Features as an emergent product of perceptual parsing:

Evidence from vowel-to-vowel coarticulation.

Bob McMurray
Dept. of Psychology
University of Iowa

Jennifer S. Cole
Dept. of Linguistics
University of Illinois

And

Cheyenne Munson
Dept. of Psychology
University of Iowa

Running Head: Feature Emergence by Parsing Processes

Corresponding Author
Bob McMurray
Dept. of Psychology
E11 SSH
University of Iowa
Iowa City, IA 52242
USA
319-335-2408 (voice)
319-335-0191 (fax)
bob-mcmurray@uiowa.edu

Features as an emergent product of perceptual parsing: Evidence from vowel-to-vowel coarticulation.

Introduction

Phonological features encode dimensions of lexical contrast among the sounds of a language, indexing the phonetic properties of speech sounds that can be used to distinguish words from one another. These phonetic properties can be quantified as continuous acoustic variables, but there are several indications that they function as discrete features in the phonological system. For example, plosive voicing in English (distinguishing word pairs like *bin* and *pin*) can be quantified as a continuous acoustic variable, Voice Onset Time (or VOT: the time difference between the release of the stop closure and the onset of laryngeal vibration). However, cross-linguistic studies of voicing (e.g. Lisker & Abramson, 1964; Keating 1984; Cho & Ladefoged 1999; Möbius 2004; Helgason & Ringen, in press) suggest that languages don't use the entire VOT continuum for the placement of voicing categories. Rather, languages exhibit clusters of VOT values that suggests two discrete features: something resembling [+/- voice] to handle pre-voicing contrasts, and something resembling [+/- spread glottis] to handle long-lag VOTs.

Sound change provides further evidence for discrete phonological features. A well-known example is German umlaut (Hock 1991:66). Umlaut arguably originates in a process of coarticulation in which a front suffix vowel alters the preceding vowel, causing a phonological sequence such as /u…i/ to be phonetically realized as [u_i … i] (where [u_i] represents a slightly fronted back vowel). At some point in the transmission of this pattern across speakers, it became codified as an alternation of the initial stem vowel: Instead of a fronted back vowel ([u_i]), the listener encodes a phonologically contrastive front vowel ([y]). Thus, the effect of coarticulation

shifts from that of variation along a continuous dimension to a discrete feature change. As this example illustrates, languages often change in a way that suggests a discrete feature system, despite the underlying gradient nature of the phonetic material through which words are realized.

Perhaps the most compelling evidence for discrete phonological features lies in their function as the units that encode contrast in lexical meaning. While the sound-meaning mapping is not entirely arbitrary (e.g. Monaghan, Chater & Christiansen, 2005), there is no debate about the fact that continuous gradations in acoustic cues or articulatory gestures do not map onto gradation of lexical meaning.[1] As any sports fan will tell you, there is nothing in between a *bunt* and a *punt*, no matter the VOT. This raises a fundamental question: how does the gradient nature of the articulatory and acoustic realization of language give way to the kind of discrete behaviors we see in phenomena like sound change and the encoding of lexical contrasts? If such discreteness is a defining property of phonological features, then in order to determine where features come from, we must first determine where discreteness comes from.

Although ultimately this question must be addressed with respect to both speech production (the ability to communicate lexical contrast) and speech perception (the ability to comprehend it), the present paper focuses exclusively on issues related to the mappings of acoustic form onto perception, asking how discrete phonological features can be perceived on the basis of speech input that exhibits variation along continuous phonetic dimensions. Perception poses a unique problem in that while discreteness would seem to be a necessary property for phonology, a growing body of work in online speech perception argues the opposite—that listeners are incredibly sensitive to fine-grained detail (Andruski, Blumstein & Burton, 1994; Utman, Blumstein & Burton, 2000; McMurray, Tanenhaus & Aslin, 2002; McMurray, Aslin, Tanenhaus, Spivey & Subik, in press) and that this sensitivity can facilitate online perception by

allowing the system to anticipate future material (Martin & Bunnel, 1981, 1982; Gow, 2001, 2003; Gow & McMurray, 2007), make use of non-contrastive detail (Salverda, Dahan & McQueen, 2003; Gow & Gordon, 1995; McLennan, Luce & Charles-Luce, 2003; Connine, 2004) and resolve prior ambiguity (Gow, 2002; McMurray, Tanenhaus & Aslin, under review). Thus, whatever mechanisms give rise to the discreteness necessary for phonology, the system must also preserve fine-grained or gradient detail for use in online perception.

We propose that discrete features emerge from a processing mechanism that *parses* the set of auditory cues that comprise the acoustic signal. Parsing effectively reduces the variation in the acoustic signal by attributing portions of the variation to properties of the context (broadly construed), and partialing out the continuous variation (sometimes described as "noise" in the signal) so that underlying features can be revealed (Fowler, 1984; Gow, 2003). We argue that parsing allows listeners to identify target sounds in the face of highly variable acoustic input, while simultaneously preserving fine-grained detail to aid in online perception. After the parsing process subtracts the effects of context from a particular acoustic cue, what remains is a more unambiguous encoding of the discrete phonological feature, while the portion subtracted away provides direct evidence for the context element and can contribute to its own featural representation. In this account, phonological features are *revealed* through the mechanism of parsing as it attributes specific qualities of the target sound to elements of the context.

This paper demonstrates the emergence of discrete features through parsing of fine-grained acoustic detail with a case study of vowel-to-vowel (V-to-V) coarticulation. In this demonstration, the acoustic parameters that encode the phonological height and backness features of a vowel are influenced by factors related to both the speaker and the local phonological context. We show that through the process of parsing, the highly variable acoustic

formant measures give way to discrete phonological features which allow correct identification of the phonologically contrastive vowel, while preserving sufficient acoustic detail to predict the context vowel in the next syllable with a high degree of accuracy.

The parsing approach developed here is not the first to address the question of how discrete features are obtained from the highly variable acoustic input nor is it the dominant paradigm for understanding the use of fine-grained detail. Section 1 discusses two historical approaches to discreteness in speech perception, as well as contemporary exemplar models as an alternative approach in which discrete elements are viewed as emergent properties of a richly detailed phonetic encoding of word forms. In Section 2, we describe parsing and our formalization of it as a simple linear model that can be applied to a concrete dataset. This model will be tested experimentally with an analysis of a dataset on vowel-to-vowel coarticulation, which is introduced in Section 3 as a prime example of context-induced variation. Section 4 presents our analyses of this dataset using our model of parsing. Finally, we return in Section 5 to the question of how features emerge, and conclude that not only does parsing result in better identification of the target sound, but also that features encoding the phonologically contrastive dimensions of a context vowel (e.g., height and backness) emerge from the same parsing process.

1. The search for discreteness in perception

With respect to perception, researchers have looked for the source of discreteness in phonology in two ways, seeking discreteness in either the acoustic signal itself (the search for acoustic invariance), or suggesting that perceptual processes impose it on the signal.

The search for acoustic invariance is perhaps the oldest question in psycholinguistics (e.g. Cooper, Liberman & Borst, 1951). Acoustic cues to phonological features, such as formants, F_0,

and VOT, tend to vary as a function of neighboring sounds, prosodic context, speaker, speaking rate and social factors, yet the premise of this undertaking was that if one looks closely enough, invariant acoustic cues can be seen amidst the noise of these extraneous factors.

Early approaches based on spectra at word onsets, for example (Stevens & Blumstein; 1978; Blumstein & Stevens, 1981; Kewley-Port & Luce, 1984) could discriminate place of articulation but could not handle positional variance (e.g. word-final vs. word-initial phonemes). Sussman and colleagues' work on locus equations (e.g. Sussman, Hilbert, Fruchter and Sirosh, 1998) also uncovered some invariant structure in the encoding of place of articulation in word-initial stops, but the equations show considerable overlap, suggesting that they may not be sufficient to distinguish place of articulation for an individual stop token (i.e., to serve as a feature). Vowels, in particular, present a problem for approaches to discreteness based on acoustic invariance. Studies like Hillenbrand, Getty, Clark & Wheeler (1995) and Hillenbrand, Clark & Nearey (2001) show quite clearly that the formant frequencies that distinguish vowels are heavily dependent on speaker and context, and that individual vowel categories overlap substantially. Thus, there is an emerging opinion that there may be no underlying invariant cues in the speech signal for certain contrasts (e.g., Lindblöm, 1996; Ohala, 1996).

Even approaches that maintain invariance have generally backed off from the strong claim that all phonological features have invariant acoustic cues. For instance, Stevens (2002) and Keyser & Stevens (2006) propose stable acoustic landmarks for major class features ("articulator-free" features) that signal the onsets or offsets of plosives, strident and non-strident fricatives, nasals, glides and vowels. Other contrasts (like place of articulation or voicing) however, are marked by "articulator-bound" features, which have a more complex set of acoustic

correlates, and which are acknowledged not to be invariant with respect to phonetic and prosodic context.

If discreteness is not to be found in the acoustic signal itself, it is possible that perceptual processes impose it on such a signal. One such process is *categorical perception* (reviewed in Liberman, Harris, Hoffman & Griffith, 1957; Repp, 1984; McMurray et al, in press). This perceptual process was suggested by the finding that listeners are unable to discriminate small acoustic differences that lie within a phonetic category (e.g. two /b/'s with different VOTs) while they are quite good at discriminating equivalently small differences that cross a boundary. This finding was taken to imply that early perceptual processing was finely tuned to discrete categories and was able to strip away unnecessary within-category variation.

While categorical perception is an attractive account of discreteness in perceptual processing, subsequent work has shown that it fails at multiple levels. First, it turns out that under many testing conditions listeners can discriminate within-category variants (Pisoni & Tash, 1974; Pisoni & Lazarus, 1974; Carney, Widen & Viemeister, 1977; Samuel, 1977; Massaro & Cohen, 1983; Gerrits & Schouten, 2004) and prior findings of poor within-category discrimination may have been the results of memory demands or biasing tasks. Second, vowels (Fry, Abramson, Eimas & Liberman, 1962) and to a lesser extent, fricatives (Healy & Repp, 1982) do not show categorical perception, even when tested under conditions that yield categorical perception in stop consonants. Finally, a host of more recent studies demonstrate that higher levels of processing (word recognition) are in fact sensitive to gradations within a category (Andruski, et al., 1994; Utman, et al., 2000; McMurray, et al., 2002; McMurray, et al., in press) so it cannot be the case that lower-level processes irretrievably eliminate a gradient representation of the signal.

Moreover, models of perceptual processing, such as categorical perception, that resolve the variability of the speech signal by discarding continuous variability are fundamentally at odds with a growing body of research suggesting that fine-grained gradient properties of the signal may facilitate upcoming processing (Gow, 2001, 2003; Gow & McMurray, 2007; Martin & Bunnell, 1981, 1982), or help resolve ambiguity that occurred in the past (Gow, 2002; McMurray, Tanenhaus & Aslin, under review). For at least some purposes, then, a discrete representation stripped of "low-level" detail may be suboptimal for perception. Any perceptual imposition of discreteness on the acoustic signal must do so in a way that also preserves fine-grained detail for this sort of processing.

Exemplar models (e.g. Goldinger, 1998; Pierrehumbert, 2003) offer a solution to the problem of deriving discrete categories while preserving phonetic detail. Such models posit that the system veridically stores (in memory) vast numbers of exemplars of the words it has been exposed to. These exemplars are stored with a very fine level of detail, and there is nothing in the encoding that differentiates non-contrastive detail (e.g. cues indexing speaker identity) from detail that cues phonological contrast. Despite this level of detail, discrete elements can emerge in these models in the form of generalization across this massively redundant set of exemplars (e.g., Lindblom, 2000). The question for exemplar models, then, is similar to the question posed in this paper: what are the perceptual or memory processes that perform this generalization? How do these processes both take into account the contextually-conditioned phonetic detail and also result in the identification of discrete phonological categories, such as segments or phonological features?

Despite the importance of this question, there have been surprisingly few formal approaches. The process of recognizing words in an exemplar system has been reasonably

spelled out (Goldinger, 1998), and there are suggestions that statistical or distributional learning mechanisms may give rise to categories in exemplar models (Pierrehumbert, 2003). However, there are no formal models that make specific claims about the origins of contrastive, discrete features from speech input containing contextual variation. Parsing, then may then offer such a mechanism, though as we will discuss in the conclusion it does not require an exemplar-based representation for words.

An additional challenge for exemplar models is to cope with patterned variation that arises from context that lies outside of the word. It is quite simple to see how variation that is conditioned by within-word context (e.g., coarticulation between vowels and consonants), would get encoded in the exemplars stored in long-term memory. But phonetic variation is often conditioned by elements beyond the word boundary. Phonetic detail conditioned by material beyond the word can not be associated with its triggering context (if the word is the unit of long-term storage), and so if left unparsed, such phonetic variability may actually compromise the contrast between the target word and its close lexical neighbors. An alternative model would be to encode words in long-term memory as underspecified along the dimensions affected by context outside the word. Such a model may not be able to harness such detail to its full advantage, and may even be unable to make lexical contrasts in some cases. However, work by Gow (2001, 2003; see also Gow & McMurray, 2007) demonstrates that listeners can take advantage of coarticulation involving place features to facilitate the perception of upcoming segments, even when the coarticulation occurs across a word boundary. Cross-word coarticulation may thus be a limiting case for exemplar models of perception, and so we focus on that pattern in the demonstration of the parsing model presented below.

In sum, the acoustic signal does not appear to cue a sufficient set of discrete features to support lexical contrast, nor does the perceptual system impose discreteness upon it by eliminating phonetic detail not relevant to making categorical distinctions. If discrete phonological features are to emerge from perceptual processes, we must look beyond acoustic invariance or categorical perception as the underlying mechanisms. Whatever perceptual process imposes discreteness, however, must also preserve a representation of the signal that includes fine-grained detail to facilitate on-line processing. Moreover, such a process must be able to cope with overlapping sources of variation in the signal, related to the phonological or phonetic context (e.g., coarticulation) and non-phonological sources (e.g., speaker), and including sources that lie outside the boundaries of the target word. While exemplar or episodic models can achieve this sensitivity to phonetic detail, it is not clear how discreteness emerges in the definition of a system of phonological contrasts, nor how such systems cope with variability across word boundaries. *Parsing* is a promising approach that may offer the ability to deal with all of these issues.

2. Parsing

Parsing is a perceptual process first proposed by Fowler (1984; Fowler & Smith, 1986) to deal with overlapping sources of variance in the speech signal, such as overlapping articulatory gestures. The idea is very simple: at any given point in the signal, the system assigns acoustic cues to causes. Fowler assumes these causes to be gestural; however, later instantiations of parsing (Gow, 2003) take a less specific stance, arguing simply that similar acoustic cues (e.g. lowered F1) are grouped via association with features like labiality or coronality. However, in both cases, because the causes can originate in the past or future (i.e., can precede or follow the

target sound), parsing can have very powerful results for speech perception.

For example, consider anticipatory vowel nasalization in English. Since English does not have contrastive vowel nasalization, oral vowels are often nasalized when they precede a nasal consonant. When the parsing process encounters a nasalized vowel, the nasal cues can be unambiguously assigned to an *upcoming* nasal gesture because they could not have arisen from the vowel itself (given the absence of nasal vowels in English). This has two useful consequences. First, it provides information that a nasal consonant is upcoming. Second, by assigning these cues to the nasal gesture, it removes them from consideration as part of the vowel, allowing the vowel to be perceived (correctly) as oral. This was indeed demonstrated by Fowler and Brown (2000) in their finding that nasal vowels sound more oral prior to a nasal consonant.

Parsing thus has the necessary properties to create discreteness while preserving gradiency. First, by removing the effects of nasalization from the vowel, it creates a more prototypical vowel with much of the "noise" removed. However, by assigning this gesture to a future segment it simultaneously is able to use the gradient coarticulation to do useful work. In a sense, by partialing out the variability in the input into a discrete category and a residual (the difference between the abstract category and the observed input), the underlying feature in the target emerges and the residual can then be used to identify other events.

This reframes the fundamental issue in speech perception. If we examined only a single feature at a time (e.g. the orality of the target vowel), we'd be faced with ambiguity and noise. However, by trying to identify both the target vowel and the subsequent context at the same time, we can simultaneously remove the nasality from the vowel (allowing its oral feature to emerge), and build evidence for a consonant (contributing to its nasal feature). Thus, while ostensibly

making the problem more difficult (by introducing simultaneous extraction of multiple features), parsing may actually solve problems that were previously insoluble.

Given its power, parsing has been proposed as a general process of speech perception that provides an explicit treatment of a number of coarticulatory phenomena: vowel-consonant coarticulation (Fowler, 1984), vowel-to-vowel coarticulation (Fowler & Smith, 1986), vowel nasalization (Fowler & Brown, 2000), F0 effects on vowels (Pardo & Fowler, 1997) as well as place and voicing assimilation (Gow, 2003; Gow & Im, 2002). However, these studies are all experimental studies of perception. At this point, there have been no systematic phonetic investigations examining whether parsing would in fact be a useful mechanism for coping with the variability actually found in large speech databases, particularly when we consider more than one source of variability simultaneously. This is in large part because of the lack of a formal or computational model of parsing—it would be difficult to answer this question without it.

We have developed a formal model of parsing using hierarchical linear regression. This model allows us to ask whether parsing processes can explain the emergence of features over a large and highly variable set of vowel productions. This surprisingly simple approach can ably model the two operations of parsing: feature identification and prediction of nearby context. Moreover, its generality allows us to ask whether similar processes could be useful for coping with variability due to non-phonological causes (e.g. speaker) and to examine the interaction of multiple sources of covariation simultaneously.

Linear regression assumes that the variability in a dependent variable (DV) can be described as simply the weighted sum of a set of independent factors. When these factors are dichotomous (e.g. if a vowel is high or not), then the weighting reflects the contribution of that

category to the continuous dependent measure. For example, if the DV is F1, the weighting on the factor (feature), *high*, would be the average change in F1 for *high* vowels vs. other vowels.

To use this to model features, we use the continuous acoustic cue (e.g. F1 frequency) as the dependent variable, and generate an equation that predicts its value as the sum of the contributions of a set of discrete features. F1, for example, will be affected by the speaker, the height of the vowel, the voicing of neighboring consonant and so on. Linear regression can easily compute the weighting for multiple sources of variance in a given dataset, as long as these features are known for each cue.

Linear regression can be used hierarchically, to systematically exclude the effects of one source of variance on a DV, and analyze the left over variability in the DV after the first source of variability has been removed (the residual). In a hierarchical regression, an initial simple model with only a few factors is first fit to the data. The residuals are then computed, and additional factors are added to determine if they are able to account for any additional variance, over and above the original model. In this way, we can first partial out the effects of speaker from a measure like F1. We can then analyze the residual and determine what other features affect the remaining variance. Importantly, we can also use the residual to identify features such as the target vowel, or features of the upcoming context.

Parsing predicts that as we partial out factors from the signal, the residual should contain a clearer and clearer instantiation of other underlying cues. These may be cues to the current segment (e.g. revealing a feature masked by variability) or cues to upcoming segments (making use of fine-grained detail to anticipate material). Thus, as variance is removed from the signal, the discrete underlying features are revealed.

This model assumes that sources of variability are additive and that the effect of features on an acoustic cues can be neatly described by this linear system. While more complex conceptualizations are clearly possible, the simpler model has some advantages. To the extent that a simple model like this can provide an appropriate characterization of the perceptual process, we may not want to posit anything more complex. Most importantly, this model can be quite straightforwardly implemented using standard statistical techniques to allow a comprehensive analysis over a corpus of data.

Given a regression implementation of this parsing model, testing is straightforward. We can apply the model to a body of phonetic measurements and make two specific predictions.

1) Identification of a target feature in the acoustic signal should be better after other sources of variation have been partialed out.
2) Partialing out some sources of variation should also improve the model's ability to make predictions about other sources of context, and ultimately to identify their underlying features.

For this initial foray into testing these ideas we chose to examine vowels, since (as discussed above) they present one of the most challenging domains in which to extract discrete features. Vowel-to-vowel coarticulation, in particular, offers an ideal domain for this undertaking. In V-to-V coarticulation, the height or backness of a vowel is influenced by a subsequent (or prior) vowel, typically across one or more consonants. Thus, vowel perception in the context of V-to-V coarticulation offers numerous sources of variability: speaker variability affects F1, the place and voicing of the following consonant will affect both F1 and F2, and the subsequent vowel will play a role. In this context, we can examine whether the parsing processes that cope with these sources of variation improve the ability to identify the features of

the target (first) vowel, and simultaneously leave enough information to predict its identity (or whether they improve the prediction). These factors allow us to test both aspects of the parsing process.

In addition, we examined V-to-V coarticulation across a word boundary, since in that context the coarticulation pattern is not part of the phonetic detail of an individual word form. Without an active process of perceptual parsing, simply lexicalizing this detail in an episodic representation would not permit either the use of context (outside the word) to recover the underlying feature, nor the use of the residuals to predict the next features.

While validating this broader undertaking requires both production and perceptual studies, this initial work focused on the production data alone. In particular, this provides the opportunity to ask whether parsing (instantiated in our model), in principle, can improve on the identification of underlying features and prediction of upcoming material, given a variable set of input in which multiple sources of variance are available. If this was not found to be the case (in this context), there would be no need to test actual listeners. Thus, in the next section, we offer a short discussion of the facts of V-to-V coarticulation, followed by a description of the dataset on which we base our analyses.

3. Vowel-to-vowel coarticulation as a test case for the parsing hypothesis

The present study applied a parsing analysis to vowel-to-vowel (V-to-V) coarticulation to test the hypothesis that parsing reduces variability in order to reveal the discrete units of lexical contrast. This represents a particularly challenging problem for parsing in that vowels exhibit variation due to coarticulation with the vowel in the following syllable as well as with the intervening consonant.

It is well known that vowel sounds exhibit a complex pattern of coarticulation, with local effects triggered by a neighboring consonant or vowel (Hillenbrand, Clark & Nearey, 2001) and long-distance effects from a vowel in an adjacent syllable, across an intervening consonant in VCV sequences (Öhman, 1966, Magen, 1997). In both cases, such coarticulation can be seen as a source of noise for vowel identification: when formant measures for a vowel phoneme are pooled across a variety of coarticulatory contexts, there is an increased variance in the formant values of the vowel (Manuel, 1990; Megan, 1997; Öhman, 1966; Recasens & Pallarès, 2000).

The increased acoustic variability of a vowel, when considered independent of its context, might be expected to contribute to an increase in perceptual confusion among contrastive vowels, especially in a language like English with a large vowel inventory and thus a densely populated vowel space. Yet there is evidence that listeners are able to compensate for the effects of coarticulation, parsing out the influence of coarticulation from the acoustic properties that cue distinctive vowel place features (Fowler, 1981, 1984; Fowler & Smith, 1986; Beddor et al., 2002). For example, Fowler & Smith (1986) show that when presented with pairs of CV_1CV_2 stimuli, listeners judge the two tokens of V_1 as similar even when they exhibit subtle coarticulatory differences, as long as the coarticulatory effect on each V_2 is appropriate for the given V_1 context vowel. In other words, in the appropriate contexts, listeners seem to parse out the portion of the variance in F1 and F2 (and possibly other acoustic parameters) that is due to coarticulation, and base their perception of the target vowel on the residual values.

However, V-to-V coarticulation does more than create ambiguity in the signal. In the case of anticipatory (i.e., regressive) coarticulation, the vowel in the earlier syllable shifts to become more similar to the vowel in the later syllable, and this shift provides a potential source of information that listeners use to infer upcoming material (Martin & Bunnell, 1982). Thus,

not only can parsing facilitate identification of the target vowel, but parsing also facilitates the prediction of the upcoming context vowel (Fowler, 1984). The strength of the prediction, and thus the potential usefulness of the parsed variance for predicting upcoming context, depends on the magnitude and consistency of anticipatory coarticulation in the language.

V-to-V coarticulation causes a shift in both the articulation and acoustic form of a vowel and is bidirectional, though the relative strength of carryover versus anticipatory effects vary in different languages (Beddor et al., 2002; Manuel, 1990; Öhman, 1966). Focusing now on anticipatory V-to-V coarticulation, prior studies on English show that coarticulation affects acoustic measures of F1 and F2, cues to vowel height and backness/roundness, respectively, which are shifted in the direction of the context vowel. Both stressed and unstressed vowels undergo coarticulation, though the effect on unstressed vowels is typically greater (Alfonso & Baer, 1982; Beddor et al., 2002; Fowler, 1981, 2005; Magen, 1997; Öhman, 1966; among others). And while V-to-V coarticulation is a significant source of variability for vowels, it is not the only source. There is also evidence of variation in vowel formants due to coarticulation with an upcoming consonant (e.g., Hillenbrand, et al., 2001; Öhman, 1966), and due to individual speaker characteristics (e.g., Hillenbrand et al., 1995). The interaction among these various sources of coarticulation has not been widely investigated.

Furthermore, while coarticulation is seen to be pervasive within syllables and words, none of these prior studies have assessed the form of V-to-V coarticulation across word boundaries. As discussed earlier, if coarticulatory variation effects only arise within words, mechanisms like parsing may be unnecessary to cope with the variability as well as take advantage of it. However, recent work that forms the basis for the present study (Cole et al, under review) shows that V-to-V coarticulation can be seen across word boundaries.

To summarize, there is ample evidence of vowel variability due to coarticulation, including anticipatory V-to-V coarticulation. Listeners appear to parse this variability, compensating for the influence of upcoming context, while at the same time using the parsed variance to predict that context. To better understand the potential benefit of parsing for speech perception, we turn now to our test case, applying parsing to the analysis of vowels that are coarticulated with the following C (within-word) and V (across a word boundary). We pose three questions. First, to what extent can variability of the F1 and F2 measures of vowels be reliably attributed to the upcoming phonological context, or to speaker voice characteristics? Second, is there a systematic pattern variation due to coarticulation from an upcoming source that crosses a word boundary? Third, can the variability of the target vowel that is due to coarticulation be used to make predictions about the upcoming vowel, in the following word?

We demonstrate the parsing analysis using a database from our previous acoustic investigation of V-to-V coarticulation (Cole et al., in review). This experiment was designed to test for effects of anticipatory coarticulation on vowels separated by a consonant and word boundary, using measures of the first two formants of naturally produced vowels in VC#V contexts across a variety of speakers. The data was collected from five males and five females (graduate or undergraduate students at the University of Illinois), all native speakers of English under 30 years old. The speakers produced carrier sentences that contained two-word test phrases with the *target* vowels (which we define as the vowel undergoing coarticulation) in the first word and the *context* vowel (the vowel triggering the coarticulation) in the second word. The vowels were separated by an intervening consonant at the end of the first word. For example, in the phrase *wet oxen* the /ɛ/ in *wet* was the target vowel and the initial /ɑ/ in *oxen* was the context vowel.

The target vowels were the two unrounded, central vowels (/ʌ/ and /ɛ/), which have the potential to show both height and front/back effects due to coarticulation. The context vowels were the three point vowels (/æ/, /ɑ/, /i/), which are maximally likely to induce coarticulation, and the matched target vowel (/ʌ/ or /ɛ/), which was expected to be neutral as a coarticulation trigger. The point vowel /u/ was excluded to avoid introducing rounding coarticulation into the design.

Test words were chosen that end in a plosive consonant. There were six plosives, combining voiced and voiceless features with three places of articulation (labial, alveolar, velar). These six consonants were combined with each target vowel in the first word[2]. There were a total of 48 phrases recorded by each speaker (2 target vowels × 4 context vowels × 6 consonants). These phrases are listed in Table 1. Each of the 48 phrases was repeated three times for a total of 144 trials.

For each of the target and context vowels F1 and F2 were measured at the midpoint with an LPC analysis and outliers were corrected based on visual examination of the spectrogram. Formant frequencies were coded in units of bark. 22 trials (out of 1440) were eliminated because of speech errors. An additional 18 trials were eliminated because participants pronounced *ecologist* with a schwa rather than the desired context vowel /i/. A total of 1400 trials were included in the analysis.

4. Testing the Parsing Model

As it has been described, parsing offers two operations during speech perception. First, by partialing out the effects of sources of variation (e.g. coarticulation), it can reveal underlying

features in the signal. Second, the residuals of this process can then be used to predict upcoming material. Thus, our analysis proceeds in two steps. First, we illustrate parsing's ability to uncover features by applying our model to the problem of identifying the target vowel. This is described in some detail in order to demonstrate the operation of the model. Next, we show that the ability of the exact same model to account for overlapping variance dramatically improves the prediction of the upcoming context vowel.

4.1 Uncovering Features of the Target Vowel.

The first analysis examined F1 and F2 and their ability to discriminate the target vowel as /ʌ/ or /ɛ/. We compared three different models. The first modeled the case of no parsing whatsoever—a sort of baseline upon which to evaluate the subsequent models. The second parsed out the effects of speaker gender from F1 and F2 before using these cues to categorize the target vowel. The third parsed out speaker gender as well as the specific speaker prior to categorizing the target vowel.

In the baseline model, raw F1 and F2 values (and an interaction term) were entered into a linear regression with target vowel (/ʌ/ or /ɛ/) as the sole predictor. Target vowel significantly affected F1 ($F(1, 473)=4.8$, $p=.029$) but only accounted for 1.0% of its variance. There was a much larger affect on F2 ($F(1, 473)=323.9$, $p<.0001$), with target vowel accounting for about 40.6% of the variance. This tells us that the two target vowels differ in terms of both F1 and F2 (and much more so for F2). However, it does not tell us how useful these raw formant values would be for identifying the vowel. That is, given the pattern of variability in F1 and F2, how many of the tokens in our dataset could be correctly identified?

To solve this problem, we used logistic regression as a simple model that maps F1 and F2 jointly onto a discrete categorical output (the correct vowel). This is similar to the approach of Jiang, Chen and Alwan (2004) who used logistic regression to determine the sufficiency of single cues to voicing. We adopt the same approach to look at both cues (and their interaction) simultaneously. The logistic regression model, in particular, allows us to compute percent correct, as well as whether each cue was used to make a given distinction. For the present analysis (the base model), the predictors are raw F1 and F2 and the dependent measure is a dichotomous /ʌ/ or /ε/ decision. However, in the subsequent models we will also use the parsed values (residuals from the linear regression) as input to this logistic regression. We can then evaluate the percentage of correct identifications of the model as a function of what was parsed.

To evaluate the baseline model, we used logistic regression alone on unparsed formant values (i.e., the linear parsing model was not used). Raw F1 and F2 values were entered directly into a logistic regression as predictors, with the target vowel's identity as the dependent variable. Overall, the model performed quite well, with classification accuracy at 90.5% correct. Each term significantly contributed to the classification individually: F1 (Wald(1)=18.2, $p<.001$), F2 (Wald(1)=4.1, $p=.042$) and the interaction (Wald(1)=13.1, $p<.001$). This high performance was expected – the model only had to make a two-alternative decision, and it was basing this decision on the strongest cues available (unlike the predictive task in the next mode, in which it must use formant-cues that occur *prior* the vowel being predicted).

Given this baseline level of performance, the next model asked if parsing variance due to the gender of the speaker could improve the model. First, we used ordinary linear regression to parse out the effects of gender on F1 and F2. This single factor significantly accounted for 63.2% of the variance in F1 ($F_{change}(1,473)=811.5$, $p<.0001$) and 35.9% of the variance in F2

($F_{change}(1,473)=264.7$, $p<.0001$) (see Tables 2 and 3). In order to parse out this factor from the continuous cues, we simply computed the difference between each F1 or F2 value from the F1 or F2 predicted by the linear regression line. Since the predictor in this case is categorical, this is equivalent to simply subtracting the mean F1 or F2 for each group (male and female) from each value. For example, the mean F1 (across vowels and contexts) for females was 6.41 bark (SD=.42 across speakers) and 5.28 bark (SD=.25) for males. Thus, if a given data point had a raw F1 of 6.0 bark, if it came from a female it was recoded as -.41 bark (low for a female), but if it was generated by a male it was recoded as +.72 (high for a male). These differences were the residuals, after the effect of gender on F1 and F2 had been removed.

These residuals were then entered as the independent variables in the logistic regression described above. This model was somewhat better, averaging 91.4% correct. As before all three covariates significantly contributed to the classification (F1: Wald(1)=12.6, $p<.001$; F2: Wald(1)=77.0, $p<.001$; F1 x F2: Wald(1)=14.2), only this time, F2 was a much stronger contributor than before (as evidenced by its increased Wald statistic).

If simply partialing out gender could improve performance, we next asked if knowing the individual speaker could further improve this. Thus, we added individual speaker codes to the linear regression model above (which already included gender). For F1, these accounted for an additional 19% of the variance ($F_{change}(8,465)=63.5$, $p<.001$), allowing the model to account for 82.4% of the total variance in F1 using only information about the speaker. For F2, individual speaker accounted for an additional 4.9% of the variance ($F_{change}(8,465)=4.8$, $p<.001$), for a total R^2 of 40.8%. The residuals were computed in the same was as before (only now these residuals included F1 and F2 values for which both the effects of gender and speaker were removed). These residuals were entered into the logistic regression model which now averaged 92.8%

correct, an improvement of 2.3% over the baseline model. Interestingly, while F1 and F2 were still significant (F1: Wald(1)=9.1, p<.001; F2: Wald(1)=77.7, p<.001), the interaction was less so (Wald(1)=4.4, p=.037), suggesting that progressively parsing out data reduces the need to keep track of higher order dependencies between cues.

Further parsing of the neighboring consonant can yield even better performance Here, place and voicing of the intervening consonant significantly accounts for an additional 1.6% of the variance in F1, over and above speaker ($F_{change}(3,462)=15.8$, $p<.0001$) and an additional 14.5% of the variance in F2 ($F_{change}(3,462)=49.9$, $p<.0001$). When these are entered into the logistic regression model the model averaged 95.2% correct. In fact, additional analyses suggest that further parsing out the influence of the context vowel can increase this to 96.2%.

This initial model illustrates that parsing out speaker-related variance (both gender and individual speaker) may yield modest improvements in the ability to classify the target vowel and parsing out the effects of the consonant (and context vowel) can have larger effects. Figure 1 illustrates this quite clearly showing a series of scatter plots of the F1 and F2 values for each measured token at each step of the foregoing analysis. Panel A shows the unparsed values: there is substantial overlap between the vowel categories, and a number of sub-clusters present. In Panel B, the effect of vowel is removed and only two clusters remain (one for each vowel). In Panels C and D the effects of individual speaker and consonant (in addition to gender) are parsed from the raw values, and by the time context vowel is added to the model (in Panel E) there is virtually no overlap between the two categories at all. Thus, by gradually removing these simple sources of variance, discrete, non-overlapping target vowel categories can be seen quite clearly.

This is a somewhat easy classification problem (as we've modeled it here): the model has the two primary cues to the contrast, and we've artificially restricted the decision space to two alternatives. Thus, it is not surprising that the base model did so well. Nonetheless, it makes a case that parsing just a few factors can improve even this simple categorization. By removing known sources of variation (speaker, and consonant) we can improve the ability of the model to reveal the underlying (discrete) vowel category. Thus, we now turn to the more complex problem: harnessing V-to-V coarticulation to facilitate perception.

4.2 Anticipating the Context Vowel

The goal of these analyses was to predict the identity of the *upcoming vowel* based solely on the formant cues of the target vowel. This is a much more difficult task as illustrated by our baseline model. This model used raw (unparsed) F1 and F2 values of the target vowel along with an interaction term to predict the upcoming vowel (/i/, /æ/, /ɑ/, or the "same" vowel /ʌ/ or /ɛ/) in a multinomial logistic regression. The model averaged only 28.6% correct (chance is 25%). It did better than expected when the prediction was for an upcoming /i/ (51.3%) or /ɑ/ (37.5), but it was much lower for /æ/ (21.7%) and virtually never identified a non-coarticulated, "same" segment (5%). The model fit as a whole was barely significant ($\chi^2(9)=17.6$, $p=.04$), and none of the individual terms (F1, F2 or the interaction) contributed significantly. Thus, this is clearly a situation in which parsing out variance has much to offer.

Our investigations of the benefits of parsing look at a number of factors which influence variability in the target vowel: speaker, the target vowel identity, and place and voicing of the neighboring consonant. We first ran complete hierarchical regression analyses examining the

effects of these factors on F1 and F2 (while simultaneously recording the residuals at each step). This linear regressions is identical to the ones on which the prior parsing model was based—parsing the effect of consonant is the same whether you are using the residuals to identify the target vowel or predict the context vowel. However, to put these factors in perspective we will briefly summarize the linear regression as a whole so that we can adequately describe the relative size of the different sources of variation in F1 and F2. Having done that, we will next evaluate the effects of parsing these factors from the speech signal prior to predicting the context vowel.

Table 2 provides a summary of the regression analysis examining F1. As we described, gender is an important factor, accounting for 63.2% of the variance, and individual speakers for an additional 19.2%. In addition, the identity of the target vowel (/ʌ/ or /ɛ/) accounted for .9% of the variance. Voicing accounted for almost double that (R^2=.018), and place, though small was also significant (R^2=.003, p=.006). While it is not surprising that these known sources of variation and coarticulation were significantly related to F1, the total amount of variation that this model explained is somewhat surprising. The total R^2 for the model was 85.5% (which increases slightly with interaction terms not reported here). Thus, there is significant variation in F1 that could be effectively dealt with by a parsing model.

The second analysis examined F2 (Table 3). Gender and speaker accounted for much less variance associated with indexical factors (Gender: R^2_{change}=.359, Speaker, R^2_{change}=.049), and gender was the bulk of this. However, target vowel was associated with substantially more variation (R^2_{change}=.413), due to the fact that the /ʌ/~/ɛ/ distinction is primarily one of frontness and thus carried in F2. Voicing and place also accounted for more variance in F2 (than F1), with voicing accounting for 3.4% of the variance, and place accounting for 5.0%. Finally, as before,

these five factors together did surprisingly well—the model overall accounted for 90.2% of the variance (and the addition of other factors can increase this to 94.0%).

Given these models, we next examined the performance of the logistic regression classifier using F1 and F2 values from which various (interesting) combinations of factors had been partialed out. Each of these models used a multinomial logistic regression to predict the context vowel (/i/, /æ/, /ɑ/, or "same") on the basis of F1, F2 and an interaction term. F1 and F2 were the residuals from the appropriate step of the linear regressions detailed above.

We examined five different models (see Table 4, and Figure 2 for a summary). The first (GENDER) assumed only that the model could parse out the effects of gender—no detailed representation of speaker was available, nor could the model cope with coarticulation from the consonant. The second (SPEAKER), had more detailed normalization and parsed out individual speaker means in addition to gender. The third model (VOWEL) normalized for speaker, and also accounted for the target vowels. From the perspective of online processing, this represents the degree of prediction that can be made at the target vowel (without any subsequent context to provide a regressive "cleaning up" of the signal). The fourth model (FULL) parsed out speaker, vowel and the place and voicing of the neighboring consonant. Finally, the fifth model (NOSPKR) represented a rather special case in which the model could only parse out coarticulation, but could not normalize for speaker factors.

The GENDER model did somewhat better than baseline, averaging 30.1% correct, and the model fit was good ($\chi^2(9)=20.24$, $p=.016$). While its performance for /i/ was reduced (44.3% compared to 51.3% at baseline) this increase came from the fact that now /æ/ was above chance (35% correct). Where the prior model tended to assign all front vowels to /i/ (a high false-alarm

rate), this model began to differentiate by height. Thus, unlike the baseline model, simply knowing the gender of the speaker allows all of the positive predictions (/i/, /æ/, /ɑ/, but not same) to be above chance. In addition, this was the first model in which F1 was significant ($\chi^2(3)=14.0$, $p=.003$), although F2 and the interaction were not.

Adding a more detailed representation of speaker added an additional 2.5% to the performance, with the SPEAKER model averaging 32.6% correct. Other than this, this model was quite similar to the GENDER model. Its pattern of performance across vowels was similar, and F1 was the only significant covariate.

Parsing out variability due to target vowel (the VOWEL model) improved the model further. This model averaged 35.2% correct and was above chance on all of the positive predictions. While "same" predictions were still below chance (12.5%) these were higher than prior models. Moreover, unlike prior models, this model appeared to make use of both F1 and F2 (F1: $\chi^2(3)=40.3$, $p<.001$; F2: $\chi^2(3)=27.7$, $p<.001$) and the interaction was not significant.

The FULL model performed best averaging 39.4% correct. Performance on /i/ was quite good (58.3% correct), and /ɑ/ (48.3%) and /æ/ (37.5) were well above chance. Even "same" responding, though below chance was markedly improved (14.2%). As in the VOWEL model, both F1 and F2 contributed significantly (F1: $\chi^2(3)=48.0$, $p<.001$; F2: $\chi^2(3)=44.9$, $p<.001$). Thus, when all sources of variance on F1 and F2 in the target vowel are accounted for, the model can predict an upcoming vowel at well above chance levels. Moreover, analyses reported in Cole et al (under review) suggest that after parsing F1 exclusively codes the height of the upcoming vowel (it is not influenced by backness) and F2 codes primarily backness (with a small influence of height).

The final model (NOSPKR) asked if speaker normalization of some kind is required to attain this sort of performance. Here, only consonant and target-vowel variation were partialed out of F1 and F2. This model suggests that accounting for speaker variation can play an important role in leveraging V-to-V coarticulation. It averaged 34.1% correct and though it did well on /i/ and /ɑ/ (50.4% and 52.5% correct, respectively) it was below chance on /æ/ (19.2%).

Across these analyses, a couple of key findings can be seen. First, none of the models did well predicting that the context vowel was the "same" as the target vowel. This suggests that at some level, the absence of coarticulation can not be interpreted as evidence that there is a neutral context coming up. Second, parsing out variability from raw values adds significantly to their predictive power. While the baseline model is barely above chance, the FULL model improves upon this by over 10%. Finally, no single source of variability is essential—even as simple a factor as gender can play a role. We've assumed that all of these sources of variability are equally easy to compute and use, however, some may be more difficult than other for the learner/perceiver. For example, keeping track of individual speaker's mean formant values may require the listener to store many more values than keeping track of means associated with a dichotomous variable like voicing (although speakers could be quickly learned). While future work should consider the processing implications when deciding whether or not to include a factor, this makes it clear that even in the absence of the FULL model, one can do quite well by parsing.

5. Discussion and Conclusions

This study offers a simple formal approach to parsing that can be applied to real speech data. However, despite this simplicity, it suggests that even the most rudimentary parsing can offer

significant power to the perceptual system. A few known sources of variability (speaker, vowel, consonant, and V-to-V coarticulation) accounts for upwards of 85% of the variance in F1 and F2. By parsing only these factors identification of the target vowel was improved by 6% (to 96%), and prediction for the subsequent vowel improved from near chance (28%) to 39.4%. This speaks to the power of attempting to account for (and exploit) multiple sources of variability simultaneously.

This approach allowed us to ask concrete questions about the benefits of parsing various elements from the speech stream, given the statistical structure present in a set of cues. For example, the improvements in identification (in terms of percentage correct) that can be had by parsing out gender (about 1.5% for anticipating the context vowel, 0.9% for the target) are similar to the improvements to be had by parsing the differences in individual speakers, over and above gender (1.5% for context vowels, 1.4% for target). Thus, some mechanism for tracking the way individual speakers use particular cues may be helpful for perception. This may even extend to relatively speaker-invariant cues such as voicing. Allen, Miller and DeSteno (2003; Allen & Miller, 2004), for example, demonstrated significant differences between speakers of the same language and dialect in their use of VOT, and that listeners were sensitive to these differences. Thus, it is possible that this parsing approach to speaker normalization may apply to many phonetic cues.

Variation due to the neighboring consonant may play an equally important role as variation due to speaker. Parsing out the consonant's effects on the target vowel improved the model's performance by 2.4% over speaker factors for target vowel identification, and by 4.2% for the context vowel. Given that speech unfolds temporally, this suggests an interesting model (Figure 3). As the utterance unfolds, the listener starts to identify the speaker (or minimally,

speaker's gender). When the target vowel arrives, the listener first parses speaker factors from the target vowel (Figure 4, Step 1). At this point the target vowel can largely be identified (Step 2), with an accuracy of 92.8%. This identification allows it to parse further variance from F1 and F2 (step 3) and start to anticipate the context vowels (Step 4) with an accuracy of 35.2%. Once the consonant is heard (Step 5), the listener can revise its decision about the target vowel, or confirm the earlier choice (since its accuracy will now be up to 95.2%), and simultaneously parse out the variance in F1 and F2 that is associated with the consonant (Step 6). This in turn allows fairly useful prediction about the consonant (Step 7) with a likely accuracy of 39.4%. Thus, parsing represents a continual interplay between anticipating the next sound, cleaning up variance in the current or prior segments and then looking forward again, and the online nature of the problem can dictate what is parsed when. Realistically, such a system is most likely not the stage-like serial process we have caricatured here. As Fowler (1984) suggests, interactive activation type architectures would seem to implement something like this quite capably.

These are just examples of the kind of findings that can be achieved with this model. In the domain of vowel perception, it leads to the somewhat obvious conclusion that virtually every source of variation can help in some way (though the timecourse over which such cues are available may affect the results). However, the results we obtain here for parsing vowel variance may not be matched in parsing other cues for other types of sounds. For example, McMurray, Jongman & Wang (in preparation) suggest that parsing the effects of voicing on a fricative offers no improvement in identifying its place of articulation.

This model has its limitations, however, as a complete treatment of parsing. In particular, it makes two assumptions which at face value may be questionable. First, parsing is only as good as the model's ability to unambiguously identify the category or feature of each source of

variance. For example, if the target vowel was identified incorrectly as an /ɛ/, what would appear to be a low F2 (for an /ɛ/) may in fact be a high F2 for an /ʌ/. This in turn could lead it to favor an /ɑ/ for the next vowel over an /æ/. Thus, miscategorizing a single feature would have ramifications downstream. However, while this may seem a challenge in the context of single feature identification, it is important to note that the system is identifying multiple sources of variation simultaneously. Thus, the ability to identify one feature (e.g. speaker), improves the ability to identify the next (e.g. vowel), which then provides information for further features (e.g. the next vowel). Here, a few stable features (e.g. landmarks: Stevens, 2002) may provide the necessary entry points. Alternatively, one could imagine the system making a preliminary decision and using that simultaneously to identify future material, and using this future material to subsequently revise the initial decision (as in Fowler's, 1984, discussion of the similarities between interactive activation models and parsing). Thus, when the consonant and context vowels are perceived, this can in turn correct ambiguous or misleading interpretations for the target vowel (or even the speaker). Under our view, the goal of the system is not just to determine a single feature, but to arrive at an optimal parse that accounts for all of the various cues and causes. Given this, there may be very few parses satisfying this constraint for a given utterance.

Second, this model assumes that during online perception, the listener has access to the mean cue values corresponding to various features (e.g. the mean F1 for high and low vowels). This is also not so unreasonable. There is clear evidence that listeners can learn the means of various categories in a brief period, via simple statistical learning mechanisms (Maye, Werker & Gerken, 2002; Maye, Weiss & Aslin, 2008), and that the structure of adult speech categories

closely resembles the statistical structure of the input cues (Miller & Volaitis, 1989). While there is clearly a great deal more to the developmental story (in particular the way that lexically contrastive meaning may help with this acquisition process), it seems clear that the relevant statistics could conceivably be extracted over the lifespan, or even in a few minutes of exposure. In fact, a number of computational models of statistical learning are based on explicitly extracting means and variances from the distribution of the input (e.g. McMurray, Aslin & Toscano, in press; Toscano & McMurray, 2007, in press), and may offer a formal platform in which to integrate statistical learning and parsing.

Beyond these limitations however, this model provides a fairly direct answer to the question of where does discreteness in phonology come from. In short, the ability to discretely identify a category from a variable signal emerges during online perception over the course of progressively parsing out sources of variation. As we've discussed, this can only happen when you attempt to identify all of the sources of variations simultaneously. If you treat the problem as identifying a single feature in a sea of noise, such operations (and their power) are not available. Only by considering everything together (as the perceptual system surely must do) can such discreteness emerge.

This has a number of important implications for speech perception. Work on the perceptual processes that cope with coarticulation has generally divided the processes into progressive effects which anticipate future material and regressive effects which resolve ambiguity in the past. However, our model suggests that these are the same thing. The same regression model was used to partial out variance for identifying the target vowels as well as for anticipating the upcoming vowel. The only differences between the analysis of anticipatory and

regressive effects are the stage at which parsing stops and the choice of which residuals are used to identify the vowel.

Moreover, this model shows that parsing is not just useful for identifying contrastive phonological features (although this paper clearly shows that it is). It can also account for other sources of variation such as speaker and gender. In this way, it is consistent with exemplar based approaches to normalization, in that it would be straightforward to extract a speaker's mean value for a cue from a set of indexically coded exemplars. However, this is not the only way to obtain such values. Speaker means can be rapidly learned (McMurray, Horst, Toscano & Samuelson, in press), or extracted as prototypes without episodically retaining the full set of exemplars. Either way, it suggests that parsing is just a generic process for dealing with variation of any kind.

This approach shares much with the gestural approaches to phonology (Browman & Goldstein, 1992; Goldstein & Fowler, 2003), but it is also distinct. Parsing originated in the gestural tradition (Fowler, 1984; Fowler & Smith, 1986) and was originally intended for interpreting overlapping gestures. Our work strongly supports this as a mechanism. However, it also points out that other sources of variance can be parsed as well. We've discussed speaker normalization, but there is also emerging evidence that the structure of the lexicon can be another source of information for parsing during the processing of place assimilated speech (Munson & McMurray, 2007; Gow & McMurray, 2007).

Exemplar approaches also overlap with our approach, in large part due to the fact that, like the exemplar approach, we stress the importance of fine-grained, continuous detail in the speech signal, and we take pains to deal with the problem posed by speaker-variability. However, unlike exemplar models in which the word is the unit that is stored, parsing can work

across word boundaries. Moreover, in order to realize the effects of fine-grained detail for anticipating future material, our model uses speech input that has been processed through parsing, rather than the raw (unprocessed) acoustic cues used in exemplar models. Furthermore, the parsing approach does not require the storage of multiple complete words in memory—means and variations of these cues (a prototype model of sorts) are sufficient to do the job.

Thus, parsing represents a somewhat novel synthesis of both gestural and exemplar-based theories. It offers a unique explanation for the origin of discreteness in perception. Features are an emergent property of a perceptual process that copes with the redundant variability in the speech signal. When gradient detail in the input is treated as signal to be accounted for and exploited, rather than noise to be ignored, perceptual processing is facilitated, and discrete features as cues to meaning can emerge.

6. Acknowledgments

The authors would like to thank Gary Linebaugh for assistance with data collection and helpful discussions throughout this project. This project was supported in part by grants NIH, DC008089, awarded to BM, and NIH HD044458 and NSF IIS 07-03624 awarded to JC.

7. References

Alfonso, P. J. & Baer, T. (1982). Dynamics of vowel articulation. *Language and Speech, 25* (2), 151-173.

Allen, J.S., Miller, J.L., & DeSteno, D. (2004) Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 115,* 3171-3183.

Allen, J.S., Miller, J.L., & DeSteno, D. (2003) Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 113,* 544-552.

Andruski, J.E., Blumstein, S.E. & Burton, M.W. (1994) The effect of subphonetic differences on lexical access. *Cognition, 52,* 163-187.

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics, 30*, 591-627.

Blumstein, S.E. & Stevens, K.N. (1981) Phonetic features and acoustic invariance in speech. *Cognition, 10,* 25-32.

Browman, C. P. & Goldstein, L. (1992) Articulatory phonology: An overview. *Phonetica 49:* 155-180.

Carney, A.E., Widin, G.P. & Viemeister, N.F. (1977) Non categorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America, 62,* 961-970.

Cho, T., & Ladefoged, P. (1999) Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics, 27,* 207–229.

Cole, J., Linebaugh, G., Munson, C., & McMurray, B. (under review) Vowel-to-vowel coarticulation across words in English: Acoustic evidence.

Connine, C. (2004) It's not what you hear but how often you hear it: on the neglected role of

phonological variant frequency in auditory word recognition. *Psychonomic Bulletin and Review, 11(6),* 1084-1089.

Cooper, F.S., Liberman, A.M. & Borst, J. (1951) The interconversion of audible and visual patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences, 37,* 318-325.

Fowler, C.A. (1981) Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech & Hearing Research, 24:* 127-139.

Fowler, C. (1984) Segmentation of coarticulated speech in perception. *Perception & Psychophysics, 36,* 359-368.

Fowler, C. A. (2005). Parsing coarticulated speech in perception: effects of coarticulation resistance. *Journal of Phonetics, 33*, 199-213.

Fowler, C., & Brown, J. (2000) Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Perception & Psychophysics, 62(1),* 21-32.

Fowler, C., & Smith, M. (1986) Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell, & D. Klatt (eds.) *Invariance and variability in speech processes*, pp. 123-136. Hillsdale, NJ: Erlbaum.

Fry, D.B., Abramson, A.S., Eimas, P.D. & Liberman, A.M. (1962) The identification and discrimination of synthetic vowels. *Language and Speech, 5,* 171-189.

Gerrits, E. & Schouten, M. (2004) Categorical perception depends on the discrimination task. *Perception & Psychophysics, 66(3),* 363-376.

Goldinger, S. (1998) Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105:2*, 251–279.

Goldstein, L. & Fowler, Carol A. (2003) Articulatory phonology: A phonology for public language use. In N. O. Schiller & A. Meyer (eds.) *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, pp. 159-207. Berlin: Mouton de Gruyter.

Gow, D. W. (2001). Assimilation and anticipation in Continuous Spoken word recognition. *Journal of Memory and Language, 45*(1), 133-159.

Gow, D.W. (2002). Does phonological assimilation create lexical ambiguity?. *Journal of Experimental Psychology: Human Perception and Performance,* ***45****,* 133–159.

Gow, D. W. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics, 65*(4), 575-590.

Gow, D. W. & Gordon, P. C. (1995) Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 344–359.

Gow, D., & Im, A. (2004) A cross-linguistic examination of assimilation context effects. *Journal of Memory and Language, 51(2),* 279-296.

Gow, D., & McMurray, B., (2007) Word recognition and phonology: The case of English coronal place assimilation. In J. Cole & J. I. Hualde (eds.) *Laboratory Phonology 9,* pp. 173-199. Berlin: Mouton de Gruyter.

Healy, A.F. & Repp, B. (1982) Context independence and phonetic mediation in categorical perception. *Journal of Experimental Psychology: Human Perception and Performance, 8(1),* 68-80.

Helgason, P., & Ringen, C. (in press) Voicing and aspiration in Swedish stops. *Journal of Phonetics*

Hillenbrand, J.M., Getty, L., Clark, M.J. & Wheeler, K. (1995) Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97(5),* 3099-3111.

Hillenbrand, J.M., Clark, M.J., & Nearey, T.M. (2001) Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America, 109(2)*, 748-763.

Hock, H. (1991) *Principles of Historical Linguistics.* Berlin: Mouton de Gruyter.

Jiang, J, Chen, M., & Alwan, A. (2006) On the perception of voicing in syllable initial plosives in noise. *Journal of the Acoustical Society of America, 119(2),* 1092-1105

Keating, P. A. (1984) Phonetic and phonological representation of stop consonant voicing. *Language*, *60,* 286–319.

Kewley-Port, D. & Luce, P.A. (1984) Time-varying features of initial stop consonants in auditory running spectra - a 1st report. *Perception & Psychophysics, 35(4),* 353-360.

Keyser, S. J. & K.N. Stevens (2006) Enhancement and overlap in the speech chain. *Language*, *82*, 33-63.

Liberman, A.M., Harris, K.S., Hoffman, H.S. & Griffith, B.C. (1957) The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54(5),* 358-368.

Lindblom, B. (1996) Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America, 99(3)*, 1683-1692.

Lindblom, B. (2000) Developmental origins of adult phonology: The interplay between phonetic emergents and the evolutionary adaptations of sound patterns. *Phonetica*, 57, 297–314.

Lisker, L. & Abramson, A.S. (1964) A cross-language study of voicing in initial stops: acoustical measurements. *Word, 20,* 384-422.

Magen, H. S. (1997) The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics, 25,* 187-205.

Manuel, S. Y. (1990). The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America, 88* (3), 1286-1298.

Martin, J.G., & Bunnell, H.T. (1981) Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America, 69(2),* 559-567.

Martin, J.G., & Bunnell, H.T. (1982) Perception of anticipatory coarticulation effects in vowel-stop consonant-bowel sequences. *Journal of Experimental Psychology: Human Perception and Performance, 8(3),* 473-488

Massaro, D.W. & Cohen, M.M. (1983) Categorical or continuous speech perception: a new test. *Speech Communication, 2,* 15-35.

Maye, J., Weiss, D.J., & Aslin, R.N. (2008) Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11.1,* 122-134.

Maye, J., Werker, J.F., & Gerken, L. (2002) Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82.3,* B101-B111.

McLennan, C., Luce, P.A. & Charles-Luce, J. (2003) Representation of Lexical Form. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29(4),* 539-553.

McMurray, B., Aslin, R., Tanenhaus, M., Spivey, M., & Subik, D. (in press). Gradient sensitivity to within-category variation in speech: Implications for categorical perception. *Journal of Experimental Psychology: Human Perception and Performance.*

McMurray, B., Aslin, R.N., & Toscano, J. (in press) Statistical learning of phonetic categories: Computational insights and limitations. *Developmental Science.*

McMurray, B., Horst, J., Toscano, J., & Samuelson, L. (in press) Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. Invited submission for Spencer, J., Thomas, M,. & McClelland, J. (eds.) *Toward a new grand theory of development? Connectionism and Dynamic Systems Theory reconsidered.* London: Oxford University Press

McMurray, B., Jongman, A., & Wang (in preparation) Linking perception and production of fricatives with a parsing approach.

McMurray, B., Tanenhaus, M., & Aslin, R. (under review) Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition

McMurray, B., Tanenhaus, M., & Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access, *Cognition, 86(2),* B33-B42.

Miller, J.L. & Volaitis, L.E. (1989) Effects of speaking rate on the perceived internal structure of phonetic categories. *Perception & Psychophysics, 46,* 505-512.

Möbius, B. (2004). Corpus-based investigations on the phonetics of consonant voicing. *Folia Linguistica 38:1-2,* 5-26.

Monaghan, P., Chater, N. & Christiansen, M.H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition, 96,* 143-182.

Munson, C., & McMurray, B. (2007, October) Perceptual features of place assimilation are continuous and contextually. *Poster presented at* *Where do Features Come From? Phonological Primitives in the Brain, the Mouth, and the Ear,* Paris.

Ohala, J. J. (1996) Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America, 99(3)*, 1718-1725.

Öhman , S. E. G. (1966) Coarticulation in VCV utterances : Spectrographic measurements, *Journal of the Acoustical Society of America, 39,* 151–168.

Pardo, J. S., & Fowler, C. A. (1997) Perceiving the causes of coarticulatory acoustic variation: consonant voicing and vowel pitch. *Perception & Psychophysics, 59(7)*, 1141-1152.

Pierrehumbert, J. (2003) Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech, 46,* 115-154.

Pisoni, D.B. & Lazarus, J.H. (1974) Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America, 55(2),* 328-333.

Pisoni, D.B. & Tash, J. (1974) Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15(2),* 285-290.

Repp, B. (1984) Categorical perception: Issues, methods and findings. In N. Lass (Ed.) Speech and Language (vol. 10): Advances in Basic Research and Practice (pp. 244-335). Orlando: Academic Press

Recasens, D., & Pallarès, M. D. (2000) A study of F1 coarticulation in VCV sequences. *Journal of Speech, Language & Hearing Research, 43*, 501-512.

Salverda, A.P., Dahan D., McQueen, J.M. (2003) The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90 (1),* 51-89.

Samuel, A. (1977) The effect of discrimination training on speech perception: Noncategorical perception. *Perception & Psychophysics, 22(4),* 312-330.

Stevens, K.N. (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America, 111(4),* 1872-1891.

Stevens, K.N. & Blumstein, S.E. (1978) Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64(5)*, 1358-1368.

Sussman, H., Fruchter, D., Hilbert, J. & Sirosh, J. (1998) Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Science, 21(2)*, 241-299.

Toscano, J., & McMurray, B. (in press) Using the distributional statistics of speech sounds for learning and combining multiple acoustic cues. *Proceedings of the Cognitive Science Society.*

Toscano, J., & McMurray, B. (2007, October) Integrating acoustic cues to phonetic features: a computational approach to cue weighting. Poster presented at *Where do Features Come From? Phonological Primitives in the Brain, the Mouth, and the Ear,* Paris.

Utman, J.A., Blumstein, S.E. & Burton, M.W. (2000) Effects of subphonetic and syllable structure variation on word recognition. *Perception & Psychophysics, 62(6),* 1297-1311.

Tables

bed	actor	tech	afternoon	web	addict
	eagle		evening		ecologist
	evergreen		elevator		educator
	ostrich		oxygen		offer
wet	afro	deck	alligator	step	admiral
	Easter Bunny		easter basket		east
	Eskimo		elephant		exit
	oxen		octopus		obstacle
mud	apple	bug	astronaut	pub	advertisement
	eater		evil		easel
	umpire		underwear		undergrad
	observation		optician		operator
cut	abdomen	duck	athlete	cup	appetizer
	evenly		eating		eavesdropping
	onion		usher		oven
	olive		officer		occupant

Table 1: Test phrases used in the experiment.

Step	Variables	R^2_{change}	F_{change}	P
1	Gender	.632	F(1,473)=811.5	.0001
2	Subjects (10)	.192	F(8,465)=63.5	.0001
3	Vowel	.009	F(1,464)=25.5	.0001
4	Voicing	.018	F(1,463)=56.5	.0001
5	Place (2)	.003	F(2,461)=5.1	.006
	Total R^2	**.855**		

Table 2: Results of a regression analysis examining all sources of variation on F1.

Step	Variables	R^2_{change}	F_{change}	P
1	Gender	.359	F(1,473)=264.7	.0001
2	Subjects (10)	.049	F(8,465)=4.8	.0001
3	Vowel	.413	F(1,464)=1066.8	.0001
4	Voicing	.034	F(1,463)=107.0	.0001
5	Place (2)	.05	F(2,461)=120.9	.0001
	Total R^2	**.902**		

Table 3: Results of a regression analysis examining all sources of variation on F2.

Model	**Parsed Out**	**% Correct**	**i**	**ɑ**	**æ**	**same**
BASELINE	-	28.6	51.3	37.5	21.7	5.0
GENDER	Gender	30.1	44.3	31.7	35.0	10.0
SPEAKER	Gender Speaker	32.6	49.6	33.3	40.0	8.3
VOWEL	Gender Speaker Target Vowel	35.2	49.6	47.5	31.7	12.5
FULL	Gender Speaker Target Vowel Consonant	39.4	58.3	48.3	37.5	14.2
NOSPKR	Target Vowel Consonant	34.1	50.4	52.5	19.2	15.0

Table 4: Percent correct for multinomial logistic regression models predicting the context vowel from various sets of F1 and F2.

Figures

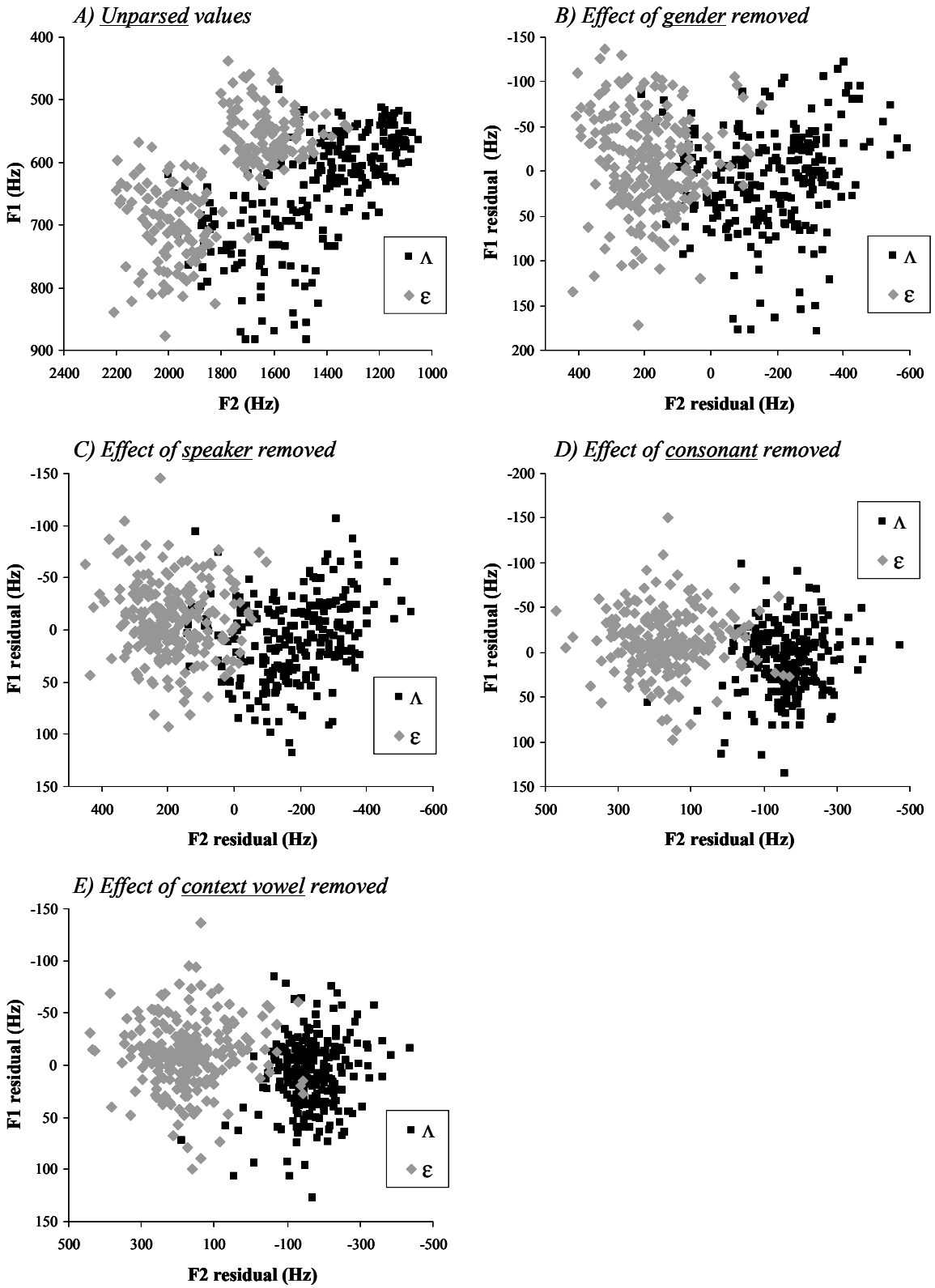
A) Unparsed values
F1 (Hz)
F2 (Hz)
B) Effect of gender removed
F1 residual (Hz)
F2 residual (Hz)
C) Effect of speaker removed
F1 residual (Hz)
F2 residual (Hz)
D) Effect of consonant removed
F1 residual (Hz)
F2 residual (Hz)
E) Effect of context vowel removed
F1 residual (Hz)
F2 residual (Hz)
Λ
ɛ

Figure 1: F1 and F2 for all tokens as a function of target vowel. Panel A shows raw values in Hz (note: regression analyses reported here were conducted on data transformed to Bark). Panel B: raw F1 and F2 frequencies after effect of gender is eliminated. Panel C: F1 and F2 after both gender and individual speaker are parsed from dataset. Panel D: Gender, speaker and now consonant have now been parsed out. Panel E: With the addition of context vowel to the parsing model there is virtually no overlap in the vowel categories.

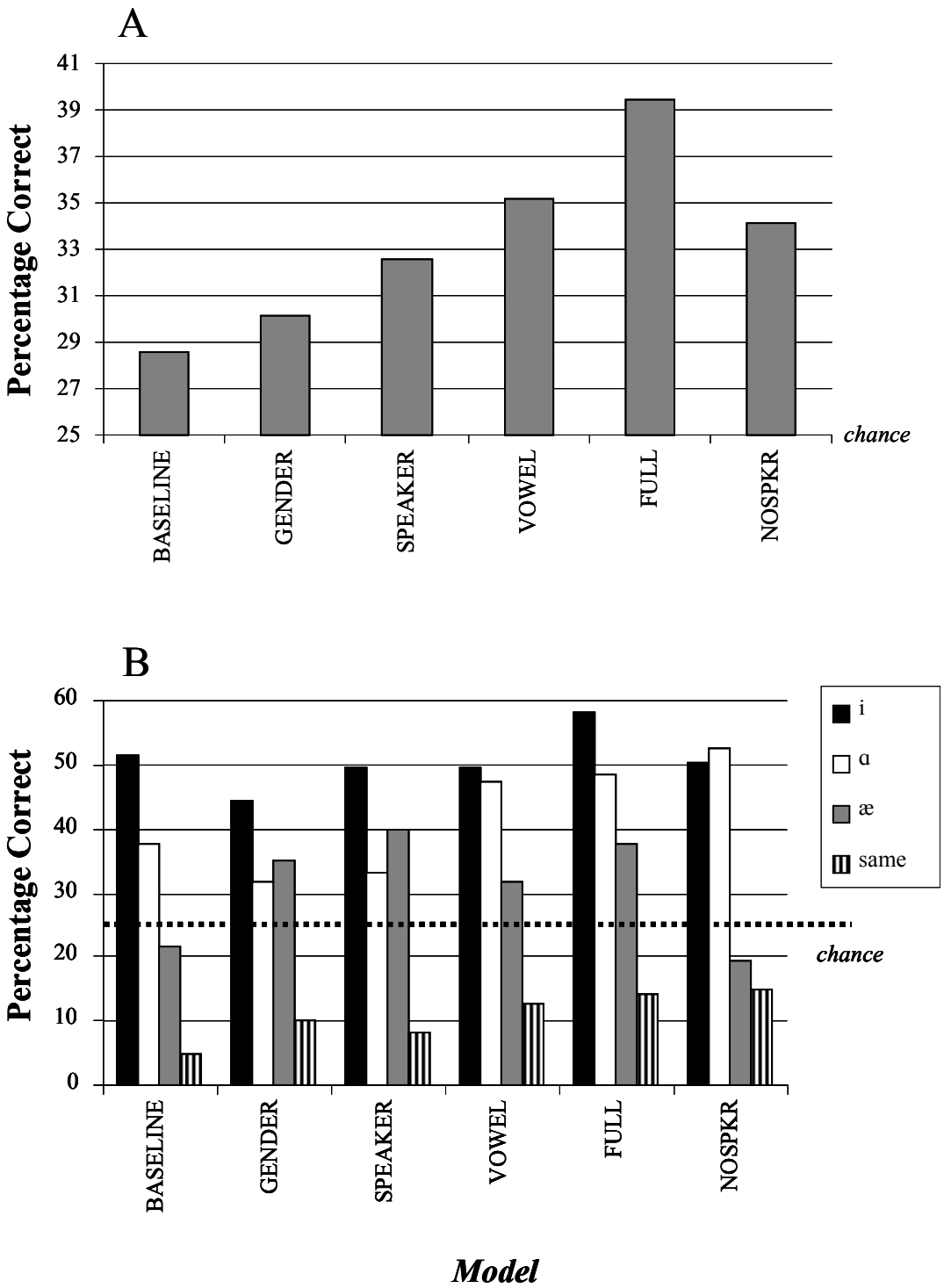

Figure 2: Summary of performance of six different parsing models on the problem of anticipating the context vowel. Panel A: overall performance averaged across all four contexts. Panel B: performance for each model, for each vowel. A clear pattern can be seen in which /i/

and /ɑ/ are consistently above chance, “same” responding is below chance for all models, and /æ/ is above only one speaker factors have been parsed from the signal.

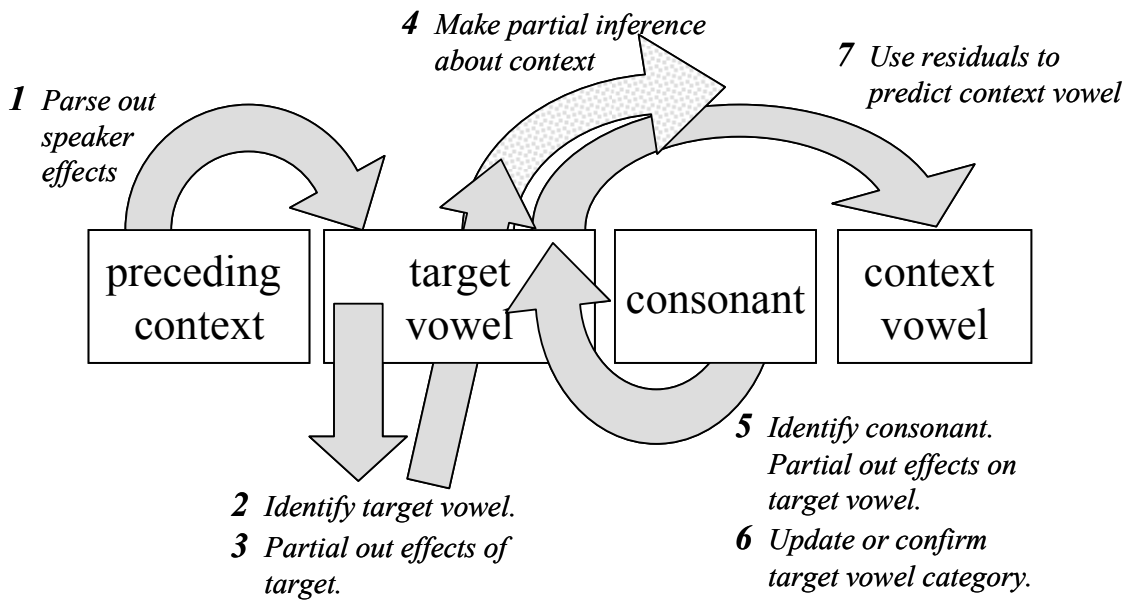

Figure 3: Some hypothesized directions of parsing for target and context vowel identification.

Notes

[1] Leaving aside F0 as a feature encoding pragmatic meaning.

[2] /ɛ/ was excluded as a target in front of /g/, as speakers often produce vowels that are higher and tenser than usual in this particular context (Hartman, 1985; Kurath & McDavid, 1961: 102, 132-133). To compensate for this elimination, /ɛ/ was recorded in the context of a second /k/ word, keeping the number of labial, velar, and alveolar contexts the same across the two target vowels.