Massive reduction in conversational American English

Keith Johnson
Ohio State University

The English are a lazy lot, and will not speak a word as it should be spoken when they can slide through it. Why be bothered to say extraordinary when you can get away with strawdiny? ... Many of the Oxford Cockneys are weaklings too languid or emasculated to speak their noble language with any vigor, but the majority are following a foolish fashion which had better be abandoned. Its ugliness alone should make it unpopular, but it has the additional effect of causing confusion. [Irish playwright St. John Ervine, quoted by H.L. Mencken (1948, p. 39)]

1. Introduction

David Stampe (1973) discussed a range of variants of the phrase *divinity fudge* three of which are shown in (1).

(1) dəvɪnəti fʌdʒ
dəvĩ̯ə̯ti fʌdʒ
dəvĩ̯ĩ̯ fʌdʒ

I will call a reduction like the one that relates [dəvɪnəti] with [dəvĩ̯ĩ̯] a "massive" reduction. By this I mean that the phonetic realization of a word involves a large deviation from the citation form such that whole syllables are lost and/or a large proportion of the phones in the form are changed. The most reduced variant in (1) has two syllables where the citation form has four, and of the eight citation segments only three [dəv] are in both the reduced form and the citation form.

The goal of this paper is to relate pronunciation variation to models of auditory word recognition. Before, addressing auditory word recognition directly, however, I will discuss how phoneticians and phonologists have approached (or avoided) pronunciation variation, touch briefly on how dictionary editors compile dictionary pronunciations, and then delve into the depths of a very large recorded corpus of conversational American English. Having considered pronunciation variation from these perspectives, the paper will conclude with a discussion of lexical representation in models of human auditory word recognition.

Recently, number of researchers have been considering the implications of pronunciation variation for theories of auditory word recognition (Connine, Blasko & Titone, 1993; Gaskell & Marslen-Wilson, 1996, 1998; Lahiri & Marslen-Wilson, 1991; Cutler, 1998). However, the focus of attention in this work has been restricted to segment-count preserving variants, either ambiguous feature information (**igarette*, Norris, 1994) or consonant place assimilation (*lea[m] bacon*, Gaskell & Marslen-Wilson, 1996). Massive reduction is not segment-count preserving.

If Stampe's reduced forms are more than a mere curiosity, that is if people actually and frequently say things like [dəvĩĩ̯ fʌdʒ] for *divinity fudge*, then auditory word recognition is very different from the visual recognition of printed words. The difficulty of seeing the word *divinity* in the sequence of phonetic symbols [dəvĩĩ̯] gives a flavor of the nature of the auditory word recognition problem if massive reduction really happens. With massive reductions, phone-by-phone segment-count preserving look up procedures analogous to Forster's (1976) approach to visual word recognition, for example, would never work.

Of course, it could be that massive reduction does occur fairly frequently in conversational speech, but as St. John Ervine suggested, it results in confusion. This would have to be the prediction of the segment-based word recognition theories discussed in section 7 below because they do not permit the recognition of massively reduced words. Though in this paper I will not present results on the perception of conversational speech, a number of other authors have reported (comfortingly enough) that listeners are generally able to understand each other in ordinary conversations.

Which is not to say that listeners *always* succeed. Massive reductions has been known to be the source of additions to the lexical stock of languages. So, for example, *ordinary* is the historical source of *ornery.* Craigie & Hulbert (1938-44) find the pronunciation *ornery* first in 1830 and later as *onery* in 1860. Kenyon & Knott (1944) list both [ɔrnərɪ] and [ɔənərɪ] as pronunciations, while Mencken (1948, p. 97) has *o'n'ry*, which is a good way to write my own pronunciation [ɑnri], a historical reduction from 4 syllables to 2.

In this paper, I describe "massive" reduction in terms of (1) syllable deletion, and (2) segmental changes because these somewhat overlapping descriptions can be tallied fairly easily in a phonetically transcribed corpus of conversational American English (the Variation in Conversation corpus, Pitt et al., 2003, will be described in section 5). That is to say, syllable deletion and segmental change are convenient descriptors given a segmentally transcribed corpus. Other ways of measuring deviation from a lexical standard may reveal that forms that appear on a segmental basis to be incredibly deviant actually do contain most of the phonetic material specified in the lexical entry.

1.1. Examples

To illustrate the type of phenomena that I wish to consider under the name "massive reduction" examples from the ViC corpus are shown in figures 1-4. These examples are some of the more extremely reduced forms in the corpus.

Figure 1 shows a zero-syllable realization of the two-syllable function word *because* in the phrase *because if*. Two segments of the word (out of five) are retained in this production, though they now form an "illegal" consonant cluster in the bimorphemic monosyllable [k^hzɪf].

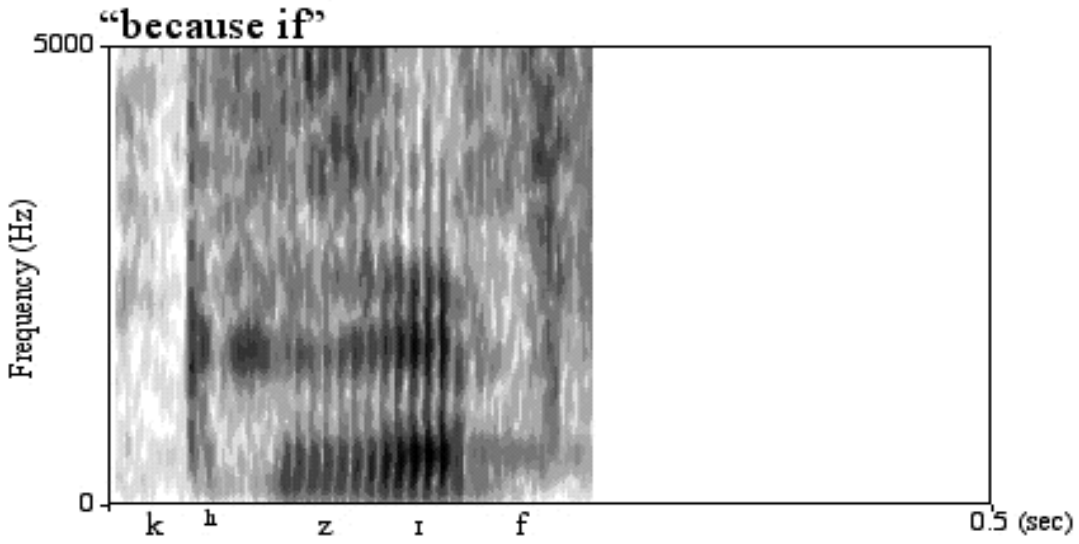

Figure 1: The two syllable word *because* is realized as [k^hz] - an illegal cluster at the onset in this instance of the phrase [k^hzɪf] *because if*.

In Figure 2 we see an instance of the phrase *apparently not*, in which the four-syllable word *apparently* is produced in only two syllables. It is difficult to give a transcription to the last syllable of *apparently* because its most dominant feature is the creaky phonation type which contrasts with the nearly falseto pitch of the emphasized word *not*. Nonetheless, of the nine segments of the citation form [əp^hɛɾəntli] only two [p^hɛ̰] survive unmodified in this production.

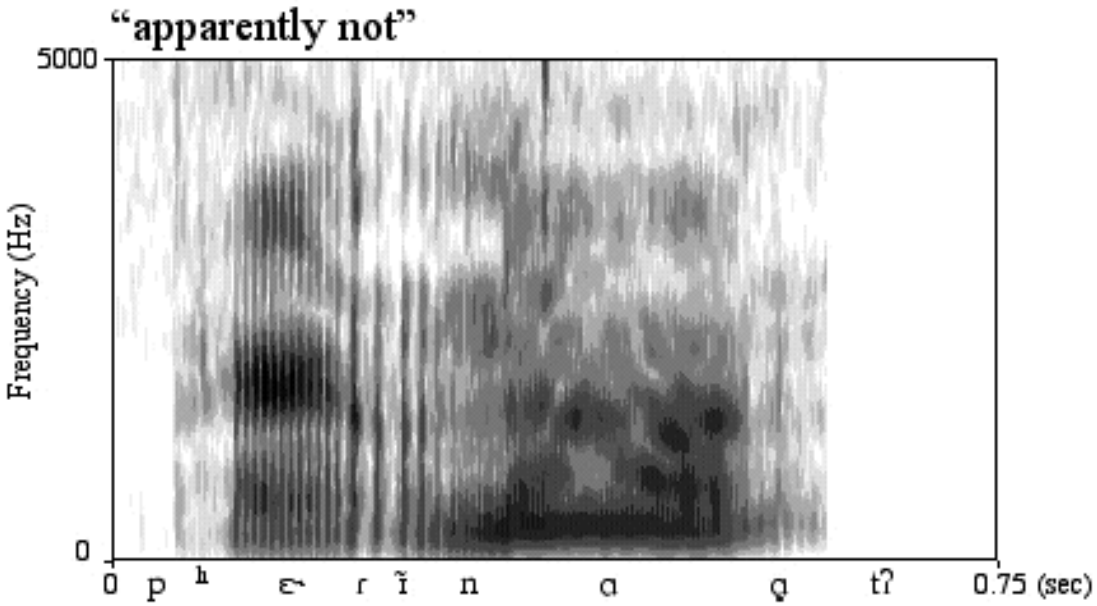

Figure 2: The four syllable word *apparently* is realized [p^hɛɾ̃ĩ] in this instance of the phrase *apparently not*.

Figure 3 shows an instance of *hilarious* which is transcribed as [hlɛrɛs]. As with *apparently*, this is a four syllable word realized with only two syllabic elements - the two instances of [ɛ]. In

this word production though, most of the phones found in the phonetic transcription match phones in the citation form [hɪlɛriʌs]. The unstressed [ɪ] of the first syllable has been deleted, and the vowels in the sequence [iʌ] have coalesced into [ɛ] being front like [i] and mid-low, lax like [ʌ].

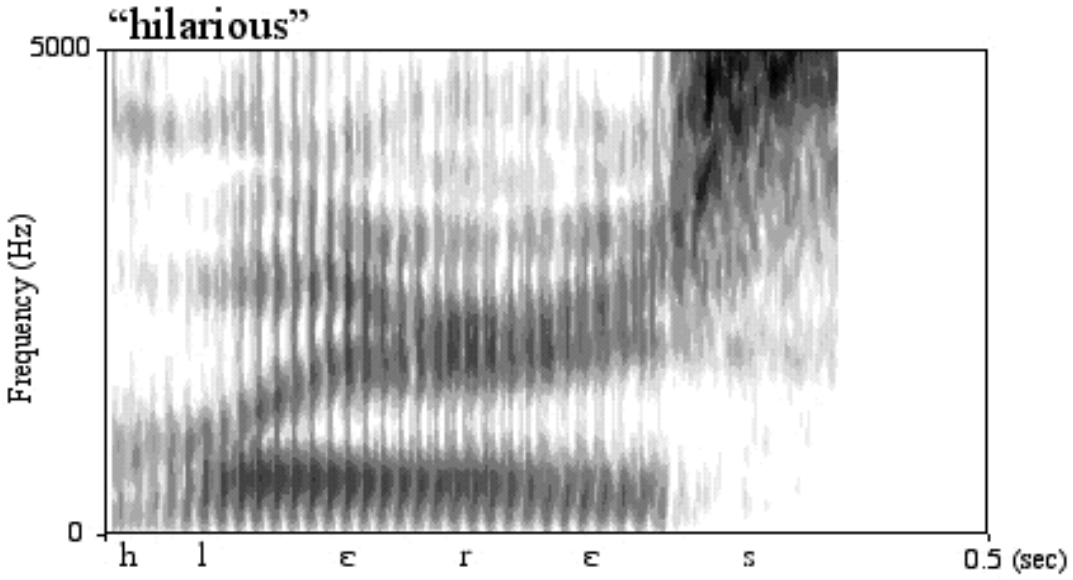

Figure 3: The four syllable word *hilarious* is realized as [hlɛrɛs] in this instance.

Figure 4 shows a final example of "massive" reduction found in the ViC corpus. In this example, *particular* is pronounced as [$p^ht^hɪk^hɚ$]. In addion to reduction from four syllables to two by the deletion of two schwas, we see in this example the deletion of [l] and the glide [j] that normally follows /k/ in the citation form [$p^hət^hɪk^hjəlɚ$].

Though these examples are interesting and perhaps even thought provoking, there is some indication in the literature on phonology and phonetics that massive reduction may be little more than a curiosity, or nusciance factor. If this is the case, then it may be unnecessary for word recognition models to trouble with such odd productions.

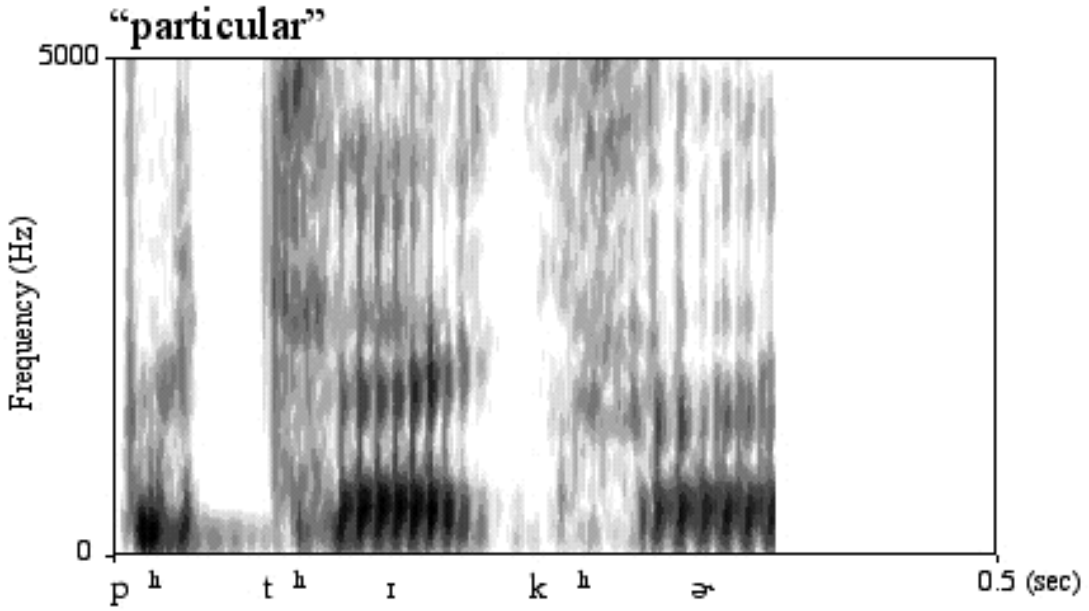

Figure 4: The four syllable word *particular* is realized as [$p^ht^hɪk^hɚ$] in this instance.

2. Phonology

One indication of the possible irrelevance of massive reduction is that many phonologists ignore these "vulgar" or "slovenly" pronunciations. There is a long tradition of this in lexicography. For example, Kenyon & Knott (1944) describe the style of speech that they represented in their "Pronouncing Dictionary of American English" as "cultivated colloquial English in the United States" which they define as "the conversational and familiar utterance of cultivated speakers when speaking in the normal contacts of life and concerned with what they are saying not how they are saying it" (pp. xv-xvi). Given this description of their domain of interest and given the fact that the productions illustrated in figures 1-4 come from the conversational speech of (generally) college educated white folks from the heart of the United States, we might expect to see these pronunciations listed in Kenyon & Knott. But they are not, and probably would not be listed in any pronouncing dictionary of American English. Lacuna of this sort might be explained by asserting that the sky is falling, i.e. that people aren't speaking correctly anymore. However, the correct explanation was given by Knott (1935) who reported that in collecting pronunciations for a dictionary the editor disregards the sounds of words in sentences, despite the description for Kenyon & Knott as "conversational", and deals only with words as spoken in isolation (by speech teachers!).

Some of the most prominent phonological theorists have also avoided heavily reduced forms. For example, Jakobson & Halle (1968) have this to say about "elliptic" speech.

> "Even such specimens as the slovenly /tem mins sem/ for 'ten minutes to seven', quoted by Jones, are not the highest degree of omission and fragmentariness encountered in familiar talk. But, once the necessity arises, speech that is elliptic on the semantic or feature level is readily translated by the utterer into an explicit form which, if needed, is apprehended by the listener in all its explicitness. The slurred fashion of pronunciation is but an abbreviated derivitive of the explicit clear-speech form that carries the highest amount of information. When analyzing the pattern of phonemes and distinctive features composing them, *one must recur to the fullest, optimal code* at the command of the given speakers." (p. 414, italics mine).

In order to delineate the system of information encoded in lingusitic sound systems (Jakobson & Halle's goal), it is indeed important to analyze the optimal code in which forms convey all of the potential contrastive information known to the speaker, so that the grammatical analysis captures what the speaker knows about how to make words distinct (or at least maximally distinct) in his/her language. This is, of course, an interesting and legitimate enterprise which procedes by first removing from consideration the reduced variants of words.

Hockett (1965) made a similar point. "In most languages, if not in all, there is a prescribed pattern for extra-clear speech, to which one resorts when normal rapid speech is not understood, or when certain social factors prescribe it" (p. 220). He notes that pronunciations like [dɪdʒə] and [wʊdʒə] for *did you* and *would you* can be pronounced more clearly as [dɪd yu] and [wud yu] and names these two end-points on the continuum of speech styles the "frequency norm" and "clarity norm". Though Hockett suggests that phonologists should "accept for analysis any utterance which is produced by a native speaker and understood, or understandable, by other native speakers", he goes on to say, "We tend to prefer the frequency norm, but we perhaps do not accept all its consequences; where we refuse to accept its consequences we are referring to the clarity norm instead" because "clarity norm analysis has the merit (if it is a merit) of considerable simplification." (p. 221).

Chomsky & Halle (1968) clearly delineated the domain of their phonological research to cover only the clarity norm/optimal code. They distinguished between a speaker-hearer's competence or "knowledge of grammar" and the implementation of that knowledge. "Performance, that is, what the speaker-hearer actually does, is based not only on his knowledge of the language but on many other factors as well - factors such as memory restrictions, inattention, distraction, nonlinguistic knowledge and beliefs, and so on. We may, if we like, think of the study of competence as the study of the potential performance of an idealized speaker-hearer who is unaffected by such grammatically irrelevant factors." (p. 3). So, by concentrating on the speaker-hearer's competence Chomsky & Halle limited their investigation (as did most other linguists) to Jakobson's "fullest, optimal code".

As we see, prominent leaders in linguistic phonology have more or less explicitly over the years disregarded "slovenly" or "slurred" forms in favor of "explicit" forms, even though "the

number of effaced features, omitted phonemes and simplified sequences may be considerable in a blurred and rapid style of speaking" (Jakobson & Halle, 1968, pp. 413-4). [It should be noted that these remarks refer mainly to phonological theory in the United States. Phonologists in Europe have been more ready to address "frequency-norm" phenomena.] As Hockett's remarks indicate though, there has been some undercurrent of worry on the part of some phonologists that thoeries based on the clarity-norm may be missing something important. Stampe dealt with this in his theory by describing the information structure found in clear-speech forms in a system of learned rules such as would be described in Chomsky and Halle's system, while also describing patterns of reduction found in conversational speech with a system of phonetically natural processes. Zwicky's work on the coding of syntactic information in casual speech phenomena (1972 and later articles) suggests that the worry about missing something important is probably well-founded.

Incidentally, while the phonological theorist's general disregard of highly reduced speech, on the grounds of studying the information content of language sound systems, makes sense, one has to wonder whether it is then folly to take theoretical phonological analysis as the starting point for a theory of auditory word recognition that aims to explicate the listener's ability to cope with phonetic variation (Stevens, 1986; Lahiri & Marslen-Wilson, 1991, 1992). While I am wholely sympathetic with Lahiri and Marslen-Wilson's emphasis on a representational solution to the "segmental bias" found in most contemporary auditory word recognition models, I find it odd to turn, for an account of casual speech phenomena, to a system of representation that has been built primarily on a foundation of clear speech.

3. Phonetics

Because the phonologists' disregard of casual speech is well justified by the nature of the enterprise, it should come as no surprise that massive reduction has not played a very important role in phonological theory (Stampe, was outside of the mainstream on this point). On the other hand, phoneticians who study casual speech phenomena have consistently noted reductions that involve extreme changes from citation forms involving deletion of segments and syllables among other changes.

For example, Cruttenden (1994) notes that "since OE (Old English) it has always been a feature of the structure of English words that the weakly accented syllables have undergone a process of gradation, i.e. the loss of phonemes or obscuration of vowels" (p. 213). He notes that this process has resulted in "established" forms in which the deleted syllable is now a feature of the word regardless of speaking style. For example, *evening, camera, Dorothy*, and *marriage* are now two syllable words in American English though earlier they had three (as reflected in the spelling). Other words show variable deletion of syllables. For example*, family, usually, easily, national*, etc. variably have three or two syllables. Cruttenden also notes a tendency to delete /ə/ and /ɪ/ before /l/ or /r/ in *police, parade, terrific, correct, collision* and others, and after voiceless fricatives as in *photography, thermometer, support, suppose*, and *satirical*. Interestingly though, like Jakobson & Halle, Cruttenden distinguishes these reductions from "vulgar" reductions like *possible* [pɑsbl] and *I'm going to* as [ɑɪŋnə]. Similarly, Shockey (2003) notes cases of "schwa absorption" in *finally*

[faˈn̩li], *Alaska* [l̩æskə], *thousand* [θaʊzn̩], *a new* [n̩u], and a highly reduced production of *that you* [ætʃ]. She also notes that, "At times, phrases which are used repeatedly reduce in ways which are extreme" (p. 46). Shockey describes these less predictable "icons" as idiosyncratic and does not discuss them further. The only examples of "icons" other than place names like [haˈstɛɪʔ] for *Ohio State* that Shockey gives are highly reduced productions of *you know* [j̃ɔ̃] and *you know what I mean?* [j̃ɔ̃w̃ʌ̃m̃ĩ].

Dalby's (1986) study of schwa deletion is particularly relevant. He gave, as an introductory example, some variations of the word *probably* (see 2 below) ranging from three syllables to one. In a study of "consultative speech" produced on television talk shows he found that 9% of all schwas deleted (~18% in medial syllables, 9% initial, and 2% in final syllables). In a second study in which talkers were asked to repeat sentences (in a laboratory setting) very quickly the average deletion rate in fast speech was 43%. One would assume that the deletion rate in ordinary conversational speech would be more like that found in consulative speech (9%) and that Dalby's fast speech results may be more relevant to how speech breaks down in the lab when talkers are given an extreme speaking task.

(2) Reduced forms of *probably*, with a syllable count and number of unchanged phones for each form.

	prɑbəbli	prɑbli	prɑli	prɑˈ
σ	3	2	2	1
# unchanged phones	8/8	6/8	5/8	2/8

Dalby (1986) suggested that the overall strategy in "fast speech" is to "reduce the number of syllables in an utterance while preserving the well-formedness of surface syllabification." (p. 67). One particularly interesting bit of evidence for this was his finding that speakers seemed to make ancillary changes to regularize illegal consonant sequences produced by schwa deletion. For example, he noted that in the phrase "seven minutes" if the schwa of "seven" deletes the potential sequence [vn] doesn't seem to appear, instead the word is [sɛm]. In his fast speech study he found that the schwa deletion rate when deletion would result in a legal consonant sequence was 52%, while the deletion rate when an illegal consonant sequence would result was only 28%. Furthermore, Dalby found that of the 877 deletions resulting in legal sequences only 16 had further consonant changes that resulted in now illegal sequences. On the other hand, of the 284 deletions resulting in illegal sequences, 173 (61%) of these were "legalized" by consonant changes of the sort illustrated in [sɛvn] → [sɛm]. As Dalby noted, this is not too surprising from the point of view of speech production where well-formed syllables may serve as production units (Fujimura and Lovins, 1978). However, from the point of view of auditory word recognition the "regularization" of ill-formed syllables may make a bad situation (schwa deletion) worse (ancillary changes).

4. Rationale for the present study

My point in this review is that there are certain types of reduced speech that are often excluded from consideration by linguists. For most theoretical phonologists studying the information structure of language sound systems, the cutoff line is quite restrictive. Only clear forms, explicitly exhibiting the optimal code, are considered relevant for the enterprise. However, phoneticians and phonologists studying conversational speech phenomena also "draw the line" at some point and don't focus on "extreme" reduced variants.

The question that this paper addresses is a rather limited one. I will examine in an initial and somewhat cursory analysis whether extreme, or as I call them here, "massive", reductions occur with any frequency in conversational speech. The rationale for this analysis is ultimately to comment on the adequacy of psycholinguistic models of auditory word recognition. So after exploring massive reduction in the Variation in Conversation corpus I will discuss the relevance of the findings for auditory word recognition theory. The patterns of reduction found in this cursory study may also be relevant for phoneticians and phonologists as well, in at least providing an estimate of the relative frequency and an initial characterization of massive reduction.

5. Method

The Variation in Conversation (ViC, Pitt, et al., 2003) corpus is a large database of recorded conversational speech. The following description of the corpus is a brief synopsis of the fuller account given in Pitt, et al. (2003).

Forty talkers were from the Columbus, Ohio community. All were natives of Central Ohio, and the sample was stratified for age (under 30 and over 40) and sex, and the sampling was limited to middle-class Caucasians. Talkers were invited to come to the Ohio State University campus to have a conversation about everyday topics such as politics, sports, traffic, schools. After the interview, talkers were debriefed on the conversation's true purpose and all consented to having their speech used in research. Interviews were conducted in a small seminar room by one of two interviewers (one male and one female) who had been trained to conduct sociolinguistic interviews. Talkers sat in a chair facing the interviewer and wore a head-mounted microphone which fed into a DAT recorder.

Talkers spoke a total of 306,652 word tokens, of which ~88,000 have been phonetically transcribed with hand-corrected phonetic labels and will be used in this study. The size of the hand-labeled corpus is approximately twice the size of the TIMIT read-speech corpus (Zue, Seneff, & Glass, 1990). Phonetic transcription proceeds in three steps. First, an orthographic transcription is produced. Second, an HMM-based recognizer performs a forced alignment of dictionary pronunciations onto the acoustic signal (Entropics Aligner). Third, a team of phoneticians (graduate students and post-docs in linguistics) hand-correct the aligner output.

So far in our phonetic transcription effort, recordings of fourteen of the forty speakers have been phonetically tagged. I examined the phonetic transcriptions in this corpus to determine the rate of occurrence of massive reduction.

To measure "massive" reduction, I conducted two analyses. In the first analysis, I counted the number of syllable peaks in the citation form and the number of syllable peaks in the token as it

was transcribed in the corpus. The point of this analysis was to measure the number of times that syllables were deleted. The examples shown earlier in figures 1-4 indicate that sometimes talkers do produce variants that have fewer syllables than the citation form has.

(3) Phonetic transcriptions of citation and massively reduced variants.

2σ ~ 0σ	[bikɔz] ~ [k^{h}z]	(Figure 1)
4σ ~ 2σ	[əp^{h}ɛɾəntli] ~ [p^{h}ɛɾ̃ĩ]	(Figure 2)
4σ ~ 2σ	[hɪlɛriʌs] ~ [hlɛrɛs]	(Figure 3)
4σ ~ 2σ	[p^{h}ət^{h}ɪk^{h}jələ̹] ~ [$p^{h}t^{h}$ɪk^{h}ə̹]	(Figure 4)

In order to find instances of syllable deletion, I defined a set of symbols as syllable peaks. These were the vowels, plus the syllabic nasals, laterals, and rhotics. Then I counted the number of syllable peaks in each citation form and in the corresponding transcribed token from the corpus.

In the second analysis, I counted how many of the segments of each token in the corpus deviated from the corresponding segments found in the citation form. So, in this analysis a form may not match very well at all (i.e. most of the segments have undergone a change of some sort) yet the token and the citation form may have the same number of syllables. This *segmental deviation analysis* then is somewhat orthogonal to the *syllable deletion analysis*, though of course if a syllable has been deleted the missing segments count as mismatched in this analysis. So, the segmental deviation analysis is a bit more sensitive than the syllable deletion analysis because it will pick out a finer grain of deviations from the citation form.

6. Results

The results for syllable deletion are presented first and then come the results for segment deviation. In both cases, the aim of the analysis is to determine whether "massive" reductions are common in conversational speech.

6.1. Syllable Deletion

For each phonetically transcribed word in the ViC corpus (49362 function words and 38560 content words) the number of syllable peaks in the citation form was compared with the number of syllable peaks in the phonetic transcription of the word. Table 1 shows the results. For content words, 35619 tokens had the same number of syllables as found in their citation forms, while 2945 tokens did not have the same number of syllables. This is a syllable mismatch rate of 7.6 %. Most mismatches (78%) involved syllable deletion rather than syllable insertion. For function words, the syllable mismatch rate was lower (5%). 46774 tokens were produced with the same number of syllables found in the citation form, while 2592 were not. As with content words, most syllable count mismatches in function words (86%) involved syllable deletion. So, 6% of the content words and 4.5% of the function words had a deletion of at least one syllable. These syllable deletion rates are more similar to the schwa deletion rates reported by Dalby (1986) in his study of "consultative"

television interview speech (9%), than the high deletion rate that he was able to induce in a fast speech task (43%).

The distributions of syllable counts in the corpus are shown in table 1. The most common deviation from the citation form was the deletion of one syllable. For the monosyllabic function words, 97% were realized with a syllabic element, but syllable deletion occurs in 27% of the two-syllable, 32% of the three-syllable, and 19% of the four-syllable function words. A similar pattern is observed with content words. Short content words, in this case one or two syllables long, tended to maintain the syllable count of the citation form (97% and 93% of the one and two-syllable content word tokens), but syllable deletion rates for longer words were quite high (32%, 26%, 33%, and 59% one-syllable deletions for 3, 4, 5, and 6- syllable words respectively).

Table 1: Comparison of the number of syllables in the citation form of words and the number of syllables in the actual pronounced word in the ViC corpus. The top panel shows percentages for content words and the bottom panel shows percentages for function words.

Content words		citation # syll						
		1	2	3	4	5	6	total
actual # syll	0	0.3	0.1					
	1	97.4	5.9	0.8				
	2	2.3	92.9	32	11			
	3		1.1	66.3	25.8	9.6		
	4			0.9	62.3	32.9	6.9	
	5				0.8	56.2	58.6	
	6						34.5	
totals		22536	11694	3080	972	249	29	38560
Function words		citation # syll						
actual # syll		1	2	3	4	total		
	0	2.2	0.5					
	1	97.1	27.2	4.3				
	2	0.7	71.6	31.5	3.2			
	3		0.6	63.7	19.1			
	4				77.7			
totals		45029	3544	695	94	49362		

So, examples (figures 1-4) like *apparently* [pʰɛɾĩ], and *particular* [pʰtʰɪkʰɚ] are surprisingly frequent in this conversational speech corpus. Eleven percent of all four-syllable content words were produced with only two syllables, suggesting that reductions of this extent are

regularly encountered by listeners. Additionally, the overall rates of syllable deletion (6% for content words and 4.5% for function words) indicate that in conversational speech one of every 20 words will have at least one syllable deletion.

6.2. Segmental deviation

In the segmental deviation analysis I was interested in determining the range of pronunciation variation by comparing the phones of each token in ViC with the phones in the citation form of the word. This analysis will obviously correlate with the syllable deletion analysis because deleted segments can't match the citation form. However, as we saw in the difference between *hilarious* and *particular* (figures 3 & 4), tokens that show the same number of syllable deletions (in these cases two) may nonetheless differ quite a bit in terms of segmental deviation from the citation form. Therefore, the results of this analysis will correlate with the syllable deletion analysis, but may also add a finer grain to our understanding of the variability found in conversational speech.

Table 2 shows an example of the segmental deviation analysis. This table shows variants of the five-segment function word *until*. All of the variants listed in table 2 occurred at least once in the corpus. Each variant is categorized according to how many of the phones in the variant deviate from the citation form. In the case of *until*, the surviving phone is [t] which is present, as a kind of island of reliability, in all variants of the word. Note that in the example transcriptions, the only variants that involve a syllable deletion happen to be those that have 3 deleted phones.

For content words and function words, I counted the number of deviating and deleted phones as illustrated in table 2. The deviation rates (percent of phones that deviated from citation form) and deletion rates (percent of phones that are absent relative to the citation form) are shown as a function of word length, in phones, for content words in figure 5 and function words in figure 6. These figures show that the transcribed phones in the ViC corpus often deviate from the phones given in the citation form.

Content words (figure 5) had deviation rates under 20% in words with three or fewer phones and deviation rates between 20% and 25% for longer words. The deletion rate was also lower for short words (under 5% for words with three phones or fewer) and relatively constant for longer words (7-12% for words longer than three phones).

Table 2: Variants of *until* classified by the number of phones that deviation from the citation form and the number of phones that are deleted relative to the citation form.

Transcription	# deviating phones	# deleted phones
ʌntɪl	0	0
ʌnt̬əl ɛ̲ntɪl	1	
ɛ̲nt̬əl ɪ̲nt̬ɪ̲w̲	2	
_n̥tɪl		1
ə̲nt̬ʌ̲	3	
_n̥t̬_l̥	4	2
__t̬ l̥ __t̬ə̲		3

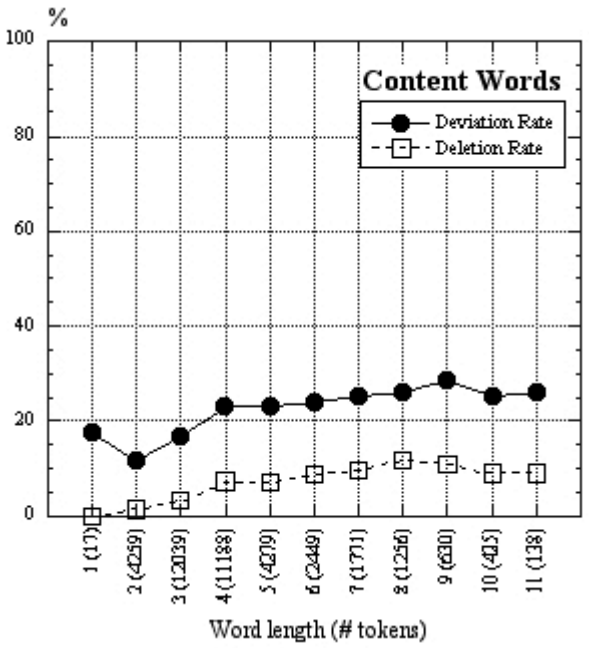

Figure 5: Segmental deviation and deletion in content words. The vertical axis shows percent of phones in content words in the ViC corpus that deviated from citation form (solid circles) or were deleted relative to the citation form (open squares). The horizonal axis shows word length in phones with the number of tokens at each length in parentheses.

Function words (figure 6) had higher deviation and deletion rates, with deviation at about 40% except for one-phone and seven-phone words. The one-phone words include the indefinite article *a* which was very rarely produced in the citation form [e^{I}]. The seven-phone function words include *anybody*, *somebody*, *sometimes*, and *everything*. For some reason, these words showed rates of deletion and change more like content words than function words. The rate of deletion in function words rose over one, two, and three-phone words just as we saw with content words, and then maintained a level between 15% and 20% in longer words.

It is interesting that the deletion rate and deviation rate assymptote for content and function words of four or three phones and longer. This may indicate that there is an upper bound on the permissible deviation from the citation form at about 20% deletion and 40% deviation for function words and 10% deletion and 25% deviation for content words. These are average rates though, with some productions showing very little deviation from citation form, and other productions with higher deviation rates than the average. Nonetheless, it may be that average deletion and deviation rates higher than these assymptotic rates may be more than the auditory word recognition system can handle, and thus we have identified a useful parameter to guide modeling efforts.

Note that 10-20% segment deletions and 25-40% segment deviations are high rates. This comes out more clearly when we look at the number of segment deviations and deletions per word. Most words in the corpus (>60%) deviate from their citation form on at least one phone, and 28% of the words deviate on two or more phones (which is a remarkable deviation rate, given that 68% of the tokens have three or fewer phones). The number of words that have a segment deletion is also quite large. We saw in the syllable deletion analysis that 5-6% of the words in the corpus have at least one syllable deletion. The number of words with a segment deletion is much higher. A little over 20% of the words in the corpus have one segment deleted and 5% have two or more segment deletions. Again, considering the short average word length in the corpus, this is a remarkable deletion rate.

A detailed examination of segmental deviation in three- and four-phone content words is instructive. The three-phone content words are usually (92%) one syllable, while the four-phone content words have one syllable 58% of the time and two syllables 41% of the time. Recall from table 1 that one and two syllable content words rarely show syllable deletion (0.3% and 6% respectively), while longer content words had syllable deletion rates at 30%. So, compared to longer words, the one- and two-syllable tokens (the vast majority of content words) seem to be relatively immune to massive reduction, when this is defined as syllable deletion.

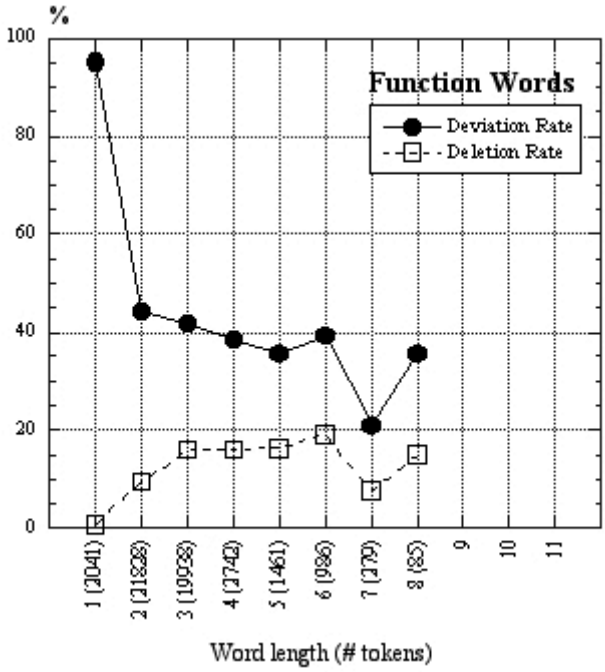

Figure 6: Segmental deviation and deletion in function words. The vertical axis shows percent of phones in function words in the ViC corpus that deviated from citation form (solid circles) or were deleted relative to the citation form (open squares). The horizonal axis shows word length in phones with the number of tokens at each length in parentheses.

Segment deviation analysis tells a different story (figure 5). Three-phone words show a deviation rate of 16% and four-phone words have a 23% deviation rate. The most common deviation in these words was the substitution of one vowel (syllable peak) for another - i.e. vowel reduction without vowel deletion (see Table 3). In 3-phone content words, plosive ([p b t d k g tʃ dʒ ɾ tʔ]) substitutions were the next most common deviation, while plosive deletion was the second most common deviation in 4-phone words followed by plosive substitution. Resonant ([m n ŋ r l ɾ̃]) deletions and substitutions also occurred, while changes involving fricatives and glides were less common. It is interesting to tabulate the deviation rates by word, asking how many words in the corpus differ from the citation form, and how different are they? For 3-phone words we find that 40% of the word tokens deviate from their citation form on one or more of the phones, with 31% differing on one phone, and 8% differing from the citation form on two of the three phones. For 4-phone words we find that 58% of the words deviate from the citation form on one or more of the phones, with 32% differing on one phone, 18.6% differing on two phones, and 6.7% differing on three of the four phones.

Table 3: Segmental deviations in three- and four-phone content words classified by the type of deviation. Percentages indicate how many of the segmental changes fall into each category. Segment classes used in the analysis were, Plosives, Vowels, Resonants, Fricatives, and Glides. 0 indicates deletion. P~0, for example, means a plosive in the citation form corresponds to nothing in the produced token (a deletion).

	3-phone	4-phone
V ~ V	42%	43%
P ~ P	28%	11%
R ~ 0	8%	8%
P ~ 0	4%	17%
R ~ V	3%	<2%
V ~ 0	3%	4%
R ~ R	3%	6%

The phone deletion rates were 3% and 7% for three- and four-phone words, respectively. In three-phone words 18% of the segmental deviations were deletions, while in four-phone words 31% were deletions. Most phone deletions were of plosives or resonants rather than vowels - vowel deletions made up only 16% and 13% of the phone deletions for the three- and four-phone words. In terms of the number of words in the corpus, we can again ask how many words have phone deletions? For 3-phone content words, 7.9% of the tokens had a one-phone deletion and 0.44% were missing two of the three citation form phones. For 4-phone content words, 22% of the tokens had a one-phone deletion and 3% were missing two of the four citation form phones. These phone deletion rates are much higher than the syllable deletion rates for these short words.

The difference between function words and content words in both the syllable deletion analysis and the segment deviation analysis is striking. The syllable deletion rate for function words was somewhat lower than for content words (4.5% vs. 6% respectively) yet in terms of segmental deviation, function words deviate more from their citation forms. This could be taken to suggest that the "citation forms" for function words are inadequate representation of their pronunciations and that function words should generally be modeled differently (perhaps with multiple citation forms). This is similar to the approach taken by Lee (1989) in the Sphinx recognition system, who used word-sized HMMs for the most common function words rather than specifying them in terms of their citation phones.

6.3. Results wrap up

The results of this analysis of reduction in conversational American English suggest that massive reduction is common in normal speech. In the syllable deletion analysis, 4.5% of all function word tokens and 6% of all content word tokens were produced with at least one syllable deleted, and two or more syllables were deleted from 22% of the four to six-syllable words. The

segmental deviation analysis, found assymptotic deletion and deviation rates of 20% deletion and 40% deviation for function words and 10% deletion and 25% deviation for content words. Over 60% of the words in the corpus deviate from their citation form on at least one phone, and 28% of the words deviate on two or more phones. In terms of segmental deletions, over 20% of the words in the corpus have one segment deleted, and 5% have two or more segment deletions. Examination of segmental deviations in short content words, which showed stable syllable counts in the syllable deletion analysis, found that 31% of the 3-phone content words differed from their citation form on one phone, and 8% differed from the citation form on two of the three phones. Approximately 8% of these tokens had one of the three phones deleted.

7. Discussion

The results presented in the previous section show that massive reduction is a regular feature of conversational speech. This finding has a number of important implications for phonetics, phonology and automatic speech recognition, but instead of focusing on these, I conclude the paper with a discussion of auditory word recognition. One reason to focus on auditory word recognition theory is that traditional dictionary assumptions continue to play a central role in this domain, while phoneticians, phonologists, and to some extent ASR researchers have been moving beyond the traditional mindset regarding the representation of the audible/speakable lexicon.

The traditional assumptions are these: (1) that lexical forms are composed of phonetic segments that are analogous to the letters of an alphabetic writing system (the *segmental assumption*), and (2) that lexical forms are stored in a single prototypical or underlying representation (the *single-entry assumption*). I will argue that the phenomenon of massive reduction, together with ancillary observations, shows that neither of these traditional assumptions about the formal representation of lexical items is tenable.

7.1. Single-entry, segmental models

Most current models of auditory word recognition fall into the category of models that adhere to both the segmental assumption and the single-entry assumption. I would put into this category most of the auditory word recogntion models that have been proposed in the last 20 years - including the Cohort model (Marslen-Wilson & Welch, 1978, Marslen-Wilson, 1987), the TRACE model (McClelland & Elman, 1986), Shortlist (Norris, 1994), and others.

Generally, auditory word recognition research has focused on theoretical issues other than lexical representation, and so has not challenged the traditional dictionary assumptions about the mental representation of auditory lexical form (except see section 7.2 below). However, one focus of attention has been on the role (or not) of top-down predictive information in the recognition of alphabetic segments, so in a sense the segmental assumption has been central in framing at least one key research question in auditory word recognition research.

Nonetheless, disregarding issues of top-down and bottom-up interaction and techniques for producing correct patterns of competition among lexical items, all single-entry segmental models operate with a two step word recognition process. In step one the segments are recognized and in

step two segmental information is used to select word candidates. As single-entry models, they also represent each lexical item with just one possible pronunciation in the mental lexicon.

Unfortunately, single-entry segmental models are not descriptively adequate. For example, Scharenborg & Boves (2002) found that the Shortlist model (Norris, 1994) had a 64% word error rate in a large-vocabulary word recognition task using input from HMM phone models. Single-entry segmental models, by virtue of their rigid segmental expectations cannot posit, for example, that [ptɪkɚ] is a possible pronunciation of *particular*, because *particular* starts with [pət] not [pt]. The requirement that segments in the input line up, left-to-right, with segments in the lexical representation prevents single-entry segmental models from being able to cope with deletion (TRACE, has a little more flexibility in this regard). Therefore, single-entry segmental models predict that deletions will short-circuit auditory word recognition, such that words with deleted segments (unless near the end of a relatively long word), let alone words with deleted syllables, could not be identified by listeners. What is more, these models predict that segmental variants, such as the two pronunciations of the indefinite article *a* [e^{I}] and [ə], should interfere with word recognition.

It might be argued that it was demonstrated long ago that speech excised from conversation IS badly perceived (Pickett & Pollack, 1963) - that listeners do have difficulty with elliptic speech just as these models predict. However, the difference between the model and the listener's behavior is that for the listener the correct lexical item remains active even though the signal does not specify that item very well. So, with additional context the correct item can be recognized (Shockey, 1998).

I am not suggesting that 20 years of research now needs to be scrapped because of the revolutionary discovery that people delete syllables in conversational speech. However, I am suggesting that the input-to-lexicon mapping which has been assumed in these models needs to be revised. As I mentioned, most researchers have taken the single-entry and segmental assumptions on authority (e.g. Chomsky & Halle, 1968) and have been concerned primarily with other issues.

7.2. Single-entry, nonsegmental models

Drawing from ideas in linguistic theory (post-1968) regarding the autosegmental representation of lexical items and the underspecification of noncontrastive or unmarked phonetic features (see Goldsmith, 1990), Lahiri (1999; Lahiri & Marslen-Wilson, 1991, 1992) proposed a single-entry nonsegmental model of auditory word recognition. One of the chief claims for this theory was that the human auditory word recognition system uses a featurally underspecified lexicon (FUL) and that this accounts for the listener's ability to disregard mismatching assimilated features such as the [m] in *lea[m] bacon*. Because place of articulation is underspecified in [n] (meaning that even though it is pronounced with the tongue touching the alveolar ridge, the lexical representation does not have any indication that the word ends with a [coronal] nasal - this feature is "unmarked" for nasals in English and filled in by default feature specification rules) the [m] in the assimilated form *lea[m]* does not mismatch any place feature in the lexical representation, and

thus the lexical representation is compatible with the assimilated form as well as the citation pronunciation.

Feature underspecification is a clever method of soaking up variance by reducing the number of features that must be matched in lexical access. By leaving [t] underspecified for [coronal] this model predicts that assimilated variants like *freigh[p] boss* and *freigh[k] guy* will be recognized as *freight* because they do not mismatch the representation of [t] in which place of articulation is not specified. In this model, the "non mismatch" of the [labial] feature of [p], or the [dorsal] feature of [k] with the underspecified place representation of [t] makes it possible to match any one of the stops, [p], [t], or [k], to the lexical representation of [t] in *freight*.

Scoring "non mismatches" where the input has some extra feature not found in the lexical representation, together with the non-segmental featural representation gives FUL the ability to deal with massive reduction. For instance, when faced with the massively reduced form of *particular* [ptɪkɚ], non mismatches occur for the "deleted" segments/features when a feature is extracted from the signal and the lexicon does not specify a contradictory feature. Because feature non mismatches arising from "deleted" segments do not actively eliminate *particular* from the cohort of lexical candidates that are consistent with the input, the model predicts that correct recognition may occur even though some segments have been deleted.

Even though this lax treatment of non mismatching features is the key to compensating for deletion in FUL, non mismatch features should be penalized in this model. In dealing with assimilation this is not obvious, but it is untenable to score deleted features as cost-free non mismatches, because, in the model as described by Lahiri (1999), the input [kæp] results in a cohort of equally viable candidates *cap, camp, crap, clap, car wrap, Karnap, catnip, catnap, cattle prod*, and *cat that Erin knew that Chris believed that the dog persued*. The features of [r] in *crap*, for example, are simply not present in [kæp] but nothing in this input actively mismatches the features of [r], so the activation of *crap* (and an infinite number of other words/phrases) is equivalent to the activation of *cap* as far as the model is concerned. If non mismatches are penalized to avoid the proliferation of deleted features, then the size of the penalty determines the behavior of the model. With a heavy non mismatch penalty the results are comparable to the results for segmental models, and with light non mismatch penalties it may be possible to posit the deletion of highly underspecified segments like [ə] or [t] which are each specified for only one feature, while ruling out more fully specified segments. This seems to be a desirable result, predicting that [ə] is likely to delete while [ɪ] is less likely to delete because [ɪ] is specified on more features.

Though Lahiri's nonsegmental single-entry model is in many ways an advance over segmental models, there are good reasons to believe that the underspecification approach is ultimately inadequate. I will briefly mention three: (1) phonological "inferencing" effects, (2) the featural unpredictability of deletion, and (3) subcategorical phonetic residue.

Phonological "inferencing" effects are a serious problem for FUL. Gaskell & Marslen-Wilson (1996) rejected FUL in favor of a phonological inferencing model because FUL predicts that *freigh[k]*, *freigh[p],* and *freight* are equally good realizations of *freight* in the phrase

freight boss. Their data suggest that the phonological viability of the variant has a strong impact on the activation of the word, so that *freigh[p] boss* is likely to sound like it has the word *freight* in it while *freigh[k] boss* is less likely to activate *freight*.

A second serious problem for FUL is that deletion is to some extent unpredictable from featural specifications alone. For instance, the realizations of *until* (see table 2) suggest that underspecified segments like [t] sometimes serve as lexical islands of reliability and are virtually undeletable. Thus, predicting when deletion (or assimilation) will be likely to happen is not as simple as saying that underspecified segments like [ə] and [t] will delete while more fully specified segments like [l] will not. It may be possible to deal with this objection by adding prosodic specification to the lexicon, though this move runs counter to the reliance of FUL on a particular kind of radically underspecified representation.

Finally, phonological processes such as deletion and assimilation leave a phonetic "residue" of cues such that supposedly identical strings such as [ʃɪpɪŋ] in "shipping" and "ship in(quiry)" (Norris, 1994) are not phonetically identical and listeners are sensitive to these subcategorical bits of phonetic information (Manuel, 1991, 1995; Whalen, 1984, 1991; Marslen-Wilson & Warren, 1994; Dahan, Magnuson, Tanenhaus & Hogan, 2001). In a model that relies on rough feature detectors (as FUL does), gradient perceptual consequences for varying degrees of subcategorical mismatch cannot be captured. The fact that subtle phonetic detail (Repp & Liberman, 1987) plays an important role in speech perception is antithetical to the design of FUL, which absorbs variation by picking up as little detail from the signal as possible.

7.3. Segmental, multiple-entry models

Pronunciation dictionaries generally list variant pronunciations when a word is pronounced differently in different regions or when consultants differ on their preferred pronunciations (Knott, 1935). For example, some people say *tom[eˈ]to* while others say *tom[a]to*, some say *roof* with [u] while others use [ʊ], and sometimes deletions figure in pronunciation variation as in the difference between *laboratory* and *lab'ratory*. As I've indicated in writing these examples, pronunciation variation can be expressed in lists of alternate pronunciations using alphabetic symbols.

Scharenborg & Boves (2002) added multiple-entry lexical representations to Shortlist and found that the word error rate dropped substantially from 64% to 48%. This result suggests that complaints about the adequacy of segmental single-entry models might be answered by simply adding variant forms like [ptɪkə˞] to the lexicon. No further modifications would be needed then. Of course a 48% error rate is too large, but this may have to do with an unusually error-prone automatic phone recognizer rather than the architecture of the model per se.

There is a large literature exploring multiple-entry methods in ASR, and some important observations come out of this literature. It should be noted at the outset that HMMs encode and make use of a huge amount of variability even in single-entry, segmental models (the most common approach). However, as researchers have attempted to build recognizers that are capable of recognizing conversational speech, many have begun to explore multiple-entry lexica (for a review see Strik & Cucchiarini, 1999). I will mention two findings from this literature.

First, segmental multiple-entry models introduce confusions into the lexicon, and these confusions result in almost as many recognition errors as improvements. An example of lexical confusion was given by Fosler-Lussier et al. (2002). When variants that have final [t] deletion are added to the lexicon the recognizer may be unable to distinguish *can't elope*, *can elope*, and *cantaloupe*. Other examples include monomorphemic confusions produced by [t] deletion (*mist ~ miss, went ~ when; mast ~ mass)*, [ə] deletion (*ago ~ go, apart ~ part, about ~ bout*), flapping (*writer ~ rider*), vowel reduction (*are ~ our*), and so on.

Kessens, Strik & Cucchiarini (2002) in a test of their segmental multiple-entry recognizer found, using a lexicon with an average of 3.7 variants per word, and, crucially, a language model trained to predict the most likely variant for any particular sentence context, that there were 489 "improvements" (words recognized that would have been missed without the additional variants), and 301 "deteriorations" (words that were no longer recognized correctly due to increased lexical confusion). So, there was a net gain for adding pronunciation variation to the lexicon but the gain was very small (188 improvements for a test set of 60,087 words!). Results showing increased recognition accuracy in this small range are the norm rather than the exception in this literature and one gets the impression that the problem of lexical confusibility has itself become the focus of research, with the aim being to find ways to include only those lexical variants that will not result in too much increased confusability (Fosler-Lussier et al., 2002), though it has apparently not been demonstrated that rule based variant generators will ever accomplish this.

Second, the problem with segmental multiple-entry models is that they are segmental. Saraçlar & Khudanpur (2000), in work on pronunciation modeling in conversational speech recognition, discussed the "intrinsic ambiguity of phone level transcriptions" which is due to "partial pronunciation change". Their point was that the lexical confusion which is introduced in the segmental multiple-entry approach is exagerated by the use of a fixed set of phonetic symbols. For example, when restricted to a small inventory of phonetic symbols, the string [ʃɪpɪŋ] matches both *shipping* and *ship in(quiry)*. The intrinsic ambiguity of the symbol [ŋ] is that it stands for acoustic signals that are systematically different in *shipping* and *inquiry*. Saraçlar & Khudanpur found that "the acoustics of a phoneme /X/, when realized as a phone [Y], lie somewhere between the average realization of the phoneme /X/ and the phone [Y]". This finding indicates that the gradient realization of acoustic cues in speech provides information about the words being spoken (see also Ellis & Hardcastle, 2003, on the articulatory basis of this effect in one common kind of assimilation). As was mentioned earlier, it has been shown that listeners are sensitive to the fine-grained phonetic residue of reduction processes (Manuel, 1991, 1995; Whalen, 1984, 1991; Marslen-Wilson & Warren, 1994; Dahan, Magnuson, Tanenhaus & Hogan, 2001). If *when* and *wen(t)* remain phonetically distinct (as these studies suggest) then a segmental multiple-entry lexicon that lists both [wɛnt] and [wɛn] as pronunciations of *went*, creates more confusion than actually exists in the signal because [wɛn] from *when* is different from [wɛn] from *went*. The key point here is that the acoustic variability that remains after deletions or substitutions is not simply noise to be removed during phone recognition, instead this variation is useful in word recognition and should not be disregarded.

7.4. Nonsegmental, multiple-entry models

In this discussion of auditory word recognition models we have been distinguishing between models based on their representations of lexical knowledge. Through our familiarity with dictionaries it is natural to think of auditory words in terms of a single-entry segmental representations. However, we have seen that neither the segmental assumption nor the single-entry assumption can be maintained. The key point is that the hearer's knowledge about the auditory forms of words is nonsegmental and includes detailed pronunciation variation. There are probably many ways to model the process of auditory word recognition such that these key aspects of lexical knowledge are maintained. I will conclude by presenting an approach that I developed some years ago within the framework of exemplar-based models of memory (Hintzman, 1986; Nosofsky, 1989, Estes, 1994; Shanks, 1995).

Klatt (1979) proposed a nonsegmetnal model of speech perception that was based on the principle that it is important to "avoid early commitments". His Lexical Access from Spectra (LAFS) model contains no level of segments or features to intervene between the signal and word forms, but instead recognizes words based on how well their acoustic/spectral shape matches stored finite-state models of possible spectral sequences representing words. Thus, the auditory memory of a word is a sequence of spectra, like a neural spectrogram (Crowder, 1981). In 1986, Klatt reported that he was unable to build a satisfactory implementation of the LAFS. He said, "I made some preliminary attempts to build a simulation, but was discouraged by the behavior of the distance metrics available to compare spectra. These metrics were as sensitive to irrelevant spectral variability as to cues to fine phonetic distinctions." (pp. 169-70).

I suggested in Johnson (1997b) that the main problem with Klatt's model was that he adopted the single entry assumption. As we have seen, in most approaches to modeling pronunciation variation using multiple-entry lexica a few alternative pronunciations are added to the lexicon. For example, *roof* might be listed in the lexicon as both [ruf] and [rʊf], but this segmental multiple-entry approach does not solve Klatt's problem - that we would like to be able to use small, fine-grained phonetic details such as can be represented in a spectrogram, while retaining the ability to disregard other phonetic details, such as those introduced by talker differences or variant pronunciations. With the X-MOD model I suggested something a bit more radical than the typical segmental multiple-entry approach.

X-MOD (Johnson, 1997a,b) is an extension of Klatt's (1979) LAFS model that assumes that lexical items are exemplar-based categories in memory (Nosofsky, 1986; implementational details are given in Johnson, 1997b). In rough outline, the model calculates an auditory spectral representation of incoming speech and sweeps this representation over an exemplar covering map (Kruschke, 1992). Exemplars in the map respond to the input in proportion to their similarity to the input and feed activation to abstract word nodes. Weights between covering map locations and word nodes are trained using back propagation of error. Additionally, if the input exemplar is not similar to any exemplar in the covering map it is added to the map.

As in LAFS, lexical items are activated with no level of representation intervening between the auditory neural spectrogram and the lexical item. This has the advantages that Klatt claimed for

LAFS (1989, p. 194). Additionally, the exemplar mode of storage permits the model to capture and use variation in the input. So, in the representation of *particular* there are variants that start with [pt] so that the acoustic phonetic details of [pt] which has been derived from /pət/ supports the identification of the word using fine phonetic details of how this sequence sounds when a schwa has been deleted. The exemplar approach thus allows phonetic detail to coexist with variation in a lexical item's representation.

The resilence of [t] in *until* (despite its phonological underspecification) and other cases of word-specific patterns of phonetic variation, are also captured in X-MOD by letting the perceptual representation be a collection of exemplars. Lavoie (2002) reported that the word *for* is often realized without /r/-coloring as [fə], while *four* always has /r/ or an /r/-colored vowel. The idea that words may have unique ranges of variation, essentially serving as their own domains for word-based phonological patterns or histories may seem odd given the segmental assumption and the single entry assumption, but the phenomenon of lexical diffusion (word-by-word spread) of sound change is a well-attested phenomenon in historical phonology (Wang, 1997). Exemplar storage of word forms correctly predicts that word-specific pronunciation variation and thus lexical diffusion will exist.

The phenomenon of phonological inferencing (Gaskell & Marslen-Wilson, 1998) is not predicted by an auditory-only exemplar-based model. In my 1997a paper I discussed the possibility that self-generated exemplars will have both auditory and articulatory representations, providing a speaker-specific mapping from acoustic/auditory output to articulatory gestures (see Johnson, Ladefoged & Lindau, 1993 concerning individual differences in this mapping). Thus, with an articulatory/gestural interpretation provided to at least some exemplars, phonological inferencing may be facilitated by reference to gestural exemplars.

8. Conclusion

The root of the problem posed by massive reduction is that, given conventional assumptions about lexical form (the segmental assumption and the single entry assumption), it seems that normal speech communication should not be possible with massively reduced forms. Examples such as those given by Stampe, if considered at all, would have to be dismissed as curiosities concocted by an over-imaginative linguist trying to make an obscure theoretical point. In this paper I have argued that massive reductions do occur frequently in conversational speech. Therefore, models of auditory word recognition, that aim to account for anything beyond laboratory speech, must abandon traditional "dictionary" assumptions about the auditory mental lexicon.

Acknowledgment

Research on the ViC corpus has been funded by the Office of Research at Ohio State University and NIH grant #R01DC04330. I thank Robin Dautricourt, Mark Pitt, and Beth Hume for their comments on an earlier draft of this paper. I thank Bill Raymond and Matt Makashay for their help in dealing with the corpus.

References

Chomsky, N. and Halle, M. (1968) *The Sound Pattern of English*. NY: Harper and Row.

Connine, C.M., Blasko, D.G. and Titone, D. (1993) Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, 32, 193-210.

Craigie, W.A. and Hulbert, J.R. (1938-1941) *A Dictionary of American English on Historical Principles*. Chicago.

Crowder, R.G. (1981) The role of auditory memory in speech perception and discrimination. In Myers, T., Laver, J. and Anderson, J. (eds.) *The Cognitive Representation of Speech*. Amsterdam: North Holland (pp. 167-179).

Cruttenden, A. (1994) *Gimson's Pronunciation of English* (5th ed.). London: Edward Arnold.

Cutler, A. (1998) The recognition of spoken words with variable representation. *Proceedings of the ESCA Workshop on the Sound Patterns of Spontaneous Speech: Production and Perception*. Aix-en-Provence, 83-92.

Dahan, D., Magnuson, J.S., Tanenhaus, M.K. and Hogan, E.M. (2001) Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, 16, 507-534.

Dalby, J. (1986) *Phonetic Strucure of Fast Speech in American English*. Bloomington, IN: Indiana University Linguistics Club.

Ellis, L. and Hardcastle, W. (2002) Categorical and gradient properties of assimilation in alveolar to velar sequences: Evidence from EGP and EMMA. *Journal of Phonetics*, 30, 373-396.

Estes, W. K. (1994) *Classification and Cognition*. Oxford: Oxford University Press.

Fosler-Lussier, E., Amdal, I. and Kuo, H-K. J. (2002) On the road to improved lexical confusability metrics. In *ISCA Tutorial and Research Workshop. Pronunciation Modeling and Lexical Adaptation for Spoken Language*. PMLA2002, Estes Park, Colorado (http://www.clsp.jhu.edu/pmla2002/cd/).

Forster K.I. (1976) Accessing the mental lexicon. In Wales, R.J. and Walker, E. (eds.) *New Approaches to Language Mechanisms*. Amsterdam: North-Holland (pp.257-287).

Fujimura, O. and Lovins, J. (1978) Syllables as concatenative phonetic units. In Bell, A. and Hopper, J.B. (eds.) *Syllables and Segments.* Amsterdam: North Holland (pp. 107-120).

Gaskell, M.G. and Marslen-Wilson, W.D. (1996) Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 144-158.

Gaskell, M.G. and Marslen-Wilson, W.D. (1998) Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 380-396.

Goldsmith, J. (1990) *Autosegmental and Metrical Phonology*. Cambridge: Blackwell Publishers.

Hintzman, D.L. (1986) 'Schema abstraction' in a multiple-trace memory model. *Psychological Review*, 94, 411-428.

Hockett, C. (1965) *A Manual of Phonology*. IJAL Monograph.

Jakobson, R. and Halle, M. (1968) Phonology in relation to phonetics. In Malberg, B. (ed.) *Manual of Phonetics*. Amsterdam: North Holland (pp. 411-449).

Johnson, K. (1997a) Speech perception without speaker normalization. In Johnson, K. and Mullennix, J.W. (eds.) *Talker Variability in Speech Processing*. NY: Academic Press (pp. 145-166).

Johnson, K. (1997b) The auditory/perceptual basis for speech segmentation. *OSU Working Papers in Lingusitics*, 50, 101-113.

Johnson, K., Ladefoged, P. and Lindau, M. (1993) Individual differences in vowel production. *Journal of the Acoustical Society of America*, 94, 701-714.

Kenyon, J.S. and Knott, T.A. (1944) *A Pronouncing Dictionary of American English*. Springfield, MA:

Merriam-Webster.

Kessens, J.M., Strik, H. and Cucchiarini, C. (2002) Modeling pronunciation variation for ASR: Comparing criteria for rule selection. In *ISCA Tutorial and Research Workshop. Pronunciation Modeling and Lexical Adaptation for Spoken Language*. PMLA2002, Estes Park, Colorado (http://www.clsp.jhu.edu/pmla2002/cd/).

Klatt, D.H. (1979) Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7, 279-312.

Klatt, D.H. (1989) Review of selected models of speech perception. In Marslen-Wilson, W.D. (ed.) *Lexical Representation and Process*. Cambridge, MA: MIT Press (pp. 169-226).

Knott, T.A. (1935) How the dictionary determines what pronunciations to use. *Quarterly Journal of Speech*, 21, 1-10.

Kruschke, J. (1992) ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 28, 43-67.

Lahiri, A. (1999) Speech recognition with phonological features. *International Congress of Phonetic Sciences*, San Francisco, 715-718.

Lahiri, A. and Marslen-Wilson, W.D. (1991) The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, 38*,* 245-294.

Lahiri, A. and Marslen-Wilson, W.D. (1992) Lexical processing and phonological representations. In Docherty, G.J. and Ladd, D.R. (eds.) *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge: Cambridge University Press (pp. 229-254).

Lavoie, L. (2002) Some influences on the realization of for and four in American English. *Journal of the International Phonetic Association*, 32, 175-202.

Lee, K-F. (1989) *Automatic Speech Recognition: The Development of the SPHINX System*. Boston, MA: Kluwer Academic Publishers.

Manuel, S.Y. (1991) Recovery of "deleted" schwa. *PERILUS XIV*: *Papers from the Symposium on Current Phonetic Research Pardigms for Speech Motor Control*. Institute of Linguistics, University of Stockholm (pp. 115-118).

Manuel, S.Y. (1995) Speakers nasalize /ð/ after /n/, but listeners still hear /ð/. *Journal of Phonetics*, 23, 453-476.

Marslen-Wilson, W. (1987) Functional parallelism in spoken word-recognition. *Cognition*, 25, 71-102.

Marslen-Wilson, W. and Warren, P. (1994) Levels of perceptual representation and process in lexical access. *Psychological Review*, 101, 653-675.

Marslen-Wilson, W. and Welch, A. (1978) Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology*, 63, 10-29.

McClelland, J.L. and Elman, J.L. (1986) The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.

Mencken, H.L. (1948) *Supplement II: The American Language, An Inquiry into the Development of English in the Unitied States*. NY: Alfred A. Knopf.

Norris, D. (1994) Shortlist: a connectionist model of continuous speech recognition. *Cognition*, 52, 189-234.

Norris, D., McQueen, J.M. and Cutler, A. (2000) Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23, 299-370.

Nosofsky, R.M. (1986) Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39-57.

Pickett, J.M. and Pollack, I. (1963) Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech*, 6, 151-164.

Pitt, M.A., Johnson, K., Hume, E., Kiesling, S. and Raymond, W. (2003) The ViC corpus of conversational speech. Manuscript submitted to *IEEE Transactions on Speech and Audio Processing: Special Issue on Spontaneous Speech Processing*.

Repp, B.H. and Liberman, A.M. (1987) Phonetic category boundaries are flexible. In Harnad, S. (ed.) *Categorical Perception: The Groundwork of Cognition*. Cambridge: Cambridge University Press (pp. 89-112).

Saraçlar, M. and Khudanpur, S. (2000) Properties of pronunciation change in conversational speech recognition. In *Proceedings of the 2000 Speech Transcription Workshop*. University of Maryland - National Institute of Standards and Technology (http://www.nist.gov/speech/publications/tw00).

Scharenborg, O. and Boves, L. (2002) Pronunciation variation modelling in a model of human word recognition. In *ISCA Tutoral and Research Workshop. Pronunication Modeling and Lexical Adaptation for Spoken Language*. PMLA2002, Estes Park, Colorado (http://www.clsp.jhu.edu/pmla2002/cd/).

Shanks, D.R. (1995) *The Psychology of Associative Learning*. Cambridge: Cambridge University Press.

Shockey, L. (1998) Perception of reduced forms by non-native speakers of English. In Duez, D. (ed.) *Sound Patterns of Spontaneous Speech*. Aix: ESCA (pp. 97-100).

Shockey, L. (2003) *Sound Patterns of Spoken English*. Cambridge: Blackwell.

Stampe, D. (1973) *A Dissertation on Natural Phonology*. PhD Diss. University of Chicago.

Stevens, K.N. (1986) Models of phonetic recognition II: A feature-based model of speech recognition. In Mermelstein, P. (ed.) *Proceedings of the Montreal Satellite Symposium on Speech Recognition*. Twelfth International Congress on Acoustics (pp. 67-68).

Strik, H. and Cucchiarini, C. (1999) Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 27, 225-246.

Wang, W.S-Y. (1977) (ed.) *The Lexicon in Phonological Change*. The Hague: Mouton.

Whalen, D.H. (1984) Subcategorical phonetic mismatches slow phonetic judgements. *Perception and Psychophysics*, 35, 49-64.

Whalen, D.H. (1991) Subcategorical phonetic mismatches and lexical access. *Perception and Psychophysics*, 50, 351-360.

Zue, V., Seneff, S. and Glass, J. (1990) Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9, 351-356.

Zwicky, A. (1972) On casual speech. *Papers from the Eighth Regional Meeting of the Chicago Linguistic Society*. Chicago: CLS. (pp. 607-615).