FREQUENCY OF OCCURRENCE OF PHONEMES IN CONVERSATIONAL ENGLISH

M. ARDUSSI MINES, BARBARA F. HANSON and JUNE E. SHOUP
Speech Communications Research Laboratory, Santa Barbara

The phoneme identification process of an automatic speech recognition system may be aided through the use of statistics of phoneme occurrence in conversational English. These statistics are also applicable to the fields of linguistics and speech, to teaching English as a foreign language and to speech pathology. In this study a data base containing 103,887 phoneme occurrences taken from casual conversational American English was obtained through interviews of sixteen adult males and ten adult females. The speech was transcribed using a quasi-phonemic system, known as ARPAbet, plus selected phoneme alternates and was analysed with computer assistance to obtain the rank order of phonemes according to frequency of occurrence. Also, the radius of the confidence interval for the observed frequency of occurrence was calculated at the 95% level for each phoneme. The top ten phonemes (in order, / ə, n, t, ɪ, s, r, i, l, d, ɛ /) account for 47% of all the data. As expected, the results of the present study correlate highly with those of one other major study of natural speech. Comparisons show some interesting differences in detail, however, that appear to be attributable to relatively minor variations in the experimental procedures.

INTRODUCTION

One of the major tasks in most automatic speech recognition and speech understanding research is the reliable identification of phonemes, or phonological units, which comprise natural speech strings. During the past two decades a great deal of effort has been devoted to developing and perfecting techniques for acoustic analysis and for automatic segmentation and labeling of phonological units. Although it has been suggested by many that the phoneme identification process can be improved by calling upon the so-called "higher" levels of linguistic knowledge such as syntax, semantics and pragmatics (Broad, 1972; Fry and Denes, 1957; Hyde, 1968; Newell, *et al.*, 1971; Peterson, 1961; Shoup, 1962), it is also important to be able to characterize certain phonological aspects of speech, such as the statistics of phoneme occurrence in natural continuous speech. It would then be possible to predict the likelihoods of phoneme sequences, and to incorporate those probabilities into a speech recognition system (Jelinek, 1976).

The purpose of this paper is to present some basic statistics on the rank order and frequency of occurrence of phonemes in conversational American English. The statistics are based on a large collection of natural speech data (comprising nearly 104,000 phoneme occurrences) which was recorded, carefully transcribed, and analysed with computer assistance. It is hoped that the information provided in this paper will be useful to speech-understanding research and to the furthering of knowledge about continuous speech. It could also be useful to teachers of English as a foreign language and to scientists studying the relation of speech pathologies and certain neurological diseases, such as dysnomia,

to phoneme frequency (Lansdell, Purnell and Laskowski, 1963).

A review of the literature on phoneme frequency of occurrence will be presented. Then the data collection and transcription process for the present study will be described, followed by an analysis of the rank order and frequency of occurrence of the phonemes. Finally, the statistics based on our data will be compared with statistics from other works based on continuous English speech.

REVIEW OF THE LITERATURE

There is one previous study of English phoneme frequencies which uses a data base similar to that used in the present study, i.e., one composed exclusively of casual, connected speech transcribed as actually pronounced (Carterette and Jones, 1974). Carterette and Jones provide and compare separate phoneme-frequency statistics for four different sets of data differentiated by school grade level and by age of the informants. The size of each of their data sets is approximately 45,000 phoneme occurrences.

A number of other phoneme-frequency studies use data taken from other than casual or conversational speech. One study (Hayden, 1950) uses data (65,122 phonemes) perceptually transcribed from six lectures given to foreign students at the University of California. It is likely that the word choice and pronunciation were affected by the fact that the speech was delivered in a "lecture" style rather than "casual" style and was addressed to an audience composed of non-native speakers of English.

A Statistical Linguistic Analysis of American English by Roberts (1965) uses a data base of sentences designed to include words from a "word count" as spoken by a single informant. An earlier work (Voelker, 1937) uses as data the perceptual transcriptions of 5,946 radio announcements. In both studies the style is not conversational and the vocabulary might be expected to vary from that in casual conversational speech.

A study by French, Carter and Koenig (1930) is based on 500 telephone conversations. The transcriptions were done, however, with the authors arbitrarily assigning pronunciations which they thought to be typical of the speech of educated New Yorkers. Thus, the data are not based on actual pronunciations of the original conversations. A feature which makes the results of this study difficult to compare with the results of other phoneme-frequency studies is the fact that articles (e.g., *the, a*), greetings, and profanities were not included. Tobias (1959) subsequently retranscribed the word lists of French, Carter and Koenig, using pronunciations based on Kenyon and Knott (1944), and published the rank order and frequency of occurrence of phonemes based on his retranscriptions.

There have also been studies of the frequency of phoneme occurrence based on perceptual transcriptions of literary materials (poetry, prose and plays) which were read aloud or recited. These include Whitney (1874), Carroll (1952) and Fowler (1957). Each was based on the pronunciations of a single informant.

Other studies have been based on dictionary pronunciations of plays and phonetic readers, cf. Denes (1963), Fry (1974) and Hultzén, Allen and Miron (1964). A study by Dewey (1923) is based on dictionary pronunciations of a variety of materials including

correspondence, newspaper articles, and speeches. A work by Trnka (1935) was based on data sampled from a British English dictionary.

It is interesting to note that in two other articles, which include correlations of rank order of phonemes as given in several of the studies mentioned above, there was shown to be high correlation among all the frequency counts based on connected speech, whether the material represented dictionary pronunciations or actual pronunciations, and whether its original form was spoken or written (Wang and Crawford, 1960; Gerber and Vertin, 1969). These authors found that only Trnka's phoneme count failed to correlate with the others, a fact which they attributed to its having been sampled from a dictionary rather than from continuous texts. A general conclusion which emerges from these correlation studies is that in English the relative frequency of occurrence of phonemes is primarily a function of the structure of the English language itself, and its relation to form (either spoken or written) or to style is very slight.

In addition to the frequency studies of single phonemes, there have been several studies which deal specifically with consonant clusters and also with the general topic of potential phoneme sequences in English. These include Carroll (1952), Denes (1963), Hultzén, Allen and Miron (1964) and Shoup (1964).

DESCRIPTION OF THE DATA

The frequency counts presented in this paper were obtained from a data base of 26 tape-recorded interviews. The data were collected in the following manner:

1. 16 interviews were conducted in a recording room at the Speech Communications Research Laboratory using an Ampex 350 tape recorder with a Turner 2302 microphone.
2. Eight interviews were conducted in the subjects' homes using a Sony Tapecorder, TC-110.
3. Two interviews were conducted under near ideal recording conditions using a sound-treated room at the University of Michigan.

Thus, the acoustic quality of the recordings varies considerably but not to a degree that would greatly affect orthographic or quasi-phonemic transcriptions of the content.

The principal speaker on each tape is the person being interviewed, and only his or her utterances have been transcribed and analysed for phonemic content. The set of utterances obtained from one interviewee is called a discourse. Although a standard questionnaire was used to elicit responses, the interviewer put the subject at ease, interrupted as little as possible, and allowed the subject to talk at length. Topics covered in the questionnaire included games played in childhood, hobbies, favourite pastimes, and films. To further ensure a data base of casual speech, the transcription of each tape was begun after the initial, possibly self-conscious, portion of the interview. There is an average of ten minutes of transcribed data for each speaker. Of the 26 speakers represented in the data, 16 are male and 10 are female. Their geographical backgrounds vary, but none uses a readily identifiable regional dialect. The subjects' ages range from 15 to over 65, although

TABLE 1

ARPAbet symbols, the quasi-phonemic notation system used to transcribe the data.

Phoneme	*Key Word*	*Phoneme*	*Key Word*	*Phoneme*	*Key Word*
i	beat	aɪ	my	ʧ	church
ɪ	bit	ɔɪ	boy	ʤ	judge
e	bait	p	pet	m	met
ɛ	bet	t	ten	n	net
æ	bat	k	kit	ŋ	sing
ɑ	Bob	b	bet	y	yes
ɔ	bought	d	den	w	wit
o	boat	g	get	r	rent
ʊ	book	f	fat	l	let
u	boot	θ	thing	h	hat
ʌ	but	s	sat	ʍ	which
ə	about	š	shut	l̩	battle
ɾ	roses	v	vat	m̩	bottom
ɝ	bird	ð	that	n̩	button
ɚ	neighbor	z	zoo	ɾ	batter
aʊ	down	ž	azure	ʔ	sen'ence [glottal stop]

most (13 males and 7 females) are between 21 and 35 years old. The total transcribed data base consists of 103,887 phoneme tokens, with the average number of phonemes per discourse being 3,996.

The basic notation system used for the transcription of the data is a quasi-phonemic alphabet. The selection of phoneme-like units was made by the Data Base Committee of the Speech Understanding Research Project of the Advanced Research Projects Agency (ARPA) for all participants in the Project. The symbol set was then called the ARPAbet and has been used during the past five years not only by members of that group but by other speech science researchers in the field. The ARPAbet system (shown in Table 1) uses most of the symbols found in the more traditional phonemic inventories, and includes the symbols for the vowel sounds /ʌ/ and /ɾ/, glottal stop /ʔ/, flap /ɾ/, and the syllabics /l̩, m̩, n̩/. These quasi-phonemic units are especially useful for the transcription of the realized forms of informal speech.

TABLE 2

Phoneme alternates. This extension of the basic notation system is an aid for a more accurate phonological representation of casual speech.

1.	ɨ	~	ə	14.	ə n	~	n̩
2.	ɚ	~	ɝ	15.	o	~	ɔ
3.	ɨ	~	ɪ	16.	ʊ	~	ə
4.	ɚ	~	ʊr	17.	ʌ	~	ɔ
5.	e	~	ɛ	18.	ɑ	~	ɔ
6.	ɪ	~	ɛ	19.	t	~	ʔ
7.	ɛ	~	æ	20.	t	~	ɾ
8.	ə	~	ʌ	21.	ʊ	~	u
9.	ɪ	~	i	22.	o	~	ə
10.	ɚ	~	ər	23.	ɪ	~	ə
11.	ɝ	~	ʌr	24.	ə	~	ɑ
12.	əl	~	l̩	25.	ʌ	~	ɑ
13.	əm	~	m̩				

In addition to the basic ARPAbet notation, a set of phoneme alternates was devised (Table 2). This was felt to be a logical and essential extension of the notation system because of the nature of the data (i.e., informal speech). The alternates were used primarily to handle sounds that fell between two phonemes, and therefore could have been classified as either phoneme. Alternates were also used when transcribers could not agree that a given sound should be represented by a certain symbol, again suggesting the possibility of classifying the sound as either phoneme. In all cases where alternates were used in transcriptions, they represent a *single* phonological occurrence. Thus the total phoneme count of the data is not changed by their use. The combination of ARPAbet and alternates, while certainly not a phonetic system, has proven to be very useful for transcription on a "realized phonemic" level. It captures, far more accurately than a standard phonemic system could, many of the phonological details that differentiate informal speech from careful, or citation, forms.

Transcription conventions include the following:

1. All utterances of the main speaker are transcribed, including hesitation sounds, false starts, partial words and gross mispronunciations.
2. Stress is not marked.
3. Pauses or silences are indicated by three dots (i.e., . . .).

4. Phonological mergers, defined as one or more phonemes shared by two words, are written as a unit (e.g., *got/to/* g ɑ r ə /).
5. Totally unintelligible utterances are enclosed by ?_??; This is rare.
6. Transcriptions are checked by at least two other transcribers, making a total of three passes. A consensus is obtained for any disputed sound.
7. Discourses, when fully transcribed both orthographically and phonologically, are coded and typed into a computer for processing and sorting on lexical and phonological levels.

ANALYSIS

The fundamental tasks were to calculate the frequency of occurrence of each phoneme and to rank the phonemes accordingly. An additional task was to calculate the radius (K) of the confidence interval for each phoneme at the 95% level of confidence, using the formula $K = 2(pq/n)^{1/2}$ where p = % of occurrence, q = 100-p, and n = 103,887 (the total number of phoneme occurrences in the data base). The factor 2 corresponds to the selection of a 95% level for the confidence interval. Seventy-two symbols were used to represent all the phoneme occurrences which make up the complete data base.

As shown in Table 3, the top ten phonemes listed in order of frequency are /ə, n, t, ɪ, s, r, i l, d, ɛ/. It is interesting that these account for nearly half of all phoneme occurrences in the data (49,126 occurrences or 47.3%). In the case of the majority of the high ranking phonemes, and even among most phonemes in general, the occurrences are spread throughout many different words and there is no close relationship with just a few particular words. However, the high frequency of a few specific phonemes is due in part to their occurrence in very frequently occurring words:

1. /ə/ – 54% of its occurrences are in realizations of six of the ten most frequent lexical items, *the, uh, a, to, of* and *was.*
2. /ð/ – 50% of its occurrences are in pronunciations of only two words, *the* and *that* both of which are among the top ten lexical items.
3. /ɑɹ/ – 40% of its occurrences are in pronunciations of the pronoun *I*, which is the third most frequent word.

In order to see how or if phoneme frequency might be related to age, sex, level of education, or early place of residence (before age fifteen) of the informant, a rank-order study was made in which the data were subdivided into these categories. Each subdivision comprised at least 5,900 phonemes. There was found to be a high degree of correlation among the rank orders in each category, which suggests that at the phonological level spoken English is not greatly affected by the informant classifications mentioned. These results support general conclusions of previous authors regarding the relationship of the frequency of occurrence of phonemes and the structure of English (Wang and Crawford, 1960; Hultzén, Allen and Miron, 1964). Gerber and Vertin (1969, p. 140) noted that ". . . variations among dialect and form [among several studies of English] were found to

TABLE 3

Frequency of occurrence of phonemes in casual conversational English.
Relative frequency is shown in percent.
Confidence intervals (K) are given at the 95% level of confidence.

Rank	Phoneme	%	No. Occ.	±K	Rank	Phoneme	%	No. Occ.	±K
1	ə	7.30	7582	0.16	37	ɨ ~ ɪ	0.64	665	0.05
2	n	6.72	6983	0.15	38	ǰ	0.56	577	0.05
3	t	5.78	6005	0.14	39	š	0.56	577	0.05
4	ɨ	5.15	5346	0.14	40	ɪ ~ ə	0.50	520	0.04
5	s	4.61	791	0.13	41	ʔ	0.50	520	0.04
6	r	3.87	4017	0.12	42	č	0.50	515	0.04
7	i	3.69	3831	0.12	43	n̩	0.46	475	0.04
8	ɪ	3.64	3777	0.12	44	ɝ	0.42	438	0.04
9	d	3.33	3463	0.11	45	ɪ ~ i	0.41	423	0.04
10	ɛ	3.21	3331	0.11	46	ɨ	0.38	399	0.04
11	ð	3.14	3264	0.11	47	ɛ ~ æ	0.37	389	0.04
12	k	3.10	3223	0.11	48	ɑ ~ ɔ	0.32	331	0.04
13	m	2.99	3108	0.11	49	n ~ n̩	0.29	305	0.03
14	a͡ɪ	2.97	3082	0.11	50	l ~ l̩	0.25	255	0.03
15	w	2.77	2878	0.10	51	o ~ ɔ	0.23	238	0.03
16	z	2.75	2855	0.10	52	ɪ ~ ɛ	0.22	226	0.03
17	æ	2.25	2339	0.09	53	e ~ ʌ	0.20	212	0.03
18	b	1.90	1971	0.09	54	ɪ ~ ə	0.19	198	0.03
19	o	1.85	1918	0.08	55	ʊ ~ ə	0.16	161	0.02
20	p	1.79	1863	0.08	56	ʊ ~ u	0.14	114	0.02
21	ɚ	1.76	1831	0.08	57	l̩	0.12	128	0.02
22	v	1.74	1804	0.08	58	t ~ ʔ	0.12	127	0.02
23	e	1.57	1633	0.08	59	ž	0.09	89	0.02
24	f	1.55	1609	0.08	60	ɔ͡ɪ	0.08	86	0.02
25	ʌ	1.46	1520	0.07	61	t͡ʃ ~ ɾ	0.08	78	0.02
26	ɑ	1.43	1484	0.07	62	m̩	0.05	52	0.01
27	h	1.31	1356	0.07	63	e ~ ɛ	0.04	44	0.01
28	g	1.18	1225	0.07	64	ʍ	0.04	43	0.01
29	u	1.13	1175	0.07	65	ɚ ~ ɝ	0.04	19	0.01
30	y	1.09	1134	0.06	66	m ~ m̩	0.02	18	0.01
31	ŋ	1.08	1125	0.06	67	o ~ ə	0.02	18	0.01
32	ɾ	1.03	1069	0.06	68	ə ~ ɑ	0.02	17	0.01
33	ɔ	0.77	797	0.05	69	ɚ ~ ʊɾ	0.01	10	0.01
34	ʊ	0.76	788	0.05	70	ʌ ~ ɑ	0.007	7	0.005
35	θ	0.70	726	0.05	71	ʌ ~ ɔ	0.004	4	0.004
36	ɑ͡ʊ	0.64	670	0.05	72	ɚ ~ əɾ	0.002	2	0.003

NOTE: Confidence intervals were calculated using the formula $K = 2(pq/n)^{1/2}$ where p = % of occurrence, q = 100-p, and n = 103,887.

have little effect upon phonemic correspondences" and they concluded that ". . . the statistical constraints upon a given language are so severe that variations in time, place and form are of little consequence" even to the point that ". . . the correlations between the spoken forms of the language, regardless of dialect . . . [are] higher than the correlations between spoken and written forms of the same dialect."

The phonemes fall naturally into four subgroups. The number of occurrences within each subgroup and the percentages of the total data are as follows:

Subgroup	Number of Occurrences	Percent
1. Consonants and consonant alternates	60,771	58.5
2. Vowels and vowel alternates	39,582	38.1
3. Retroflex vowels and retroflex vowel alternates	2,300	2.2
4. Syllabics and syllabic alternates	1,234	1.2

The individual members of subgroups three and four are not included in subgroups one and two. Detailed information about the proportionate makeup of each subgroup is shown in Appendix A. Included is the percent of occurrence of each phoneme in each of the three positions within a word (word-initial [WI], word-medial [WM] and word-final [WF]) and the percent of occurrence of each phoneme as a single item (e.g., words such as *I, uh,* or false starts, etc.). It can be noted that almost one-third of the vowel occurrences are either /ə/ or /ɪ/ and that one-half are one of four vowels /ɔ, ɪ, i, ɛ/. Seven consonants (n, t, s, r, l, d, ð/ account for over half of all consonant occurrences. If all the vowel and consonant alternates are taken together as a subgroup, they comprise 5,187 occurrences, 4.8% of the total data. They occur only in the lower half of the rank order list.

A look at Appendix A reveals some unexpected occurrences in the word-initial column: nine instances of word-initial /ž/ and one instance of initial /ŋ/. One of the instances of initial /ž/ occurred in the loanword, *genre,* but the rest appear to be the result of assimilation to the final phoneme of the preceding word, which in all cases was either /z/ or /s/. Examples are *suppose you'd,* realized as /səpoz žud/, and *friends is,* realized as /frɛnz žiz/. Initial /ŋ/ occurred in an irregular pronunciation of the word *oh* (/ŋo/).

Many of the isolated or single occurrences of phonemes are due to stutters and false starts of words. However, the large number of single occurrences of /ə/ (1,606 occurrences) is accounted for by the very frequent use of the pause sound *uh* (the fourth highest ranking lexical item) and of the indefinite article *a* (the fifth highest ranking lexical item). The 1,224 single occurrences of /ɑɹ/ are almost completely accounted for by the pronoun *I,* the third highest ranking lexical item.

The distribution of phonemes in various articulatory categories is shown in Appendix B. Table B–1 shows the consonants classified according to voicing, manner of articulation, and place of articulation. Table B–2 classifies the vowels according to articulation. Tables B–3 and B–4 show the distributions of retroflex vowels and syllabics.

When the phonemes are categorized in this way, the following information concerning

articulatory distribution is made apparent.

1. Almost two-thirds of the consonants (64.73%) in the data are voiced.
2. Nearly one-third of the consonants (29.21%) are plosives. The next two most common manners of articulation are sonorants (19.42%) and nasals (18.46%). This is an interesting distribution when one considers that three of the six plosives, /t, d, k/, are among the top ten ranking consonant phonemes, three of the four sonorants, /r, l, w/, are also ranked in the first ten consonant phonemes, and two of the three nasals, /n, m/, are also among the ten highest. Indeed, /n/ is the highest ranked consonant and second only to /ə/ in overall frequency.
3. The vast majority of consonants are articulated at the front area of the mouth. Dental and alveolar sounds (60.93%) and labial and labiodental sounds (21.58%) comprise over four-fifths of the consonant occurrences.
4. As with the consonants, the majority of vowels are articulated near the front of the mouth. Combining front and central vowels, it can be seen that this category accounts for nearly three-fourths, 72.40%, of the vowels in the data.
5. The most frequently occurring vowels are also articulated in the high and mid-sections of the mouth. Again, nearly three-fourths, 71.25%, of the vowels in the data are high (36.39%), high-mid (2.91%), or mid (31.95%).

COMPARISON WITH OTHER STUDIES

Because the earlier studies cited above indicated that phoneme frequency in English is primarily a function of the structure of English, it could be predicted that the phoneme frequency counts reported here would correlate highly with counts from any of the previous phoneme studies based on English data. (It has been pointed out, however, that this generalization could not have been readily drawn from the previous statistical analyses of "informal" speech.) It was especially felt that a comparison with the data of Carterette and Jones (1974) for adult speakers might show a particularly close correlation since there are many similarities in data collection procedures between the two studies. It should be mentioned that we are not comparing the results of this study with the Carterette and Jones data for the younger age groups, for we only used adults.

Present Study	*Carterette and Jones*
1. Casual, conversational speech in American English obtained by interview.	1. Same, but obtained in group discussions.
2. Adult informants.	2. Same.
3. Both male and female informants.	3. Same.
4. Informants from a variety of dialect areas of the U.S.	4. Same.

5. 26 speakers in 26 interviews.	5. 24 speakers in eight 3-way conversations.
6. Topics: personal values, goals, experiences, games, movies.	6. Topics: personal interests and experiences.
7. Speakers were from a variety of occupational and educational backgrounds, mostly college educated.	7. Speakers were students from elementary college-level psychology classes.
8. 72 *symbols* in phonemic alphabet: 27 consonants, 2 consonant alternates, 16 vowels, 16 vowel alternates, 2 retroflex vowels, 3 retroflex vowel alternates, 3 syllabics, 3 syllabic alternates.	8. *41 symbols:* 14 vowels, 24 consonants plus word mark and sentence mark.
9. 103,887 phoneme occurrences.	9. 48,708 phoneme occurrences.
10. All types of recording conditions.	10. Stereophonic recordings.

Tables 4a and 4b give frequency-of-occurrence information for the top 15 vowels and top 20 consonants from the present data and for the top 14 vowels and top 20 consonants from the data of Carterette and Jones. Ninety-five percent confidence intervals are provided for the phonemes of both data sets and also for the percentage differences. Percentage differences which are statistically significant are indicated with asterisks. When the two sets of data are compared in this way, a high degree of correlation in phoneme rank order and relative frequency of occurrence can be seen. However, looking at the last column in the table, it will be noticed that most of the differences in percent are statistically significant at the 95% level of confidence. In order for the percentages not to be significantly different at this level the two sets of data would have to match much more closely than they do. Assuming it is true that in English the relative frequency of occurrence of phonemes is *primarily* a statistical feature of the language, the differences in phoneme percentages between two sets of data would still be expected to be significant at the 95% level of confidence unless the same notation system and conventions of transcriptions were used, and this was not the case for the two studies being compared. Other factors, such as mode of speech and topics of conversation would also be likely to affect the statistics to some degree.

In Table 4a (Vowels), it should be noted that Carterette and Jones used only 14 vowel symbols, all of which are shown, and the present study used 32 (including alternates), of which 15 are shown. For the data reported here the vowel alternate / ɤ ~ ɪ / (the fifteenth highest ranking vowel) was included in the figure because it was so close to the fourteenth ranking vowel.

Five consonants / n, s, l, d, z / and six vowels (ə, ɪ, i, æ, o, ʊ / rank in corresponding positions between the two studies. Eight phonemes rank in adjacent positions / ɛ, ɑɪ, e, ɑ, k, m, h, ŋ/ and, except for /u, p, y, f, r /, the remaining phonemes differ in only two rank positions.

In some instances differences in rank position become less meaningful when one looks at the percent figures. For example, /w/, which ranks ten in the present study and eight

TABLE 4a

Phoneme frequency data of the present study compared with Carterette and Jones data: *top 15 vowels* of the study. Relative frequency within total data base is given in percent. Confidence intervals (K) are at the 95% level of confidence. Asterisks indicate differences in percent that are significant at the 95% level.

		PRESENT STUDY			CARTERETTE AND JONES			DIFFERENCE	
Rank	Phon.	No. Occ.	Percent Occ.	±K	No. Occ.	Percent Occ.	±K	Percent	±K
1	ə	7582	7.30	0.16	6325	12.99	0.30	*5.69	0.34
2	ɪ	5346	5.15	0.14	2489	5.11	0.20	0.04	0.24
3	i	3831	3.69	0.12	1835	3.77	0.09	0.08	0.15
4	ɛ	3331	3.21	0.11	1551	3.18	0.08	0.03	0.14
5	a͡ɪ	3082	2.97	0.11	1555	3.19	0.08	*0.22	0.14
6	æ	2339	2.25	0.09	1229	2.52	0.07	*0.27	0.11
7	o	1918	1.85	0.08	1139	2.34	0.07	*0.49	0.11
8	e	1633	1.57	0.08	755	1.55	0.06	0.02	0.10
9	ʌ	1520	1.46	0.07	not applicable				
10	ɑ	1484	1.43	0.07	596	1.22	0.05	*0.21	0.09
11	u	1176	1.13	0.07	869	1.78	0.06	*0.65	0.09
12	ɔ	797	0.77	0.05	737	1.51	0.06	*0.74	0.08
13	ʊ	788	0.76	0.05	230	0.47	0.03	*0.29	0.06
14	a͡ʊ	670	0.64	0.05	367	0.75	0.04	*0.11	0.06
15	ɪ͡ɪ	665	0.64	0.05	not applicable				
	ɔ͡ɪ	not applicable			43	0.09	0.02		
TOTAL		36,162	34.82		19,720	40.47			

NOTE: For the confidence interval (K) of the individual studies, $K = 2(pq/n)^{1/2}$ where p = % of occurrence, q = 100-p, and n = total number of phoneme occurrences in data base. For the confidence interval of the differences in percentages between the two studies $K = (K_1^2 + K_2^2)^{1/2}$ where K_1 = present study confidence interval and K_2 = Carterette and Jones confidence interval. All of the vowel occurrences from the Carterette and Jones data, but not all from the present study, are represented.

TABLE 4b

Phoneme frequency data of the present study compared with Carterette and Jones data: *top 20 consonants* in this study. Relative frequency within total data base is given in percent. Confidence intervals (K) are at the 95% level of confidence. Asterisks indicate differences in percent that are significant at the 95% level.

		PRESENT STUDY			CARTERETTE AND JONES			DIFFERENCE	
Rank	Phon.	No. Occ.	Percent Occ.	±K	No. Occ.	Percent Occ.	±K	Percent	±K
1	n	6983	6.72	0.15	3464	7.11	0.23	*0.39	0.27
2	t	6005	5.78	0.14	2248	4.62	0.18	*1.16	0.23
3	s	4791	4.61	0.13	2264	4.65	0.18	0.04	0.22
4	r	4017	3.87	0.12	2806	5.76	0.21	*1.89	0.24
5	l	3777	3.64	0.12	1850	3.80	0.17	0.16	0.21
6	d	3463	3.33	0.11	1827	3.75	0.17	*0.42	0.20
7	ð	3264	3.14	0.11	1354	2.78	0.15	*0.35	0.18
8	k	3223	3.10	0.11	1414	2.90	0.15	*0.20	0.18
9	m	3108	2.99	0.11	1199	2.46	0.14	*0.53	0.18
10	w	2878	2.77	0.10	1397	2.87	0.15	*0.10	0.18
11	z	2855	2.75	0.10	1107	2.27	0.13	*0.48	0.18
12	b	1971	1.90	0.09	877	1.80	0.12	*0.10	0.14
13	p	1863	1.79	0.08	694	1.43	0.11	*0.36	0.14
14	v	1804	1.74	0.08	738	1.52	0.11	*0.22	0.14
15	f	1609	1.55	0.08	692	1.42	0.11	0.13	0.14
16	h	1356	1.31	0.07	793	1.63	0.11	*0.32	0.14
17	g	1225	1.18	0.07	598	1.23	0.10	0.05	0.12
18	y	1134	1.09	0.06	941	1.93	0.12	*0.84	0.14
19	ŋ	1125	1.08	0.06	516	1.06	0.09	0.02	0.12
20	ɾ	1069	1.03	0.06	not applicable				
	ʔ	not applicable			988	2.03	0.13		
TOTAL		57,520	55.37		27,767	57.02			

See NOTE to Table 4a regarding computation of confidence intervals. Notice that only the top twenty consonants from each data set are represented.

in the Carterette and Jones consonant lists, has a 2.77% of occurrence in the data reported here and only a slightly greater percent of occurrence (2.87%) in the Carterette and Jones data. Similar examples are /b, ŋ, ɛ, ɑʊ/.

Some variation in rank position and percentage is obviously due to differences in the transcription systems used. Carterette and Jones did not use /ɾ/, /ɬ/, or any symbols for the syllabics, retroflex vowels, or alternates transcribed in the present study. One example of the effect of this is the difference in the percent figures for /ə/. That the percent of occurrence of /ə/ on our vowel list is only 7.30 compared to a percent of 12.99 on the Carterette and Jones list seems attributable to the fact that in our transcriptions there were 23 symbols for reduced vowel sounds, most of which would probably be transcribed /ə/ by Carterette and Jones. These 23 reduced vowel symbols, as an aggregate, account for 14.27% of our data.

The smaller percent of occurrence figures for the phonemes /n, r, l, d/ in the present data as compared with the Carterette and Jones data are probably due to our having additional symbols for syllabics, retroflex vowels, flap, and alternates. For example, we have available seven possible symbols for retroflex sounds, /r, ɚ, ɝ, ɚ ~ ɝ, ɚ ~ ʊr, ɝ ~ ər, ɝ ~ ʌr/ while Carterette and Jones use only /r/. Pooling the retroflex symbols gives 6.08%, which compares much more closely to Carterette and Jones' 5.76% than our /r/ alone does (3.87%).

The higher frequencies of /w/ and /h/ in the Carterette and Jones data might be explained to some degree by the fact that Carterette and Jones used a combination of these two symbols for the sound which would be transcribed /ʍ/ in our data.

Some differences in rank position and percent of occurrence were possibly caused by differences in topics and modes of speech (resulting in different vocabularies and word frequencies). For example, the fact that the data transcribed by Carterette and Jones were taken from three-way conversations placed the word *you* much higher on their word frequency list than on ours. This significantly affected the number of occurrences of the phonemes /y/ and /u/.

Final Remarks

Using a data base comprised of 103,887 phoneme occurrences taken from casual conversations in English, the frequency of occurrence of each phoneme was calculated and the phonemes were ranked accordingly. The rank listing and frequency counts were analysed and compared with the data of Carterette and Jones (1974). The analysis and comparison show:

1. The top ten phonemes /ə, n, t, ɪ, s, r, i, l, d, ɛ/ account for nearly half of all phoneme occurrences.
2. The high frequency of occurrence of some phonemes is due in part to the fact that they occur in one or more very frequently occurring words, such as /ə/ in *a, uh* and *the*.
3. Frequency of occurrence of phonemes was found to be related only slightly to

the age, sex, level of education or early place of residence of the informant.

4. Vowels and vowel alternates account for 38.1% of the total data, consonants and consonant alternates account for 58.5%, retroflex vowels and retroflex vowel alternates, 2.2%, and syllabics and syllabic alternates, 1.2%.

5. When our phoneme frequency data were compared with the Carterette and Jones data, a very high degree of correlation was found. This is due to similar procedures in data collection and, more importantly, it seems to show that the two studies are representative of spoken English. This would support the findings of Wang and Crawford (1960), Hultzén, Allen and Miron (1964), Gerber and Vertin (1969) and others, who concluded that frequency of occurrence of phonemes is primarily a function of the language as opposed to a function of style.

REFERENCES

BROAD, D.J. (1972). Basic directions in automatic speech recognition. *Int. J. Man-Machine Studies*, **4**, 105-18.

CARROLL, J.G. (1952). Progress report on Project 52, Transitional probabilities of English phonemes, March 15, 1952; Supplemental information of Project 52, October 17, 1952; hectograph materials privately distributed.

CARTERETTE, E.C. and JONES, M.H. (1974). *Informal Speech/Alphabetic and Phonemic Texts with Statistical Analyses and Tables* (Berkeley, California).

DENES, P.B. (1963). On the statistics of spoken English. *J. acoust. Soc. Amer.*, **30**, 892-904.

DEWEY, G. (1923). *The Relative Frequency of English Speech Sounds* (Cambridge).

FOWLER, M. (1957). Herdan's statistical parameter and the frequency of English phonemes, *Studies Presented to Joshua Whatmough* (The Hague).

FRENCH, N.R., CARTER, C.W., Jr. and KOENIG, W. (1930). The words and sounds of telephone conversations. *Bell System Tech. J.*, **9**, 290-324.

FRY, D.B. (1947). The frequency of occurrence of speech sounds in Southern English. *Arch. néerl. de Phon. expér.*, **20**, 103-6.

FRY, D.B. and DENES, P. (1957). On presenting the output of a mechanical speech recognizer. *J. acoust. Soc. Amer.*, **29**, 364-67.

GERBER, S.E. and VERTIN, S. (1969). Comparative frequency counts of English phonemes. *Phonetica*, **19**, 133-41.

HAYDEN, R.E. (1950). The relative frequency of phonemes in general American English. *Word*, **6**, 217-23.

HULTZÉN, L.S., ALLEN, J.H.D. and MIRON, M.S. (1964). *Tables of Transitional Frequencies of English Phonemes* (Urbana, Illinois).

HYDE, S.R. (1972). Automatic speech recognition: a critical survey of the literature. In E.E. David and P.B. Denes (eds.), *Human Communication: A Unified View* (New York), 399-438.

JELINEK, F. (1976). Continuous speech recognition by statistical methods. *Proc. IEEE*, **64**, 532-56.

KENYON, J.S. and KNOTT, T.A. (1944). *A Pronouncing Dictionary of American English* (Springfield, Massachusetts).

LANSDELL, H., PURNELL, J.K. and LASKOWSKI, E.J. (1963). The relation of induced dysnomia to phoneme frequency. *Language and Speech*, **6**, 88-93.

NEWELL, A., BARNETT, J., FORGIE, J., GREEN, C., KLATT, D., LICKLIDER, J.C.R., MUNSON, J., REDDY, R. and WOODS, W. (1971). *Speech-Understanding Systems: Final Report of a Study Group* (Carnegie-Mellon University, Pittsburgh).

PETERSON, G.E. (1961). Automatic speech recognition procedures. *Language and Speech*. **4**, 200-19.

ROBERTS, A.H. (1965). *A Statistical Linguistic Analysis of American English* (The Hague).

SHOUP, J.E. (162). Phoneme selection for studies in automatic speech recognition. *J. acoust. Soc. Amer.*, **34**, 397-403.

SHOUP, J.E. (1964). *The Phonemic Interpretation of Acoustic-Phonetic Data*. Ph.D. dissertation, University of Michigan.

TOBIAS, J.V. (1959). Relative occurrence of phonemes in American English. *J. acoust. Soc. Amer.*, **31**, 631 (Letter).

TRNKA, B. (1935). *A Phonological Analysis of Present-Day Standard English (Studies in English by Members of the English Seminar of the Charles University*, Prague, Vol. 5).

VOELKER, C.H. (1937). A comparative study of investigations of phonetic dispersion in connected American speech. *Arch. néerl. de Phon. expér.*, **13**, 138-57.

WANG, W. S.-Y., and CRAWFORD, J. (1960). Frequency studies of English consonants. *Language and Speech*, **3**, 131-39.

WHITNEY, W.D. (1874). The proportional elements of English utterance. *Proc. Amer. Philol. Assn.*, 14-17.

APPENDIX A

Phoneme Distribution by Subgroup and Word Position

The phonemes can be categorized into four subgroups: consonants and consonant alternates, vowels and vowel alternates, retroflex vowels and retroflex vowel alternates, and syllabics and syllabic alternates. In Tables A–1 through A–4 all the phonemes are ranked within their respective subgroups and the proportionate makeup according to frequency of occurrence of each subgroup is given. Also shown are the actual frequencies of each phoneme in word-initial, word-medial, word-final and single positions.

TABLE A–1

Rank order and frequency of occurrence of consonants and consonant alternates (58.50% of all phonemes) in word-initial (WI), word-medial (WM), word-final (WF), and single (S) positions.

Rank	Phoneme	Percent of Consonants	No. WI	No. WM	No. WF	No. S
1	n	11.49	867	2,861	3,246	9
2	t	9.88	1,242	1,804	2,953	6
3	s	7.88	1,777	1,406	1,473	135
4	r	6.61	588	2,541	886	2
5	l	6.21	896	1,961	918	2
6	d	5.70	872	850	1,731	10
7	ð	5.37	2,945	276	41	2
8	k	5.30	1,108	1,346	766	3
9	m	5.11	1,177	871	1,054	6
10	w	4.74	2,536	333	4	5
11	z	4.70	17	306	2,467	65
12	b	3.24	1,286	646	36	3
13	p	3.07	890	670	298	5
14	v	2.97	190	711	902	1
15	f	2.65	1,016	378	207	8
16	h	2.23	1,250	106	0	0
17	g	2.02	835	292	96	2
18	y	1.87	894	236	0	4
19	ŋ	1.85	1	346	778	0
20	ɾ	1.76	75	600	394	0
21	θ	1.19	417	149	155	5
22	ǰ	0.95	284	182	111	0
23	š	0.95	189	336	48	4
24	ʔ	0.85	82	107	321	9
25	č	0.85	96	149	270	0
26	t ~ ʔ	0.21	0	9	118	0
27	ž	0.15	9	75	5	0
28	t ~ ɾ	0.13	14	36	28	0
29	ʍ	0.07	43	0	0	0
TOTAL		100.00	21,596	19,583	19,306	286

TABLE A–2

Rank order and frequency of occurrence of vowels and vowel alternates (38.1% of all phonemes) in word-initial (WI), word-medial (WM), word-final (WF), and single (S) positions.

Rank	Phoneme	Percent of Vowels	No. WI	No. WM	No. WF	No. S
1	ə	19.15	1,427	2,632	1,917	1,606
2	ɪ	13.51	1,664	3,595	62	25
3	i	9.68	98	1,396	2,299	38
4	ɛ	8.41	688	2,558	77	8
5	a͡ɪ	7.79	241	1,238	379	1,224
6	æ	5.91	717	1,584	35	3
7	o	4.85	204	788	830	96
8	e	4.13	63	1,051	491	28
9	ʌ	3.84	221	1,248	8	43
10	ɑ	3.75	329	1,112	26	17
11	u	2.97	0	597	577	2
12	ɔ	2.01	281	498	18	0
13	ʊ	1.99	11	543	229	5
14	a͡ʊ	1.69	120	441	108	1
15	ɨ ~ ɪ	1.68	104	549	10	2
16	ɨ ~ ə	1.31	68	416	32	4
17	ɨ ~ i	1.07	5	346	72	0
18	ɨ	1.01	66	306	22	5
19	ɛ ~ æ	0.98	205	144	38	2
20	ɑ ~ ɔ	0.84	99	220	11	1
21	o ~ ɔ	0.60	17	219	2	0
22	ɪ ~ ɛ	0.57	87	138	0	1
23	ə ~ ʌ	0.54	57	125	2	28
24	ɪ ~ ə	0.50	32	157	9	0
25	ʊ ~ ə	0.41	1	57	103	3
26	ʊ ~ u	0.36	0	51	93	0
27	ɔɪ	0.22	4	50	32	0
28	e ~ ɛ	0.11	4	32	8	0
29	o ~ ə	0.05	2	11	4	1
30	ə ~ ɑ	0.04	11	0	1	5
31	ʌ ~ ɑ	0.02	2	0	0	5
32	ʌ ~ ɔ	0.01	0	4	0	0
TOTAL		100.00	6,828	22,106	7,495	3,153

TABLE A–3

Rank order and frequency of occurrence of retroflex vowels and retroflex vowel alternates (2.2% of all phonemes) in word-initial (WI), word medial (WM), word-final (WF), and single (S) positions.

Rank	Phoneme	Percent of Retroflex Vowels	No. WI	No. WM	No. WF	No. S
1	ɚ	79.61	1	534	1,124	172
2	ɝ	19.04	9	398	31	0
3	ɚ ~ ɝ	0.83	0	10	9	0
4	ɚ ~ ur	0.43	0	3	7	0
5	ɚ ~ ər	0.09	1	1	0	0
TOTAL		100.00	11	946	1,171	172

TABLE A–4

Rank order and frequency of occurrence of syllabics and syllabic alternates (1.2% of all phonemes) in word-initial (WI), word-medial (WM), word-final (WF), and single (S) positions.

Rank	Phoneme	Percent of Syllabics	No. WI	No. WM	No. WF	No. S
1	n̩	38.49	5	122	105	243
2	n ~ n̩	24.72	0	80	71	154
3	l ~ l̩	20.67	2	71	182	0
4	l̩	10.37	0	36	91	1
5	m̩	4.21	5	4	20	23
6	m ~ m̩	1.54	0	9	6	4
TOTAL		100.00	12	322	475	425

APPENDIX B

Distributions of Articulatory Categories

It is interesting to observe where the phoneme occurrences fall when the sounds are divided into classes according to place and manner of articulation. Table B–1 gives data on the consonants; Table B–2 on the vowels; Table B–3 on the retroflex vowels and Tabel B–4 on the syllabics. In Table B–5 the totals from the previous sections are added together to give the aggregate frequencies for the categories. It should be noticed that in Table B–1, part c, the labial and labiodental phonemes are grouped together, as are the dentals and alveolars. The velars, glottal stop, and /h/ also share one category.

TABLE B-1

Distributions of consonants by (a) voicing, (b) manner of articulation, and (c) place of articulation.

	Frequency	Percent of Consonants	Percent of all Phonemes
(a) *Voicing*			
Voiceless (p, t, k, ʔ, ʍ, f, θ, h, s, š, č, t ~ ʔ)	21,354	35.14	20.56
Voiced (b, d, g, ɾ, m, n, ŋ, v, ð, z, ž, ǰ, l, r, w, y)	39,339	64.73	37.87
Voiceless ~ Voiced (t ~ ɾ)	78	0.13	0.07
	60,771	100.00	58.50
(b) *Manner of Articulation*			
Plosives (p, b, t, d, k, g)	17,750	29.21	17.09
Plosive Glottal (t ~ ʔ)	127	0.21	0.12
Plosive Flap (t ~ ɾ)	78	0.13	0.08
Glottal Stop (ʔ)	519	0.85	0.50
Flap (ɾ)	1,069	1.76	1.03
Fricatives (ʍ, v, f, ð, θ, h)	8,802	14.48	8.47
Sibilants (s, z, š, ž)	8,312	13.68	8.00
Affricates (č, ǰ)	1,092	1.80	1.05
Nasals (m, n, ŋ)	11,216	18.46	10.80
Sonorants (l, r, w, y)	11,806	19.42	11.36
	60,771	100.00	58.50
(c) *Place of Articulation*			
Labial and Labiodental (p, b, m, ʍ, f, v, w)	13,276	21.85	12.78
Dental and Alveolar (t, d, ɾ, n, θ, ð, s, z, l, r, t ~ ɾ)	37,028	60.93	35.64
Alveopalatal (š, ž, č, ǰ, y)	2,892	4.76	2.79
Velar and Glottal (k, g, ʔ, ŋ, h)	7,448	12.25	7.17
Dental Glottal (t ~ ʔ)	127	0.21	0.12
	60,771	100.00	58.50

TABLE B–2

Distributions of vowels by (a) horizontal, and (b) vertical place of articulation.

	Frequency	Percent of Vowels	Percent of All Phonemes
(a) *Horizontal Place of Articulation*			
Front (ɪ, i, ɛ, e, ɪ ~ i, ɪ ~ ɛ, e ~ ɛ)	14,834	37.48	14.28
Front ~ Central (ɪ ~ ɨ, ɛ ~ æ, ɪ ~ ə)	1,252	3.16	1.21
Central (æ, ɨ, ə, ʌ, ɨ ~ ə, ə ~ ʌ)	12,572	31.76	12.10
Central ~ Back (ʊ ~ ə, o ~ ə, ə ~ ɑ, ʌ ~ ɑ, ʌ ~ ɔ)	210	0.53	0.20
Back (u, ʊ, o, ɔ, ɑ, ɑʊ, ɑ ~ ɔ, o ~ ɔ, ʊ ~ u)	7,546	19.07	7.26
Back to Front (ɑɪ, oɪ)	3,168	8.00	3.05
	39,582	100.00	38.10
(b) *Vertical Place of Articulation*			
High (i, ɪ, ɨ, e, ʊ, u, ɪ ~ i, ɪ ~ ɨ, ʊ ~ u)	14,405	36.39	13.87
High ~ Mid (ɪ ~ ɛ, e ~ ɛ, ɪ ~ ə, ɨ ~ ə, ʊ ~ ə)	1,152	2.91	1.11
Mid (ɛ, ə, ʌ, ə ~ ʌ)	12,645	31.95	12.17
Mid ~ Low (ɛ ~ æ, ʌ ~ ɑ, ʌ ~ ɔ, ə ~ o, ə ~ ɑ)	435	1.10	0.42
Low (æ, ɑ, o, ɔ, ɑ ~ ɔ, o ~ ɔ)	7,107	17.95	6.84
Low to High (ɑɪ, ɑʊ, oɪ)	3,838	9.70	3.69
	39,582	100.00	38.10

TABLE B–3

Distributions of the retroflex vowels and their alternates.

Phoneme	Frequency	Percent of Retroflex Vowels	Percent of All Phonemes
ɚ	1,831	79.61	1.76
ɜ˞	438	19.04	0.42
ɚ ~ ɜ˞	19	0.83	0.02
ɚ ~ ʊ r	10	0.43	0.01
ɚ ~ ə r	2	0.09	0.002
	2,300	100.00	2.21

TABLE B–4

Distribution of syllabics and their alternates.

Phoneme	Frequency	Percent of Syllabics	Percent of All Phonemes
n̩	475	38.50	0.46
n ~ n̩	305	24.72	0.29
l̩	128	10.37	0.12
l ~ l̩	255	20.66	0.25
m̩	52	4.21	0.05
m ~ m̩	19	1.54	0.02
	1,234	100.00	1.19

TABLE B–5

Distributions of the pooled categories:
vowel, consonant, retroflex vowel, and syllabic.

	Frequency	Percent of all Phonemes
Vowels and Vowel Alternates	35,744	38.10
Consonants and Consonant Alternates	60,771	58.50
Retroflex Vowels and Retroflex Vowel Alternates	2,300	2.21
Syllabics and Syllabic Alternates	1,234	1.19
	103,887	100.00