codas (syllable-final consonants) were classified in the same way as above into three classes: nil, cluster and simple. Our hypothesis was similar to the one for the onsets: that the high frequency words over represent nil and simple codas. Table 4 shows the relative distribution by frequency classes. The results indicate that while the hypothesis is true, the pattern of distribution is very different from the onset. The proportion of the clusters among the low frequency words ranges from 50% to 2%, while the comparable statistics for the onsets ranged from 30% to 8%. At the same time, the nil coda ranged from 6% to 77% for the same population, while the onsets ranged from 4% to 15%. Another striking fact is that the simple codas decrease in proportion to length in all frequency classes, in addition to the fact that their proportion for one-syllable length is lower than those for the onsets (except for the mid frequency words). The data on one-syllable length is important because there is no chance for stress resyllabication.

We demonstrated that there are structural differences among words of different frequencies along three dimensions: onset types, coda types, and syllable lengths. We have been able to show that there is a correlation between these properties and word frequencies.

LEXICAL INFORMATION AND LEXICAL ACCESS

There are several ways in which such lexical information can contribute to the lexical access problem in a speech recognition system. For example, syllable length of a word is potentially a very powerful device especially when a word is long. The length constraint was proposed and demonstrated to be effective [1, 3]. However, these proposals centered around phoneme length. The advantage of syllable over phoneme length is that the phoneme insertion and deletion errors can be avoided altogether. The disadvantage is that the cohort size is much larger.

Another possible constraint that can be used is the information on the type of onset. We have been able to identify 68 unique onsets over all the syllables of the complete set of sample words. We saw that the majority of English words favors the CV type of syllables. One might, for example, assign a probability associated with the types of onset prior to identifying the onset itself. It remains to be seen how powerful this constraint might be when this information is used even partially, e.g. at the beginnning of a word.

CONCLUSION

What is the relationship between word frequency and the phonological structure? We examined some of the phonological properties of English words which were not discussed before. We proposed a metric of simplicity to account in part for the structural differences between high and low frequency words. We also suggested that syllabic structural information might be used to organize the lexicon into equivalence classes in a speech recognition system.

REFERENCES

[1] Pisoni, D.B., Nusbaum, H.C., Luce, P.A., and Slowiaczek, L.M. Speech Perception, Word Recognition and the Structure of the Lexicon. *Speech Communication*, 1985, 4, 75-95.

[2] Landauer, T.K., Streeter, L.A. Structural Differences Between Common and Rare Words: Failure of Equivalence Assumptions for Theories of Word Recognition. *Journal of Verbal Learning and Verbal Behavior*, 1973, 12, 119-131.

[3] Huttenlocher, D.P. and Zue, V.W. A Model of Lexical Access Based on Partial Phonetic Information. *Proceedings of ICASSP-84*, 1984, 2.

[4] Kucera, F. and Francis, W. *Computational Analysis of Present Day American English*. Brown University Press, 1967.

[5] Makino, S., Wakita, H. and Applebaum, T.H. Lexical Analysis for Word Recognition Based on Phoneme-Pair Differences. Talk delivered at ASA meeting, Minneapolis. October 1980.

[6] Denes, P.B. On the Statistics of Spoken English. *J. Acoust. Soc. America*, 1963, 35, 6 892-904.

[7] Greenberg, J.H. and Jenkins, J.J. Studies in the Psychological Correlates of the Sound System of American English. *Word*, 1964, 20, 157-177.