SPEAKING IN SHORTHAND - A SYLLABLE-CENTRIC PERSPECTIVE FOR UNDERSTANDING PRONUNCIATION VARIATION

Steven Greenberg

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704, USA
steveng@icsi.berkeley.edu

ABSTRACT

Current-generation automatic speech recognition (ASR) systems model spoken discourse as a linear sequence of words and phones. Because it is unusual for every phone within a word to be pronounced in a standard ("canonical") way, ASR systems often depend on a multi-pronunciation lexicon to match an acoustic sequence with a lexical unit. Since there are, in practice, many different ways for a word to be pronounced, this standard approach adds a layer of complexity and ambiguity to the decoding process which, if modified, could potentially improve recognition performance. Systematic analysis of pronunciation variation in a corpus of spontaneous English discourse (Switchboard) demonstrates that the variation observed is systematic at the level of the syllable. Syllabic onsets are realized in canonical form far more frequently than either coda or nuclear constituents. Prosodic stress also plays an important role in pronunciation. The governing mechanism is likely to involve the informational valence associated with syllable elements, and for this reason pronunciation variation offers a potential window onto the mechanisms responsible for the production and understanding of speech.

****

"The little things are infinitely the most important"
- Arthur Conan Doyle [5]

1. INTRODUCTION

No two speakers utter the same words in precisely the same way, and it is rare for the speech of even the same individual to repeat precisely over the course of a day (or even a lifetime), despite the apparent ease with which the acoustic waveform is linguistically decoded. And as aware as the listener may be of the subtle (and not-so-subtle) acoustic variations in the signal, they rarely interfere with the ability to understand spoken language. On the contrary, such variability, whether it be a consequence of the speaker's gender, age, geographical dialect or emotional state, often provides additional information with which to shape the interpretation of the signal's linguistic message. This seeming paradox of semantic precision and complexity transmitted via an inherently ambiguous and variable acoustic source is a central property of spontaneous speech, one that offers potentially keen insights into the mechanisms underlying pronunciation variation, as well as into the processes germane to the organization and representation of spoken language in general.

The variation of spoken language pronunciation has traditionally received scant attention from the linguistic community except within the context of regional dialects (e.g, [19]) or sociological factors (e.g., [2, 20]). Other sources of pronunciation variation are often attributed to such factors as speaker idiosyncrasies or "economy of effort" [18], with little concerted effort devoted to delineating the specific parameters underlying its generation or linguistic expression (but cf. [23]).

The introduction of large-vocabulary, speaker-independent speech recognition systems has stimulated considerable interest in pronunciation variation since a significant proportion of the performance errors in current-generation systems are likely to be the consequence of such factors. Special-purpose lexica, incorporating some of the most commonly observed variations in word pronunciation, have increased the performance of such systems by a modest amount [3, 6], but not nearly to the level characteristic of human listeners.

One problem with the current multi-pronunciation approach to automatic speech recognition is its emphasis on a phonemic representation for lexical elements. Individual words are represented solely as sequences of phonetic elements, akin to a pronouncing dictionary [19]. In such lexica all elements in the phonetic sequence receive equal weight relative to others, and little attempt is made to provide alternative lexical representations based on organizational units above or below the phone. Within such a monolithic approach lurks potentially dire consequences for recognition performance when things go wrong (as they often do).

Human listeners typically rely on several, if not dozens, of different representational tiers to decode the speech signal during the course of a typical conversation [12, 22]. Variations in the spectrum, speech envelope, fundamental frequency, segmental duration, movement of the lips and jaw, and as well as detailed knowledge of the statistical properties of spoken language are all utilized to deduce the linguistic message embedded in the acoustic signal. As of yet, ASR systems take little advantage of such extra-phonetic information in decoding the speech stream (but cf. [24]) and it is therefore unsurprising that such features have not been systematically investigated with respect to pronunciation variation.

One means by which to rectify this representational imbalance is through systematic analysis of the *phonetic* properties of spontaneous speech in an effort to ascertain precisely how much of the variation in spoken language pronunciation can be accounted for on the basis of such narrow linguistic criteria. Such knowledge can then be used to delineate the extra-phonetic factors involved in the patterning of pronunciation variation.

2. THE PHONETIC TRANSCRIPTION OF SPONTANEOUS SPEECH

The Switchboard corpus [11] is currently one of the primary sets of material with which to assess the reliability and accuracy of automatic speech recognition for spoken language. In contrast to such corpora as TIMIT [30] or Wall Street Journal [10], in which a speaker reads prepared written material, Switchboard comprises informal, unscripted, telephone dialog on a wide range of topics spoken by individuals of both genders and encompassing a wide range of dialectal variation, age and educational background.

Four hours of material from this corpus was phonetically labeled by linguistically trained, highly experienced transcribers over the course of a year's time [15] and made

N	phonetic transcription					N	Phonetic Transcription					N	Phonetic Transcription					
82	ae	n				2	ax	nx				1	ix	dcl	d			
63	eh	n				2	q	ae	ae	n	d	1	ae	eh	n			
45	ix	n				2	q	ix	n			1	hh	n				
35	ax	n				2	ix	n	dcl	d		1	ix	n	t			
34	en					2	ih					1	ae	ax	n	dcl	d	
30	n					2	eh	eh	n			1	iy	eh	n			
20	ae	n	dcl	d		2	q	eh	nx			1	m					
17	ih	n				2	ix	d	n			1	ae	ae	n	d		
17	q	ae	n			1	eh	m				1	nx					
11	ae	n	d			1	ax	n	dcl	d		1	q	ae	ae	n		
7	q	eh	n			1	aw	n				1	q	ae	ae	n	dcl	d
7	ae	nx				1	ae	q				1	q	ae	eh	n	dcl	d
6	ae	ae	n			1	eh	dcl				1	q	ae	ih	n		
6	ah	n				1	ah	nx				1	aa	n				
5	eh	nx				1	ae	n	t			1	q	ae	n	d		
4	uh	n				1	eh	d				1	?	nx				
4	ix	nx				1	ah	n	dcl	d		1	q	ae	n	q		
4	q	ae	n	dcl	d	1	ey	ih	n	dcl	d	1	eh	n	m			
3	eh	n	d			1	ae	ix	n			1	q	eh	en	dcl		
3	q	ae	nx			1	ae	nx	ax			1	eh	ng				
3	eh					1	ax	ng				1	q	eh	n	q		
2	ae	n	dcl			1	ay	n				1	em					
2	ae					1	ih	ah	n	d		1	q	eh	ow	m		
2	ax	m				1	ae	hh				1	q	ih	n			
2	ax	n	d			1	ih	ng				1	q	ix	en			
2	ae	eh	n	dcl	d	1	ix					1	er					
2	eh	n	dcl	d		1	ae	n	d	dcl								

Table 1. 80 pronunciation variants of the word "and" from the Switchboard Transcription Corpus. The variants are listed in order of their frequency. The phonetic symbols are from a transcription system based on Arpabet. The segment [q] denotes a glottal stop. The symbol set and transcription methods are described in [15].

available through the Johns Hopkins' Center for Language and Speech Processing to the ASR community for developing future-generation recognition systems and for improving current methods for modeling pronunciation variation [3, 25, 28].

Three quarters of the material was labeled at the phone level and segmented at the syllabic boundaries. The remainder (72 minutes) was labeled and segmented at the phonetic segment level, but also segmented at the syllabic level to insure compatibility with the three other hours of material. Both portions of the corpus were transcribed using a custom-designed variant of the Arpabet phonetic symbol set [30]. A small portion of this material was transcribed in common by all of the transcribers in order to ascertain the interlabeler agreement (ca. 75-80%). A detailed description of this project is provided in [15] and various statistical analyses of the phonetic transcription material are described in [14] and [16]. The transcription material are available via the World Wide Web [27].

Such analyses provide striking testimony for the ephemeral quality of the phonetic segment at the lexical level (e.g., Table 1) by virtue of the large proportion of phones that either change their identity (i.e., substitutions) or are altogether missing in action (i.e., deletions). Occasionally, entire words are swallowed whole, with only a short pausal junction marking their (perceived) location [15]. At the level of the phone, pronunciation variation is a difficult beast to comprehend or tame. It is for this reason that the focus of the current analyses is aimed at a higher level, that of the syllable.

3. ANATOMY OF A SYLLABLE

The syllable can be likened to a linguistic "wolf" in phonetic clothing. It is perhaps the "sheepish" nature of its outer lining that has led many to believe that it is but a mere sequence of phones, and therefore can be simulated via a multi-phone (i.e., tri- or quinta-phone) approach to ASR. What distinguishes the syllable from this phonetic exterior is its structural integrity, grounded in both the production and perception of speech, and wedded to the higher tiers of linguistic organization.

The syllable is structurally divisible into three parts, the onset, nucleus and coda (the nucleus and coda, taken together, are often referred to as the "rime"). Although many syllables contain all three elements, a significant proportion contain only one or two. With rare exception, when a single component is present, it is the nucleus. Generally (though not always), the nucleus is vocalic, while the onset and coda are typically consonantal in form. For example, the word (and syllable) "cat," can be phonetically represented as three distinct segments, [k] [ae] [t], each associated with a specific structural element of the syllable. The [k] is the onset, followed by the nucleus [ae], with the coda, [t], bringing up the rear. A second means by which to characterize the syllable is in terms of its consonantal-vocalic composition. Within this framework [k ae t] is classified as a CVC syllable (at least in terms of its canonical representation - see below).

If all words were spoken in canonical form there would be little reason to prefer the syllable over some other form of multi-phone representation for automatic speech recognition.

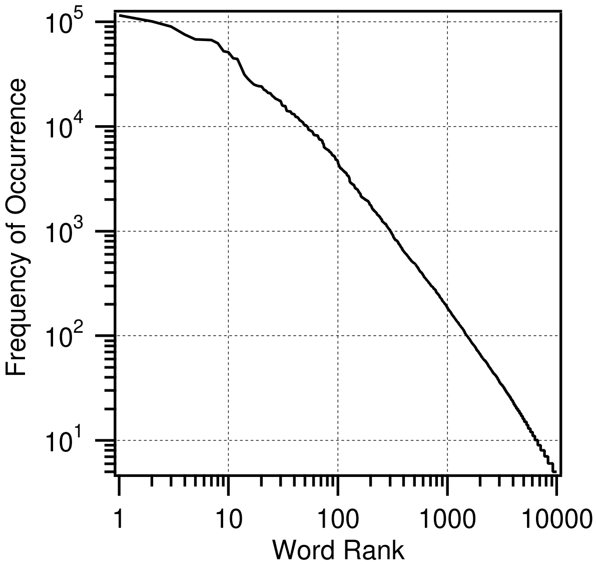

Figure 1. The frequency of occurrence for the 10,000 most frequent words in the Switchboard corpus, organized in rank order of frequency. Total number of distinct words in the corpus is 25,923. Reprinted from [14].

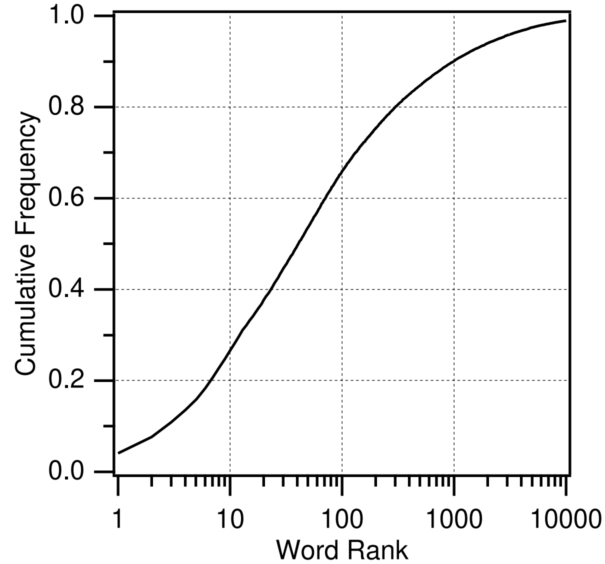

Figure 2. Cumulative frequency of occurrence as a function of word frequency rank for the 10,000 most frequent lexical items in the Switchboard corpus. Reprinted from [14].

However, the pattern of pronunciation variation observed in spontaneous speech is far from egalitarian. The onset portion of the syllable is generally a "survivor," maintaining its canonical identity regardless of speaking conditions, while the nucleus is a "chameleon," capable of assuming a wide range of vocalic appearances. And the coda often gets no respect, as a consequence of its disposable quality. Why should this be so?

The answers are likely to be various and to reside at several levels of analysis, from the acoustic-phonetic and auditory at the lower end, to the lexical and prosodic at the upper end of linguistic function. This veil of representational tiers is governed by the requisites of information transmission and bound into a coherent, functioning whole by virtue of the syllable.

We shall first examine some of the statistical properties of the Switchboard corpus from the perspective of both the syllable and the word in order to gain some insight into the linguistic foundations of pronunciation variation before turning our attention to how these linguistic representations might be encoded at the level of the auditory pathway and higher cortical centers of the brain.

4. ALL WORDS, GREAT AND SMALL

Since the days of Dewey [4] and Zipf [29] it has been known that words differ greatly in terms of their frequency of occurrence in written language. French and colleagues [8] were the first to demonstrate a comparable pattern for spoken English dialog.

A frequency analysis of the Switchboard lexicon illustrates the magnitude of this effect. The most common words occur far more frequently than the least (Figure 1). The ten most frequent words account for approximately 25% of all the lexical instances in the corpus. One hundred words account for fully 66% of the individual tokens (Figure 2). A perusal of these most frequently occurring words (Table 2) indicates that most come from the so-called "closed" or "function" class words such as pronouns, articles, conjunctions and modal/auxiliary verbs. Many of the remainder stem from just a few basic nominal, adjectival or verbal forms. Clearly, mastery of these 100 most common words goes a long way towards facilitating comprehension of spoken discourse. The perceptual criteria for recognizing such common words are likely to be very different from those associated with their infrequent lexical counterparts.

5. THE SYLLABIC REPRESENTATION OF THE LEXICON

Although a mere list of common words does not provide sufficient data with which to interpret the speech signal, it could be used in conjunction with other knowledge sources to prune the number of likely lexical alternatives. One potentially useful representation is at the level of the syllable.

The 30 most common words in the Switchboard corpus are monosyllabic (Table 2), and of the 100 most frequent lexical items only ten are not (and all of these contain but two syllables). This decided lexical preference for syllabic brevity among the most frequently occurring words is largely representative of the corpus as a whole. Although only 22% of the Switchboard lexicon is composed of monosyllabic forms, fully 81% of the corpus tokens are just one syllable in length (Table 3). The portion of the lexicon consisting of three or more syllables (38%) rarely gets to strut its stuff in spontaneous language, accounting for less than 5% of the spoken instances (Table 3). This statistical skew towards short syllabic forms provides a potentially powerful interpretative constraint on the decoding of the speech stream.

For the three hundred most frequently occurring words the cumulative statistical distribution is remarkably similar to their syllabic counterparts (Figure 3). Thus, word and syllable segmentation and recognition reduce to virtually the same thing for this most favored portion of the lexicon.

	Word	N	#Pr.	Most Common Pronunciation	% Tot
1	I	649	**53**	ay	**53**
2	and	521	**87**	ae n	**16**
3	the	475	**76**	dh ax	**27**
4	you	406	**68**	y ix	**20**
5	that	328	**117**	dh ae	**11**
6	a	319	**28**	ax	**64**
7	to	288	**66**	tcl t uw	**14**
8	know	249	**34**	n ow	**56**
9	of	242	**44**	ax v	**21**
10	it	240	**49**	ih	**22**
11	yeah	203	**48**	y ae	**43**
12	in	178	**22**	ih n	**45**
13	they	152	**28**	dh ey	**60**
14	do	131	**30**	dcl d uw	**54**
15	so	130	**14**	s ow	**74**
16	but	123	**45**	bcl b ah tcl t	**12**
17	is	120	**24**	ih z	**50**
18	like	119	**19**	l ay kcl k	**46**
19	have	116	**22**	hh ae v	**54**
20	was	111	**24**	w ah z	**23**
21	we	108	**13**	w iy	**83**
22	it's	101	**14**	ih tcl s	**20**
23	just	101	**34**	jh ix s	**17**
24	on	98	**18**	aa n	**49**
25	or	94	**23**	er	**36**
26	not	92	**24**	m aa q	**24**
27	think	92	**23**	th ih ng kcl k	**32**
28	for	87	**19**	f er	**46**
29	well	84	**49**	w eh l	**23**
30	what	82	**40**	w ah dx	**14**
31	about	77	**46**	ax bcl b aw	**12**
32	all	74	**27**	ao l	**24**
33	that's	74	**19**	dh eh s	**16**
34	oh	74	**17**	ow	**61**
35	really	71	**25**	r ih l iy	**45**
36	one	69	**8**	w ah n	**78**
37	are	68	**19**	er	**42**
38	right	61	**21**	r ay	**28**
39	uh	60	**16**	ah	**41**
40	them	60	**18**	ax m	**23**
41	at	59	**36**	ae dx	**8**
42	there	58	**28**	dh eh r	**22**
43	my	58	**9**	m ay	**66**
44	mean	56	**10**	m iy n	**58**
45	don't	56	**21**	dx ow	**14**
46	no	55	**8**	n ow	**77**
47	with	55	**20**	w ih th	**35**
48	if	55	**18**	ih f	**41**
49	when	54	**18**	w eh n	**31**
50	can	54	**28**	kcl k ae n	**15**
51	then	51	**19**	dh eh n	**38**
52	be	50	**11**	bcl b iy	**76**
53	as	49	**16**	ae z	**18**
54	out	47	**19**	ae dx	**22**
55	kind	47	**17**	kcl k ax nx	**21**
56	becaue	46	**31**	kcl k ax z	**15**
57	people	45	**21**	pcl p iy pcl l el	**44**
58	go	45	**5**	gcl g ow	**83**
59	got	45	**32**	gcl g aa	**15**
60	this	44	**11**	dh ih s	**47**
61	some	43	**4**	s ah m	**48**
62	i'm	42	**9**	q aa m	**26**
63	would	41	**16**	w ih dcl	**29**
64	things	41	**15**	th ih ng z	**52**
65	now	39	**11**	n aw	**69**
66	lot	39	**9**	l aa dx	**47**
67	had	39	**19**	hh ae dcl	**24**
68	how	39	**11**	hh aw	**53**
69	good	38	**13**	gcl g uh dcl	**27**
70	get	38	**20**	gcl g eh dx	**13**
71	see	37	**6**	s iy	**80**
72	from	36	**10**	f r ah m	**28**
73	he	36	**7**	iy	**39**
74	me	35	**5**	m iy	**87**
75	don't	35	**21**	dx ow	**14**
76	their	33	**19**	dh eh r	**25**
77	more	32	**11**	m ao r	**56**
78	it's	31	**14**	ih tcl s	**20**
79	that's	31	**20**	dh eh s	**16**
80	too	31	**6**	tcl t uw	**60**
81	okay	31	**17**	ow kcl k ey	**45**
82	very	30	**11**	v eh r iy	**36**
83	up	30	**11**	ah pcl p	**34**
84	been	30	**11**	bcl b ih n	**51**
85	guess	29	**8**	gcl g eh s	**42**
86	time	29	**8**	tcl t ay m	**62**
87	going	29	**21**	gcl g ow ih ng	**13**
88	into	28	**20**	ih n tcl t uw	**14**
89	those	27	**12**	dh ow z	**42**
90	here	27	**11**	hh iy er	**25**
91	did	27	**13**	dcl d ih dx	**23**
92	work	25	**8**	w er kcl k	**66**
93	other	25	**14**	ah dh er	**26**
94	an	25	**12**	ax n	**28**
95	I've	25	**7**	ay v	**46**
96	thing	24	**9**	th ih ng	**52**
97	even	24	**7**	iy v ix n	**40**
98	our	23	**9**	aa r	**33**
99	any	23	**11**	ix n iy	**23**
100	I'm	23	**9**	q aa m	**26**

Table 2. Pronunciation variability for the 100 most common words in the phonetically segmented portion of the Switchboard Transcription Corpus. "N" is the number of instances each word appears in the 72-minute corpus. "#Pr." is the number of distinct phonetic expressions for each word. "%Tot" is the percentage of the total number of pronunciations accounted for by the single most common variant. The phonetic representation is derived from a variant of the Arpabet orthography. Further details concerning both the pronunciation data and the transcription orthography may be found in [15]. Reprinted from [14].

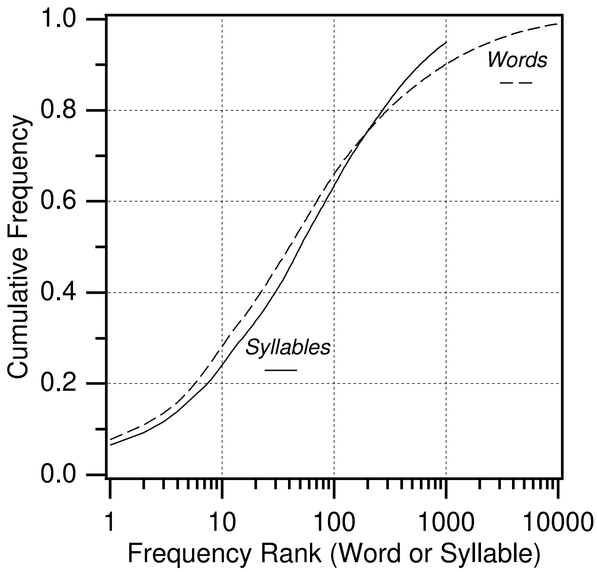

Figure 3. The cumulative frequency of syllables in the entire Switchboard corpus as a function of syllable frequency rank compared with the cumulative frequency of occurrence for words in the same corpus. Reprinted from [14].

#Syllables	*Usage (%)*	*Lexicon (%)*
1	81.04	22.39
2	14.30	39.76
3	3.50	24.26
4	0.96	9.91
5	0.18	3.21
6	0.02	0.40

Table 3. The proportion of words consisting of n-syllables for the entire Switchboard corpus (i.e., tokens or "usage") and lexicon (i.e., type). Comparable data from a telephone dialog corpus study performed in the 1920's [8] shows a virtually identical frequency pattern as a function of syllabic length for lexical items. Reprinted from [14].

6. THE INNER LIFE OF A SYLLABLE

Many languages of the world (such as Japanese and those of the Malayo-Polynesian family) possess a relatively transparent ("simple") syllable structure consisting of just several canonical forms. Most of the syllables in such languages contain just two phonetic segments, typically a consonantal onset followed by a vocalic nucleus (CV). The remaining syllabic forms are generally of the V (nucleus) or VC (nucleus+coda) variety. Such "syllable-timed" languages tend to exhibit an agglutinative grammatical morphology and are thought to possess a relatively even tempo (but see Figure 7).

In contrast, English and German (as well as many other Indo-European languages) possess a more highly variegated syllable structure by virtue of incorporating "complex" patterns into their syllabic repertoire. In such forms, the onset and/or coda elements contain two or more consonants, resulting in thousands of distinct syllabic entities (a consequence of the combinatorial potential of consonantal sequences) and tend to exhibit either an inflectional or synthetic (but rarely an agglutinative) morphology. Such languages tend to informationally highlight (i.e., "stress") a certain proportion of syllables via selective lengthening of segmental durational, resulting in a higher variability of syllable duration than observed among the syllable-timed languages.

The distinction between syllable- and stress-timed languages is illustrated through a comparison of the syllabic properties of Japanese (Table 4) and English (Table 5).

The most salient property shared in common by English and Japanese is the preference for CV syllabic forms in *spontaneous* speech. Nearly half of the forms in English, and over 70% of the syllables in Japanese are of this variety. There is also a substantial proportion of CVC syllables in the spontaneous speech of both languages.

Japanese and English contrast principally in terms of the proportion of complex syllables. In English nearly 15% of the syllabic elements are of the complex variety, containing clusters of two or more consonants, while less than 4% of the Japanese forms are of this type (and in the *canonical* version of the Japanese syllable, the mora, there is no provision for consonantal clusters whatever). Such a distinction is of potential significance in considering the sources of pronunciation variation.

7. THE SYLLABIC BASIS OF PRONUNCIATION

The importance of the syllable as an organizational unit of spoken language becomes manifest when considering pronunciation variation. In spontaneous speech the phonetic realization often differs markedly from the canonical, phonological representation. Entire phone elements are frequently dropped or transformed into other phonetic segments. These patterns of deletions and substitutions appear rather complex and somewhat arbitrary when analyzed at the level of the phonetic or phonological segment. However, this variation becomes systematic when placed within the framework of the syllable.

Several principles of pronunciation variation can be discerned for spontaneously spoken English from analyses of the transcription corpus, as illustrated in Tables 5-8:

(1) *Syllable onsets are generally preserved.*

The phonetic realization of syllabic onsets tends to approximate the canonical (i.e., be "preserved") for most lexical instances and to a far greater degree than nuclear and coda elements. This preference for the canonical is particularly marked for instances of complex onsets containing two or more consonantal segments, and is most easily discerned in the absence of a (canonical) syllabic coda, as in the case of CCV and CCCV syllabic forms. For example, the proportion of CCV forms in the canonical lexicon is 2.6%, but rises to nearly double this quantity in terms of their phonetic realization.

2) *Coda elements are often dispensed with.*

The coda element is often deleted or transformed into a segment that is phonetically homo-organic with that of the following syllable's onset (i.e., it is assimilated). The proportion of syllables classified as of the canonical CVC (31.6%) variety drops by nearly a third when classified on the basis of phonetic pronunciation (22.1%), indicative of coda deletion. An even greater decline in the realization of the canonical coda element is observed for the VC, VCC and CVCC syllabic forms. This reduction in the proportion of syllables that are phonetically realized in

Syllable	n	%
CV	3238	60.4
CVV	626	11.7
CVC	961	17.9
CVVC	71	1.3
VC	64	1.2
VVC	6	0.1
V	154	2.9
VV	29	0.5
CCV	89	1.7
CCVV	72	1.3
CCVC	19	0.4
CCVVC	8	0.1
other	21	0.3
Total	5358	100.0
Mora		
V	1148	15.3
CV	5589	74.7
CjV	182	2.4
N	384	5.1
Q	183	2.4
Total	7486	100.0

Table 4. Frequency of occurrence of syllabic and moraic forms in Japanese. Data are based on manual phonetic transcription of 15 minutes of spontaneous Japanese speech recorded over the telephone. In Japanese, each vocalic element is a separate mora, but often adjacent vocalic morae coalesce into a dipthongal nucleus (VV forms). Such data are discussed further in [1], from where this table is adapted.

their canonically complex coda form is accompanied by a corresponding increase in the number of syllables that are realized without coda. The proportion of CV syllables is 47.2% of the corpus, despite the fact that only 34% of the canonical (phonological) forms are of this variety. And the proportion of V syllables (containing a nucleus only) is 11.2%, even though but 6.3% of the canonical instances are of this form. The complexity of the coda, therefore, has little or no impact on the likelihood of canonical pronunciation.

(3) *The nucleus often deviates from the canonical.*

The nucleus is the syllable's "bedrock," forming its core, and is virtually always vocalic in nature. Thus, any deviation from the canonical is likely to preserve the vocalic form of the nucleus, and therefore such departures are likely to be substitutions (in contrast to those of the coda, which tend to be deletions).

Syllable Type	*Lexicon(%)*	*Corpus(%)*	*Phn. Tr(%)*
CV	36.2	34.0	47.2
CVC	28.8	31.6	22.1
VC	5.3	11.7	4.8
V	4.8	6.3	11.2
Subtotal	*75.1*	*83.6*	*85.3*
"Complex"			
CVCC	7.3	6.3	2.9
VCC	0.5	4.3	0.5
CCV	7.4	2.6	5.1
CCVC	5.0	2.2	2.5
CCVCC	2.2	0.6	0.4
CVCCC	1.0	0.4	0.2
CCCVC	0.5	<0.1	0.1
CCCV	0.4	<0.1	0.3
CCVCCC	0.3	< 0.1	< 0.1
CCCVCC	0.2	< 0.1	< 0.1
VCCC	< 0.1	< 0.1	< 0.1
CCCVCCC	< 0.1	< 0.1	< 0.1

Table 5. The relative frequency of occurrence for various syllable types in both the lexicon and spoken usage of the Switchboard corpus. The data are derived from canonical pronunciations of dictionary sources, and are compared with the syllable structure for actual pronunciation derived from phonetic transcription (Phn. Tr.). Reprinted from [14].

(4) *The likelihood of canonical expression percolates through the syllable.*

The probability of canonical pronunciation for a given constituent is influenced, to a certain degree, by the pronunciation of the elements in the syllable. Thus, the probability of the nucleus or coda being pronounced canonically is higher when the onset is also articulated in the standard manner. Furthermore, the coda is more likely to be pronounced in canonical fashion if the nucleus is as well, and vice versa. Such a pattern of pronunciation variation implies that the specific mechanism responsible for crafting pronunciation looks beyond the individual constituent and almost surely reflects control at the syllable or supra-syllabic level. The factors potentially governing this syllabic linkage are discussed in Sections 8-10.

(5} *The linguistic factors governing pronunciation variation are likely to reflect exceedingly high-level processing*

Specific patterns of pronunciation probably reflect the information valence of the utterance and is indicative of the speaker's projection of the listener's internal knowledge model.

Syllable Constituent	All Instances	Percent Canonical
Onset (total)	39,221	81
Simple [C]	32,853	79
Complex [CC(C)]	6,368	88
Nucleus	49,185	63
Coda (total)	32,408	61
Simple [C]	20,178	61
Complex [CC(C)]	12,230	60
Total (O+N+C)	120,814	68

Table 6. The frequency with which the phonetic pronunciation corresponds to the lexicon's canonical pronunciation, as a function of syllabic constituent. The onset element is far more likely to be preserved in canonical form than either the nucleus or the coda.

8. THE ROLE OF PROSODIC STRESS IN PRONUNCIATION VARIATION

8.1 Durational Properties of Syllables

The relation between prosodic stress and pronunciation variation can be illustrated in several different ways. We first examine the durational properties of syllables in both English and Japanese, and apply these insights to understanding the pattern of pronunciation variation in a stress-timed language (such as English), where syllabic duration plays an important role in the prosodic delineation of semantically important words and syllables [21].

The distribution of syllabic duration for four hours of spontaneous English discourse is illustrated in Figure 4. Of interest is the broad dispersion of syllable lengths (s.d. = 103 ms) and the extended tail of the right-hand portion of the distribution. Approximately 15% of the syllables are longer than 300 ms. Most of this population is likely to be prosodically stressed (see below, Figure 5 and Table 9).

As a consequence of this durational distribution the probability of a shorter syllable following a longer one (and vice versa) is quite high. Such a pattern of durational relations follows directly from the shape of the statistical distribution and does not imply any form of conscious alternation between long and short syllables (as some forms of metrical phonological theory [e.g., 17] would imply). The nature of this process is illustrated in Figure 5, which plots the data of Figure 4 in terms of durational relation of successive syllables. Syllables whose durations are relatively close to the mean of the distribution will tend to be followed and preceded by syllables that do not differ all that much in length. Syllables appreciably longer or shorter than the mean tend to be bracketed by syllables whose durations are significantly different. However, the conditional dependence of duration on prior and following syllable length is remarkably even [15]. The only basis for the skew in odds is the sheer weight of numbers in the distribution's core.

Syl	Con	Onset			Coda		
	Can	+	-	+/-	+	-	+/-
Nuc	+	.567	.109	**.676**	.398	.230	**.628**
Nuc	-	.226	.098	**.324**	.212	.160	**.372**
	+/-	**.793**	**.207**		**.610**	**.390**	
Co	+	.498	.097	**.595**			
Co	-	.317	.088	**.405**			
	+/-	**.815**	**.185**				
Ri	+	.463	.090	**.553**			
Ri	-	.330	.117	**.447**			
	+/-	**.793**	**.207**				

Table 7. The frequency with which different constituents of a syllable are phonetically realized as the canonical form. The frequencies associated with each 2x2 matrix sum to 1. For example, the presence of a canonical onset associated with a canonical nucleus occurs for 56.7% of all instances of syllables containing both an onset and a nucleus. The frequency with which an onset is canonically realized is 0.793 across all forms of nuclear realizations. The frequency with which a nucleus is realized as the canonical form *for the same syllabic population* is .676, and so on. Abbreviations - Can = Canonical, Co = coda, Nuc = nucleus, Ri = Rime. Canonical pronunciations are denoted by a '+' and the non-canonical variety by a '-'.

P|Q	P	On	Nuc	Co	Ri
	Can	**P(+)**	**P(+)**	**P(+)**	**P(+)**
Q					
On	**P(+)**	**.794**	.715	.611	.584
On	**P(-)**		.527	.522	.434
Nuc	**P(+)**	.839	**.628**	.634	
Nuc	**P(-)**	.699		.430	
Co	**P(+)**	.837	.713	**.558**	
Co	**P(-)**	.782	.520		
Ri	**P(+)**	.838			**.553**
Ri	**P(-)**	.738			

Table 8. The conditional probabilities associated with whether a specific syllabic constituent is realized as canonical (+) or non-canonical (-), conditioned on the canonical status of the other constituents within the same syllable. The probability that a specific syllabic constituent is realized canonically *for the specific group of syllables* is indicated in bold-face type along the diagonal. Abbreviations as in Figure 7.

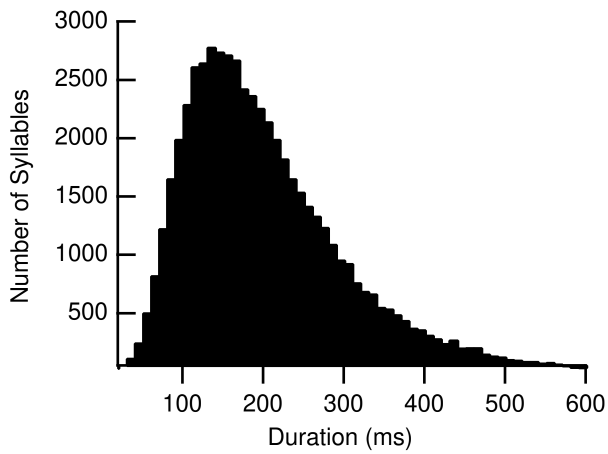

Figure 4. Frequency distribution of syllables from a corpus of spontaneous English discourse (Switchboard). Durations are derived from manual segmentation of syllabic boundaries by phonetically trained individuals. The mean of the distribution is 200.5 ms and the standard deviation is 103 ms. N = 56,400 syllables.

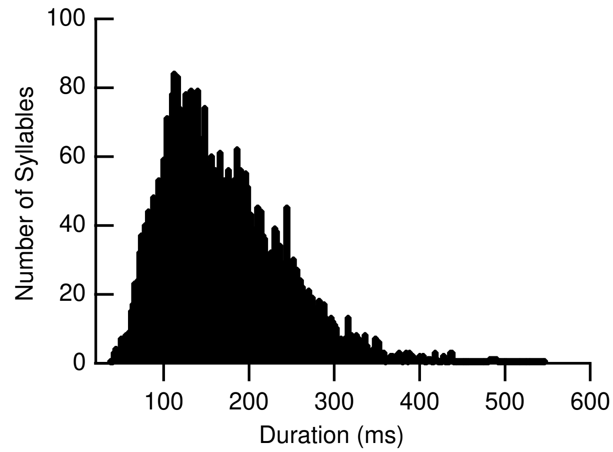

Figure 6. Frequency distribution of syllables from a corpus of spontaneous Japanese speech (OGI-TS). Durations were derived from manual segmentation of syllabic boundaries by a phonetically trained, native speaker of Japanese. The mean of the distribution is 166 ms and the standard deviation is 73 ms. N = 5,358 syllables. Reprinted from [1].

Japanese provides an instructive contrast with English. This language has traditionally been considered a language of even-tempo, with relatively little variation in the length of the syllable. Statistical analysis of the syllabic durations of a 15-minute portion of a spontaneous Japanese corpus calls this assumption into question (Figures 6 and 7). Over much of the course of the distribution, the duration of Japanese syllables is as variable as English.

However, Japanese lacks the expansive tail in the right-hand branch of the distribution characteristic of English syllable durations. This relatively subtle feature constitutes perhaps the primary statistical distinction between a stress- and syllable-timed language. The relative paucity of syllables with complex onsets and codas in Japanese may also be related to this property. Thus, it is likely that prosodic stress in a language, such as English, is tied to the presence of such complex syllabic constituents.

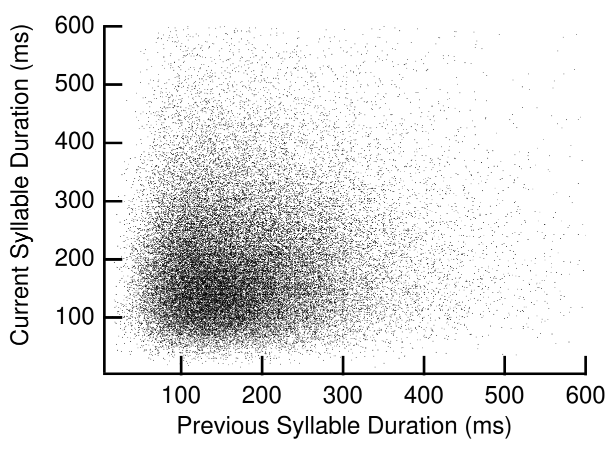

Figure 5. Conditional dependence of syllable duration between successive syllables using the same portion of the Switchboard corpus as in Figure 4. N = 47,061 syllable pairs. Adapted from [15].

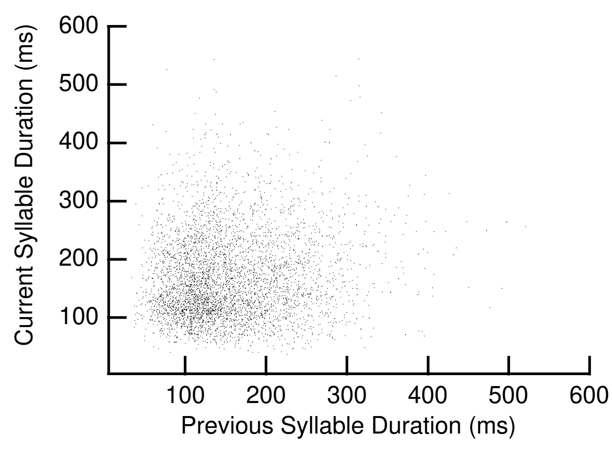

Figure 7. Conditional dependence of syllabic duration between successive syllables using the same portion of the OGI-TS corpus as in Figure 6. N = 4,674 syllable pairs. Reprinted from [1].

8.2 The Acoustic Bases of Prosodic Stress

The relation between prosodic stress and syllabic duration is well established in the experimental literature [21] but has not previously been tested on a corpus of spontaneous English discourse.

To ascertain duration's role in prosodic stress, several minutes of the Switchboard corpus were labeled in terms of primary and secondary stress by a linguist, and this material used to train and develop an automatic algorithm to accomplish the same objective. The automatic procedure ultimately is capable of correctly distinguishing stressed from unstressed syllables 96% of the time (though its ability to distinguish primary and secondary stress is not as fine) [Table 9].

The algorithm operates on three separate parameters of the acoustic signal - duration, amplitude and fundamental frequency. An optimization routine was developed to ascertain the best combination to yield optimum discrimination between stressed and unstressed syllables. The results indicate that the product of amplitude and duration (i.e., energy) of the syllabic nucleus yields the performance closest to that of the linguistic transcriber. Fundamental frequency is relatively unimportant for distinguishing between the presence and absence of stress (though it does appear to play a somewhat more important role in distinguishing primary from secondary stress).

	Primary	**Secondary**	**Unstressed**
+ Stress	48	36	19
- Stress	2	10	379
Total	50	46	398
%Correct	**96**	**78**	**95**

Table 9. Performance of an automatic stress labeling algorithm for a small portion of the Switchboard corpus. A total of 398 syllables were scored with respect to the stress labeling performed by a linguist.

8.3 The Relation between Prosodic Stress and Pronunciation Variation

Prosodic stress serves to informationally highlight specific lexical and syllabic elements. Roughly one quarter of the syllables in spontaneous discourse receive some form of stress (cf. Table 9) and this proportion is likely to reflect some intrinsic division of the lexicon into informationally significant classes.

The data described in Tables 6-9 imply a positive relation between canonical pronunciation and prosodic stress. The two properties appear to travel together oftentimes, though they are not inseparable under many conditions. Syllables whose entire suite of constituents are canonically stressed are more than likely to receive primary or secondary stress. Two-thirds of the phonetic segments are pronounced in canonical fashion, two and a half times the proportion of stressed elements, indicating that stress is only one of several factors underlying the patterning of pronunciation variation. However, it is difficult to quantitatively ascertain with greater precision the relationship between stress and canonical pronunication without a larger amount of reliably transcribed material than is currently available.

9. INFORMATION'S ROLE IN PRONUNCIATION VARIATION

Words of high information valence (typically infrequently occurring referential constituents of a nominal phrase [i.e., nouns or adjectives]) tend to be pronounced in canonical fashion, while common lexical items, particularly pronouns, conjunctions and articles, generally depart from canonical form with regularity [7, 14]. Such patterning suggests that the information valence associated with specific words and syllables may play a decisive role within an utterance.

This is of potential significance to the design of ASR systems, since their lexica usually contain but a single canonical, as well as several alternative pronunciations for each lexical entry. The task of going from sound to meaning for such a system would, in principle, be far simpler if each spoken instance were of the canonical form. Since the probability of a word being spoken in canonical fashion increases as the speaking rate declines [7] it is likely that the negative relationship between recognition performance and speaking rate is primarily a consequence of this factor.

Speaking rate, per se, may not be the truly governing factor guiding the pronunciation of a spoken utterance. The degree to which a syllable deviates from the canonical is a function of both speaking rate and word frequency [7]. The slope of the function is far steeper for words of high frequency than for low. If speaking rate were the governing factor the slope of the function would be relatively constant across unigram frequency. However, the slopes differ by roughly a factor of two [7: Figure 3], suggesting that low-frequency words, irrespective of their rate of articulation, are more likely to be realized in canonical form than their commonly occurring counterparts.

The information valence of frequent words is more likely to fluctuate as a function of phrasal and sentential context than less common lexical items, thus providing a potential basis for the greater variability of pronunciation under differential speaking conditions. It is likely that frequently occurring words tend to be spoken faster and in more reduced fashion because of their inherent predictability. Under extreme conditions words of high frequency and predictability may be entirely deleted from the utterance, but without the listener's conscious awareness [15].

10. THE AUDITORY BASIS OF PRONUNCIATION VARIATION

Why are the onsets of syllables relatively well preserved while the codas and nuclei so highly variable in pronunciation? Most accounts of pronunciation variation cite biomechanical constraints imposed by the vocal apparatus [22] as the controlling parameter. However, such production-based accounts do not explain why a mechanical system capable of such versatile behavior under a wide range of speaking conditions can also tailor its performance at will to deviate in systematic fashion when circumstances dictate. Might articulation serve as the handmaiden of higher linguistic function guided in its descent by the informational demands of the regime?

The auditory system is particularly sensitive and responsive to the beginnings of sounds, be they speech, music or noise. Our sense of hearing evolved under considerable selection pressure to detect and decode constituents of the acoustic signal possessing potential biological significance. Onsets, by their very nature, are typically more informative than medial or terminal elements, serving both to alert, as well as to segment the incoming acoustic stream. For this reason the majority of auditory neurons, from periphery to cortex are most highly responsive to the initial portion of a signal. Complex, multi-level chains of neural adaptation and inhibition reinforce and enhance this bias towards the onset [13].

The syllable can be thought of as the structural instantiation of this auditory process, shaping the encoding of linguistic information so as to maximize its probability of reception and decoding. Over the course of a lifetime, control of pronunciation is beveled so as to take advantage of the ear's (and the brain's) predilection for onsets (and other transients) and to tailor the meter of the speech to the low-frequency rhythm of the auditory cortex

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation, the Johns Hopkins Center for Language and Speech Processing and the U.S. Department of Defense. The computational assistance of Takayuki Arai, Eric Fosler, Rosaria Silipo and Su-Lin Wu is gratefully acknowledged. The labeling of linguistic stress was done by Collin Baker.

REFERENCES

[1] Arai, T. and Greenberg, S. "The temporal properties of spoken Japanese are similar to those of English," *Proceedings of Eurospeech*, Rhodes, Greece, 1997, pp. 1011-1014.

[2] Bernstein, B. B. *Class, Codes And Control.* London: Routledge, Kegan, Paul, 1974.

[3] Byrne, W., Finke, M., Khudanpur, S., McDonnough, J., Nock, H., Saraclar, M., Wooters, C. and Zavaliagkos, G. "Pronunciation modelling for conversational speech recognition - A status report from WS97," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 26-33.

[4] Dewey, G. (1923) *Relative Frequency of English Speech Sounds*, Cambridge: Harvard University Press.

[5] Doyle, A. C. *The Adventures of Sherlock Holmes*, New York: Harper, 1892.

[6] Fosler , E., Weintraub, M., Wegmann, S., Kao, Y.-H., Khudanpur, S., Galles, C. and Saraclar, M. "Automatic learning of word pronunciation from data," in the *Proceedings of the International Conference on Spoken Language Processing*, S28-29.

[7] Fosler-Lussier, E. and Morgan, N. "Effects of speaking rate and word frequency on conversational pronunciation," *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Speech Recognition.*, Kekrade, Netherlands, 1998.

[8] French, N. R., Carter, C. W. and Koenig, W. "The words and sounds of telephone conversations," *Bell System Tech. J.*, 9: 290-324, 1930.

[9] Ganapathiraju, A., Goel, V., Picone, J., Corrada, A., Doddington, G., Kirchhoff, K., Ordowski, M. and Wheatley, B. "Syllable - A promising recognition unit for LVCSR," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 1996, pp. 207-214.

[10] Gauvain, J., Lamel, L., Adda, G. and Adda-Decker, M. "The LIMSI continuous speech dictation system: evaluation on the ARPA Wall Street Journal task," IEEE ICASSP94, 1994, pp. 557-560.

[11] Godfrey, J. J., Holliman, E. C. and McDaniel, J. "SWITCHBOARD: Telephone speech corpus for research and development," IEEE ICASSP92, 1992, pp. 517-520.

[12] Goldinger, S. D., Pisoni, D. B. and Luce, P. "Speech perception and spoken word recognition: research and theory," in *Principles of Experimental Phonetics*, N. Lass (ed.), St. Louis: Mosby, 1996, pp. 277-327.

[13] Greenberg, S. "Auditory function," in *Encyclopedia of Acoustics*,, M. Crocker, editor. New York: John Wiley, 1997, pp. 1301-1323.

[14] Greenberg, S. "On the origins of speech intelligibility in the real world," ESCA Workshop on *Robust Speech Recognition for Unknown Communication Channels,* Pont-a-Mousson, France, 1997, pp. 23-32.

[15] Greenberg, S. "The Switchboard Transcription Project," in Research Report #24, *Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1997.

[16] Greenberg, S., Hollenback, J. and Ellis, D. "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus." *International Conference on Spoken Language Processing*, Philadelphia, 1996, pp. S32-35.

[17] Hayes, B. (1995) *Metrical Stress Theory,* Chicago: University of Chicago Press.

[18] Jespersen, O. *Language; Its Nature, Development and Origin.* London, Allen and Unwin, 1922.

[19] Kenyon, J. S. and Knott, T. A. *A Pronouncing Dictionary of American English*. Springfield, MA: Merriam, 1953.

[20] Labov, W. *Sociolinguistic Patterns*, Philadelphia: University of Pennsylvania Press, 1972.

[21] Lehiste, I. "Suprasegmental features of speech," in *Principles of Experimental Phonetics*, N. Lass (ed.), St. Louis: Mosby, 1996, pp. 226-244.

[22] Levelt, W. *Speaking*. Cambridge: MIT Press, 1989.

[23] Lindblom, B. "Explaining phonetic variation: A sketch of the H & H theory," in *Speech Production and Speech Modeling*, W. Hardcastle and A. Marchal (eds.), Dordrecht, Kluwer, 1990, pp. 403-439.

[24] Niemann, H., Noth, E., Kiessling, A., Kompe, R. and Batliner, A. "Prosodic processing and its use in Verbmobil," IEEE ICASSP-97, pp. 75-78, 1997.

[25] Ostendorf, M., Byrne, B., Macchiani, M., Finke, M., Gunawardana, A., Ross, K., Roweis, S., Shriberg, E., Talkinm, D., Waibel, A. Wheatley, B. and Zeppenfeld, T. "Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode," in Research Report #24, *Large Vocabulary Continuous Speech Recognition Workshop Technical Report Series*. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1997.

[26] Riley, M. "A statistical model for generating pronunciation networks," in ICASSP91, 1991, pp. 737-740.

[27] Switchboard Transcription Project. ICSI, Berkeley, CA. http://www.icsi.berkeley.edu/real/stp/, 1997.

[28] Weinstraub, M., Fosler, E., Galles, C., Kao, Y.-H., Khudanpur, S., Saraclar, M. and Wegmann, S. "WS96 Project Report: Automatic Learning Of Word Pronunciation from Data," in Research Report #24, *Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1997.

[29] Zipf, G. K. "The meaning-frequency relationship of words." *J. Gen. Psych.*, 33: 251-256, 1945.

[30] Zue, V.W. and Seneff, S. "Transcription and alignment of the TIMIT database." in *Recent Research Towards Advanced Man-Machine Interface Through Spoken Language,* H. Fujisaki (ed.), Amsterdam: Elsevier, 1996, pp. 515-525.