Word-level phonetic variation in large speech corpora

Patricia A. Keating
Phonetics Lab, Linguistics Department, UCLA

1. Introduction

The phonetic word is of crucial importance for continuous speech recognition. This is because the word is both a basic unit that is recognized (i.e. pieces of the speech signal are matched to words in a recognition lexicon) and a basic unit used in higher-level language models. The pronunciation variability of words is very important, since such variability makes it harder to match signals to lexical items. It is particularly problematic in large vocabulary systems, since variation in the pronunciation of one word will likely make it confusable with some other word.

The problem of pronunciation variability of words has become acute as recognition has turned to more casual, unscripted, speech. Word and acoustic models built from careful speech, especially read speech, have not generalized well to more natural speech. It is thought that more natural speech is more variable in two ways:

- **phonetically** (more realizations of phonemes or other sub-word units of recognition)
- **phonemically** (more realizations of each word expressed in such units)

The typical solutions to these problems are:

- **phonetic:** use more training data to get better statistical models of acoustic variation
- **phonemic:** use more pronunciations per word in the recognition lexicon

In most automatic speech recognition systems, words are entered into a lexicon with one pronunciation ("word model") -- either from a dictionary, or some estimate of the "Most Common Pronunciation", or a baseform designed specifically as input to a phonology. Phonological rules or networks can then be used to generate alternate pronunciations from any one of these types of lexical entries. Or, alternatively, alternate pronunciations can be entered directly into a lexicon. For example, a working group at the 1996 speech recognition summer workshop reported in Fosler et al. (1996) that they tried putting pronunciations actually found in their training data (pronunciations found at least seven times) into their lexicon. There is a clear trade-off between allowing few vs. many pronunciations for each word. Cohen (1989) estimated that for careful (e.g. read) speech, a single pronunciation for each word covers (on average) about 80% of its tokens, but to cover the other 20% of tokens, multiple pronunciations are required. Thus a recognition system which performed at 59% correct using only a Most Common Pronunciation for each lexical item, improved to 66% correct under one scheme of multiple pronunciations (weighted for probability) generated by rule from a single base form. At the same time, Cohen also showed that it is crucial not to generate too many alternate pronunciations of lexical entries, else the recognizer can be overwhelmed by false alarms.
In this paper I will test the hypothesis that the pronunciation of words is more variable in unscripted speech than in read speech. If this is so, then this confounding of hits by false alarms in a lexicon with multiple pronunciations would be more problematic for unscripted

speech. If only a small number of pronunciations is allowed (because of the false-alarm problem) then many pronunciations of many words will be necessarily unrepresented in a lexicon, leading to misses. It will then become important to understand which words or word classes are likely to be more variable, so that different strategies can be applied to different parts of the lexicon.

2. Method

2.1. Speech materials

2.1.1. Corpora

The two most important large corpora of recorded American English speech are TIMIT[1] and Switchboard[2]. TIMIT consists of 6300 read sentences, 10 each from 630 speakers, totaling about 5100 word types and about 54391 word tokens. Switchboard consists of about 3 million (orthographic) word tokens of unscripted telephone conversations from 550 speakers. TIMIT was for some time the resource most used in developing and testing continuous speech recognition systems; as a result, recognizers got very good at read speech. Problems arise when everything learned from and based on TIMIT is carried over to recognizing speech from Switchboard - recognition error rates, while no longer as disastrous as they were even two years ago, are much higher.

All of TIMIT could be used for this study since it is available at little cost. A randomly chosen subset of Switchboard was available from a previous project (Keating et al. 1994).

2.1.2. Words (lexical items)

A set of words that occur in both corpora was chosen, and pronunciations of each word were compared across the corpora. For practical reasons, by "word" here is meant the orthographic word, i.e. delimited by spaces or punctuation. Thus, while "no" and "know" count as different words, "that" (determiner) and "that" (complementizer), or "like" (preposition or interjection) and "like" (verb), would count as the same word; and while "it" and "it's" would both count as single words, "it is" would count as two words. It is quite possible that some pronunciation variation of lexical items counted in this way arises from the fact that different linguistic words are being collapsed together.

To study pronunciation variability a large number of tokens is required for each word. Frequency counts for words in TIMIT (*sa*, *sx*, and *si* sentences) were made from our database (REF). Frequency counts for the 160 most common words in Switchboard, and frequency bins for about 100 other words of variable frequency in Switchboard, were made available by Mark Liberman (p.c.). An arbitrary threshold for inclusion was set at 33 tokens per word, that is, a word must occur at least 33 tokens in each corpus. A further criterion, which applied only to the TIMIT sample, was that no more than 3 of the tokens for a word could come from the same speaker or the same (orthographic) sentence. No attempt was made to eliminate tokens that occurred within identical word strings shorter than the sentence. (This means that in Switchboard, more than in TIMIT, some tokens may have come from similar contexts. So this would work to reduce the apparent variability in Switchboard, and thus make Switchboard and TIMIT more alike in variability (thus going against the hypothesis)).

---

[1] DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) October 1990, NIST Speech Disc 1-1.1 (1 disc); http://www.nist.gov/itl/div894/894.01/corpora/timit.htm

[2] http://www.ldc.upenn.edu/ldc/catalog/html/speech_html/scr.html

Table 1. Comparison sample: words sampled from both Switchboard (SWB) and TIMIT, in alphabetical order, with total number of tokens of each word in each corpus. There are about 3 million word tokens in Switchboard, about 54000 in TIMIT. Therefore to compare the two figures very approximately, multiply the TIMIT figure by 50. To compare Switchboard with Kučera and Francis (1967), a 1 million-word corpus, divide the Switchboard figure by 3.

WORD	*# tokens in SWB*	*# tokens in TIMIT*
a	72924	1168
about	12362	50
and	106833	667
are	14024	349
as	10141	197
at	10791	134
be	14321	263
but	28291	136
don't	18641	668
for	19867	377
had	11033	709
have	30394	149
he	9594	341
I	121443	127
in	40532	1260
is	26182	517
it	55571	236
like	23441	697
my	15007	117
not	14977	158
of	56340	640
on	17010	267
one	12728	78
or	16851	117
out	11091	82
so	26417	65
that	67035	827
the	98301	2202
them	10468	58
there	13290	59
they	33212	179
this	9862	210
to	73147	1370
up	9973	89
was	24187	321
we	25672	187
well	22024	37
what	14933	62
with	14044	244
you	80241	362

A total of 40 of the 60 most common words in Switchboard, words that also occur in TIMIT, were selected by these criteria and are listed in Table 1. (High-frequency words of Switchboard that occur infrequently or not at all in TIMIT are: *uh, yeah, uh-huh, that's, think, oh, really, right, um, I'm*, and words which occur fewer than 33 times, or which failed the second criterion, are: *know, it's, don't.*) To insist on more than 33 tokens greatly limits the number of words that can be studied in TIMIT, and of course those that do occur this often are all high-frequency function words. So as to include some lower-frequency words, including content words, in the study, an additional 32 words were selected from Switchboard only. These are shown in Table 2. The pronunciations of these words cannot be compared to TIMIT, but they can be compared to the high-frequency function words in Switchboard.

Table 2. Sample of other words from Switchboard only (not enough tokens occur in TIMIT). Exact frequencies not available, only frequency ranges.

WORD	*frequency in SWB*
after	between 1000 and 1400
cases	between 40 and 50
chips	between 40 and 50
could	between 3000 and 5500
down	between 3000 and 5500
facts	between 40 and 50
glass	between 180 and 240
goal	between 40 and 50
island	between 100 and 140
know	47560 (included here because too few in TIMIT)
market	between 180 and 240
metric	between 100 and 140
must	between 300 and 500
okay	between 3000 and 5500
once	between 1000 and 1400
paint	between 180 and 240
played	between 300 and 500
probably	between 3000 and 5500
road	between 180 and 240
simple	between 100 and 140
since	between 1000 and 1400
stick	between 180 and 240
system	between 1000 and 1400
taken	between 300 and 500
there's	between 3000 and 5500
under	between 300 and 500
upon	between 100 and 140
very	between 3000 and 5500
weeds	between 40 and 50
weekend	between 300 and 500
what's	between 1000 and 1400
years	between 3000 and 5500

2.1.3. Tokens

Matched numbers of tokens of each word were selected at random from Switchboard and TIMIT. This number was determined by whichever corpus yielded the smaller number of tokens (usually TIMIT). The number of tokens from each corpus was capped at 40.

2.2. Transcriptions

Phonetic and phonemic (dictionary-style) transcriptions of each token were obtained. Throughout this paper these transcriptions are shown in the ARPAbet-style symbols of the TIMITbet (Zue and Seneff 1988), listed in Table 3.

For TIMIT, the phonetic transcriptions used were those provided with the corpus: the "TIMITbet" transcriptions which are narrower than phonemic, but not especially narrow. Phonemic transcriptions were derived from these by a set of collapsing rules which collapsed the phonetic categories into fewer, broader, categories. The general approach of the collapsing rules is to map each more-specific symbol into the phonetically most similar more-general symbol. These collapsing rules do not take into account what the word is.

Table 3. TIMITbet symbols and nearest IPA equivalents; phonemic symbols used here. Case is not distinctive for TIMITbet symbols

TIMITbet symbols	*nearest IPA symbol*	*phonemicized here as*
pcl	p˺ (closure only)	p
p	p (release only)	p
b	b (release only)	b
bcl	b˺ (closure only)	b
t	t (release only)	t
tcl	t˺ (closure only)	t
d	d (release only)	d
dcl	d˺ (closure only)	d
k	k (release only)	k
kcl	k˺ (closure only)	k
g	g (release only)	g
gcl	g˺ (closure only)	g
f	f	f
v	v	v
th	θ	th
dh	ð	dh
s	s	s
z	z	z
sh	ʃ	sh
zh	ʒ	zh
ch	tʃ (release only)	ch
jh	dʒ (release only)	jh

h (or hh)	h	h
hv	ɦ	h
m	m	m
n	n	n
ng	ŋ	ng
em	m̩	ax m
en	n̩	ax n
eng	ŋ̩	ax ng
r	ɹ	r
l	l	l
er (also listed below)	ɻ	ax r
el	l̩	ax l
w	w	w
y	j	y
dx	ɾ	d
nx	ɾ̃	n
q	ʔ	-
iy	i	iy
ih	ɪ	ih
ey	eɪ	ey
eh	ɛ	eh
ae	æ	ae
aa	ɑ	aa
ay	aɪ	ay
aw	aʊ	aw
ao	ɔ	ao
ow	oʊ	ow
oy	ɔɪ	oy
uh	ʊ	uh
uw	u	uw
ah	ʌ	ah
er	ɝ	ax r
ux	ʉ	uw
ix	ɨ	ih
ax	ə	ax
ax-h	ə̥	ax
axr	ɚ	ax r

For Switchboard, the initial phonetic transcriptions were done at UCLA and were narrower still, in "UCLAbet" symbols (Keating et al. 1994). Some of the Switchboard transcriptions were done by two or more transcribers. Agreement between these transcribers was good overall for unscripted telephone speech. Therefore additional Switchboard transcriptions were done by the author alone. It should be noted that in general it seems harder to get transcribers to agree when transcribing rapid fluent speech like Switchboard, than when

transcribing read speech like TIMIT. Thus, pronunciation variability is probably necessarily confounded with transcription variability in studies such as the one here (with human transcribers).

These narrow transcriptions have been done for the purpose of studying phonetic variation in more detail than TIMITbet transcription would allow. For present purposes, however, these were converted into TIMITbet by a second set of collapsing rules. The Switchboard phonemic transcriptions were then derived from these TIMITbet transcriptions as was done for TIMIT. Table 4 schematizes the levels of transcription.

Table 4. Levels of transcription produced by collapsing rules.

	UCLAbet narrow	--->	*TIMITbet phonetic*	--->	*phonemic*
TIMIT	(not available)		yyy	--->	zzz
SWB	xxx	--->	yyy	--->	zzz

Another difference between the corpora relevant to the transcriptions is that while Switchboard is telephone speech, TIMIT is not (at least, not the original TIMIT used for the transcriptions). So to the extent that Switchboard is degraded speech relative to TIMIT, that could also make the pronunciations seem more variable -- it is simply harder to ascertain what the speaker said. In fact though this is probably not a big factor here: when a sample was really noisy we didn't use it, and the difficult issues of transcription were not generally related to bandwidth or noise. (They were about syllabicity and vowel reduction.)

2.3. Analyses

From the set of transcriptions, the Most Common Pronunciation was determined for each word in each corpus at each level of transcription. The *Most Common Pronunciation*, or MCP, is that pronunciation that occurs most frequently in the sample of 33-40 tokens, and its *coverage* is the percentage of the sample with that pronunciation. For example, 39 of 40 tokens of "stick" have the phonemic transcription /s t ɪ k/, so that is its MCP (phonemic), and the coverage of that MCP is 98%.

A number of different counts and calculations were also done. These will be described along with their results in sections below.

3. Results

3.1. Number of pronunciations per word

The raw number of distinct pronunciations was counted for each word. These are summarized in Table 5 for the 40 words available for both corpora.

Table 5. Average numbers of pronunciations per word, comparison sample of 40 words.

	in TIMIT	*in SWB*
phonetic transcriptions	9.5	14.3
phonemic transcriptions	5.8	9.5

It can be seen that there are fewer different phonemic pronunciations than phonetic in both corpora (this is almost definitionally so), and that there are fewer different pronunciations at both levels in TIMIT than in Switchboard, as hypothesized. These results can be compared with those in Table 6, which shows the same counts for the sample of 32 other words from Switchboard, mostly low-frequency content words. The figures for these words in Switchboard are remarkably similar to those for the higher-frequency words in TIMIT.

Table 6. Average numbers of pronunciations per word, lower-frequency words (SWB only)

phonetic transcriptions	10.0
phonemic transcriptions	5.7

3.2. Phonemic variation

It is quite striking that even in a phonemic (dictionary-style) transcription, there are almost 10 different pronunciations per high-frequency word for samples of only 33-40 words, and over 5 different pronunciations even for lower-frequency words. Phonemic transcriptions were tabulated because it is sometimes suggested that if only the phonemes could be reliably recovered from the signal, then the word recognition problems would be minor. The results in the previous section show that this is not true. (In a similar vein, Fosler et al. (1996) compared (hand-done) Switchboard transcriptions with dictionary baseforms, and found that on average, one out of eight phones (phonemes) from the baseforms were deleted in the transcriptions.) However, the figures in Tables 5-6 are averages, and it is certainly the case that some words do not vary much in phonemic transcriptions. For those words, which are listed in Table 7, successful recognition of the phonemes would ensure ready recognition of the words. While such words are generally from the low-frequency sample, it can be seen that not all 32 low-frequency words have this property, as there are only 10 such words here.

Table 7. Words in Switchboard (out of 72) which do not vary much at phonemic level.

WORD	*# phonemic pronunciations*
bear	2
facts	2
glass	2
goal	2
like	2
metric	3
must	3
my	3
simple	3
stick	2
system	2
very	2

For those words which do vary at the phonemic level, several generalizations can be made, which hold for the content words too. All phonemic pronunciations which occurred four or more times were examined and the following patterns found.

3.2.1. The 2-schwas problem: TIMITbet distinguishes between a lower [ax] and a higher [ix] reduced vowel (basically, IPA [ə] vs. [ɨ]). The criterion for deciding between them is whether F2 is closer to F1 vs. F3. These two reduced vowels were phonemicized differently, as /ax/ vs. /ih/. In general this accords with the underlying vowels, but not always. For some words individual tokens were found to vary in the F2 frequency, and this difference was then carried up to the phonemicization. Note that these phonemicizations are determined only by the signal; it would be circular to restore underlying segments on the basis of lexical knowledge. Words with this variation included *a, and, as, at, but, cases, in, is, of, system, taken, that, the, was, what, with.*

3.2.2. Vowel reductions: In general, all vowels in function words can reduce. There were some general tendencies in these reductions, as follows (in IPA symbols): /i/ /u/ /ʊ/ often reduce to /ɪ/; /ʌ/ /o/ /ɛ/ often reduce to /ə/; /æ/ often reduces to /ɛ/. But there was enough variation beyond these patterns to give rise to multiple pronunciations, in words such as *and, as, be, but, could, don't, one, she, so, that, them, under, we, what, what's, you.*

3.2.3. Flapping: Both underlying /t/ and /d/ were often flapped. However, all flaps were phonemicized as /d/, since that is the phonetically closer quality.

3.2.4. Final /t d n l/ loss: These anterior coronal consonants tend to not be heard/seen word-finally, but not consistently so. Words with this variation included *and, at, don't, down, in, it, must, not, out, paint, road, that, weekend, well, what.*

3.2.5. Dialect variation in vowels: Some words contain vowels that seem to vary greatly across speakers, including *my, on, our, the, well, I.*

3.2.6. Weak syllable loss: Stressless syllables are vulnerable in vowel-initial iambs (*upon, about*) and word-medially (*probably*), but not consistently so.

3.2.7. Initial /dh/ loss in function words: Words like *them, they, this* may appear to lose their initial consonant in some, but not all, contexts. They are particulary vulnerable when following another function word ending in a nasal.

3.2.8. Final -(r)z devoicing: Word-final /z/ is sometimes devoiced in *there's, years.*

Table 8. Phonemic MCP and its coverage in the two corpora; dictionary pronunciation (converted to phonemic transcription used here). In the dictionary consulted, some special r-colored vowel symbols were used; these have been converted here to our usual transcriptions. Where the MCP for a given word is different in the two corpora, the coverage of each MCP in the other corpus is given in parentheses.

WORD	*MCP in SWB*	*its coverage*	*MCP in TIMIT*	*its coverage*	*dictionary form*
a	ax	29	ax	55	ey, ax
about	ax b aw /	8	ax b aw t	39	ax b aw t
	ax b aw t /		(ax ba w	6)	
	b aa / ih b aw d		(b aa	0)	
and	eh n	25	ih n	35	ae n d, ax n

	(ih n	15)	(eh n	28)	ax n d, en
are	ax r	62	aa r	54	aa r, ax r, axr
	(aa r	30)	(ax r	30)	ax
as	ih z	53	ih z	32	ae z, ax z
at	ih t	24	ae t	29	ae t, ax t
	(ae t	5)	(ih t	20)	
be	b iy	67	b iy	100	b iy, b ih
but	b ah t	26	b ah t	38	b ah t, b ax t
don't	d ow n	33	d ow n t	45	d ow n t
	(d ow n t	0)	(d ow n	39)	
for	f ax r	61	f ax r	61	f ao r, f ax r
had	h ae d	51	h ae d	38	h ae d
have	h ae v	69	h ae v	67	h ae v
he	h iy	66	h iy	87	h iy
I	ay	64	ay	92	ay
in	ih n	45	ih n	79	ih n
is	ih z	71	ih z	87	ih z
it	ih t	36	ih t	62	ih t
(know)	n ow	86	n ow	91	n ow
like	l ay k	97	l ay k	100	l ay k
my	m ay	71	m ay	87	m ay
not	n aa t	43	n aa t	55	n ao t
of	ax v	32	ax v	49	ah v, ao v, ax v
on	ao n	32	ao n	41	ao n
one	w ah n	50	w ah n	84	w ah n
or	ax r	64	ao r	41	ao r, ax r
	(ao r	5)	(ax r	31)	
out	aw / aw t	20	aw t	54	aw t
			(aw	14)	
so	s ow	61	s ow	87	s ow
that	dh ih t	15	dh ae t	23	dh ae t, dh ax t
the	dh ax	35	dh ax	40	dh iy, dh ax, dh ih
them	ax m /	26	dh eh m	72	dh eh m, dh ax m
	dh ax m /		(ax m	3)	
	dh eh m		(dh ax m	8)	
there	dh eh r	69	dh eh r	36	dh eh r
they	dh ey	68	dh ey	95	dh ey
this	dh ih s	69	dh ih s	89	dh ih s
to	t ih	31	t uw	31	t uw, t ax
	(t uw	23)	(t ih	28)	
up	ah p	59	ah p	95	ah p
was	w ih z	39	w ih z	33	w ah z, w ao z, w ax z
we	w iy	66	w iy	89	w iy
well	w eh l	34	w eh l	86	w eh l
what	w ax d/w ax t	15	w ah d	35	w ao t, w ah t
	(w ah d	13)	(w ax d	5)	
			(w ax t	8)	
with	w ih th	41	w ih th	54	w ih th, w ih dh
you	y uw	35	y uw	78	y uw

3.3. Most Common Pronunciation

Recall that Cohen (1989) found that the MCP covers, on average, about 80% of tokens for words in read speech. Table 8 gives the phonemic MCP, and its coverage, for each word in our comparison samples. It also gives a pronunciation for each word taken from a dictionary (Harcourt, Brace, & World's *Standard College Dictionary*, 1963). Table 9 gives the phonetic MCPs and their coverage. For this sample, the MCP is often the same for the two corpora. Phonemically, it is the same for 80% of the words, while phonetically it is the same for 65% of the words. That is, a phonemic lexicon based on the MCPs in TIMIT is a reasonable starting point for a Switchboard lexicon, since the agreement here is 80%. Furthermore, when the MCP's coverage is greater than 50% in both corpora (that is, just the cases where the MCP is doing the most work), the two corpora almost always have the same MCP. Exceptions to this generalization are phonetic *this* (TIMIT [dh ih s], Switchboard [dh ix s]) and phonemic *are* (TIMIT /aa r/, Switchboard /ax r/).

Table 9. Phonetic MCP and its coverage in the two corpora. Format as in previous table.

WORD	*MCP in SWB*	*its coverage*	*MCP in TIMIT*	*its coverage*
a	ix	26	ax	47
	(ax	24)	(ix	21)
about	ax bcl b aw q	8	ax bcl b aw tcl	25
	(ax bcl b aw tcl	3)	(ax bcl b aw q	3)
and	eh nx / en	13	ix n	18
	(ix n	8)	(eh nx	5)
			(en	8)
are	axr	41	aa r	38
	(aa r	11)	(ax r	19)
as	ix z	47	ix z	23
at	ix tcl	15	ae tcl	15
	(ae tcl	2)	(ix tcl	10)
be	bcl b iy	49	bcl b iy	59
but	bcl b ah dx	15	b ah tcl	23
	(b ah tcl	0)	(bcl b ah dx	0)
don't	dcl d ow n	15	dcl d ow n tcl	24
	(dcl d ow n tcl	0)	(dcl d ow n	6)
			(dcl d aw nx	5)
for	f ax r	51	f ax r	56
had	hv ae dx	16	eh dcl/hv ae dx	16
	(eh dcl	0)		
have	hv ae v	38	hv ae v	38
he	hv iy	37	hh iy	79
	(hh iy	21)	(hv iy	8)
I	ay	32	ay	54
in	ih n / ix n	16	ix n	39
			(ih n	21)
is	ix z	42	ix z	53
it	ih q	12	ih tcl	24
	(ih tcl	0)	(ih q	2)
(know)	n ow	57	n ow	91
like	l ay kcl k	49	l ay kcl k	59

my	m ay	61	m ay	87
not	n aa tcl	33	n aa tcl	53
of	ax	30	ax v	41
	(ax v	27)	(ax	5)
on	ao n	24	ao n	35
one	w ah n	25	w ah n	70
or	axr	38	axr	26
out	aw / aw tcl	11	aw tcl	46
			(aw	0)
so	s ow	61	s ow	87
that	dh ae dx	13	dh ae tcl	18
	(dh ae tcl	5)	(dh ae dx	5)
the	dh ax	30	dh ax	35
them	dh eh m	26	dh eh m	72
there	dh eh r	67	dh eh r	36
they	dh ey	65	dh ey	95
this	dh ix s	60	dh ih s	86
	(dh ih s	9)	(dh ix s	3)
to	t ix / tcl t ix / tcl t ux	13	tcl t ix	21
			(t ix	8)
			(tcl t ux	10)
up	ah pcl p	38	ah pcl p	49
was	w ix z	39	w ax z	31
	(w ax z	25)	(w ix z	28)
we	w iy	63	w iy	89
well	w eh l	31	w eh l	86
what	w ax dx	15	w ah dx / w ah tcl	33
	(w ah dx	13)	(w ax dx	5)
	(w ah tcl	8)		
with	w ix th	31	w ix th	44
you	y ix	20	y ux	63
	(y ux	20)	(y ix	10)

It can readily be seen also that for most words the MCP has better coverage in TIMIT than in Switchboard: this is so for 78% of the words considered phonemically, and 80% of the words considered phonetically. There are some exceptions, however; the words *are, as, for, have, had, or, there, was* are more consistently reduced in Switchboard, so that the MCP is this reduced form.

The average coverages are given in Table 10. At both levels of transcription there is about a 15% difference in coverage. That means that, although a TIMIT-based lexicon in general will provide a good base form for Switchboard, the coverage offered by that form will be less. It will be noted that these coverages are quite low in general; in particular, the 62% phonemic coverage in TIMIT is much lower than Cohen's 80% figure for read speech. This is in part because the sample here is limited to a set of very high-frequency function words, whereas Cohen's figure was derived over a larger set of words. In addition, Cohen's data were not from TIMIT, but from a study of the DARPA Resource Management Database[3], which involves only a subset of the speakers from TIMIT, reading database query sentences.

[3] http://www.itl.nist.gov/div894/894.01/corpshrt.htm

Table 10. Coverage of MCP (% of sample) -- comparison sample of 40 words		
	in TIMIT	*in SWB*
phonetic	48	33
phonemic	62	47

Table 11 shows that the average coverage of the phonemic MCP for the lower-frequency words in Switchboard is 70%, much closer to Cohen's 80%. These low-frequency words in Switchboard are more like the high-frequency words in TIMIT above. So we would expect a lexicon derived from TIMIT to work reasonably for the lower-frequency content words of Switchboard, but not for the high-frequency function words. These two tables also show that the difference between the two samples from Switchboard (higher frequency words in Table 10, lower frequency words in Table 11) is greater when phonemic transcriptions are counted.

Table 11. Coverage of MCP (% of sample) -- Switchboard-only sample of 32 words	
phonetic	47
phonemic	70

The phonemic MCP can be compared to a dictionary entry, shown in the last column of Table 8. The dictionary consulted here included alternate reduced pronunciations for function words. In general these pronunciations correspond to the observed MCP (plus some British-like variants given in the dictionary): they are the same for 90% of the 40 words for TIMIT, and for 75% of the (same) 40 words for Switchboard.

Finally, it is interesting to see whether any words within these samples share their MCP, or look as if they might share their MCP with some other word not in the sample. Such cases would pose obvious problems for recognition. There are a few, whether the phonetic or the phonemic transcriptions are considered. In the TIMIT sample, *and/in*, *as/is*, and *are/our* share their MCP, and in Switchboard *as/is* and *are/or* do (see tables for specific forms).

3.4. Other schemes for inclusion of pronunciations

3.4.1. Pronunciations occurring 7 or more times

Fosler et al. (1996) attempted to improve recognition performance by constructing a recognition lexicon from observed pronunciations. Pronunciations observed at least 7 times in the training data, a sample of 2116 sentences, were used. What kind of coverage would this criterion give for the present Switchboard samples? While the number of word tokens in the 2116 sentences that they sampled is larger than the number of tokens in the present study, the number of high-frequency words is probably roughly similar. For the samples here of 33-40 tokens per word, a pronunciation that occurs 7 times would cover about 18-21% of the tokens.

Table 12 gives the number of pronunciations occurring 7 or more times for each word. It can be seen that there are usually 1 or 2 per word; the average is 1.45 such pronunciations per word. For those words where there is one such pronunciation, or two which are tied in coverage, it is the same as the MCP. For other words with 2 such pronunciations, their combined coverage will necessarily be better than that of the MCP pronunciation alone. But for a few words, there is no such pronunciation - no single pronunciation occurs at least 7 times - and for these words, this criterion would hurt, not help, coverage.

Table 12. Counts of phonemic transcriptions, high-frequency Switchboard sample only.

WORD	*# prons 7+ times*	*coverage (%)*	*# prons 2+ times*	*coverage (%)*	*#prons 50% coverage*
a	2	55	7	92	2
about	0	0	10	67	5
and	2	48	6	85	3
are	2	92	2	92	1
as	1	48	7	93	2
at	1	25	9	90	3
be	1	67	3	87	1
but	2	44	7	92	3
don't	1	33	3	61	3
for	1	63	4	93	1
had	1	51	5	89	1
have	1	69	3	82	1
he	1	66	5	92	1
I	2	83	3	93	1
in	1	45	4	79	2
is	1	71	5	97	1
it	2	60	5	88	2
like	1	97	1	97	1
my	2	92	3	100	1
not	2	75	4	95	2
of	3	89	4	95	2
on	2	57	5	92	2
one	2	75	3	83	1
or	1	64	5	90	1
out	2	40	7	74	4
so	2	82	4	95	1
that	0	0	8	83	4
the	2	60	5	90	2
them	3	77	5	90	2
there	1	69	4	86	1
they	1	68	4	100	1
this	1	69	3	89	1
to	2	54	8	92	2
up	1	59	5	85	1
was	2	58	5	89	2
we	2	92	2	92	1
well	1	34	6	86	2
what	0	0	7	78	4
with	1	41	6	82	2
you	2	58	5	85	2

3.4.2. Pronunciations occurring more than once

Table 12 also shows the number of pronunciations per word when a less restrictive criterion is applied: eliminate only pronunciations that occur only once (the presumed outlier pronunciations). These pronunciations cover, on average, 88% of the tokens for the 40

words in the Switchboard sample (with an average of five pronunciations per word), and virtually the same coverage, 89%, for the second Switchboard sample (with an average of three pronunciations per word). The coverage of these pronunciations ranges from 61% to 100%, but is generally high. Still, this result means that the outlier pronunciations which have been excluded account for over 10% of the tokens. Furthermore, this figure of 3-5 pronunciations per word, which seems to be necessary to get even this moderately acceptable level of covereage, is a high number, when the false alarm problem of high-vocabulary recognition is considered.

3.4.3. Pronunciations giving 50% coverage of samples

Finally, Table 12 shows the number of pronunciations per word needed for 50% coverage. Here we see numbers of pronunciations per word that would not cause a large false alarm problem.

4. Conclusion

This study has compared pronunciation variability, for a set of 40 lexical items, in the read speech of TIMIT vs. the non-read speech of Switchboard. The read speech of TIMIT is less variable on every measure. The Most Common Pronunciations of the lexical items are often the same in the two corpora (the same for 80% of lexical items sampled, in phonemic transcription), but their coverage is much reduced -- only 57% of the individual tokens for the 72 words in the two Switchboard samples presented here. More pronunciations beyond this one Most Common Pronunciation must be allowed to get reasonable coverage of the tokens of at least the high-frequency lexical items. Even if phonemes can be recognized completely accurately, there will still be much pronunciation variability to deal with.

The results presented here show that this variability is not the same for all words, however. The low-frequency content words of Switchboard vary no more than do words in TIMIT; therefore these words should present no new difficulties. It is the high-frequency function words of Switchboard that vary so much more and which must be the focus of new efforts. Even with these, not all of them vary greatly, or if they do, not always in ways that would make them potentially confusable with other lexical items. Therefore it would seem that research should focus on strategies for just the most variable and confusable words. For example, since these are function words, perhaps better language models for the structures they occur in could help.

Another possible approach would be to focus more on phonetic variation that distinguishes one sequence of phonemes from another. Even the TIMITbet-style transcriptions studied here collapse over phonetic details that could be useful in distinguishing lexical items, details that can be spread out over a larger span of speech. Some of these details are well-known to phoneticians: vowel nasalization that distinguishes *and* from *at*, or *in* from *it*, even when final consonants are deleted; glottalization that also distinguishes *in* from *it*; vowel duration differences that preserve voicing distinctions, or reflect the number of underlying consonants in a word. Other differences are less well-known, being either idiosyncratic or prosodic: for example, that *our* and *are* are generally distinguished by nasalization in *our*; or by the presence of a full glottal stop at the beginning of *our*. Such differences as these are not reflected in the transcriptions compared in this study. Furthermore, for many such useful properties, there are currently no good acoustic models that would allow their recognition. I

hope phoneticians and phonologists will get to work on this challenge, which provides a chance to show that our knowledge of sound structure can help with a practical problem.

References

Cohen, M. H. (1989). *Phonological structures for speech recognition*. Ph.D. dissertation, Dept. Electrical Enginerring and Computer Science, UC Berkeley.

Fosler, E., M. Weintraub, S. Wegmann, Y.-H. Kao, S. Khudanpur, C. Galles, M. Saraclar (1996). Automatic Learning of Word Pronunciation From Data. *ICSLP-96*.

Keating, P., M. MacEachern, A. Shryock, and S. Dominguez (1994). A manual for phonetic transcription: Segmentation and labeling of words in spontaneous speech. Manual written for the Linguistic Data Consortium, *UCLA Working Papers in Phonetics 88*, 91-120.

Kučera, H. and N. Francis (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Zue, V. and S. Seneff (1988), Transcription and alignment of the TIMIT database, *Proc. Second Meeting on Advanced Man-Machine Interface through Spoken Language*, pp. 11.1-11.10.

Acknowledgments

This work was supported by the UCLA Academic Senate Committee on Research. I thank Chai-Shune Hsu and Narineh Hacopian for much-appreciated assistance.