July 2007

Commentary: Modeling listeners

Edward Flemming
flemming@mit.edu

1. Introduction

The precise pronunciation of speech sounds is highly variable even within the speech of a single individual. A complete theory of phonetic realization must account for this variability, identifying any systematic conditioning factors. It is natural and expected that details of phonetic realization should depend on phonetic/phonological factors such as segmental and prosodic context and speech rate. Rathcke & Harrington (this volume) demonstrate an effect of this kind, providing evidence that the precise realization of a German nuclear pitch accent depends on the number of unstressed syllables that follow it. It is perhaps more surprising that phonetic realization is also affected by non-phonological factors such as word frequency, lexical neighborhood structure (e.g. Wright 2004, Munson & Solomon 2004), and contextual predictability of words (e.g. Bell et al 2003). Why do these non-phonological factors affect phonetic realization? As discussed by Wright (2004), Lindblom's (1990) Hypo- and Hperarticulation (H&H) theory of speech production provides a parsimonious account of these effects based on the observation that all of these factors also affect the ease with which words are recognized. H&H theory connects production to word recognition via the hypothesis that speakers try to facilitate communication by speaking more clearly where the listener is likely to have greater difficulty with word recognition. Consequently any factor that makes a word more difficult to recognize is expected to make a speaker more inclined to produce it clearly. This general model of production is investigated by Pluymaekers et al (PEBB) and Scarborough.

This commentary addresses two issues raised by these 'listener-oriented' models of speech production. First, the hypothesis that speakers are listener-oriented explains the effect of variables such as neighborhood density on phonetic realization in terms of the effects of the same variables on word recognition, but this in turn raises the question why these variables affect recognition. We will develop an analysis of the effects on word recognition according to which they are consequences of listeners adopting a Bayesian approach to the problem of word recognition, building on proposals by Jurafsky (1996) and Norris (2006). We will see that a Bayesian analysis can account for the importance of frequency, neighborhood density and predictability and has the further implication that these effects are not independent. Fundamentally they all concern the contextual predictability of words – the target word and its competitors. This conception leads to the correct prediction that the effects of frequency and neighborhood density on word recognition should interact with contextual predictability.

Once the basis for the word recognition effects has been clarified, we can turn to the question of how close the match is between production and recognition – how closely do speakers track listener difficulty in modulating the clarity of their speech? In broad outlines, there is obviously a good correspondence between production and recognition effects since this is what motivated applying H&H theory to the analysis of the production phenomena in the first place (e.g. Wright 2004), but Scarborough tests a novel prediction that the effects of

word frequency and neighborhood density should be reduced where contextual predictability is higher. This interaction is observed in word recognition, so if speakers track listener difficulty closely the same effect should be observed in production. Scarborough failed to confirm this prediction, and we will consider the implications of this result.

PEBB raise a rather different question about how closely speakers track listener difficulty. So far we have been discussing the possibility that speakers vary the clarity of words according to anticipated speaker difficulty, but the analysis of word recognition implies that different parts of a word can vary in their importance to lexical access depending on the role they play in distinguishing words from competitors. So we can ask whether speakers modulate clarity segment by segment rather than word by word. We address this development of H&H theory at the end of the commentary.

2. A Bayesian analysis of word recognition

The problem of spoken word recognition can be construed as identifying the word that is most likely to be represented by a given speech signal. Listeners have two sources of information in this process: bottom-up information from the speech signal itself, and top-down information about what words are likely to occur in the context. Bayes' Rule specifies how to combine these two sources of information to derive an assessment of the probability of a word *w* given evidence *E* from the speech signal (1). $p(w|E)$ is the probability of word *w* given *E*, $p(w)$, is the prior probability of word *w* based on top-down information, $p(E)$ is the prior probability of the evidence, i.e. the probability of the speech signal itself, and $p(E \mid w)$ is the probability of word *w* being realized by the signal *E*.

Assuming that the relevant speech signal is the realization of a word, then the prior probability of the signal, $p(E)$, is obtained by summing, for all the words in the lexicon, the probability of each word being realized by that signal, multiplied by the prior probability of that word, so (1) can be expanded as shown in (2)[1]. The probabilities of words being realized by particular speech signals, $p(E \mid w_i)$, must be derived from a generative model of the phonetic realization of words – i.e. a stochastic grammar of phonetic realization. The prior probability of a word, $p(w)$, can depend on a wide variety of contextual factors, but in tasks involving recognition of isolated words, prior probability can often be estimated based on relative frequency of words.

$$(1) \qquad p(w \mid E) = \frac{p(E \mid w)p(w)}{p(E)}$$

$$(2) \qquad p(w \mid E) = \frac{p(E \mid w)p(w)}{\sum_{w_i \in lexicon} p(E \mid w_i)p(w_i)}$$

The Bayesian approach to word recognition is standard in automatic speech recognition (e.g. Rabiner & Juang 1993:434f.) and Jurafsky (1996) and Norris (2006) argue that it provides a good characterization of human word recognition as well. As they show, it

[1] To allow for novel words, some probability mass should be reserved for unknown words. However this modification would not affect any of the reasoning below.

accounts for observed effects of frequency, neighborhood density and contextual predictability on word recognition performance. The primary measures of performance are accuracy and reaction time in word recognition tasks. Predictions about accuracy follow directly from the Bayesian formulation of the word recognition task, but to derive predictions about reaction times it is necessary to make some minimal assumptions about the nature of the processes involved in word recognition. We follow Norris (2006) in assuming that evidence is accumulated over time as the signal is produced and perceptually analyzed, and that listeners will generally identify a signal as word *w* when the probability of that word exceeds some threshold, so in general where less signal-dependent evidence is required to reach the threshold, we expect faster reaction times.

2.1 Frequency effects in word recognition

It is very well established that more frequent words are identified more rapidly and accurately (e.g. Goldinger et al 1996). In terms of the Bayesian analysis, higher frequency of a word *w* implies a higher prior probability of that word, $p(w)$. Accordingly less bottom-up evidence is required to reach a given threshold probability and reaction times are faster. The difference in accuracy arises because Bayesian listeners are biased to respond with more frequent words, so given ambiguity between a higher frequency word and a lower frequency word they will respond with the higher frequency word, which results in a higher percentage of correct identifications for high frequency words, on average.

2.2 Neighborhood density effects in word recognition

Luce, Pisoni & Goldinger (1990) show that frequency alone is a relatively poor predictor of word recognition performance because it neglects the effects of competition between phonetically similar words, i.e. neighbors. In identifying a spoken word, a listener has to eliminate all other words in the lexicon as candidates for the identity of the uttered word. This task is expected to be more difficult where there are many words that are phonetically similar to the target word, i.e. where it is in a dense lexical neighborhood. It has also been found that the strength of this competition from neighbors depends on their frequencies: higher frequency neighbors impede word recognition more than lower frequency neighbors. For example, Luce et al (1990) found effects of word frequency and frequency-weighted neighborhood density on the accuracy of identification of CVC words presented in noise: Higher frequency words and words from sparse neighborhoods were identified more accurately. They also found that reaction times in a lexical decision task were longer for non-words that had more higher frequency neighbors than for words with fewer or lower frequency neighbors.

These competition effects are represented by the denominator of the expression in (2). The denominator sums the products of the prior probabilities of each word in the lexicon multiplied by the probability of that word giving rise to the observed speech signal, $p(E|w_i)$. The latter probability is highest for words that are most similar to the word being spoken – i.e. its neighbors – since similar words are more likely to give rise to the same signal. Multiplying $p(E|w_i)$ by the prior probability of w_i, $p(w_i)$, means that the effect of neighbors is weighted by their frequencies, as observed by Luce et al. So the more high frequency neighbors a word has, the larger the denominator of (2) is going to be, reducing the posterior

probability of the target word $p(w|E)$. Under these circumstances there is greater probability of error in word identification, and it will take more time to accumulate sufficient evidence to reject the real word neighbors of a non-word in lexical decision.

As Jurafsky (1996) observes, Luce's (1986) neighborhood probability rule for predicting the probability of correctly identifying a word stimulus is similar to the Bayes Rule formulation in (2), but substitutes probabilities of confusion between words in place of the likelihood terms, $p(E|w_i)$. The qualitative predictions are thus similar, but the Bayesian formulation has a principled basis, and can be generalized to encompass effects of context, as we will see below.

The Bayesian analysis implies that all words should contribute to competition effects to the extent that they are similar to the target word. In other words, all words are neighbors, but some are closer neighbors than others. This differs from the operational definition of neighborhood adopted by Scarborough according to which the neighbors of a word are words that differ from it by the addition, deletion or substitution of one phoneme. However, as Scarborough notes, this definition has always been regarded as a crude approximation to a more general measure of word similarity of essentially the kind implied by the Bayesian formulation (Luce 1986:6).

2.3 Contextual predictability effects in word recognition

When words are more predictable due to their context in a sentence they are identified more accurately (e.g. Boothroyd & Nittrouer 1988, Sommers & Danielson 1999) and are identified earlier in a gating task (Craig et al 1993). To model these predictability effects, it is necessary to augment the Bayesian model in (2) to recognize that the probability of a word depends on context. For example, the word *phonetics* has a low frequency in general, but it could be much more probable in the context of a phonetics conference. The probabilities of words depend on a variety of contextual factors such as sentence topic and syntactic context implied by the preceding words, so an ideal listener should take these context effects into account in estimating the prior probabilities of words. Accordingly, the word probabilities are conditioned on the context C in the revised Bayesian formulation in (3).

$$(3) \qquad p(w \mid E,C) = \frac{p(E \mid w)p(w \mid C)}{\sum_{w_i \in lexicon} p(E \mid w_i)p(w_i \mid C)}$$

This model implies that effects of predictability on word recognition generally involve the prior probabilities of both the target word and its competitors: where a word is more predictable $p(w|C)$ is higher and the probabilities of most competitors, $p(w_i|C)$, are likely to be lower, both of which raise the posterior probability of the target word for a given quantity of evidence from the signal, $p(E|w)$.

2.4 Interactions between predictability, frequency and neighborhood density

What is particularly interesting about the analysis in (3) is that it predicts interactions between the effects of predictability, frequency and neighborhood density on word recognition. The expression in (3) refers to the prior probabilities of words given the context,

not to their frequencies. So word frequency should only be relevant to the extent that it is used to estimate prior probabilities of words, $p(w|C)$. In the absence of contextual constraints, word frequencies provide the best estimates of prior probability, but as contextual constraints on word probability become stronger, we should see a lower correlation between word frequency and word recognition performance. This applies both to the frequency of the target word and to frequencies of lexical neighbors because a context that makes one word more likely must also make other words less likely, and in general this will include most of the neighbors of that word. So the effect of competition from neighbors should be reduced where *w* is contextually predictable (i.e. the denominator in (3) is smaller under these conditions). In other words, the model predicts that the effects of frequency and neighborhood density should decrease in contexts where the target word is more predictable. So Bayesian analysis provides a basis for Scarborough's prediction that neighborhood density effects on pronunciation should be reduced in high predictability contexts.

These predictions are of considerable interest, because they imply that measures of the effects of word frequency on speech processing that are obtained from tasks that use isolated words or uninformative sentence contexts are liable to overestimate the magnitude of these effects in running speech where rich contextual information is almost always available. In addition, when this model of word recognition is embedded in a model of listener-oriented speech production, as outlined above, it implies that a speaker's estimate of the listener's difficulty with word recognition must be calculated online rather than being based purely on lexical statistics like frequency and neighborhood density.

There is evidence confirming these predicted interactions with respect to word recognition. Evidence that the effect of word-frequency on recognition decreases as words become more contextually predictable comes from a gating study by Grosjean & Itzler (1984). They found that the gate where a word could be reliably identified was earlier in more frequent words, but that this frequency effect was reduced where words were more predictable from preceding sentence context, almost to zero in the most constraining contexts. A similar pattern is observed in electrophysiological measures of reading. In an event-related potentials study of silent reading of meaningful sentences, van Petten & Kutas (1990) found that less frequent words were associated with larger N400 components when they occurred early in sentences, but this frequency effect disappeared later in the sentence. They suggest that 'frequency does not play a mandatory role in word recognition but can be superseded by the contextual constraint provided by a sentence' (p.380).

The effect of neighborhood density on word recognition is also reduced where words are more predictable from context. In an auditory word identification task, Sommers & Danielson (1999) found that the difference in accuracy of identification between words from sparse and dense neighborhoods decreased when words were highly predictable from preceding sentential context. Along similar lines, Sommers, Kirk & Pisoni (1997) found a reduction in effects of neighborhood density on word identification accuracy when subjects had to pick from a closed set of words. This can be interpreted as showing that a closed response set results in reduced competition from neighbors that are not in the response set because the experimental context gives these words low prior probabilities.

3. Testing the correspondence between production and recognition

H&H theory hypothesizes that speakers modulate the clarity of their speech according to the difficulty the listener is expected to have with word recognition. To pursue this strategy, speakers must have a model of listener difficulty. We have seen that a Bayesian model of word recognition provides an accurate characterization of the qualitative effects of frequency, neighborhood density and contextual predictability on word recognition performance. So if speakers' internal models of listeners are accurate then variation in clarity should follow a corresponding pattern. Scarborough's experiments provide a novel test of this idea, investigating the combined effects of contextual predictability and frequency-weighted neighborhood density on vowel production. In recognition the effects of frequency and neighborhood density on performance are reduced in high predictability contexts, but Scarborough's study did not reveal a corresponding interaction in production: the effects of neighborhood density and predictability on vowel dispersion were independent. So clarity of vowel production does not seem to be tracking difficulty of word recognition in this case.

Taken at face value, Scarborough's results show that either the listener-oriented model of production outlined above is wrong, or at least that speakers are not accurately tracking listener difficulty in this regard. However, this conclusion depends on a null result, so it is worth considering whether the experiment was suitably designed to detect the hypothesized interaction.

A basic problem with studying the effects of neighborhood density on phonetic realization is that it is not possible to manipulate neighborhood density independently of segmental context – in order for words to be in different neighborhoods they must differ segmentally, so the effects of this confound can only be minimized, not eliminated (cf. Wright 2004: 79, Munson & Solomon 2004: 1050). Scarborough studies vowel realization, and surrounding consonants can have significant effects on the realization of vowels so this confound is potentially problematic.

Scarborough's materials raise particular concerns in this respect because the consonantal contexts of the vowels are clearly not balanced: all of the low neighborhood density words have onset clusters whereas none of the high density words do. This confound could be the source of the apparent effect of neighborhood density on vowel duration, since most of the clusters are obstruent-liquid clusters which have a shortening effect on following vowels, compared to a singleton obstruent or liquid (Van Santen 1992:531-2). Scarborough's figure 2 provides some evidence that this effect was present in her data (e.g. compare *frame* vs. *fame*, *plump* vs. *pump*, *trunk* vs. *dunk*). Previous studies of the effects of neighborhood density on vowel realization only studied CVC words and found no effect of neighborhood density on vowel duration (Wright 2004, Munson & Solomon 2004, Munson in press). In addition the liquids [l] and [ɹ] that appear in many of the onset clusters can have significant coarticulatory effects on vowel quality.

H&H theory also predicts that segmental context could affect the magnitude of the effect of predictability on vowel quality. H&H theory does not posit a direct link between predictability and vowel centralization, rather it posits a conflict between clarity and duration/effort: speakers are hypothesized to expend less time and/or effort on words that are easier to recognize, given the context. Reduced effort and duration results in assimilation of vowels to their surrounding consonants, which tends to lead to centralization but need not (Lindblom 1963, Moon & Lindblom 1994). For example [i] in a word like *yeast* will not be

centralized if it assimilates to the preceding palatal glide. Consequently a reduction in duration or effort is expected to result in a greater change in vowel quality where there is greater conflict between the articulatory requirements of the vowel and its adjacent consonants. In other words, there is reason to expect item-specific variation in the effect of the predictability manipulation on vowel centralization, depending on the specific segmental make-up of each item. So the fact that segmental contexts are not closely matched across the high and low neighborhood density words means that the segmental differences could conceivably obscure differences in the magnitude of predictability effects on the measure of vowel centralization. In other words it is possible that an interaction between neighborhood density and predictability could have been obscured by confounding differences in segmental contexts. Looking at the results for individual items (Scarborough's fig. 4), it is apparent that there is variation in the predictability effect between items from the same neighborhood density group, but on the other hand it is not immediately obvious that this variation follows from segmental context.

It should also be noted that the same considerations imply that Scarborough's use of ANCOVA analysis to factor out the effect of vowel duration on vowel centralization is probably insufficient. The ANCOVA analysis with vowel duration as a covariate employs a model in which vowel duration has the same linear effect on vowel centralization for all items, and so cannot capture context-dependent or non-linear duration effects.

In summary, further tests of the interaction between predictability and neighborhood density with better control of segmental context are warranted. However, the current results do not show any indication of the predicted interaction (figure 3), so it is also worth considering whether there are any models of the effects of neighborhood density and predictability on phonetic realization that predict independent effects, and whether there are other ways to test competing models that would avoid the thorny problem of trying to control segmental context while varying neighborhood density.

4. Alternative models

Some mismatches between listener difficulty and speaker clarity could be accommodated in a model according to which speakers do consider the needs of the listener, but only make an approximate assessment of those needs. As Scarborough observes, an approximate model of the listener could be motivated by the need to simplify the processing involved in tracking listener difficulty (cf. Bard et al 2000). Calculating prior probabilities in the Bayesian model of the listener involves estimating the contextual probabilities of all the words in the lexicon, which might well be too demanding for speakers. However it is unclear why the best approximation to this ideal should be an additive combination of predictability and neighborhood density. A very simple approximation to the Bayesian model would be to take context into account only in evaluating the probability of the word to be spoken. If the probability of the target word is increased in the context compared to its unconditioned probability, then the probabilities of all competitors are reduced proportionately to ensure that all word probabilities sum to 1, but without further regard for context. Even this simplistic model of context effects predicts reduced competition effects in contexts where the target word is more predictable because the effects of competition would be reduced by the lower prior probabilities of neighbors, while the prior probability of the target word is

increased. So while it is very plausible that speakers should employ a computationally simple estimate of listener difficulty, this does not in itself lead us to expect independent effects of neighborhood density and predictability.

A lack of correspondence between listener difficulty and speaker clarity would be unsurprising if speakers are in fact not generally listener-oriented – i.e. the H&H account is wrong. For this alternative to be viable, it is necessary to provide alternative accounts of the effects of predictability and neighborhood density on phonetic realization. Pierrehumbert (2002) proposes two explanations for the effects of neighborhood density, neither of which appeals to listener-orientation. The first is based on the hypothesis that if a word is more difficult for a speaker to access in production, then the speaker pronounces that word more clearly (although the reasons for this link are not clear – p. 107f.). Consequently any factor that impedes lexical access is predicted to result in more reduced pronunciation. Pierrehumbert speculates that high neighborhood density results in clear pronunciation because it impedes lexical access in production as well as recognition. Scarborough (this volume) and Munson (in press) discuss the problems facing this account – not least of which is that subsequent investigation has shown that high neighborhood density can actually facilitate lexical access in production, as indicated by faster responses in a picture naming task (Vitevitch 2002, Vitevitch & Sommers 2003).

The second account of neighborhood density effects proposed in Pierrehumbert (2002) appeals to the nature of word learning in an exemplar-based model of speech production. According to this model, words are represented in the lexicon by multiple phonetically detailed exemplars rather than a single, more abstracted representation. The exemplars of a word are remembered utterances of that word as experienced by the speaker. But for an utterance to be stored as an exemplar of a word, it must be recognized as an instance of that word. Pierrehumbert suggests that since words from dense neighborhoods are more difficult to recognize, less clear renditions of these words are more likely to be misperceived than words from sparse neighborhoods. Accordingly the exemplars of words from dense neighborhoods will tend to be skewed towards more hyperarticulated tokens compared to words from sparse neighborhoods because reduced realizations are less likely to be encoded as exemplars. This difference in lexical representation is then reflected in speech production: Pierrehumbert proposes that production proceeds by selection of a random exemplar, which is then averaged together with nearby exemplars to yield a production target. So the distribution of vowel qualities in production reflects the distribution in the stored exemplars, with the result that vowels will, on average, be more hyperarticulated in hard words than in easy words.

This line of analysis obviously does not extend to the effects of predictability on pronunciation since it relies on lexical representations whereas predictability effects must be derived online since they depend on context. Accordingly predictability effects must be accounted for by some other mechanism. But if the two kinds of effect arise from independent mechanisms, that could well lead to the prediction that the effects should be independent. That is, if differences in vowel quality due to neighborhood density are represented in the lexicon and higher predictability results in a uniform reduction of the representation drawn from the lexicon, the two factors would yield independent effects on vowel centralization. The question is what mechanism could account for predictability effects in such a model? Pierrehumbert (2002) hypothesizes that the effects of predictability on pronunciation arise from the proposed relationship between difficulty in lexical access and

hyperarticulation, discussed above: predictable words are easier to access in production (e.g. Griffin & Bock 1998) so they are given reduced pronunciations. However lexical access is also affected by neighborhood density, and the relationship between predictability and neighborhood density has not been investigated in speech production, so it remains possible that the two interact.

So none of the alternative models reviewed here really predicts independent effects of neighborhood density and contextual predictability on clarity of speech, but Pierrehumbert's exemplar-based model of speech production at least predicts one source of neighborhood density effects that should be unaffected by context. The exemplar-based model is sharply at odds with the H&H account in placing the effects of neighborhood density offline, encoded in the lexical representations of words, where the H&H model hypothesizes that speakers assess the extent of competition from neighbors online, based on the context. This contrast suggests another way to test for the interaction between neighborhood density and predictability predicted by the listener-oriented model, by manipulating the predictability of neighbors as well as the target word. If speakers are tracking anticipated word recognition difficulty online then they should reduce words less if a close neighbor is probable in the same context, whereas if neighborhood density effects are offline, then the contextual predictability of neighbors should have no effect on speech production. This prediction can be tested with complete control over segmental contexts since predictability and competition are both manipulated via context, so the same words can be used in all conditions.

5. Tracking listener difficulty within words

Scarborough's experiment can be viewed as asking how closely speaker clarity tracks listener difficulty with lexical access, testing for a counterpart in production of the interaction between neighborhood density and predictability that is observed in word recognition performance. PEBB can be viewed as pushing a similar research program in a different direction, looking inside the word. So far we have discussed the difficulty of recognizing words, and the clarity of word pronunciation, but the analysis of word recognition implies that we could be more precise in localizing the points of difficulty within a word, and that speakers could accordingly allocate more effort to the production of these particular segments or features. For example, the neighbors of a word are similar to it in particular respects: e.g. *mad* and *bad* are similar in onset and vowel, but differ in their codas. If a word has no neighbors that differ in vowel only then hyperarticulation of the vowel might be expected to be less useful in overcoming the effects of competition.

PEBB investigate a hypothesis along these lines, investigating the effects of neighborhood density on the realization of specific segments in the suffix (or suffix combination) *–igheid*. It is not simple to derive predictions about the expected realizations of particular segments given the hypothesis that speakers allocate production effort segment by segment to maximize the probability of successful word recognition. We do not have direct evidence that the structure of lexical neighborhoods affects which parts of words are important for word recognition – in this case production studies have preceded perceptual studies – so the predictions concerning segment-specific hyperarticulation have to be derived from models of the relative importance of segments within a word for word recognition.

One line of reasoning cited by PEBB holds that the importance of a segment is inversely related to its predictability given preceding segments. This approach has been developed most explicitly by van Son & Pols (2003a), who calculate the probability of a segment by considering the frequencies of all words that continue the preceding segment string – i.e. the members of the current cohort[2]. In the context of a cohort-based model of word recognition where candidate words are eliminated as they become inconsistent with the segment string identified from the signal (Marslen-Wilson & Welsh 1978), a lower probability segment does more to reduce the size of the current cohort and is in that sense can be regarded as more informative.

A model of this kind does not account for PEBB's data on the realization of *–igheid* in Dutch. They distinguish four kinds of *–igheid* words based on the range of morphologically related words, as summarized in (4). According to the model sketched above, [h] should be most predictable in underived and +igheid words since these words lack +ig forms. So, for example, once the initial string [vastix] of the word *vastigheid* has been heard, it should be predictable that the next segment is [h] since there is not *vastig(e)*. So [h] should be most reduced in these word classes. In fact the [xh] cluster is longer in +igheid than in +heid words, and underived words are intermediate, not significantly different from the other two classes.

(4)

type	stem	-ig form	-igheid form
+ig+heid	baas	bazig	bazigheid
+heid	*no*	zuinig	zuinigheid
+igheid	vast	*no*	vastigheid
underived	*no*	*no*	saamhorigheid

There have been a number of other attempts to test for segment-specific reduction as a function of segment predictability, mostly with negative or inconclusive results. Van Son & Pols (2003a, b) tested the segmental predictability hypothesis using a substantial corpus of speech. Aylett & Turk (2004, 2006) also used large corpora to test the related hypothesis that more predictable syllables should be more reduced (shorter and with more centralized vowels). Employing a corpus of spontaneous speech obviously raises the problem of controlling for confounding variables that could affect segment reduction. In particular, Aylett & Turk (2004) found in a study based on a large corpus of spontaneous dialogues that predictability of syllables was confounded with prosodic properties (phrasing and accentuation). Although syllable predictability was correlated with syllable duration, predictability only accounted for a small proportion of the variance once prosodic factors were taken into account as well. This lead Aylett & Turk to suggest that predictability influences the assignment of prosody but has little direct effect on the realization of segments. Given this finding, it is very important to control for prosodic factors when looking for direct effects of predictability on realization. Although van Son & Pols (2003a, b) find significant correlations between their cohort-based measure of segment predictability and some measures of segmental reduction, including vowel duration, it is not clear that

[2] Van Son & Pols actually quantify informativeness of a segment in information theoretic terms as the surprisal of the segment, i.e. the negative log of the probability of the segment.

prosody was adequately controlled since prominence and phrasing were predicted from the text rather than being measured or transcribed.

Aylett & Turk (2006) found that more predictable syllables are shorter and have more centralized vowels, even after controlling for stress, accent and position in phrase, but their measures of predictability only made reference to syllables, not words. Specifically, measures of predictability were based on the preceding three syllables without regard to word boundaries, so the measures potentially combined conditional probabilities of words with phonotactic predictability. Accordingly it is not clear that the observed effects are syllable predictability effects as opposed to word predictability effects.

Billerey-Mosier (2000) looked for segmental reduction after the uniqueness points of words. The uniqueness point of a word is 'the segment which identifies the word uniquely against all the other words sharing the same initial string'. This hypothesis is closely related to the proposal of van Son and Pols, since segments after the uniqueness point should be highly predictable, assuming that the preceding segments are accurately identified. However Billerey found no reduction in segment duration nor any vowel centralization after the uniqueness point.

The absence of clear effects of segment predictability given preceding segments is perhaps unsurprising given that this approach to measuring the importance of segments makes some rather strong simplifications. In particular, it implicitly assumes that segments are correctly identified, so the probability of a segment can be conditioned on the actual string of preceding segments, and so there is no need to identify earlier segments based on later information. In fact identification of segments is typically uncertain, so later segments can help to disambiguate the intended word. For example, the [t] of *shoot* would be counted as relatively uninformative according to van Son & Pols's measure since no other consonants can follow [ʃʊ] (at least in common words), but if the vowel were somewhat ambiguous between [u] and [ʊ], the final [t] would provide evidence for the correct word since there is no word [ʃʊt] in most dialects of English. In other words, there is no reason to privilege statistical dependencies of later segments on earlier segments to the exclusion of dependencies running in the reverse direction, but that is exactly what van Son & Pols's metric of informativeness does.

In terms of the Bayesian analysis of word recognition, producing a particular segment can facilitate recognition by increasing the likelihood of the signal given the target word, $p(E|w)$, and decreasing the likelihood of the signal given competitors, $p(E|w_i)$, but if the segment in question appears in a similar position in a competitor, as with the vowel in *mad* and *bad*, then clearer cues to the vowel do not decrease the likelihood of the competitor. Hyperarticulation should be most effective where it yields better cues to differences between the target word and its closest, highest frequency neighbors. Accordingly, it should be more efficient to allocate effort where it has the greatest effect on differentiating words rather than hyperarticulating difficult words uniformly. However there is no evidence so far that speakers adopt this strategy either.

Scarborough (2003) directly tested for segment-specific hyperarticulation based on the structure of lexical neighborhoods. She studied CVN words with comparable neighborhood densities but where the neighbors were of different kinds: some of the words had many neighbors that differed by having a non-nasal coda and others had few neighbors that differed in this respect. Her earlier studies had shown that in CVN words, nasal coarticulation on the vowel is greater in words with higher neighborhood density, and that this increased

coarticulation facilitates word recognition (Brown 2001, Scarborough 2004). Presumably nasal coarticulation provides cues for the presence of a following nasal, and so should be particularly useful where the CVN word has many neighbors that differ by lacking a nasal coda, but the follow-up experiment showed that the increased nasal coarticulation obtained regardless of the segmental structure of the neighbors.

This is a single negative results, so the possibility of localized hyperarticulation within words remains open, but it seems that any such effects are much less robust than word-level effects, which have been demonstrated repeatedly. Furthermore, the results of Scarborough (2003) indicate that if there is segment-to-segment variation in clarity, it is in addition to the word-level effects since neighborhood density seems to affect segmental realization regardless of the precise segmental make-up of the neighborhood. This could indicate that speakers do not track the contribution of individual segments to word recognition, but it also might indicate that hyperarticulation generally involves a global shift in articulation, and cannot easily be modulated on a segment-by-segment basis.

PEBB ultimately argue that the pattern of variation in the duration of [xh] clusters in *–igheit* reflects the number of forms in the paradigm of each type of word – the more forms there are, the less reduced the [h] is (although it is not clear that underived words fit this generalization since they have the smallest paradigms, but don't have the shortest clusters). They argue that this is essentially a neighborhood density effect on the grounds that morphological derivatives of a stem are phonologically similar, although they are not neighbors in the sense of differing in a single segment. But this reasoning does not predict that the [h] in particular should be reduced in smaller paradigms – we might expect that the whole syllable *–heid,* or even the whole word, should be reduced. PEBB did not examine the duration of other segments so we do not in fact know whether they have observed an instance of segment-specific reduction.

6. Summary

Lindblom's H&H theory implies that speakers maintain an internal model of listeners in order to predict difficulties with word recognition. We have seen that a model based on the ideas that listeners adopt a probabilistic approach to word recognition and that they employ Bayes' rule in combining top-down and bottom-up evidence for words provides a qualitative account of a variety of effects on word recognition performance, including the effects of word frequency, neighborhood density, contextual predictability, and interactions between them. This model provides a basis for investigating to what extent speakers model (and respond to) expected listener difficulty. The papers by PEBB and Scarborough contribute to this research program, going beyond the basic effects of word predictability, frequency and neighborhood density to investigate interactions between predictability and neighborhood density (Scarborough) and the possibility that hyperarticulation might be targeted on the most important segments within a word (PEBB). The results so far are inconclusive, but they suggest further avenues for research.

7. References:

Brown, Rebecca A. (2001). *Effects of Lexical Confusability on the Production of Coarticulation*. M.A. thesis, UCLA.

Aylett, M. & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence and Duration in Spontaneous Speech. *Language and Speech* 47, 31-56.

Aylett, M. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America* 119, 3048-3058.

Bard, E.G., Anderson, A.H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. Journal of Memory and Language, 42, 1–22.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113, 1001-1023.

Billerey-Mosier, R. (2000). Lexical effects on the phonetic realization of English segments. *UCLA Working Papers in Phonetics* 100.

Boothroyd and Nittrouer (1988). Mathematical treatment of context effects in phoneme and word recognition. *JASA* 84(1):101.

Craig, Chie H., Kim, Byoung W., Pecyna Rhyner, Paula M., & Bowen Chirillo, Tricia K. (1993). Effects of word predictability, child development, and aging on time-gated speech recognition performance. *Journal of Speech and Hearing Research* 36, 832-841.

Goldinger, S.D., Pisoni, D.B., & Luce, P.A. (1996). Speech perception and spoken word recognition: Research and theory. In N.J. Lass (Ed.), *Principles of Experimental Phonetics*. St. Louis: Mosby. 277-327.

Griffin, Z.M., and Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language* 38, 313-338.

Grosjean, F., & Itzler, J. (1984). Can semantic constraint reduce the role of word frequency during spoken word recognition? *Bulletin of the Psychonomic Society* 22, 180-182.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20, 137-194

Lindblom, Björn (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35, 1773-1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W.J. Hardcastle and A. Marchal (eds.) *Speech Production and Speech Modeling*. Kluwer: Dordrecht

Luce, P., and Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19. 1-36.

Luce, P.A., Pisoni, D.B. & Goldinger S.D. (1990). Similarity neighborhoods of spoken words. G. Altman (ed.) *Cognitive Models of Speech Processing*. MIT Press, Cambridge, 122-147.

Luce, P.A. (1986). *Neighborhoods of Words in the Mental Lexicon*. Ph.D. dissertation, Indiana University, Bloomington.

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

Moon, S-J, & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America* 96, 40-55.

Munson, B. (in press) Lexical access, lexical representation, and vowel production. J. Cole and J. I. Hualde (eds.), *Papers in Laboratory Phonology IX*. Mouton de Gruyter, New York.

Munson, B., and Solomon, N.P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research* 47, 1048-1058.

Norris, D. (2006) The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review* 113, 327-357.

Pierrehumbert, J.B. (2002). Word-specific phonetics. In C. Gussenhoven and N. Warner (eds.) *Papers in Laboratory Phonology VII*, Mouton de Gruyter, New York, 101-139.

Rabiner, Lawrence & Juang, Bing-Hwang (1993). *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs.

Scarborough, Rebecca A. (2004) Degree of Coarticulation and Lexical Confusability. Nowak, P.M., Yoquelet, C., and Mortensen, D., eds., *Proceedings of the 29th Meeting of the Berkeley Linguistics Society*.

Scarborough, Rebecca A. (2003). The word-level specificity of lexical confusability effects. Poster presented at the 146th Meeting of the Acoustical Society of America, Austin, TX.

Sommers, M., Kirk, K., Pisoni, D. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I. The effects of response format. *Ear and Hearing* 18, 89–99.

Sommers, M.S., and Danielson, S.M. (1999). Inhibitory processes and spoken word recognition in young and older adults: the interaction of lexical competition and semantic context. *Psychology and Aging* 14, 458-472.

Van Petten, C. and Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory and Cognition* 18, 380-393.

Van Santen, J.P.H. (1992). Contextual effects on vowel duration. *Speech Communication* 11, 513-546.

Van Son, R.J.J.H. and Pols, L.C.W. (2003a). Information structure and efficiency in speech production. *Proceedings of Eurospeech 2003*, Geneva, Switzerland, 769-772.

Van Son, R.J.J.H. and Pols, L.C.W. (2003b). An acoustic model of communicative efficiency in consonants and vowels taking into account contextual distinctiveness. *Proceedings of ICPhS*, Barcelona, Spain, 2141-2144.

Vitevitch, M. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28, 735-747.

Vitevitch, M.S. & Sommers, M. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production. *Memory & Cognition, 31,* 491-504.

Wright, R. (2004) Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, and R. Temple (eds.), *Papers in Laboratory Phonology VI*. CUP, Cambridge, 75-87.