Word-specific phonetics

Janet B. Pierrehumbert

Abstract

In standard models of phonetic implementation, surface phonological representations arise as words are retrieved from the lexicon and assembled in a buffer, where the phrasal intonation and prosody are added. These (categorical and hierarchical) representations provide the input to the phonetic implementation rules, which map them into motor gestures and acoustic outcomes. The model has been highly successful in handling across-the-board effects on phonetic outcomes, including language-specific phonetic patterns of allophony and shifts in overall voice level or force of articulation. The very causes of this success render it unable to handle instances of word-specific phonetic detail, which have now come to light through large-scale experimental and sociolinguistic studies. This paper summarizes the evidence that long-term representations of words include more phonetic detail than previously imagined. It sketches a hybrid model of speech production, in which exemplar theory is used to model implicit knowledge of the probability distributions for phonological elements as well as of word-specific phonetic patterns. Production goals for specific phonological elements are biased by stronger activation of exemplars associated with the current word. Thus, experience with specific words influences the exact production goals for those words, even as the phonological decomposition plays the dominant role. The consequences of this model for the production of morphologically complex words are also explored. The model also provides a mechanism for the subphonemic paradigm uniformity effects which other authors have recently documented.

1. Introduction

A long-standing forte of the Laboratory Phonology series has been work on phonetic implementation of phonological representations. Numerous studies in this series have elucidated the patterns of variation in the realization of phonological categories in different segmental and prosodic contexts, and such studies now provide one of the main lines of evidence about the cognitive representation of sound structure.

In a consensus view of phonetic implementation, lexemes (the phonological representations of words) are abstract structures

made up of categorical, contrastive elements. The phonetic implementation system relates this abstract, long-term, categorical knowledge to the time course of phonetic parameters in particular acts of speech. In fluent mature speakers, the phonetic implementation system is a modular, feed-forward system, reflecting its nature as an extremely practiced and automatic behavior. Lexemes are retrieved from the lexicon, and assembled in a phonological buffer in which phrasal prosody and intonation are also assigned. The fully formed hierarchical structures thus assembled provide the input to the phonetic implementation rules, which compute the degree and timing of articulatory gestures. The model is feedforward because no arrows go backwards, from articulatory plans to phonological encoding, or from the phonological encoding to the lexical level (apart from some post-hoc monitoring which permits people to notice and correct speech errors). It is modular because no lexeme information can influence the phonetic implementation directly, bypassing the level of phonological buffering.

Though highly successful in explaining a wide range of data, such models are now challenged by a number of studies demonstrating the existence of word-specific phonetic detail. In modular feed-forward models, the (categorical) form of the lexeme wholly determines the phonetic outcome. If two words differ at all in their phonetics, then they differ categorically, and accordingly one job of the phonology is to identify a category set which captures all systematic differences amongst words. Another feature of these models is they do not take on the job of describing systematic phonetic variation related to sociostylistic register. Though the authors of such models would no doubt acknowledge the existence of such variation, they have not undertaken to provide a formal treatment of the cognitive capabilities which permit it. These limitations in formal models of speech production are related, because some cases of word-specific phonetic detail can be traced to the typical patterns of word usage in different social contexts. Developing the next generation of speech production models which can handle such variation is an important goal, because control of subphonemic variation is an important aspect of the human language capability. Its interaction with the more categorial

aspects of linguistic competence appears to be highly structured, and it promises to be a rich source of information about the architecture of language. In the theoretical stance taken here, categorical aspects of phonological competence are embedded in less categorical aspects, rather than modularized in a conventional fashion. The reader is referred to Pierrehumbert (2000) and Pierrehumbert, Beckman & Ladd (2001) for a more detailed defence of this stance.

Production models based on exemplar theory can readily capture findings of word-specific allophonic detail. In such models, each word can be associated with an empirically determined frequency distribution over phonetic outcomes. The distributions are continuously updated based on experience, and nonphonemic differences in these experiences accrue in the representations. For example, if some word is most often produced in leniting contexts, its long-term representation will show more lenition. If it is most often produced in a particular sociostylistic register, its long term representation will show the hallmarks of that register. The chief drawback of this approach is that it handles none of the data which motivated the standard modular feedforward models.

In this paper, I sketch out a hybrid model which generates word-specific allophony while still retaining the insights of the modular feedforward models. I will first review the major lines of evidence for the modular and exemplar-based models. Then I will show how each line of evidence is handled in the hybrid model.

2. Modular Feedforward Models

Modular feedforward models of phonetic implementation were developed on the basis of experimental work in psycholinguistics and phonetics. Psycholinguistic studies have concentrated on the speed and accuracy of word production in various tasks. Experiments on induction of speech errors provide one of the earliest lines of evidence that lexemes are first copied into a phonological buffer before being pronounced. In Shattuck-Hufnagel's highly influential account (Shattuck-Hufnagel, 1979) errors in the copying and checkoff procedures explain the statistical patterns of anticipation,

perseveration, and transposition errors. Competing models which lack buffering have difficulties in explaining transposition errors. A set of experiments by Sternberg et al. (1978) and Sternberg et al. (1980) on the latency to begin speaking provides evidence that assembly of longer phonological plans takes longer than assembly of shorter plans – further evidence that such an assembly process is critically involved. A long series of experiments by Levelt and colleagues (see Levelt, 1989; and also the WEAVER model presented in Roelofs, 1997) used a variety of speech production tasks. A critical finding from this work is that some predictable features of word prosody are computed on-line rather than stored in long-term representations.

This general class of experimental results brings home the fact that in both speech perception and speech production, a correspondence is established between events which unfold in time (the speech signal) and the metatemporal long-term representations of words. The long-term representations are metatemporal in the sense that they are about sequences of phonological events that unfold in time. They describe such events, but they themselves are not events that occur in time. This means that there is a discrepancy in logical type between lexical representations (which are long-term – or nearly permanent – memories), and speech events (which occur in time and and which are as evanescent and irreversible as other other physical events). In the terms of psychology, lexical representations are examples of declarative memories: "I know **that** the word *evanescent* has such-and-such articulatory and acoustic events in such-and-such order." The phonetic implementation rules provide an example of procedural knowledge: "**how** to say /v/." The mere fact of this distinction provides a basic argument for a modular theory.

Detailed studies of phonetic implementation rules have concentrated on phonetic variability related to the phrasal context of a word. The original goal of these studies was to explore the psychological reality of hierarchical structures and/or to delineate the architecture for fluent and natural-sounding speech synthesis systems. Quantitative studies of f0 showed that the phonetic realization of any given tonal element can only be computed if a complete

phrasal phonological structure is available (see Pierrehumbert & Beckman, 1988, and literature reviewed there). More recent work has extended these findings to the domain of segmental allophony. In addition to the well-known case of phrase-final lengthening, phrasal prosody is now known to affect aspiration and glottalization (Pierrehumbert & Talkin, 1992; Pierrehumbert, 1994; Pierrehumbert & Frisch, 1996; Dilley, Shattuck-Hufnagel & Ostendorf, 1996), as well as the force, accuracy, and duration of the other aspects of consonantal articulation (de Jong, Beckman & Edwards, 1993; de Jong, 1995; Keating et al., forthcoming).

Such studies provide a prima-facie case for a level of phonological encoding which is distinct from the lexeme level. Phrasal prosody depends on the syntactic structure and pragmatic force of a sentence, which are productively determined when words are combined. In intonation languages such as English, the tonal elements also are assigned at the phrasal level. The phonological buffer proposed by Shattuck-Hufnagel and Sternberg et al. provides a locus for the calculation of phrasal prosody and intonation. Detailed phonetic effects of phrasal phonology are then taken to arise from the way that the fully parsed contents of this buffer are executed by the motor system. The execution is known to be language-particular (since allophonic details for even the most comparable phonemes differ from one language to another); however, it is of course constrained by human capabilities for articulation and perception.

Such studies of phonetic implementation bring home the success of modular feed-forward models in explaining across-the-board effects of all types. Across-the-board effects are ones which pertain to all words which share the triggering phonological context. There are several different kinds of examples of such effects. One is allophonic rules which are peculiar to a language or dialect. For example, in American English, intervocalic word-internal /t/s in a falling stress context (as in the word *pretty*) is typically produced as a voiced flap. In many dialects of British English, the voicelessness of the /t/ is preserved in the same context even if the closure is reduced. A related fact is the outcome of Neogrammarian sound changes (which enter the language as allo-

phonic processes and may eventually become fossilized across the entire vocabulary). For example, the Germanic affrication of Indo-European stops affected all words containing the target stops. Changes in stylistic register are also across-the board, in that they affect all words throughout a phrase. In a clear speech style, the speaker produces all words more slowly and with more articulatory effort. Raising the voice causes all words to be louder and have a higher f_0.

However, as we will see below, the assumption that such effects are across-the-board is not fully correct. The effects are across-the-board in that broad classes of words are eligible to undergo them, and they can even apply to novel words. A person who flaps the /t/ in *pretty* and *Betty* will find that an invented *Bretty* is also eligible for flapping. But detailed studies have shown that the probability and extent of reduction processes is word-dependent. Dependence on both word frequency and on morphological relatives has been documented. In the next session, I review reports of word-specific subphonemic detail.

3. Long term word-specific phonetic patterns

Reports of allophonic effects related to word frequency and/or contextual predictability go back to Zipf or earlier. They are usefully reviewed in the contribution by Jurafsky, Bell & Girard to this volume, and I will not repeat this review here. Both in experiments and in corpora of natural conversation, words which are highly expectable are produced faster and less clearly than words which are rare or surprising. What does this mean for the architecture of the cognitive model? An important issue is whether such effects are generated on-line (and if so, by what means) or whether they result from long-term storage of different phonetic patterns for different words. The primary concern of this paper is the relationship of long-term representations to production patterns, and so my primary focus will be patterns which do not plausibly result from on-line control of speech style and therefore implicate long-term memory.

Consider first the effects of contextual predictability, the main topic of Jurafsky et al. (this volume). For any given word, contextual predictability results in differential lenition rates depending on how much novel information the word contributes (above what is in any case inferrable from the thread of the conversation and the neighbouring words). Such effects can be viewed as an on-line modification of speech style. The standard modular model does not generate such effects, but it can be readily modified to do so. When a lexeme is retrieved and loaded into the phonological buffer, assume that a gradient value representing the ease of retrieval is passed to the buffer as a quantitative attribute of the Prosodic Word node. This parameter would control, or rather play a part in controlling, an overall parameter of articulatory clarity and effort. This would be a direct formal analogue of the gradient pitch range parameters which full-blown f_0 synthesis algorithms, such as Pierrehumbert & Beckman (1988), employ to generate tonal outcomes under varying conditions of intonational subordination.

A similar line of reasoning may be available for the finding by Wright (1997) that CVC words produced in citation form show a dependence of formant values on lexical neighbourhood density. Words with a high neighbourhood density have many lexical neighbours which differ in just one phoneme (according to the metric used). Words with a low neighbourhood density have few such neighbours. Wright (1997) found that the vowel space was expanded in words with high neighbourhood density. Since high neighbourhood density slows word recognition in perception experiments (an effect attributed to lexical competition), it is at least possible that neighbourhood density would also slow lexeme retrieval in production. If so, Jurafsky et al. (this volume) would predict an on-line effect on speech clarity.

The pervasiveness and automaticity of such reduction effects suggests, of course, that this description is a piece of formalism in search of an explanation. Jurafsky and colleagues are actively seeking to identify the underlying cognitive or neural factor which creates such an intimate connection between the ease of lexical retrieval and speech style.

Whatever this factor may prove to be, it predicts the existence of reduction effects related to word frequency. First, word frequency is uncontroversially related to resting activation levels, so that even out of context, frequent words are retrieved faster. Second, word frequency is correlated with contextual predictability. Since (by definition) high frequency words occur more often in running speech than low frequency words, any given high frequency word is more likely or unsurprising in the average context than a low frequency word. Furthermore, a high frequency word is more likely to occur a short time after a previous mention of the same word. Thus, on-line tracking of predictability would have the result that the aggregate statistics for high-frequency words would display higher average predictability, and hence more lenition on the average. A surface pattern of reduction related to word frequency is not enough in itself to argue for long-term storage of word-specific allophonic detail. As far as long-term storage goes, then, the most telling phenomena are ones which do not involve the connection of frequency or neighbourhood density to lenition.

One example of such a phenomenon is provided by an experiment reported in Goldinger (2000). Goldinger carried out a speech production experiment in order to elucidate the nature and role of long-term memory of specific voices. Goldinger's prior work on speech perception had indicated that long-term memory of words includes traces of the specific voices in which the words were spoken (Goldinger, 1996). In the production experiment, subjects first made a baseline recording of a set of test words by reading the words from a screen. The next day, they carried out a task in which they heard words in various voices and located the word in a visual display on a computer screen. Five days later, they returned to the lab and read the words for a second time, providing a set of test utterances for a second experiment. AXB stimuli for the second experiment were constructed from the baseline recordings, the test recordings, and the stimuli of the first experiment. The word in the X position was one which had been played to the first group of subjects for the visual search and identification task. The words in the A and B positions are baseline and test utterances by the first group of subjects (balanced for position in the pre-

sentation). Then, a new group of 300 subjects listened to the AXB stimuli and made judgments of "which utterance was a better imitation of the middle word." Overall performance was well above chance on this task, indicating that the test utterances resembled the speech stimuli of the first experiment more than the baseline stimuli did. More importantly, the word frequency and the number of repetitions of the word on Day 2 had a strong impact on the success rate in the AXB task. Low frequency words which had been heard many times were most reliably identified as imitations. This specific pattern in the data indicates that the subjects success in the AXB task could not be an artifact of some global effect, such as overall fluency on the word set. Instead, it is word-specific.

The modular feed-forward model summarized above has a certain capability for handling the finding that long-term perceptual memories of words include voice information. Clearly, words have numerous associations or connotations, which would figure in the lexical model as a set of links to the words. Nothing in the model prevents specific words from evoking specific voices or sets of voices. However, the modular feed-forward model cannot handle the finding that these voice memories for words impact phonetic details in production. In this approach, retrieval of the lexeme means that a categorical encoding of that lexeme is loaded into the phonological buffer for execution. If the voice memories do not result in a categorically distinct form of the lexeme, they can have no impact on the production of the form. With the data in Goldinger (2000) showing gradient effects of word frequency and repetition count on perceived accuracy of imitation, the strict modularity of the standard model does not appear to be viable.

Bybee (2001) reviews a considerable body of literature on leniting historical changes, including both her own work and the landmark paper Phillips (1984). An example of such a change would be reduction of full vowels to schwa, with eventual loss of the schwa and the syllable it projected. Such changes are typically more advanced in high-frequency words than in low frequency words; the effects of word frequency appear to be gradient. Another example is provided by English doubly-marked past tense verbs (such as *left*, past tense of *leave*). As shown in Bybee (2000b),

the rate of /t/ reduction and/or deletion in such forms is a function of their frequency. As a static state of affairs, such a difference can, in principle, be accounted for by an on-line factor, as discussed above, However, this proposal fails to account for the fact that historical leniting changes advance on a scale of decades. They advance in two senses. Phonological sequences which are at first lenited become more and more lenited until they disappear entirely. Leniting changes which first become evident in high-frequency words typically spread to low-frequency words, in the end affecting the entire vocabulary. An example is provided by the history of French. In French, the post-tonic syllables of Latin eventually disappeared entirely, leaving an entire lexicon with word-final stress.

To model the progress of such effects requires a model in which the lexical representations of words include incrementally updated information about the phonetic distribution for each word. An across-the board effect, in the form of a consistent leniting bias on the productions, explains why all words are affected. However, high frequency words are affected more, because they are produced more often and so more memories of them in their lenited form accrue, once the lenition gets underway. Exactly such a model is developed in Pierrehumbert (2001), on which more below.

Word-specific effects in historical change are not confined to word-frequency effects. Yaeger-Dror & Kemp (1992) and Yaeger-Dror (1996) document a vowel shift in progress in Quebecois French. They found that a particular group of words failed to shift despite exhibiting the phonological sequence which was targeted in the change. These words were a group of semantic associates, representing organs of the church, the military, and the schools. Yaeger-Dror was not able to identify any phonological properties shared by these words which distinguished them from words which did undergo the shift.

A phonetic pattern with a morphosyntactic component is discussed in Hay (2000). Hay examined the production of /t/ in words such as *swiftly* and *listless*. Target word pairs in the study were phonologically matched, and differed in their degree of morphological decomposibility, as predicted by Hay's model. Taking into

account psycholinguistic results on morphological parsing, Hay predicts that complex words of equal frequency will be perceived as more decomposed when the base is more frequent than the word itself, and less decomposed when the base is infrequent compared to the word itself. For example, *swiftly* is highly decomposible because *swift* is readily perceived inside *swiftly*. However, *list* is not perceived inside *listless*. Hay found a significant effect of decomposibility on the /t/ allophony. (It appears to be gradient, though a larger data set would be desirable.) The more decomposible the form is, the more strongly the /t/ is pronounced. Note that this effect is in the opposite direction from an effect of base frequency per se. Given Hay's experimental design, the bases of the more decomposible words were **more** frequent than the bases of the less decomposible words. Nonetheless, they were produced with a **stronger** /t/. Thus, the pattern could not result from an on-line effect of the frequency of stem (as related to ease of access of the stem). Nor do they relate to word frequency of the complex form, since this factor was controlled in the experiment. The pattern can be generated in the model described below, in which the long-term representations of words include probability distributions over phonetic outcomes.

Mendoza-Denton (1997) and Hay, Jannedy & Mendoza-Denton (1999) hint at how word-frequency effects and lexical field effects may come together in the cognitive model. Mendoza-Denton (1997) reports the degree of raising and fronting of /ɪ/ in the speech of Latina gang girls in California. Hay et al. (1999) studied the degree of monophthongization of the diphthong /aɪ/ (a characteristic of African-American Vernacular English) in the speech of the African American TV personality Oprah Winfrey. Monophthongization of /aɪ/ might be viewed as a lenition, but raising and fronting of a lax front vowel to its tense variant is clearly not a lenition. However, in both studies, the shift is most marked in high-frequency words which serve as markers of sociolinguistic register. Fronting and raising of /ɪ/ was greatest on the word *nothing*, which acts as a discourse marker in the dialect in question. Monophthongization of /aɪ/ was strongest on the word *I*. In addition to interacting with word frequency, these words reflected the

sociolinguistic situation of the speaker. For the gang girls, the shift was most advanced for core gang members. Oprah Winfrey displayed the ability to shift her speech style between a more AAVE influenced style to a more mainstream style, depending on the subject matter she was speaking about. Another striking example of gradual adaptation of sociolinguistic style is provided by Harrington, Palethorpe & Watson (2000). This study, based on decades of radio broadcasts by Her Majesty Queen Elizabeth II, showed that she has gradually shifted her pronunciation in the direction of the Southern British English which has become fashionable with younger speakers.

4. Exemplar production models

The results reviewed in the last section all point to a model in which speakers learn implicit frequency distributions over phonetic outcomes, these distributions are stored in long-term memory, and they are subject to incremental updating. The psychological literature on categorization in perception provides precedents for models of this class. The approach I will be working with here is exemplar theory.

The critical ingredients in exemplar theory are a map of the perceptual space and a set of labels over this map. A clear example of a map is provided by the lowest level of encoding in visual perception. This is a sort of mental movie screen on which the neural signals from the retina are displayed. An example of a long-term memory of a visual map would be a long-term memory of a visual environment, such as one's own bedroom. The labels would be objects in this scene, such as *bed*, *lamp*, *bookcase*.

For phonetics, the relevant physical domain is the articulatory/acoustic space, whose dimensions are the relevant dimensions of contrast in articulation and acoustics. This domain provides the perceptual map for phonetic encoding and memory. The familiar F1−F2 space for vowels shows part of the information encoded in this map, but the real map is of course much higher dimensional. The higher dimensional space is still a space, however, in

the sense that a metric is defined along each dimension. Thanks to this metric it is possible to quantify the distance between any two stimuli in some single respect, or in all respects. The labels over the map are the inventory of phonological primitives, e.g. phonemes, features, or other phonological units.

According to exemplar theory, people have detailed long-term memories of particular percepts, and these are stored as locations on the map. These are the "exemplars" of the theory. Exemplars are categorized using the label set, and this has the result that each label is associated with a large set of remembered percepts. These implicitly define the region of the map which corresponds to that label. For example, the set of exemplars labelled with /i/ implicitly defines the region of the formant space which corresponds to that vowel; at the center of this distribution, the exemplars are numerous whereas towards the margins of the distribution, the exemplars become much sparser. A fresh stimulus is classified as follows. The perceptual encoding of the stimulus causes it to be placed at some location on the map. Given that location on the map, a statistical choice rule determines its most probable classification, given the number, location, and labelling of the previously stored exemplars in the region of the fresh stimulus. As discussed in Johnson (1997) and Pierrehumbert (2001a), this approach is highly successful in capturing the interaction of similarity and frequency in perceptual classification. It is also successful in handling prototype effects. Of course, we view the model as a logical schema rather than taking it as a literal picture of activity in the brain. Any model which stores implicit and incrementally updatable frequency distributions over a cognitive map will show similar behaviour; it is not important that all percepts are individuated as separate memories in the long term. The statistical choice rule is presumably physically implemented through activation and lateral inhibition of labels competing over a neighbourhood of the map.

The phenomena described in the last section are all patterns of speech perception. Classical exemplar theory says nothing whatsoever about production. Therefore, the model must be extended if it is to be applied. Goldinger (2000), Pierrehumbert (2001a), and Kirchner (forthcoming) all adopt a similar viewpoint on how to

obtain productions from an exemplar model. The basic insight, which appears to originate with work on motor control by Rosenbaum et al. (1993), is that activating the group of exemplars in a subregion of the perceptual map can specify a production goal which corresponds to the aggregate or average properties of the members of the group. To produce an /i/, for example, we activate the exemplars in some area of the /i/ region in the vowel space. This group of /i/s serves as a goal for the current production, much as a perceived object can serve as a goal for a reaching motion.

Models of this general class predict strong effects of degree of language exposure on production. Acquiring a fully native accent in a language involves building up probability distributions for all the different phonological elements in their various contexts, a task of empirical estimation which requires hearing and encoding a very large amount of speech. A variety of consequences is predicted from low levels of exposure, for example in the phonetic patterns which result from attempting to imitate a different dialect. This imitation will succeed only if the speaker has some amount of exposure to the dialect, and it is to be expected that the most frequent and perceptually salient features of the dialect would be imitated the most accurately, since utterances exhibiting these features would serve to establish labels and phonetic distributions characteristic of the dialect. However, without very extensive experience with the dialect, errors in establishing the label set and effects of undersampling would combine to predict various kinds of over- and under- generalization in phonetic outcomes. It is also important to note that the exemplar space itself provides a strong cabability for generalization based on phonetic similarity. Other types of generalizations are also supported by the model, because the model has multiple levels of representation. For example, the characteristic stress pattern of nouns and verbs differs in English, and learning this generalization requires access to the syntactic level, at which the variables N (noun) and V (verb) are defined. Although such phenomena are not the focus of the present paper, nothing about the model prevents relations of abstract levels of description to be established with other, even more abstract, levels.

In formalizing the model, I will adopt the specifics of Pierrehumbert (2001a). In this model, production of the phonolog-

ical category represented by any specific label involves making a random selection from the exemplar cloud for that label. The selection is random because of the kind of variability which is displayed in productions. If the production model always selected the single best exemplar (by any measure), then the production goal would be invariant. In fact, however, the outcomes vary with variables at nonphonological levels (such as speech rate, style, and speaking conditions). The aggregate effect of such variation as viewed from within the phonological model is random variation over the exemplar cloud; I will return below to the hidden systematicity which a more complete model should capture. The mathematical nature of random sampling does of course entail that the location selected is more likely to be in a densely populated part of the exemplar cloud than in a sparse part.

The specific equations of this model are as follows, repeated from Pierrehumbert (2001a). The exemplar list $E(L)$ consists of the list of exemplars $\{e_1^L, \ldots, e_n^L\}$ associated with label L. To decide which label to assign to a new utterance with phonetic characteristic x, we define a score for each label by the equation

$$(1) \qquad score\ (L,\ x) = \sum_{i\ =\ 1\ldots n} W\,(x - e_i^L)\ exp\left(-\frac{t - T_i}{\tau}\right)$$

where W is a window function, t is the current time, T_i is the time at which the i^{th} exemplar was admitted to the list, and τ is the memory decay time. Different exemplars have different strengths in the model, because memories are assumed to decay in time. An exponential decay of the exemplar strength is used to model this effect. The window function is a square in Pierrehumbert (2001a) but other choices are possible.

In production, a target x_{target} is obtained by picking an exemplar randomly from the exemplar list of the desired label. Since the probability (or strength) of each exemplar is time-dependent, old exemplars are only rarely used as targets. The actual production target is formed by taking a group of exemplars around this random element. This is necessary for the system to behave correctly as experience increases. If just a single exemplar is chosen

as the target, and if it is produced with some probabilistic degree of error (arising as random variation in the motor system), then the phonetic distribution for any given label will spread out more and more. In fact, experience tends to make distributions sharpen up, a phenomenon known as entrenchment. Using a region around x_{target} to control productions puts an anti-diffusive factor in the model, which causes productions to be biased towards the center of the distribution. Specifically, n_{trench} closest exemplars to x_{target} are selected using the memory-weighted distance

$$d_i = \left| x_{target} - e_i^L \right| exp\left(-\frac{t - T_i}{\tau}\right) \tag{2}$$

A new target is formed by taking the memory-weighted mean of these n_{trench} values. In the limit of very large n_{trench}, the production target becomes fixed at the memory weighted mean of the exemplar list. The final x_{target} is then produced with some random error ϵ.

$$x = x_{target} + \epsilon \tag{3}$$

In the case of a leniting bias, this equation has the form:

$$x = x_{target} + \epsilon + \lambda \tag{4}$$

On the assumption that exemplar clouds are associated with phonological units – such as phonemes – models of this class readily handle phonologization of phonetic tendencies. The model discussed in Kirchner (forthcoming) uses spreading activation in a connectionist framework to derive specifics. Clearly, an exemplar production model has to associate exemplars with phonological units, either directly or indirectly. Otherwise, it would be impossible to pronounce novel forms, such as words learned through reading.

One might also assume that exemplar clouds are directly associated with words. Clearly, rather complex memories can be associated with particular labels; for example, I associate a mental image

of the photograph on Keith Johnson's web site with the label *Keith Johnson*. Equally, I could associate a recollection of a sizable speech fragment with the word that it instantiates. To do this, it is necessary to impute a temporal dimension to the perceptual map; but this is probably necessary even for modeling phonological units, since phonological units have characteristic dynamics. On the assumption that exemplar clouds contain longer perceptual traces which are directly associated with word labels, the approach readily handles most of findings of the last section; only Hay's results on morphological decomposibility require further apparatus which will be provided below.

Specifically, in Goldinger's (2000) experiment, the exemplar distribution associated with each word would be impacted by the repetitions of the word encountered on Day 2 of the experiment. The more repetitions encountered, the more the distribution would be impacted. Furthermore, for low frequency words, the proportion of exposures which occurred in the context of the experiment would be higher than for more common words. Thus, the proportional effect of the target voices on the mental representation would be higher for low frequency words than for high frequency words, as Goldinger actually found.

Wright's (1997) findings would fall out from the fact that words with a low neighbourhood density are (all else equal) more readily recognized than words with a high neighbourhood density. If a word has no similar competitors, then even a rather slurred example of it will be recognized as a token of the word. As a result, the exemplar distribution for successfully recognized instances of low density words will include more reduced tokens than for high density words. This account leaves us with two different mechanisms for explaining Wright's data, and in fact, both could be involved.

The findings about Quebec French in Yaegor-Dror & Kemp (1992) and Yaegor-Dror (1996) would fall out if words in a particular semantic domain are dominantly used in a social group dominated by older speakers and/or in a formal speech register. In this case, the frequency distributions for words used colloquially in everyday interactions would drift while those in the exceptional semantic field would stay in place.

Pierrehumbert (2001a) assumes that exemplar clouds are associated with phonological units as exhibited in words. Consider the process of vowel reduction in the context of sonorants, one of the initial cases for which Bybee established a relationship between word frequency and degree of reduction. (See Bybee, 2001) To model this effect, it is necessary to assume that the change in progress refers to a structural description within each word, namely the vowel-sonorant combination targeted by the change. A persistent tendency to hypoarticulation of this combination is a language-specific instantiation of broad tendencies to hypoarticulate as discussed in Lindblom (1983). It induces the persistent production bias represented by the variable λ in equation (4). It is also necessary to assume that phonetic distributions for individual words are maintained. The perceptual memories of the lenited word tokens accumulate, incrementally updating the distribution for the word. Since high frequency words are produced more often than low frequency ones, the listener encounters more numerous word tokens which have been affected by the leniting change. As a result, the frequency distribution of outcomes for high frequency words is shifted further in the direction of the historical change than for low frequency words. Obviously, this treatment is not confined to lenition; any systematic bias on the allophonic outcome would incrementally impact high frequency words at a greater rate than low frequency words. In short the model is applicable to any Neogrammarian sound change, by which I mean sound changes which get started in the phonetic implementation and eventually sweep through the vocabulary. (Analogical sound changes, in which words shift their pronunciations categorically through pressure from morphological relatives are generally agreed to arise a a different level in the system).

In this treatment, the exemplar distributions associated with particular phonological units arise as the union of the relevant temporal subparts of exemplars associated with words. For example, each words containing the vowel /o/ would contribute the region which manifests the /o/ to the exemplar distribution for that phoneme. Of course the allophony of this vowel depends on the segmental context and other factors, such as word frequency. This

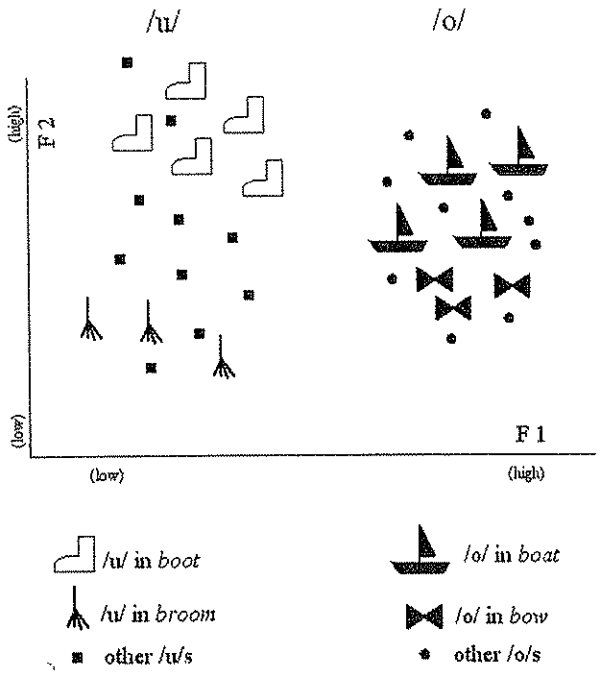

Figure 1: Hypothetical distributions of exemplars for /o/ and /u/.

situation is illustrated by Figure 1, showing exemplar locations for /o/ in *bow* and *boat*, as well as /u/ in *broom* and *boot*. The perceptual memories of in *boot* and *boat* have generally higher F2 values than most other words with the respective vowels, because the /t/ causing fronting of the round back vowels. Since some instances of the word *bow* occur in coronal contexts, some of these /o/s could be rather fronted, too.

However, some awkward and crucial issues about the relationship of word-level and phoneme-level labelling of the exemplar space are swept under the rug in Pierrehumbert (2001a). Specifically, the entire production model presupposes that the perceptual labelling of the space is simply used "in reverse" in production. When a label is activated through the speaker's lexical choice, then a region of the exemplars associated with that label is activated and guides the production. With word-level labels being associated directly with the exemplar space, it is unclear what enforces a phonological decomposition of the word in the first place. Why couldn't the exemplars of any given word provide a holistic plan for the production of that word? If this were possible, two awkward consequences would ensue. First of all, there would be no necessary reason why productions of a word would actually be

subject to the persistent bias which encapsulates the historical change in progress. Recall that the allophonic principle has a structural description, and therefore entails a phonological decomposition of the word. Direct linking of words to phonetic outcomes bypasses this level of analysis entirely, and so it is unclear why the leniting bias would impact every production of the word. In fact, the general assumptions of spreading activation models (to which exemplar models are closely related) would be that frequently used activation pathways tend to be more active. High word frequency would tend to enhance the "holistic" route of word production, if such were available, thus exempting high frequency words from a historical change in progress. Obviously, this is contrary to the finding that high frequency words tend to lead Neogrammarian sound changes.

A second, and conceptually related, problem relates to the fact that Neogrammarian sound changes normally sweep through the vocabulary. Even if allophonic rules are not absolutely across the board, they are nearly across the board. Cases such as that of vowel change in Quebec French are unusual, and even in this case there are only two classes of words – the majority that undergoes the change and a semantic group that resists it. The long-term results of a system in which words activated exemplar clouds directly (without implicating a phonological decomposition) would, however, be arbitrary phonetic dispersion. Because of the cumulative nature of the perception-production loop in this model, tiny differences between words build up over time as a function of differences in their social and linguistic contexts of use. Thus, the words would disperse through the available regions of the phonetic hyperspace. The notion that words can be mapped directly to phonetic outcomes is at odds with the phonological principle, according to which human languages are characterized by the repeated use of a small number of phonological elements which are found in many different combinations to make words. Though this characterization of language is not absolutely true (as we have seen), it is approximately true and the fact that language can be approximated in this way is an important one.

In the next section, I sketch and compare two alternative models for overcoming these problems.

5. Models

The two models I will discuss share a number of features. First, in both the production system is closely tied to the perceptual system, with the same levels of representation appearing in both perception and production. As discussed, these include the lexical network itself (in which only lexeme information interests us here) and the exemplar space over which frequency distributions are built up. Secondly, in both there is an intermediate level of representation, that of phonological encoding, which has a privileged position amongst the labels over the exemplar space. It is the level at which time actually unfolds (in contrast to the lexicon and the exemplar space, which are long-term memories of linguistic events).

In production, this intermediate level corresponds directly to the level of phonological encoding and buffering found in modular feed-forward models, as discussed in section 2 above. In particular, I assume that this level represents procedural knowledge; that phonological representations of words are incrementally loaded into this buffer; that a complete phonological parse including metrical and intonational structure up to the phrasal level is assigned here. Thus, the major deviation from previous views is the radically representational concept of phonetic implementation. The contents of this buffer are not subject to phonetic implementation rules in the traditional sense. Instead, they probabilistically evoke regions of the exemplar space as production goals.

In perception, there is increasing experimental evidence for an analogous level of processing, which is termed the Fast Phonological Preprocessor (or FPP) in Pierrehumbert (2001b). This level uses language-specific but still general knowledge of the phonotactic and prosodic patterns of a language to parse the incoming speech stream. The critical function of this level is hypothesizing possible word boundaries, that is, identifying temporal locations in the speech stream at which a lexical search should be initiated. All current models of word recognition (such as Norris 1994, Vitevich & Luce 1998, Norris, McQueen & Cutler, 2000) describe word recognition in terms of words being incrementally activated as a

reflex of their similarity to the speech stream. Simultaneously activated words compete through mutual inhibition, until one candidate wins over the others. There is clear evidence that staggered or partially overlapping candidates are implicitly considered during word recognition. For example, during processing of the word *festoon*, the word *tune* might be activated beginnning at the /t/. In short, multiple desynchronized word candidates are maintained in parallel as the competition plays out. However, it would be problematic to assume that fresh lexical searches are launched absolutely continuously – for example, every 5 msec, representing the temporal resolution of some digital speech processing systems. The kind of partial overlaps which have been reported arise when junctural statistics or prosody suggest that a word boundary might be present (See e.g Cutler & Norris, 1988; McQueen, 1998; Content, Dumay & Frauenfelder, 2000). The activation of *tune* in *festoon* occurs because most English words begin in a stressed syllable, and the stress is accordingly a probabilistic cue for a word boundary. The word *emu* is not necessarily activated in processing the word *honeymoon*, even though it is possible to find an acoustic subregion of this word which sounds a great deal like *emu* when played in isolation. By hypothesizing possible word boundaries, the FPP places practical bounds on the number of number of possibly staggered candidates maintained in parallel.

According to Norris et al. (2000), decontextualized phoneme decisions (as in a phoneme monitoring experiment which requires subjects to push a button for any word containing /p/) are made in a module which is distinct from the FPP and which is influenced by lexical matches (if any). This suggestion, which I take to be well-supported, liberates the FPP from the task of providing a phonemic transcription. This is a welcome result, since the FPP is a bottom-up processor and the efforts of phoneticians and speech engineers show bottom-up phonemic transcription to be extremely problematic. It is statistically fragile and it discards allophonic information which is demonstrably used in recognizing words. For example, Dahan et al. (2000) show that misleading coarticulatory information affects the time course of lexical matching as reflected

in eye-tracking data. Furthermore, if subphonemic information were discarded during lexical access, then there would be no way that it could accumulate in long-term memories associated with particular words.

My conclusion, then, is that the FPP maintains extremely detailed phonetic encoding, and that its primary contribution is to add parsing information. A mental image of a grainy spectrogram decorated with a prosodic parse can provide a mnemonic for this conclusion. The results of lexical access and even post-lexical decisions can continue to add labelling to this same structure. Fragments of phonetic streams labelled in this way provide material for the labelled exemplar space.

Now, I come to two alternatives on how detailed phonetic outcomes can be associated with particular words under this model. The examples I will be using come from a pilot experiment on glottalization at a morpheme boundary, in the words *preoccupied*, *high-octane*, *overarching*, *realignment*, and *reenact* as produced in sentence contexts. Such glottalization is not contrastive in English, and it shows a considerable range of phonetic variation, from a full glottal stop to creaking voicing to a merely pressed voice quality. The target words all have somewhat idiosyncratic meanings, with the least familiar to foreign readers possibly being *high-octane* on the meaning of "forceful". Baseline data were also collected on the rate of glottalization at word boundaries for vowel initial words following a function word, in sequences such as "to Egypt" and "are available". In this experiment, stress was manipulated through design of the materials, and speech style was manipulated through direct instructions to the five subjects, who were Northwestern undergraduates enrolled in introductory linguistics courses. The summary data show a cumulative interaction of stress, morphosyntactic status, and speech style, as shown in the following table. The patterns shown were also found within the speech of individuals. The lexical issue which appears in these data is the relationship of allophony in a base form to allophony in a derived form. In the discussion, I will also return to some of the phenomena summarized above which are not exemplified in this data set.

Table 1:

	Stress-Clear	Stress-Normal	Unstress-Clear	Unstress-Normal
BaseL	100%	80%	52%	8%
Mor. Complex	75%	20%	12%	0%

The most straightforward extension of the modular feed-forward model would seek to model the probability distributions for the phonetics of individual words via links from the lexemes to units of phonological encoding. Note that all of the words in the data set have idiosyncratic meanings, and thus must be lexical entries. In order to generate the various outcomes, this model requires that glottalization be available as a category in the phonological encoding. I will transcribe it as a glottal stop, /ʔ/, despite its wide phonetic range.

Figure 2 sketches how this would generate the contrasting outcomes for the two words *high-octane* and *reenact*, each of which has glottalization at the VV hiatus some of the time. In order to illustrate the point, rates of glottalization in this figure are taken from data on these stress configurations in the clear speech condition, since no glottalization of stem-initial schwa was found in the normal speech condition.

Note that the final distribution of glottalization for *reenact* shows a less frequent and less extensive glottalization than for *high-octane*. The prosodic parse for *high-octane* shows stress on the second vowel, and the exemplar cloud associated with this position has more frequent and more forceful glottalization than for the unstressed case. In a classic modular feed-forward model, this regularity would be described by having two different phonetic implementation rules, one for the stressed condition and one for the unstressed condition. In the present model, these rules are replaced by associations between phonological fragments (including relevant prosodic structure) and probability distributions over the degree of glottal constriction.

Now, let us consider the speaker who glottalizes more in *high-octane* than in *reenter*. This outcome can be encoded, if not necessarily explained in the model, provided that the mental representa-

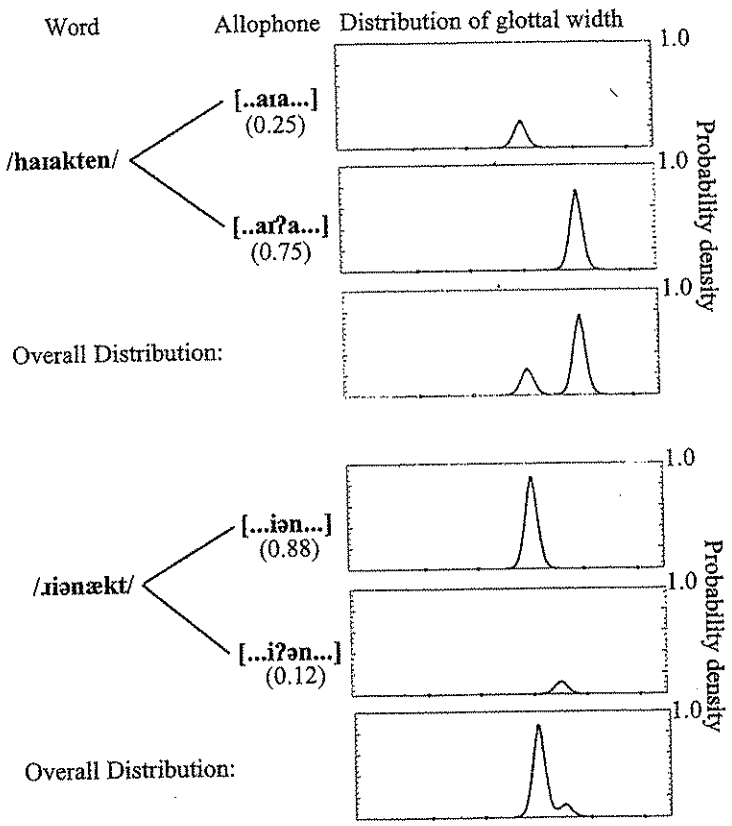

Figure 2: Contributing and total frequency distributions for degree of glottal adduction (spread ... constricted), with two rules for glottal stop insertion.

tions maintain implicit frequency counters on the pronunciations as encountered in perception. (Rates of glottalization in this figure are taken from pooled data on these words in the normal speech condition.)

By comparing the displays for *high-octane* in Figures 2 and 3, we can also see how clear speech style would be modelled under this approach. It must affect the probabilities for the rules mapping to allophonic outcomes, in order to model the fact that glottal stop insertion is most frequent in clear speech. In addition it must affect the force and speed of articulation, a factor not illustrated here.

Speaking more generally, assume that the model maintains for each word a probability distribution over some number of categorically different phonetic outcomes. For words such as *cat*, these might be an aspirated released /t/, a plain /t/, and unreleased plain /t/, a glottalized /t/, and a glottal stop, providing in effect a five-step scale along the dimension of [spread glottis] – [constricted

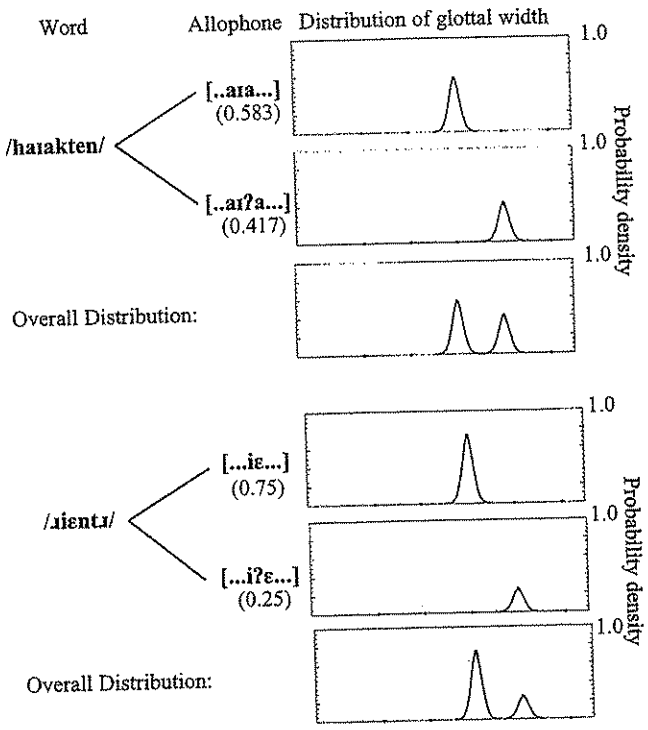

Figure 3: Contributing and total frequency distributions for degree of glottal adduction (spread ... constricted) for two words with idiosyncratically different rates of glottalization.

glottis]. Production of each outcome in turn relates to a probability distribution over the phonetic space. In this case, the total space of outcomes for each word is a weighted sum of the distributions for the variants. This can be conceptualized as a set of mountains which maintain their location but differ in their height, as in Figure 4a. However if the distributions are wide compared to the separation between them, the peaks in the result need not correspond to the peaks in the underlying distributions. This is shown in 4b, which provides a strong appearance of gradient effects over the phonetic space. In particular, a variable mixture of two distributions can cause the mean and even the mode to exhibit a gradient dependence on the proportion describing the mixture. An approach in which superficial distributions such as 4b are uniformly attributed to mixtures of underlying categories will be termed the "secret categories" approach. Since a distribution such as 4b can arise mathematically either as a wide distribution for a single label, or a mixture of the different distributions for a set of labels, further considerations must be brought to bear to evaluate this approach.

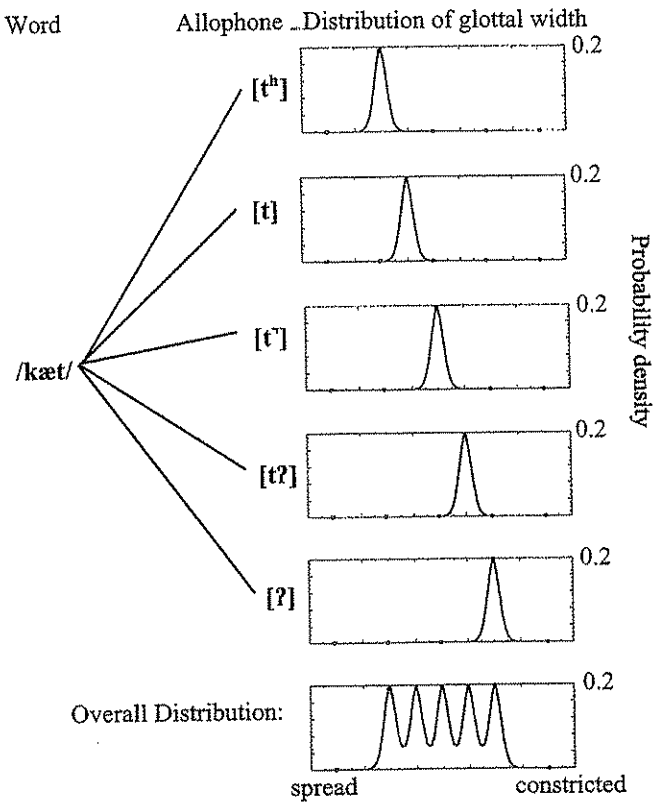

Figure 4 a: Mixtures of sharp categories.

A second model, and the one I will advocate, draws on the proposals made in Johnson (1997) about speaker normalization. Johnson (1997) builds on findings that people have long term memories of the specific voices in which words were spoken to propose that the exemplar space is labelled by speaker as well as phonologically. Speaker normalization occurs through attentional weighting of the exemplars. For example, in attempting to classify an incoming stimulus as /ɛ/ or /ɪ/, the basic statistical choice rule would be sensitive to all /ɛ/s or /ɪ/s in the neighbourhood of the incoming stimulus. If I know I am listening to my 11-year old daughter, however, I can weight more highly exemplars which originated from her speech. Since her vocal tract is shorter than that of an adult, the net effect would be to shift the F2 boundary for /ɛ/ versus /ɪ/ in this perceptual classification, a typical example of successful speaker normalization. Extended to production, this would mean that activating memories of a particular speaker does not in itself cause speech to come pouring out. However, if one is

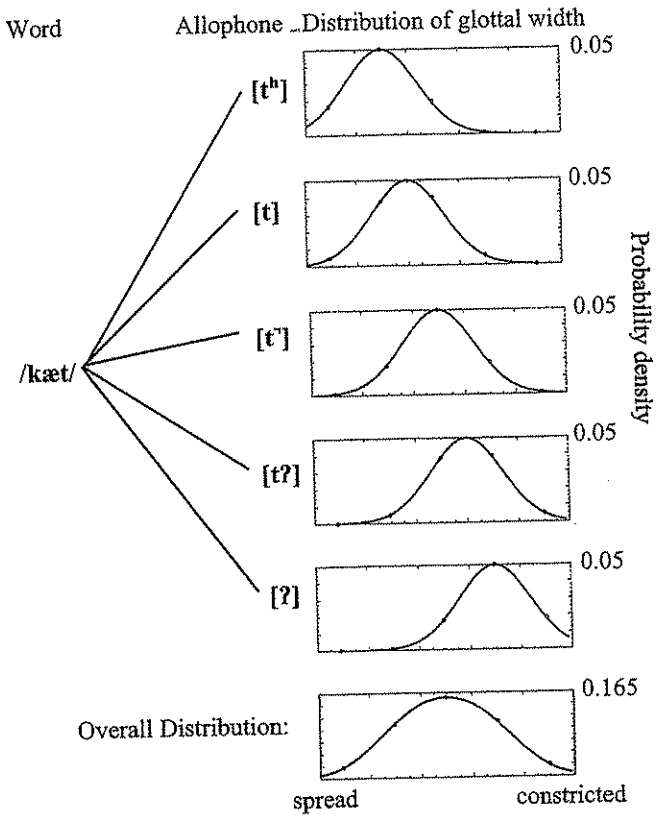

Figure 4 b: Mixtures of soft categories.

speaking, then activating memories of a particular speaker can bias the productions which occur. The productions would be biased towards phonetic imitations of the speaker in question, because exemplars of that speaker's productions would be more activated and thus acquire a disproportionate role in shaping production targets.

Implementing this model is a very straightforward extension of the equations introduced above. For each exemplar e_i, define a weighting coefficient A_i^t. This is the extent of activation of exemplar e_i at time t due to its originating in the current word. The value of this coefficient depends on time because the activation is temporary. Now, recall that a weighting coefficient related to the temporal decay of the exemplars was already included in the model, namely.

$$(5) \qquad exp\left(-\frac{t-T_i}{\tau}\right)$$

The overall weighting of the exemplar can then be treated as a product of our two coefficients (for some appropriate choice of units).

$$(6) \qquad exp\left(-\frac{t - T_i}{\tau}\right) A_i^t$$

This overall weighting will then be active at all three critical points in the dynamics of the model: Highly weighted exemplars will play a stronger role in classification. They will be more frequently selected as the core of a production target. And they will be more influential in the aggregate target computed from the neighbourhood of the core. An important corollary of the weighting scheme is that the influence of particular words on phonetic outcomes is secondary, with the actual phonological makeup of the words providing the primary influence. That is, individual words can be shifted or biased within the space provided by their phonological makeup, but not into regions of phonetic hyperspace which are not used generally in the language. This is correct, and goes towards explaining why many of the patterns in question have only been found recently as large scale studies become possible.

This model will tend to transfer allophony from a base form to a morphologically related complex form, without requiring an allophone to be projected as a category. For the glottalization, this works as follows, assuming that word-initial glottalization is provided by a direct mapping from the triggering context (the word-initial vowel) to dimensions of the exemplar space such as vocal fold adduction. Consider the word *realign*. Consider first the evolution of the situation for an initial condition in which *realign* has no glottal attack on the second syllable, just like *realize*. (1) When *realign* is activated, activation spreads to *align*. (2) Producing *realign* involves producing a schwa in the second syllable. Examples of the schwa which originate from this word are weighted in establishing the production plan. In the initial condition, these have no glottal attack. (3) Because of the spreading activation of *align*, exemplars originating from this word as also weighted. These exemplars have a glottal attack. (4) *realign* is therefore prob-

abilistically produced with a glottal attack, though on average less than *align*. (5) Tokens of *realign* which have glottalization update the exemplar distributions (of other speakers), as do ones which lack glottalization.

Now consider the contrary case, in which *realign* is productively generated as a neologism from *re-* and *align*. Under this scenario, the initial condition is that in which the word *align* is actually produced, and therefore one would find exactly the same rate of glottal attacks in both words. If *realign* becomes stored in the lexicon as a unit, then its production entails activation of the exemplar space for the VV hiatus, and this exemplar space will include exemplars of unanalyzed words such as *realize*. Though tokens of *realign* (insofar as they are available) are positively weighted in establishing the production goal, they are not the sole factor. In this case, *realign* will evolve to shower fewer and weaker glottal attacks than *align*. To summarize, then, the pronunciation of nondecomposed *realign* will tend to evolve towards the phonetic pattern of *align*, to the extent to which the word *align* is perceived within *realign*. The pronunciation of transparently decomposed *realign* will evolve toward the phonetic pattern of *realize*, to the extent that the word becomes lexicalized as a whole. If the word is partially or sporadically decomposed, the phonetic pattern will end up in the middle. The underlying assumption, is that morphological decomposition is gradient. This assumption follows from current morphological processing models notably Caramazza, Laudanna & Romani (1988), Frauenfelder & Schreuder (1992), Schreuder & Baayen (1995), Wurm (1997), Baayen & Schreuder (1999), and Hay (2000). Unlike generative linguistic models, in which a given word either is or is not morphologically decomposed, the processing models suggest that examples can be found all along the scale from fully simplex to fully decomposed. The glottalization results in Table 1 are what the model would predict for semi-decomposed words – for each stress and speech style condition, the rate of glottal attacks within the complex words is less than that at a full-fledged word boundary. The model also predicts gradient allophonic results as a reflex of gradient decomposability. This is exactly the finding of the Hay (2000) experiment on /t/ allophony discussed above.

Of course, one would not wish to deny that transfer can also occur at a categorical level. Examples are provided by speech errors such as *fat* for the past tense of *fit* (under the pressure of *sit/sat*); and the paradigm levelling which is widely attested in historical morphology. In such cases, however, one observes two qualitatively different outcomes without examples of any in between. Speech errors such as *fat* do not necessarily imply the existence of a set of intermediate cases along the vowel height dimension.

The model predicts a cumulative effect between the probability/degree of glottalization, and any bias represented by the parameter λ. λ was introduced to describe a persistent leniting bias. However, the work by Lindblom to which it hearks back proposes a continuous scale of hypo- to hyper-articulation (Lindblom, 1984) The clear speech style in the data set obviously provides an example of hyper-articulation. The results of Table 1 are broadly in line with prediction. They give percentages of tokens which crossed a threshhold of glottalization. All of these percentages are shifted up under the clear speech condition, preserving, however, the rank ordering of the cases. More detailed consideration of these numbers raises some issues, since λ only shifts the degree of some gesture which is already planned, and does not bring it into being if it did not exist in the first place. The shift of 80 to 100 percent (for the case of a stressed word boundary), or 0 to 12 percent (for unstressed morpheme boundaries) would thus need to be interpreted as threshholding artifacts. Such numbers would follow under the assumption that a sampling of x_{target}s for the former case are all glottalized to some degree in normal speech (with 20 percent of targets showing such slight glottalization that it is below threshhold), and that the sampling x_{target}s for the latter case includes at least some number of tokens with a glottal adduction gesture, which is amenable to being strengthened. These assumptions would need to be validated with more conclusive measures, such as stereofibroscope pictures of the vocal folds. There is also a hint in the numbers that the influence of morphological relatives actually increases for the clear speech condition. If this proves to be the case, then the parameter λ is in the wrong place in the model; an effect of this type would require λ to bias the underlying sam-

pling for the production plan, rather than shifting the production plan post-hoc.

Thus, the second model readily captures the tendency for morphologically complex forms to be influenced by the allophony of the base. It predicts that complex forms are **more** influenced by the allophony of an embedded morphological relative than by the allophony of a phonologically embedded word which is unrelated. For example, insofar as the word *mislay* is decomposed, it would be more influenced by *lay* (predicting a well-voiced /l/) than by *sleigh* (predicting a largely devoiced /l/). This prediction follows because *mislay* activates *lay*, which would then bias the set of exemplars contributing to the production goal. In contrast, *mislay* does not activate *sleigh* (in fact, it competes with it in perception). Thus *sleigh* has no particular privilege to affect the pronunciation of *mislay*, with any commonalities coming about solely from the common phonological content.

Further cases of allophonic transfers from morphological relatives are discussed in Steriade (2000) and Rialland (1986). The model predicts in particular the existence of cases in which relationship of phonetic outcomes to morphological relatedness is gradient. More large-scale experiments are needed to evaluate this prediction.

Another line of argument for the second model over the secret categories approach depends some general observations about categorization systems. When the equations presented above are applied iteratively in a simulation of the production-perception loop, two qualitatively different outcomes readily arise for cases like Figure 4b. In one outcome (representing a parameter range with a high degree of entrenchment), the distributions are gradually sharpened up until they become well separated, as in Figure 4a. The other outcome (representing a lesser degree of entrenchment), the distributions spread out. Over more and more intervals of the phonetic space, a competition arises between a more frequent label and a less frequent one. In this case, more and more tokens are assimilated to the more frequent categories until the less frequent labels are gobbled up and the distinctions in the system have collapsed. Notice that the slightest difference in frequency tends to

become amplified over iteration, since the perceptual classification is de facto biased towards the higher frequency label in any neighborhood. I have not actually been able to find a parameter range for this model which shows stable overlapping distributions. Given the complex nonlinear character of this model, there is at present no mathematical proof that all cases such as 4b are unstable. However, phonetic typology strongly suggests that situations such as 4b evolve towards either a sharper category system or a category collapse. The most studied cases of overlapping phonetic distributions are the "near-mergers" discovered by Labov, Karan & Miller (1991) in speech communities with varied dialects. In these cases, the overlapping categories carry a much higher functional load than those discussed here, because they distinguish words for some speakers and also provide socioeconomic information about speakers. Nonetheless, the actual perceptual discriminability of the labels is less than a statistical phonetic analysis would support, and the labels tend to collapse. The suggestion that the secret and non-meaning-bearing categories of the phonetic implementation system show stability properties which are not found even for lexically distinctive phonological units appears to be highly problematic. The second model makes it possible to reserve the projection of categories for situations when when a phonetic contrast is plainly bimodal and/or carries a high functional load.

More generally, the secret categories model relies on something like an IPA fine transcription to achieve coverage of the phonetic gradients which are observed. This is a level I have attacked elsewhere on the grounds that it has difficulty modelling the gradient cumulative effects in phonetic implementaion which are observed in experiments on continuous speech. (See Pierrehumbert & Beckman, 1988 regarding f_0, and Pierrehumbert & Talkin, 1992, regarding aspiration and glottalization). Direct mapping of more abstract entities to quantitative parameters meets with more success. A similarly broad issue is the reliance of the secret categories model on multiple representations for the same word. As argued in Bybee (2000a), there are strong cognitive pressures to maintain only a single representation for any given word. Lastly, the second model is more parsimonious. Findings on speaker normalization and long-term memory of particular voices strongly suggests that we

need attentional weighting on the exemplar space. In fact, a weighting scheme is needed to describe any kind of contextual effect which gradiently shifts category boundaries in either production or perception. Reusing this independently motivated device would appear to be preferable to proliferating categories over anything that would otherwise be viewed as a phonetic continuum.

6. Conclusion

The empirical studies which gave rise to modular feed-forward models of speech production provide strong evidence for distinguishing three levels in the cognitive system: lexemes, phonological encoding, and quantitative knowledge of phonetic outcomes. These levels are found in both perception and production. The models are successful in capturing the productiveness of speech processing and the existence of across-the-board effects.

More recent and detailed studies show that phonetic outcomes are not as across-the-board as they originally appeared. A number of cases have come to light in which allophonic details are systematically associated with words. Some of these effects (most notably those involving word accessibility in context) may arise on-line, but others are difficult to explain without assuming that individual words have associated phonetic distributions.

Exemplar-based production models provide a method for integrating these findings with prior work. Phonetic implementation rules are modelled through a correspondence between phonological labels and frequency distributions over the phonetic space. Individual words can bias the set of exemplars which serve as production goals. By assuming that words bias productions – rather than providing holistic production goals – the approach captures the fact that word-specific phonetic effects are second-order effects. The approach also makes it possible to capture cases of allophonic transfer between morphological relatives. Because of averaging in the system, it can capture gradient effects related to degree of morphological decomposibility. It establishes a connection between the likelihood of effects and their degree. Interactions at more abstract levels of representation must also be posited to handle nongradient effects.

Weighting of exemplars by individual words is not the only weighting in the system. Sociostylistic register and other contextual and attentional factors are clearly important. For this reason, the system does not predict an exhaustive match between perception and production. If speech tokens were perceived but not committed to long-term memory, they would fail to influence production. If the contexts for perception and production differed, differences in the exemplar sets activated for any given label would also ensue. Likewise, the connection of personal identity to sociostylistic register could also give rise to differences.

References

Baayen, R. H. & Schreuder, R.
1999 War and Peace: Morphemes and Full Forms in a Noninteractive Activation Parallel Dual-Route Model. *Brain and Language*, **68**, 213–217.

Bybee, J.
2000 a Lexicalization of sound change and alternating environments. In Broe, M. and Pierrehumbert, J. (eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, (pp. 250–269). Cambridge UK: Cambridge University Press.

Bybee, J.
2000 b The phonology of the lexicon; evidence from lexical diffusion. In Barlow, M. & Kemmer, S. (eds.), *Usage-Based Models of Language*, (pp. 65–85). Stanford: CSLI.

Bybee, J.
2001 *Phonology and Language Use*. Cambridge UK: Cambridge University Press.

Caramazza, A., Laudanna, A. & Romani, C.
1988 Lexical access and inflectional morphology. *Cognition*, **28**, 297–332.

Content, A., Dumay, N. & Frauenfelder, U.
2000 The role of syllable structure in lexical segmentation: Helping listeners avoid mondegreens. In A. Cutler, J. M. McQueen & R. Zondervan, *Proceedings of SWAP (Spoken Word Access Processes)*, Nijmegen, Max Planck Institute for Psycholinguistics. 39–42.

Cutler, A. & Norris, D.
1988 The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 113–121.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K. & Hogan, E. M.
2000 Tracking the time course of subcategorical mismatches on lexical access: Evidence for lexical competition. In A. Cutler, J. M. McQueen & R. Zondervan (eds.), *Proceedings of SWAP (Spoken Word Access Processes)*, Max-Planck Institute for Psycholinguistics, Nijmegen, (pp. 67–70).

de Jong, K., Beckman, M. E. & Edwards, J.
1993 The Interplay between prosodic structure and coarticulation. *Language and Speech*, **36**, 197–212.

de Jong, K. J.
1995 The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J. Acoust. Soc. Am.*, **97**, 491–504.

Dilley, L., Shattuck-Hufnagel, S. & Ostendorf, M.
1996 Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, **24**, 423–444.

Frauenfelder, U. H. & Schreuder, R.
1992 Constraining Psycholinguistic Models of Morphological Processing and Representation: The Role of Productivity. In G. Mooij & J. van Marle, (eds.), *Yearbook of Morphology 1991*. Dordrecht: Kluwer Academic Publishers. 165–185.

Goldinger, S. D.
1996 Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 1166–1183.

Goldinger, S. D.
2000 The role of perceptual episodes in lexical processing. In A. Cutler, J. M. McQueen & R. Zondervan, *Proceedings of SWAP (Spoken Word Access Processes)*, Nijmegen, Max Planck Institute for Psycholinguistics. 155–159.

Harrington, J., Palethorpe, S. & Watson, C. I.
2000 Does the Queen speak the Queen's English? *Nature*, **408**, 927–928.

Hay, J. B.
2000 *Causes and Consequences of Word Structure*. PhD dissertation, Northwestern University. (Downloadable from http://www.ling.canterbury.ac.nz/jen/)

Hay, J. B., Jannedy, S. & Mendoza-Denton, N.
1999 Oprah and /ay/: Lexical Frequency, Referee Design and Style. Paper R3TEL1, *Proceedings of the 14th International Congress of Phonetic Sciences*. 1389–1392.

Johnson, K.
1997 Speech perception without speaker normalization. In K. Johnson & J. W. Mullennix (eds.), *Talker Variability in Speech Processing*, (pp. 145–166). San Diego: Academic Press.

Keating, P., Cho, T., Fougeron, C. & C. Hsu
forthcoming Domain-initial articulatory strengthening in four languages. In R. Odgen, J. Local & R. Temple (eds.), *Papers in Laboratory Phonology VI*. Cambridge: Cambridge University Press.

Kirchner, R.
forthcoming Preliminary thoughts on "phonologisation" within an exemplar based speech processing model. *UCLA Working Papers in Linguistics* Volume 6.

Labov, W., Karan, M. & Miller, C.
1991 Near mergers and the suspension of phonemic contrast. *Language Variation and Change*, **3**, 33–74.

Levelt, W. J. M.
1989 *Speaking*. Cambridge MA: MIT Press.

Lindblom, B.
1983 Economy of speech gestures. In MacNeilage, P. (ed.), *The Production of Speech*. (pp. 217–245). New York: Springer-Verlag.

McQueen, J. M.
1998 Segmentation of Continuous Speech Using Phonotactics. *Journal of Memory and Language*, **39**, 21–46.

Mendoza-Denton, N.
1997 *Chicana/Mexicana Identity and Linguistic Variation: An Ethnographic and Sociolinguistics Study of Gang Affiliation in an Urban High School.* Ph.D dissertation, Stanford University.

Norris, D., McQueen, J. M. & Cutler, A.
2000 Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, **3**, 299–325.

Phillips, B. S.
1984 Word Frequency and the actuation of sound change. *Language*, **60**, 320–42.

Pierrehumbert, J.
1994 Prosodic Effects on Glottal Allophones. In O. Fujimura & M. Hirano (eds.), *Vocal Fold Physiology: voice quality control*. (pp. 39–60). San Diego: Singular Publishing Group.

Pierrehumbert, J.
2000 The phonetic grounding of phonology. *Les Cahiers de l'ICP, Bulletin de la Communication Parlé*, **5**, 7–23.

Pierrehumbert, J.
2001 a Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (eds.), *Frequency Effects and the Emergence of Linguistic Structure*. (pp. 137–157). John Benjamins, Amsterdam.

Pierrehumbert, J.
2001 b Why phonological constraints are so coarse-grained. In J. M. McQueen & A. Cutler (eds.), *SWAP* special issue, *Language and Cognitive Process*, **16**, 691–698.

Pierrehumbert, J. & Beckman, M. E.
1988 *Japanese Tone Structure*. LI Monograph, **15**. Cambridge, MA: MIT Press.

Pierrehumbert, J., Beckman, M. E. & Ladd, D. R.
2001 Conceptual Foundations of Phonology as a Laboratory Science. In Burton-Roberts, N., Carr, P. & Docherty, G. (eds.), *Phonological Knowledge*, (pp. 273–304). Oxford, UK: Oxford University Press.

Pierrehumbert, J. & S. Frisch,
1996 Synthesizing Allophonic Glottalization, J. P. H. van Santen, R. Sproat, J. Olive, & J. Hirschberg, (eds.), *Progress in Speech Synthesis*, (pp. 9–26). New York: Springer-Verlag.

Pierrehumbert, J. & D. Talkin,
1992 Lenition of /h/ and glottal stop. In G. Doherty & D. R. Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. (pp. 90–179. Cambridge, UK: Cambridge Univ. Press.

Rialland, A.
1986 Schwa et syllabes en Français. In L. Wetzels & E. Sezer (eds), *Studies in Compensatory Lengthening*. (pp. 187–226). Dordrecht: Foris Publications.

Roelofs, A.
1997 The WEAVER model of word-form encoding in speech production. *Cognition*, **64**, 249–284.

Rosenbaum, D. A., Engelbrecht, S. E., Bushe, M. M. & Loukopoulos, L. D.
1993 A model for reaching control. *Acta Psychologica*, **82**, 237–250.

Schreuder, R. & Baayen, R. H.
1995 Modeling Morphological Processing. In I. B. Felman (ed.), *Morphological Aspects of Language Production*. (pp. 131–156). Hillsdale, NJ: Lawrence Erlbaum Associates,

Shattuck-Hufnagel, S.
1979 Speech errors as evidence for a serial order mechanism in sentence production. In W. E. Cooper & E. C. T. Walker (eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. Hillsdale, NJ: Lawrence Erlbaum.

Steriade, D.
2000 Paradigm Uniformity and the phonetics-phonology interface. In M. Broe & J. Pierrehumbert (eds.), *Papers in Laboratory Phonology V: Acquisition and the Lexicon*. (pp. 313–335). Cambridge UK: Cambridge University Press.

Sternberg, S., Monsell, S., Knoll, R. L. & Wright, C. E.
1978 The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (ed.), *Information Processing in Motor Control and Learning*. New York: Academic Press.

Sternberg, S., Wright, C. E., Knoll, R. L. & Monsell, S.
1980 Motor programs in rapid speech: Additional evidence. In R. A. Cole (ed.), *Perception and Production of Fluent Speech*. Hillsdale NJ: Lawrence Erlbaum Associates.

Vitevich, M. & Luce, P.
1998 When words compete: Levels of processing in perception of spoken words. *Psychological Science*, **9**: 4, 325–329.

Wright, R.
1997 Lexical competition and reduction in speech: A preliminary report. Research on spoken language processing: Progress report **21**. Bloomington, IN: Indiana University. (Related paper also forthcoming in Papers in Laboratory Phonology VI).

Wurm, L. H.
1997 Auditory Processing of Prefixed English Words is Both Continuous and Decompositional. *Journal of Memory and Language*, **37**, 438–461.

Yaeger-Dror, M.
1996 Phonetic evidence for the evolution of lexical classes: The case of a Montreal French vowel shift. In G. Guy, C. Feagin, J. Baugh & D. Schiffrin (eds.), *Towards a Social Science of Language* (pp. 263–287). Philadelphia: Benjamins.

Yaeger-Dror, M. & Kemp, W.
1992 Lexical classes in Montreal French. *Language and Speech*, **35**, 251–293.