English obstruent perception: The content of the p-map and its relationship to natural class structure

Noah Silbert[a]
nsilbert@umd.edu

Kenneth de Jong[b,c]
kdejong@indiana.edu

a. Center for Advanced Study of Language
University of Maryland
7005 52nd Ave.,
College Park, Md. 20742

b. Department of Linguistics, Indiana University
Memorial Hall 322
1021 E. Third Street
Bloomington, IN
47405-7005

c. Department of Cognitive Science, Indiana University
Eigenmann 819
1910 E. 10th St.,
Bloomington, IN
47406-7512

Abstract.

This paper reports a series of analyses of consonant identification data from Cutler, Weber, Smits, & Cooper (2004) and which empirically derive perceptual similarity to determine how it compares with feature-based models of phoneme similarity. A model in which minimal contrast structure is equivalent to perceptual structure provides a baseline 'null hypothesis'. This baseline model is compared to 1) an empirically fitted choice model that does not require intersegment similarity to bear any particular relation to feature structure, 2) models which add an explicit interaction between features, and 3) a set-theoretic model in which feature structure relates to perceptual structure via the number of each pair of segments' shared and non-shared natural classes (the Natural Class Similarity Model, Frisch 1996). These model comparisons indicate that features interact pervasively in perception and that syllable position modulates these interactions. Pairs of similarity parameters were then compared across the listeners in order to identify contrasts that systematically differ from one another in similarity (which we define as a failure in *Feature Equivalence*), and contrasts that systematically differ in similarity depending on the segments in which the contrast appears (which we define as a failure in *Feature Independence*). Some of these identified differences could be accounted for by the set theoretic model, though this model systematically underestimates the distinctiveness of sibilants from other segments, does not account for the robustness of coda voicing contrasts, and overestimates the distinctiveness of dental fricatives. Finally, multidimensional scaling analyses indicate that the expected phonological contrast structure was present in the confusion patterns, despite the pervasive and robust presence of irregularities. The results robustly document a pattern of perceptual irregularity in the face of an overall pattern of featural contrast structure, and are discussed in terms of a number of factors involved in relating perceptual confusability to phonological systems.

Keywords: perception, phonological features, confusions, syllable position, perceptual similarity

1. Introduction

1.1. Similarity and the P-map

In a very commonly cited manuscript, Steriade (2001) argues for a perceptual component of phonological grammar called the p-map. The p-map is taken as a repository of information concerning the relative similarity of pairs of contrastive phones in different prosodic locations. While this proposal has been commonly cited, the content of this map is not commonly discussed in detail, and systematically determining the precise content of the p-map is not a standard part of the phonological research employing it. The current paper seeks to outline a method for determining the content of the p-map, and to apply this method to determine the contents of the p-map for anterior obstruents in English, as they appear in pre-vocalic and post-vocalic position.

Because the p-map is meant to account for irregularities in the similarity of phonological contrasts in different contexts, the approach taken here is to identify cases in which measures of perceptual similarity between minimally contrastive phones deviate from a model in which contrasts are simple categorical distinctions unaffected by the specification of other contrasts or prosodic position. That is, we take as a starting point a model in which segments contrast in dichotomous featural units which are all of the same 'size', and which combine with other units orthogonally to create contrasting segments. Assuming *only* representations made of dichotomous and orthogonal contrasts, we can define a phonological 'space' with 'dimensions' corresponding to contrasts. An auditory phonetic space defined solely by such dichotomous contrasts would be an extremely simple version of the p-map, one in which the p-map would not account for any phonological irregularities beyond traditional representations, which might allow for having more than one feature contrasting segments.

We will call one important attribute of this perceptual space *Feature Equivalence* (*FE*). *FE* is the property that each contrast is 'equally contrastive'. Couched in terms of a spatial representation, *FE* states that the opposing elements across a given contrast are no further apart nor closer together than the elements across any other contrast.[1] Two implicit corollaries to *FE* are readily apparent. First, contrasts are independent. *Feature Independence* (*FI*) implies, e.g., that voicing does not affect place of articulation, nor does place affect voicing.[2] Second, *FE* holds regardless of prosodic context. A

[1] Feature independence here is distinct from perceptual independence in the General Recognition Theory framework (Ashby & Townsend, 1986).

[2] At first glance, the hierarchical structure of feature geometry seems to predict a failure in *FI*, but this is illusory, at least with regard to terminal nodes in a feature 'tree'. A given (terminal-node) feature takes its value regardless of the values taken by other (terminal-node) features within the same segment. Feature geometric representations allow for grouping via hierarchical dependency of features, but not for segment-internal (sister) feature interaction. For example, a voiced segment is no more or less voiced if it is coronal or labial. Whether a feature is an element in a vector or a (subordinate) node in a tree, it is specified independently of other elements in the vector or (sister) nodes in the

contrastive feature in onset position is no different than that same feature in coda position.[3]

There is plenty of evidence that *FE* does not always hold, that is that not all features are equally distinctive in perceptual behavior. For example, *FE* was addressed directly and found to fail in judged similarity data for stops varying in voicing and place of articulation (Greenberg & Jenkins, 1964). Evidence for the failure of *FE* in speech perception can also be found in Miller and Nicely's seminal study of consonant identification in noise (1955; see also Shepard, 1972). Similar research on other data sets has found a number of different patterns of feature non-equivalence (and some limited agreement in such patterns) across a variety of tasks, noise masking conditions, languages, and prosodic contexts (Ahmed & Agrawal, 1969; Benkí, 2003; Graham & House, 1970; Singh & Black, 1966; Singh, Woods, & Becker, 1972; Wang & Bilger, 1973; Wilson, 1963). In addition, cross-linguistic variation in phoneme inventories hints that feature interaction can play a role in diachronic change. For example, voiceless labial stops and voiced velar stops tend to be missing from consonant inventories more often than are voiced labials and voiceless velars (Hayes, 1999).

In addition to fitting stipulated feature sets to confusion data (e.g., Miller & Nicely, 1955[4]), attempts have also been made to *deduce* feature sets from perceptual confusions. Further evidence for the failure of *FE* and *FI* is found here, as well, suggesting that interactions are not due simply to the use of the wrong features. Klatt (1968) employed an optimization routine in a re-analysis of Wickelgren's (1966) short term memory consonant confusion data and found manner, place, and voicing interactions; contrasts in one feature are confused in a manner that depends on the presence of other features. Non-hierarchical clustering and individual-differences non-hierarchical clustering analyses provide still more evidence of such interactions, whether couched explicitly in terms of phonetic categories (Shepard & Arabie, 1979) or in terms of mixed categorical and gradient properties of the phones in question (Soli, Arabie, & Carroll, 1986).

A major aspect of our method of evaluating the content of the p-map concerns the comparison of individual listeners. Since the content of the p-map is taken to be part of a system that is common to members of a linguistic community, the information contained in it should be evident in most, if not all, of the individuals in the linguistic community. Thus, we look for regularities in deviations from the simple feature counting model across individual listeners.

The p-map, of course, is not the first proposal for integrating perceptual structure into explanations of aspects of phonological systems. Frisch (1996) works out a proposal

---

tree. Thus, neither an unordered bundle nor a hierarchical feature tree provides a mechanism by which contrasting elements can interact.

[3] Metrical and prosodic phonological models that explicitly address super-segmental structure do not provide mechanisms for feature-by-prosodic-context interactions. These must be added with phonetic interpretation rules.

[4] The feature set employed by Miller and Nicely was, at the time of publication, a set chosen somewhat arbitrarily for the particular consonant stimuli used in their experiments.

concerning how similarity might be derived systematically from a traditional feature representation, which does not require a specific repository of information about confusability such as the p-map. In the natural classes similarity model (NCSM), the similarity of a pair of segments is determined by the number of natural classes that the two segments share relative to the number that they do not share, not directly in terms of the features which distinguish them. The NCSM has been used successfully to explain gradient patterns of co-occurrence restrictions between consonants in Arabic triconsonantal roots (Frisch, Pierrehumbert & Broe, 2004), as well as attendant well-formedness judgments of non-words (Frisch & Zawyadeh, 2001), suggesting an important role for class-determined perceptual similarity in at least some phonological patterns.

A second goal, then, of the current paper is to examine the degree to which patterns of similarity derived from perceptual confusions in a phoneme identification task differ from measures of similarity yielded by the NCSM. Three outcomes are possible. If both methods agree, this would provide powerful converging evidence for the general role of natural class similarity in phonological structures and in perceptual confusability. Bailey & Hahn (2005) have conducted such a comparison previously, and proposed a simple feature counting model of similarity as the best measure of phoneme similarity across a number of production, perception, and relative similarity judgment data. They claim that this simple feature-counting model outperforms the NCSM. To the extent that this conclusion is supported in the current analysis, it would suggest that natural class similarity is distinct from simple feature-counting (perceptual) similarity. To the extent that empirically-derived perceptual similarity exhibits a different structure from both a feature-counting baseline and that predicted by the NCSM, our results motivate a third type of information, reasonably interpreted as something like the specific content of the p-map in Steriade's proposal.

1.2. Mathematical models of psychological space

1.2.1. A data-based measure of psychological similarity

The basic task at hand is to determine, from perceptual data, the psychological proximity of categories of objects and the psychological space in which such objects reside. Perceptual data can take a number of forms. For example, such data may take the form of explicit judgments of (dis)similarity of pairs of stimuli. The data analyzed below, however, are identification confusions. Raw confusions are of limited utility, though, requiring some link between the likelihood that two objects are confused with one another and a measure of the similarity between the two.

In explicit (dis)similarity judgment tasks, finding inter-stimulus similarities is reasonably straightforward. In identification confusion tasks, on the other hand, obtaining inter-stimulus similarities is not so simple. A powerful, well-motivated formulation of similarity is given by the similarity choice model (or SCM; Shepard, 1957; Luce, 1963). The SCM models the probability of responding, say, "t" when presented with [d] as the normalized product of the similarity between [t] and [d] and a bias toward responding "t", as expressed in (1). Here, $s(t,d)$ represents the estimated (symmetric) similarity between stimuli [t] and [d], which is weighted by $b(t)$, the bias toward responding "t" regardless of

what the stimulus is. The probability then is obtained by expressing the weighted similarity of the pair relative to the weighted similarities of [d] to all of the other stimuli in a given experiment.

$$(1) \qquad p(t \mid d) = \frac{b(t)s(t,d)}{\sum_{k=1}^{n} b(k)s(t,k)}$$

One of the SCM's greatest pragmatic strengths is that, to fit the model, the maximum likelihood similarity and bias parameters can be calculated via a closed form solution (i.e., there is no need to search for the best fitting parameters). The formula for the model parameter representing the similarity between [t] and [d] is given in (2); the similarity between any other pair of stimuli would simply require substitution of the appropriate identification-confusion probabilities.

$$(2) \qquad s(t,d) = \sqrt{\frac{p(t \mid d)p(d \mid t)}{p(t \mid t)p(d \mid d)}}$$

Although using SCM similarity parameters may have implications for a spatial or clustering model of a data set, the SCM's similarity parameters need not be interpreted with regard to perceptual space. For example, consideration of the ordinal relationships between SCM similarity parameters need not involve any assumptions about the number or nature of spatial, psychological dimensions the stimuli are situated in. Thus, use of the SCM allows for more or less direct assessment of the relationships between inter-stimulus proximities. *FI* and *FE* (or their failure) should be evident in these relationships, so there is no pressing theoretical need to transform the data and represent the inter-stimulus proximities spatially or in terms of clusters, although this will be done below for the sake of exposition and illustration.

1.2.2. The natural class model of similarity

The SCM provides a simple method for estimating inter-segment similarity based on observed confusion frequencies, which in turn allows for a straightforward method for evaluating *FE* and *FI*. However, predictions concerning similarity can also be obtained directly from the feature structure of segments. For example, Frisch (1996) proposed a measure of inter-segment similarity based on the number of shared and non-shared natural classes (cf. the feature counting model of similarity in Pierrehumbert, 1993). This is given in (3). Here, $s(t,d)$ represents the estimated (symmetric) similarity between [t] and [d], $n(t \cap d)$ represents the number of natural classes shared by [t] and [d], and $n(\overline{t \cap d})$ represents the number of natural classes *not* shared by [t] and [d].

$$(3) \qquad s(t,d) = \frac{n(t \cap d)}{n(t \cap d) + n(\overline{t \cap d})}$$

As mentioned above, this model of similarity has proven very successful in accounting for the gradient distribution of triconsonantal roots in Arabic (Frisch, Pierrehumbert, & Broe, 2004), and there is some evidence that it accounts for various forms of behavioral data, as well (e.g., speech error 'confusions', Frisch, 1996; speech error and perceptual confusions, Bailey & Hahn, 2005).

The simpler feature counting model of similarity can be expressed in the same general form as the natural class model given in (3) by substituting the number of shared and non-shared features for the number of shared and non-shared natural classes. Because we are concerned with minimal contrasts in our evaluation of *FE* and *FI* (see *Methods* section below), this simple feature counting model predicts the null hypothesis with regard to *FE*, namely that similarities across any and all minimal contrasts are identical. The feature counting and natural class models of similarity provide grammar-based sets of predictions concerning *FE* and *FI*, while the similarity choice model provides estimates based on perceptual behavior. Comparison of the three models' predictions should allow us to pinpoint aspects of the perceptual behavior that are not accounted for by a feature-counting model or the NCSM (i.e., that are candidates for an additional perceptual component in patterns of similarity, providing the content of the p-map, as posited).

1.2.3. Individual versus group data

There is one further point that needs to be addressed before quantifying *FE* and *FI*, and this concerns potential differences between individuals. When considering a phonological system (or, even more, what Steriade would call the content of a grammar), it is clear that the system must be a property of the set of people using the system. However, there is good reason to be wary of confusion and similarity judgment data that are pooled across participants (Ashby, Maddox, & Lee, 1994; although see Lee & Pope, 2003, for one way to mitigate the negative effects of using averaged data). Of primary concern when averaging data is the artificial inflation of symmetry in the averaged confusion matrix.

To avoid these problems, a more conservative approach to comparing individuals is taken in the present study. The data is analyzed to test for the failure of *FE* and *FI* in rank-order differences between individual participants. Consonant confusion data from Cutler et al (2004) were chosen in large part because the availability of data for each individual participant enabled reasonably powerful tests of *FE* and *FI* as properties of a shared phonological system. However, because the data under consideration were not gathered with these issues or tests in mind, it seems unwise to make strong assumptions about the relationship between similarity and distance in perceptual space in our analysis. Furthermore, because MDS and hierarchical clustering models would take the SCM similarity parameters (or some other measure of similarity) as input, and because the rank order of similarities is the focus of our analysis, taking the extra steps of transforming similarities to distance and fitting a clustering or spatial model is unnecessary for

addressing our primary questions. Thus, our main analysis focuses primarily on the ordinal relationships between measures of similarity.

2. Methods

2.1. Data

The data under consideration in the present study are a subset of the consonant confusion data presented in Cutler, et al., (2004), in which 16 native American English and 16 native Dutch listeners identified 24 consonants and 15 vowels in two prosodic positions and 3 signal-to-noise (multi-talker babble) ratios (0, 8, and 16 dB). Thus, our data consist of consonant identifications for consonants appearing either before or after each of 15 English vowels in simple consonant-vowel or vowel-consonant syllables at each of the three signal-to-noise ratios. For more detailed description of the participants, materials, and procedure employed in data collection, see Cutler, et al., (2004).

We have restricted our analysis to the native American English listeners' perceptual confusions of the anterior obstruent consonants: [p], [b], [f], [v], [θ], [ð], [t], [d], [s], and [z]. These consonants represent a reasonably neat, pseudo-factorial combination of contrasts in voicing, place, and manner of articulation. This arrangement of consonants (i.e., feature specifications) allows for the direct assessment of *FE* and *FI* with regard to a fairly large number of minimal contrasts across a number of features. A graph indicating the general contrast structure tested here is given in Figure 1, in which consonants connected by single lines are taken to be minimally contrastive.

In some standard feature sets employed in phonological analysis, a number of the consonant pairs connected in Figure 1 are not considered minimally contrastive, because they contrast on more than one feature. The labiodental fricatives differ in both manner and place from the bilabial stops, the interdental fricatives bear an analogous relationship to the alveolar stops, the alveolar fricatives may be specified as 'strident' while the interdental (and possibly the labial) fricatives may be specified as 'distributed,' and 'voicing' contrasts in stops are often taken to involve features specifying properties of the glottal aperture. The assumed feature structure of these consonants (one of a number of possibilities) is given in Table 1. Note that the labial sounds are specified as 'distributed.' Because labial obstruents involve contact of articulators over a relatively large area, one could argue that, under a phonetic interpretation of this feature, they are, in fact, distributed. However, under at least some feature systems, (e.g., the feature geometric system outlined in Clements & Hume, 1995), 'distributed' is only specified for coronal consonants. Within the consonant set described here, any feature that distinguishes between interdental and alveolar fricatives (e.g., 'distributed', 'strident', 'grooved', etc.) must be partially redundant with at least one of the other features employed.

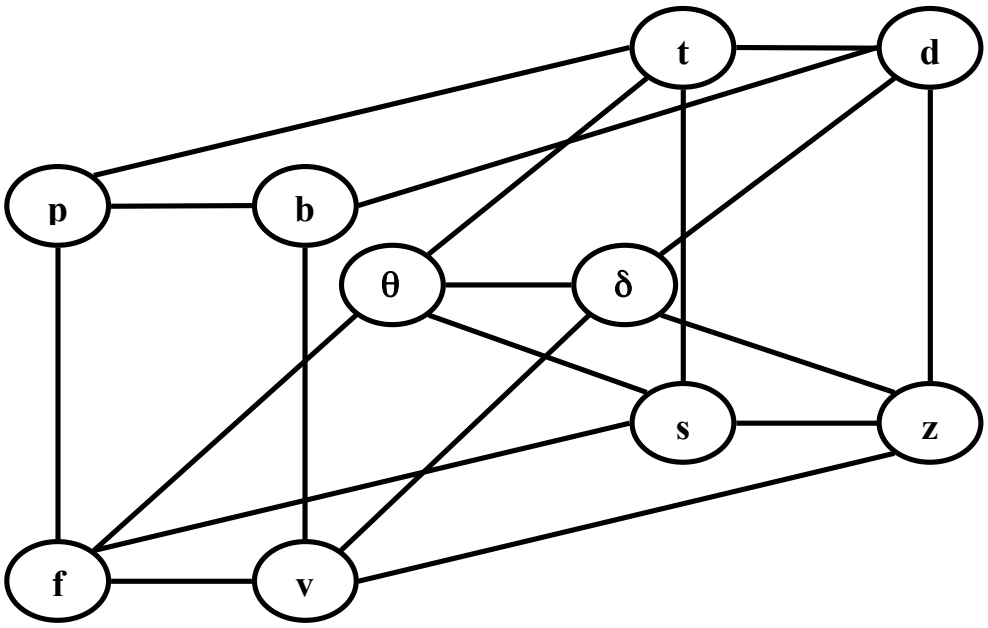

Figure 1: Contrast structure of English anterior obstruents. Lines represent minimal contrasts between connected segments.

Nonetheless, the contrast structure given in Figure 1 is a sound baseline for the current analysis. First, and foremost, there are no consonants 'in between' any of these contrasts in English (e.g., English does not have bilabial fricatives, interdental stops, etc.). In this sense, each contrast represented in Figure 1 is minimal. Second, features such as 'distributed,' 'strident,' and 'spread glottis' are contrastively redundant to other features, and so potentially allow for a minimal contrast to be considered as 'twice as large', by dint of having two specified features for one contrast. For example, it is well understood that 'voicing' contrasts in stops are very different (in production) than the 'same' contrast in fricatives (i.e., the relative timing and degree of oral constriction gestures and glottal oscillation; see Stevens, 1998), and 'distributed' and 'strident' capture the fact that place of articulation for fricatives has dramatic implications for the acoustic signal. (The same is true, although perhaps less dramatically, for stops.) Some of the contrasts, then, in Figure 1, might fail to exhibit *FE*, but this is not evidence for perceptual irregularity beyond what can be captured by employing additional features. Examining the available analyses for each failure has to be done on a case by case basis.

Table 1: Binary feature specifications employed in hierarchical model fitting

	p	t	f	θ	s	b	d	v	ð	z
voiced	–	–	–	–	–	+	+	+	+	+
coronal	–	+	–	+	+	–	+	–	+	+
continuant	–	–	+	+	+	–	–	+	+	+
distributed	+	–	+	+	–	+	–	+	+	–

2.2. Analysis

Three analytic procedures are used to investigate the relationship between the SCM model of the perceptual confusions and a simple minimal-contrast model with *FI* and *FE* and between the SCM model and the NCSM model predictions.

First, to determine whether the overall confusion data is adequately modeled in terms of a simple minimal-contrast model with *FE* and *FI*, we constructed a number of SCM models of the raw data which encode different degrees of *FE* and *FI* (described in section 2.2.1), and compared their fit to the data. The models differed systematically in how much independent variation is allowed for the various segments; thus, comparing the fit of the different models gives an overall measure as to the necessity allowing various degrees of interaction between the features.

The second analysis takes a different tack in examining the consistency across individuals in the SCM similarity estimates for specific pairs of segments. This analysis allows us to determine specific cases where there is evidence for failures of *FE* and *FI.*

Finally, to allow for visualization of the overall structure of the similarity space, the similarity parameters were submitted to an individual-differences multidimensional scaling analysis.

2.2.1. Analysis 1, SCM Model Comparisons

For the purposes of model comparison, the consonants under consideration were specified on the four binary feature dimensions given in Table 1 above. In the set of models the similarity between a given pair of consonants was defined as the product of a set of parameters indicating (mis)match on each feature dimension. If stimuli *i* and *j* match on feature *k*, then $f_k = 1$, whereas if stimuli *i* and *j* mismatch on feature *k*, then $0 < f_k < 1$. Thus, any mismatch represents a decrease in similarity. Nested models encoding different feature interactions varied the number of feature 'dimensions' *n* on which pairs of stimuli (mis)match.

The most restricted form of this model disallows any interaction. In this case, there are four f_k parameters – one for each of the 'plain' feature dimensions given in Table 1. A slightly more general model could, for example, allow for interaction between voicing and manner. In this case, the product of *f* parameters determining the similarity between a pair of fricatives that mismatch on voicing would include a unique mismatch parameter $f_{v\text{-}fric}$, a pair of stops that mismatch on voicing would include a different unique mismatch parameter $f_{v\text{-}stop}$, and a stop-fricative pair that mismatch on voicing would include a third unique mismatch parameter $f_{v\text{-}fs}$. An interaction between voicing, place, and manner would require 9 mismatch parameters, and an interaction between voicing and all three other features would require 27. Thus, the most general restricted (relative to the full SCM) model allowed each of the features to interact with each of the other features, in which case 81 unique *f* parameters were required. Each unique set of mismatch parameters (i.e., each restricted model) has fewer similarity parameters than does the full SCM. The most restricted model produced 8 similarity parameters, while the most general produced 31. With 10 stimuli, the full SCM has 45 similarity parameters.

Restricting the number of similarity parameters cannot improve the fit of the SCM, although it can fail to worsen it. Goodness of fit for a pair of nested models (i.e., a pair in which one model is a restricted version of a more general model) is measured by the difference in G^2 between the two. This difference is distributed as a χ^2 random variable with degrees of freedom equal to the difference in number of free (similarity) parameters between the two models. The full SCM for an *n*-by-*n* confusion matrix has $n(n-1)/2$ similarity parameters and $n - 1$ bias parameters, and the saturated model, which predicts cell frequencies exactly, has $n(n-1)$ parameters. The G^2 statistic for the full SCM is first evaluated by checking it against the critical χ^2 for degrees of freedom equal to the difference in number of parameters between the full and restricted model, which in this case is 90 – 54 = 36. Fits of the full SCM indicate that all but one participants' G^2 is below the critical value in the present analysis. Thus, the full SCM fits the data quite well overall, and provides a good platform for the model comparisons. The single poor fit could be a real departure from the constraints put on the data by the SCM, or it may well be a false alarm, a reasonably likely occurrence given the large number of statistical tests performed.

To execute the model comparisons, data for each of the 16 native English participants, an optimization method of estimation called iterative proportional fitting was used to find maximum likelihood confusion frequencies based on SCM models that differed in their ability to encode differences between segments. The model fitting procedure addresses all inter-phoneme similarities simultaneously. This provides strong statistical evidence for the utility of allowing for failures of *FE* and *FI* in modeling the perceptual confusions.

2.2.2. Analysis 2, Group Consistency Analysis

The second analytic procedure takes a different tack at testing for failures of *FE* and *FI*, which has the added benefit of allowing us to identify segments in which such failures appear. For each of the 16 native English participants, iterative proportional fitting was used to find maximum likelihood confusion frequencies based on the full SCM as described in equation (1). These frequencies were converted to probabilities and transformed into similarity scores according to equation (2). Similarity parameters were calculated for the connected pairs of segments in Figure 1. The parameters for each individual for each pair were then compared with a) pairs differing in the dimension of contrast (e.g. voicing vs. manner), or b) with other pairs differing in the same dimension of contrast (e.g. voicing in labial stops vs. voicing in labial fricatives). Although stronger assumptions about the magnitude of the similarity parameters might have allowed for more informative tests, we have restricted our analysis to ordinal (i.e., greater or less than) relationships, because we are wary of making strong assumptions about the magnitude of the similarity parameters because the data were not collected with this sort of modeling in mind.

Specific comparisons were chosen as follows. We relate *FE* to the SCM similarity parameters in the following way. If *FE* holds, $s(i, j)$, the similarity of consonants *i* and *j* contrasting in one dimension, should not be consistently greater or less (across individuals) than $s(i, k)$, the similarity of consonants *i* and *k* contrasting in a different dimension. Thus, the similarity parameter for a given minimally contrastive pair of consonants, e.g., [p] and [b], should equal to the similarity parameter for a

'neighboring' pair of consonants contrasting on a different dimension, e.g., [p] and [f]. Given the contrast structure in Figure 1, any consistent ordinal relationship between these two pairs would serve as evidence of the failure of *FE*. If, on the other hand, [p] and [f] are taken to differ in more than one feature specification, if [p] and [f] are consistently more similar than are [p] and [b], this is strong evidence that *FE* does not hold at the level of feature specifications, either – in this case, two features together would appear to be 'less contrastive' than a neighboring minimal, single feature contrast. Similar arguments can be constructed for the other locations in which the contrast structure of Figure 1 differs from contrast structures based on other feature specifications. Of course, the opposite is not true; if we find that [p] and [b] are consistently more similar than [p] and [f], we cannot be sure that this asymmetry due to the failure of *FE*, rather than the presence of additional, redundant features in the case of [p] and [f].

We relate *FI* to the SCM similarity parameters in the following way. If *FI* holds, $s(i, j)$, the similarity of consonants *i* and *j* contrasting in one feature dimension, should not be consistently greater or less than $s(k, l)$, the similarity of consonants *k* and *l* contrasting in the same dimension but differing from *i* and *j* on a second dimension. So, if *FI* holds, 'parallel' similarity parameters, e.g., between [p]-[b] and [t]-[d], should not differ consistently across individuals. If there is a consistent ordinal relationship, it reflects a failure of *FI* at the contrast level, such that the confusability of the voicing contrast depends on the value of the place contrast. It is possible that some such differences in confusability might correspond to cases in which redundant features are involved in one pair of segments, but not the other, e.g. the 'distributed' feature is redundant in the [p]-[f] contrast, but not in the [t] – [s] contrast.

Of course, due to random error and the effect(s) of any unmeasured, related variables, equality between pairs of inter-stimulus similarities can be expected to fail occasionally even if *FE* and *FI* actually hold. Given the problems associated with group data discussed above, and given that *FE* and *FI* are taken to be properties of a language's (i.e., not an individual speaker's) phonology, we treat each participant as an independent 'experiment' testing these predictions of *FE* and *FI*. If the failure of *FE* or *FI* is a general property of the shared English system, we should expect a large preponderance of the participants to exhibit the same ordinal relationship between a given pair of similarity parameters. We then test the reliability of the participants' patterns of ordinal similarity parameter relations by conducting a series of binomial tests. That is, we compare the probability that *n* out of 16 participants' data exhibit the same ordinal relation between a given pair of similarity parameters to the critical values under the *FE/FI* null hypothesis that the probability of one similarity parameter being greater than another is 0.5.

Thus, to test the relationship between the similarity between [p] and [b] and the similarity between [p] and [f] in a given prosodic position, the SCM similarity parameters $s(p, b)$ and $s(p, f)$ are calculated for each of the 16 participants. The number of participants exhibiting a particular relationship between these two similarity parameters is then tallied and compared to the 'critical number', the number taken to be sufficiently unlikely to have occurred purely by chance under the assumption that the null hypothesis is true. Because each test involves 16 total 'trials,' and because the order of each of the specific relationships tested is arbitrary, the same (two-tailed) critical values serve as cutoffs for every test. With $\alpha = 0.05$, the critical values are 4 and 12. That is, under the

assumption that the null hypothesis is true, it is taken to be sufficiently unlikely that either less than 5 or more than 11 participants would exhibit a particular ordinal relationship between two minimally contrastive pairs of consonants if chance is the only operative factor.

Because of the large number of tests being carried out – there are 42 comparisons of pairs involving one shared consonant and 26 comparisons of 'parallel' pairs – getting a statistically significant result by chance becomes a near certainty. That is, with $\alpha = 0.05$, we expect 1 out of every 20 tests to be falsely statistically significant. Given 68 tests, then, we can reasonably expect 3 or 4 false alarms. To mitigate this risk, we perform a Bonferroni adjustment, dividing the alpha level by the number of tests conducted, making the significance criterion more conservative. In this case, $\alpha / 68$ requires 15 or 16 of the 16 participants to exhibit the same ordinal relationship between a given pair of similarity parameters to obtain a statistically significant result.

2.2.3. Analysis 3, Individual Differences MDS

In order to visualize the similarity relationships tested here, we also submitted matrices of similarity parameters (16 each for onset and coda positions) to an individual-differences MDS routine. Individual differences MDS employs a 'group space' – a set of coordinates that all the participants share – and individual participant weights that selectively exaggerate or diminish the importance of the group space dimensions for each participant. The individual participant weights from the MDS solution can also serve to augment the binomial test results by providing additional information regarding how consistent the participants were.

We can gauge how good the fit of the model is with a measure of how much variance in the data is accounted for – R^2 – and a 'Stress' statistic (see, e.g., Kruskal, 1964a, b; Shepard, 1962), which measures the divergence from a strictly monotonic relationship between each participant's psychological proximities and the distances between the points in the appropriately weighted space. In addition, it is also standard practice to evaluate a non-metric MDS model by whether we can interpret the dimensions of the derived space. However, because the distribution of the Stress statistic is not particularly well understood (compared to, say, commonly used *t* and *F* statistics), and because the original experiment was not designed with spatial modeling in mind, the spatial results of the MDS analysis will be used primarily for visualization of the data and to help interpret the other, more rigorous analysis.

3. Results

3.1. Analysis 1, SCM Model Fits

For onset position, *any* restriction on the number of similarity parameters resulted in an appreciable reduction in the goodness of fit; only the full SCM fit the observed onset confusion frequencies well (i.e., had a statistically non-significant G^2 fit statistic). For coda position, restricted models fit the observed confusion frequencies reasonably well (i.e., restrictions did not statistically significantly worsen the fit) for only six of the sixteen participants, while the remaining ten participants required the full SCM. For all

six of these participants, the best balance between reduction in the number of similarity parameters and goodness of fit was found with models requiring interaction between at least two features. All six participants' best fitting restricted models required interaction between manner and 'distributed' (the partially redundant feature employed here). Five of the six participants' best fitting restricted models required interaction between 'distributed' and place. Four of the six participants' best fitting restricted models required voicing to interact with all three other features; the other two required lower-order voicing interactions. A different set of four participants' best fitting restricted models required two-way interactions involving place.

Thus, fitting contrast based models with varying degrees of interaction between the different contrasts supports the conclusion, suggested by previous literature, that perceptual confusions are persistently and pervasively infused with featural interactions. There is no sense in which one can say that the consonant confusions in the data set are created simply by the neat, minimal dichotomous and orthogonal contrasts in Figure 1 plus some noise (cf. the feature counting model in Bailey & Hahn, 2005). There is much evidence for structure in the perceptual confusions beyond that captured by a simple orthogonal contrast model.

To examine the localization of the failures of *FE* and *FI* indicated here in more detail, as well as to determine their consistency across the corpus, we turn now to binomial tests of specific comparisons.

3.2. Analysis 2, Group Consistency Analysis

3.2.1. Binomial tests of ordinal relationships between similarity parameters

For the onset stimuli, 55 out of the 68 total tests were statistically significant using the non-adjusted α level (i.e., $\alpha = 0.05$). With the Bonferonni adjustment, 34 of the 68 tests were statistically significant. Within the test results indicating failure of *FE* (e.g., a statistically significant difference between the similarity of [t] and [d] and the similarity of [t] and [p]) there are 36 statistically significant ($\alpha = 0.05$) cases of failure of *FE*, and 21 of these were significant after Bonferonni adjustment. Within the test results indicating failure of *FI* (i.e. comparisons of the confusability of a pair of segments with that of another pair contrasting in the same feature), we see 19 failures of *FI* with $\alpha = 0.05$, with 13 of these still significant after Bonferonni adjustment. Even if 3 or 4 of these statistically significant results are false alarms (i.e., due to chance alone), the data indicate very strongly that neither *FE* nor *FI* hold in general among these onset English obstruents.

A similar, though clearly not identical, pattern of results holds for the coda stimuli. For the coda stimuli, 51 out of the 68 tests were statistically significant using the non-adjusted α level. With the Bonferonni adjustment, 36 tests were statistically significant. Of the tests indicating failure of *FE*, 29 were statistically significant with $\alpha = 0.05$, and 19 of these were statistically significant after Bonferonni adjustment. Of the tests indicating failure of *FI*, 22 were statistically significant with $\alpha = 0.05$, and 17 of these were statistically significant after adjustment. As with the onset data set, the danger

of false alarms explaining these results is minimal.[5] The differences between onset and coda patterns are discussed in detail further below (section 3.2.3)

3.2.2. Inequalities predicted by the NCSM

This section explores the degree to which the natural class model of similarity (NCSM) can explain the consistent inequalities found in the previous section. The form of the NCSM ensures that the rank order of similarities within a set of segments is constant whether the set is considered alone or as part of a larger set. Thus, previously published estimates of similarity for the set of consonants considered here (Frisch, 1996) are used to predict where the NSCM would expect inequalities. The pairwise similarity values predicted by the NSCM for the ten consonants analyzed here, taken from Frisch, 1996, are given in Table 2; similarity for the minimal contrasts are shown in bold.

Table 2: Natural Class Similarity Measures (from Frisch, 1996)

	p	b	f	v	t	d	T	D	s	z
p	1	**0.4**	**0.26**	0.15	**0.3**	0.14	0.11	0.07	0.1	0.06
b		1	0.13	**0.3**	0.14	**0.28**	0.06	0.12	0.05	0.11
f			1	**0.38**	0.1	0.05	**0.43**	0.19	**0.18**	0.09
v				1	0.06	0.11	0.19	**0.39**	0.1	**0.19**
t					1	**0.39**	**0.2**	0.12	**0.3**	0.17
d						1	0.11	**0.23**	0.15	**0.33**
T							1	**0.38**	**0.4**	0.19
D								1	0.2	**0.44**
s									1	**0.37**
z										1

From Table 2, the NSCM predicts 46 of the similarity inequalities that were observed in the similarity parameters from the Cutler, et al. (2004) perceptual data. In 33 other cases, the NCSM predicts inequalities where no consistent inequality is observed in the perceptual data, and in 49 cases, the NCSM predicts a relationship *opposite* that

[5] Wilcoxon signed-rank tests were employed to compare the rank order of similarities of all pairs of consonants (i.e., not just the minimally contrastive pairs) across onset and coda position within subjects. Eight subjects had statistically significantly different (rank ordered) similarities in onset and coda ($p < 0.05$; five with $p < 0.01$), suggesting a moderate modulation of failures of *FE* and *FI* due to syllable position.

observed in the perceptual data. In 2 cases, the NCSM predicts equality between two similarities where an inequality was actually observed.

Dividing tests between those examining *FE* and those examining *FI*, we find a roughly equal distribution of matches across the sets. With regard to the onset *FE* results, the NCSM had 18 matches (predicted what is observed), 19 opposite mismatches (prediction is opposite in direction from what is observed), and 7 failures to match (cases where a predicted inequality is not observed). With regard to the onset *FI* results, the NCSM had 12 matches, 7 opposite mismatches, and 8 failures to match. With regard to coda *FE* results, the NCSM predicts 14 correct ordinal relationships, 15 opposite relationships, and 14 unobserved relationships. With regard to coda *FI* results, the NCSM correctly predicts 12 inequalities, and predicts the opposite of what is observed in 10 inequalities.

Overall, about a third of the *FE* and *FI* test results match the predictions of the NCSM. The divergence between the NCSM and the fitted SCM model was confirmed by Wilcoxon signed-rank tests comparing the rank order of the similarities of all pairs of consonants (i.e., not just the minimally contrastive pairs) for the NCSM and each individual subject. In every case, for both onset and coda data, the signed rank tests were statistically significant with $p < 0.001$. Thus, though some of the feature interactions might be explained with the NSCM, much of the current observed patterns cannot be so explained.

3.2.3. Surveying feature interactions.

In order to make the specific results of these analyses available for reference in studies of perception in phonology, such as targeted by the p-map hypothesis, this section discusses the specific pairs where consistent differences in similarity are found. Surveying the matches and mis-matches of the similarity estimates and those predicted by the NCSM also allows us to evaluate where the NCSM might account for the inequalities in the Cutler et al data set.

Tables 3 – 7 list specific ordinal differences in perceptual similarity found to be significant using the binomial tests. In the right hand column are inequalities not predicted by the NCSM. Also ‘chains’ of inequalities in these tables indicate cases in which all of the ordinal differences obtain. So, e.g. Table 3 indicates the following chain: t-d > p-t > t-θ, meaning that /t/ and /d/ are less similar than /p/ and /t/ are, and also than /t/ and /θ/ are, and in addition, /p/ and /t/ are less similar than are /t/ and /θ/. Results with 12 – 14 participants in agreement (i.e., results that are statistically significant with $\alpha = 0.05$) are indicated with ‘>’, and results with 15 or 16 participants in agreement (i.e., results that are statistically significant after Bonferonni adjustment) are indicated with ‘>>’. Tables 3 and 5 list the binomial test results indicating failure of *FE*; pairs always contrast in a different feature, since we are comparing similarities for two different features. Tables 4 and 6 list the binomial test results indicating failure of *FI*; pairs always involve the same feature, since we are looking to see if the same feature in a different pair of segments has the same degree of similarity. Following, we will go over the patterns for onsets first, followed by those for codas.

Table 3 indicates that the contrast between sibilants and other segments is systematically the largest (always appearing on the left of each inequality), and the voicing and place contrasts involving dental fricatives are the smallest (always appearing to the right of each inequality). There is no obvious overall pattern of certain contrasts being generally larger than others, though, since the stop – non-sibilant contrasts are smaller than the stop – sibilant contrast (e.g. t-θ vs. t-s). This could indicate the presence of a double (two-feature) contrast in the cases involving sibilants. However, we note also that place contrasts in stops are generally less confusable than place features in non-sibilant fricatives (e.g. p-t vs. p-f, both of which are less confusable than f-θ). Similarly, voice contrasts in t-d are among the least confusable, while voice contrasts in θ-ð are the most confusable.

Table 3. Summary of failures of *FE* for consonants in onset position.

Observed, not predicted by NCSM	Observed and predicted by NSCM
	Θ-s > > Θ-ð
t-d > b-d >> d-ð	d-ð > > Θ-ð
t-d >> d-ð	d-ð > v-ð
d-z >> b-d	b-d > > b-v
p-b >> b-v	v-z > > b-v
d-z >> v-z	v-z > f-v
v-ð > Θ-ð	f-v > > v-ð
f-Θ > Θ-ð	
t-d > p-t > t-Θ	t-Θ > > f-Θ
t-d >> t-Θ	t-Θ > > Θ-ð
p-b >> p-f	f-s > > p-f
	p-f > f-Θ
	f-s > > f-v
	f-v > > f-Θ
ð-z > > Θ-ð	f-s > s-z
t-s > f-s	s-z > ð-z
t-s >> p-t	t-s > t-d
p-t >> p-f	t-s >> s-z
s-z > v-z	d-z >> s-z

Thus, since the similarity of contrasting segments depends on which segments are being contrasted, Table 3 also shows indirect evidence for failure of *FI*. These indirect indicators of feature interaction are corroborated by the direct comparisons, listed in Table 4. Voicing in θ-ð is more confusable than in f-v, s-z, and t-d. In addition to this, there is a general interaction between place and voicing; voicing in labials tends to more confusable than in coronals. Manner contrasts also are systematically more confusable in labials. Also as noted with respect to Table 3, sibilants are systematically less confusable with stops than are non-sibilants.

Table 4. Summary of failures of *FI* for consonants in onset position.

Observed, not predicted by NCSM	Observed and predicted by NSCM
	t-d > p-b
f-v >> Θ-ð	s-z > f-v
t-d >> Θ-ð	s-z > > Θ-ð
t-s >> t-Θ	
t-s >> p-f	t-Θ > p-f
d-z >> d-ð	d-ð > b-v
d-z >> b-v	t-s > d-z
	f-s >> f-Θ
	p-t >> f-Θ
	Θ-s > > f-Θ
	Θ-s > ð-z
ð-z > > v-ð	v-z >> v-ð
	b-d >> v-ð

Focusing on place contrasts reveals that non-sibilant fricatives systematically have more place confusions than other segments; f-θ and v-ð are consistently more similar than p-t and b-d, respectively. Note that these place interactions are where many traditional phonological representations would employ multiple, redundant features to encode the differences in fine point of articulation between the stop and the fricative. This might be taken to point to the utility of extra redundant features distinguishing stops that are not distinguishing the non-sibilant fricatives. However, note also two parallel patterns of similarity inequality: the consistently higher similarity between f-θ relative to f-s and θ-s and the consistently higher similarity between v-ð relative to v-z and ð-z. The similarity of the sibilants to either of the non-sibilants is less than their mutual similarity, regardless of how place features would distinguish the non-sibilants.

Tables 5 and 6 list failures of *FE* and *FI* for codas. As can be seen in Table 5, the coda *FE* results also exhibit 'tiers' of contrast types with many of the patterns found in the onset data. Place contrasts involving labial and dental fricatives are more similar than many of the manner contrasts, and one of the laryngeal contrasts (f-v), which are often, in turn, more similar than *other* place contrasts. A number of these 'middle tier' place contrasts are more similar than the laryngeal contrasts involving stops.

Table 5. Summary of failures of *FE* for consonants in coda position.

Observed, not predicted by NCSM	Observed and predicted by NSCM
t-d > d-z > b-d > d-ð	
t-d >> d-ð	d-ð > > v-ð
	v-z > > b-v
p-b >> b-v	b-v > > v-ð
	v-z > f-v
	f-v > > v-ð
t-d >> t-Θ	t-Θ > > f-Θ
t-s >> f-s	f-s > p-f
t-s >> p-t	p-f > > f-Θ
p-b >> b-d	f-s > f-v
p-b >> p-f	f-v > > f-Θ
t-d >> p-t	f-s >> s-z
	t-s >> s-z
Θ-s > Θ-ð	d-z >> s-z
ð-z > Θ-ð	t-Θ > > Θ-ð

In Table 6, the coda *FI* results also show many of the interactive effects found in onset position. Again, voicing interacts with manner, and voicing in stops is less confusable than in fricatives. There is also a difference between coronal and labial non-sibilants, where, again, the contrasts in dentals are smaller. Manner interacts with both voicing and place, though this effect seems largely indicative of a larger contrast for sibilants than for non-sibilants. Place interacts with manner, sibilants being least confusable, stops more confusable, and non-sibilants even more confusable. Place also interacts with voicing, though less systematically; f-θ is more confusable than v-ð, but p-t is not systematically more confusable than b-d.

Table 6. Summary of failures of *FI* for consonants in coda position.

Observed, not predicted by NCSM	Observed and predicted by NSCM
p-b >> f-v	
f-v >> Θ-ð	
t-d >> Θ-ð	s-z > > Θ-ð
t-d >> s-z	
d-z >> b-v > d-ð	t-Θ > > d-ð
t-s >> t-Θ	t-s >> d-z
t-s >> p-f	
	f-s > p-t >> f-Θ
	Θ-s > > f-Θ
ð-z > > v-ð	Θ-s > ð-z
	v-z > b-d >> v-ð > > f-Θ

Comparing Tables 3 and 4 with Tables 5 and 6 shows that, although the gross patterns found in the onset and coda data are quite similar, there are a number of interesting differences. Among the tests of *FE* (Tables 3 and 5), place contrasts tend to exhibit relatively more confusions than other contrasts in the coda position. For example, in coda position, b-v and p-f are more confusable than δ-v and θ-f, respectively, but this is not true in onsets. Also, in onset position the d-z and t-s manner contrasts are both consistently less confusable than any other contrast analyzed, and t-s is less confusable than the t-d voicing contrast. In coda position, however, this relationship is reversed; voicing contrasts tend to be larger in coda position, relative to manner contrasts.

There are also fewer 'transitive' connections in coda position (chains), providing less evidence of feature interaction. Thus, changes in prosodic context also cause changes in the pattern of results directly indicating failure of *FI* (Table 4 vs. Table 6). While voicing interacts primarily with place and sometimes with manner in onset position, both interactions appear quite robust in coda position. The only evidence of voicing by manner interaction in onset position is in the difference between θ-δ and t-d, whereas a voicing by place interaction is evident within each manner class. In coda position, on the other hand, two of the voicing by place interactions are 'replaced' by voicing by manner interactions – the fricative voicing contrasts are more similar than their homorganic stop voicing contrasts. Additionally, no voicing interaction relationship is exhibited by less than 15 of the participants in coda position, whereas in onset position two voicing interactions are.

A similar 'strengthening' of interaction patterns can be observed for the place by manner and place-by-place interactions in coda position. The parallel structures in onset position are complicated in coda position by the appearance of place by voicing interactions. The indirect evidence of interaction discussed above is made more direct by

manner of articulation interacting more heavily with both place and voicing in coda position. Again, the parallel structure across voicing specification observed in onset position is more complicated in coda position, in which the voiced alveolar manner contrast is more similar than its voiceless counterpart. The symmetry of manner and place interactions across voicing specifications among onset consonants is lost in coda position, where a more complicated three-way interaction appears. Finally, the place interactions are also more complex in coda than in onset position. All of the place interactions exhibited in onset position were exhibited in coda position, as well, and additional place by manner and place by voicing interactions occur in coda position.

3.2.4. Feature interaction mismatches with predictions by the NCSM.

Finally, we will relate these patterns of inequalities to those predicted by the NCSM, as encoded in the column segregation in Tables 3 - 6. Evaluating the relationship between NCSM and observed similarity parameters, as in section 3.2.2, might lead one to believe that there is no systematic relationship between confusability and natural-class similarity. However, a closer examination of which interactions are mismatched with NCSM predictions suggests that there may be a discernable relationship between the two.

Since it is much easier to characterize the divergences between the NCSM predictions and what is observed by examining failures of *FI*, we begin with Tables 4 and 6. In Table 4, 3 of the 5 observed voicing interactions, 6 of the 7 place interactions, and 3 of the 7 manner interactions matched. Thus irregularities in place similarity are almost exactly predicted by the NCSM (except that δ-z is systematically less similar than δ-v). Mismatches in manner are due to the NCSM systematically overestimating the confusability of sibilants and stops. Mismatches in voicing are due entirely to the NCSM underestimating the confusability of the two dental fricatives.

Turning to Table 3, a number of the mismatches appear to be due to the same or related issues. For example, the similarities for t-s, d-z and δ-z are observed to be less than neighboring contrasts, contrary to predictions of the NCSM, suggesting, again, that the NCSM tends to underestimate the size of sibilant – non-sibilant contrasts. Complementing this are a set of mismatches where the NCSM is overestimating the size of the stop – non-sibilant contrast, thus getting mismatches between manner (b-v and p-f) and voicing (f-v and p-b), as well as between manner (d-δ and t-θ) and place (b-d and p-t). Also, as found in Table 4, the similarity for θ-δ relative to place contrasts is consistently predicted by the NCSM to be less than what is observed. Also t-d are more similar in the NCSM than they are in the current data, relative to nearby contrasts.

Turning to codas (Table 6), we again find that place interactions are generally as predicted by the NCSM (with the same exception found for onsets involving δ-z and v-δ). However, most of the voicing (4 of 5) and manner (5 of 7) inequalities are mismatches. For voicing, the NCSM does not predict the observed reduction in similarity of voicing contrasts for stops relative to fricatives. Also, as with onsets, the dental fricative contrasts are smaller than the NCSM would predict. For the manner contrasts, 4 of the 5 mismatches are, again, due to the NCSM systematically overestimating the similarity of sibilants to other segments, an effect already noted for onsets.

In Table 5, many of the mismatches, similarly are due to the NCSM overestimating voicing similarities in coda stops (8 mismatches), overestimating the similarity of sibilants and stops (3 mismatches), and underestimating the similarity the dental fricatives (2 mismatches). These three account for all of the mismatches in direction, except for 1 case, d-ð relative to b-d.

3.3. Individual differences MDS analysis

The present data show numerous interactions between contrasts and numerous inequalities in the 'size' of contrasts on different dimensions. Indeed, one might have the impression, as suggested by Wang and Bilger (1973), that the structure of our baseline featural representation has nothing to do with perceptual confusability at all. It turns out that, for both onset and coda positions, MDS analysis indicates that the observed confusion frequencies are, in fact, reasonably well characterized by a 'three dimensional' contrast structure, similar to that given in Figure 1, despite the frequent failures of *FI* and *FE*.

The results of the MDS fitting are presented in Figures 2 - 5. For each prosodic position, a three-dimensional group space is displayed in three two-dimensional plots, plotting in turn dimension 1 vs. 2 and 2 vs. 3. Figures 2 and 4 present the group space for onset and coda consonants, respectively, and Figures 3 and 5 plot the individual participant weights.

The solution plotted in Figure 2 is a good fit, producing Stress = 0.147 and R^2 = 0.770, and thus accounts for nearly 80% of the variance in the similarity parameters. In addition, examining Figure 2 reveals that the three dimensions created by the model-fitting algorithm are readily interpretable: dimension 1 corresponds to voicing, dimension 2 to manner, and dimension 3 to place specifications. Thus, the plot of dimension 1 against dimension 2 has all of the voiceless segments to the right and voiced segments to the left, and has stops in the lower half and fricatives in the upper half. There is, of course, some divergence in that the solution conflates the sibilant fricative contrast with manner of articulation, in a sense treating the sibilant [s] and [z] as something like 'super-fricatives', and [θ] does not come out in the fricative half of the manner dimension relative to voicing. Overall, though, the patterns in the solution are readily interpretable in terms of featural contrasts.

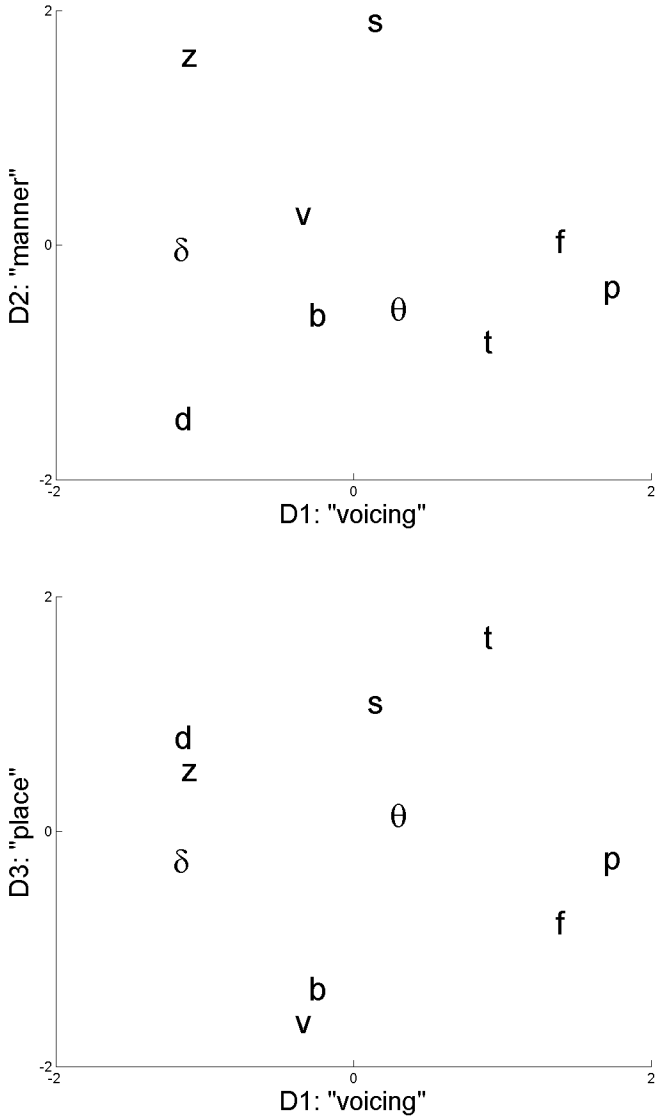

Figure 2. Three-dimensional group space for MDS model of onset consonant confusions from Cutler, et al. (2004).

Figure 3 shows one more notable aspect of the model, namely a high degree of similarity in the individual participant weights. This provides further evidence, along with the remarkable number of effects in which 15 or 16 participants agree in the analyses above, that the listeners behaved very consistently with respect to each other.

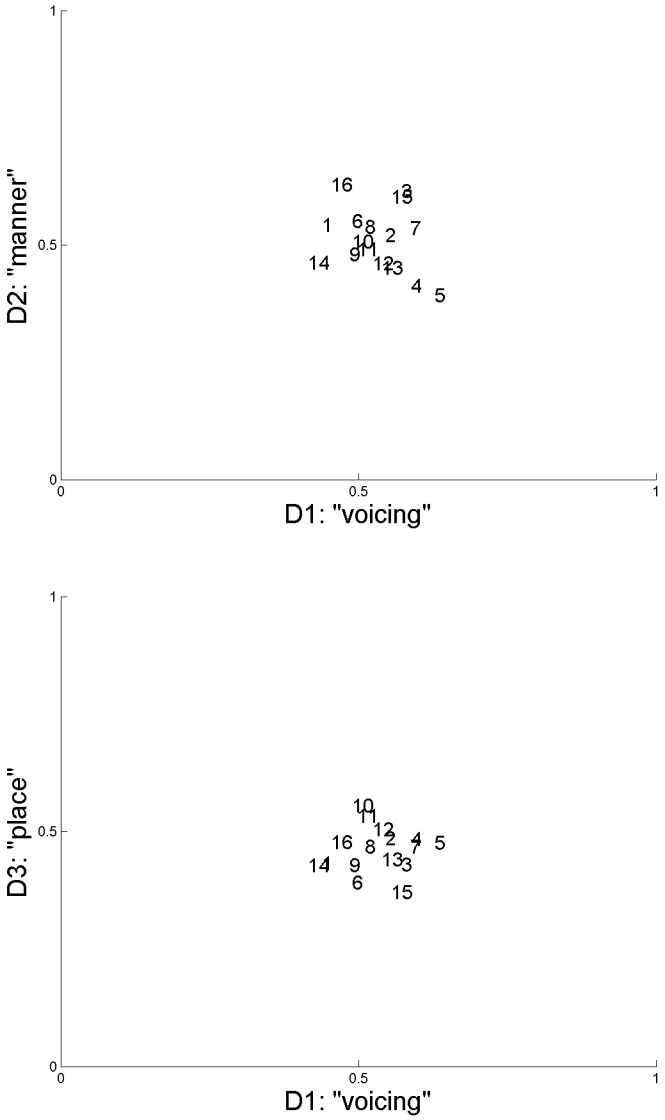

Figure 3. Individual weights for each dimension in an MDS model of onset confusions from Cutler et al. (2004).

Figures 4 and 5 present the group space and participant weights, respectively, for consonants in coda position. For coda position, the three dimensional solution was also a good fit, producing Stress = 0.151 and R^2 = 0.769, again accounting for nearly 80% of the variance in the data. As with onsets, the three dimensions are readily interpretable as the three featural dimensions of contrast, dimension 1 corresponds again to voicing, dimension 2 to place, and dimension 3 to manner. Here the voicing division in dimension 1 is noticeably enlarged. Again, there are some divergences from this one-to-one relationship: the sibilant contrast is grouped with the place contrast, making [s] and [z] 'super-coronals,' [b] appears very near its non-sibilant fricative neighbors, and [p] is very near its coronal counterpart. Still, as with the onset consonants, the model is readily interpretable in terms of feature structure.

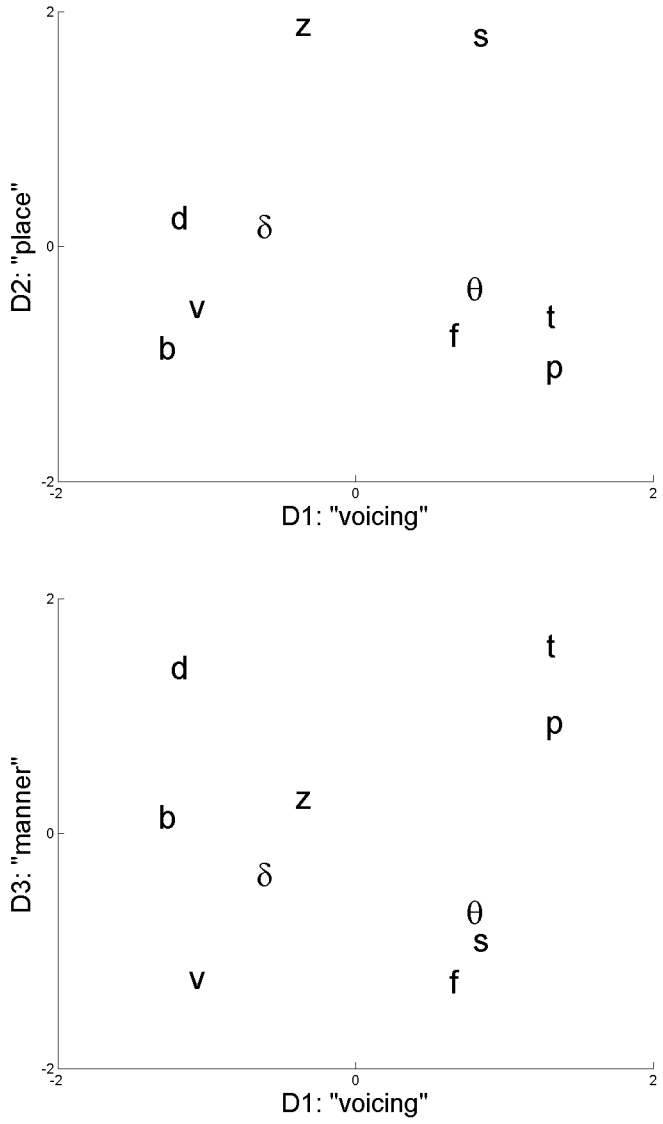

Figure 4. Three-dimensional group space for MDS model of coda consonant confusions from Cutler, et al. (2004).

Figure 5 also indicates a high degree of inter-listener consistency in the weighting of each dimension across the participants. Comparing Figure 3 and 5, the onset weights indicate that each of the three dimensions is roughly equally important (the cluster of individuals sitting in the center of the space), whereas the coda weights indicate that voicing was emphasized (the cluster lying to the right of center on D1) and place deemphasized relative to the other features (the cluster lying below the mid-line in the bottom panel). The participant weights in conjunction with the group space indicate that obstruents contrasting minimally in 'voicing' in coda position are very dissimilar, whereas those contrasting minimally in place tend to be very similar.

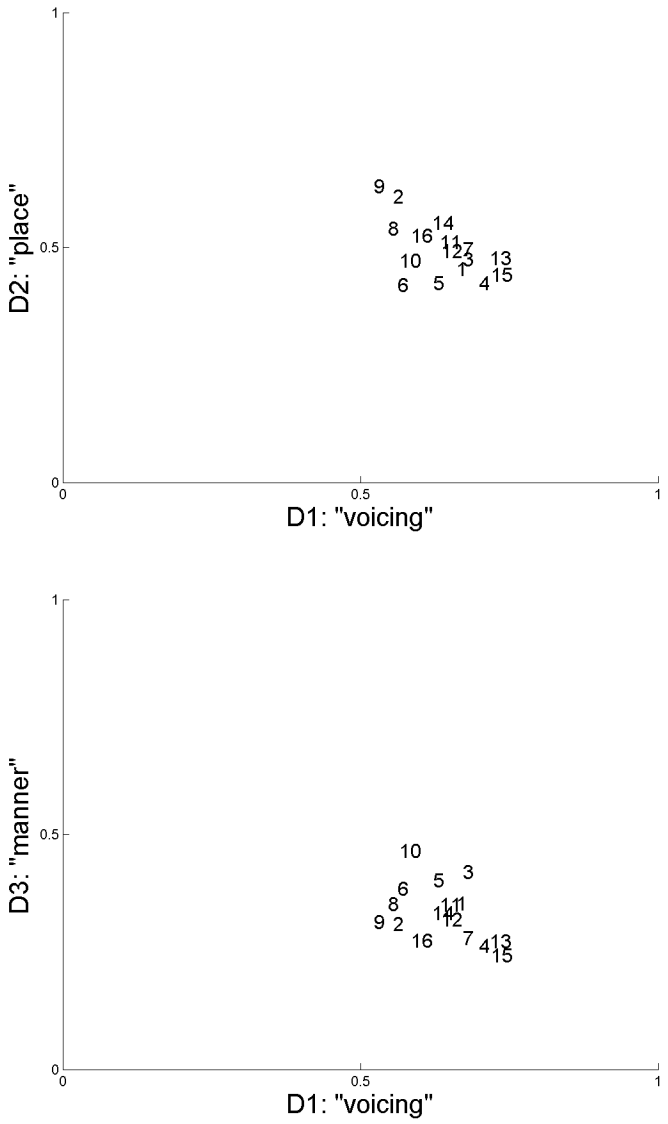

Figure 5. Individual weights for each dimension in an MDS model of coda confusions from Cutler et al. (2004).

What is remarkable about these results is that, for both onset and coda position, the three dimensional MDS solutions correspond as well as they do to the traditional descriptions of the consonant system as residing in three dimensions. We underscore the point that there is no sense in which the MDS algorithm is given a featural classification of the segments; the MDS routine finds a solution based only on the inter-stimulus similarities. This makes the presence of categorical structure along featural lines all the more striking.

4. Discussion and Conclusion

Applying well tested models of perceptual similarity to phoneme identification data provides strong evidence for structure in the consonant system beyond the simple unitary and dichotomous contrast structure encoded in traditional phonological descriptions of the system. Two properties of this description were tested – *Feature Equivalence* (*FE*, the property that all contrasts are equally contrastive) and *Feature Independence* (*FI*, the property that a contrast is equally contrastive regardless of other contrasts inhering in the segments) – and found often not to hold in perceptual phoneme identification. Also, prosodic context modulates failures of both *FE* and *FI*. Importantly, these failures of *FE* and *FI* are exhibited consistently across individual listeners; most, often all, of the native listeners analyzed here exhibited the same patterns of feature interaction. This high degree of consistency across experiment participants indicates that failures of *FE* and *FI* are robust properties of the system, not merely behavioral idiosyncrasies of individuals. Thus, although the basic feature structure is apparent in the MDS representations of the Cutler, et al. (2004) data, there is considerable gradient structure beyond that suggested by a simple contrast space defined by dichotomous, orthogonal dimensions.

Thus, one basic contention of the p-map hypothesis, that there is a systematic component of perceptual confusability that is additional to featural representations, is amply borne out in the current analysis. Other contentions of the p-map hypothesis regarding how such structure relates to phonological patterning go well beyond the scope of the current paper. However, our analyses and the general analytic technique outlined here provide an empirical foundation for investigations of the phonological implications of feature interaction. For example, one claim in Steriade (2001, based on Walden & Montgomery, 1975), namely that continuancy is systematically more perceptually salient than voicing and place (a failure of *FE*) is not experimentally verified here. Rather, features interact with one another such that the salience of a given feature may well vary from one pair of segments to another. If perceptual saliency is to be cited as a cause for phonological effects, such as lexical borrowing patterns, there must be a systematic method of evaluating the likelihood of such perceptual structure. The analyses presented here are one such method.

Some brief comments on the relationship between the SCM structures found here and phonological patterning are in order at this juncture. One may hold the position that phonological structure has no bearing on the similarity relations evident in identification confusions (i.e., that perceptual confusions have nothing to do with phonological structure). This is a problematic position to hold, though, given the close correspondence between our MDS analysis and the well-established traditional laryngeal, place, and manner contrasts. Clearly, the perceptual confusions indicate a gross structure very similar to the expected contrast structure. In addition, the parallel patterns of feature interaction and 'tiers' found in the patterns of failure of *FE* indicate, indirectly, the function of these contrast specifications. It is worth emphasizing that neither the similarity parameters nor the MDS solution based on these parameters is predisposed to find evidence for an assumed structure. Furthermore, as mentioned above, the similarity choice model (SCM) has proven extremely well suited to modeling identification

confusions across a large number of stimulus types. The ability of the SCM to fit the present data should come as no surprise, then, and the evidence of contrast structure within the SCM and MDS analyses here indicate a close relationship between perceptual similarity and the contrast structure in typical phonological representations. Finally, the comparative model fitting analysis shows that, in at least a few cases, restricted versions of the SCM explicitly encoding feature structure fit the data statistically significantly better than did the full SCM.

Another piece of evidence that the SCM structure relates closely to phonological contrast structure is the degree to which the similarities in perception reflect claims in the phonological literature concerning typological commonality ('markedness'). In each of our analyses, the most similar pairs are consistently the most typologically uncommon. For example, in Table 3, every consonant pair lying consistently to right of the inequalities involves contrasts between non-sibilant fricatives and stops or other non-sibilants, which are very uncommon across the world's languages. By contrast, the pairs lying systematically to the left of inequalities include the contrast between sibilants and stops, which exists in virtually all of the world's languages. To the extent that such markedness considerations are integrated into phonological systems, our results suggest that the graded structure of the perceptual space might, in fact, be an underlying source of such effects, though how such effects are to be encoded in phonological systems remains an open question.

Similar comments can be made concerning the differences found here between prevocalic and postvocalic consonants. The phonological literature has noted that post-vocalic assimilation of place features is more common than pre-vocalic, and a comparison of Tables 3 and 5 shows a tendency for contrasts between labials and coronals to 'drift leftward' in the similarity hierarchy for codas. Similarly for *FI*, the phonological literature notes cases in which voicing is maintained in stops but not in fricatives. This is the pattern found in Table 4. Hence, hints of the structure of phonological patterning are found throughout the confusion data analyzed here, as suggested by Steriade's p-map hypothesis.

The exact source and nature of this perceptual component is liable to be extremely complicated. Our current analyses examine one possible source of the confusion structure, that the specifications of the segments are best evaluated in terms of how informative contrasts are, as defined by the relative number of shared and non-shared natural classes between segments. While the NCSM predictions concerning similarity differences are only found in the data a third of the time, there are many aspects of the irregularities in confusions that might fit with the NCSM. More specifically, similarity across place distinctions conforms well to NCSM predictions. Concerning the remainder of mismatches between the NCSM predictions and observed interactions, we might be able to account for half of them as being due to a failure of the binomial tests to reject the null-hypothesis in similarity due to sample noise; it is possible that the predicted NCSM differences have perceptual consequences, but that there is sufficient noise in the data to obscure them. However, the other half of the tests where the NCSM predicts similarity inequalities opposite those derived from the perceptual data are clearly troublesome for any explanation that uses structured specifications to account for the confusion structure.

Of these areas of mismatch, three stand out. First, the NCSM systematically overestimates the similarity of contrasts between sibilants and non-sibilants. While it might be possible to account for this sort of effect by adding additional features for sibilants, this approach runs directly counter to the assumption of the NCSM that what matters is the system of contrast, not the substance that expresses the contrast (e.g., number of features). What this systematic mismatch suggests is the simple brute fact that sibilants are very perceptually different from the other obstruents.

Second, another obvious effect in the data that the NCSM does not predict is the exceptionally low similarity of post-vocalic stops contrasting in voicing. There are two parts to the problem. First, the NCSM, by itself, does not predict differences due to prosodic context, as long as there are no neutralization rules removing contrasts. The second problem appears to be similar to the problem with sibilants. The post-vocalic voicing contrast with stops is exceptionally large. One explanation of this effect is that of redundancy of the voicing specification on the preceding vowel. Silbert & de Jong (2008) noted that, while voicing differences in the preceding vowel duration are very large before stops, they appear to be substantially reduced in fricatives. The presence of this large durational difference for the stops makes them very perceptually salient.

The third systematic effect is the exceptionally high similarity of the voiced and voiceless dental fricatives. While, again, it could be simply that laryngeal contrast is not well encoded in the signal for dental fricatives, it should be noted that this voicing contrast is very small, even in post-vocalic position, where other voicing contrasts are very robust. Also, where the bio-physical constraints on production of voicing contrasts would seem to be very similar for the dental and labial non-sibilants, there is a systematically greater similarity for the dentals. This effect would appear to be due to either an orthographic peculiarity (i.e., the fact that the dentals are spelled with the same characters) or the relatively low (type) frequency of dental fricative minimal pairs. The specific contribution of these two factors – related orthographic-cognitive representations of the segments and contrast frequency effects – deserve more rigorous investigation, but are beyond the scope of the current work.

We hope that the present findings can serve, in part, as a question-generating catalog of perception-specific similarity effects. In any such further research, progress would be well served if data collection were to be guided by specific hypotheses and modeling approaches. Although the data collected by Cutler et al. (2004) served well in these analyses, some of the experimental factors in the original study (e.g., vowel quality) had to be ignored in order to make use of the SCM. Further work employing, and guided by, the SCM and related models should produce data precise and reliable enough to enable analysis of the interval scale (i.e., magnitudes, as opposed to ordinal scale) relationships between inter-segment similarities. Analyses like those employed here can also be extended to study the role of other contrastive features in perception. The largest potential obstacle is difficulty in finding appropriately 'factorial' distributions of features within a class of segments. One straightforward application of these tests would be to identification confusions of voiced stops and nasal segments varying across place of articulation. Using similar techniques, the role of distinctive features in second language speech perception has also been investigated (Silbert & de Jong, 2007a).

There are also a number of other perceptual models that may profitably be applied to the study of speech perception and the role of distinctive features therein. General Recognition Theory (GRT), a multidimensional extension of Signal Detection Theory, provides rigorous mathematical definitions of perceptual independence, perceptual separability, and decisional separability (see, e.g., Ashby & Townsend, 1986). Perceptual separability relates directly to the *FI* construct addressed here. More precisely, testing for perceptual separability in speech perception would provide valuable information regarding the perceptual effect of a given feature at various specifications of another, distinct feature. For example, finding that voicing is not perceptually separable from manner of articulation would indicate that the perceptual effect of voicing changes depending on the manner specification in a given segment. Perceptual independence, or lack thereof, pertains to the stochastic relationship between two (or more) features within a given stimulus (i.e., without regard to variation in the specification of either feature). We are in the process of applying GRT to consonant perception, and preliminary results suggest that perceptual independence holds regularly, whereas perceptual (and decisional) separability does not.

Both GRT and the SCM (as well as other models not considered here) also allow for the direct consideration of decisional mechanisms in perceptual behavior. Although the issue was not addressed in the present undertaking, analyses of, e.g., the relationships between different kinds of frequency (type, token, and contrast) and decision bias represent other promising avenues of research. We may reasonably expect that listeners, when unsure of the identity of a segment, tend to be biased toward more frequent segments. Alternatively, it may be that segment frequency plays very little role in speech perception; word frequency effects may swamp any effect of segment frequency. Analyses of SCM bias parameters indicate that segment frequency correlates with response bias for coda, but not onset, consonant identification (Silbert & de Jong, 2007b). Finally, contrast frequency, or functional load (e.g., the number of minimal pairs contrastive on a given feature) may interact in various ways with similarity and decision bias. These and many thus far unanswered questions are well suited to investigation via mathematical models of perception and decision making. It is clear that such modeling of perceptual behavior provides a set of tools that can, and should, prove invaluable to the development of our understanding of phonological structure, auditory phonetics, and human perception in general.

Acknowledgements

This work is supported by the NSF (grant #BCS-04406540; "Prosody in Cross-language Production and Perception"). We would like to express our appreciation to Anne Cutler, Andrea Weber, Roel Smits, and Nicole Cooper for making their data publicly accessible, without which this analysis could not have been carried out. We would also like to thank Todd Bailey, Stefan Frisch, Beth Casserly, and three anonymous reviewers who provided thorough, thoughtful comments and criticisms of earlier versions of this ms.

References

Ahmed, R., & Agrawal, S. S. (1969). Significant features in the perception of (Hindi) consonants. *Journal of the Acoustical Society of America*, 45 (3), 758-763.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, 5 (3), 144-150.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, 93, 154-179.

Bailey, T., & Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, 52, 339-362.

Benkí, J. R. (2003). Analysis of English nonsense syllable recognition in noise. *Phonetica*, 60, 129-157.

Boersma, P. (1998). *Functional Phonology: Formalizing the Interaction Between Articulatory and Perceptual Drives*. The Hague: Holland Academic Graphics.

Brannen, K. (2002). The role of perception in differential substitution. *Canadian Journal of Linguistics*, 47: 1 – 46.

Clements, N., & Hume, E. V. (1995). The internal organization of speech sounds. In J. A. Goldsmith (Ed.) *The handbook of phonological theory* (pp. 245-306). Cambridge, MA: Blackwell.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America*, 116 (6), 3668-3678.

Flemming, E. (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, 18(1), 7-44.

Frisch, S. (1996) Frequency and similarity in phonology. Northwestern University dissertation. Retrieved February 17, 2007, from http://www.cas.usf.edu/~frisch/Frisch96.pdf

Frisch, S., Pierrehumbert, J., & Broe, M. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, 22, 179-228.

Frisch, S., & Zawaydeh, B.A. (2001). The psychological reality of OCP-place in Arabic. *Language*, 77: 91-106.

Graham, L. W., & House, A. S. (1970). Phonological oppositions in children: A perceptual study. *Journal of the Acoustical Society of America*, 49 (2), 559-566.

Greenberg, J., & Jenkins, J. (1964). Studies in the psychological correlates of sound system of American English. *Word*, 20, 157-177.

Hayes, B. (1999). Phonetically-driven Phonology: the Role of Optimality Theory and Inductive Grounding. In M. Darnell, E. Moravscik, M. Noonan, F. Newmeyer & K. Wheatly (Eds.), *Functionalism and formalism in linguistics, vol. 1: general papers* (pp. 243-285). Amsterdam: John Benjamins.

Klatt, D. H. (1968). Structure of confusions in short-term memory between English consonants. *Journal of the Acoustical Society of America*, 44 (2), 401-407.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29 (1), 1-27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29 (2), 115-129.

Lee, M. D., & Pope, K. J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology*, 47, 32-46.

Lindblom, B., & I. Maddieson (1988). Phonetic universals in consonant systems. In L. Hyman & C. Li (eds.), *Language, Speech, and Mind* (pp. 62 – 80). London: Routledge.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (pp. 103-189). New York: John Wiley & Sons.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusion among some English consonants. *Journal of the Acoustical Society of America*, 27 (2), 338-352.

Peters, R. W. (1963). Dimensions of perception for consonants. *Journal of the Acoustical Society of America*, 35 (12), 1985-1989.

Pierrehumbert, J. (1993). Dissimilarity in the Arabic verbal roots. *Proceedings of North East Linguistic Society*, 23, 367-381.

Pierrehumbert, J. (1994). Knowledge of variation. In K. Beals, J. Denton, R. Knippen, L. Melnar, H. Suzuki, & E. Zeinfeld (Eds.), *Papers from the parasession on variation, 30th meeting of the Chicago Linguistic Society* (pp. 232-256). Chicago: Chicago Linguistic Society.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22 (4), 325-345.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function, II. *Psychometrika*, 27 (3), 219-246.

Shepard, R. N. (1972). Psychological representation of speech sounds. In E. E. David, Jr., & P. B. Denes (Eds.), *Human Communication: A Unified View* (pp. 67-113). New York: McGraw-Hill.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86 (2), 87-123.

Silbert, N., & de Jong, K. (2007a). Laryngeal feature structure in 1st and 2nd language speech perception. In J. Trouvain and W.J. Barry (eds.), *ICPhS XVI* (pp. 1901 - 1905).

Silbert, N. & de Jong, K. (2007b). Response bias, type and token frequency, and prosodic context in segment identification. *Journal of the Acoustical Society of America, 122*: 2019, abstract: 2aSCa15.

Silbert, N.H., and K.J. de Jong (2008). "Focus, Prosodic Context, and Phonological Feature Specification: Patterns of Variation in Fricative Production." *Journal of the Acoustical Society of America,* **123**: 2769 – 2779.

Singh, S., & Black, J. W. (1966). Study of twenty-six intervocalic consonants as spoken and recognized by four language groups. *Journal of the Acoustical Society of America*, 39 (2), 372-386.

Singh, S., Woods, D. R., & Becker, G. M. (1972). Perceptual structure of 22 prevocalic English consonants. *Journal of the Acoustical Society of America*, 52 (6), 1698-1713.

Soli, S. D., Arabie, P., & Carroll, J. D. (1986). Discrete representation of perceptual structure underlying consonant confusions. *Journal of the Acoustical Society of America*, 79 (3), 826-837.

Steriade, D. (2001). The phonology of perceptibility effect: The P-map and its consequences for constraint organization. Ms. University of California, Los Angeles. Retrieved April 14, 2007, from www.linguistics.ucla.edu/people/steriade/papers/P-map_for_phonology.doc

Stevens, K. N. (1998). Acoustic Phonetics. Cambridge, Massachusetts: The MIT Press.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America*, 54 (5), 1248-1266.

Wickelgren, W. W. (1966). Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America*, 39 (2), 388-398.

Wilson, K. V. (1963). Multidimensional analyses of confusions of English consonants. *American Journal of Psychology*, 76 (1), 89-95.