Features as an emergent product of computing perceptual cues relative to expectations

Bob McMurray[1], Jennifer S. Cole[2] & Cheyenne Munson[1]
[1]University of Iowa/[2]University of Illinois

The product of speech perception is contrast between discrete units of meaning, words, which are contrasted by features. While traditional approaches argued that discreteness is imposed by mechanisms like categorical perception that discard within-category detail, recent research suggests that fine-grained detail is preserved throughout processing. We develop an alternative that argues that discreteness emerges from processes that parse overlapping sources of variance from the signal. These need not discard acoustic detail and may make it more useful to listeners. We develop a computational implementation (Computing Cues Relative to Expectations, C-CuRE) and test it on a corpus of vowel productions. It shows how C-CuRE reveals underlying vowel features despite contextual variance, and simultaneously uses the variance to better predict upcoming vowels.

1. 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 in part 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, although there are a large number of other cues that contribute to voicing). However, cross-linguistic studies of voicing (e.g. Lisker & Abramson 1964; Keating 1984; Cho & Ladefoged 1999; Möbius 2004; Helgason & Ringen 2008) suggest that across languages, voicing categories are not distributed uniformly along the range of possible VOT values. Rather, languages exhibit clusters of VOT values that suggest two discrete features for describing crosslinguistic variation: 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 movement 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 discrete 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 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 mapping 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 (see Gafos & Benus 2006, for analogous issues in production). Perception poses a unique domain in which to address these issues, in that while discreteness would seem necessary for phonology, a growing body of work in online speech perception argues that perception functions gradiently. This work suggests that listeners are sensitive to fine-grained detail (Andruski, Blumstein & Burton 1994; Utman, Blumstein & Burton 2000; McMurray, Tanenhaus & Aslin 2002; McMurray, Aslin, Tanenhaus,

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

Spivey & Subik 2008) and that this sensitivity can facilitate online perception by allowing listeners to anticipate future material (Martin & Bunnell 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 2009). Thus, whatever mechanisms give rise to the discreteness necessary for phonology, these mechanisms must also preserve fine-grained or gradient detail for use in online perception.

We propose that discrete features emerge from a processing mechanism that computes cues from the acoustic signal as continuous values. However, these values reflect the difference between the actual cue that was heard and listeners' expectations about how cues behave given contextual factors such as the speaker or neighboring phonemes. We term this process *Computing CUes Relative to Expectations (C-CuRE),* and as we describe, this is a more general, and formally implemented, version of parsing mechanisms that have been previously proposed (Fowler 1984; Gow 2003). C-CuRE effectively reduces the variation in the acoustic signal by attributing portions of the variation to properties of the context (broadly construed). Once this continuous variation has been attributed to contextual factors, underlying features can be revealed more directly (c.f., Ohala 1981; Fowler 1984, 2005; Gow 2003). Of course, factors that are considered context for one feature are likely to be the target of another. Thus, if we allow that listeners simultaneously account for variance due to multiple features, the listener may be able to properly assign much of the so-called "noise" to its underlying phonetic causes.

Computing relative cues or C-CuRE 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 this 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 evidence for the context element and can contribute to its own featural representation. In this account, phonological features are *revealed* through the relative encoding of acoustic cues, as specific acoustic properties are attributed to the target sound or to elements of the context.

This paper demonstrates the emergence of discrete features through this process with a case study of vowel-to-vowel (V-to-V) coarticulation. In this demonstration, the acoustic parameters that are primary cues to the phonological height and backness features of a vowel also reflect both the speaker and the local phonological context, in part due to the fact that the gestures for consonant and vowel overlap. We show that through the by encoding cues relative to expectation based on this context, 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 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 in which discrete elements are viewed as emergent properties of a richly detailed phonetic encoding of word forms. In the remainder of this section, we describe the theoretical underpinnings of parsing and the C-CuRE approach, and our formalization of C-CuRE 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 2 as a prime example of context-induced variation. Section 3 presents our analyses of this dataset using our model. Finally, we return in Section 4 to the question of how features emerge, and conclude that not only does C-CuRE result in better identification of the target sound, but it also preserves fine-grained detail that contributes in identifying the phonologically contrastive dimensions of a context vowel (e.g. height and backness).

1.1 The search for discreteness in perception

In speech perception, researchers have looked for the source of discreteness in phonology in two ways. First, they have sought discreteness in the acoustic signal itself (the search for acoustic invariance). If phonological features can be mapped deterministically from acoustic cues that are relatively invariant to context, discrete features can be found at this level of analysis. Second, they have examined the possibility that perceptual processes impose it on the signal. That is, they have asked if there are discontinuities in perception such that highly variable input can be mapped to the same discrete representation.

The search for acoustic/auditory invariance is perhaps one of the oldest research issues in psycholinguistics (e.g. Cooper, Liberman & Borst 1951). Acoustic cues to phonological features, such as formant frequencies, 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 to invariance focused on stop consonants. Acoustic spectra at word onset, for example (Stevens & Blumstein 1978; Blumstein & Stevens 1981; Kewley-Port & Luce 1984) were found to discriminate place of articulation word-initially, but fared somewhat worse with word-final stops, particularly unreleased stops (Blumstein & Stevens 1979). Sussman and colleagues' work on locus equations (e.g. Sussman, Hilbert, Fruchter & Sirosh, 1998) also uncovered some invariant structure in the encoding of place of articulation in word-initial stops, but the equations show some 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. Numerous phonetic studies of the acoustics of vowels (e.g. Hillenbrand, Getty, Clark & Wheeler 1995; 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, for certain contrasts, there do not appear to be any underlying cues in the speech signal that can be mapped directly to phonological features (see Lindblom 1996; Ohala 1996), a fact that led some to look for invariance at the articulatory level (Liberman & Mattingly 1985; Fowler 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 there is no invariance (with respect to linguistic features) in the acoustic signal itself, it is possible that perceptual processes impose discreteness on the signal. One such process is *categorical perception* (Liberman, Harris, Hoffman & Griffith 1957; reviewed in Repp 1984; McMurray et al. 2008). Categorical perception was suggested by the finding that it is difficult for listeners to discriminate acoustic differences that lie within a phonetic category (e.g. two /b/'s with different VOTs) while they are good at discriminating equivalently small differences that cross a boundary. This finding was taken to imply that early perceptual processing is finely tuned to discrete categories and strips 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), and often can do so at levels equivalent to between-category discrimination (Massaro & Cohen 1983; Gerrits & Schouten 2004). In fact, prior findings of poor within-category discrimination may have been the results of memory demands or biasing tasks (Gerrits & Schouten 2004; Schouten, Gerrits & Van Hessen 2003). Second, listeners do not show categorical perception for vowels (Fry, Abramson, Eimas & Liberman 1962) and show weakened categorical perception in fricatives (Healy & Repp 1982), even when tested under conditions that would normally yield categorical perception in stop consonants. Finally, a host of more recent studies demonstrate that levels of processing that are presumed to operate down-stream from speech perception are in fact sensitive to gradations within a category. For example, activation for lexical candidates is modulated by differences within a category (Andruski, et al. 1994; Utman, et al. 2000; McMurray, et al. 2002; McMurray, et al. 2009c). Thus, 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 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, et al. 2009c). For at least some purposes, then, a discrete representation stripped of "low-level" detail may be suboptimal for perception. If discreteness is to emerge via perceptual processes, it 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 one solution to the problem of deriving discrete categories while preserving phonetic detail. Such models posit that listeners veridically store (in memory) vast numbers of exemplars of the words they have been exposed to. These exemplars are stored with high fidelity to the veridical acoustic signal, and there is nothing in the encoding that differentiates non-contrastive detail (e.g. cues indexing speaker identity) from detail that cues phonological contrast – both are stored with the exemplar and, a priori, both participate equally in retrieval (Goldinger 1998). 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; Pierrehumbert 2003). 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 with enough scale to look for such emergence. Goldinger (1998) focuses on word recognition and the role of speaker/indexical detail – his model does not focus on the variability in the speech signal. Pierrehumbert (2003) emphasizes acoustic variability in the speech signal and offers a number of ways to operationalize perception in an exemplar model; yet, these ideas were illustrated only with simple models of single cues. Johnson (1997) reports on a more complex implementation of an exemplar model, which categorizes 11 vowels produced in single-word utterances from a multi-talker database (39 talkers), using five acoustic parameters. Johnson's model is similar to the parsing model proposed here in that speaker variability is exploited to identify vowels without discarding or changing acoustic information, but the model is not developed to account for context effects due to coarticulation. Most importantly, the range of perceptual processes proposed in exemplar models seems restricted to comparisons and similarity metrics – more complicated inference processes have not been considered. As a result, there are no formal models that make specific claims about the origins of contrastive, discrete features from speech input containing contextual variation. The C-CuRE approach, then may offer such a mechanism, though as we will discuss in the conclusion it does not require an exemplar-based representation in the lexicon.

An additional challenge for exemplar models is coping with patterned variation that arises from context outside of the word. It is 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 in long-term memory. But phonetic variation is often conditioned by elements beyond the word boundary. In most exemplar models, such detail can not be associated with the context from which it derives if the word is the unit of long-term storage. Moreover, if the storage of this lawfully variable acoustic detail was left in its raw (exemplar) encoding such phonetic variability may actually compromise the contrast between the target word and its close lexical neighbors (e.g. Gow 2003). Perhaps this limitation could be dealt with if words are encoded in long-term memory as underspecified along the dimensions affected by context outside the word, but such a model may not be able to harness phonetic detail to its full advantage (which seems to violate the spirit of the exemplar enterprise), and may even be unable to make lexical contrasts in some cases. Moreover, 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 difficult phenomenon for exemplar models to either cope with or take advantage of, and so we focus on that pattern in the demonstration of the parsing model presented below.

In sum, it does not appear that invariant cues for many linguistic features are available in the acoustic signal. That is, linguistic features are not transparently reflected in the signal. Moreover, the perceptual system does not appear to 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 gives rise to 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 effects of disparate phonetic events, effects related to the phonological or phonetic context (e.g. coarticulation) and non-phonological sources (e.g. speaker), and that may include sources that lie outside the boundaries of the target word. While exemplar models can achieve this sensitivity to phonetic detail, it is not clear how discreteness emerges in this framework, nor how such systems cope with variability across word boundaries. C-CuRE is a promising approach that may offer an account of all of these issues.

1.2 Computing cues relative to expectations

The C-CuRE approach derives from a class of perceptual processes loosely identified as parsing. Parsing was first hinted at by Ohala (1981) and later developed 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 simple: at any given point in the signal, listeners assign acoustic cues to causes. Fowler's version of parsing 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) in different positions (e.g. syllable-finally, and at the onset of the next syllable) are grouped via association with features like labiality or coronality. Whether parsing is in terms of gestures or feature-cues, the "causes" can originate in the past or future (i.e. can precede or follow the target sound), or can be cotemporaneous with the target, with the result that parsing can have very powerful results for speech perception.

For example, consider anticipatory vowel nasalization in English. English does not have contrastive vowel nasalization, and oral vowels are often nasalized when they precede a nasal consonant. When the parsing process encounters a nasalized vowel, the nasal cues can be 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 coming up. 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) who found that nasal vowels sound more oral prior to a nasal consonant than to an oral one.

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 some of the potentially confounding information (the irrelevant nasality) eliminated. However, by assigning the nasal cues or gestures to a future segment it simultaneously can 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 examine only a single feature at a time (e.g. the orality or nasality of the target vowel), we are faced with ambiguity and a noisy signal. However, by trying to identify both the target vowel and the conditioning context at the same time (in this case, the subsequent context, but this could also be cotemporaneous or prior), 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 quite difficult.

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). These studies have established that listeners appear to process the signal in a way that is consistent with parsing (broadly defined) to cope with conditioned variation due to a wide range of sources.

Parsing has traditionally been framed either in terms of gestures, or acoustic feature cues, and been largely applied to the problem of coarticulation. However, it is clearly a more general principle of attributing information in the signal to various sources, and using these attributions as the basis of interpreting what is left. Our C-CuRE approach attempts to capture this generality by extending parsing in several ways, and by developing a formal implementation. First, C-CuRE is based on easily measurable and discretely defined cues, like formant frequencies, and durations of components of the signal, rather than more abstract entities like gestures. Second, CERE assumes that any sort of expectation can be the basis of parsing-like operations. For example, a low F0 could be attributed to phonetic events (e.g. a voiced sound) or other factors (e.g. a male speaker). Finally, we assume that the actual cue-value is encoded relative to the expectations driven by this event. That is, if the listener knows the speaker is male, F0 is now coded as high (or low) for a male, the difference between the actual F0 and the expected F0 for men. Thus, CERE posits that one of the outcomes of the attribution process is that the perceptual encoding of the signal is changed.

A critical property of this generalized approach to parsing is its computational tractability. Evidence for parsing has largely come from perceptual work and there have been few systematic investigations of speech production 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 (for a parallel approach in the domain of statistical word segmentation, see Yang 2005). Such work could reveal if the statistical structure of actually occurring acoustic cues is one that could benefit from mechanisms like parsing in the C-CuRE framework, or offer insight into the amount of improvement that C-CuRE could offer (over simply using unprocessed cues). For example, given the total variability in the signal, it is possible that relative cue encoding offers only a marginal benefit. Such a finding might motivate more complex models or other approaches to coping with variability. Conversely, it is possible that unprocessed acoustic measurements are sufficient to support both accurate identification of the target segment and anticipation of future context – in this case, C-CuRE may not be necessary to account for listeners' abilities.

While perceptual experiments can reveal the presence of processes like parsing, they can only assess a small number of tokens in the lab, and cannot assess the broad applicability of such mechanisms. Computational work on a real corpus then can offer an important complement, by showing how such processes can scale up. The lack of studies that test the viability of such mechanisms is due in large part because of the lack of a formal or computational model of them – it would be difficult to answer this question without it.

C-CuRE offers a very general approach to this problem by using easily measurable cues and flexible commitments as to what can be used to generate an expectation. This permits a simple formal model that can be applied to existing phonetic datasets using hierarchical linear regression. This model allows us to ask whether C-CuRE processes can explain the emergence of features over a large and highly variable set of vowel productions. This can ably model the two highly overlapping operations that fall out of parsing in the C-CuRE approach: identifying a target feature and predicting the nearby context (we only separate them here operationally – both are fundamentally about feature extraction). Moreover, the generality of this model 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, consider a model in which the DV is F1 frequency, one of the cues for recognizing vowel identity. In this case, the weighting on the factor (feature), *high*, would literally be the average difference in F1 between *high* vowels and other vowels. The standard regression equation is given below as

$$F1 = \beta_0 + \beta_1 \cdot X$$

Where F1 is the DV (in this example), X is a dichotomous independent variable (height), β_1 is the slope, β_0 the y-intercept. In this case, since height is dichotomous, we would let X be 1 for high vowels and 0 for non-high vowels. As a result, the weight on X (β_1) will be the difference between high and non-high vowels and β_0 will be the mean of the non-high vowels. To illustrate, if $X = 0$ (a non-high vowel), F1 will be $\beta_0 + \beta_1 \times 0$, or β_0; similarly, if $X = 1$ (a high vowel), F1 will be $\beta_0 + \beta_{1.}$

To use this regression model to implement parsing, we use the continuous acoustic measurement as the dependent variable, and generate an equation that predicts its value as the sum of the contributions of a set of discrete features, essentially the expected value of the cue given what is known about the contextual factors. To continue our example of F1 as a DV, F1 will be affected by the speaker (which affects numerous other acoustic parameters as well), the height of the vowel, the voicing of neighboring consonant and so on.

$$F1 = \beta_0 + \beta_1 \cdot \text{speaker} + \beta_1 \cdot \text{height} + \beta_2 \cdot \text{voicing} + \beta_3 \cdot \text{place} \ldots$$

Linear regression can easily compute the weighting (mean differences) for multiple sources of variance in a given dataset, as long as the values of these factors are known for each measurement in the database, that is, as long as we know the values of speaker, height, voicing (etc) corresponding to each measurements of the DV. Similar operations can be used to deal with variance in the other acoustic measures that contribute to the percept (e.g. F2, pitch or duration), allowing a large number of sources of information to be dealt with simultaneously (see McMurray & Jongman, under review, for an example with over 20 measures).

Once we've computed these weightings, the linear regression equation becomes a way of predicting what the value of that cue would be given those contextual factors. That is, what cue value would listeners expect given the speaker, vowel height etc. This can then be used to systematically *remove* the effects of one source of variance on a DV. For example, if we are only considering the effect of speaker, the regression equation is given by

$$F1 = \beta_0 + \beta_1 \cdot \text{speaker} + \varepsilon$$

But once we know that, we can also compute the difference between the predicted F1 and the actual F1, in a sense, the value of F1 relative to its expected value based on the speaker.

$$F1_{\text{resid}} = F1_{\text{measured}} - (\beta_0 + \beta_1 \cdot \text{speaker})$$

This tells us if a given F1 value is higher or lower than it should have been for that speaker; across the dataset, we can examine the variance in this new cue (the residual) to determine how much variability is left over after the effect of speaker as been removed, and further, if the residual variance can be accounted for by or predict them.

This removal of variance can be done progressively in a hierarchical regression to look at multiple features (Cohen & Cohen 1983). In this technique an initial simple model with only a few factors is first fit to the data. The residuals are then computed, and then used as the DV in a new analysis in which 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 residuals as input to a second model whose task is not just to analyze the variance in F1 (or F2 or any other measurement) but rather to use it to identify features such as the target vowel, or features of the upcoming context. As we will discuss, we use logistic regression for such a model, using the residualized (parsed) F1 and F2 values as the input on which it must identify the target and context vowels.

Thus, hierarchical regression operates conceptually like parsing process by assigning variance to factors like speaker, vowel height, voicing and place, and removing that variance from the original acoustic measurement. These residualized measures can then be used to identify the features of both the current and other segments. Importantly, while our example as focused on F1 and a small number of factors, it should be clear that it applies to any acoustic cue with any number of conditioning sources of variation.

C-CuRE predicts that as we partial out more factors from the signal, the residual cues should contain a clearer and clearer instantiation of the remaining factors affecting the cue. This may reveal feature of the current segment (e.g. revealing a feature masked by variability) or may provide hints to upcoming segments (making use of fine-grained detail to anticipate material) or prior ones. 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 cue 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.

The third, and most important, goal is to quantify this benefit. If identification of the target feature is better for relative than raw cues, by how much? If C-CuRE offers only a marginal benefit (or if raw cue-values are sufficient to support perception), it may not be worth the cost.

For our 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, the problem of recognizing vowels in the context of V-to-V coarticulation offers a domain in which there will be numerous sources of variability in the acoustic signal at any given point in time: speaker variability, the place and voicing of the following consonant (Hillenbrand, et al. 2001), and the height, backness and/or roundness of the subsequent vowel may all affect both F1 and F2 (as well as other cues). In this context, we can examine whether C-CuRE 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). This allows 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. In the case of cross-word coarticulation, without an active process of perceptual parsing as in the C-CuRE approach, simply lexicalizing the acoustic effects of coarticulation 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.

A thorough test of our hypothesis ultimately requires both production and perception studies. However, our initial work here focuses on production data alone. This provides the opportunity to ask whether C-CuRE (as instantiated in our model), can in principle 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 we find this not to be the case (in the current context), there would be no need to test actual listeners in a future study. 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.

2. Vowel-to-vowel coarticulation as a test case

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

It is well known that vowel sounds exhibit a complex pattern of coarticulation: the acoustic cues at any one point in time can show the influence of neighboring consonants or vowels (Hillenbrand, et al. 2001), long-distance effects from a vowel in an adjacent syllable, and even effects of vowels 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; Magen 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 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 coarticulation, the vowel in the earlier syllable becomes more similar to the vowel in the later syllable, and this 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 can be observed as a dynamic change in both the articulatory and acoustic form of a vowel. It is bidirectional, a result of the overlap of the gestures giving rise to each vowel, 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 changed 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 effects only arise within words, mechanisms like parsing in the C-CuRE framework 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 Linebaugh, Munson & McMurray 2010) 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 compensate for the influence of upcoming context, while at the same time using that variance to predict the context. To better understand the potential benefit of the C-CuRE approach for speech perception, we turn now to our test case, applying our formal model 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 of 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?

2.1 The corpus

We address these questions using a database from our previous acoustic investigation of V-to-V coarticulation (Cole et al. 2010). 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 driving 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.[2]

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 (i.e. they are not in edge or corner positions of the vowel space). The context vowels were three of the four "corner" vowels (/æ/, /ɑ/, /i/), which are most likely to induce coarticulation for these target vowels, and the matched target vowel (/ʌ/ or /ɛ/), which was expected to be neutral as a source of coarticulation. The fourth corner vowel /u/ was excluded to avoid introducing rounding coarticulation into the design.

Test words were chosen whose final consonant was a plosive. 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.[3] 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.

[2] F1 values of /a/ were examined to verify that speakers were producing the low vowel [ɑ] and not the rounded [ɔ], which is present in the speech of some Midwesterners. Data reported in Linebaugh (2007) show that the speakers in this experiment consistently produced all /ɑ/ tokens in a single cluster which was in the extreme low, back region of the vowel space of each speaker

[3] /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.

Table 1. Test phrases used in the experiment

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

For each of the target and context vowels F1 and F2 were measured at the midpoint with an LPC analysis. 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/. This left, a total of 1400 trials were included in the analysis.

In the primary report on this data (Cole et al. 2010), both F1 and F2 of target vowels were analyzed in a series of repeated measures ANOVAs to test the effects of V-to-V coarticulation, C-to-V coarticulation, and speaker on these formant frequencies. To briefly summarize the results, we found that voicing of the following C, but not place of articulation, influenced the F1 frequency of the target vowel, and that both voicing and place affected F2. An effect of target vowel identity (/ʌ/ vs. /ɛ/) was seen on both F1 and F2, though it was (predictably) much greater for F2.

Most importantly, the context vowel in the following syllable significantly affected both formants across all three target vowels tested (i.e. the context vowels conditioned changes in the target vowel F1 and F2 frequencies such that the target vowel more closely approximated the context vowel). Follow-up analyses showed that F1 was primarily affected by the height of the context vowel and F2 was affected primarily by its backness (though there were also effects of context vowel height on the target vowel's F2). Thus, it appears that both F1 and F2 are considerably affected by coarticulatory context from a number of sources.

3. Testing the parsing model

As it has been described, parsing in the C-CuRE approach offers two overlapping 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 (the identification of a second feature). Thus, our analysis proceeds in two steps. We first illustrate the 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 use the same model to show how accounting for overlapping variance dramatically improves the prediction of the upcoming context vowel.

3.1 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 four different models. The first, BASELINE, modeled a situation in which the listener engaged in no parsing whatsoever, and serves as a baseline upon which to evaluate the subsequent models. The second, GENDER, used the C-CuRE approach to parse out the effects of speaker gender from F1 and F2 before using these cues to categorize the target vowel. The third, SPEAKER, parsed out speaker gender as well as the specific speaker prior to categorizing the target vowel. Finally, the FULL model parsed out the place and voicing of the intervening consonant as well as the speaker identity.

In the baseline 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 effect 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 technique to map F1 and F2 jointly onto a discrete categorical output (the correct vowel) (see Jiang, Chen & Alwan 2004, for an example with single voicing cues). Logistic regression simply computes a weighted sum of the independent factors (identical to linear regression) and then transforms this through a nonlinear function that converts this linear value to a probability. It is a direct extension of the linear model to a problem in which the DV is a discrete event (e.g. a category) (see Hosmer & Lemeshow 2000, for a tutorial). Once we've estimated the equation of the model, we can then compute percent correct, as well as whether each cue was used to make a given distinction. Functionally logistic regression can be used similarly to discriminant analysis to categorize sets of continuous measurements (though there are important reasons to prefer logistic regression: Morrison & Kondaurova 2009). Thus, logistic regression can serve functionally as a sort of all-purpose classifier to convert continuous measurements to estimated categories. This parallels Nearey's (1997) use of it as the basis of the normalized a posteriori probability (NAPP) model.

For the present analysis (the base model), the predictors are raw F1 and F2 and the dependent measure is a dichotomous /ʌ/ or ɛ/ 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 on raw formant frequencies. 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, $p = .042$)[4] 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 based 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* to the vowel being predicted).

Given this baseline level of performance, the next model asked if accounting for variance due to the gender of the speaker using the C-CuRE approach could improve the model (the GENDER model). First, we used ordinary linear regression to generate equations predicting F1 and F2 from the gender. 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). We then recoded F1 and F2 as the difference from these expected formant frequencies (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

[4] The Wald Statistic is a standard statistic for computing whether the coefficient of a single factor in a logistic regression is significantly different than 0; that is whether or not that factor contributed significantly to the decision. This is much like the use of the t-statistic to test the hypothesis that a coefficient in a linear regression was significantly different than zero.

(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.

Table 2. Summary of the analyses identifying the target vowel (/ʌ/ or /ɛ/) from F1 and F2 frequencies after various contextual factors have been partialed out. "New variance" refers to the R^2_{change} statistic of only the new factors in that model (e.g. for the full model, the R^2_{change} of place and voicing, over and above speaker)

Model	**%Correct**	**New variance accounted for by context**	
		F1	**F2**
Baseline	90.5		
Gender	91.4	63.2%	35.9%
Speaker	92.8	19%	4.9%
Full	95.2	1.6%	14.5%

Table 3. Results of a regression analysis examining all sources of variation on F1. In parenthesis is the number of variables added to the model (when more than one was used). Note that variables with more than two values are recoded into N-1 dummy coded variables in standard regression practice. Thus, there were three places of articulation but only two variables needed to account for it

Step	**Variables**	**R^2_{change}**	**F_{change}**	**P**
1	Gender	.632	F(1,473) = 811.5	.0001
2	Subjects (9)	.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**		

These residuals were entered as the independent variables in the logistic regression described above. This model was somewhat better than the baseline model, 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, p < .001), only this time, F2 was a much stronger contributor than before (as evidenced by its increased Wald statistic).

We next asked if knowing the individual speaker could further improve performance with the SPEAKER model. Thus, we added individual speaker codes to the linear regression model 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 way 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 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 accounting for the effect of the neighboring consonant (the FULL model) yields even better performance. Here, the 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. Moreover, additional analyses suggest that further parsing out the influence of the context vowel can increase this to 96.2%.

This initial model illustrates that compensating for 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 (see Table 2 for a summary). 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 raw values: there is substantial overlap between the vowel categories, and a number of sub-clusters present. In Panel B, once these values are recoded in terms of their difference from the expected values for the gender, only two clusters remain (one for each vowel). In Panels C and D, formant values are now relative to expected values based on the individual speaker and the consonant (in addition to gender), 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 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 using just a few factors as the basis of encoding cue-values 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.

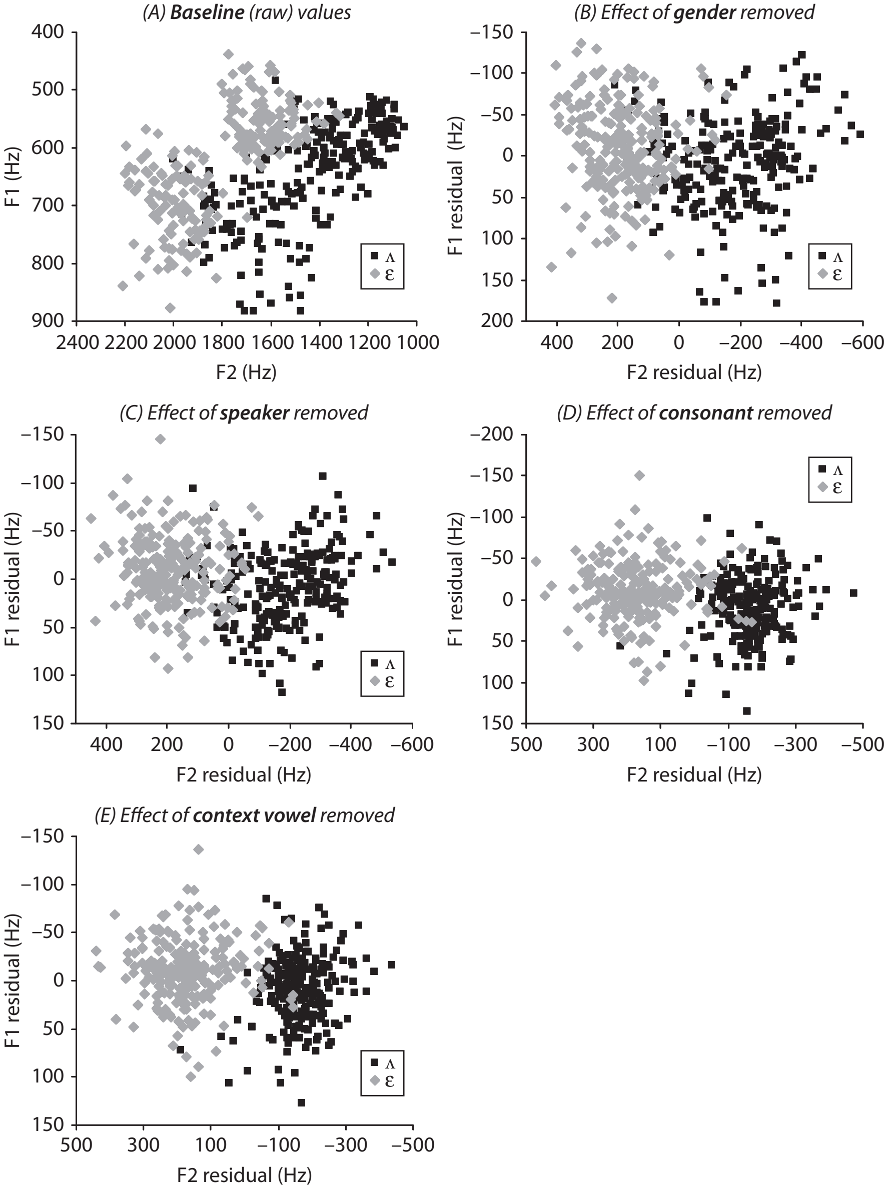

Figure 1. F1 and F2 for all tokens as a function of target vowel. (A) Raw values in Hz (note: regression analyses reported here were conducted on data transformed to Bark). (B) F1 and F2 frequencies after they have been recoded relative to expectations based on gender eliminated. (C) F1 and F2 after both gender and individual speaker are parsed from dataset. (D) Gender, speaker and now consonant have now been parsed out. (E) With the addition of context vowel to the parsing model there is virtually no overlap in the vowel categories. Note: Grey triangles (/ɛ/) are overlaid on black squares (/ʌ/) thus, some tokens of /ʌ/ may be masked on these figures

3.2 Anticipating the context vowel

The goal of the next 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 the raw F1 and F2 frequencies of the target vowel (the same dataset as before) along with an interaction term to predict the upcoming vowel (/i/, /æ/, /ɑ/, or the "same" vowel as the target 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, the relatively poor performance of this BASELINE model suggests that harnessing V2V coarticulation for anticipation is clearly a situation in which compensating for sources of variance in the signal has much to offer.

Our investigations of the benefits of the C-CuRE approach for predicting the context vowel examined 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). These linear regressions are identical to the ones on which the prior model was based – recoding F1 and F2 relative to the consonant is the same whether you are using the residuals to identify the target vowel or, as we are doing now, to predict the context vowel. However, to put these factors in perspective we 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 3 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 analysis of F2 (Table 4) showed that there was 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 backness 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%).

Table 4. Results of a regression analysis examining all sources of variation on F2

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**		

Given these models, we next examined the performance of the logistic regression classifier using F1 and F2 values from which various 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 the neutral context vowel which is the same as the target vowel) 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 5, and Figure 2 for a summary). The first (GENDER) assumed only that cues were computed relative to gender expectations – no detailed representation of speaker was available, nor could the model cope with coarticulation from the consonant. The second (SPEAKER), parsed out individual speaker means in addition to gender. The third model (VOWEL) accounted for the effect of target vowel in addition to speaker and gender. From the perspective of online processing, this represents the degree of prediction that can be made at the point at which the target vowel has been heard (without any subsequent context to provide a regressive "cleaning up" of the signal). The fourth model (FULL) recoded F1 and F2 relative to 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 account for coarticulation, but not for speaker factors.

Table 5. Performance of multinomial logistic regression models predicting the context vowel from various sets of F1 and F2. Percent correct refers to the number of tokens (in the corpus) for which the context vowel was correctly identified (chance would be 25%)

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

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.

Recoding cues relative to the 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: $\chi^2(3) = 27.7$, $p < .001$) though 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

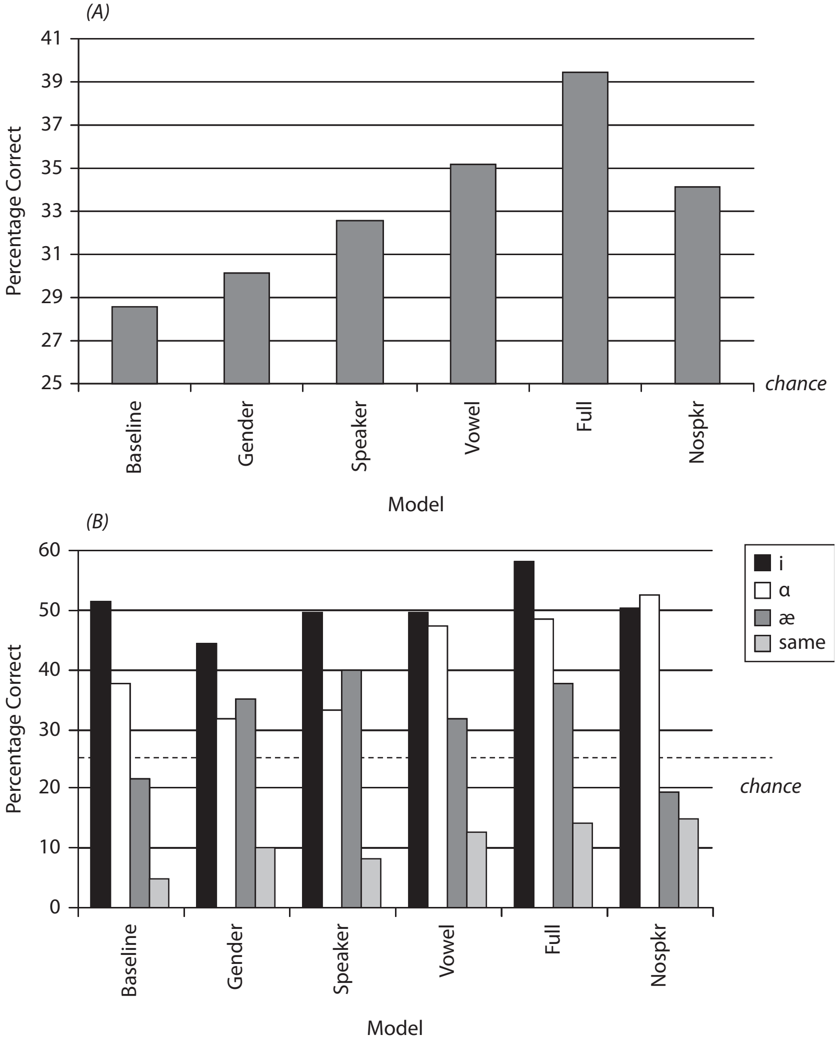

Figure 2. Summary of performance of six different parsing models on the problem of anticipating the context vowel. (A) Overall performance averaged across all four contexts. (B) Performance for each model, for each vowel. A clear pattern can be seen in which /i/ and /ɑ/ are consistently above chance, "same" responding is below chance for all models, and /æ/ is only above chance for models in which speaker factors have been parsed from the signal (GENDER, SPEAKER, VOWEL and FULL)

(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 the upcoming vowel (at least among these options) at well above chance levels. Moreover, analyses reported in Cole et al. (2010) suggest that after parsing in the

C-CuRE framework, 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 did not perform well, averaging 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%). This suggests that accounting for speaker variation can play an important role in leveraging V-to-V coarticulation.

Across these analyses, a couple of key findings are seen. First, prediction was quite good (particularly in the full model) and clearly well enough to be of substantial benefit in perception. This is largely due to our encoding of cue values relative to their expected values. Figure 3 shows an analogous vowel space to Figure 1. Panel A shows that without any compensation, the raw formant values cannot be clearly distinguished on the basis of context vowel. However, with after values are encoded relative to expectations based on speaker, consonant and the target vowel (Panel B), vowels preceding /i/ are generally higher and fronter; while those preceding /ɑ/ are lower and backer. Thus, parsing in the C-CuRE approach can not only uncover the target vowel, but it can also uncover information for other phonetic events lurking in the very same cues.

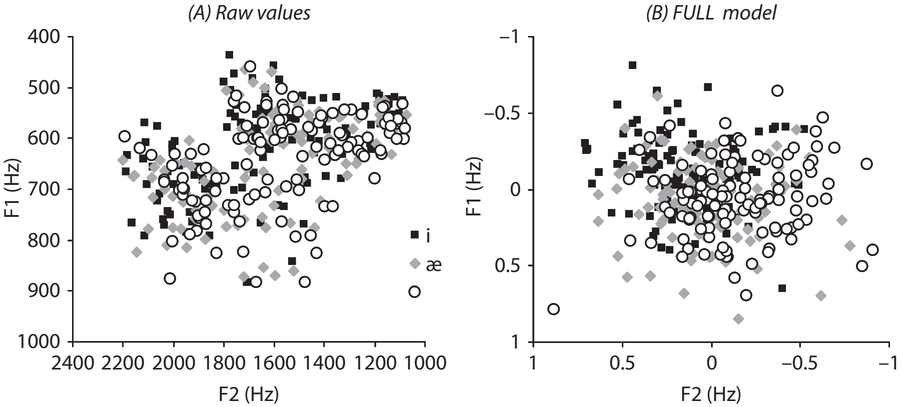

Figure 3. F1 and F2 for all tokens as a function of the context vowel. (A) Raw values in Hz (note: regression analyses reported here were conducted on data transformed to Bark). Very little systematicity can be seen with tokens for all three vowels scattered throughout the space. (B) The same F1 and F2 frequencies after the effects of gender, speaker, target vowel and consonant have been parsed out (the full model). Here, tokens preceding an /i/ (dark squares) generally appear in the top left, those preceding an /ɑ/ (open circles) are on the bottom right, and /æ/ on the bottom left. Note, we've left out the "same" condition for ease of presentation

Second, 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.

Third, compensating for 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%. Similarly, it is not sufficient to deal only with context effects due to coarticulation: the NOSPKR model performed 5% worse than the full model and was not able to predict all three vowels above chance. This has likely not been observed in prior studies, because most prior perceptual studies of parsing have used stimuli from a single speaker – listeners' ability to anticipate upcoming vowels may be attenuated in the context of several speakers.

Finally, no single source of variability is irrelevant – 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 also be quickly estimated on the fly from a few sentences). 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 relativizing cue values against a subset of the possible factors.

4. Discussion and conclusions

The C-CuRE framework 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) account 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 – even in a relatively general framework such as C-CuRE.

This approach allowed us to ask concrete questions about the benefits of computing acoustic cues relative to expectations driven by various factors that affect 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 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 4). As the utterance unfolds, the listener starts to identify the speaker (or minimally, speaker's gender) and generates expectations for how it will affect various acoustic cues. When the target vowel arrives, the incoming cues are compared to these expectations (Figure 4, Step 1), at which point the target vowel can largely be identified (Step 2), with an accuracy of 92.8%. This identification allows the listener (or the model) to refine expectations about how F1 and F2 should behave (step 3) and to use the residual differences to start to anticipate the upcoming consonant and the next vowel

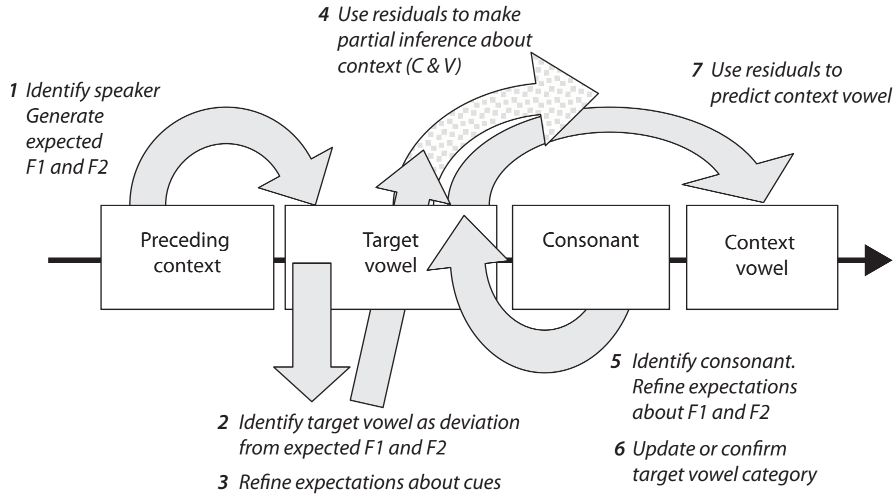
4 Use residuals to make partial inference about context (C & V)
1 Identify speaker Generate expected F1 and F2
7 Use residuals to predict context vowel
Preceding context
Target vowel
Consonant
Context vowel
5 Identify consonant. Refine expectations about F1 and F2
2 Identify target vowel as deviation from expected F1 and F2
3 Refine expectations about cues
6 Update or confirm target vowel category

Figure 4. Some hypothesized directions of parsing for target and context vowel identification

(Step 4) with an accuracy of 35.2%. Once the consonant is heard (Step 5), the listener can further revise their expectations for the values of the relevant cues, to revise their earlier decision about the target vowel or confirm that choice (since its accuracy will now be up to 95.2%), parsing out the variance in F1 and F2 associated with the consonant (Step 6). At this point, expectations for how F1 and F2 should behave are quite specific, and this allows listeners to capitalize on very small deviations to make a useful prediction about the upcoming vowel (Step 7) with a likely accuracy of 39.4%. Thus, parsing in the C-CuRE framework 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. Critically all this happens by comparing actual cue values (e.g. formant frequencies) to expectations driven by various aspects of the context: the speaker, the consonant and so forth. However, 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 what 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). Nonetheless, that does not seem to require us to go to exotic extremes – by only considering relatively straightforward sources of variance: speaker, place, voicing and the vowel one segment away, we were able to count for upwards of 85% of the variance in the acoustics. Incorporating additional factors may offer diminishing returns. Thus, it may be that considering a handful of simple factors is sufficient to reap large benefits in both identification and anticipation – as long as they are considered simultaneously. Moreover, the advantage we see here for parsing vowel variance may not be matched in parsing other cues for other types of sounds – this will largely be a function of the statistical properties of the cues in question and their relative dependence on phonetic context (though see McMurray & Jongman, submitted, for an application with fricatives).

Our model is consistent with a number of perceptual effects, such as Alfonso and Baer's (1982) finding that listeners can identify upcoming vowels at above chance levels from a preceding /ə/ alone; and Martin and Bunnell's (1981) finding that mismatching V-to-V coarticulation could delay perception. However, at face value, the importance of partialing out speaker factors would seem to conflict with findings that listeners are quite good at recognizing vowels when speaker varies randomly throughout the experiment (e.g. Andruski & Nearey 1991), even for "silent-center" vowels where steady state information has been eliminated (Jenkins, Strange & Miranda 1994). Such data, do not conflict with our model: even without parsing, our formant data was quite sufficient to identify the target vowel (and our logistic regression approach is formally similar to the NAPP model that Andruski and Nearey apply to their data). Parsing is not required for identification perception, and may only be needed to deal with mapping problems in which the information is much more ambiguous. Thus, evidence for using speaker information as a source of expectations for computing relative cues may need to come from more difficult perceptual tests (such as anticipation).

This model does have limitations, however. In particular, it makes two assumptions which at face value may be questionable. First, the ability to compensate for a factor like speaker 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 instance of the phoneme /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/ as its prediction of 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 listeners are 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 listeners 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. That is, the listener should have access to things like the mean F1 for high and low vowels; the mean F1 for different speakers and so forth.[5] This is critical for generating expectations against

[5] It remains to be seen whether listeners must track combinations of these means (e.g. an individual speaker's mean F1 for high vowels). Our model did not do this, simply parsing each individual factor sequentially. It performed quite well, suggesting that combinations may not be necessary, greatly reducing the overhead of such a model.

which to evaluate cues. This is also not so unreasonable (and requires substantially less long-term memory than an exemplar model). 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). There is also evidence more specifically that listeners track speaker's mean values of cues like VOT (Allen & Miller 2004). In fact, such information would be readily available from prototypes (as the mean is definitional for a prototype) in prototype accounts of speech categories (e.g. Miller 1997) or could be statistically extracted across exemplars in exemplar accounts (e.g. Goldinger 1998). Thus, the relevant information is already present in two popular accounts of phonetic categorization (McMurray & Farris-Trimble, in press).

Moreover, the parsing model is broadly consistent with work by Kluender and colleagues (e.g. Kluender, Coady & Kiefte 2003; Kiefte & Kleunder 2001) arguing that the auditory system processes incoming sounds in terms of their difference from long-term properties of the signal. For example, when spectral tilt was constant across a carrier sentence and a vowel, listeners ignored it as a cue for the vowel's identity; however when it changed between the carrier and the target it was an effective cue. Thus, it is not clear whether the "means" used as the basis of parsing could reflect linguistic or auditory cues, or whether they are determined over the course of seconds or years. Thus, while there is substantial work to be done on how (and if) listeners can track these measures of central tendency over different cue/category combinations. it is also clear that tracking such factors is something that listeners are likely to be capable of.

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 2009a; Toscano & McMurray 2010), and may offer a formal platform in which to integrate statistical learning and parsing.

Beyond these limitations however, the C-CuRE approach 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 and take advantage of 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, the generality of the C-CuRE model shows that parsing is not just useful for identifying and coping with the effects of contrastive phonological features. 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 2009b), 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). In some sense, the C-CuRE approach is relatively agnostic to the nature of the units, and this is on purpose. Any source of information that variation can be potentially assigned to is likely to be useful and we see no reason why strong theoretical considerations need to constrain the nature of the units that can participate in such operations.

Exemplar approaches also overlap considerably with C-CuRE, 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 (though in a way that could almost be called normalization, cf., Johnson 1997). However, unlike exemplar models in which the word is the unit that is stored, parsing can work across word boundaries. This offers considerably more flexibility than restricting the analysis and use of fine-grained detail to factors that can be lexicalized. 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.

Finally, at one level, our approach would seem to relate to enhancement theory (Stevens & Keyser, in press; Keyser & Stevens 2006). Partialing out a factor like speaker does indeed enhance the listener's ability to use that cue. However, this is quite orthogonal to the type of enhancement Stevens and Keyser discuss. They emphasize the way that additional gestures can enhance phonetic contrast (e.g. lip rounding for back vowels) or how pairs of acoustic cues can enhance each other (e.g. a high F0 enhancing perception of aspiration). In both cases, this concerns enhancement that arises from additional cues, not from interpreting them in light of context. However, enhancing gestures also pose an interesting challenge for parsing, as prior knowledge of enhancing relations b/w gestures could result in parsing that attributes part of a contrastive acoustic cue to the enhancing gesture (not the primary one). This could be particularly problematic when the enhancing gesture is not actually produced (as it would be expected). Further computational work, may be needed to understand if this is a problem. Nonetheless, at the broadest level, the incorporation of such enhancing factors could significantly enhance parsing models, by offering additional sources of bottom-up information.

Thus, C-CuRE as a general approach to compensation represents a novel synthesis of both gestural and exemplar-based theories that is also consistent with the notion of acoustic or gestural enhancement. It offers a unique explanation for the origin of discreteness in perception. Features are an emergent property of real-time perceptual processes that cope with the redundant variability in the speech signal. That is, discrete phonological contrast appears in real time as the variable speech signal contacts listeners' stored knowledge of phonetic categories and speakers. Contrast cannot be found in the listener or the signal, but in how the listener copes with the signal (see McMurray & Farris-Trimble, in press, for a discussion of the emergence of categories). Crucially, achieving this discreteness does not require processes like categorical perception that explicitly discard information. 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.

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 a grants NIH, DC008089, awarded to BM, and NIH HD044458 and NSF IIS 07–03624 awarded to JC.

References

Alfonso, Peter J. & Thomas Baer. 1982. Dynamics of vowel articulation. *Language and Speech,* 25 (2). 151–173.

Allen, Sean J. & Joanne L. Miller. 2004. Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America* 115. 3171–3183.

Allen, Sean J., Joanne L. Miller & David DeSteno. 2003. Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America* 113. 544–552.

Andruski, Jean E., Sheila E. Blumstein & Martha W. Burton. 1994. The effect of subphonetic differences on lexical access. *Cognition* 52, 163–187.

Andruski, Jean E. & Terrance M. Nearey. 1992. On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. *Journal of the Acoustical Society of America* 91 (1). 390–410.

Beddor, Pattrice S., James D. Harnsberger & Stephanie Lindemann. 2002. Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics* 30. 591–627.

Blumstein, Sheila E. & Kenneth N. Stevens. 1979. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America.* 66 (4). 1001–1017.

Blumstein, Sheila E. & Kenneth N. Stevens. 1981. Phonetic features and acoustic invariance in speech. *Cognition* 10. 25–32.

Browman, Catherine P. & Louis Goldstein.1992. Articulatory phonology: An overview. *Phonetica* 49. 155–180.

Carney, Arlene E., Gregory P. Widin & Neal F. Viemeister. 1977. Non categorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America* 62. 961–970.

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

Cohen, Jacob & Patricia Cohen. 1983. *Applied multiple regression/correlation analysis for the behavioral sciences,* 2nd *Edition*. Hillsdale, NJ: Lawrence Earlbaum.

Cole, Jennifer S., Gary Linebaugh, Cheyenne Munson & Bob McMurray. 2010. Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics* 38 (2). 167–184.

Connine, Cynthia. 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, Franklin S., Alvin M. Liberman & John M. Borst. 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, Carol A. 1981. Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech & Hearing Research* 24. 127–139.

Fowler, Carol A. 1984. Segmentation of coarticulated speech in perception. *Perception & Psychophysics* 36. 359–368.

Fowler, Carol A. 1996. Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America,* 99 (3). 1730–1741.

Fowler, Carol A. 2005. Parsing coarticulated speech in perception: effects of coarticulation resistance. *Journal of Phonetics* 33. 199–213.

Fowler, Carol A. & Julie M. Brown. 2000. Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Perception & Psychophysics* 62 (1). 21–32.

Fowler, Carol A. & Mary R. Smith. 1986. Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In Joseph Perkell & Dennis Klatt (eds), *Invariance and variability in speech processes*, 123–136. Hillsdale, NJ: Erlbaum.

Fry, Dennis B., Arthur S. Abramson, Peter D. Eimas & Alvin M. Liberman. 1962. The identification and discrimination of synthetic vowels. *Language and Speech* 5. 171–189.

Gafos, Adamantios & Stefan Benus. 2006. Dynamics of phonological cognition. *Cognitive Science* 30. 905–943.

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

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

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

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

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

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

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

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

Gow, David & Bob McMurray. 2007. Word recognition and phonology: The case of English coronal place assimilation. In Jennifer S. Cole & Jose I. Hualde (eds) *Laboratory Phonology 9,* 173–199. Berlin: Mouton de Gruyter.

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

Helgason, Petur & Catherine Ringen. 2008. Voicing and aspiration in Swedish stops. *Journal of Phonetics* 36 (4). 607–628.

Hillenbrand, James M., Laura Getty, Michael J. Clark & Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97 (5). 3099–3111.

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

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

Hosmer, David W. & Stanley Lemeshow. 2000. *Applied Logistic Regression, 2nd Edition*. New York: John Wiley & Sons, Inc.

Jenkins, James, Winifred Strange & Salvatore Miranda. 1994. Vowel identification in mixed-speaker silent-center syllables. *Journal of the Acoustical Society of America* 95 (2). 1030–1043.

Jiang, Jintao, Marcia Chen & Abeer Alwan. 2006. On the perception of voicing in syllable initial plosives in noise. *Journal of the Acoustical Society of America* 119 (2). 1092–1105.

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

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

Keyser, Samuel J. & Kenneth N. Stevens. 2006. Enhancement and overlap in the speech chain. *Language* 82. 33–63.

Kiefte, Michael J. & Keith R. Kluender. 2001. Spectral tilt versus formant frequency in static and dynamic vowels. *Journal of the Acoustical Society of America* 109 (5). 2294–2295.

Kluender, Keith R., Jeffrey A. Coady & Michael J. Kiefte. 2003. Sensitivity to change in perception of speech. *Speech Communication* 41 (1). 59–69.

Liberman, Alvin M., Katherine S. Harris, Howard S. Hoffman & Belver C. Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54 (5) 358–368.

Liberman, Alvin M. and Ignatius G. Mattingly. 1985. The motor theory of speech perception revised *Cognition 21 (1).* 1–36.

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

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

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

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

Manuel, Sharon 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, James G. & Timothy H. Bunnell. 1981. Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America* 69 (2). 559–567.

Martin, James G. & Timothy H. Bunnell. 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, Dominic W. & Michael M. Cohen. 1983. Categorical or continuous speech perception: a new test. *Speech Communication* 2. 15–35.

Maye, Jessica, Daniel J. Weiss & Richard N. Aslin. 2008. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science* 11 (1). 122–134.

Maye, Jessica, Janet F. Werker & LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82 (3). B101-B111.

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

McMurray, Bob, Richard N. Aslin, Michael K. Tanenhaus, Michael J. Spivey & Dana Subik. 2008. Gradient sensitivity to within-category variation in speech: Implications for categorical perception. *Journal of Experimental Psychology: Human Perception and Performance* 34 (6). 1609–1631.

McMurray, Bob, Richard N. Aslin & Joseph Toscano. 2009a. Statistical learning of phonetic categories: Computational insights and limitations. *Developmental Science* 12 (3). 369–379.

McMurray, Bob & Ashley Farris-Trimble. in press. Emergent, yet unintended, coupling of perception and production: General processing principles and statistical learning. In Abigail Cohn, Cécile Fougeron & Marie Huffman (eds) *The Oxford Handbook of Laboratory Phonology*, Oxford, UK: Oxford University Press.

McMurray, Bob, Jessica Horst, Joseph Toscano & Larissa Samuelson. 2009b. Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. Invited submission for John Spencer, Michael Thomas & James McClelland (eds) *Toward a new grand theory of development? Connectionism and Dynamic Systems Theory reconsidered*. London: Oxford University Press.

McMurray, Bob & Allard Jongman. under review. What information is necessary for speech categorization? Assessing the informational assumptions of models of speech perception with a corpus of fricatives.

McMurray, Bob, Michael Tanenhaus & Richard N. Aslin. 2009c. Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition, *Journal of Memory and Language* 60 (1). 65–91.

McMurray, Bob, Michael Tanenhaus & Richard N. Aslin. 2002. Gradient effects of within-category phonetic variation on lexical access, *Cognition* 86 (2). B33–B42.

Miller, Joanne L. & Lydia E. Volaitis. 1989. Effects of speaking rate on the perceived internal structure of phonetic categories. *Perception & Psychophysics* 46. 505–512.

Miller, Jonanne L. 1997. Internal structure of phonetic categories. *Language and Cognitive Processes 12.* 865–869.

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

Monaghan, Padraic, Nick Chater & Morten H. Christiansen. 2005. The differential role of phonological and distributional cues in grammatical categorisation. *Cognition* 96. 143–182.

Morrison, Geoffrey S. & Maria V. Kondaurova. 2009. Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis (L). *Journal of the Acoustical Society of America* 126. 2159–2162.

Munson, Cheyenne & Bob McMurray. 2007. 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.

Nearey, Terrance M. 1997. Speech perception as pattern recognition. *Journal of the Acoustical Society of America* 101. 3241–3256.

Ohala, John J. 1981. The listeners as a source of sound change. In C.S. Masek, R.A. Hendrick, & M.F. Miller (eds), *Papers from the Parasession on Language and Behavior*, 178–203. Chicago: Chicago Linguuistics Society.

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

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

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

Pierrehumbert, Janet. 2003. Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech* 46. 115–154.

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

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

Repp, Bruno. 1984. Categorical perception: Issues, methods and findings. In Norman Lass (ed) *Speech and Language (vol. 10): Advances in Basic Research and Practice*, 244–335. Orlando: Academic Press.

Recasens, Daniel & Maria D. Pallarès. 2000. A study of F1 coarticulation in VCV sequences. *Journal of Speech, Language & Hearing Research* 43. 501–512.

Salverda, Anne Pier, Delphine Dahan & James M. McQueen. 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition* 90 (1). 51–89.

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

Schouten, Bert, Ellen Gerrits & Arjan Van Hessen. 2003. The end of categorical perception as we know it. *Speech Communication* 41 (1). 71–80.

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

Stevens, Kenneth N. & Samuel J. Keyser. 2010. Quantal theory, enhancement, and overlap. *Journal of Phonetics* 38 (1). 10–19.

Sussman, Harvey, David Fruchter, Jon Hilbert & Joseph Sirosh. 1998. Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Science* 21 (2). 241–299.

Toscano, Joseph & Bob McMurray. in press. Using the distributional statistics of speech sounds for learning and combining multiple acoustic cues. *Cognitive Science.*

Toscano, Joseph & Bob McMurray. 2007. 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, Jennifer A., Sheila E. Blumstein & Martha W. Burton. 2000. Effects of subphonetic and syllable structure variation on word recognition. *Perception & Psychophysics* 62 (6). 1297–1311.

Yang, Charles. 2004. Universal Grammar, statistics or both? *Trends in Cognitive Sciences* 8 (10). 451–456.