Exemplar dynamics: Word frequency, lenition and contrast

Janet B. Pierrehumbert
Northwestern University
Evanston, IL

June 8, 2000

Abstract

To appear in J. Bybee & P. Hopper (eds.), Frequency effects and emergent grammar. Amsterdam: John Benjamins.

Exemplar theory was first developed as a model of similarity and classification in perception. In this paper, the theory is extended to model speech production as well as speech perception. Straightforward extension of the model provides a formal framework for thinking about the quantitative predictions of usage-based phonology, as proposed by Bybee. A model is proposed which allows us to derive the finding that leniting historical changes are more advanced in frequent words than in rarer ones. Calculations using this model are presented which reveal the interaction of production noise, lenition and entrenchment. A realistic treatment is also provided for the time course of a phonological merger which originates from lenition of a marked category.

1 Introduction

Over the last decades, a considerable body of evidence has accumulated that speakers have detailed phonetic knowledge of a type which is not readily modelled using the categories and categorical rules of phonological theory. One line of evidence is systematic differences between languages in fine details of pronunciation. For example, it is known that Spanish and English differ systematically in the exact formant patterns typical of their point vowels (Bradlow 1995). Canadian French differs from both Canadian English and European French in the distribution of VOT times of voiced and voiceless stops (Caramazza and Yeni-Komshian, 1974). These are just two of many examples, with more reviewed in Pierrehumbert (in press) and Pierrehumbert et al. (in press); at this point, it is not possible to point to a single case in which analogous phonemes in two different languages display exactly the same phonetic targets and the same pattern of phonetic variation in different contexts. Exact phonetic targets and patterns of variation must accordingly be learned during the course of language acquisition. The usage-based framework readily accomodates such findings by proposing that mental representations of phonological targets and patterns are gradually built up through experience with speech.

A particularly interesting and challenging result is the discovery that learned phonetic detail may be associated not just with languages or dialects, but even with specific words in the lexicon of a given dialect. This observation is made most convincingly in a series of studies by Bybee which explore the relationship of word frequency to lenition. Bybee (Hooper 1976) explored the process of schwa reduction and desyllabification which applies variably before sonorants such as /r/ and /n/ in English. She found that in high frequency words, such as *every* and the noun *evening* , the schwa was completely absent and the syllable it originally projected had vanished. In mid-frequency words, such as *memory* and *salary* , the modal outcome is a syllabic /r/. In rare words, such as *mammary* and *artillery* , the modal outcome is a schwa plus /r/. Another example is provided by so-called t/d-deletion, which is widely acknowledged to be a case of variable undershoot of the coronal articulation of the /t/ or /d/. Bybee (2000) found that deletion – defined as the inability of the transcriber to hear the stop on a tape-recording – is more prevalent in high-frequency words than in low-frequency words. The set of double-marked past tense verbs (such as *told* and *left*) provides a way to control for the morphological factors which could play a part in this pattern. Within the set of double-marked pasts alone, Bybee's data showed a statistically significant relationship of word frequency to the rate of /t/ deletion, with the most frequent word (*told*) having /d/ deleted in 68% of cases while the least frequent (*meant*) never had the /t/ deleted. Further documentation of the association between word frequency and leniting historical change is provided in Phillips (1984, this volume).

Although these frequency effects will be the main focus in this paper, is is also important to acknowledge that word-specific allophony has been found in a number of other situations as well. For example, Yaeger-Dror and Kemp (1992) and Yaeger-Dror (1996) demonstrate that words in a particular cultural/semantic field in Montreal French have resisted a historical shift in the vowel system and as a result display idiosyncratic vowel quality. Hay (2000) also presents data relating degree of morphological decomposibility to degree of /t/ lenition in words such as "shiftless".

These results challenge standard models of phonology and phonetics at two levels. First, in all standard models, the lexicon is distinguished from the phonological grammar. The exact phonetic details of a word's pronunciation arise because the word is retrieved from the lexicon, and processed by the rules or constraints of the grammar whose result (the surface phonological form of the word) is fed to a phonetic implementation component. The phonetic implementation component computes the articulatory and/or acoustic goals which actualize the word as speech. The phonetic implementation component applies in exactly the same way to all surface phonological representations, and the outcome depends solely on the categories and prosodic structures displayed in those representations. As a result, there is no way in which the phonetic implementation can apply differently to some words than to others. If a phonetic implementation rule is variable and gradient, then the same probability distribution of outcomes would arise for all words which meet the structural description of the rule. This generic feature of modular generative models with phonetic implementation rules is developed at more length in Pierrehumbert (1994).

A second challenge arises from the fact that the differential phonetic outcomes relate specifically to word frequency. Standard generative models do not encode word frequency. They treat the word frequency effects which are so pervasive in experiments involving priming or lexical decision tasks as matters of linguistic performance rather than linguistic competence. Thus

the intrusion of word frequency into a traditional area of linguistics, namely the conditioning of allophony, is not readily accommodated in the classical generative viewpoint.

If each word corresponded to a completely idiosyncratic phonetic signal, then results such as Bybee's could be readily formalized in a highly transparent scientific model. We would simply assume that holistic gestural or acoustic templates are associated with word meanings. The real challenge arises from the fact that the classical view does provide important insights about the mental representation of phonology. Although a word may have idiosyncratic phonetic properties, it is perceived as made up of units of sound structure which are also shared with other words. The existence of these subparts – whether phonemes, moras, or syllables – is reflected in productive behaviors such as pronunciation of neologisms and loan word assimilations. It is also reflected in the tendency of historical changes to sweep through the vocabulary. Thus, the correct model must describe the interaction of word-specific phonetic detail with more general principles of phonological structure.

In this paper, we will develop a formal architecture which is capable of capturing these regularities. This formal architecture is "generative" in the sense that it provides explicitly for phonological representations and processes; it predicts that some outcomes are possible and others are not. Like a generative grammar, it is informed by the goal of specifying all and only the outcomes which are possible in human language. It represents a considerable departure from generative models, however, in the way the lexical representations are organized and the consequences of lexical representation for speech production. Specifically, the model assumes that detailed phonetic memories are associated with individual words and it implicitly defines word specific probability distributions over phonetic outcomes. Whereas the classic models define a strong separation between the lexicon and the grammar, in the present model these represent two degrees of generalization over the same memories and are thus strongly related to each other. Furthermore, in the present model, frequency information plays an intrinsic role in the system because it is implicitly encoded by the very nature of the memory system. These general properties of the model all originate from the psychological model of memory and classification from which the proposal derives, namely exemplar theory. From its origins as a model of perception and classification only, it is extended to be a model of perception, production, and the consequences of the perception-production loop over time.

2 Exemplar Theory

Exemplar theory was first introduced in psychology as a model of perception and categorization. It was subsequently extended specifically to speech sounds by Johnson (1996) and Lacerda (in press), providing a highly successful model of vowel categorization in particular. Goldinger (1996) also applies the strongly related model of Hintzman (1986) to model the identification and recognition of words. I will adopt some key assumptions from this previous work, indicating briefly the empirical motivation for these assumptions.

In an exemplar model, each category is represented in memory by a large cloud of remembered tokens of that category. These memories are organized in a cognitive map, so that memories of highly similar instances are close to each other and memories of dissimilar instances are far apart. The remembered tokens display the range of variation that is exhibited in the physical manifestations of the category. For example, the remembered tokens of the

vowel /ϵ/ would exhibit a variety of formant values (related to variation in vocal tract anatomy across speakers, variation along the dimension of hypo-hyperarticulation, and so forth) as well as variation in f0 and in duration. The entire system is then a mapping between points in a phonetic parameter space and the labels of the categorization system. The labels constitute a level of representation in their own right, or else they may be viewed as functional links to other levels of representation.

It is important to note that the same remembered tokens may be simultaneously subject to more than one categorization scheme, under such a model. For example, a recollection of the phrase *Supper's ready!* could be labelled as "Mom" and "female speech", in addition to exemplifying the words and phonemes in the phrase.

If every encountered token of a category is stored as a separate exemplar, then frequent categories will obviously be represented by numerous tokens and infrequent categories will be represented by less numerous tokens. The difference in token count is one ingredient of the model's explanations of frequency effects, as we will see below. The mind's capacity for long-term memories of individual examples is in fact astonishingly large, as experiments reviewed in Johnson (1996) indicate. Nonetheless, the volume of speech which a person processes in a lifetime is so great that we would not wish to assume individual memories of every use of every word.

Exemplar theory responds to this problem in two ways. First of all, we assume that memories decay. Memories of utterances that we heard yesterday are more vivid than memories from a decade ago. Second, the parameter space in which the exemplars are represented is assumed to be granularized. Examples whose differences are too fine to show up under the granularization are encoded as identical (see Kruschke, 1992). For example, the ear cannot distinguish arbitrarily fine differences in f0. The JND (just noticeable difference) for f0 in any given part of the range is determined by the resolution of the anatomical and neural mechanisms which are involved in encoding f0. Thus, it is reasonable to suppose that speech tokens differing by less than one JND in f0 are stored as if they had identical f0s. Similar constraints on the resolution of all other perceptual dimensions would motivate granularization of the phonetic parameter space as a whole. As a result, an individual exemplar – which is a detailed perceptual memory – does not correspond to a single perceptual experience, but rather to an equivalence class of perceptual experiences.

This said, it becomes reasonable to propose that each exemplar has an associated strength – which may be viewed as a resting activation level. The exemplars encoding frequent recent experiences have higher resting activation levels than exemplars encoding infrequent and temporally remote experiences.

When a new token is encountered, it is classified in exemplar theory according to its similarity to the exemplars already stored. Perceptual encoding of the new token locates it in the relevant parameter space. Its similarity to any single stored exemplar can be computed as its distance from the exemplar in the parameter space. To classify the new token, the most probable labelling given the labelling of the exemplars in the neighborhood is computed. The model implemented here follows the specifics of Lacerda (in press). A fixed size neighborhood around the new token determines the set of exemplars which influence the classification. The summed similarities to the exemplars for each label instantiated in that neighborhood is computed, with the similarity to each given exemplar weighted by the strength (or activation) of that exemplar. Recall that the strength is a function of the number and recency of phonetic

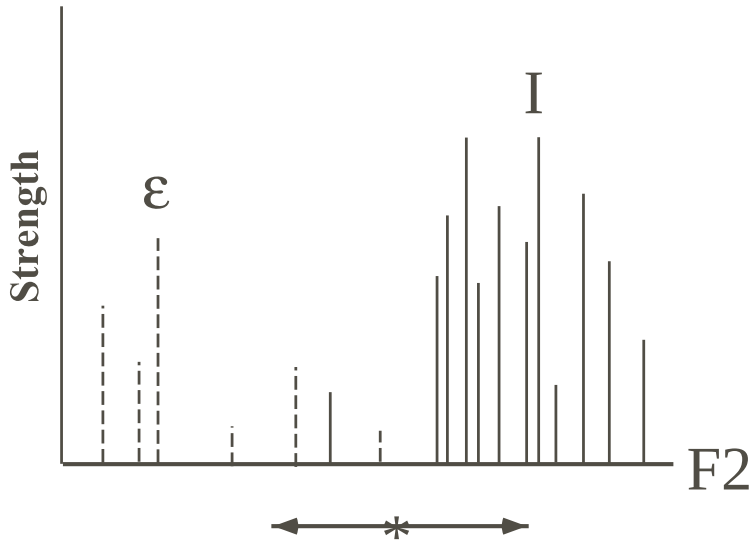


Figure 1:

tokens at that location in the exemplar space.

Figure 1 illustrates the operation of the choice rule for a hypothetical case in which the labels /I/ and /ϵ/ are being considered for an unknown vowel token. For the sake of exposition only, I assume that the only relevant dimension is f2 (the value of the second formant); this is the x-axis. In a realistic situation, the input would of course have higher dimensionality. The y axis is the activation level for each of the stored exemplars. Exemplars of /ϵ/ are shown with dashed lines towards the left, and exemplars of /I/ with solid lines towards the right, a consequence of the fact that the vowel /I/ generally exhibits higher f2 than /ϵ/. However, a few individual tokens of /ϵ/ have a higher f2 than a few tokens of /I/. This overlap of the phonetic distributions for /ϵ/ and /I/ really does arise in practice, because of dialect differences, speaker differences, and random variation in production. The unidentified vowel has an f2 which places it in a region of ambiguity, as shown by the location of the star under the x-axis. The window in which the comparison is being made is shown by arrows. Within this window, there are seven exemplars of /I/, of which six are highly activated. There are only two (less activated) exemplars of /ϵ/. Hence, the winning label is /I/. The equation specifying this classification rule is given in the appendix.

In other approaches (e.g. Kruschke, 1992), all exemplars with all labels contribute to the classification, but an exponentially decaying weighting by distance has the result that the exemplars nearest to the stimulus dominate the calculation. As a result, the overall behavior of the model is substantially similar to that of the model reported here. We note also that attentional weights may be imposed to model how different contexts, expectations, and task requirements influence classification; however these effects are not at issue in the present paper.

Note that the labelling depends on the relationship amongst the exemplar clouds in the neighborhood; the winning label is the one which is overall more probable than the competitors. A label which has more numerous or more activated exemplars in the neighborhood of the new

token has an advantage in the competition. Given that high frequency labels are associated with more numerous exemplars (whose resting activations are, on the average, higher), they will have more dense and more activated exemplar clouds. In situations involving ambiguity, the model thus predicts a bias towards a high-frequency label. This prediction is supported by the experimental literature.

The classification rules just discussed have no temporal scale, summarizing only the end result of the decision process. Of course this does not mean that the brain has for each perceptual classification process a separate little pocket calculator, which it employs to compute the values of the relevant formuli over the relevant exemplar clouds. Instead, the decision rules may be viewed as representing synoptically the behavior of an activation/inhibition system. The sums of the exemplar strengths represent the fact that exemplars spread activation to labels, so that the activation of any given label is a cumulative function of the number and activation level of the exemplars associated with it. The comparison amongst the scores for different labels reflects the results of reciprocal inhibition amongst labels, with the winning label being the one which succeeds in suppressing the activation of its competitors. The model is consistent with the standard assumption that reaction times for phonological and lexical decisions reflect the time required for activation to build up and cross a decision threshhold. Thus, the model is consistent with, and can even serve to elucidate, results on the speed of phonological and lexical decisions.

To summarize, the exemplar approach associates with each category of the system a cloud of detailed perceptual memories. The memories are granularized as a function of the acuity of the perceptual system (and possibly as a function of additional factors). Frequency is not overtly encoded in the model. Instead, it is intrinsic to the cognitive representations for the categories. More frequent categories have more exemplars and more highly activated exemplars than less frequent categories.

Let us now review the most obvious successes of this approach, as it applies to speech, before passing on to extensions of the model.

Exemplar theory provides us with a way to formalize the detailed phonetic knowledge that native speakers have about the categories of their language. Since exemplar theory stores directly the distribution of phonetic parameter values associated with each label, it provides us with a picture of the "implicit phonetic knowledge of the speaker". The acquisition of this knowledge can be understood simply in terms of the acquisition of a large number of memory traces of experiences. There is no competing model which achieves the same level of descriptive adequacy. Notably, the assumption that there exists a universal symbolic alphabet which provides an interface to a universal sensori-motor phonetic implementation component (as in Chomsky and Halle, 1968; Chomsky and Lasnik, 1995) provides no means of representing the extremely fine differences across languages in values and probability distributions of phonetic properties. Therefore, it yields no insight into how the knowledge of such details might be acquired.

Another obvious success of the model is its treatment of prototype effects, handling with a single mechanism two major findings. One is the finding that a new token which is well-positioned with respect to a category can actually provide a better example of that category (in being recognized quickly and rated highly) than any actual example of that category that has been previously experienced. This phenomenon, sometimes taken as an argument for the abstraction of prototypes, follows from the exemplar model if "goodness" is interpreted in

terms of the probability of the winning label (with the probability arising from the relative score in relation to the scores of competitors). This probability does not necessarily reach a maximum on a position in the parameter space which is actually occupied by an exemplar; a position which is centrally positioned in a dense neighborhood of exemplars will receive a very high probability even if there is no exemplar at that exact point. Thus the abstract prototype need not be explicitly computed and stored in advance. A second success of the model, as noted by Lacerda, is its the ability to explain the fact that extreme examples of phonological categories are sometimes judged to be better than modal examples. For example, as shown in Johnson, Flemming, and Wright (1993) the perceptually best examples of the corner vowels /i/ and /u/ have more extreme formant values than typical productions. This outcome follows from the fact that the probability for a label is influenced both by the activation of exemplars having that label, and by competition from other labels having exemplars in the same area of the cognitive map. Increasing the distance of a novel token from all exemplars with competing labels will thus raise the subjective goodness.

A last strength of exemplar models is that they provide a foundation for modelling frequency effects, since frequency is built in to the very mechanism by which memories of categories are stored and new examples are classified. It is not necessary to posit special frequency counters whose cognitive and neural status are dubious. Indeed, exemplar models can be fleshed out with assumptions about neural encoding so as to capture the main experimental findings about frequency effects, including an understanding of why frequency affects both the outcome of decisions and the speed with which decisions are taken.

3 Production

3.1 Model 1

As is evident from the last section, exemplar models were developed to model perceptual data. Real language use in communication involves both perception and production. In this section, we undertake an extension of the model in order to handle production. By modelling the complete perception-production loop using exemplar theory, we will show that facts about the reflexes of word frequency in production which were discovered by Bybee and Phillips can be modelled. No other current theoretical approach can handle these facts.

In perception, the encoded phonetic character of an incoming stimulus locates it in the parameter space. Activation of exemplars in the neighborhood is passed upwards to the labels, with the most probable label winning in competition with alternatives. Production proceeds in the opposite direction. Following Levelt (1989) and others, assume that the decision to produce a given category is realized through activation of that label. The selection of a phonetic target, given the label, may be modelled as a random selection of an exemplar from the cloud of exemplars associated with the label. It will not be important here whether the exemplars have a dual acoustic-motor nature, or whether the motor program is computed on the fly in order to match the acoustic goals represented by the exemplar. Similarly, we will not attempt to model the deeper causes which may figure in the choice amongst possible exemplars. Although social and stylistic factors may select for different parts of the exemplar cloud in different situations, the aggregate behavior of the system over all situations may be modelled as a repeated random

sampling from the entire aggregate of exemplars. The likelihood that a particular exemplar will be selected is proportionate to its strength. Production is taken to be sensitive to strength in exactly the same way that perceptual classification is. Thus, this first model of production is a minimal extension of previous work on how exemplars function in perception.

Now, a phonetic target is not necessarily achieved exactly. Even for a speaker who is merely talking to himself, one may assume random deviations from the phonetic target due to noise in the motor control and execution. For a community of multiple speakers, there would be random differences amongst the stored memories of different members of the community. Thus if a listener hears a speech token produced by a different speaker than himself, that speech token could be randomly different from the exemplars in his own stored memories. In sum, new tokens being added to an existing exemplar cloud may be viewed (to a first order approximation) as a random sampling from that cloud with added noise.

Figure 2 shows the consequences of this simple approach for the evolution of a single category from a single token to a distribution of exemplars. As in Figure 1, the situation is simplified to one phonetic dimension for expository purposes. The x-axis of the figure represents a relevant phonetic parameter, such as second formant value (if we are considering categories of vowel frontness), or f0 (if we are considering tonal categories). A nominal scale is indicated. The single token of the category which seeded the cloud is located at x = 1. That is, the very first speech token which the listener associates with the category label in question displays a phonetic value of 1, and this value serves as the starting point for the development of the new category. (We have said nothing about **why** a listener may posit a new category, as this question involves functional issues which exceed the scope of the paper). The production noise is unbiased with a uniform distribution of width 0.2. The y-axis is the count of memory-weighted count exemplars in each small interval of the phonetic scale. The e-folding time of a memory is 2000 time steps (e.g the parameter controlling the exponential decay of memories is 2000 production/perception iterations. See appendix for further details). Three superimposed curves show the situation after 10,000, 50,000, and 100,000 iterations. Thus, the figure is essentially like three superimposed histograms, except that the area under each curve is not normalized to 1.0 as a probability would be. As discussed above, the total representation of the category is strengthened as more and more memories are stored; temporal decay of older memories, not normalization, is responsible for the gradual lowering of the peak in the figure.

Figure 2 is based on the idealization by which every single production is accurately classified as a member of the category. Note that the variance of the distribution along the phonetic dimension displayed increases with usage. It is important to model this increase in variance, since mature categories do display variation. (They do not have spike-like distributions showing only phonetic properties which correspond exactly to the first token of the category which is internalized by the listener.) The overall shape approaches a Gaussian distribution as the number of tokens increases. This limiting behavior arises from the fact that the production-perception loop is an additive random process.

3.2 Model II: Systematic Bias

Figure 2 showed the case where there is no systematic bias in production. Recent work by Lindblom and colleagues on hypo- and hyper- articulation (Lindblom, 1984) argues for systematic

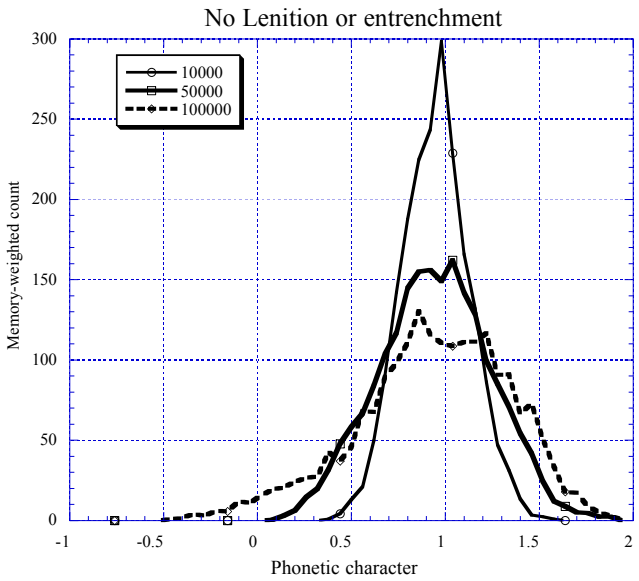


Figure 2:

production biases. The case which will interest us here is hypo-articulation, or the tendency to undershoot articulatory targets in order to save effort and speed up communication. This tendency is arguably the cause of leniting historical changes, such as schwa reduction and /t/-deletion. Of course, in a complete model of historical change it will be necessary to offer some explanation of why certain languages at certain times begin to permit particular leniting changes while not permitting others. But given that a historical leniting change is in progress, its phonetic consequences may be represented as a systematic bias on the production process in the model we are developing here.

Figure 3 presents results of a calculation identical to Figure 2, except that a systematic bias has been introduced in the production process. The bias applied is -0.01, or leftwards along the phonetic scale which serves as the x-axis. This means that each token is produced slightly lenited compared to the exemplar of the category which has been randomly selected as a production goal. No matter how lenited the production goal may be, the production is that little bit more lenited. This is one concrete interpretation of Lindblom's general observations. Lindblom is claiming that speakers undershoot targets to the extent possible – e.g. to an extent that still permits communication. It would not be consistent with Lindblom's general line of thought to think that speakers underarticulate to the point that their target words become unrecoverable. As before, the distributions shown represent the results of 10,000, 50,000 and 100,000 interations of the model. By comparing Figure 3 to Figure 2, we see that a systematic lenition bias causes the distribution of exemplars to shift. In addition, it causes an increase in variance, much as a photograph of a moving object shows a blur.

One way to view this figure is diachronically. It shows how the distribution of a category evolves over time after a leniting historical change is first introduced. The mode of the distribution gradually moves towards the left (or lenited) end of the phonetic axis. The graph also has a synchronic interpretation, provided that we add a key assumption – namely, that

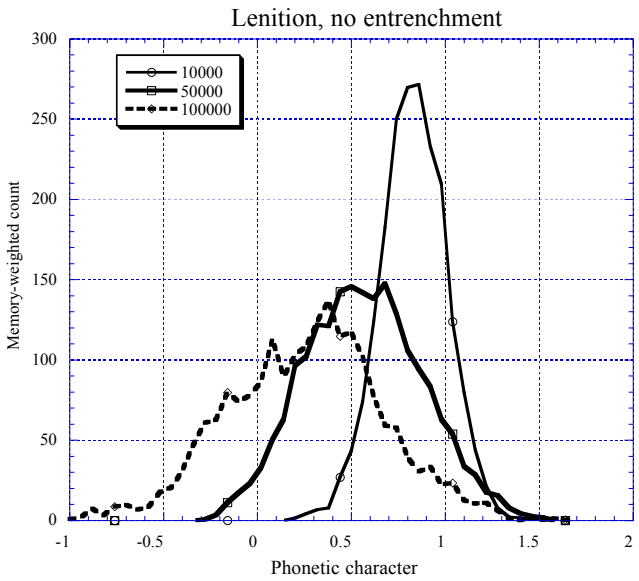


Figure 3:

not just phonemes, but individual words, have associated exemplar clouds. For example, we assume that each of the words *bet*, *bed* , and *bend* has an exemplar cloud, and that the exemplar cloud for the phoneme /ϵ/ is the union of the /ϵ/ sections of the exemplar clouds for these words and for all other words containing an /ϵ/. With this added assumption, the figure may be viewed as displaying a synchronic comparison amongst words of different frequencies which are impacted by the same historical change in progress. Since the high frequency words are used more often than the low frequency words, their stored exemplar representations show more numerous impacts of the persistent bias towards lenition. As a result, they are further to the left on the axis than the low frequency words.

The result displayed in Figure 3 is exactly the result documented by Bybee, Philips, and others. Some detailed predictions of the model include: 1) Each individual word displays a certain amount of variability in production. 2) The effect of word frequency on lenition rates is gradient. 3) The effect of word frequency on lenition rates should be observable within the speech of individuals; it is not an artifact of averaging data across the different generations which make up a speech community. 4) The effect of word frequency on lenition rates should be observable both synchronically (by comparing the pronunciation of words of different frequency) and diachronically (by examining the evolution of word pronunciations over the years within each person's speech.) The exemplar model is the only current model which has these properties. An additional prediction is that probability distributions for words undergoing a historical change should be skewed, with the extent of the skew being slight or great according to the velocity of the change. Even with recent advances in speech processing technology, it would require an extremely ambitious amount of data analysis to evaluate this prediction.

Two further observations may be made on the cognitive interpretation of this model. First, note that speakers immersed in a new speech environment find that their pronunciation pat-

terns shift over a relatively long time span, of several months or more. (For example, see the longitudinal phonetic study reported in Sancier and Fowler, 1997). The time span for historical changes is on the order of decades or more. Thus, the extremely high number of iterations used in making the calculations in the figures is not unrealistic. Consider, for example, a leniting change affecting the vowel in the preposition "of". The present paper alone has over 200 examples of this word, and 10,000 examples would probably occur in less than one month of speech. Second, it is often noted that historical changes impact the speech of older people less than younger people, so that a change in progress results in a divergence between the speech patterns of different generations. The model suggests two possible factors in this finding. First, older people may have more exemplars than younger ones for the same pattern, so that the parameter values displayed in older exemplars dominate the production statistics. This line of explanation depends on the assumption that memories decay slowly. A second possibility is that older people are less likely to add new exemplars than young ones; because the formation of new memories becomes less rapid and robust with age, the production statistics are dominated by exemplars stored at a younger age. Differences in attention or in feelings of social affiliation could impact formation of exemplar memories in an analogous way. Both of these lines of explanation predict that the speech patterns of older adults could shift to some extent, just not as rapidly as for younger people.

3.3 Model III: Entrenchment

Figure 3 has a serious problem which is already foreshadowed in Figure 2. In a model with production noise, the variance for any given category steadily increases with usage; when there is a systematic production bias, the velocity the bias imparts to this distribution aggravates the spread. However, practice is often reported to have the opposite effect of decreasing the variance, a phenomenon known as "entrenchment". For example, a child who takes up the cello produces highly variable tuning of notes at the beginning, and more and more accurate tuning over years of practice. The phonetic variability associated with a typical phonological category decreases gradually up through late childhood (Lee et al. 1999). The bare exemplar model provides no way to model entrenchment. There is no combination of parameter settings for the model which allows a category to fill out after being seeded by a single example, without simultaneously predicting that the spreading out will go on indefinitely.

The model must be further elaborated in order to model entrenchment effects. The model of entrenchment for which we present calculations is broadly inspired by work by Rosenbaum et al. (1993) on reaching movements. The understanding of production is modified so that production does not depend only on a single target exemplar (selected at random). Instead, a target location in the exemplar cloud is selected at random, and the exemplars in the neighborhood of this location all contribute to the production plan, to a degree which reflects their activation level. The neural interpretation of this proposal is that a region in the brain, not merely a single point, is activated when planning a production. Activation-weighted averaging over a group of exemplars results in entrenchment, because averaging mathematically causes reversion towards the mean of a distribution.

Calculations of a leniting change in progress which include this treatment of entrenchment are displayed in Figure 4. A neighborhood of 500 exemplars is used in calculating the distributions displayed in this figure. A comparison of Figure 3 and Figure 4 shows that the

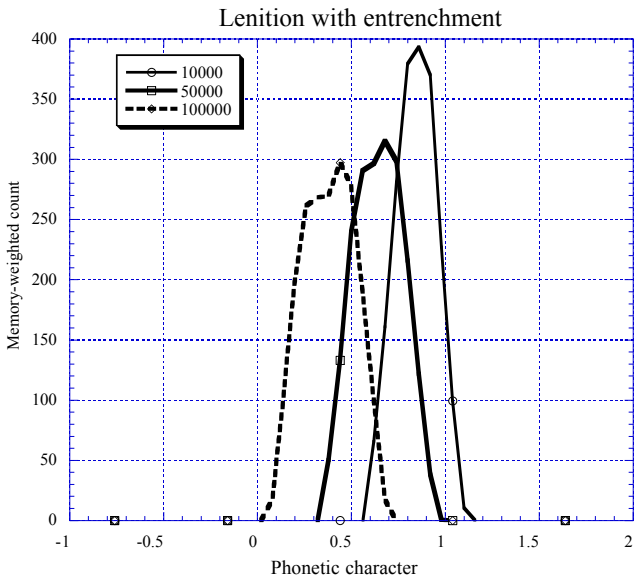


Figure 4:

entrenchment narrows the distributions, so that the distribution width for the case of 100,000 iterations is roughly comparable to that for 10,000 iterations. With the particular parameter settings selected here, the spreading effects arising from production noise and lenition and the anti-diffusive effect of entrenchment have essentially cancelled out in determining the variance. If a larger neighborhood were used in the treatment of entrenchment, then the high count case would have less variance. In a situation involving high production noise or a high degree of systematic bias, the high count case would display more variance than the low count case.

The issue of entrenchment is a complicated one, and the treatment we have presented is only one of many possible ones. The Hintzman/Goldinger model proposes an entrenchment effect on the perception side rather than the production side; when any given stimulus is classified, it sets up an "echo" which reflects not only its own properties but also the properties of the exemplars in the stimulus neighborhood which contributed to the classification. The echo is what is stored in memory, not the stimulus itself. Since the echo combines information over a neighborhood, it shows reversion towards the mean just as our production model does. The type of data we are considering here – patterns of historical change – involve the entire perception-production loop and they do not tell us whether entrenchment occurs in perception, in production, or in both.

In the Hintzman/Goldinger model, the neighborhood which influences the echo of a stimulus has a fixed size on the exemplar map. This means that there are few influences on the echo if the neighborhood of the stimulus is sparsely populated. In the production model presented here, the neighborhood contains a fixed number of exemplars; it expands its size in regions which are sparsely populated. The Hintzman/Goldinger treatment has the result that more episodic information is encoded in memory for rare events than for frequent ones; for example, one is more likely to remember that a word was spoken in a particular voice if the word is rare than if it is common. However, we were unable to make a fixed neighborhood work out

in the production model since it creates too much instability in the exemplar dynamics at the beginning of the calculation when there are very few examples of a category. This is why an n-nearest-neighbors model is offered here. An integrated model which handles all known neighborhood effects simultaneously remains to be developed.

A third issue is whether entrenchment critically involves feedback from other levels, and if so, what kind of feedback. Notice that self-organizing systems can in principle form and sharpen peaky distributions without any type of feedback at all, much as the lumpiness in the energy distribution after the Big Bang eventually evolved into the universe we know with concentrated physical objects, such as galaxies and viruses. All that is needed is some type of anti-diffusive factor, such as gravity, which causes unevenness in the parameter distributions to become exagerated. Equally, however, people sharpen categories faster and to a greater degree if they receive feedback, particularly if the feedback provides functionally important rewards or penalties. Speech patterns appear to fall into an intermediate situation, in that people adapt their speech patterns to their speech community even without overt pressures and rewards, but that communicative success and social attunement provide implicit feedback which is certainly important. The model presented here does have feedback, in that it has an informational loop between the stimulus encoding and the abstract level of representation represented by the labelling. If an incoming stimulus is so ambiguous that it can't be labelled, then it is ignored rather than stored. That is, the exemplar cloud is only updated when the communication was successful to the extent that the speech signal was analyzable (As in real life, there is no guarantee that the listener's analysis is the speaker's, however.) In addition, the model automatically generates social accommodation of speech patterns, since speech patterns which are heard recently and frequently dominate the set of exemplars for any given label, and therefore guide the typical productions. This effect arises from the feedback loop from production to classification to production which is set up by the "speech chain" of conversational interaction. To model the more specific feedback effects which occur in different social contexts, it is necessary to introduce attentional weighting as a further factor. For example, if a child emulates the speech patterns of a particularly admired role model, this would be modelled by weighting of the exemplars in that particular voice. This weighting represents the net positive effect of feedback from the other levels of representation involved in the child's understanding of his social situation.

4 Neutralization

In the calculations presented so far, it has been assumed that every single production of a label is accurately classified as an example of that category. Under this assumption, a leniting change causes an unbounded drift in the phonetic distribution for each word exemplifying a category. In fact, however, historical changes have natural completion states. When the change is complete, the new situation is stable.

To model this situation, we need to look at two labels which are competing over a phonetic parameter range. We consider the case of a marked phonological category competing with an unmarked one. Following Greenberg and others, we take the unmarked category to be more frequent than the marked one (see papers in Greenberg et al. 1978). In the calculation presented, the unmarked category is three times as frequent as the marked one. The marked

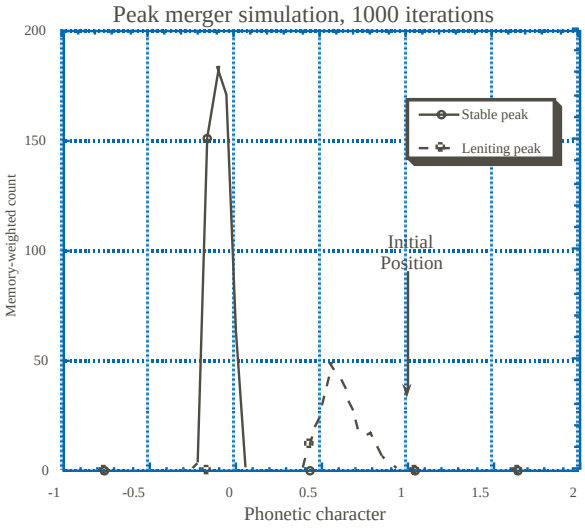
category is also the phonetically unstable one which is subject to a persistent bias. The unmarked one is assumed to be phonetically stable. An example of this situation would be the collapse of a phrase-final voicing contrast. Phrase-final voiced obstruents are typically less frequent than voiceless ones. Lack of articulatory effort results in poor voicing in final position, e.g in tokens which are subject to being misperceived as voiceless. Historically, voiced and voiceless obstruents are reported to collapse to the unvoiced category in this position.

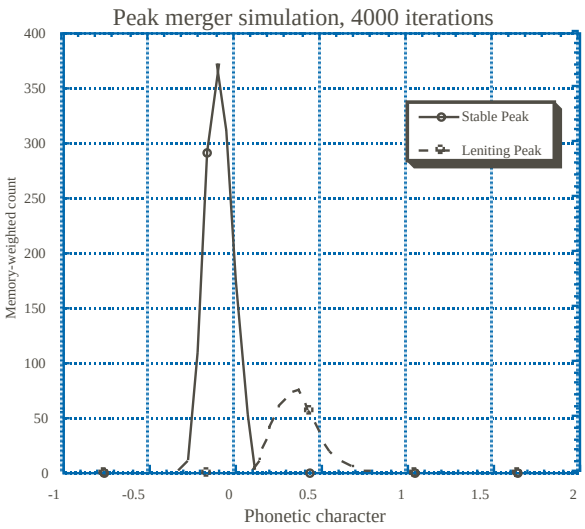
In Figure 5, the right hand distribution represents the marked category which is subject to a persistent leftwards bias. The left hand distributrion is a stable unmarked distribution competing for labelling of the same phonetic parameter. The successive panels represent four time slices in the evolution of the situation. Because the marked distribution is subject to a persistent bias, it drifts to the left. When it approaches the unmarked distribution, some individual tokens which were intended as examples of the marked case are perceived and stored as examples of the unmarked case. This happens more often than the reverse. Insofar as it does happen, the disproportion in frequency between the two categories increases. In the end, the marked category is completely gobbled up by the unmarked one. Note that the distribution of the unmarked category does show some influence of the marked category it absorbed. Although the location of the distribution is still closer to the original location of the unmarked category than that of the marked category, the mode of the distribution is a bit to the right from where it was. This is not necessarily unrealistic. One could imagine a situation in which the distinction between final voiceless aspirated stops and final voiced stops is neutralized to final voiceless unaspirated stops. To evaluate this general type of prediction, detailed statistical distributions of parameter values for changes in progress will need to be collected. Modelling such distributions will require serious consideration of the relationship between the phonetic scales which are readily susceptible to measurement and the scale of effort on which the persistent leniting bias is presumed to be defined. The physics of speech production exhibits many nonlinearities, including ceiling and floor effects, and these will shape the assymptotic behavior of the system in a way which circumscribes the possibilities for stable outcomes.

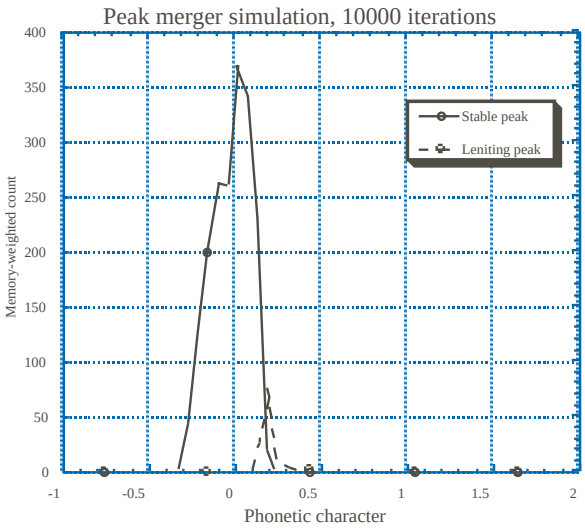
5 Conclusion

In conclusion, exemplar dynamics provides an incisive model of the main findings of usage-based phonology. The assumption that people learn phonological categories by remembering many labelled tokens of these categories explains the ability to learn fine phonetic patterns of a language. It also explains why patterns are incrementally modified over long periods of time in adult speech, and why leniting historical changes are typically more advanced for high-frequency words than for low frequency words. A realistic treatment of the neutralization which results when a marked category collides with an unmarked category is also provided.

Model calculations using exemplar theory yield a number of predictions whose validation provides an area for future research. Documentation of the variance as well as the means of phonetic distributions is critical to a full understanding of entrenchment. Similarly, the documentation of mergers-in-progress is also signalled as an important topic.


Peak merger simulation, 1000 iterations
Stable peak
Leniting peak
Initial
Position
Memory-weighted count
Phonetic character


Peak merger simulation, 4000 iterations
Stable Peak
Leniting Peak
Memory-weighted count
Phonetic character


Peak merger simulation, 10000 iterations
Stable peak
Leniting peak
Memory-weighted count
Phonetic character

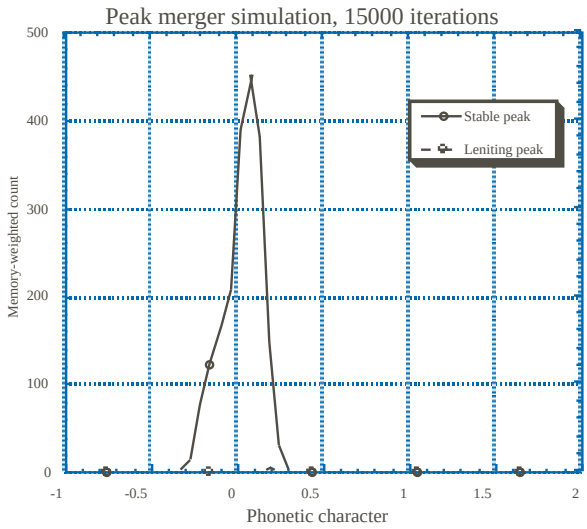

Peak merger simulation, 15000 iterations
Stable peak
Leniting peak
Memory-weighted count
Phonetic character

Figure 5:

Appendix: Model Description

The fundamental entity in our model is the exemplar list $E(L)$, which consists of the list of exemplars $\{e_1^L, ..., e_n^L\}$ 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

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

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. Currently we are using a square window function, with $W = 1$ if its argument has absolute value below .05, and $W = 0$ otherwise. If, for example there are two labels A and B in contention, we compute $score(A, x)$ and $score(B, x)$ and assign x to the label with the greatest score. In the case of a tie, the utterance is discarded. In the case of the successful classification, x is put at the head of the exemplar list corresponding to its label.

The exemplar list is also used in the production step. First a production target x_{target} is obtained by picking an exemplar randomly from the exemplar list of the desired label. In picking an exemplar, we assign each exemplar a probability which decays according to its age, specifically $exp(-(t-T_i)/\tau)$. This implements memory decay in the production process, as old exemplars are only rarely used. Without entrenchment, the token produced is then obtained by adding a performance noise and a lenition bias to the target Thus

$$x = x_{target} + \epsilon + \lambda \tag{2}$$

where ϵ is a random number chosen from a uniform distribution ranging from $-.1$ to .1 and λ is a constant lenition bias. In the one-peak cases shown in Figure 3 and Figure 4, we used $\lambda = -.01$. In the neutralization case shown in Figure 5, we used $\lambda = -.1$ for the (infrequent) leniting peak. Note that the noise and the lenition bias is applied once per utterance, so that infrequent utterances evolve on a slower time base. An additional effect, however, is that if the memory time τ is held fixed for all labels, infrequent labels access an effectively smaller portion of the exemplar list in production and classification, owing to a greater impact of memory decay. In all calculations reported above, we used a fixed memory time $\tau = 2000$ for both production and classification.

To implement entrenchment, the production target was modified as follows, prior to addition of noise and bias. We picked the n_{trench} closest exemplars to the trial x_{target}, using the memory-weighted distance

$$d_i = |x_{target} - e_i^L| exp(\frac{t - T_i}{\tau}) \tag{3}$$

and then formed a new target by taking the memory-weighted mean of the 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 simulations reported above were carried out with $n_{trench} = 500$

In the case of a single label, the production-iteration loop proceeds as follows. First, we seed the examplar list with a single value. Subsequently, we alternate between producing a new token according the protocol described above, and adding the new token to the exemplar list *provided its score is nonzero.* In the case of two labels A and B, we seed each exemplar list

with a single value, then randomly produce a token x of A or B with probability p and $1-p$ respectively, compute $score(A, x)$ and $score(B, x)$ and finally append x to the exemplar list of the higher scoring label. This procedure generalizes in the obvious way to arbitrary numbers of labels.

Acknowledgments

I'm grateful to Bjorn Lindblom, Stefan Frisch, and Gary Dell for their useful comments during the development of this model. I'm grateful for the stimulating discussion at the symposium from which the present volume results, and I'd also like to thank audiences at University of Arizona, University of North Carolina, and Northwestern University for their feedback after subsequent presentations of the work.

References

Bradlow, A. 1995. "A comparative acoustic study of English and Spanish vowels". *Journal of the Acoustical Society of America.* 97(3), 1916-1924.

Bybee, J. 2000. "The phonology of the lexicon; evidence from lexical diffusion". In M. Barlow and S. Kemmerer (eds.) *Usage-Based Models of Language.* Stanford: CSLI. (See also Hooper).

Caramazza, A. and Yeni-Komshian, G. H. 1974. "Voice onset time in two French dialects". *Journal of Phonetics* 2, 239- 245.

Chomsky, N. and Halle, M. 1968. *The Sound Pattern of English.* New York: Harper and Row.

Chomsky, N. and Lasnik, H. 1995. "The Theory of Principles and Parameters". In Chomsky, N. (ed.) *The Minimalist Program.* Cambridge, MA: MIT Press, 13-128.

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.

Greenberg, J. H., Ferguson, C. A. and Moravscik, E. A. (eds.) 1978. *Universals of human language.* Stanford University Press, Stanford CA.

Hay, J. B. 2000. "Causes and Consequences of Word Structure". Ph.D dissertation, Northwestern University.

Hintzman, D. L. 1986. ""Schema abstraction" in a multiple-trace memory model". *Psychological Review* 93, 328-338.

Hooper, J. D. 1976. "Word frequency in lexical diffusion and the source of morphophonological change". in Christie, W. (ed.) *Current Progress in Historical Linguistics.* Amsterdam:

NorthHolland 96-105.

Johnson, K. 1996. Speech perception without speaker normalization. in Johnson, K. and Mullennix (eds.) *Talker Variability in Speech Processing.* San Diego. Academic Press.

Johnson, K., Flemming, E. and Wright, R. 1993. "The hyperspace effect: phonetic targets are hyperarticulated". *Language* 69, 505-528.

Kruschke, J.K. 1992 " ALCOVE: An exemplar-based connectionist model of category learning". *Psych. Review* 99, 22-44.

Lacerda, F. In press. "Distributed memory representations generate the perceptual-magnet effect". *Journal of the Acoustical Society of America*.

Lee, S., Potamianos, A. and Narayan, S. 1999. "Acoustics of children's speech; developmental changes of temporal and spectral paramaters". *Journal of the Acoustical Society of America* 105, 1455-1468.
Levelt, W. J. M. (1989) *Speaking.* Cambridge MA: MIT Press.

Lindblom, B. 1984. "Economy of speech gestures". In MacNeilage, P. (ed) *The Production of Speech.* 217-245

Pierrehumbert, J. 1994. "Knowledge of variation". *Papers from the parasession on variation*, 30th meeting of the Chicago Linguistic Society. Chicago: Chicago Linguistic Society, 232-256.

Pierrehumbert, J. In press. "What people know about sounds of language". *Studies in the Linguistic Sciences* 29(2).

Pierrehumbert, J., Beckman, M. E. and Ladd, D.R. In press. "Conceptual Foundations of Phonology as a Laboratory Science". Burton-Roberts, N. Carr, P., and Docherty, G. (eds. *Phonological Knowledge*, Oxford UK: Oxford University Press.

Phillips, B. S. (1984) "Word Frequency and the actuation of sound change". *Language* 60, 320-42.

Phillips, B. S. (this volume)

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

Sancier, M.L. and Fowler, C.A. 1997 "Gestural drift in a bilingual speaker of Brazilian Portuguese and English", *Journal of Phonetics* 25, 421-436.

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, and D. Schiffrin (eds.) *Towards a Social Science of Language*, 263-287. Philadelphia: Benjamins.

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