*Phonetic Diversity, Statistical Learning, and Acquisition of Phonology**

Janet B. Pierrehumbert

Northwestern University

Key words

categorization

exemplars

phonotactics

prosody

statistical learning

Abstract

In learning to perceive and produce speech, children master complex language-specific patterns. Daunting language-specific variation is found both in the segmental domain and in the domain of prosody and intonation. This article reviews the challenges posed by results in phonetic typology and sociolinguistics for the theory of language acquisition. It argues that categories are initiated bottom-up from statistical modes in use of the phonetic space, and sketches how exemplar theory can be used to model the updating of categories once they are initiated. It also argues that bottom-up initiation of categories is successful thanks to the perception-production loop operating in the speech community. The behavior of this loop means that the superficial statistical properties of speech available to the infant indirectly reflect the contrastiveness and discriminability of categories in the adult grammar. The article also argues that the developing system is refined using internal feedback from type statistics over the lexicon, once the lexicon is well-developed. The application of type statistics to a system initiated with surface statistics does not cause a fundamental reorganization of the system. Instead, it exploits confluences across levels of representation which characterize human language and make bootstrapping possible.

1 Introduction

Infants show evidence of phonetic categorization and of perceptual parsing of the speech stream before they learn to speak, before they have large vocabularies, and possibly before they even understand that words are referential. Some parts of the speech processing system are initiated early. However, the system takes a long time to develop, not achieving adult levels even at the age of 12, according to some recent results by Hazan and Barrett (2000). In adults, it achieves astounding levels of speed, accuracy and robustness in parsing complex, language-specific, phonetic patterns. In this article, I present some ideas about how the system is initiated and subsequently refined. These ideas are based on an integration of the research literature in linguistic phonetics, psycholinguistics, and phonological acquisition.

* *Address for correspondence:* Janet Pierrehumbert, Laboratoire de Sciences Cognitives et Psycholinguistique, Ecole Normale Supérieure, 46, Rue d'Ulm, 75005 Paris, France. e-mail: <jbp@northwestern.edu>.

As background, I assume that the ultimate target of phonological acquisition is a cognitive architecture with multiple levels of representation. These levels minimally include: (1) *Parametric phonetics:* A quantitative map of the acoustic and articulatory space, on which proximity in multiple dimensions can be defined. (2) *Phonetic encoding:* Low-level categorization of the phonetic space. (3) *The lexicon:* Lexical representations of word-forms, which provide a locus for association between form and meaning. (4) *The phonological grammar:* General constraints on word-forms in the lexicon such as constraints on metrical structure or segmental sequencing. (5) *Morphophonological correspondences:* Phonological relationships amongst morphologically related words which are not independently predictable from constraints on word form.

Consider, for example, the lexicalized compound, the verb *blindfold* (to prevent someone from seeing by fixing a cloth over the eyes). The parametric phonetic space provides a way to represent the time course of spectral and/or articulatory parameters on any individual occasion of the word being uttered. In speech perception, it represents the perceptual capture of the speech which makes it possible for the speech to be submitted to cognitive processing of any kind. In speech production, it represents a motor plan with appropriate specification in time and space of motor gestures. The facts captured by the phonetic encoding system would include the difference between the clear /l/ in the initial /bl/ cluster and more vocalic /l/ in the final /ld/ cluster, as well as the contrast between the obligatory release of the initial /b/ and the optional release of the medial and final /d/s. These are language particular subphonemic details which nonetheless have their role in production or perception of the word.

The phonological grammar assigns a phonological word boundary in the middle of the word *blindfold*, despite its semantic opacity. Two crucial factors in the phonological parse are the medial /df/ cluster and the superheavy first syllable, which contains a diphthong plus two consonants. These features are rare or impossible in the absence of a word boundary. Lastly, knowledge of morphophonological correspondences yields the prediction that the neologism *blindfoldee* would exhibit a shift of the primary stress to the last syllable, just as in *examine*, *examinee*. In contrast, the stress would remain on the stem in the neologism *blindfolder*, as in *employ*, *employer.*

There are systematic logical dependences amongst these levels, and these dependences must be both exploited and created during language acquisition. At the periphery of the system, encoding the speech signal depends on capturing it in the first place. Development of the lexicon depends on the existence of a system for encoding lexical items. Generalizations about word-forms depend on knowing a sufficient number of words. Knowledge of morphophonological relations likewise depends on having a sufficient vocabulary, and a sufficient knowledge of syntactic and semantic relations amongst words, for relevant word pairs to be identified and for generalizations to be formed over these pairs.

It is thus no surprise that these levels of representation are manifested in the order given, from peripheral to abstract, as shown by the review in Vihman (1996). Infants appear to be innately predisposed to attend to speech, and evidence of basic phonetic encoding is found almost from birth, as shown by Mehler, Jusczyk,

Lambertz, Halsted, Bertoncini, and Amiel-Tison (1988) and subsequent work. Werker and Tees (1984) found that infants in the first six months attend to a wide variety of dimensions of phonetic variation. But later in the first year, they show reduced sensitivity to variation in parts of the phonetic space which are not utilized in the ambient language. These results may be understood in terms of the projection of phonetic encoding units from experienced speech. These units are available for use in learning word forms. A lexicon becomes evident by the age of one year, and knowledge of morphophonological alternations begins subsequently, with a strong dependence on vocabulary development. Thus, extremely regular and productive alternations (such as the voicing and vowel epenthesis in the English plural) appear early, while the knowledge of unproductive, irregular, or opaque alternations found only in erudite vocabulary continues to develop into adulthood. The focus in this paper will be the development of the first four levels (the parametric phonetic levels through the phonological grammar). Morphophonological alternations will not be considered further here.

The phonological system is built while being used. Since the knowledge that can be acquired at any time is dependent on the processing capabilities at that time, we can only understand acquisition in terms of the relationship between processing and knowledge. Thus, adult models of speech perception provide an important reference for models of language acquisition by children. In what follows, I will therefore presuppose some of the consensus features of current speech processing models, such as Norris, McQueen, and Cutler (2000) and Vitevich and Luce (1998). These models include the following features: (1) A fast, automatic encoding system which exploits general features of a language to decompose the speech stream. A key role of this encoding system in adults is identifying possible word boundaries, with lexical access demonstrably facilitated when strong cues are available in the speech stream. (2) A lexicon, which is the locus of the association between word meanings and word forms (3) The ability to form higher level abstractions over lexical items. In the MERGE speech perception model of Norris et al. (2000), this ability is implicated in phoneme identification, which takes place after lexical access rather than before. In the model of Vitevich and Luce (1998), it is involved in the way that lexical properties and frequencies interact in decisions about lexical items.

The present paper has three interconnected themes. One theme is the terrible complexity of phonetic patterns. The problem of phonological acquisition is far worse than generally supposed by psycholinguists, because of the large amount of language-particular phonetic detail which must be acquired. Both phonological categories and prosodic structures have language-specific phonetic characteristics. A bit of the speech signal which counts as voiced in one language might count as unvoiced in another, and a bit which counts as stressed in one language might count as unstressed in another. These observations point to the conclusion that categories are acquired from statistics of the speech signal (as opposed to being made available a priori by universal grammar).

Models for describing such learning have been developed for perceptual categories. These models rely on the understanding of categories as labels over a phonetic map, with the frequency distribution for each label being incrementally updated

through ongoing exposure to speech. The recent results by Maye and Gerken (2000) and Maye, Werker and Gerken (2002) on learning by infants and adult are propitious for this class of model, indicating that categories may be initiated bottom up on the basis of statistical modes in the speech signal. However, the high separability of categories in the adult system, and the reflexes of lexical contrastiveness in the phonetics, also point to a role for feedback. Thus, feedback is a second theme of this paper. However, I will not be concerned with on-line feedback from individual lexical items to the speech encoding system, the most hotly disputed feature of the TRACE model proposed in McClelland and Elman (1986). Instead, I will consider two other types of feedback. One is community feedback (e.g., the feedback loop set up by speech communication in the population). The existence of the feedback loop through the population is undeniable, and the key issue is thus whether it is sufficient to explain the maturation of the categorization system. The alternative for sharpening the category system is internal feedback from the general properties of the lexicon, that is, from the phonological grammar to the encoding system. I will argue that community feedback is more powerful than might be supposed, but that there is still some evidence for internal feedback from the phonological grammar as the system matures.

The third theme is the confluence amongst levels of representation in the phonological system. The phonological system appears to be initiated bottom-up from surface statistics over the speech stream, but refined using type statistics over the lexicon. Nonetheless, learning appears to proceed incrementally and the use of type statistics does not require any fundamental reorganization. This is only possible because of subtle but systematic relationships across levels. These relationships characterize human language and play a part in distinguishing actual human languages from conceivable but unnatural language systems.

2 Some terminology

In what follows, I will use some common technical terms in very specific ways.

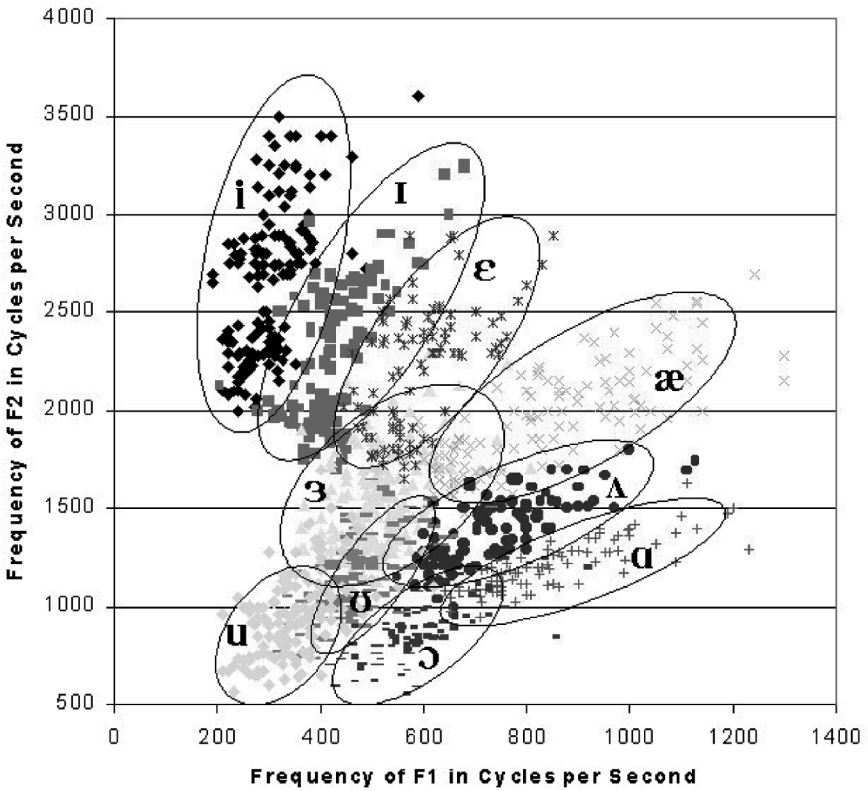
Segment: I will use the term segment for a temporally minimal unit of encoding or analysis, regardless of the level at which it appears in the system. Thus, phonemes and allophones both count as segments, even though phonemes are more abstract than allophones. Thanks to acoustic landmarks (see Stevens, 1998), the decomposition of the speech stream into segments can be presupposed in some cases. In other situations, it is much less clear and different languages or different children may impose different segmentations on the same speech signal. This usage is consistent with that of the International Phonetic Alphabet (the IPA), in which broad phonemic and fine phonetic transcriptions are both taken to be segmental transcriptions, despite plain differences in the level of abstraction represented.

Phoneme: The term phoneme will be used in a narrow sense as a minimal unit of contrast in the lexicon. Following the traditional literature, I will also take equivalence across contexts as a key characteristic of phonemes. That is, the phonemic level is one at which the start of the word *pat* is the same as the end of word *cap*, and the end of word *cap* is the same as the start of the second syllable in *capital*. As we will see, these requirements substantially curtail the role that the phoneme could possibly play in phonological bootstrapping.

Category: I will use the term category in a broad sense which harks back to the foundational works of mathematical psychology, such as Luce (1963) and Luce and Galanter (1963). A category is a mental construct which relates two levels of representation, a discrete level and a parametric level. Specifically, a category defines a density distribution over the parametric level, and a category system defines a set of such distributions. Using the density distributions for categories in a category system, incoming signals may be recognized, identified, and discriminated through statistical choice rules. This understanding of categories has been generally adopted in experimental phonetics and sociolinguistics. An example is provided by the standard representation of a vowel space as a set of density distributions in F1-F2 space, as in Figure 1.

Figure 1

The vowel space of Peterson and Barney (1952), illustrating the concept of categories as density distributions in a parametric space. Figure created by Stef Jannedy, and reproduced from Pierrehumbert (2003)



On this understanding, the system of phonological categories includes not only segments, but also other types of discrete entities in the phonological grammar, such as tones, syllables, and metrical feet. Each of these has phonetic correlates in its own

right. Since a category is a statistical relationship between a discrete level and a parametric level, it follows that two categories are identical only if they are identical at both levels. Analogously, although the French word *marron* is sometimes translated into English as *brown*, it is not actually the same category. The percepts which would be described as *brown* in English are divided amongst *marron*, *brun*, and *doré* in French, with *doré* in some cases glossed as *golden* in English. Thus, the relationship of the categorical label to the parametric level is not actually the same in the two languages, even if a certain similarity can be discerned. If two categories are not identical, they may still be equivalent (if an appropriate equivalence relation can be defined, in the mathematical sense), or analogous (if they are comparable in some looser sense).

It is also important to be clear on what constitutes a small phonetic difference. A main theme of my paper will be the scientific challenges raised by language-specific categorization of the parametric phonetic level. Within-category differences are typically much smaller than what is described as a "fine phonetic difference" in the psycholinguistics/acquisition literature. This term is commonly used to refer to a phonologically minimal categorical distinction in the adult system. Such differences are objectively and perceptually large compared to within-category differences. For example, the difference between /b/ and /d/, explored in Werker and Stager (2000), is lexically contrastive in English and would be recognized with extremely high accuracy by English-speaking adults, thanks to its many phonetic cues (including the formant transition, the burst amplitude and spectrum, and the ratio of the closure to the voice onset time). Most of the contrasts explored in Swingley (this issue) are of a similar nature. To date, only a minority of experiments in the language acquisition literature explore differences as small as those explored in the research literature on psychoacoustics, sociolinguistics, or phonetic typology.

3 Phonetic learning

In classical surveys of phonetic typology, similar items appearing in different languages are treated as members of the same category. For example, surveys reported that both French and Finnish have a voiceless unaspirated labial stop, /p/, or that both English and Finnish have a trochaic foot structure. These reports were based on impressionistic data, inevitably influenced by the category system of the person making the transcriptions. Since the introduction of high-powered computer stations, it has become possible to gather large and objective data sets on the quantitative properties and exact patterns of variation of phonological categories in different languages. Such studies have revealed that superficially analogous categories have different quantitative properties in different languages. These detailed differences must be learned by native speakers, because they have consequences for category boundaries in perception and because they must be accurately reproduced to achieve a native accent in production.

Such results have been found for segments, prosodic features, and intonation. A few of the many relevant findings are summarized here. Additional references may be found in Pierrehumbert, Beckman, and Ladd (2001).

3.1 *Segments*

In American English, glottalization (produced by adducting the vocal folds) occurs on voiceless stops, especially /t/, variably in coda position. Glottalization can also, in effect, provide a null consonant onset for words beginning in a stressed vowel, especially when they follow a vowel-final word. Ramifications of the vocal fold adduction include reduction of amplitude and disturbance of the F0, as discussed in Hillenbrand and Houde (1996), and Pierrehumbert and Frisch (1996). In Coatzospan Mixtec, glottalization is a contrastive feature of vowels (Gerfen & Baker, in press). A comparison of quantitative results on these two languages shows that the phonetic ranges of the phonological categories overlap; some instances of amplitude reduction and F0 disturbance which would be attributed to a vocalic feature in Mixtec would count as instances of a consonant in English.

Engstrand and Krull (1994) report measures of vowel lengths in conversational speech in Swedish, Finnish, and Estonian. In Swedish, there is more extensive overlap in the distributions of durations for long and short vowels than in Finnish or Estonian. The authors relate this finding to the fact that Swedish vowel length distinctions are reinforced by formant structure differences to a greater extent than in Finnish and Estonian.

Overlap of articulatory gestures means that sequences of stops in many languages are produced with partially overlaid closures. That is, the second closure is formed before the first is fully released. Experiments by Kochetov (2002) show that the degree of overlap is less in Russian than in English.

In American English, word-final stops are often unreleased. The distinction between voiced and voiceless stops, which would be compromised by the lack of information in the release, is effectively cued by the length of the preceding vowel. In Indian English, vowels are not reliably lengthened before voiced stops. But the final stops are always released, effectively cueing the voicing distinction. This aspect of Indian English phonetic implementation is illustrated in Figure 2 (overleaf) which contrasts the words "pup" and "pub" produced in the same context *The pup/pub deserved a prize*. This treatment of voicing in Indian English reflects that found in Hindi (Hussain & Nair, 1995).

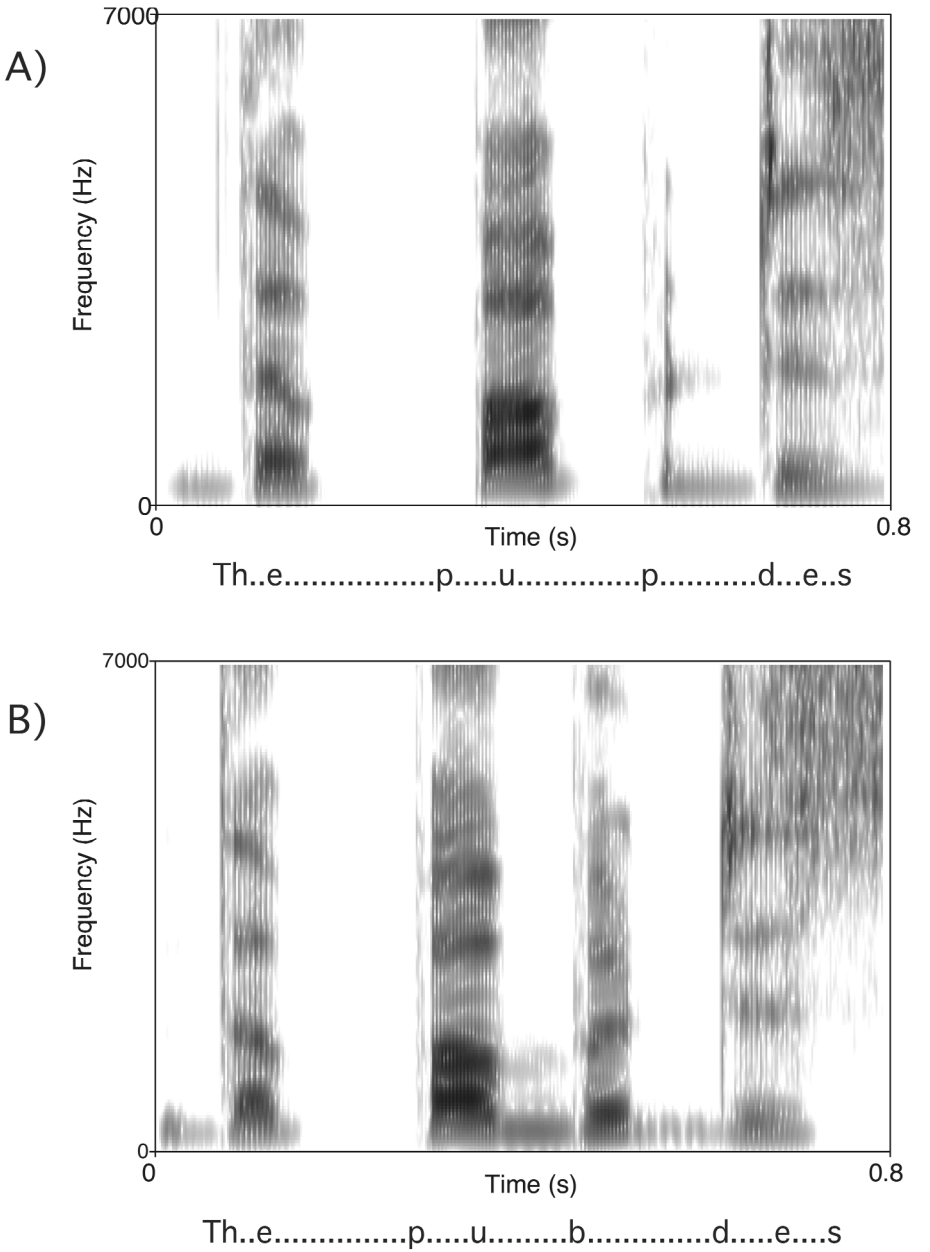
All of these differences may be viewed as differences in the phonetic implementation of segments. On deeper thought, most of them also have a prosodic aspect in that they involve the phonetic properties of segments as a function of their position in the syllable structure, the metrical structure, or the word. Thus it is unsurprising that similar variation has also been found for the standard reflexes of prosodic structure and intonation. I now turn to some examples.

3.2 *Prosody and intonation*

In linguistic theory, the term prosody is used to cover all aspects of grouping, rhythm and prominence from subparts of the syllable up through the organization of words in the phrase. Intonation refers to tonal/melodic features assigned at the phrasal level. In some languages, including English, the melody is entirely assigned at the

Figure 2

The first three syllables of the minimally different sentences *The pup deserves a prize* and *The pub deserves a prize*, produced by a native speaker of Indian English. The vowel lengths of the second syllable are the same, and the voicing contrast for the word-final stop is carried by the properties of the stop gap and release. Fuller data set to appear in Beckman and Pierrehumbert (forthcoming)



phrasal level as a function of pragmatic meaning. In languages with lexical tone or lexical pitch accents, such as Mandarin or Japanese, these lexical features interact with intonational features to determine the fundamental frequency contours assigned to phrases. The prosodic and intonation systems of different languages present broad analogies at an abstract level. These analogies have inspired some authors to propose universal categories in these domains. For example, McCarthy and Prince (1993), building on Prince's previous work in metrical phonology as well as McCarthy's work in prosodic morphology, proposes a universal system of bona fide prosodic categories including the mora, the syllable, and the metrical foot.

Despite these analogies however, the quantitative reflexes of prosodic and intonation categories in the speech stream are surprisingly diverse. For example, it is often assumed that word or phrase level stress lengthens the stressed syllable. However, this lengthening is really only characteristic of so-called "dynamic stress" languages such as English. In these languages, stress is phonetically implemented as a broad modulation of the force and duration of the articulatory gestures for the segments. Furthermore, word stress and phrasal stress are implemented uniformly in English, with the result that phrasal stress is associated with an increase in the duration and extent of the articulatory gestures for the most stressed syllable of the head word of the phrase (see Beckman, 1986; de Jong, 1995). The situation is different in languages such as Finnish and Aleut that use vowel length distinctively in lexical contrasts without a supporting distinction in vowel quality. In these languages, lengthening of stressed vowels is claimed by Berinstein (1979) to be rare or nonexistent. Though the measured durations of Aleut vowels in Taff, Rozelle, Cho, Ladefoged, Dirks, and Wegelin (2001) do show some influence of stress, they also display considerable overlap between the durations of stressed and unstressed vowels of the same phonological length. The result is that duration could not be an effective cue to stress. A further twist on how a language can protect a vowel length contrast is provided by Lehtonen's (1970) study of phrasal stress in Finnish. According to Lehtonen's data, phrasal stress is associated with lengthening, but not necessarily lengthening of the syllable with the main word stress. If the stressed syllable is long, the lengthening is localized on that syllable, but if it is short, the lengthening is displaced to the following syllable.

It is also commonly assumed that stressed syllables are marked with a fundamental frequency (F0) peak or rise, and that the existence of this feature peak can be used bottom-up as a cue for stress. However, in the rich intonation systems of English and German, a lexical stress may be marked with many different F0 features. This point is illustrated in Figure 3 overleaf. Note in particular the downstepped accent on the word *ballgown* in Figure 3B and the low accent on *orange* in Figure 3C. Figure 4 illustrates the L* + H L H% contour, which is used in English to convey reservations about the propositional content of the discourse, for example, uncertainty or incredulity (see Pierrehumbert & Hirschberg, 1990). This contour has a marked F0 depression on the stressed syllable, and the peak can occur as much as two syllables later, as shown in the example. Note also that there is no F0 marking of the primary stress of the word *monomorphemic* which is postnuclear in the phrase due to focus on *rigamarole*. However, the F0 rises on the final (unstressed) syllable due to the choice of phrasal contour. Thus all peaks and rises in this contour fall on unstressed syllables.

Figure 3

A, B, C: Three distinct intonation patterns on the phrase *An orange ballgown*. See Pierrehumbert and Hirschberg (1990) for discussion of the pragmatic meanings of the contours. Figure reproduced from Beckman and Pierrehumbert (1986)

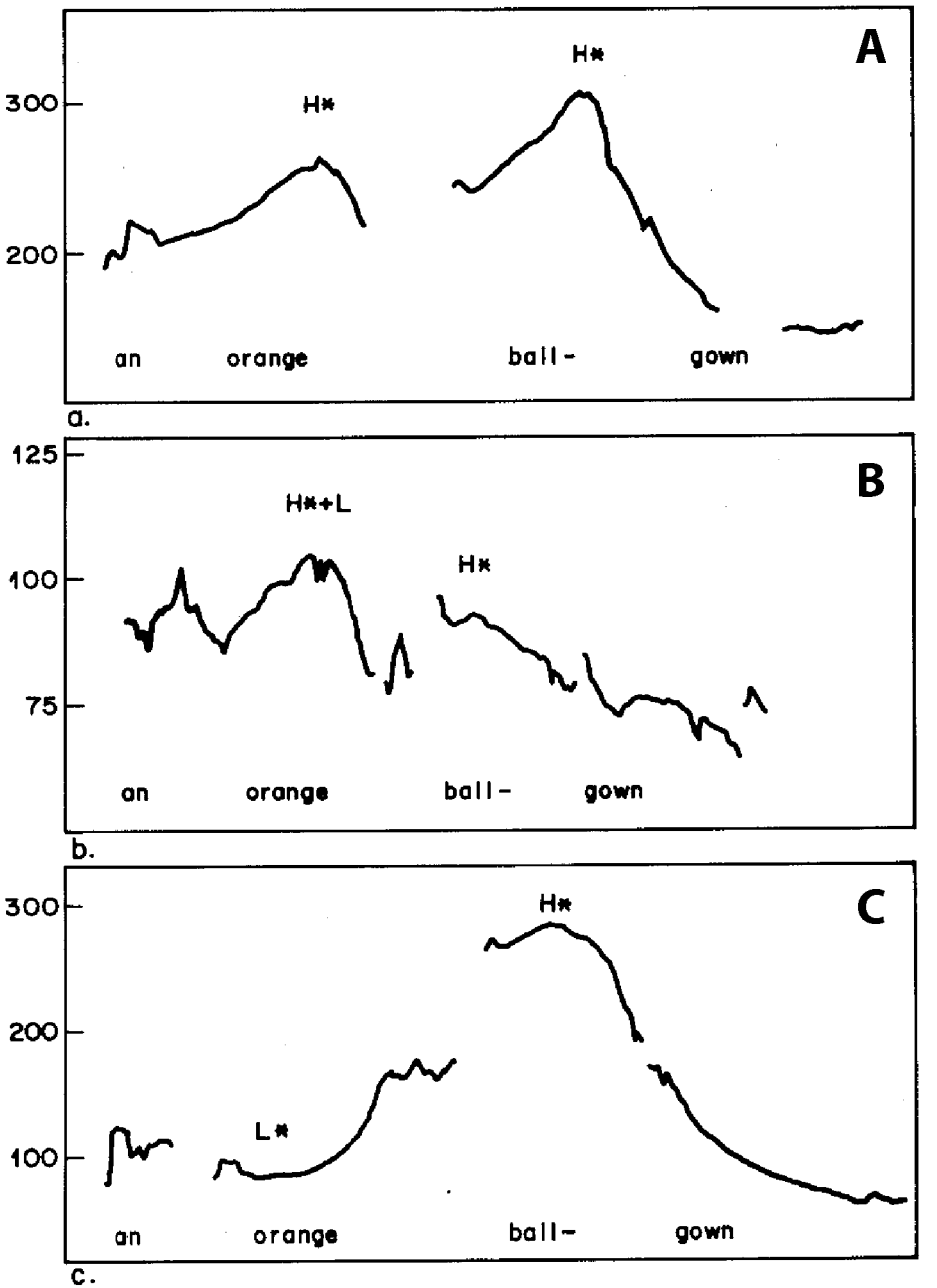
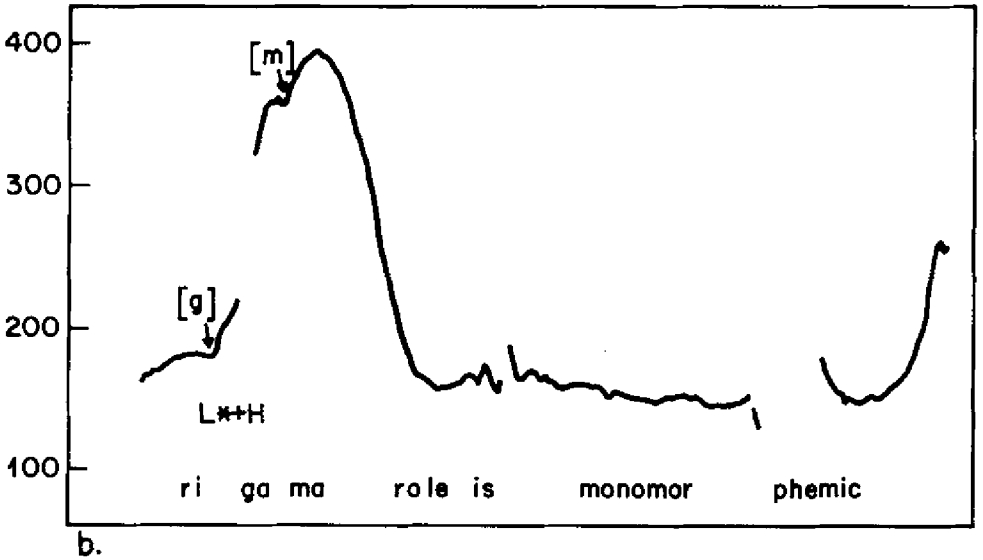


Figure 4

L + H* L H% on the phrase *rigamarole is monomorphemic*. In this example, the F0 rises and peaks are temporally aligned with unstressed syllable. Figure reproduced from Beckman and Pierrehumbert (1986)

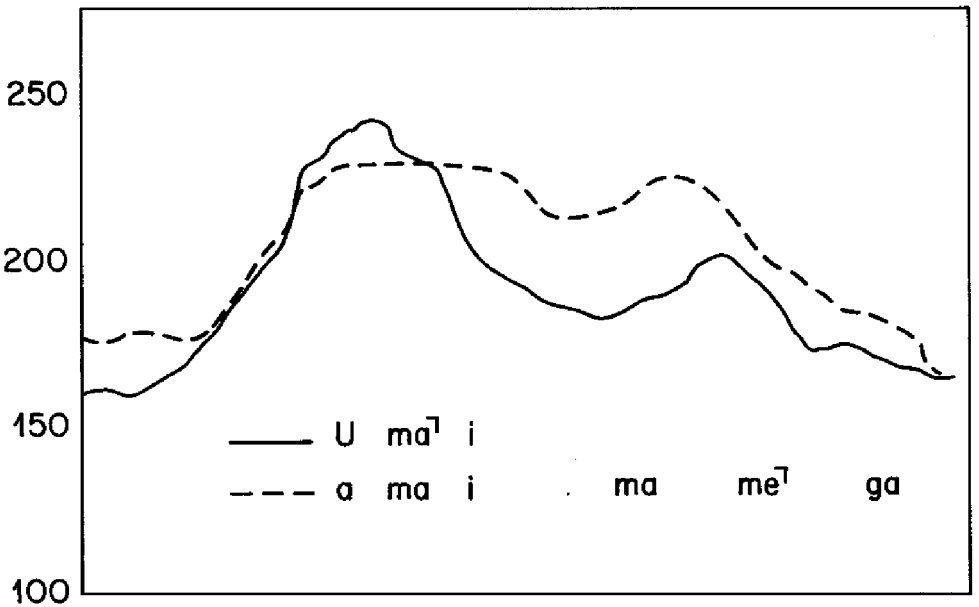


Typological studies of intonation have also identified a number of languages in which the default pitch accent type is associated with a distinctively low F0 on the stressed syllable. For example, the exceptionally thorough study of Danish intonation described in Grønnum (1992) reveals a low-rising accent (extremely similar to that in Fig. 4) to be the norm for Danish. Hussain (1997) also shows that stressed syllables in Urdu are lower than comparable unstressed syllables rather than higher. In Japanese, the metrical patterns are completely decoupled from the pitch accent patterns, with the result that F0 is not a cue for metrical prominence. (Beckman, 1986.)

Many of the world's languages make use of an F0 pattern which incrementally steps down at key positions in the phrase, forming a downward staircase. Such downstepped F0 patterns exist in both Japanese and English. It is the most common pattern in Japanese (Pierrehumbert & Beckman, 1988) and British English (Crystal, 1969). It is pragmatically marked, but still fairly common, in American English. (Pierrehumbert & Hirschberg, 1990.) Figure 5 illustrates the contrast between the downstepped pattern in Japanese (triggered by the presence of a lexical accent) with a nondownstepped pattern (found in phrases with no lexical accent). The downstepped pattern is extremely similar to that shown for English in Figure 3B. The relationship of the downstepped pattern to phrasal stress is problematic. Nespor, Guasti, and Christophe (1996) suggest a tight connection between phrasal prominence and basic syntactic word order, which children could presumably exploit in bootstrapping the abstract syntactic system from more superficial characteristics of the speech. This account would imply that the downstepped contour in Japanese is characterized by

Figure 5

Contrast between two F0 contours in Japanese, one with and one without downstep. Downstep is triggered by the H + L lexical accent. Figure reproduced from Pierrehumbert and Beckman (1988)



initial prominence (related to the basic SOV order of Japanese) whereas the extremely similar contour in English has final prominence (related to the basic SVO order of English). Given how extremely similar the contours are, it is not clear how this differential phonological analysis could be arrived at bottom-up. It appears more likely that the location of the nuclear stress in the English downstepped contour is inferred from a more complex analysis of the patterns of variation in the language.

In summary, some languages use length to realize stress, whereas other languages use length primarily for other purposes. Even when length is used, the way it is used can depend on language-specific treatment of the interaction between word stress and phrasal stress. Overall force of articulation is a correlate of stress only in dynamic stress languages such as English. F0 is not a universal cue to stress because the F0 associated with stress may be distinctively high, distinctively low, or variable. In addition, some languages dissociate F0 from stress entirely. The extent of such variability means that the relationship of phonetic cues to the hierarchical prosodic and intonational representation must be learned in each individual language.

3.3
Relation of perception to production

The studies I have just summarized reveal language-specific phonetic details in production. Some models of speech processing have separate modules for production

and perception, and in early language acquisition, perception leads production. Thus, the connection between production and perception needs to be examined. A number of outstanding studies have taken up exactly this issue. These studies generally show that the perception system in adults in attuned to the way that phonological contrasts are manifested in each individual language. The weighting and interaction of cues in perception thus depends in language-specific fashion on the strength, reliability, and phonological importance of the cues.

For example, a universal tendency to lengthen vowels before voiced consonants is manifested to different degrees in different languages. Flege and Hillenbrand (1986) show that English and French listeners make differential use of the length of /i/ and /z/ in pairs such as *peace*, *peas*. English speakers are more influenced by the length of the vowel than the French speakers, a finding which accords with the production patterns in the two languages. Overlap of articulatory gestures leads to a universal tendency to nasalize vowels before nasals, and to anticipate in one syllable the vowel quality of the next. However, the extent of these effects depends on the system of lexical contrasts. Beddor and Krakow (1999) show that English and Thai listeners differ in perceptual compensation for coarticulatory nasalization. In production, Thai has less anticipatory coarticulation in VN sequences than English does. Beddor, Harnsberger, and Lindemann (2002) show that English and Shona listeners differ in perceptual compensation for V-V coarticulation. Shona differs from English both in the degree and direction of V-V coarticulation. It has more coarticulation (possibly due to its smaller vowel system) and anticipatory effects exceed carryover effects, the opposite pattern from English. Beddor et al. (2002) found that these subtle differences are reflected in perception. They also provide an extensive review of related findings in other languages.

Systematic relations between patterns in production and sensitivity in speech perception have also been found in the area of prosody and intonation. Results presented by Dupoux, Pallier, Sebastian-Galles, and Mehler (1997) on stress perception in adult speakers of French and Spanish show that stress perception also reflects attunement to linguistic experiences. I will return to this work below. A perception-production experiment by Pierrehumbert and Steele (1990) shows that English has two categories of peak alignment for L + H pitch accents (L* + H vs. L + H*, where * indicates the tone aligning with the stress). These results may be compared to Kohler's (1987a, b) results on perception of peak alignment in German. Kohler's experiments show evidence for three peak alignments, rather than two. This difference clearly reflects differences between the intonational systems of the two languages.

3.4
Statistical modes and their separation

Results such as those just summarized indicate that categories, defined as relations between a discrete level and a parametric phonetic level, cannot be universal. That is, the picture of human phonetic resources as pegs in an IPA-like phonetic pegboard cannot be sustained. Instead, learning the phonetic patterns of a language involves learning probability distributions over the parametric phonetic space. This space

can be understood as a high-dimensional cognitive map on which a metric of proximity or similarity is defined. Relevant dimensions on this map include both acoustic and articulatory properties. Categories are labels over this map, and thus each label has associated with it a probability distribution over the space defined by the map.

These probability distributions are acquired from experience with a language, and empirical studies indicate that a very substantial amount of experience is necessary to master these distributions in full detail. Although infants show evidence of categorization of the speech stream extremely early, Nittrouer (1996) demonstrates further learning by three-year olds, and Hazan and Barrett (2000) show that categorization of consonants in minimal pairs such as *boat*, *goat* continues to develop between 6 and 12 years. At age 12 it still has not reached adult levels. Thus the model of phonological bootstrapping needs to explain how precocious initiation of the relevant levels of representation is combined with a lengthy process of elaboration and refinement.

A fully explicit model of how categories are acquired must address at least three aspects of the issue. First, what phonetic dimensions are initially available to define the phonetic map, and how do these dimensions evolve with experience? Second, how is the speech signal segmented? The signal must be segmented in some way for similar regions to be identified despite differences in their context. Third, what conditions must be met for a region of the phonetic map to be treated as a category in the encoding system? Current work in this area has tackled the third question by taking up some cases in which phonetic theory provides working assumptions about the first two. For example, the acoustic theory of speech production and the properties of the auditory system together mean that the formant map can be taken as a starting point for vowels. As discussed in Stevens (1998), articulatory, aerodynamic, and psychoacoustic considerations also make available some acoustic landmarks as segmentation points in the speech signal. I return below to some less straightforward cases.

When acoustic dimensions and segmentation are not problematic, recent results indicate that categories can be initiated bottom-up on the basis of modes in the statistical distribution of phonetic properties. In a set of experiments on infants and adults, Maye and Gerken (2000) and Maye, Werker, and Gerken (2002), created a synthetic phonetic continuum between /d/ and unaspirated /t/, which are not phonemically distinct in English in the position they examined. The subjects in the experiment were English speaking adults and babies being raised in an English-speaking environment. Adult subjects were told that they were learning a novel language. All subjects heard some examples of all values along the continuum. However, some subjects were exposed to a unimodal distribution over the continuum and others were exposed to a bimodal distribution. Those who heard the bimodal distribution were more likely to impute categorization to the variation in the stimuli. The thrust of these results echoes the earlier computational analysis of English vowels discussed in Kornai (1998). Analyzing vowel data from Peterson and Barney (1952), Kornai claimed that English vowels correspond to rather distinct clusters in the formant space, and as a result can be acquired from unsupervised cluster analysis (e.g., cluster analysis carried out without any feedback from higher levels about category mem-

bership). Kornai's claims are strengthened by the more recent survey of English vowels in Hillenbrand, Getty, Clark, and Wheeler (1995). In their survey of upper-Midwestern American English, vowels that overlapped in the F1-F2 space were shown to be statistically discriminable when additional dimensions of vowel length and formant motion are considered.

The results of Maye and colleagues also lead us to expect that bottom-up category formation should be most successful for categories which actually do present distinct modes in phonetic distributions in ordinary speech. This consideration immediately leads us away from the classic concept of the phoneme, towards a much less abstract level of characterization. A phoneme is a unit of sound structure which is equivalent across the various positions it appears in: the words *tap* and *pat* have the same phonemes in different orders. However, the phonetic outcome for any given phoneme is affected by its prosodic position and segmental context. The large amount of variation which results from these factors means that an outcome which represents one phoneme in one context may represent a different phoneme in a different context. For example, a devoiced /ə/ in *classify* is essentially the same as the /h/ in *hit*. The /z/ in *matches* is similar to the /s/ in *sit*. In general, the probability distributions associated with phonological categories are rather well distinguished within context (with some exceptions), but less well distinguished when measurements are combined across contexts. This point is developed at more length in Pierrehumbert (1993). Illustrative data on the duration of /s/ and /z/ as a function of position in the word are presented. Given positional devoicing of /z/, duration relations provide the single most reliable cue for /s/ versus /z/. The ranges of durations for the sounds are heavily overlapped, but for each given position, there is no overlap.

Thus, the most promising units for bottom-up category building appear to be positional variants of phonemes rather than phonemes in the classical sense. This is not to say that they are phones in the sense of the IPA. IPA phones are taken to be equatable across contexts just as phonemes can be; they do not include any representation of segmental, syllabic, or metrical position and they differ from phonemes only in the extent to which fine-grained detail is included. In fact, phones are worse candidates than phonemes for bottom-up category projection because they are much more numerous and thus their density distributions are therefore on the average more crowded together. For positional variants to be learnable as distinct modes in the parametric space, however, strong assumptions must be made about the context-dependence of perception. With regard to the /s/-/z/ distinction, a first step is to distinguish initial /s/ as in a phrase beginning with *Sue*, from initial /z/ as in a phrase beginning with *zoo*; and similarly to distinguish final /s/ from final /z/. Only subsequently can initial /s/ and final /s/ be associated, or initial /z/ and final /z/ be associated, possibly through an analysis of homologies. The position-dependent distinctions appear to be tractable if the final variants are inactive when initial variants are being perceived, and initial variants are inactive when final variants are being perceived. The distinctions are intractable if all four of initial /s/, initial /z/, final /s/, and final /z/ are simultaneous candidates for encoding any incoming fricative region in the speech stream. The combined distribution of these four positional variants does not provide distinct modes, and the overlap in phonetic properties between initial /s/ and final /z/ would tend to undermine the conceptual structures which we are seeking

to explain. Thus, extending results such as Maye and Gerken's to the more general situation requires the assumption that perception of segments occurs relative to their prosodic context, and that prosodic encoding is learned concurrently with segmental encoding. Fortunately, a large body of experimental work by now indicates that perception of prosody is just as precocious as attunement to the ambient segmental phonetics (cf. Mattys, Jusczyk, Luce, & Morgan, 1999; Mehler et al., 1988; review in Vihman, 1996).

Assuming that lexical representations consist of what the phonetic encoding system provides, the claim that positional variants are the workhorses of phonetic encoding is in line with the research reviewed in Pierrehumbert (2002). This body of work indicates that long term representations of words in the lexicon are more detailed than previously believed. Important findings include the fact that people automatically imitate speaker-particular details of words they have heard many times in a particular voice. (Goldinger, 1996, 2000.) Leniting historical changes differentially impact high- and low-frequency words over long periods of time. (Phillips, 1984, review in Bybee, 2001.) Semantic gangs of words can be left behind in a historical phoneme shift, if they are used predominately in a certain situationally evoked speech register (Yaeger-Dror, 1996; Yaeger-Dror & Kemp, 1992). Detailed lexical representations are also needed to explain cases of phoneme splits as discussed in Labov (1994); see in particular the discussion of "lexical splits," in which complex conditioning of an allophonic pattern combines with a fringe of irregular cases in order to yield a lexicalized difference in what was previously subcategorical variation. Lastly, morphological paradigm effects can include subphonemic detail. Steriade (2000) reports paradigm effects for flapping in English. She shows that use of an optional flap in words such as *positive* tends to be carried over to optional flapping in derived forms such as *positivistic*. Scobbie, Turk, and Hewlett (1999) also provide evidence for morphological paradigm effects on vowel length in Scottish English. If such details are retained in long-term memory, they must have been encoded in the first place. Accordingly, the initial phonetic encoding of the speech stream, which is used to access the lexicon, appears to be quite detailed.

Results such as Maye and Gerken (2000) and Kornai (1998) leave open, however, many issues about how category formation occurs in the natural setting. One issue is segmentation. Though some acoustic landmarks are obvious, other segmentation issues are so subtle that even linguistic theorists are not agreed on the correct answer. Are affricates sequences of a stop and a fricative? Or merely a stop with a strident release? Is English /e/ (phonologically long and pronounced with an offglide) a single vowel occupying two structural positions? Or is it rather a sequence /ej/ containing a vowel and a glide? These questions are vexed, indeed so vexed that children learning the same language may not answer them in the same way. It is not a problem for one speaker to posit two units where another speaker has one, as long as words produced by one can be understood by another.

A more subtle issue related to segmentation is how the perceptual system copes with the articulatory and acoustic overlap amongst segments. An important set of case studies involves the interaction of nasals with neighboring segments: Vowels preceding nasals tend to be nasalized, and nasals in turn tend to share place of artic-

ulation with a following obstruent. Lahiri and Marslen-Wilson (1991) report that objectively nasalized vowels in the context of a nasal consonant are not perceived as nasal in Bengali. An initial interpretation of such results is that contextually predictable variation is simply ignored in perception. However, more recent results in Beddor et al. (2002) indicate that compensation for coarticulation is incomplete, and shows quantitative, language-specific effects. Amplifying this more detailed view are the findings by Bradlow (2002) and Dahan, Magnuson, Tanenhause, & Hogan (2001). Bradlow shows that coarticulation patterns are preserved in clear speech. Dahan et al., (2001) show that that misleading anticipatory coarticulation cues slow lexical access. These results support a view in which segment boundaries are soft, and temporally distributed information in the signal tends to be perceptually aggregated with information located at points of highest discriminability. For example, nasalized vowels are weakly distinct from oral vowels, but nasal stops are extremely distinct from oral stops. Thus, nasalization on the vowel would be parsed with nasalization of a stop if a nasal stop is present.

This slippery view of segmentation suggests that the segmentation of temporally distributed phonetic properties can be quite unstable, and this is indeed the case. For example, if nasal stops are shortened or lenited (degrading their discriminability from other consonants or from the lack of a consonant) nasalization can be reassigned to the vowel in a well-attested line of historical development. Hajek and Maeda (2000) explore the complex interactions with vowel height and length in the development of distinctively nasalized vowels. Scobbie (personal communication) also points to the surprising segmentations which are entertained by children with perceptual or phonological deficits. These unexpected segmentations could come about because the initial perception of some phonetic properties is not the same as for other children, leading to different loci of statistical discriminability.

A third important issue is the dimensional organization of the phonetic map. The VOT dimension explored in Maye and Gerken (2000) is unproblematic to both scientists and subjects. It has been heavily investigated and is known to be highly salient in the auditory system. Furthermore, it was no doubt salient to the subjects as the only dimension of variation in the experiment. In natural speech, the dimensions of variation are extremely numerous. One of the reasons for the diverse phonetic typology discussed above is that different languages bundle the phonetic dimensions differently for purposes of phonological categorization and lexical contrast. As discussed in Hussain (1997), Urdu differs from English in that the voicing information is carried primarily by the properties of the stop closure, not by the region after the stop burst. The region after the burst carries a breathiness feature which is distinctive for both voiced and voiceless stops. Thus, the four-way contrast in Urdu amongst /p/, /b/, /p^h/, /b^h/ requires children learning Urdu to maintain two phonetic dimensions of closure voicing and breathiness which English speaking children conflate into a single dimension of overall voicedness. Similarly, English speaking children maintain an encoding dimension corresponding to the third formant (critical to distinguishing /ɹ/ from /l/). The difficulty Japanese learners of English experience in attending to F3 reflects the loss of this dimension in Japanese (Strange & Dittman, 1984). In fact, learning to attend to this dimension in learning a second language induces observable changes in patterns of brain activation (Zhang, Kuhl,

Imada, Iverson, Pruitt, Kotani, & Stevens, 2000). Thus, understanding how children learn which parameters are functioning together in which contexts is an extremely important and challenging issue.

3.5 *Exemplar theory*

Exemplar theory, as discussed for speech by Johnson (1997) and Pierrehumbert (2001a), provides a way to model category formation and refinement at the level under discussion. Johnson (1997) lays out the concepts of the approach as applied in speech perception. The input is an auditory coding of the speech signal. A covering map provides an analog representation of the phonetic space, with the dimensions being the many phonetic parameters which are relevant to speech perception. Any particular speech stimulus defines a location on the map by virtue of its perceptual properties. Category nodes are labels over the map. Each category label is associated with a frequency distribution of remembered instances of that label. In a straightforward computational implementation of the approach, as in Pierrehumbert (2001a), these frequency distributions are established by simply storing in memory every encoded percept (or exemplar); the strength of the representation at a location on the map depends merely on the number and recency of the exemplars at that location. Of course, this is merely a mathematical schema and any cognitive model which supported the acquisition and ongoing updating of frequency distributions for categories would retain the important features of the model. Each incoming stimulus is categorized through a standard statistical choice rule, which compares the competing distributions in the neighborhood of the incoming stimulus and selects the most probable one (for details see Johnson, 1997; Luce & Galanter, 1963). A more neurolinguistic interpretation of this choice process is that each new token activates exemplars as a function of proximity in the parameter space. The strength of the activated exemplars cumulates in activating category labels. Labels compete through mutual inhibition.

Clearly, this approach provides a way to capture the results of typological studies showing that languages differ in arbitrarily fine phonetic detail, as reviewed above. With a sufficient level of exposure, a distribution can be learned at an arbitrary location on the phonetic map. Since the frequency distributions for labels are not normalized separately, the approach is also successful in modeling frequency effects. High frequency labels are advantaged in perception because they are more strongly represented on the cognitive map. The approach also explains how categories can continue to evolve and sharpen for a long time after they are first initiated. Through incremental experience, listeners acquire more and more accurate estimates of both the center of any given category distribution and the behavior of the tails of the distribution. This leads to more and more adult-like boundaries and patterns of variation.

A critical innovation in Pierrehumbert (2001a) is to extend the model of perception to also cover production. In production, a label is first selected. A random sampling of the exemplar distribution is taken for that label. (Though as discussed in Pierrehumbert, 2001a, and Pierrehumbert, 2002, this sampling can be biased to model sociostylistic options.) The neighborhood of the selected exemplar is activated,

and the average properties of this neighborhood constitute the production goal. The production goal is executed, with noise, corresponding to noise in the planning and execution of motor gestures.

With this schematic account of both perception and production, it is possible to explore the effects of the perception-production loop (e.g., of the feedback through the speech community which occurs as a listener encodes the speech he hears, and speaks in turn to the person he was listening to). Some calculations of this type are discussed in Pierrehumbert (2001a, 2002) and Pierrehumbert and Gross (2003). The result of primary importance here concerns the fate of categories whose separation is poor compared to the amount of variability for each. This situation is illustrated in Figure 6, comprised of black and white stills from the color movies posted on the web site for Pierrehumbert and Gross (2003). Both columns in Figure 6 illustrate the evolution of a three-category system in a two dimensional phonetic space, for a single speaker operating in a uniform speech community. The speaker begins with three categories shown at the respective locations of the squares at the top of the page. Each category is continually updated as the speaker perceives and encodes incoming examples; the updated distributions then provide the basis for productions by the speaker. Distributions are displayed as two-dimensional histograms, with pixel saturation coding the density at each location in the phonetic space. The time span from Time 0 to Time 3 represents 8000 iterations of the loop, so that each step displayed shows the state after 2000 iterations. Moving down the left column, we see that the middle category, squeezed by competition on both sides, is disadvantaged in perception and production and eventually disappears. In the right column, the noise in the system is the same but the separation of categories is greater. Thanks to the greater separation, all categories can survive in the perception-production loop. As discussed in Pierrehumbert and Gross (2003), categories with marginal spacing, as in the first row, can also survive if random variation in production happens to increase their spacing. At this particular spacing, three categories survive about half the time, and when they do, they have drifted to an average of 133% of the original spacing. As the spacing is decreased to zero in the initial conditions, the probability of all categories surviving also goes to zero.

Realistic examples of such situations occur for both vowels and consonants. For vowels, /ɛ/ is squeezed in the F2 dimension between /ɪ/ and /æ/. These vowels have merged in Stockholm Swedish though not in English. American English /t/ occupies an enormous range of different glottal openings, from a highly adducted position in /tʔ/ to a highly abducted position in /t^h/. Figure 7 (overleaf) explores the hypothesis that there are five distinct categories along this dimension, comparing the situation for slight versus considerable random variation in each category. In 7A, the distributions for the five allophones overlap only slightly, so the total distribution for glottal width in this position shows five distinct peaks. This situation is predicted to be stable. In Figure 7B, the greater variability means that distributions overlap much more and the summed distribution has only a single broad peak. The situation in Figure 7B is unstable, as can be seen by considering effects of a perception-production loop on the upper three allophones. For the plain released but unaspirated /t/, all production outcomes in the horizontally striped area on the left hand side of the distribution will be misperceived as aspirated /t/s. All production outcomes in

Figure 6

Density distributions over time for three categories as they evolve through a perception-production loop in a uniform speech community. Distributions are displayed as two dimensional histograms, with pixel saturation coding the density at each location in the phonetic space. The time span from Time 0 to Time 3 represents 8000 iterations of the loop, so that each step displayed shows the state after 2000 iterations. The circumference for each category is hand-drawn through the outermost exemplars of that category. Left Column: Unfavorable spacing in comparison to the variability or noise in the system results in erosion and loss of the medial category. Right Column: All three categories survive when the spacing is adequate in comparison to the variability

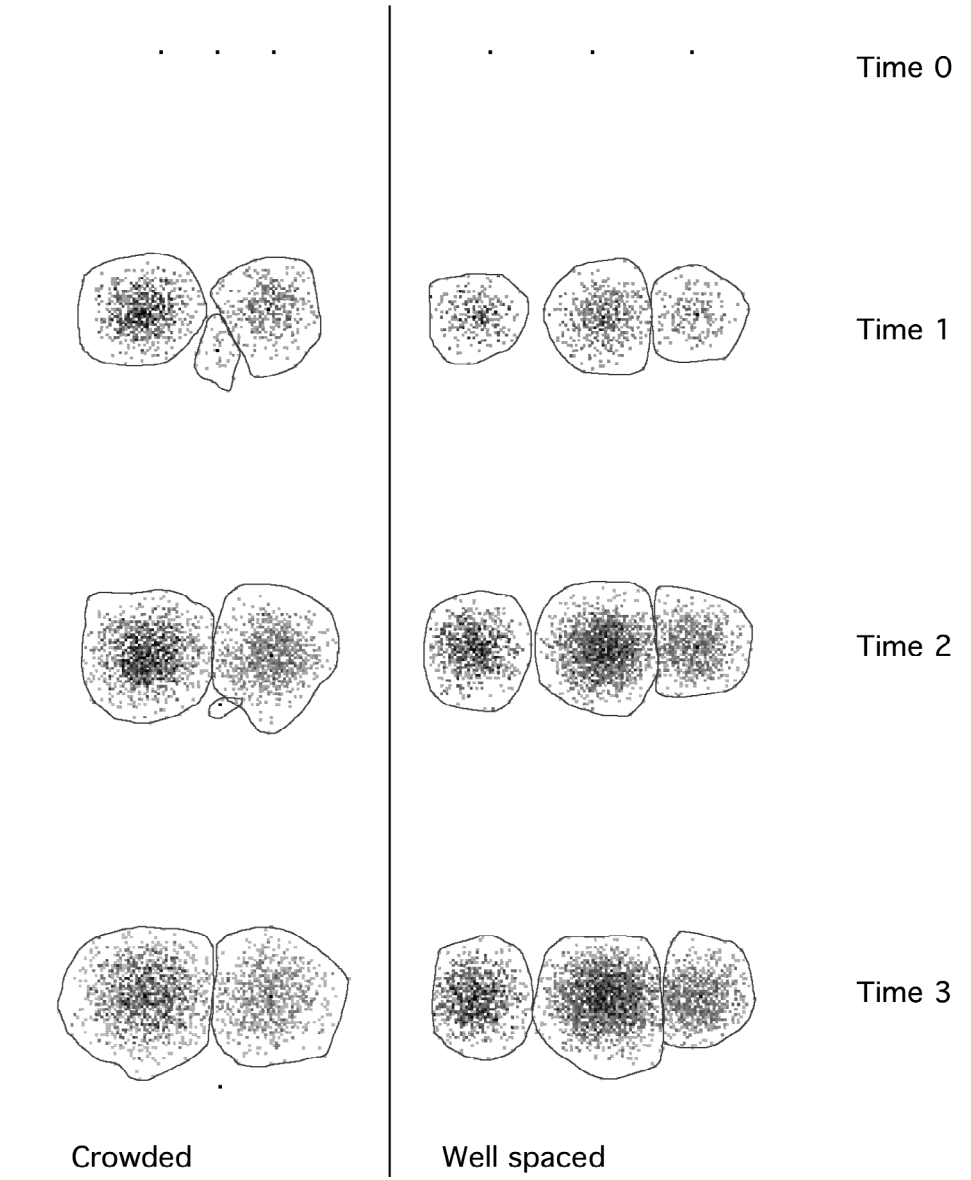


Figure 7

Stable versus unstable configurations for figure hypothetical categories along the glottalization … aspiration dimension for /t/

Figure 7A

Hypothetical allophonic categorization involving well-separated distributions which sum to a multimodal distribution

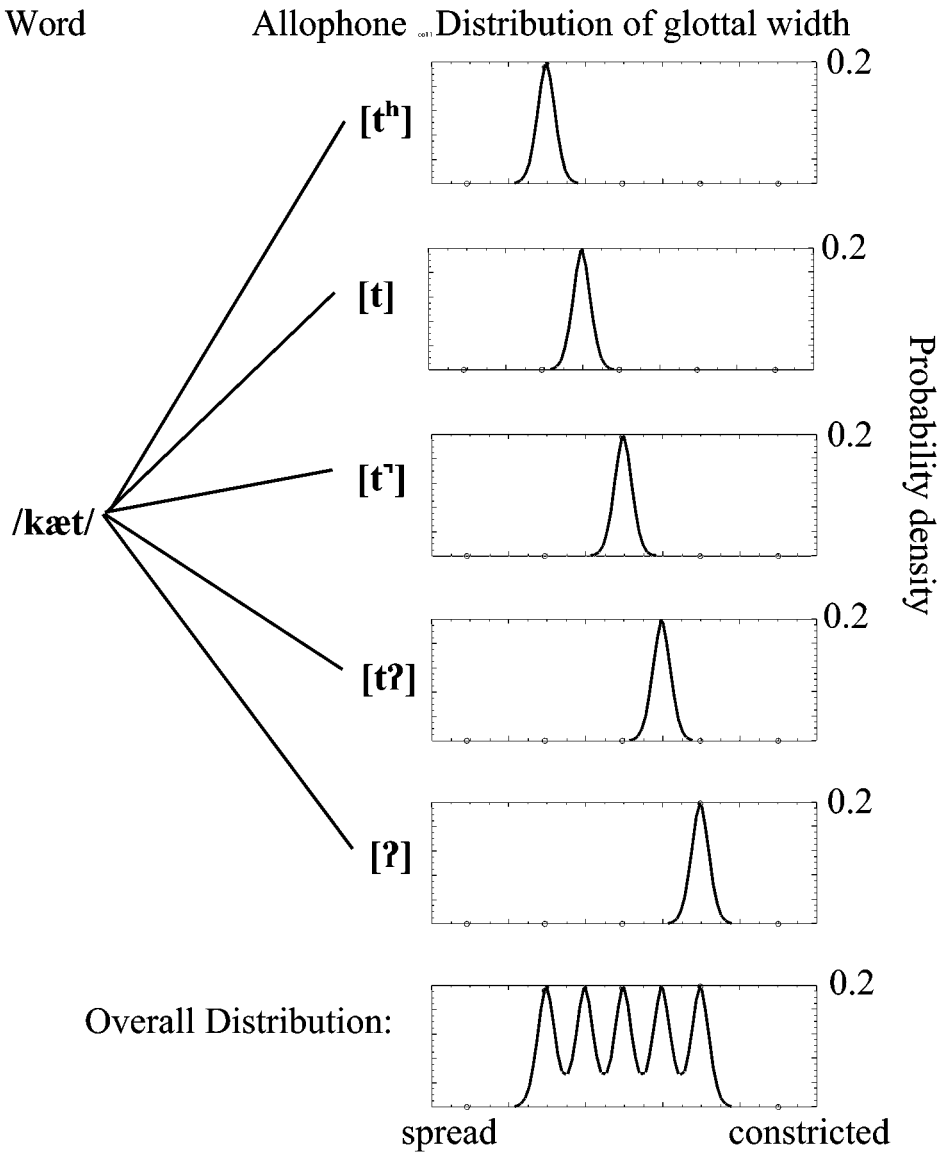
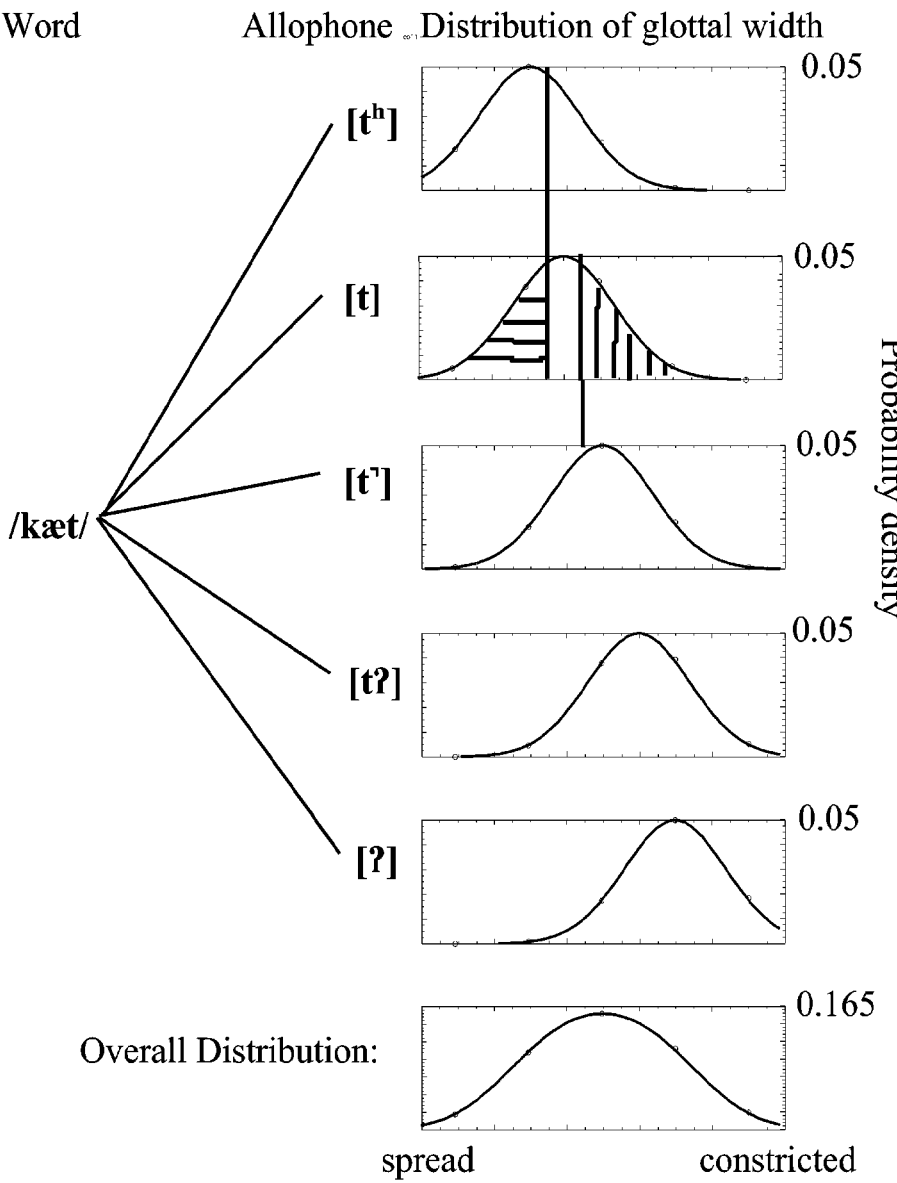


Figure 7B

B: Analogous situation involving heavily overlapped distributions which sum to a unimodal distribution. Striping highlights the instability of the second category. The horizontally striped region to the left of one indicated threshhold would always be classified as the allophone above. The vertically striped region to the right of the other indicated threshhold would always be classified as the allophone beneath. As a result of the low frequency with which the allophone is perceived, its peak would eventually fall below both of the competing distributions and it would never be perceived



the vertically striped area on the right hand side of the distribution will also be misperceived, as unreleased /t/s. As a result, a frequency differential will develop between /t/ and its competitors, and once the peak of the /t/ distribution falls below the intersection of the distributions for its competitors, it will never more be the most likely analysis of any production.

Realistic phonetic situations involve multiple phonetic dimensions rather than the one or two dimensions used in these tutorial figures. As noted above, the phonetic hyperspace has numerous dimensions, and languages differ in how these dimensions are bundled in the categorization system. A number of experiments have found that in cases in which a language uses multiple dimensions to cue a phonological contrast, there is poorer statistical separation in any one dimension than if that one dimension carried a phonological distinction on its own. For example, in Swedish, as in English, vowel quality and vowel length function together in the system of vowel categories. Finnish, in contrast, has a pure length distinction with extremely little impact on vowel quality. As reported by Engstrand and Krull (1994), Swedish vowels show more broad and overlapped phonetic distributions for length than Finnish vowels. In a similar vein, Ham (2001) found that geminate consonants are more differentiated (by length) in languages in which they are lexically contrastive than in languages in which they are phonologically predictable. If the geminate is phonologically predictable (e.g., from the stress pattern) then the phonetic cues for the predictor presumably function together with the consonantal length in cueing contrasts. Such results are predicted from the model sketched above. The statistical discrimination between competing categories operates over distributions in all relevant dimensions. If the production distributions for two categories overlap when all dimensions are taken into account, the categories will tend to merge. Categories that risk merger can be saved by any means that improves the relationship of spacing to variation. These include drifting apart in any single dimension, and moving into additional dimensions. One may also speculate that sheer improvement of accuracy would improve the survival rate of dense category systems, by improving the relationship of variability to spacing. However at present the longitudinal typological data needed to evaluate this idea are not available.

In summary, systems of categories reside in the perception-production loop in the speech community. Despite the continuous nature of the phonetic hyperspace, any individual language must have probability distributions over the parametric phonetic space which exhibit reasonably distinct modes. These distinct modes mean that the categories are well separated in comparison to the amount of variation for each; this relationship of spacing to variance is what permits the high levels of discrimination needed for reliable classification of each production. Even infants who do not have a well-developed lexicon can exploit the indirect reflexes of lexical regularities which result from the involvement of adult lexica in the perception-production loop at the community level. Phonetic modes do not correspond to phonemes in the traditional sense, but do provide a categorization of the signal. The infant can exploit these modes to initiate a system of phonetic encoding, and a statistical schema for how this may be accomplished also supports incremental updating on the basis of further experience. Incremental updating is critical to explain how much and for how long children's speech processing performance improves.

3.6
Internal lexical and phonological feedback

The previous sections have described how sharp, discriminable categories can evolve from perceptual input only, given the feedback which exists through speech communication in the community. However, there are some findings on speech perception which are difficult using only bottom-up category formation and community-level feedback. These suggest an influence of the lexical patterns on phonetic encoding in adults. Specifically, there is some evidence that general properties of the lexicon, or constraints in the phonological system, help to refine the speech encoding as the system matures. Similar conclusions are reached in Nittrouer (1996) and Boersma, Escudero, and Hayes (2003).

A number of experiments indicate that adult speakers are insensitive to cues which always, or else never, occur in their native language. For example, listeners of Japanese never encounter F3 values as low as those in English /ɹ/. This leads to a deterioration of their ability to perceive the third formant at all. A similar result is reported for stress in an experiment by Dupoux et al. (1997). Speakers of French, whose native language has completely invariant location of stress with respect to the word boundary are unable to perceive the (variably located) stress of Spanish.

In information theory, ubiquity and nonexistence have in common the characteristic of uninformativeness. Therefore, results like those just mentioned have been interpreted by some (such as Lahiri & Marslen-Wilson, 1991) as providing evidence for underspecification theory, according to which uninformative properties are simply omitted from lexical representations. This conclusion is at odds with findings summarized above which show lexical representations to include a considerable amount of subphonemic detail, as well as with various psycholinguistic results in Broe and Pierrehumbert (2000). It also creates serious technical problems for formal modeling, as discussed in Steriade (1995). However, the findings can also be understood in another way. The Japanese speaker's lack of experience with low F3 values entails that a statistical mode over that region of the phonetic space would never develop. With regard to invariant stress, there are distinct phonetic modes reflecting the phonetic correlates of stressed versus unstressed syllables only if these phonetic correlates are tabulated without regard to context. Contextualizing the phonetic parameters to the word boundary again yields the situation where there are no distinct modes (since the stress is invariably present, or invariably absent, for each given context). Since the contextualization system (or prosodic parser) is the only aspect of the system that sees two modes, the prediction is that it will categorize on stress and use this categorization to distinguish contexts — in this case, to locate word boundaries. Thus, we see that analysis of statistical modes in and of itself incorporates a notion of contrastiveness, while still retaining the explicit and detailed representation of variation which seems to be required.

A related case is that in which information appears to be ignored because it is too variable to be useful. This situation has been extensively studied in the sociolinguistics literature under the rubric of "near-mergers," reviewed in Labov (1994). Among other examples, he discusses the production and perception of *source* and *sauce* in New York City. In this r-less dialect, such words pairs have different, but poorly separated, distributions of vowel quality. Speakers who pronounce the two words

differently cannot, however, reliably judge which word they have heard. A plausible explanation of this apparent dissociation between production and perception is that speakers have learned not to pay attention to phonetic cues that are extremely variable in the speech community, and which therefore do not distinguish these words on the average. A particularly striking case of this effect is found in Janson and Schulman's (1983) study of bilingual speakers of northern Swedish and English. These speakers, unlike Stockholmers, distinguish /ɛ/ from /æ/ in pronunciation. In perception, they can distinguish these vowels when the experiment is conducted in English (a language in which the vowels are reliably distinct). But they cannot perceive the difference when the same experiment is conducted in Swedish (a language in which the historical merger in the dominant dialect renders the distinction unreliable). The performance in English provides prima facie evidence that the /ɛ/-/æ/ distinction is within the encoding capabilities of the listeners. Their lower performance when tested in Swedish accordingly indicates situationally induced downweighting of cues which do not reliably distinguish amongst words of Swedish. Such downweighting makes sense under the common understanding that attention is a limited resource which is functionally allocated on a situation-by-situation basis.

Apparent lexical effects are also seen in the results of an experiment by Hay, Pierrehumbert, and Beckman (in press). They report transcription data and well-formedness judgments for nonsense words involving nasal-obstruent clusters of varying frequency, with frequency established from type statistics over monomorphemic words. To control for acoustic properties, clusters were created by cross-splicing. Clusters which are infrequent or impossible within monomorphemic words (such as /np/) were either misperceived as acoustically similar likely clusters (e.g., /mp/), or if accurately perceived, the words were rated as if they contained an internal word boundary. The viable perceptual interpretations of /np/ were thus /mp/ and /n#p/; the interpretation /np/, lacking an internal word boundary, was not apparently available. It should be noted that the /mp/ interpretation is immediate and vivid, much like the visual perception of a Necker cube as one or another type of physical object.

This pattern of results can be readily explained in a model if adult listeners exploit lexical type statistics in parsing the speech stream and estimating the locations of possible word boundaries. Note that this feedback depends crucially on the morphosyntactic analysis of entries in the lexicon. In addition to monomorphemic words, the lexicon also includes many compounds, fixed phrases and frequent collocations. These stored complex forms contain internal word boundaries, which are reflected in their prosody, phonotactis, and allophony, as discussed above in connection with the example *blindfold*. Thus, the distinction between segmental sequences which do and do not occur in the lexicon is not very sharp; any segmental sequence that occurs at all in speech could occur in the lexicon in some or another complex form. The distinction between segmental sequences which occur in monomorphemic words and those which do not is much sharper, and this distinction is what eliminates the parse of /np/ without a word boundary as a viable percept.

Clearly, refined lexical type statistics cannot be computed from bottom-up analysis of the speech stream. They depend on the morphosyntactic decomposition of words, which in turn depends on the type statistics in the lexicon as well as on the application of syntactic and semantic representations. This level of analysis exceeds

the capabilities of infants, since refined type statistics require a large lexicon, and syntactic and semantic development unfolds over many years.

In summary, this tour of possible lexical effects on speech perception provides some evidence that the adult system has gone beyond what bottom-up projection of categories from modes in the phonetic distributions alone could provide. Bottom-up projection of modes, combined with feedback via speech communication in the population, can go surprisingly far in explaining basic observations about contrastiveness. However, it appears unable to explain the readjustment of cue weighting in response to situational differences in cue reliability, as in Janson and Schulman (1983). It also appears unable to explain the application of lexical type statistics by adults to word boundary perception (Hay et al., in press). The results of Dupoux et al. (1997) on stress deafness are consistent with those of Hay et al. (in press), because stressed syllables would project a distinct phonetic mode when viewed from the perspective of running speech. The unavailability of this mode only ensues when the perception is contextualized by the word boundary. The clear implication is that in languages with invariant stress, stress deafness will increase from infancy to adulthood as the concept of a phonological word develops and is used to improve the detection of word boundaries in speech.

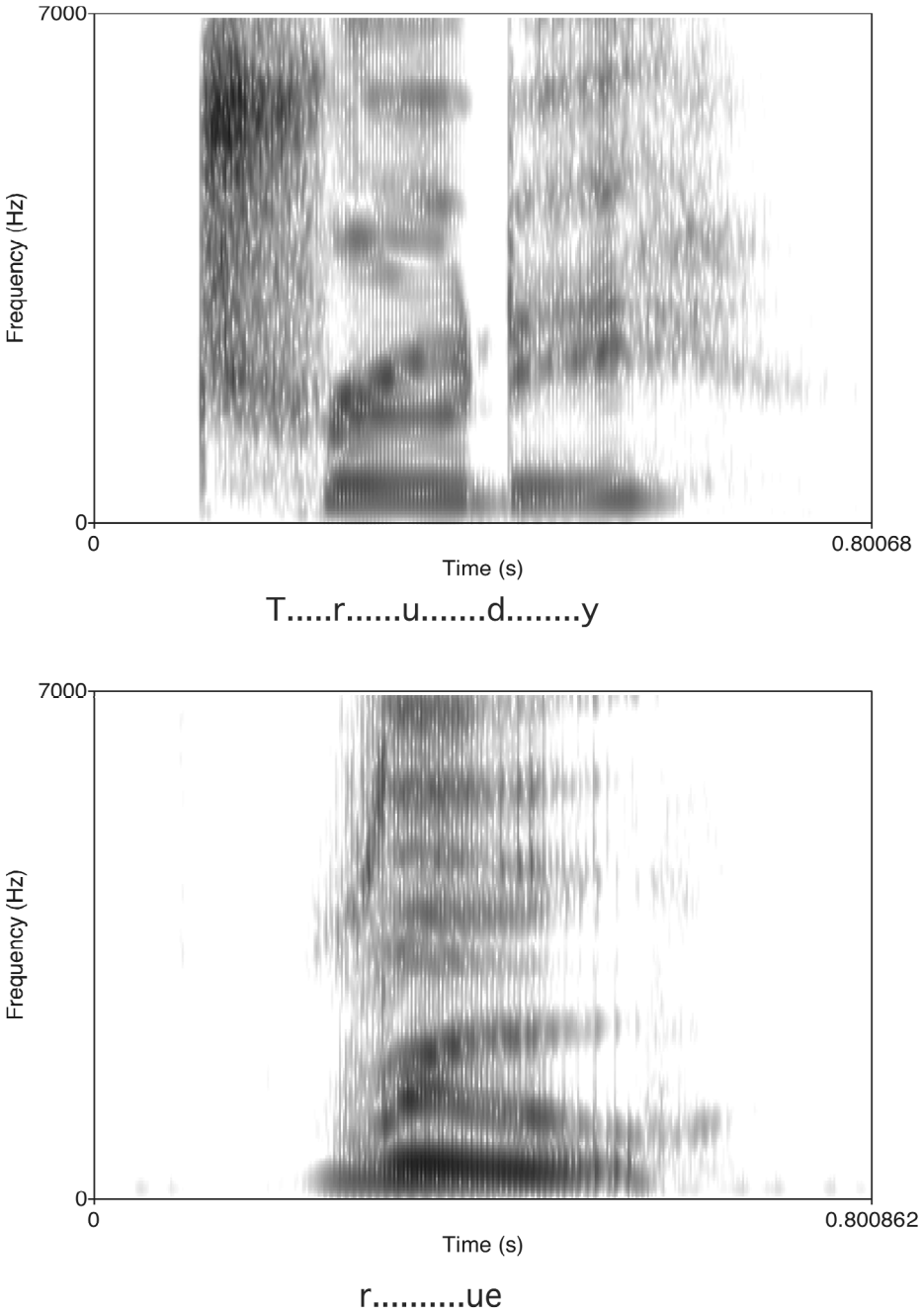
4 Bootstrapping. Confluence

In summary, the engine of adult speech perception appears to be positional segmental variants. The density distributions of these variants are plausibly initiated on the basis of modes in the statistical use of the phonetic space, and updated over a long period of subsequent experience. The modes are only distinct if the available set of segmental contrasts is contextualized by the prosodic structure. Therefore, prosodic encoding must develop at the same time or prior to segmental encoding, and swift and robust prosodic perception is crucial to the speed and accuracy of speech perception in adults. Statistical modes in the overall phonetic patterning for the language can go far in initiating the system, but there is evidence that the mature system has been sharpened by feedback from the phonological grammar, arising as generalizations over the lexicon. The ability to initiate the system bottom-up, and refine it using feedback, depends, I will argue, on subtle confluence across levels of representation. These confluences are key characteristics of human language, distinguishing the systems we find from alternative systems which are mathematically conceivable but unnatural.

The general picture I have sketched is consistent with a large body of work in speech engineering, which has long sought a passage between the Scylla of phonemic transcription and the Charybdis of spectral template matching. It has been known at least since Cole and Jakimik (1980) that lexical access from a phonemic transcription of the speech signal does not perform well. It is inaccurate, because phonemes cannot be reliably recognized bottom-up. It creates spurious ambiguity because the phonemic transcription loses information which is crucial for accurate identification of words. This point is illustrated in Figure 8, showing spectrograms for the words *Trudy* and *rue* (/tɹudi/ and /ɹu/). Though *rue* is phonemically embedded in *Trudy*, it is not phonetically embedded, and there is no subpart of *Trudy* which sounds like

Figure 8

Comparison of spectrograms for *Trudy* and *rue*. Even if *rue* is phonemically embedded in *Trudy*, no subpart of *Trudy* can be perceived as *rue*



rue. As shown in the spectrogram, there are obvious differences in the initial portion of the /ɪ/, as well as in the trajectory of the second formant.

Partially in response to such problems, Klatt (1980) proposed a model of lexical access based on direct spectral template matching. However, this approach runs afoul of phenomena such as those illustrated in Figure 9. Figure 9A is a natural recording of the word *fest*, and 9B, of the word *best*. As is obvious, a perfect example of the word *best* can be excised from *fest* by cutting off all but the last bit of the fricative region. If every single centisecond in the signal is treated as a possible word beginning, one finds many embeddings which do not appear to be psychologically relevant. Thus, this approach, too, leads to spurious ambiguities. The problem arises because there is no treatment of prosodic structure and context, so that all phonetic alternatives are active competitors at all time points. Further problems traceable to the lack of contextual indexing include poor handling of durational cues, a poor ability to generalize to new voices, and the need to learn systematic patterns of variation afresh for each individual word.

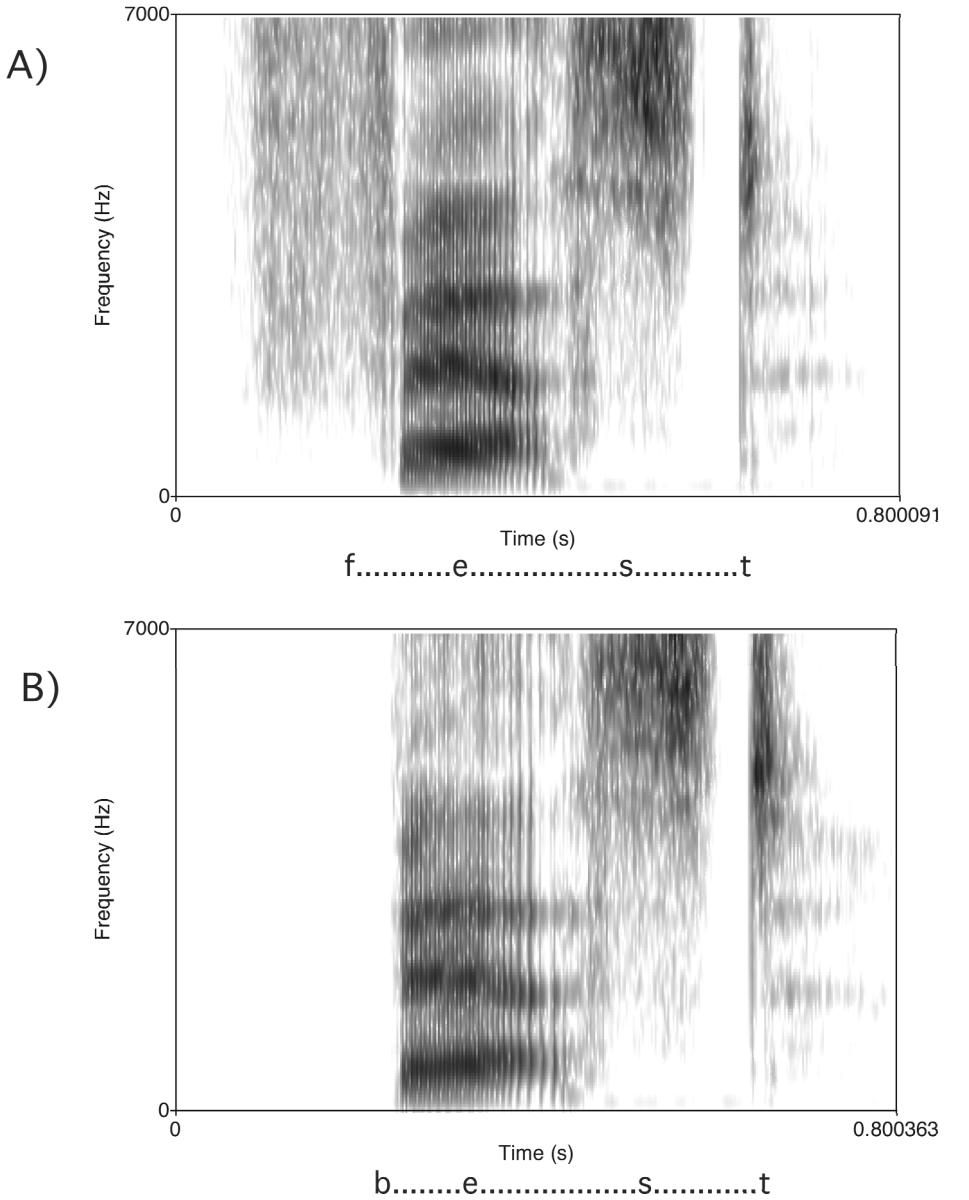
Rejecting both bottom-up phonemic transcription and direct phonetic parameter matching, my position here is closest to the line of work represented by Church (1987), Lehiste (1960), Nakatani and Dukes (1977), Shafran and Ostendorf (to appear), and Shafran, Ostendorf, and Wright (2001). Lehiste (1960) and Nakatani and Dukes (1977) both note that word boundaries are much better marked at the allophonic than at the phonemic level. However, they provide no formal architecture for managing allophones in relation to hierarchical prosodic structure. Church (1987) takes a major step in this direction by developing formal methods to handle the relation of allophones to syllables. Shafran and colleagues go much further by carrying out a large-scale cluster analysis of phonetic properties in relation to a large set of prosodic variables.

Though such research in speech signal processing and machine learning bodes well for models of linguistic bootstrapping, it still falls short of a complete model. Interactions of prosodic structure and phonetic categorization have only been explored in supervised learning models. That is, the starting point for the statistical analysis is a large tagged database in which the prosodic structure is already coded. Results on unsupervised cluster analysis, such as Kornai (1998), are more limited. They deal with segmental inventories in a single context and do not explore the variation which occurs across contexts. I have suggested above that phonological knowledge is initiated bottom-up, but is subsequently updated and refined with lexical feedback. In machine learning terms, this means the system shifts from unsupervised learning to supervised learning. A key challenge for a bootstrapping theory is explain how this shift is possible. Why does the results of unsupervised learning provide an adequate platform for supervised learning? How can supervised learning set in without a radical reorganization of the system?

I take up these questions in relation to phonotactic learning (i.e., learning of what sequences of segments are more and less well-formed within words). Though this does not exhaust the issue of phonological bootstrapping, it is an important locus of the interaction between segmental learning and word boundary detection in speech perception. Phonotactic cues are known to be used for segmentation both

Figure 9

Comparison of spectrograms for *best* and *fest*. *Best* is not phonemically embedded in *fest*, but a subpart of *fest* provides a perfect example of *best*



by adults (McQueen, 1998) and by infants as young as nine months (Mattys & Jusczyk, 2001; Mattys et al., 1999.) In addition, a large number of studies reviewed in these papers show that infants prefer words with likely phonotactics and that

likely phonotactics also enhance lexical access in adults. (Jusczyk, Luce, & Charles-Luce, 1994; Vitevich & Luce, 1998.)

In the experiments by Mattys et al. (1999) and Mattys & Jusczyk (2001) as in most related experiments on infant speech perception, probabilities of segmental combinations were computed with respect to phonemes or phoneme sequences. Under the assumptions I have sketched, the interpretation of these results might seem problematic, because the encoding units claimed to be available prelexically are positional variants, which are less abstract than phonemes. Inspection of the materials in the experiments, however, suggests that this problem is illusory. The gross distinctions in frequency which are manipulated in the experimental designs would apparently have been the same if calculations were made on positional variants rather than on phonemes proper. For example, a /zb/ sequence (as in *Frisbee*) always contains a coda /z/ plus onset /b/. A phonemic combination such as /tw/, which is possible word-initially but not possible word finally (cf. *twill*, **litw*) would, when pronounced, also supply possible and impossible sequences of phonetic variants.

A deeper issue relates to the distinction between type statistics (statistics over the lexicon) versus token statistics (surface statistics in running speech). Lacking a lexicon, the new infant must begin with surface statistics. However, results such as Hay et al. (in press) and Bailey and Hahn (2001) indicate that type statistics are important in the adult system. Type and token statistics are not necessarily the same, as in the well-known discrepancy between the type and token frequencies for the phoneme /ð/. So, why are token statistics useful to the infant, providing an avenue to more mature use of type statistics?

An important observation related to this issue is advanced in Hay (2000). As she notes, it is a trivial formal exercise to design a language in which high frequency phoneme transitions (rather than low frequency ones) occur across word boundaries. For example, imagine a language with eight consonants (let's say, {p, t, k, b, d, g, f, s}) and five vowels {a, e, i, o, u}. Assume that all words have the structure C (V C)+, subject to the constraint that /t/ occurs invariably and only at word onsets, and /s/ occurs invariably and only at word ends. In this language, all word boundaries would be marked with the sequence /st/. Assuming typical patterns for the other phonemes, /st/ would be the only consonant cluster, the most frequent diphone, and a totally reliable cue to the word boundary. This is a simple and mathematically possible phonology. It is also nonhuman. Human languages tend to have maximal contrast sets (maximal statistical perplexity) at the word onset, whereas this artificial language has minimal perplexity at the word onset. Maximal perplexity has the consequence that segmental transitions across word boundaries tend to be low frequency. It thus supports an acquisition trajectory in which transitions with low surface frequency are reliable indices to boundaries, can be used in developing a lexicon, and can survive as cues in the adult system. The relationship between the surface statistics available to the infant and the location of perplexity in the lexicon provides the first example of a confluence which supports bootstrapping.

In this light, the specific results of Mattys et al. (1999, 2001) are somewhat bewildering. In designing their experimental stimuli, they controlled for the surface frequency of the crucial diphones in their stimuli. They varied only whether these

diphones were more likely to arise within words (e.g., /ŋk/), or between words (e.g., /ŋt/). The design constraint that clusters that are rare within-in word be reasonably frequent across a word boundary, and vice versa, effectively confined the experiment to the middle of the frequency range. It did not include clusters which are hugely frequent thanks to within-word constraints, since no clusters of comparable frequency across a boundary exist. Nor did it include clusters which are extremely rare in all positions, as these cannot display a big difference in frequency within and between words. One prediction from Hay's (2000) analysis is that even stronger effects would have been found in an experiment which used the full range of type frequencies. The fact that an effect was found, despite the highly controlled design, appears to suggest that even at nine months, infants have a large enough lexicon for feedback from the lexicon to the phonetic encoding system to have occurred. This suggestion must be viewed with caution, however. An alternative possibility is that the adult speakers who recorded the stimuli perceived word boundaries in relation to type statistics, and that infants were sensitive to the manifestations of these boundaries in the duration and formant structure. If so, the experiment speaks yet again to the exquisite sensitivity of infants to the phonetic distributions used to implement the adult lexicon.

The perceptual relevance of type statistics also has a second line of implications, which relates to what statistics matter in parsing the speech stream in the first place. Any statistic relates to a phonological descriptor, and it is estimated by finding the frequency with which that descriptor is met. Surveying the linguistic and psycholinguistic literature suggests that the relevant phonotactic descriptors have no preferred unit and cross-cut each other. Phonotactic constraints for which implicit stochastic knowledge have been demonstrated include constraints governing syllable onsets, syllable rhymes (Treiman, Kessler, Kneewasser, Tincoff, & Bowman, 2000) syllable junctures (Hay et al., in press), stress templates (Cutler & Butterfield, 1992), vowel projections (as in languages with vowel harmony, see Suomi, McQueen, & Cutler, 1997) and consonantal projections (as in languages with OCP effects on place or manner, see Frisch & Zawaydeh, 2001). Even the infant literature, which still lacks comparable typological coverage, shows that infants are sensitive both to segmental transitions and stress templates (Mattys et al., 1999; Mattys et al. 2001; and works reviewed there). Is it the case that relevant templates includes any arbitrary fragments of phonological description for which lexical statistics can be established? A unifying observation is that that the assortment of phonological constraints just described are all relatively simple or coarse-grained. Results on syllable onsets and rhymes have only been obtained for onsets and rhymes consisting of one or two segments. More complex structures, such as the ternary onset of *string* (/stɹ/) or the quaternary rhyme of *grind* (/aɪnd/) are apparently perceived or judged as a function of their subparts. Long distance constraints, such as vowel harmony and OCP effects, either target two segments separated by arbitrary intervening material, or else they abstract across phonemes, or both.

Pierrehumbert (2001b, 2003) undertakes to explain this observation on the basis of learnability considerations. She asserts that relevant statistics are those which can be effectively trained as type statistics given realistic levels of lexical development. This consideration places an effective limit on the complexity of descriptors which are

relevant. Words provide a much smaller sample size than word tokens in running speech, because the number of words in an adult lexicon is some three orders of magnitude smaller than the number of word tokens an adult has encountered. Given the small sample size, overly complex or detailed constraints are not reliably learnable across individual variation in vocabulary. Above, we saw that the perception-production loop in the speech community has the result that only robust category systems are viable. Here, we are applying the same line of reasoning to the use of phonotactic constraints in parsing the speech stream. Lack of agreement within the speech community in sequential encoding and decoding would lead to lexical access errors and failures of understanding, through mismarking or misperception of word boundaries. Thus the only viable systems are ones which are uniformly learned by speakers in a community.

This point is supported by a Monte Carlo simulation of vocabulary learning over the monomorphemes in CELEX. (The monomorphemes define the class of lexical items lacking internal word boundaries, see discussion above). One result of this calculation is that the relative goodness of nasal-obstruent clusters in English is learnable from 1/3 of the total word list, a realistic vocabulary level for late childhood. In contrast, some unrealistically complex and detailed constraints are not uniformly learnable, even though they can be stated as statistical regularities on the full word list.

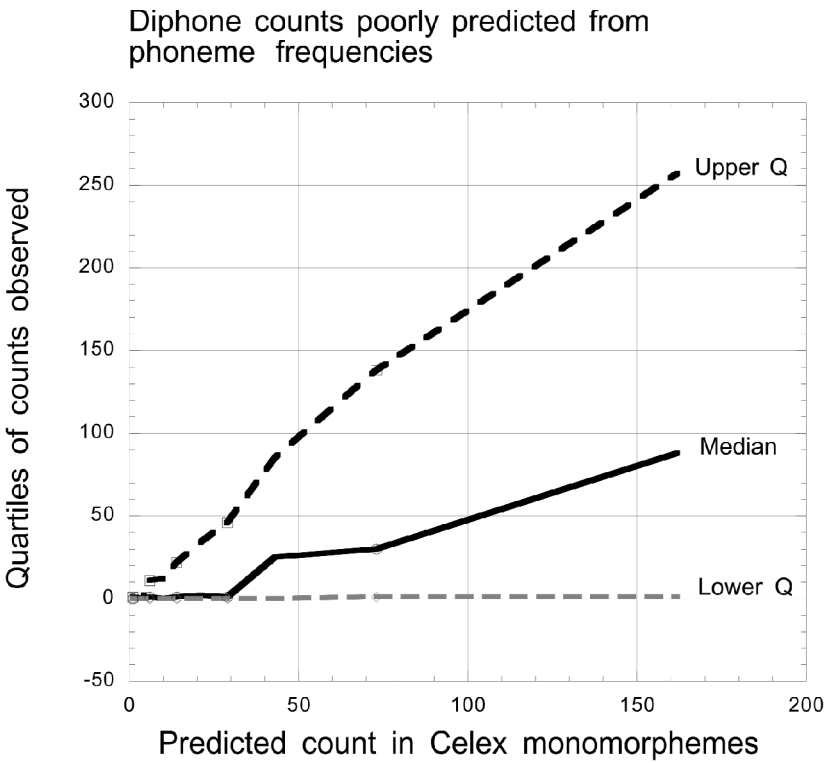
Further pursuing this line of reasoning, Pierrehumbert (2003) explores the extent to which diphone and triphone statistics can be, and must be, learned from the lexicon. (Calculations in this paper are also made using the CELEX monomorphemes.) The statistics can be learned if the lexicon is big enough to reliably estimate the frequency of each combination. They must be learned if the frequency of the sequence cannot be effectively estimated from the independent combination of the subparts. Pierrehumbert finds that diphone statistics can be, and must be learned. In general, triphone statistics cannot be, and need not be, learned.

Diphone statistics must be learned, because the prediction of diphone statistics from segmental statistics is not at all successful. Although diphones which are predicted not to exist (e.g., have expected counts of under 1.0 in a lexicon of the size of the Celex monomorphemes) are indeed rare this generalization only covers 48 cases, or 4% of the relevant set. Meanwhile, 40% of the combinations which are predicted to exist are actually absent. This is unsurprising given that basic facts of sonority sequencing are not captured. The overall trend is brought out by Figure 10, which shows the median, upper and lower quartiles for rate of occurrence in the lexicon as a function of the predicted rate of occurrence. Though the median is upwards trending, the lower quartile remains right at the bottom of the graph.

The situation is different for triphone statistics as estimated from diphone statistics. Ninety-five percent of the combinations which are expected to be absent are indeed absent, covering 42,776 of the relevant cases. Figure 11, constructed in the same way as Figure 10, also shows a much better prediction of rate of occurrence for combinations which do occur. Thus, it is not in general necessary to learn triphone statistics. It is of course possible that there is learning in connection with a certain number of highly overrepresented or underrepresented triphones. The overall approach does not predict a single preferred temporal scale for encoding, but rather a defining relationship between scale, detail, and frequency. Specifically, a positive or negative

Figure 10

Poor prediction of rate of occurrence of diphones from the frequencies of the component phones. Lines indicate the upper quartile, median, and lower quartile of actual frequency in relation to the frequency predicted from independent combination of the phone subparts. Reproduced from Pierrehumbert (2003)

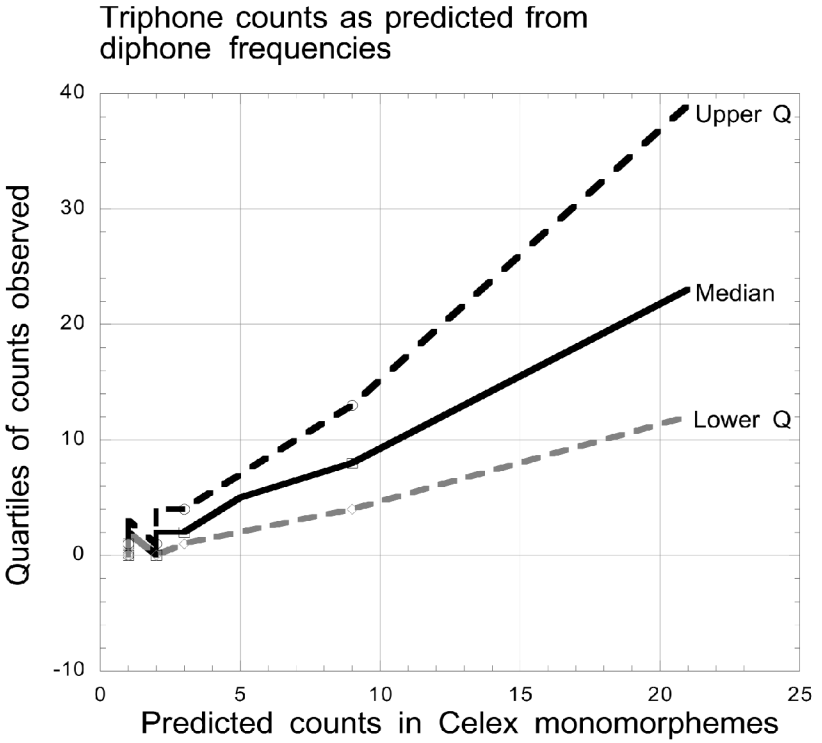


generalization about a long and detailed descriptor can be learned if there is a sufficiently large discrepancy between its expected frequency (as a random combination of subparts) and its actual frequency. This situation would arise if it fails to occur despite its hugely frequent components (a negative generalization) or the combination occurs fairly frequently (despite its infrequent components). See Manning and Schutze (2002) for full formal development of this point. In either case, the combination would be an outlier in the overall distribution of triphones, since most triphones are expected to be rare, and are indeed rare.

Diphone statistics can be learned from a lexicon of realistic size, since there are 1369 possible diphones, but more than 50,000 examples of diphones in the training set. In contrast, triphone statistics are not, in general, learnable from a lexicon of realistic size. Pierrehumbert (2003) reports 50,653 possible triphones, but just under 40,000 examples of triphones in the training set. With an average of less than one example per target triphone, it is obvious that the sample size is too small to estimate the target frequencies.

Figure 11:

Superior prediction of rate of occurrence of triphones from the frequencies of the component diphones. Reproduced from Pierrehumbert (2003)



These results were created using CELEX transcriptions, and they therefore inherit the segmentation assumptions of CELEX. A reviewer raises the question of whether the results would have been different if some of the more disputable segmentations had been handled differently — in other words, whether a child who made different choices about these marginal cases would have been in a different situation. Consider first the case of a diphone which might be analyzed as a triphone. For the diphone analysis to be entertained at all, the corresponding triphone would need a high degree of cohesion between two of its components, in the sense of having a strong statistical correlation and/or a distinctive articulatory program for the two parts taken together. These considerations all imply that it would be a high frequency triphone, hence a strong candidate for learnability. The case of a triphone which might be analyzed as a diphone is not problematic, because in that case its properties are claimed to be learnable. Thus, the general force of the analysis presented in Figures 10 and 11 is relatively insensitive to specific decisions about segmentation.

The privileged statistical status of diphones provides a further example of confluence across levels. There are phonetic reasons to expect that diphones are perceptually salient. As discussed in Stevens (1998), diphones, especially transitions from obstruent consonants to sonorant sounds, provide acoustic landmarks. The laws of physics mean

that these landmarks exhibit rapid spectral changes due to changing shape of the vocal tract. Pressure buildup during occlusion makes releases loud. Psychoacoustically, the automatic gain control of the ear makes vowel onsets perceptually salient after obstruents. Considerations on the articulatory side indicate that much language-particular detail about diphones must be learned. Overlapping of gestures leads to assimiliations, occlusions, and neutralizations, as discussed in (Browman & Goldstein, 1986, 1992). Different languages respond to these phonetic pressures differently, as discussed above under **Phonetic learning**. These considerations mean that the infant can begin with a short encoding span, and pay most attention to acoustic cues which are superficially apparent in the signal. The relationship of acoustics, articulation, and lexicon size means that this start does not lead the child astray about the adult grammar. On the contrary, it causes the child to initiate a large set of phonotactic descriptors which are exactly the sort needed to host type statistics as the system matures.

The discussion of these two cases does not, of course, exhaust the issue of confluences across levels which characterize human language and support phonological bootstrapping. I hope that similar connections will become evident for the acquisition of foot structure, word stress, and phrasal prosody and intonation.

5 Conclusion

In conclusion, learning a language involves learning a tremendous amount of phonetic detail, including language-specific categories of phonetic encoding and language-specific principles of sequencing, grouping, and prominence. The systems of categories which evolve through the perception-production loop in the speech community appear to have desirable properties of discriminability and robustness. The level of representation which has the best properties of discriminability and robustness will be the one that language learners acquire first through bottom-up analysis of the speech signal. I have argued that positional variants are stronger candidates for this level than the phoneme as traditionally conceived. This level of encoding supports initialization of the lexicon. There is some evidence that the phonetic encoding system is then refined, in response to the general patterns found in a more mature lexicon. This feedback from the phonological grammar would enhance capabilities for detecting word boundaries and as to increase attention paid to lexically contrastive phonetic cues. On-line manipulation of this feedback would permit listener adjustment to languages and dialects with different systems.

Human languages exhibit confluences across levels. These confluences make incremental development of the speech processing system possible. Levels of representation can be initiated bottom-up, and refined through further experience. Two examples of confluences were presented. One was the relationship of surface statistics to positional contrastiveness (or perplexity) in words. This relationship makes it possible for the infant to initially assume that low frequency transitions are word boundaries. The second was the relationship of acoustic landmarks to diphone statistics. This relationship brings together phonetic capabilities and pressures with the complexity of the grammar and the size of the lexicon. Both relationships only become evident when statistical issues in learnability are considered, and they help to delineate the nature of human language.

References

BAILEY, T. M., & HAHN, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, **4**, 568–591.

BECKMAN, M. E. (1986). *Stress and non-stress accent*. (Netherlands Phonetic Archives No. 7). Foris. (Second printing, 1992, by Walter de Gruyter.)

BECKMAN, M. E., & PIERREHUMBERT, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook* III, 15–70.

BECKMAN, M. E., & PIERREHUMBERT, J. (forthcoming). *A textbook in laboratory phonology*. Oxford: Basil Blackwell.

BEDDOR, P. S., HARNSBERGER, J., & LINDEMANN, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, **30**, 591–627.

BEDDOR, P. S., & KRAKOW, R. A. (1999). Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation. *J. Acoustical Society of America*, **106**(5), 2868–2887.

BERINSTEIN, A. (1979). A cross-linguistic study on the perception and production of stress. *UCLA Working Papers in Phonetics*, **47**, Dept. of Linguistics, UCLA.

BOERSMA, P., ESCUDERO, P., & HAYES, R. (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, August 3–9, 2003.

BRADLOW, A. (2002). Confluent talker-and listener-oriented forces in clear speech production. In C. Gussenhoven, & N. Warner, (Eds.), *Laboratory Phonology 7* (pp. 237–274). Berlin: Mouton de Gruyter.

BROE, M., & PIERREHUMBERT, J. (Eds.). (2000). *Papers in laboratory phonology V: Acquisition and the lexicon*, Cambridge U.K.: Cambridge University Press.

BROWMAN, C., & GOLDSTEIN, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, **3**, 219–252.

BROWMAN, C., & GOLDSTEIN, L. (1992). Articulatory phonology: An overview. *Phonetica*, **49**, 155–180.

BYBEE, J. (2001). *Phonology and language use*, Cambridge U.K.: Cambridge University Press.

CHURCH, K. W. (1987). *Phonological parsing in speech recognition*. Boston, MA: Kluwer Academic Publishers.

CUTLER, A., & BUTTERFIELD, S. (1992). Rythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, **31**, 218–236.

COLE, R., & JAKIMIK, J. (1980). A model of speech perception. In R. Cole (Ed.), *Perception and production of fluent speech* (pp. 133–164). Hillsdale, NJ: Lawrence Erlbaum.

CRYSTAL, D. (1969). *Prosodic systems and intonation in English*. London: Cambridge University Press.

DAHAN, E., MAGNUSON, J. S., TANENHAUSE, M. K., & HOGAN, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. In J. M. McQueen, & A. Cutler, (Eds.), Spoken word access processes. Special Issue of *Language and Cognitive Processes*, **16**(5–6), 507–534.

De JONG, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J. Acoustical Society of America*, **91**(1), 491–504.

DUPOUX, E., PALLIER, C., SEBASTIAN-GALLES, N., & MEHLER, J. (1997). A destressing "deafness" in French? *Journal of Memory and Language*, **36**, 406–421.

ENGSTRAND, O., & KRULL, D. (1994). Durational correlates of quantity in Swedish, Finnish, and Estonian: Cross-language evidence for a theory of adaptive dispersion. *Phonetica*, **51**, 80–91.

FLEGE, J. E., & HILLENBRAND, J. (1986). Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. *J. Acoustical Society of America*, **79**(2), 508–517.

FRISCH, S. A., & ZAWAYDEH, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, **77**, 91–106.

GERFEN, C., & BAKER, K. (in press). Production and perception of glottalized vowels in Coatzospan Mixtec. *Journal of Phonetics*.

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. Paper presented at *Spoken Word Access Processes*, May 2000, Max Planck Institute for Psycholinguistics, Nijmegen.

GRØNNUM, N. (1992). *The groundworks of Danish intonation. Museum Tusculanum Press*. University of Copenhagen.

HAJEK, J., & MAEDA, S. (2000). Investigating universals of sound change: The effect of vowel height and duration on the development of distinctive nasalization. In Broe & Pierrehumbert (Eds.), 52–69.

HAM, W. (2001). *Phonetic and phonological aspects of geminate timing*. Routledge.

HAY, J. B. (2000). *Causes and consequences of word structure*. Ph.D. dissertation, Northwestern University.

HAY, J. B., PIERREHUMBERT, J., & BECKMAN, M. E. (in press). Speech perception, well-formedness, and the statistics of the lexicon. In R. Ogden, J. Local & R. Temple (Eds.), *Papers in Laboratory Phonology VI*, *Cambridge University Press*. Cambridge U.K.

HAZAN, V., & BARRETT, S. (2000). The development of phonemic categorization in children aged 6 to 12. *Journal of Phonetics*, **28**, 377–396.

HILLENBRAND, J., GETTY, L. A., CLARK, M. J., & WHEELER, K. (1995). Acoustic characteristics of American English vowels. *J. Acoustical Society of America*, **97**, 3099–3111.

HILLENBRAND, J., & HOUDE, R. (1996). Role of F0 and amplitude in the perception of intervocalic glottal stops. *J. Speech and Hearing Research*, **39**, 1182–1190.

HUSSAIN, S. (1997). *Phonetic correlates of lexical stress in Urdu*. Ph.D. dissertation, Dept. of Communication Sciences and Disorders, Northwestern University.

HUSSAIN, S., & NAIR, R. (1995). Voicing and aspiration contrasts in Hindi and Urdu, *Papers from the 31st. Meeting of the Chicago Linguistics Society*. Chicago: University of Chicago.

JANSON, T., & SCHULMAN, R. (1983). Nondistinctive features and their use. *Journal of Linguistics*, **19**, 252–265.

JOHNSON, K. (1997). Speech perception without speaker normalization. In Johnson & Mullennix (Eds.), *Talker variability in speech processing* (pp.145–166). San Diego, Academic Press.

JUSCZYK, P. W., LUCE, P. A., & CHARLES-LUCE, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, **33**, 630–645.

KLATT, D. H. (1980). SCRIBER and LAFS: Two approaches to speech analysis. In W. A. Lea (Ed.), *Trends in speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.

KOCHETOV, A. (2002). *Production, perception, and emergent phonotactic patterns: A case of contrastive palatalization*. New York: Routledge.

KOHLER, K. J. (1987a). Categorical pitch perception. *Proceedings Eleventh International Congress Phonetic Sciences* (Talliin).

KOHLER, K. J. (1987b). The linguistic functions of F0 peaks. *Proceedings Eleventh International Congress Phonetic Sciences* (Talliin).

KORNAI, A. (1998). "Analytic models in phonology", J. Durand & B. Laks (Eds.). *The organization of phonology: Constraints, levels and representations* (pp. 395–418). Oxford U.K.: Oxford University Press.

LABOV, W. (1994). *Principles of linguistic change: Internal factors.* Oxford U.K.: Blackwell.

LAHIRI, A., & MARSLEN-WILSON, W. (1991). The mental representation of a lexical form: A phonological approach to the recognition lexicon. *Cognition*, **38**, 245–294.

LEHISTE, I. (1960). An acoustic-phonetic study of internal open juncture. Supplement to *Phonetic*, 5.

LEHTONEN, J. (1970). Aspects of quantity in Standard Finnish. *Studiia Philologica Jyvaskylaensia* VI, Jyvaskyla, Finland. Lexical neighborhoods? *Journal of Memory and Language*, **4**, 568–591.

LUCE, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: John Wiley and Sons, Inc.

LUCE, R. D., & GALANTER, E. (1963). Discrimination. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 191–243). New York: John Wiley and Sons, Inc.

MANNING, C., & SCHUTZE, H. (2002). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

MATTYS, S. L., JUSCZYK, P. W., LUCE, P. A., & MORGAN, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, **38**, 465–494.

MATTYS, S. L., & JUSCZYK, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, **78**, 91–121.

MAYE, J., & GERKEN, L. (2000). Learning phonemes without minimal pairs. *Proceedings of the 24th Annual Boston University Conference on Language Development*, 522–533.

MAYE, J., WERKER, J. F., & GERKEN, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, B101–B111.

McCARTHY, J. J., & PRINCE, A. (1993). Generalized alignment. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology*, 1993, 79–153.

McCLELLAND, J. L., & ELMAN, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1–86.

McQUEEN, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, **39**, 21–46.

MEHLER, J., JUSCZYK, P. W., LAMBERTZ, G., HALSTED, N., BERTONCINI, J., & AMIEL-TISON, C. (1988). A precursor of language acquisition in young infants. *Cognition*, **29**, 143–178.

NAKATANI, L. H., & DUKES, K. D. (1977). Locus of segmental cues for word juncture. *J. Acoustical Society America*, **62**, 714–719.

NESPOR, M., GUASTI, M. T., & CHRISTOPHE, A. (1996). Selecting word order: The Rhythmic Activation Principle. In Kleinhenz, U. (Ed.), *Interfaces in phonology. Studia grammatica*, **41**, 1–26.

NITTROUER, S. (1996). Discriminability and Perceptual Weighting of some acoustic cues to stop perception by three-year-olds. *J. Speech and Hearing Research*, **39**, 278–297.

NORRIS, D., McQUEEN, J., & CUTLER, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences* **23**(3), 299–324.

PETERSON, G. E., & BARNEY, H. L. (1952). Control methods used in a study of the vowels. *J. Acoustical Society America*, **24**, 175–184. Data downloadable from: <http://www-2.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/speech/database/pb>.

PHILLIPS, B. S. (1984). Word Frequency and the actuation of sound change. *Language*, **60**, 320–342.

PIERREHUMBERT, J. (1993). Prosody, Intonation, and Speech Technology. In M. Bates & R. Weischedel (Eds.), *Challenges in natural language processing* (pp. 257–282). Cambridge U.K.: Cambridge University Press.

PIERREHUMBERT, J. (2001a). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 137–157). Amsterdam: John Benjamins.

PIERREHUMBERT, J. (2001b). Why phonology is so coarse grained. In J. M. McQueen, & A. Cutler, (Eds.), Spoken word access processes. Special Issue of *Language and Cognitive Processes*, **16**(5–6), 691–698.

PIERREHUMBERT, J. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology VII* (pp. 101–140). Berlin: Mouton de Gruyter.

PIERREHUMBERT, J. (2003). Probabilistic theories of phonology. In R. Bod, J. B. Hay, & S. Jannedy (Eds.), *Probability theory in linguistics* (pp. 177–228). Cambridge, MA: MIT Press.

PIERREHUMBERT, J., & BECKMAN, M. E. (1988). *Japanese tone structure.* Linguistic inquiry Monograph 15, MIT Press, Cambridge.

PIERREHUMBERT, J., BECKMAN, M. E., & LADD, D. R. (2001). Conceptual foundations of phonology as a laboratory science. In N. Burton-Roberts, P. Carr, & G. Docherty, (Eds.), *Phonological knowledge* (pp. 273–304). Oxford, U.K.: Oxford University Press.

PIERREHUMBERT, J., & FRISCH, S. (1996). Synthesizing allophonic glottalization. In 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., & GROSS, P. (2003). Community phonology. Paper presented at the 39th Annual Meeting of the Chicago Linguistic Society, April 10–12, 2003. Movies (QuickTime). <http://www.ling.nwu.edu/jbp/conference.html>.

PIERREHUMBERT, J., & HIRSCHBERG, J. (1990). The meaning of intonation in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in communication*, Cambridge, MA: MIT Press.

PIERREHUMBERT, J., & STEELE, S. (1990). Categories of tonal alignment in English, *Phonetica*, 181–196.

SCOBBIE, J. M., TURK, A. E., & HEWLETT, N. (1999). Morphemes, phonetics and lexical Items: The case of the Scottish vowel length rule. *Proceedings of the XIVth International Congress of Phonetic Sciences*, **2**, 1617–1620.

SHAFRAN, I., & OSTENDORF, M. (to appear). Acoustic model clustering based on syllable structure. *Computer Speech and Language.*

SHAFRAN, I., OSTENDORF, M., & WRIGHT, R. (2001). Prosody and phonetic variability: Lessons learned from acoustic model clustering. *Proceedings ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding.*

STERIADE, D. (1995). Markedness and underspecification in the study of segments. In J. Goldsmith (Ed.), *A handbook of phonological theory* (pp. 114–175). Oxford U.K.: Basil Blackwell.

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. 313–332). Cambridge U.K.: Cambridge University Press.

STEVENS, K. N. (1998). *Acoustic phonetics.* Cambridge, MA: MIT Press.

STRANGE, W., & DITTMANN, S. (1984). Effects of discrimination training on the percpetion of /r-l/ by Japanese adults learning English. *Perception and Psychophysics*, **36**, 131–145.

SUOMI, K., McQUEEN, J. M., & CUTLER, A. (1997). Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Language*, **36**(3), 422–444.

SWINGLEY, D. (2003). (this issue). Phonetic detail in the developing lexicon. *Language and Speech*, **46**, 265–294.

TAFF, A., ROZELLE, L., CHO, T., LADEFOGED, P., DIRKS, M., & WEGELIN, J. (2001). Phonetic structures of Aleut. *Journal of Phonetics*, **29**, 231–272.

TREIMAN, R., KESSLER, B., KNEEWASSER, S., TINCOFF, R., & BOWMAN, M. (2000). English speakers' sensitivity to phonotactic patterns. In Broe & Pierrehumbert (Eds.), 269–283.

VIHMAN, M. (1996). *Phonological development: The origins of language in the child.* Oxford: Blackwell Publishers.

VITEVICH, M., & LUCE, P. (1998). When words compete: Levels of processing perception of spoken words. *Psychological Review*, **9**, 325–329.

WERKER, J. F., & STAGER, C. L. (2000). Developmental changes in infant speech perception and early word learning: Is there a link? (pp. 181–193). In M. Broe & J. Pierrehumbert (Eds.), Cambridge U.K.: Cambridge University Press.

WERKER, J. F., & TEES, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, **7**, 49–63.

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.

ZHANG, Y., KUHL, P. K., IMADA, T., IVERSON, P., PRUITT, J., KOTANI, M., & STEVENS, E. (2000). Neural plasticity revealed in perceptual training of a Japanese adult listener to learn American /l-r/ contrast: A whole-head magnetoencephalography study. *International Congress of Spoken Language Processing.* (Proceedings on CD).