Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Memory and Language 52 (2005) 339–362
Journal of Memory and Language
www.elsevier.com/locate/jml

Phoneme similarity and confusability

Todd M. Bailey*, Ulrike Hahn

School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff CF10 3AT, UK

Received 12 August 2004; revision received 17 December 2004

Abstract

Similarity between component speech sounds influences language processing in numerous ways. Explanation and detailed prediction of linguistic performance consequently requires an understanding of these basic similarities. The research reported in this paper contrasts two broad classes of approach to the issue of phoneme similarity—theoretically driven measures based on phonological features, and empirically derived measures based on confusability. Two experiments are used to test a variety of measures from both classes for their ability to predict judgments of similarity between English consonants. The simplest featural measure is the best on these tests. This recommends it as the basis for further explorations, and improvements to the basic feature model are identified. The paper concludes by examining the extent to which confusability depends on phoneme similarity.

Keywords: Phoneme similarity; Phoneme confusability; Phonological features; Magnitude estimation; Speech errors; Auditory confusion

Phoneme similarity and confusability

The similarity between phonemes seems to influence linguistic behavior in a wide range of contexts. More or less independently of each other, many areas of psycholinguistics have identified effects of similar phonemes. For example, speech errors are influenced by the similarity between phonemes (e.g., Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1990, 1991a), as is confusability in short-term memory (Wicklegren, 1965, 1966), and similarity also influences the strength of phonotactic constraints (Frisch, 1996). Furthermore, speech sounds are the components of words, and the degree of similarity between phonemes influences the overall similarity between whole words (see Bailey & Hahn, 2001; Goldinger, Luce, & Pisoni, 1989; Connine, Blasko, & Titone, 1993; Greenberg & Jenkins, 1964; Hahn & Bailey, in press). Thus, 'tuck' (/tʌk/) is more similar to 'duck' (/dʌk/) than to 'luck' (/lʌk/), because /t/ is more similar to /d/ than to /l/ . Phoneme similarity consequently is potentially important for the many psycholinguistic contexts that have been shown to be sensitive to the similarity between words. For example, in demonstrations of phonological priming, word primes have been found to influence processing rates of similar sounding targets (e.g., Connine et al., 1993). Similarity between words influences short-term memory performance (Baddeley, 1966; Mueller, Seymour, Kieras, & Meyer, 2003 and references therein). Also, the perceived typicality of a sequence is influenced by the number of similar sounding words in the lexicon (Bailey & Hahn, 2001 and references therein; Greenberg & Jenkins, 1964). Latencies in word naming and lexical decision are sensitive to the influence of lexical neighbors (e.g., Luce,

* Corresponding author. Fax: +44 29 2087 4858.
E-mail address: BaileyTM1@cardiff.ac.uk (T.M. Bailey).

0749-596X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2004.12.003

1986). Finally, the chosen inflection for novel sequences is influenced by the inflection patterns of their lexical neighbors (e.g., Albright & Hayes, 2003; Bybee, 1995; Prasada & Pinker, 1993).

Given this range of findings, two questions immediately pose themselves:

1. Do all these domains involve a single, unitary phenomenon 'phoneme similarity?'
2. How is phoneme similarity best measured?

These questions are intimately connected in that finding different 'best measures' across different tasks would provide compelling evidence against a common phonology. Both questions have important theoretical and methodological implications.

The first question has theoretical implications for the overall architecture of the language processing system and the way different psycholinguistic tasks relate. The second question is important because the accuracy of measures of phoneme similarity necessarily influences the accuracy of the predictions and the completeness of explanation provided for those numerous tasks in which it is seemingly involved. For example, Mueller et al. (2003) have argued that past studies of verbal short-term memory have reached the wrong conclusion with regards to its architecture, because similarity was only poorly controlled due to very crude measures of the similarity between words. In a study of wordlikeness, Bailey and Hahn (2001) found that even their best measure of lexical influence left most of the item variance unexplained. Though word similarity poses additional problems with regards to the question of how various matches and mismatches across two words combine into an overall degree of similarity, poor measures of the similarities between component segments themselves may partly explain Bailey and Hahn's finding.

Both questions (1) and (2) are empirical issues that can only be resolved through research in a large number of psycholinguistic domains. In the absence of clear evidence to the contrary, parsimony argues for the default assumption that phoneme similarity is indeed the same across these numerous domains. Indeed, a majority of studies across domains have used the same so-called major class features—voicing, place of articulation, and manner of articulation—to manipulate and measure phoneme similarity and this would seem to reflect a wide-spread assumption of a single underlying notion of phoneme similarity. However, theoretically motivated major class features are not the only approach in the literature. Luce (1986) used the auditory confusability of different phonemes under masking noise as a measure of their similarity for models of lexical influence in word recognition. This allowed him to model interactions between neighborhood density, frequency and the frequency of the probe in word recognition. Auditory confusabilities were also used by Goldinger et al. (1989) in their study of priming.

These two contrasting approaches—features and confusabilities—have never been compared directly. Such a comparison would not only be desirable given the importance of the measurement issue. It also has the potential to illuminate further the nature of confusabilities themselves. This is of interest because confusabilities have figured in attempts to uncover the nature of the mental representations underlying language processing, in particular in the domain of speech production (e.g., Dell & Reich, 1981).

In this paper, we report comparisons between similarity measures based on phonological features and similarity measures based on confusabilities from speech perception, speech production, and short-term memory in their ability to predict choices on a phoneme similarity task. We also examine the ability of featural measures and confusabilities to predict confusabilities themselves; and finally, we examine more directly the relationships among different kinds of confusions and the extent to which they are based on a common underlying notion of phoneme similarity.

Background: Features and confusability

The classic approach to systematically predicting the similarity between phonemes draws on the linguistic analysis of phonemes in terms of sub-phonemic linguistic features. Given a theory of how phonemes decompose into sub-phonemic linguistic features, measures of phoneme similarity based on degree of featural overlap can straightforwardly be derived: the more features two phonemes share the more similar they are. However, linguistics and psycholinguistics have seen a range of competing feature systems without any one having emerged as an outright winner, and some researchers have abandoned attempts to theoretically derive metrics of phoneme similarity in favor of the use of phoneme confusability as an indicator of similarity (e.g., Luce, 1986).

Confusability is widely linked to similarity across the whole of cognition (e.g., Shepard, 1987; Tversky, 1977) by the assumption that two things are more likely to be confused the more similar they are. So, conversely, the degree of confusability between objects should be a means of gauging their similarity. However, a fundamental limitation of confusability data is that they do not constitute an explanation or a *theory* of what makes two things similar, and therefore cannot be used to *derive* or predict similarities. On the other hand, confusability data may be the only practical option if no empirically adequate theory of similarities exists.

Why might one think that theoretically motivated linguistic feature systems do not, or even could not, pro-

vide empirically adequate measures of the similarity between phonemes? One reason might be findings of discrepancies between standard linguistic feature systems and confusability data. The classic study of perceptual confusability is Miller and Nicely's (1955) analysis of auditory confusion of English consonants under noise. Miller and Nicely found differential effects of signal distortion (high pass and low pass filtering) on the perception of major articulatory features such as voicing and place of articulation. Features and confusability were next linked in the studies of Wicklegren (1965, 1966). In contrast to Miller and Nicely, Wickelgren specifically sought to examine the nature of the mental representations underlying phonemes, and he assessed the psychological validity of several linguistic feature systems according to their ability to predict confusability in short-term memory. Wickelgren's first study concerned vowels and compared four systems of distinctive features for English vowels (Wicklegren, 1965). Sub-phonemic features based on either articulatory or acoustic analyses of vowels provided good predictions of confusability between vowels, and Wickelgren interpreted this as evidence for a feature-based representation of phonemes in short-term memory. However, the more abstract featural analysis proposed in Chomsky and Halle's (1968) classic monograph did not correspond well to the observed pattern of vowel confusability. In a further study, Wicklegren (1966) compared Miller and Nicely's feature system, Chomsky and Halle's system, and a system of his own in their ability to predict short-term memory confusions of English consonants. The Wickelgren system was found to be the most accurate, and the Chomsky and Halle system the least accurate. If one assumes that empirical confusabilities are determined primarily by similarity, then Wickelgren's results could easily give rise to pessimism about feature-based measures of phoneme similarity. Chomsky and Halle's feature system, which was based on extensive linguistic analysis and which was to become a standard in linguistics for many years, performed poorly in both of Wickelgren's studies. Even the best predictor of consonant confusions, Wickelgren's own feature system which was introduced with no independent motivation, performed only mid way between chance and perfect performance.

However, further work is necessary to determine the merits of feature-based measures of phoneme similarity relative to measures based on confusability, because a mismatch between the predictions of a featural measure and confusability data does not unequivocally isolate the feature system as the culprit. The fact that a particular feature system is met with only limited success in predicting a particular set of confusability data needs to be interpreted with care for several reasons. First, it might be that it is not the posited linguistic features per se that are to blame, but rather the similarity function by which the various featural matches and mismatches between two phonemes are aggregated into a single judgment. Similarity functions have been the focus of the general cognitive literature on similarity (e.g., Goldstone, 1994; Shepard, 1987; Tversky, 1977) and, there, simple counts of the number of matching or mismatching features have typically been regarded as empirically inadequate ways of computing similarity (see also Frisch, 1996). In general, alternative similarity functions should be explored before a particular set of linguistic features is dismissed.

Second, the confusability data itself might be at fault. If a data set is very noisy, then even the best possible predictor will achieve only comparatively low overall levels of accuracy. Wicklegren (1966), for example, notes that a very large number of trials are necessary before the random error is so low that all individual pairwise confusabilities can be trusted; this means that the accuracy of the best feature system in his study may be as accurate a prediction as the variance in his confusability data permits. Further difficulties arise for confusability under noise (e.g., Miller & Nicely, 1955) as there is no a priori way of knowing which kind and what levels of noise are most appropriate. These considerations suggest that the picture painted by early results might be overly bleak: there might be comparatively simple adjustments to the way featural theories derive similarities, or to what is deemed benchmark confusability data, that would give rise to far greater levels of match between the two.

Finally, and crucially, an apparent mismatch between featural predictions and confusability does not in and of itself indicate that confusability data provide a more accurate way of measuring the relevant similarities. The promise of using confusability data in the absence of an empirically tried and true theory of phoneme similarity lies in the possibility that confusability provides a data oriented alternative—a "bottom up" means of measuring similarity in contrast to the "top down" approach of linguistic description. However, it is an empirical question whether or not confusability is a better predictor of phoneme similarity than phonological features. A mismatch between featural prediction and confusability says only that (trivially) confusability is better than features at predicting confusability. It does not imply that confusability will be a better predictor of the similarity between phonemes *in any other context*, such as the activation of similar sounding words in the lexicon (e.g., Luce, 1986).

In particular, confusion frequencies are likely to be at least partly determined by factors other than similarity, factors that may reflect idiosyncrasies of the task and modality in which confusion data are collected. For example, confusability under noise may be determined primarily by the resilience of the particular cues that enable the detection of given phonemes within the speech

signal. Moreover, one would expect systematic differences between different kinds of confusability such as perceptual confusability, confusion errors in speech production, and confusion in long- or short-term memory. Models of language comprehension and production typically assume a hierarchy of levels of processing. If this is correct, then it seems likely that confusability is determined by contributions from all levels involved. The crucial question is which level dominates for a particular kind of confusion: speech error data, for example, might be influenced to a considerable extent by properties of the articulatory and motor system that are unrelated to phoneme similarities.

The only way to resolve the question of whether or not confusability provides a useful data driven alternative to linguistically motivated derivations of phoneme similarity is through empirical comparisons. We must test whether the range of available confusability measures provide better predictions of phoneme similarity than do available linguistic descriptions. A comprehensive test would have to provide an exhaustive pairing of the range of confusability measures with the range of currently available linguistic theories tested in their ability to predict phoneme similarity in a wide range of contexts such as lexical activation, word and nonword priming, recall and more. Such a test is clearly beyond the scope of a single paper. Given the current lack of data regarding this issue, however, even a comparatively small scale investigation can shed important light on the merits of confusability data as empirical measures of phoneme similarity.

The tests we report in this paper compare linguistic derivations of phoneme similarity and three kinds of confusability data—perceptual confusability under noise, STM errors, and speech production errors—with respect to their ability to accurately predict participants' preferences in an explicit similarity judgment task involving contrasts between English consonants, and also with respect to their ability to predict consonant confusions. As a practical result of this work, a table with the pairwise phoneme similarities of the best metric is provided for future use. We also provide an empirical and theoretical analysis of the general relationship between phoneme similarity and confusability that addresses the extent to which phoneme confusions of different sorts measure a single underlying similarity construct, and how this construct relates to phonological features.

Specifically, the paper proceeds as follows: we first compare the different measures in their ability to explain judgments of phoneme similarity collected in two experiments. We then test the robustness of our results by considering variants of the measures, and also evaluate ways of making them better. The 'best' measure is a simple extension of phonological major class features, incorporating the sonorant–obstruent contrast in addition to voicing, place, and manner of articulation. We test the robustness of our conclusions from similarity judgments (Experiments 1 and 2) by examining the ability of the various similarity measures to predict confusions in speech production, short-term memory, and auditory confusability under noise. These tests confirm the conclusions of Experiments 1 and 2. In the final part of the paper, we examine the possibility of a common similarity-based 'core' across different kinds of confusions.

Similarity measures

The various measures of similarity considered in the rest of the paper are introduced in the next two sections.

Featural similarity

Featural representations of speech sounds were developed originally to explain natural classes, that is, sets of phonemes that behave like members of a single category in phonological patterns (see Kenstowicz, 1994 for an introduction; see Padgett, 2002 for a recent linguistic analysis of features and phonological patterns).

We take as our baseline measure of phoneme similarity the common method of counting shared major class features, including place of articulation, manner of articulation, and voicing. These familiar features provide convenient descriptions of phoneme contrasts, but perhaps more importantly they figure prominently in feature geometry theories of phonological representations (e.g. McCarthy, 1988). These features are used often in all manner of psycholinguistic studies (e.g., Bailey & Plunkett, 2002; Gerken, Murphy, & Aslin, 1995; Jusczyk & Aslin, 1995; Marslen-Wilson, Moss, & van Halen, 1996; Stemberger, 1990; and many other works on speech errors cited in Frisch, 1996). Nevertheless, they are articulatory, not acoustic features, so it is possible that they are more relevant to short-term memory and speech production tasks than to word recognition tasks.

In addition to the three familiar major class features, later in the paper we will also include the sonorant–obstruent contrast in our list of major class features (cf. Chomsky & Halle, 1968; McCarthy, 1988). Sonorant sounds are those articulated with relatively little constriction in the vocal tract so that they allow spontaneous voicing—basically, they can be sung on a held pitch (O'Grady, Dobrovolsky, & Aronoff, 1989). For example, the consonants /m/ and /l/ are sonorant, while /b/ and /z/ are non-sonorant, or obstruent. Sounds that differ in sonority necessarily differ in manner of articulation as well.

A metric of similarity, S_{PMV}, based on place, manner, and voicing, is computed by simply counting the number

of these features shared by two phonemes of interest. This provides only a very coarse measure of similarity, distinguishing among just four levels of similarity (0, 1, 2, or 3 shared features).

We also consider an alternative metric of phoneme similarity proposed by Frisch (1996). In contrast to S_{PMV}, which is based on a comparison of multi-valent features representing two phonemes, Frisch's metric is based on mono-valent, unary features which are either present or absent in the representation of each phoneme (e.g., /d/ includes the feature ORAL, whereas /n/ includes the feature NASAL, among others). Frisch has related the natural class metric to behavioral data including English speech errors, phonotactic constraints in Arabic, and acceptability judgments of Arabic pseudo-words (Frisch, 1996; Frisch & Zawaydeh, 2001; Frisch, Pierrehumbert, & Broe, 2004). Frisch found that the natural class metric was superior to the S_{PMV} metric in predicting rates of phonological speech errors, and also in predicting the strength of phonotactic constraints between adjacent and non-adjacent phonemes. Like S_{PMV}, Frisch's metric is based on articulatory features, so although Frisch obtained good results across a range of tasks, the Frisch measure of similarity could conceivably be more relevant to short-term memory and speech production tasks than to word recognition tasks.

Frisch's metric of similarity, S_{Frisch}, is based on a comparison of the natural classes to which two phonemes belong, where natural classes are sets of phonemes that share a particular combination of features. The Frisch similarity between two phonemes is the proportion of natural classes they share compared to the sum of those that they share and those that they do not. This metric provides much finer degrees of similarity than S_{PMV}.

Phoneme confusabilities

There are a number of published tables of phoneme confusions that might be used as a basis for estimating phoneme similarities given the assumption that relative confusion probabilities reflect the similarities of the items in question. We consider three sets of confusion data from quite different language processing tasks, including short-term memory, speech production, and syllable perception (all pertaining to the English language).

As a measure of phoneme confusability in short-term memory we use Wicklegren's (1966) table of consonant confusions, obtained from a controlled serial recall task involving lists of consonant–vowel (CV) syllables (in which the vowel was always /a/). In recalling a syllable list, participants would sometimes substitute a consonant from one position in the list for another, writing consonant y instead of x at a particular place in the list. We computed a metric of similarity, S_{Wick}, by comparing the number of observed confusions with the number which would be expected to occur by chance, following the procedure of Frisch (1996) and Pierrehumbert (1993). The chance probability of a target phoneme being replaced by a particular intrusion phoneme was estimated as the product of the separate probabilities of the target and confusion phonemes, except that the expected probability of a target phoneme being replaced in error by itself was assumed to be zero. Thus, the probability of phoneme y being substituted for x is estimated as

$$p(x \to y) = p(\text{target} = x) \cdot p(\text{intrusion} = y) \text{ for } x \neq y.$$

This method of estimating chance probabilities compensates for differences in presentation frequency across consonants, and also factors out response biases. The expected number of such confusions in a corpus of N total confusions is given by the equation

$$\text{Expected}(x \to y) = \frac{p(x \to y)}{\sum_{i \neq j} p(x_i \to y_j)} \cdot N.$$

The ratio between observed and expected confusions of one phoneme for another yields a similarity score from 0 (not at all confusable/similar) to positive infinity. A second estimate of the similarity between two phonemes is computed from the probability of substituting x for y rather than vice versa, and the two estimates are averaged to obtain S_{Wick}.

As a measure of phoneme confusability in speech production we use speech error data from the MIT corpus of single consonant intrusion errors in spontaneous speech (Shattuck-Hufnagel & Klatt, 1979). This corpus includes substitution errors (*mell* for *well*) and also exchange errors (*pade mossible* for *made possible*). We compute a metric of similarity, S_{MIT}, by comparing the number of observed errors with the number expected by chance, as described above.

To measure perceptual phoneme confusability we use Luce's (1986) tables of identification errors for CV or VC syllables under signal-to-noise ratios (SNR) of +15 dB, +5 dB, and −5 dB. There are thus six sets of Luce consonant confusabilities (from separate tables for onset and coda consonants, each at three levels of noise). We compute metrics of similarity, S_{L+15}, S_{L+5}, and S_{L-5} by comparing the number of observed and expected errors, as described above. Each of these metrics can be computed either for onset consonants or coda consonants, and it will be clear from the context below whether we are referring to onset or coda similarity.

Experiment 1

Experiment 1 compares the similarity metrics introduced above in their ability to explain judgments of rel-

ative similarity reported in Hahn and Bailey (in press, Experiment 1A). Participants performed an auditory two-alternative forced-choice task, similar in relevant respects to tasks widely used in the general literature on similarity (e.g., Tversky, 1977). In our task, a target word was paired with two choice words (e.g., /pʌsp/, /bʌsp/ and /pʌsp/, /gʌsp/). Participants indicated which pairing—target-and-choice-word-A (TA), or target-and-choice-word-B (TB)—sounded more similar. One choice word differed from the target by a single phonological feature (place, manner, or voicing), either in the initial consonant or the final consonant, and the other choice word differed by two phonological features in the same syllable position. Although participants were comparing the relative similarity of whole words, these words differ by a single phoneme only. We assume that any systematic response preferences are due to phoneme similarities.

The main question of interest is how well the various similarity metrics explain participants' judgments of relative similarity in onsets and codas. Though the choice items differed from each other by only one phonological feature, there is considerable variability across items in this data set. This variability between items suggests that there are some aspects of similarity that the major class features are not discriminating. The question is whether this variation will be explained any better by the more fine-grained featural metric or by the confusability-based measures.

Method

We present here a relatively brief description of the experimental method. A full description is given in Hahn and Bailey (in press).

Participants

Twenty eight phonetically naïve psychology undergraduates at Cardiff University took part in the study for course credit.

Stimuli

Stimuli were single-syllable nonwords. There were 20 target-choice triads of nonwords which differed in the initial consonant (the *onset* stimuli), and 20 triads which differed in the final consonant (the *coda* stimuli). Each target ended with the same consonant with which it began (e.g., /pʌsp/). A choice candidate, A, was derived from each target by changing a single major class feature (e.g., /bʌsp/). A second choice candidate, B, was derived from the first by changing an additional second feature of the same phoneme (e.g., gʌsp). The stimuli are listed in Hahn and Bailey (in press). The syllables were spoken by a male speaker of British English, for digital recordings (16 bit, 20 kHz) in a professional recording studio.

Procedure

Participants were tested one at a time in a sound-insulated room, listening to the stimuli on headphones. On each trial, participants first heard a target word followed by one of the two corresponding choice candidates. Participants then heard the target again followed by the other choice candidate, and judged whether the first or second candidate was more similar to the target syllable. The order of presentation of the choice candidates was counterbalanced across participants. In addition to the 180 test trials (of which 40 are relevant for present purposes), two foil trials were included to identify inattentive participants. The structure of the foil trials was identical to the test trials, except that on foil trials one of the choice items was identical to the target, providing an obvious correct answer. Also, there were four rest trials which instructed participants to take a short break. The presentation order of trials was randomized for each participant.

Results

Analyses are based on data from 20 attentive participants. Responses for each target-candidate triad are given in Hahn and Bailey (in press, Appendix A). Among the 400 onset comparisons (20 items $\times$ 20 participants), 62% of responses favored the TA pairs, which differed by a single feature, over the TB pairs, which differed by two features. Among the coda comparisons, the TA preference was 70%. As reported in the original study, these results indicate that, on average, participants were sensitive to the number of featural differences between target and candidate syllables.

In this study, the TA pairs always differed by a single feature (and TB pairs differed by two features), but they included a wide variety of phonemes contrasting many different features. Across items, the fraction of participants selecting the TA pair rather than the TB pair varied from 25 to 90% in onset comparisons, and from 50 to 90% in coda comparisons (see Hahn & Bailey, in press, Appendix A). In the original study, we did not assess whether these item differences simply reflect random noise, or whether some TA–TB contrasts really differ in their relative similarity. This question arises in the present context because of the possibility that alternative similarity metrics might explain item differences that are not explained by major class features. The items variable was a significant factor in analyses of variance (ANOVA), $F(19, 361) = 2.8$ and 1.7 for onset and coda comparisons, respectively, $MSE\text{s} = 0.21$ and 0.20, $p < .001$ and $p = .030$. This indicates that TA–TB comparisons did not all involve the same degree of difference, and suggests that there is systematic variation in similarity above and beyond the single-feature difference along which the items were chosen. Evidently major class

features are not sufficient to fully account for these similarity judgments.

Each of our metrics of similarity gives two similarity values for each triplet, S(TA) and S(TB), comparing the phoneme in the target word (T) to the corresponding phonemes in the two choice words (A and B). At the coarsest level, each metric predicts that the fraction of TA responses in the judgment task should be greater than, equal to, or less than 50%, depending on whether S(TA) is greater than, equal to, or less than S(TB). Our analysis assessed the extent to which each model simultaneously predicted this three-way distinction and also predicted the rank order deviations from 50% responding. We analyzed ranks rather than raw scores because analysis of raw scores would require strong assumptions about the form of the function relating a particular measure of similarity to the behavioral data. Moreover, it will be the relative ranks that are of interest in most practical contexts.

We computed the experimental TA response fraction for each triplet, and assigned ranks 1 to n to triplets in ascending order of response fractions. We then adjusted the ranks by subtracting the rank corresponding to chance performance (i.e., 50%). This resulted in negative ranks for triplets with less than 50% TA responses, a rank of 0 for triplets with exactly 50% TA responses, and positive ranks for triplets with greater than 50% TA responses. To compute predicted ranks for each similarity metric, we ranked the $S(TA) - S(TB)$ differences for each triplet, and then subtracted the rank corresponding to $S(TA) = S(TB)$. Regression through the origin assessed the extent to which the predicted ranks matched the experimental ranks.[1]

Results for the various similarity metrics are given in Table 1, which shows the fraction of variance explained by each metric, for both onset and coda triplets, as well as the combination of the two (n.b. data for the Luce metrics are based on 19 rather than 20 onset triplets because the Luce confusion matrices do not include /ʒ/). For onset positions, S_{MIT} explained the most variance (53%), followed by the major class feature measure, S_{PMV} (45%). Most metrics did better in coda position than in onsets. Here, the best metric was S_{PMV} (67%) followed by S_{Wick} and S_{Frisch} (63 and 59%, respectively). The best metric overall was S_{PMV} (57%).

As described above, in computing the S_{Wick}, S_{MIT}, and S_{Luce} metrics we combined data from $x \to y$ and $y \to x$ confusions, and the procedure normalized confusion probabilities relative to chance by dividing observed by expected probabilities. To guard against the possibility that some aspect of this computational procedure biased the results, we also evaluated $x \to y$ and $y \to x$ confusions separately, with and without the normalization procedure. Two metrics, S_{L+5} and S_{MIT}, performed marginally better when only $x \to y$ confusions were considered, explaining 36 and 44% of the overall variance, respectively (with normalization for chance). Otherwise, alternative computational assumptions resulted in worse fits to the data. Because the difference was very small in the case of S_{MIT}, and the differences were not consistent across the three S_{Luce} metrics, these variants are not used in further analyses below.

Table 1
Performance of similarity metrics (adjusted R^2%) in predicting responses for Experiment 1

Metric	Onset	Coda	Total
Confusabilities			
Wick	26	63	47
MIT	53	32	42
L + 15	17	36	30
L + 5	22	32	29
L − 5	33	52	42
Feature metrics			
PMV	45	67	57
Frisch	26	59	45

These statistics estimate the amount of variance in ranks relative to chance performance explained by each metric.

Discussion

The simplest measure of similarity, based on major class features, was the best predictor for codas and the best predictor overall. This result is surprising given the simplicity of the S_{PMV} measure, and given the variability between items that S_{PMV} is unable to account for. In the present sample (originally chosen to address a different question) choice word A always shared two features with the target, and choice word B shared one feature. S_{PMV} predicts that participants should always prefer choice A, and that the size of the preference for A over B should be the same across all stimuli. However, neither of these predictions is fully supported by the data. Participants sometimes preferred choice B over choice A, and the fraction of participants selecting choice word A varied from 25 to 90%. The significant effect of the items variable suggests that at least some of this variation is due to genuine differences between item similarities and not just experimental noise. This indicates the presence of fine-grained differences in phoneme similarity that the simple S_{PMV} metric does not explain. Nevertheless, the other, more detailed similarity metrics were unable even to match the predictive accuracy of S_{PMV}, let alone to improve on it.

[1] Our analysis of ranks is similar to Spearman's correlation for ranked data except that we are interested in deviation from a reference value specified a priori (50 from average ranks computed from the data).

With regards to Frisch's natural class metric this result is disappointing. This metric is theoretically well-motivated and was more successful than S_{PMV} in predicting speech error phoneme confusions in Frisch (1996). The contrast with the present result is a reminder that any measure of similarity must ultimately be tested across a range of data sets and tasks. At present, it is unclear what makes Frisch's metric so well suited to speech errors. One possibility, suggested to us by a reviewer, is that the Frisch metric includes additional articulatory-based features, which should be particularly relevant to speech production errors.

Whatever the merit of the two featural measures, the results suggest that empirically derived measures of similarity based on confusability do not provide a universal data driven shortcut to phoneme similarity. Only the MIT corpus of speech errors gave better predictions of participants' responses, and this only for onset comparisons. All of the confusability measures that we tested potentially measure subtle shades of similarity between phonemes. Consequently, it is remarkable that none of these measures performed better overall than S_{PMV}.

Our judgment task has a format that seeks to determine which of two choice words is more similar to the target word. This is a so-called directional comparison, 'how similar is *a* to *b*,' as distinguished from symmetric comparisons such as 'how similar are *a* and *b*?' The general literature on similarity has found small, but systematic asymmetries for directional comparisons depending on which way they are made, that is whether participants are asked to compare *a* to *b* or *b* to *a* (e.g., Tversky, 1977). While the featural measures are inherently symmetric, the confusability based measures could capture such asymmetries if they existed for phoneme similarity as well. We examined whether the potential asymmetry of the empirical confusability metrics conferred any advantage in explaining participants' judgments in our study. There was no consistent benefit of asymmetric confusability measures over their symmetrified counterparts based on averaged values. Indeed, the asymmetric versions were almost always worse. This indicates that any potential gain in accuracy from using asymmetric matrices is outweighed by the reduction in noise afforded by averaging.[2] This result fits with the findings of Benki (2003) who failed to discern any consistent explanation for the asymmetries in his consonant confusion data (see also Stemberger, 1991b on speech errors). Although one cannot rule out the presence of systematic asymmetries in phoneme similarity altogether, our results suggest that they are at best comparatively small, a conclusion in line with findings on similarity in other domains (e.g., Ashby, Maddox, & Lee, 1994; see also Nosofsky, 1991 for discussion as to how asymmetries might arise). As a consequence, it seems likely that for most practical purposes symmetrical measures will suffice.

Finally, measures of empirical confusability potentially allow positional information specific to either onsets or codas to enter the metric. In contrast, S_{PMV} (and Frisch's metric) makes no distinction between phonemes according to their position within the syllable. With respect to the relative performance of our measures on onsets and codas, the picture is not entirely clear. Whatever their relative levels of performance against each other, most metrics were better at predicting participants' coda responses than onsets. This was true regardless of whether they distinguished onsets and codas (Luce's matrices), or whether they applied a single measure to both onsets and codas (S_{PMV}, Frisch, and Wickelgren's confusions which are on onset position only). One possibility is that there is more noise in judgments of onsets than in judgments of codas, due to greater perceptual salience for codas than onsets in similarity comparisons. Support for this conclusion comes from the experimental investigations of word similarity reported in Hahn and Bailey (in press). That study found greater sensitivity to phoneme differences in coda contexts than in onsets across a series of five experiments. Nevertheless, in the present study there was one exception to the general pattern of better coda performance in that S_{MIT} performed better on onset response rates than on coda response rates. The MIT corpus combines phoneme intrusion errors without regard to their syllable positions. However, such errors involve onset consonants more frequently than consonants in other syllable positions (Shattuck-Hufnagel & Klatt, 1979). Therefore, the confusion counts on which S_{MIT} is based primarily reflect onset similarities. The poorer performance of S_{MIT} on coda similarities might suggest that phoneme similarities vary depending on the syllable positions of the phonemes. This could be examined by collating speech error data that separated coda and onset confusions (assuming the corpus of speech errors was large enough to yield reliable counts even for the less frequent coda intrusions). If position-specific confusabilities improved the prediction of coda similarities, this could be interpreted as evidence for systematic differences in phoneme similarity between onsets and codas (see Goldstein & Fowler, 2003). Further interest in such a test stems from the fact that S_{MIT} performs better than S_{PMV} in explaining degrees of similarity in onsets. It is the fact that it accounts for only half as much variance as does S_{PMV} on the corresponding coda comparisons which makes it only the third best performer overall. Consequently, it is possible that position specific speech errors might prove a better predictor of phoneme similarity

[2] Averaging necessarily collapses two sets of observations, confusions of *x* with *y* and confusions of *y* with *x*, into a single cell. The underlying sample size for this cell is therefore larger than those of the directional observations.

than our current featural metrics that do not take syllable position into account.

We return to the question of positional specificity in the third part of this paper. For our next experiment, however, we focus on comparisons in coda position to take advantage of the greater sensitivity participants seem to have for comparisons in this position. This next experiment evaluates the featural and confusability-based metrics on a broader range of phoneme comparisons.

Experiment 2

The range of phoneme comparisons in Experiment 1 was restricted to comparisons between items differing in one and two place, manner or voicing features from the target. The results obtained for our metrics in Experiment 1 consequently do not necessarily extend to performance across the whole range of possible phoneme comparisons as might be encountered, for instance, when trying to capture the influence of a word's similar sounding neighbors within models of lexical activation. The way to estimate performance across all possible comparisons is through evaluation on a randomly chosen subset of sufficient size. Such a sample of phoneme comparisons will include contrasts which are likely to involve greater dissimilarities than those sampled in Experiment 1, such as contrasts involving simultaneous changes to place, manner, and voicing; at the same time it will also contain phoneme comparisons of potentially greater similarity than those sampled in Experiment 1 as it will include contrasts such as *dusp* and *gusp* which both differ from the target *busp* by a single major class feature (place of articulation). The PMV feature metric predicts that these two words will be exactly equal in their similarity to *busp*; they need not, however, be of equal confusability and hence similarity for the empirically derived metrics.

Experiment 2 retains the task used in Experiment 1, with participants choosing which of two choice words is most similar to a given target word. However, Experiment 2 focuses exclusively on comparisons involving final consonants.

Method

Participants

Thirty-four psychology undergraduates at Cardiff University took part in the study for course credit. All were phonetically naïve.

Stimuli

Sixteen potential final consonants were included in this study, including /pkbgʃčfvθðjʒmnŋl/. This excluded consonants that might be interpreted as past tense or plural inflections (/tdsz/), and consonants that do not occur word-finally in British English (/r/ and /h/). From the 1680 possible combinations of a target and two different choice phonemes, we selected a random sample of 180 target-choice triplets. We then identified word triplets that contrasted the 180 target-choice triplets in word-final position. Words were chosen from a set of 11,000 pronounceable nonwords identified in preparation for Experiment 1, as described in Hahn and Bailey (in press). Each word began with one or more consonants and ended with a single consonant. The items in each triplet differed only in the final consonant (e.g., /stɪg/, /stɪp/, and /stɪb/). The stimuli are listed in Appendix A. We recorded (16 bit digital samples at 20 kHz) the stimulus words in the same recording studio, pronounced by the same speaker who recorded the words for Experiment 1. Waveforms were scaled to the same peak amplitude.[3]

Procedure

The procedure was the same as in Experiment 1, except that there were 10 foil trials to identify inattentive participants. There were thus 194 trials in all, including four rest trials. On the 180 test trials, the order of presentation of the two choice words for each triplet was counterbalanced across subjects.

Results

Inspection of responses to the 10 foil trials identified four participants with two or more errors on these trials. Data from these apparently inattentive participants were dropped from further analyses. Responses were averaged across participants to determine the fraction of participants choosing the TA word pair for each triplet. For summary purposes, we identified the 'winning' response (TA or TB) for each comparison, that is, the response given by the majority of participants. In the six cases in which equal numbers of participants chose both alternatives, the TA choice was arbitrarily designated the winner. Winning response rates varied from 50 to 100%, with a median value of 70%, and interquartile range of 20%. Rates for winning items are listed in Appendix A.

The agreement between participants increased with the difference in featural differences between TA and TB. Thus, on trials where the difference in featural differences was 0, 1, or 2 features, the fraction of participants choosing the winning pair was 65, 71, and 88%, respec-

[3] Experience with the RMS amplitude normalization used in Experiment 1 suggested that it was at best marginally better than peak amplitude normalization at equalizing perceived loudness levels across stimuli, so the simpler procedure was used in Experiment 2.

tively. The level of agreement observed here for the 1–2 feature contrast are in good agreement with the results of Experiment 1 (70%). Agreement levels were lower for trials involving two pairs of words differing by the same number of features. These trials involved fine degrees of difference in similarities, and judgments of relative similarity between these pairs were evidently fairly difficult.

As in Experiment 1, we computed experimental ranks from the TA response fractions, and subtracted the rank corresponding to chance performance. Again, the predicted ranks for each metric are based on the similarity differences TA–TB. Regression through the origin assessed the extent to which the predicted ranks matched the experimental ranks. Results for the various metrics are given in Table 2, which shows the fraction of variance explained by each metric. The Wickelgren confusion data are based on 147 rather than 180 stimuli because 33 of our stimulus triplets involved the phoneme /ŋ/, which was not included in Wickelgren's data.

The two featural metrics, PMV and Frisch, performed about equally well, predicting 55% or more of the variance. This is substantially more than any of the metrics based on confusability data, which predicted at best (L + 5) 38% of the variance.

Discussion

In contrast to Experiment 1, the current study estimates the performance of the various metrics across the whole range of possible phoneme comparisons by testing them on a representative sample of all possible contrasts. This provides a more stringent test of the metrics' performance and the fairest kind of comparison between measures. On this test, the main conclusion of Experiment 1 is confirmed. The theoretical, featural metrics of similarity, S_{PMV} and S_{Frisch}, are substantially better than the empirical metrics at predicting judgments of similarity—here, specifically similarity between coda consonants.

Table 2
Performance of various metrics (adjusted R^2%) in predicting responses for Experiment 2

Metric	TA–TB
Confusabilities	
Wick	28
MIT	32
L + 15	28
L + 5	38
L − 5	35
Feature metrics	
PMV	56
Frisch	55

These statistics estimate the amount of variance in ranks relative to chance performance explained by each metric.

Again, the simplest featural measure performs very well. The best empirical measure in this experiment was the Luce matrix of auditory confusability at a signal to noise ratio of +5 dB. However, S_{L+5} falls far short of the simple featural measure, accounting for almost 20 percentage points less variance. The evidence from both Experiments 1 and 2, then, suggests that data-driven measures based on confusability do not offer any universal shortcut for measuring phoneme similarities. At least on this explicit judgment task, featural measures yield better predictions of similarity.

Improving the measures

In this section we consider and test variants of the featural and confusability-based measures. Our strategy is to consider an alternative response rule, and also to consider measures of dissimilarity rather than similarity. These assessments allow us to test the generality of the conclusion that the featural measures are better predictors of similarity than the confusabilities. At the same time, qualitative characteristics of phoneme similarity are examined, and finally, we identify measures that are even better predictors of similarity judgments than the variants considered above.

In seeking a function relating psychological similarities to behavioral data, one must bear in mind that perceived similarities operate in the context of a particular task—in our case forced choice decision, in other cases, a recognition task for example, or a classification task. Conceptually, perceived psychological similarities can be distinguished from the decision rule under which participants determine their actual responses on a particular task. A range of possible decision rules exists. In an old-new recognition task, participants might, for example, simply adopt a threshold and always produce an "old" response when that threshold is exceeded because the stimulus is sufficiently similar to previously seen items. Alternatively, their response rule might be based on probability matching, whereby the probability with which they produce a particular response is matched to the assumed probability that this response is indeed correct.

Our analyses above implicitly assume probability matching in that the fraction of TA responses in the aggregate data is assumed to be indicative of the similarities perceived at the individual level. This implies a response rule whereby the probability of a TA response directly relates to the degree by which the chosen comparison item exceeds the other in similarity to the target; that is, responses are assumed to follow the magnitude of the difference Similarity(TA) − Similarity(TB). However, instead of taking the difference between scale values as a measure of response probability, a plausible alternative is to take a ratio of scale values, as in

R. D. Luce's (1959) choice rule. Luce's choice rule is related to normative theories of rational choice, and has been used widely in similarity-based modeling, including exemplar-based categorization (e.g., Nosofsky, 1986), P. A. Luce's neighborhood activation model of word recognition (Luce, 1986; Luce, Pisoni, & Goldinger, 1990), and inflectional morphology (Hahn & Nakisa, 2000). In its general form, Luce's choice rule links the probabilities of participants' responses in a forced choice task to an underlying 'response strength' scale. Given a choice among response alternatives, the probability of a particular response, x, is the ratio of the response strength for x divided by the sum of the response strengths for all alternatives. Applied to the present study, it is natural to interpret 'response strength' in terms of the phonological similarity between two words. Thus, the probability of a TA response is the ratio

$$\frac{S(TA)}{S(TA) + S(TB)},$$

for some measure of psychological similarity S applied to TA and TB. The response probabilities predicted by similarity ratios (Luce's choice rule) are not, in general, the same as those predicted by similarity differences (probability matching). Therefore, it is an empirical question which response rule provides better predictions of similarity judgments.

Our discussion thus far has focused on similarity rather than the dissimilarity between two phonemes. Similarity and dissimilarity (or distance) are often assumed to be equivalent ways of expressing the same concept, since one can readily be turned into the other. For example, instead of counting shared features, as in S_{PMV}, we could count featural differences to obtain a measure of dissimilarity between two phonemes, D_{PMV}. Despite the close conceptual link between similarity and dissimilarity scales, when combined with a particular decision rule to predict behavioral responses, similarities and dissimilarities generally produce different patterns of predicted responses. To assess which combination best characterizes performance in the present study, we factorially applied Luce's choice rule and the probability matching rule to both similarity and dissimilarity scales.

Results

From the seven similarity metrics we derived dissimilarity metrics and combined them with choice rules as follows. For featural metrics, we defined a dissimilarity metric based on major class features, $D_{PMV}(XY) = 3 - S_{PMV}(XY)$. Also, we followed Frisch (1996) in deriving a dissimilarity metric by reversing the S_{Frisch} scale, so that $D_{Frisch}(XY) = 1 - S_{Frisch}(XY)$. When entered into Luce's choice rule, D_{PMV} and D_{Frisch} produce different patterns of predicted responses compared to the corresponding similarity scales, so dissimilarity ratios

$$\frac{D_{PMV}(TB)}{D_{PMV}(TA) + D_{PMV}(TB)} \text{ and } \frac{D_{Frisch}(TB)}{D_{Frisch}(TA) + D_{Frisch}(TB)}$$

must be considered separately from the corresponding similarity ratios. In contrast, when entered into the probability matching rule the resulting dissimilarity differences $D_{PMV}(TB) - D_{PMV}(TA)$ and $D_{Frisch}(TB) - D_{Frisch}(TA)$ are functionally equivalent to their similarity counterparts, and need not be considered separately. Thus, for each of the feature metrics there are three distinct combinations to consider: similarity differences, similarity ratios, and dissimilarity ratios.

For confusabilities, dissimilarity metrics were derived from the various similarity metrics by taking the multiplicative inverse, so that $D_{Wick}(XY) = 1/S_{Wick}(XY)$, and so on.[4] When entered into the probability matching rule these dissimilarity scales produce different patterns of predicted responses compared to the corresponding similarity scales, so for these metrics dissimilarity differences must be considered separately from the corresponding similarity differences. In contrast, when entered into Luce's choice rule, the dissimilarity scales derived from confusabilities are functionally equivalent to their similarity counterparts. Thus, for each of the confusability metrics there are three distinct combinations to consider: similarity differences, similarity ratios, and dissimilarity differences.

We tested these alternatives on the data from Experiment 2 above. For each distinct combination of metric and choice rule, regression through the origin assessed the extent to which the predicted ranks matched the experimental ranks, relative to chance. Results for the various metrics are given in Table 3, which shows the fraction of variance explained by similarity differences and similarity ratios for each metric, along with results for either dissimilarity differences or dissimilarity ratios, whichever is distinct from the corresponding similarity scores. Table 3 also shows results for a new metric, PMVS, which will be discussed below. Again, results for the Wickelgren confusion data are based on 147 rather than 180 stimuli because 33 of our stimulus triplets involved the phoneme /ŋ/, which was not included in Wickelgren's data.

For PMV and Frisch, the best predictions came from dissimilarity ratios, though similarity differences were nearly as good. Similarity ratios performed substantially worse. Among the empirical, confusability-based

[4] A complication arises in taking the inverse when the similarity metric has a value of zero. This happens whenever none of the corresponding phoneme confusions were observed in a particular corpus. In each of these cases, we estimated a dissimilarity value by taking a number slightly larger than the maximum value observed in the other cases. The estimated values were 5, 18, 10, 14, and 20, for D_{Wick}, D_{MIT}, D_{L+15}, D_{L+5}, and D_{L-5}, respectively.

Table 3
Performance of various metrics (adjusted R^2%) in predicting responses for Experiment 2

Metric	Similarity metrics		Dissimilarity metrics	
	S(TA) − S(TB)	$\frac{S(TA)}{S(TA)+S(TB)}$	D(TB) − D(TA)	$\frac{D(TB)}{D(TA)+D(TB)}$
Confusabilities				
Wick	**28**	23	18	
MIT	**32**	25	29	
L + 15	28	24	**29**	
L + 5	**38**	31	32	
L − 5	**35**	29	28	
Feature metrics				
PMV	56	48		**58**
Frisch	55	49		**56**
PMVS	62	49		**67**

These statistics estimate the amount of variance in ranks relative to chance performance explained by each metric. Results are given for similarity differences and ratios. Results are also given for dissimilarity scores where these are distinct from similarity scores. The best result for each metric is shown in bold.

metrics, similarity differences generally provided the best predictions, except that one of the Luce data sets (L + 15) achieved slightly better results for dissimilarity differences. The best empirical metric, again, came from one of the other Luce data sets (L + 5). Similarity differences from this data set explained 38% of the variance, substantially less than the 58% explained by ratios of PMV dissimilarities.

Discussion

The first conclusion from these tests is that they support our earlier results. In this sense, the relative performances of the theoretical and empirical measures are robust to changes in the particular functions relating them to the data, and whether they are used in the form of similarity or dissimilarity measures.

The second conclusion from the present tests is that phoneme similarity seems to be based on psychological estimates of difference as opposed to estimates of commonality. Ratios of PMV dissimilarities gave better predictions of similarity judgments than either ratios or differences of similarities. These ratios can be interpreted as a case of Luce's choice rule (R. D. Luce, 1959), applied to a response scale measured in featural differences:

$$p(\text{Response A}) = \frac{\text{Featural Diffs(TB)}}{\text{Featural Diffs(TA)} + \text{Featural Diffs(TB)}}.$$

One implication of Luce's choice rule is that the task should be easier when the choice words are similar to their target (few featural differences), and harder when the choice words are less similar to their target (more featural differences). Thus, if choices A and B differ from target T by one and two features, respectively, the predicted fraction of choice A responses is $2/3 = .67$, which is much better than chance. In contrast, if choices C and D differ from target U by three and four features, respectively, the predicted fraction of choice C responses is $4/7 = .57$, which is better than chance, but by a much smaller margin. This prediction is borne out in the data from Experiment 2, as shown in Fig. 1. The figure shows the response probability for the most common (winning) response for each target-choice triplet as a function of the total distance to the target. Circles connected by the solid line show results for triplets in which both choice words differed from the target by the same number of features. Crosses connected by the dotted line show results for triplets in which one choice differed from the target by one more feature than the other choice did. For completeness, the figure also shows the results for triplets in which one choice differed from the target by two more features than the other choice did. The downward slope of the two lines indicates that as the total dissimilarity between choice and target words increased, participants found it harder to determine which choice was more similar to the target. This pattern of responses is predicted by the ratio of PMV dissimilarity scores.

In contrast, the ratio of PMV *similarity* scores predicts exactly the opposite pattern of results.[5] It predicts that response probabilities should approach chance as the total similarity between choice words and targets increases. This would produce an upward slope for the lines in Fig. 1. This prediction is not consistent with our data, and as a consequence the ratio of PMV similarity scores does not fit the data nearly as well as the ratio of PMV dissimilarity scores (see Table 3). Turning

[5] If the decision rule is based on shared features, the numerator must refer to TA rather than TB, so that the decision rule is

$$p(\text{Response A}) = \frac{\text{Shared Feats(TA)}}{\text{Shared Feats(TA)} + \text{Shared Feats(TB)}}.$$

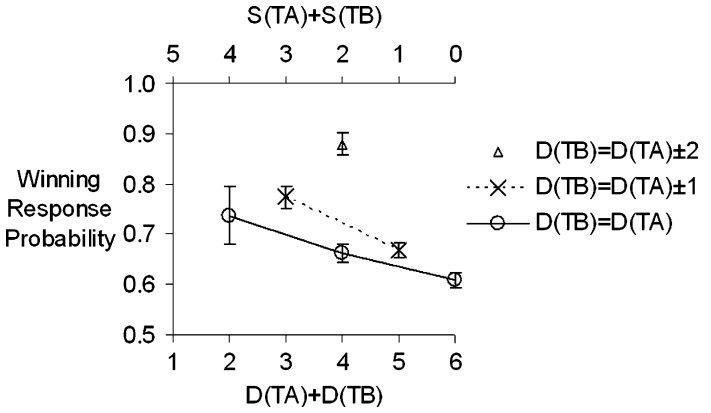

Fig. 1. Results of Experiment 2, showing the probability of the winning response for each stimulus triplet, plotted as a function (lower horizontal axis) of increasing total featural differences between a target and its two choice words, or (upper horizontal axis) decreasing total features shared by a target and its two choice words. D(TA) and D(TB) represent PMV dissimilarity between winning and losing choices, resp. S(TA) and S(TB) represent corresponding PMV similarity. Error bars show ± 1*SE*.

attention to the arithmetic difference between PMV similarity scores (or, equivalently, dissimilarity scores), S(TA) − S(TB) predicts that response probabilities should depend only on the difference between the numbers of features the two choice words share with the target. Responses should not depend on the total numbers of features shared or not shared. Thus, the difference of PMV similarity scores predicts no slope for the lines in Fig. 1. Again, this prediction is not consistent with the data.

The finding that sensitivity to featural differences declines as the number of differences increases is consistent with the findings of Hahn and Bailey (in press). In that study, participants were more certain in their similarity judgments when choice words differed from their target by a single phoneme substitution (e.g., skub–skup, skub–skuck) than when choice words differed from their target by a phoneme insertion as well as a substitution (skub–skusp, skub–skusk). Indeed, it is well-documented across many psychophysical domains that perceptual sensitivities go down as the magnitudes under comparison increase (Weber, 1835/1978). Thus, a difference of a few grams is readily detectable when comparing two sheets of paper but not when comparing two heavy suitcases. Luce's choice rule explains this result (R. D. Luce, 1959), on the assumption that the mental comparisons are based on representations of weight, not lightness. The same explanation appears relevant to comparisons between words, where a given difference in similarity to the target is hard to perceive when the choice words differ substantially from the target, but easier to perceive when the choice words are minimally different from the target. This finding can be related to Weber's magnitude effects, but only under the assumption that word comparisons involve a psychological scale of dissimilarity (e.g., featural differences) rather than similarity.

If word comparisons involve psychological dissimilarities rather than similarities, we might wonder why the various sets of confusability data gave the best predictions when incorporated into similarity metrics rather than dissimilarity ones (Table 3). This apparent discrepancy might well be attributed to the particular metrics of dissimilarity we have considered. For example, instead of computing confusion dissimilarities as the multiplicative inverse of the various confusion similarities, $D_{Wick}(TA) = 1/S_{Wick}(TA)$, and so on, we might equally well have computed dissimilarities as a decaying exponential function of the similarity values, $D'_{Wick}(TA) = \exp\{-\beta \cdot S_{Wick}(TA)\}$. In that case, we would find that the dissimilarity ratio, $D'(TB)/(D'(TA) + D'(TB))$, performs just as well as the similarity difference, S(TA) − S(TB). In fact, one is a simple logistic function of the other, and they produce exactly the same rank order of predicted response probabilities, so they are indistinguishable in the non-parametric analyses presented above.[6] In this sense, our results for the confusion-based metrics are fully consistent with the conclusion that participants' judgments of word similarity reflected an underlying scale of psychological dissimilarity.

Refining the basic featural measure

In light of the preceding results and discussion it seems fair to say that D_{PMV} is: (a) the simplest measure

[6] In parametric regression analyses (whose details are not reported here), the ratios of D′ dissimilarities fit the similarity judgment data much better than alternative metrics based on confusabilities (but not nearly as well as feature-based metrics).

of phoneme similarity in that it is trivial to derive for any language whatsoever, and that (b) none of the more costly metrics offer any real benefit in predicting similarity judgments (nor, as it turns out below, in predicting confusability across various tasks). The empirical measures perform considerably worse than D_{PMV}; the performance of Frisch's featural model is comparable to D_{PMV}, but it involves much more complexity in its computation. It is consequently D_{PMV} that recommends itself as a starting point in the search for better measures of phoneme similarity.

Making a feature scheme more fine-grained does not automatically lead to improvement. Wicklegren's (1966) study obtained relatively poor predictions of short-term memory errors from the Chomsky and Halle set of binary features. In Experiment 1 above, Frisch's metric based on detailed monovalent features performed substantially worse than the PMV metric based on coarse major class distinctions, and even in Experiment 2, Frisch's metric failed to achieve better results than PMV. This suggests the need for a guided search. Post hoc analyses revealed that, of the three major class features included in the PMV metric, manner of articulation was particularly important: allowing manner to carry more weight than voicing or place of articulation led to significantly improved data fits (also see Carter, 1987; Denes, 1963; Peters, 1963). Further inspection of the phoneme comparisons that were most poorly explained by the PMV metric suggested that differences in sonority might be relevant. We therefore added to the basic PMV metric a fourth, binary feature that expands the manner feature. It distinguishes the sonorant consonants from the obstruents. We abbreviate the new metric as PMVS, with S_{PMVS} and D_{PMVS} referring specifically to measures of similarity and dissimilarity, respectively. For convenience, a list of English consonants and their values on the PMVS featural dimensions of place, manner, voicing, and sonorant/obstruent is given in Appendix B along with the corresponding interphoneme distances.

As shown in Table 3, the inclusion of the sonority feature gave better predictions of similarity judgments than any other metric yet examined. Ratios of PMVS dissimilarities explained 67% of the variance, a significant improvement over PMV dissimilarities according to Steiger's test of differences between correlated correlations, $Z = 2.85$, $p = .005$. The ratios of PMVS dissimilarities also provided a significantly better characterization of participants' judgments than similarity differences did, $Z = 3.38$, $p = .001$.

Because the sonorant/obstruent distinction is nested within manner of articulation, one interpretation of its usefulness in predicting participants' similarity judgments is that some differences in manner of articulation are more salient than others. In particular, differences in manner which correspond to differences between sonorants and obstruents are evidently more salient than manner differences among sonorants or among obstruents. Additional distinctions could well be relevant among various manners of articulation, or indeed, among various places of articulation (e.g., is the difference between labials and velars more or less salient than the difference between alveolars and velars?). Nevertheless, it is notable that none of the four PMVS features is more important than any of the others in explaining our similarity judgments. With the inclusion of the sonorant/obstruent distinction in PMVS, the four features made roughly equal contributions, as indicated by parametric regressions that incorporated a weighting parameter for each feature (details not reported here). The best-fit parameters assigned roughly equal weights to each of these four major class features, and a regression model without parameters fit the data nearly as well as a model with separate weight parameters for each feature. This suggests that any further refinements of the PMVS metric involving more detailed features will have to preserve the overall balance between features representing place, manner, voicing, and sonorant/obstruent distinctions.

Other tasks: Predicting confusability

D_{PMVS} was the best measure at predicting the perceived similarities in the data from Experiment 2. However, explicit similarity tasks might simply not tap into the representations, and hence similarities, relevant to online processing. If that were the case, the fact that D_{PMVS} is the best current predictor of similarity judgments would tell us little about the relative merits of various similarity metrics applied to other tasks. Even if the auditory forced choice task of Experiment 2 is an ideal way to elicit similarity data, the kinds of similarity in operation here may not be the same as those relevant for word recognition, verbal short-term memory, or speech production.

The way to examine the generality of our findings is to test similarity data from several different online processing contexts. Indeed, we already have such data in the form of the confusability matrices from speech production, auditory perception under noise, and verbal short-term memory. We can simply examine the various metrics in their ability to predict the various kinds of confusability data. Although confusability data should not be considered to be a *privileged* standard against which to test measures of similarity, it is widely assumed that phoneme confusability is explained, at least to some extent, by the underlying similarity between phonemes. If PMVS is *generally* the best predictor of similarity, then it should be better than the other metrics at predicting the various phoneme confusabilities.

Results and discussion

We computed bivariate Spearman rank correlations for the various metrics of phoneme similarity, across the 171 pairs of phonemes common to all sets of confusion data. Our analyses considered the Luce data for initial and final consonants separately. For clarity, we present results here only for Luce's data obtained at a signal-to-noise ratio (SNR) of +5 dB (this SNR showed the highest correlation with the similarity data of Experiment 2). The results reported for these Luce matrices are representative of the others. Also, the correlation of one Luce matrix with another is of limited interest in the present context since the matrices constitute only minor variations of the same type of confusion.

Squared correlation coefficients are shown in Table 4, with the highest coefficient in each column in bold for emphasis. Both of the theoretical metrics, PMVS and Frisch, predict the various sets of confusion data better than one type of confusability predicts another. The only exception to this generalization is that Luce confusion data for final consonants is a better predictor of Luce confusabilities for initial consonants than is either of the theoretical metrics. However, this exception is arguably irrelevant to the question of how well a metric generalizes across different tasks, since the Luce confusion data for both initial and final consonants is from a single task applied to different stimuli. In comparisons between the theoretical metrics, PMVS is substantially better than the Frisch metric at predicting Luce confusabilities, but slightly worse at predicting confusabilities from the Wickelgren and MIT data sets.

The Frisch metric provides substantially better predictions of the MIT speech error data than of other types of confusability. This suggests the possibility that Frisch's measure might be picking up factors of importance to speech errors above and beyond any form of phoneme similarity that is common across various tasks.

Table 4
Spearman's R^2% between the metrics of similarity and various confusabilities

Predictor	Confusability			
	1	2	3	4
Confusabilities				
1. Wick		12	18	18
2. MIT	12		24	14
3. L + 5 Final C	18	24		(35)
4. L + 5 Initial C	18	14	(35)	
Feature metrics				
5. Frisch	**31**	**42**	31	17
6. PMVS	30	38	**42**	**30**

The largest value in each column is in bold.

Finally, in contrast to Experiment 2, the data examined here is not confined to just codas. As we noted in the discussion of Experiment 1, the MIT corpus primarily reflects onset confusions. The results of Experiment 1 raised the possibility that onset and coda similarities might differ. Comparing the relationship of the MIT corpus to the Luce measures derived from initial and final consonants, it is notable that the onset measure performs worse than the coda measure at predicting the MIT confusions. There is consequently no evidence from these tests for systematic differences between onset and coda similarities.

In summary, the two featural measures of similarity recommend themselves across the range of tests conducted, from explicit judgments through three very different kinds of confusability data. We think it likely that this pattern extends to the prediction of phoneme similarity across the wide range of psycholinguistic contexts in which similarity is relevant.

The relationship between similarity and confusability

In this final section, we examine the relationship between similarity and confusability more generally. As detailed in the Introduction, confusability is widely assumed to reflect similarity. However, it is evident from the results above that the individual sets of confusion data measure, to some extent, something other than what the theoretical metrics of similarity assess. It is also evident that the confusability metrics measure something other than what the similarity judgments of Experiments 1 and 2 assess. One thing that is not yet clear, is whether the various confusability metrics are measuring a single common core of confusability, which might be construed as similarity, or whether they are each measuring something different and idiosyncratic about the various tasks from which the confusabilities were drawn. The answer to this question is of interest because it constrains how well *any* measure of phoneme similarity could do in predicting confusabilities. However, the answer to this question also has implications for attempts to draw inferences about the general functioning of the language system on the basis of confusability data (e.g., Dell & Reich, 1981; Garrett, 1975; Harley, 1993; Levelt, 1992).

Given our current understanding of language processing, it is plausible, but by no means certain, that the various sets of confusabilities reflect a common underlying core. For illustration purposes, the reader might consider the organization for the language system proposed by Trevor Harley (Harley, 2001, Fig. 13.1). In Harley's scheme, the phonological output buffer emerges as a common point of contact between the various language processing tasks relevant to Wickelgren's confusions in short-term memory, Luce's perceptual

confusions of nonsense syllables, and the MIT corpus of speech errors. Harley's phonological output buffer is essential to speech production, so it could well influence speech errors like those in the MIT corpus. Neither Wickelgren's nor Luce's confusability tasks involved speech production, because responses in both studies were written. Nevertheless, responses to the auditory stimuli in these studies would have to pass through the phonological output buffer before being operated on by sound-to-letter rules on their way to the orthographic output buffer. If Harley's structure even approximates key aspects of the language system, then it is at least possible that the various confusabilities all reflect the extent to which phonemes get confused in the phonological output buffer, or some other part of the language system that is common to the various confusion tasks.

Of course, the various tasks differ in numerous ways, and there is ample opportunity for confusions in the different tasks to arise in different parts of the language system. However, confusion probabilities in different parts of the language system are unlikely to be independent. Parts of the language system necessarily interact with each other. What is more, the perceptual system makes sense of acoustic signals produced by the production system. As Goldstein and Fowler (2003, p. 174) put it, "there must be a common currency among knowers of language forms, producers of the forms and perceivers of them." To the extent that different sets of confusion probabilities reflect the common currency, they should be measuring the same thing. It is an empirical question to what extent this is so.

In this section, we test whether the three types of confusion data reflect a common core, the extent to which similarity judgments also measure the same core as the confusion matrices, and the extent to which core similarities are predicted by major class features (place, manner, voicing, and sonority). In particular, we determine the amount of variance in the confusability metrics that can be explained by a single latent variable (representing the common core of confusabilities). We also determine the amount of total variance in the similarity judgments as well as the confusability metrics that can be explained by a single latent variable representing a general notion of core similarity.

Results

Common core

We assume that each set of confusion data reflects several independent factors, including a core that is common to all of them, other factors specific to each type of confusability, and noise. Thus the various sets of confusion data can be described by the following models:

$$\text{MIT} = \text{Common Core} + \text{Speech Error Factors} + \text{Noise}$$

$$\text{Wick} = \text{Common Core} + \text{Memory Confusion Factors} + \text{Noise}$$

$$\text{Luce} = \text{Common Core} + \text{Perceptual Confusion Factors} + \text{Noise}$$

The relative contribution of the common core determines the extent to which the various confusabilities measure the same thing. We extracted a variable representing the common core by performing a factor analysis on MIT, Wickelgren and Luce metrics based on their ranked similarity difference scores for the stimulus triplets of Experiment 2.[7] There were 147 triplets for analysis, excluding triplets for which the Wickelgren matrix made no predictions. To represent the various Luce matrices in these analyses, we used the +5 dB matrix for final consonants since it is the most representative (as indicated by relatively high pairwise correlations with the other Luce matrices). The factor analysis identified the linear combination of confusability metrics that minimized total squared differences between the combination and the individual confusability metrics. Essentially, regression analysis fit an equation of the form

$$\text{Common Core} = \text{A} \times \text{MIT} + \text{B} \times \text{Wick} + \text{C} \times \text{Luce},$$

by choosing parameters A, B, and C so as to minimize the sum of squared differences

$$\text{SS} = (\text{U} \times \text{Core} - \text{MIT})^2 + (\text{V} \times \text{Core} - \text{Wick})^2 + (\text{W} \times \text{Core} - \text{Luce})^2.$$

The resulting Common Core factor extracts (explains) variance common to the three confusability metrics. The common core explained 66% of the overall variance within the three confusability metrics, including 76, 70, and 44% within the MIT, Luce, and Wickelgren metrics, respectively.

The common core explained substantially less variance in the Wickelgren metric than in the other two. That result indicates that the Wickelgren metric is relatively uncorrelated with the other two metrics (also see Table 4). Further analysis suggested that the Wickelgren data include a much higher proportion of noise than the

[7] These analyses were based on the sample of comparisons included in Experiment 2 instead of on all phoneme pairs common to the confusability matrices in order to facilitate comparisons with analyses below involving the similarity judgments from Experiment 2. Because our stimuli are a random sample of the possible comparisons, the results reported here should be representative of the full set of possible comparisons. The representativeness of our stimuli was confirmed in preliminary analyses with both the sample and the full set of possible comparisons.

other confusability data, that is, the Wickelgren data are more nearly what would be expected by chance.[8]

Similarity judgments and core similarity

We next examine whether the confusabilities contain any kind of common core that can be construed as similarity. That confusabilities reflect similarity is implicit in the widespread use of confusability data to estimate similarities. The common core that we find for the various confusabilities in our first analysis, however, need not reflect similarities. If it does, then we would expect shared variance among the confusabilities and our similarity judgment data. Consequently, we repeat the preceding analyses with a fourth variable, the judgments from Experiment 2. Again we test the degree to which a single underlying latent variable—core similarity—can explain the variance across all variables. We assume the following models:

$$\text{MIT} = \text{Core Similarity} + \text{Speech Error Factors} + \text{Noise}$$
$$\text{Wick} = \text{Core Similarity} + \text{Memory Confusion Factors} + \text{Noise}$$
$$\text{Luce} = \text{Core Similarity} + \text{Perceptual Confusion Factors} + \text{Noise}$$
$$\text{Sim Judgments} = \text{Core Similarity} + \text{Sim Judgment Factors} + \text{Noise}$$

The resulting Core Similarity factor extracts variance common to the four variables. Core Similarity explained 52% of the overall variance within the four variables, including 66, 45, and 68% within the MIT, Wickelgren, and Luce metrics, respectively, and 75% for similarity judgments.

Phonological features and core similarity

Finally, can Core Similarity be explained in terms of shared major class features along the lines of the PMVS metric, or does this metric leave a substantial amount of Core Similarity unexplained? Here, we decompose Core Similarity into independent components corresponding to PMVS, similarity due to other phonological features, and similarity not obviously related to phonological features at all (though we make no attempt to separately quantify the last two):

$$\text{Core Similarity} = \text{PMVS Similarity} + \text{NonPMVS Featural Sim} + \text{NonFeatural Sim}$$

We are interested in the direct relationship between Core Similarity and PMVS similarity. Because PMVS similarity is derived from theory, as opposed to being estimated from data, there is no latent variable mediating its relationship to Core Similarity. Accordingly, we simply asked what fraction of the variance in Core Similarity was explained by PMVS. We found that 71% of the rank variance in Core Similarity was explained by PMVS, using ratios of PMVS dissimilarities.

Discussion

These analyses give rise to four main conclusions. First, the common core explains a large fraction of the variability in the various confusability metrics, even though bivariate correlations among the metrics were fairly small (Table 4). Two thirds of the variability in confusabilities can be attributed to an underlying common core. The remainder reflects some combination of measurement error (noise) and idiosyncrasies of the particular tasks in which the confusions arise.

Second, the confusability matrices reflect phoneme similarity, in the sense that much of the variance within them is explained by a latent factor in common with metalinguistic judgments of similarity. This confirms that the confusability data provide good testing material for measures of similarity. A better measure of similarity should be able to explain relatively more variance across the range of confusability data, though the absolute levels of accuracy in predicting individual sets of confusability data will probably always be low, due to a combination of noise and idiosyncratic factors affecting individual confusability matrices. This combination of noise and idiosyncratic factors also means that individual sets of confusability data make relatively poor general-purpose predictors of phoneme similarity.[9] It is unclear at present exactly what the idiosyncratic factors

[8] We can compute a complexity coefficient to gauge the extent to which a confusion matrix reflects interesting structure rather than noise. Absolute deviation from chance can be measured by computing the usual log-likelihood chi-square statistic, $\chi^2 = 2\sum O_{ij}\ln(O_{ij}/E_{ij})$, where O_{ij} is the number of observed confusions between two phonemes, and E_{ij} is the number expected by chance. Factoring out the size of the confusion matrix and bearing in mind we have symmetrified it, we derive a normalized complexity 'coefficient, $C = \sqrt{\frac{\chi^2}{2N\ln(k-1)}}$, where N is the total number of confusions in the matrix, and k is the number of rows and columns in the matrix. The complexity coefficient ranges from 0 (exactly the expected number of confusions of all types) to 1 (maximum departure from expected confusions). Assessed by this metric for the 19 phonemes which are common to all, the Wickelgren data are much less complex than either the MIT or L + 5 Final C data, Cs = .133, .568, and .525, respectively.

[9] That the remainder is not simply noise can be inferred for L + 5 from the fact that it has a much higher R^2 correlation with L + 15 ($R^2 = 74$) confusabilities. In principle, correlations could also be computed between the MIT corpus and other speech error corpora, and it would be devastating for speech error-based research if such comparisons failed to produce high R^2 values.

are. However, there are other studies that confirm the considerable influence of idiosyncratic information. For example, Wang and Bilger (1973) found that auditory confusions under noise were not only sensitive to absolute levels of noise, SNRs, the consonant set under consideration, and the choice of vowel with which they were combined, but also that there were numerous interactions between these factors which seemed neither systematic nor readily interpretable.

One further possibility that would restrict the size of the core similarity in our estimates is that of task specific, but systematic, shifts in similarity. Up to a point, task specific similarities are compatible with a notion of 'phoneme similarity' as a single explanatory construct. It is possible, for example, that a common set of features receives differential weighting or importance across tasks (see also Wang & Bilger, 1973). In the wider cognitive literature, Nosofsky (1986) was able to provide an explanation of both identification and categorization data for simple artificial objects through a model based on similarity to previously encountered exemplars. The attention paid to the individual dimensions of these exemplars, and with that their similarities, changed across the two tasks. They were nevertheless systematically related through a single underlying set of relevant dimensions.

This notion of task-specific similarities is appealing. However, several considerations must be borne in mind. First, if present, task specific similarities are unlikely to be the only factor limiting the size of the core similarity in the analyses above, because Wang and Bilger (1973) were unable to fully relate different matrices of auditory confusions alone. Second, establishing task specific similarities without exploding similarity as an explanatory construct altogether requires an underlying theory about what the relevant commonalities or differences are from which similarities are derived. Only through such a set of relevant units can the similarities across tasks be related to each other. Only through such a set of relevant units, whether these be phonological features or something else, can similarity be anchored and made explanatory once it varies across tasks. The danger here is circularity. As detailed in the general introduction, phoneme similarity is widely used to explain patterns of behavior in psycholinguistic processing. For example, "similar sounding words are more likely to be confused in short-term memory." This requires some way of determining whether or not two words are similar in this task, other than the mere fact that they were confused. If 'similar words' is defined only as 'words that were confused', then the seeming explanation that words in short-term memory are confused *because* they are similar is entirely circular and explanatorily vacuous. Likewise, the fact that auditory confusions can predict auditory confusions seems neither useful nor interesting. In summary, the idea of task specific similarities places constraints on what a similarity measure must be like, and interacts with the degree to which similarity can be viewed as an explanatory construct at all.

Across the various data sets we examined D_{PMVS} was consistently the most robust predictor. In principle, D_{PMVS} could accommodate task specific notions of phoneme similarity, because features could be weighted differently across tasks. Nevertheless, it is not clear at present whether this would lead to better prediction.

Conclusions

Confusability data do not offer a shortcut which solves, at least on a practical level, the problem of measuring the similarity between phonemes. That is the conclusion of the comparisons between measures reported in this paper. Confusability-based measures offer no benefit over theoretical, feature-based measures in predicting phoneme similarity judgments in two auditory forced choice tasks. They are not even superior to these measures in predicting confusabilities. This was found to be the case despite investigating a considerable range of transformations to the basic confusability matrices in question, which examined asymmetrical and symmetrified versions, normalizations for frequency and response biases, consideration of different decision rules, as well as the aggregation of all matrices. There are, of course, always other things one could try, but the whole appeal of using confusability data lay in the idea that they provide a ready solution. If considerable amounts of work are required in order to obtain such a solution, then investing this effort into the development of a theory of phoneme similarity will always be more desirable: theoretically derived, featural measures are not just practically useful metrics—they offer an explanation of the similarity relationships obtained.

Our analyses found the confusability matrices share a common, core similarity. The degree of correlation of this core similarity with major class features suggests that it can indeed be construed as "phoneme similarity." This supports the use of phoneme similarity as an explanatory notion across a wide range of psycholinguistic contexts. On a practical level, it suggests that considerable progress in understanding the processes of verbal short-term memory, speech production or auditory perception might be achieved by focusing on those aspects of these processes which are unique; practically this means seeking to explain patterns in the data which remain once phoneme similarity has been factored out (e.g., Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1991a).

With regards to the general properties of phoneme similarity, our results suggest that phoneme similarity is based psychologically on an assessment of differences rather than commonalities. Though it is often assumed

that quantifying the degree of difference between two objects and quantifying the degree of commonality amounts to two equivalent ways of expressing the same thing, the two can deviate considerably in the degree of similarity they assign to a pair of objects (Hahn & Bailey, in press; Tversky, 1977). Degree of difference (or distance) takes as its point of departure identity between two objects and determines the amount of deviation from that; degree of commonality takes as its point of departure that two objects are different and seeks to determine how much they nevertheless share. It is an empirical matter which of these approaches underlies similarities of a given kind. The best predictions were achieved by variants that imply an underlying difference scale, and patterns in the raw data in line with general principles of magnitude estimation supported this view. Finally, phoneme similarity seems to be more influenced by differences in manner than in voicing or place of articulation, in line with the fact that manner carries the bulk of the information in distinguishing between words (Carter, 1987; Denes, 1963). Open questions concerning the nature of phoneme similarity, which can only be resolved through further research include the degree to which the similarities between vowels can equally well be captured by featural metrics and whether they are also based on differences, furthermore whether there are small but reliable asymmetries in phoneme similarities, and whether there are small but systematic differences in the way that onset and coda similarities are determined.

With regards to the measurement of phoneme similarity our current best measure of phoneme similarity, PMVS, is based on simple counts of the number of major articulatory features—place of articulation, voicing, and manner of articulation (within which is nested the sonority feature)—in which the two phonemes fail to match. Major class features provided the best overall predictor across the range of tests considered, that is, prediction of the similarity data from our two experiments as well as prediction of the speech errors held in the MIT corpus, auditory confusions under noise obtained by Luce (1986), and Wickelgren's data on confusions in short-term memory. Frisch's (1996) natural class measure is also better than the empirically derived measures based on confusability, but does noticeably worse than PMVS. There is one exception to this overall pattern in that the Frisch metric is noticeably better than PMVS in predicting the speech error data of the MIT corpus. In practical terms the success of PMVS is encouraging as it is so easy to calculate and use (indeed, for English consonants one can simply look up the number of mismatching features between two phonemes in Appendix B). The predictive accuracy of PMVS should be sufficient for most practical applications. It correctly predicts 67% of the rank variance in Experiment 2. Given the nature of the sample, this figure should extend to any representative selection of phoneme comparisons.

In absolute terms, of course, this measure can undoubtedly still be improved upon. The experience of the research reported in this paper suggests to us that the combination of more detailed linguistic analysis and insights from research on similarity in other domains provides the most promising route to better predictors of phoneme similarity.

Acknowledgments

We would like to thank Stefan A. Frisch, Emma Laing, Josie Briscoe, Gordon Harold, and two anonymous reviewers for helpful comments. The order of the authors is arbitrary.

Appendix A. Experiment 2 stimulus triplets, including a target consonant and two choice consonants, and the pseudo-words in which they were embedded. The choice judged most similar to the target is the 'winning' consonant or word. The final column gives the fraction of participants who agreed on the winning choice

	Tgt C	Win C	Lose C	Tgt Wd	Win Wd	Lose Wd	Win%
1	č	ǰ	ŋ	gɪč	gIǰ	gɪŋ	100
2	m	ŋ	k	ǰʌm	ǰʌŋ	ǰʌk	100
3	ž	š	p	guž	guš	gup	100
4	ð	θ	č	veð	veθ	več	100
5	θ	f	ǰ	jʌθ	jʌf	jʌǰ	97
6	v	ð	θ	biv	bið	biθ	97
7	f	θ	ŋ	glɛf	glɛθ	glɛŋ	97
8	ǰ	č	ð	θɛǰ	θɛč	θɛŋ	93
9	š	č	b	kwoš	kwoč	kwob	93
10	š	ž	v	biš	biž	biv	93
11	ǰ	v	k	ji ǰ	jiv	jik	93
12	v	ž	m	skev	skež	skem	93
13	θ	f	k	gruθ	gruf	gruk	93

(*continued on next page*)

Appendix A (*continued*)

	Tgt C	Win C	Lose C	Tgt Wd	Win Wd	Lose Wd	Win%
14	ǰ	č	l	dæǰ	dæč	dæl	93
15	v	ð	m	glev	gleð	glem	93
16	ǰ	ž	θ	gluǰ	gluž	gluθ	93
17	m	n	ð	zɪm	zɪn	zɪð	93
18	č	θ	v	blɛč	blɛθ	blɛv	90
19	f	θ	b	kruf	kruθ	krub	90
20	θ	f	p	dæθ	dæf	dæp	90
21	n	m	ǰ	zæn	zæm	zæǰ	90
22	ž	v	k	tuž	tuv	tuk	90
23	ŋ	m	l	pɛŋ	pɛm	pɛl	87
24	v	θ	g	hɪv	hɪθ	hɪg	87
25	θ	v	k	wʌθ	wʌv	wʌk	87
26	g	ǰ	l	snog	snoǰ	snol	87
27	g	k	š	preg	prek	preš	87
28	v	f	p	twɛv	twɛf	twɛp	87
29	k	g	m	zɪk	zɪg	zɪm	87
30	ǰ	v	ŋ	krIǰ	krɪv	krɪŋ	87
31	č	š	g	frɪč	frɪš	frɪg	87
32	č	š	m	več	veš	vem	87
33	θ	v	ŋ	prʌθ	prʌv	prʌŋ	87
34	ŋ	m	č	frɪŋ	frɪm	frɪč	83
35	p	č	l	zɛp	zɛč	zɛl	83
36	č	ǰ	f	pɛč	pɛǰ	pɛf	83
37	m	ð	p	fa^{j}m	fa^{j}ð	fa^{j}p	83
38	ð	g	p	krɪð	krɪg	krɪp	83
39	f	ð	l	pra^{j}f	pra^{j}ð	pra^{j}l	83
40	θ	š	m	droθ	droš	drom	83
41	θ	č	ŋ	prɛθ	prɛč	prɛŋ	80
42	ð	θ	k	zɪð	zɪθ	zɪk	80
43	v	ž	š	suv	suž	suš	80
44	m	b	k	prʌm	prʌb	prʌk	80
45	ǰ	č	θ	blɛǰ	blɛč	blɛθ	80
46	p	k	f	bla:p	bla:k	bla:f	80
47	v	b	l	zuv	zub	zul	80
48	m	ð	l	vum	vuð	vul	80
49	š	ǰ	ð	griš	griǰ	grið	80
50	č	p	l	smɪč	smɪp	smɪl	80
51	š	v	ŋ	grɪš	grɪv	grɪŋ	77
52	g	ǰ	ð	plog	ploǰ	ploð	77
53	š	f	p	væš	væf	væp	77
54	k	n	l	gɛk	gɛn	gɛl	77
55	ð	g	l	ǰoð	ǰog	ǰol	77
56	ž	š	č	fuž	fuš	fuč	77
57	b	g	n	stub	stug	stun	77
58	ǰ	g	k	ðʌǰ	ðʌg	ðʌk	77
59	n	ð	k	gla^{j}n	gla^{j}ð	gla^{j}k	77
60	θ	š	ž	siθ	siš	siž	77
61	g	v	p	trig	triv	trip	77
62	p	k	v	gɜp	gɜk	gɜv	77
63	ž	š	f	jež	ješ	jef	77
64	m	g	š	flɔ:m	flɔ:g	flɔ:š	73
65	v	š	ŋ	spʌv	spʌš	spʌŋ	73
66	ž	ð	n	huž	huð	hun	73
67	ǰ	g	θ	ploǰ	plog	ploθ	73
68	ǰ	θ	p	ǰuǰ	ǰuθ	ǰup	73
69	θ	ž	l	snuθ	snuž	snul	73
70	θ	ž	m	peθ	pež	pem	73
71	k	ǰ	l	skok	skoǰ	skol	73

Appendix A (*continued*)

	Tgt C	Win C	Lose C	Tgt Wd	Win Wd	Lose Wd	Win%
72	n	p	č	gɛn	gɛp	gɛč	73
73	f	k	l	muf	muk	mul	73
74	l	ǰ	š	snæl	snæǰ	snæš	70
75	š	v	m	skoš	skov	skom	70
76	f	ǰ	ð	hɪf	hIǰ	hɪŋ	70
77	č	θ	m	zič	ziθ	zim	70
78	č	k	n	θič	θik	θin	70
79	p	k	ž	ša^{j}p	ša^{j}k	ša^{j}ž	70
80	θ	ð	v	ma^{j}θ	ma^{j}ð	ma^{j}v	73
81	p	θ	ž	flup	fluθ	fluž	70
82	š	č	ǰ	miš	mič	miǰ	70
83	v	ž	ǰ	nuv	nuž	nuǰ	70
84	ð	š	ŋ	frɪð	frɪš	frɪŋ	70
85	v	b	m	ǰuv	ǰub	ǰum	70
86	ð	g	č	gið	gig	gič	70
87	č	ǰ	θ	wɔːč	wɔːǰ	wɔːθ	70
88	g	m	č	jug	jum	juč	70
89	m	v	l	prem	prev	prel	70
90	n	ǰ	š	skʌn	skʌǰ	skʌš	70
91	č	f	b	šɜč	šɜf	šɜb	70
92	č	f	k	tuž	tuf	tuk	70
93	č	v	ŋ	grʌč	grʌv	grʌŋ	67
94	ǰ	m	p	na^{j}ǰ	na^{j}m	na^{j}p	67
95	f	ž	ǰ	flef	fležˇ	fleǰ	67
96	ǰ	g	ŋ	stɛǰ	stɛg	stɛŋ	67
97	ž	g	m	wiž	wig	wim	67
98	f	č	b	grɪf	grɪč	grɪb	67
99	f	š	č	zof	zoš	zoč	67
100	p	g	v	gop	gog	gov	67
101	b	ð	n	frɪb	frɪð	frɪn	67
102	ð	ǰ	l	la^{w}ð	la^{w}ǰ	la^{w}l	67
103	p	b	v	gip	gib	giv	67
104	l	ž	v	čil	čiž	čiv	67
105	b	k	ž	hub	huk	huž	67
106	ž	č	g	než	neč	neg	67
107	ŋ	g	š	glɪŋ	glɪg	glɪš	67
108	k	v	ŋ	frɪk	frɪv	frɪŋ	63
109	ŋ	š	p	zɛŋ	zɛš	zɛp	63
110	č	p	b	stɜč	stɜp	stɜb	63
111	f	g	n	tef	teg	ten	63
112	p	č	ð	sep	seč	seð	63
113	č	ǰ	š	zæč	zæǰ	zaeš	63
114	š	v	n	još	jov	jon	63
115	n	g	b	vin	vig	vib	63
116	ð	n	š	brɪð	brɪn	brɪš	63
117	v	ǰ	p	mʌv	mʌǰ	mʌp	63
118	g	ð	m	frig	friðˇ	frim	63
119	v	b	g	ziv	zib	zig	63
120	f	b	l	bæf	bæb	bæl	63
121	ŋ	g	b	frʌŋ	frʌg	frʌb	63
122	č	v	m	prʌč	prʌv	prʌm	63
123	ŋ	f	k	gʊŋ	gʊf	gʊk	63
124	š	ð	g	teš	teð	teg	63
125	g	k	ŋ	gʌg	gʌk	gʌŋ	63
126	m	θ	p	vem	veθ	vep	63
127	n	θ	f	jin	jiθ	jif	63
128	ŋ	f	p	glæŋ	glæf	glæp	63
129	š	m	b	zeš	zem	zeb	63

(*continued on next page*)

Appendix A (*continued*)

	Tgt C	Win C	Lose C	Tgt Wd	Win Wd	Lose Wd	Win%
130	ž	č	f	pež	peč	pef	63
131	v	š	č	gɛv	gɛš	gɛč	60
132	m	ð	f	skum	skuð	skuf	60
133	ǰ	v	g	fræǰ	fræv	fræg	60
134	θ	n	l	wæθ	wæn	wæl	60
135	f	č	g	θɔːf	θɔːč	θɔːg	60
136	m	ž	f	hjum	hjuž	hjuf	60
137	n	ð	f	klen	kleð	klef	60
138	θ	ð	g	rɛθ	rɛŋ	rɛg	60
139	ŋ	ð	š	prɪŋ	prɪð	prɪš	57
140	l	θ	š	zɛl	zɛθ	zɛš	57
141	ž	č	b	spež	speč	speb	57
142	v	p	m	prev	prep	prem	57
143	l	θ	p	trɛl	trɛθ	trɛp	57
144	ŋ	f	č	trʌŋ	trʌf	trʌč	57
145	n	g	p	pren	preg	prep	57
146	k	v	š	zuk	zuv	zuš	57
147	p	θ	n	zep	zeθ	zen	57
148	k	b	ž	vaːk	vaːb	vaːž	57
149	p	ǰ	ŋ	krɪp	krIǰ	krɪŋ	57
150	ŋ	ǰ	š	spɪŋ	spIǰ	spɪš	53
151	l	g	š	zul	zug	zuš	53
152	l	č	š	spʌl	spʌč	spʌš	53
153	ǰ	b	ŋ	šʌǰ	šʌb	šʌŋ	53
154	l	č	ǰ	zɪl	zɪč	zIǰ	53
155	m	ŋ	b	čɛm	čɛŋ	čɛb	53
156	m	č	m	dɛp	dɛč	dɛm	53
157	m	k	č	glem	glek	gleč	53
158	g	θ	f	ǰig	ǰiθ	ǰif	53
159	m	k	l	grom	grok	grol	53
160	g	m	p	gɜg	gɜm	gɜp	53
161	ǰ	b	v	floǰ	flob	flov	53
162	š	f	ǰ	gluš	gluf	gluǰ	53
163	g	n	č	frɛg	frɛn	frɛč	53
164	č	b	m	preč	preb	prem	53
165	ŋ	θ	k	lɛŋ	lɛθ	lɛk	53
166	p	ð	ž	fup	fuð	fuž	53
167	n	p	k	θaʲn	θaʲp	θaʲk	53
168	g	θ	š	flig	fliθ	fliš	53
169	n	ð	ǰ	trɪn	trɪð	trIǰ	53
170	ŋ	ð	b	gɪŋ	gɪð	gɪb	53
171	č	p	f	jɛč	jɛp	jɛf	53
172	b	ǰ	f	glæb	glæǰ	glæf	53
173	ŋ	ǰ	l	gɛŋ	gɛǰ	gɛl	53
174	b	č	l	vub	vuč	vul	53
175	f	ŋ	k	wɛf	wɛŋ	wɛk	50
176	š	m	g	juš	jum	jug	50
177	p	ǰ	č	fep	feǰ	feč	50
178	ǰ	b	g	snIǰ	snɪb	snɪg	50
179	ð	ž	f	jið	již	jif	50
180	f	ǰ	b	trɛf	trɛǰ	trɛb	50

Appendix B. Major class features (place, sonority-obstruent, manner, and voicing) and number of featural differences for English consonants

C	Place	Son	Man	Vce	p	b	f	v	m	w	θ	ð	t	d	s	z	n	l	r	č	ǰ	š	ž	j	k	g	ŋ	h
p	Lab	Obs	Stop	Vls	0	1	1	2	3	3	2	3	1	2	2	3	4	4	4	2	3	2	3	4	1	2	4	2
b	Lab	Obs	Stop	Vcd	1	0	2	1	2	2	3	2	2	1	3	2	3	3	3	3	2	3	2	3	2	1	3	3
f	Lab	Obs	Fric	Vls	1	2	0	1	3	3	1	2	2	3	1	2	4	4	4	2	3	1	2	4	2	3	4	1
v	Lab	Obs	Fric	Vcd	2	1	1	0	2	2	2	1	3	2	2	1	3	3	3	3	2	2	1	3	3	2	3	2
m	Lab	Son	Nas	Vcd	3	2	3	2	0	1	4	3	4	3	4	3	1	2	2	4	3	4	3	2	4	3	1	4
w	Lab	Son	Gld	Vcd	3	2	3	2	1	0	4	3	4	3	4	3	2	2	2	4	3	4	3	1	4	3	2	4
θ	Dent	Obs	Fric	Vls	2	3	1	2	4	4	0	1	2	3	1	2	4	4	4	2	3	1	2	4	2	3	4	1
ð	Dent	Obs	Fric	Vcd	3	2	2	1	3	3	1	0	3	2	2	1	3	3	3	3	2	2	1	3	3	2	3	2
t	Alv	Obs	Stop	Vls	1	2	2	3	4	4	2	3	0	1	1	2	3	3	3	2	3	2	3	4	1	2	4	2
d	Alv	Obs	Stop	Vcd	2	1	3	2	3	3	3	2	1	0	2	1	2	2	2	3	2	3	2	3	2	1	3	3
s	Alv	Obs	Fric	Vls	2	3	1	2	4	4	1	2	1	2	0	1	3	3	3	2	3	1	2	4	2	3	4	1
z	Alv	Obs	Fric	Vcd	3	2	2	1	3	3	2	1	2	1	1	0	2	2	2	3	2	2	1	3	3	2	3	2
n	Alv	Son	Nas	Vcd	4	3	4	3	1	2	4	3	3	2	3	2	0	1	1	4	3	4	3	2	4	3	1	4
l	Alv	Son	Lat	Vcd	4	3	4	3	2	2	4	3	3	2	3	2	1	0	1	4	3	4	3	2	4	3	2	4
r	Alv	Son	Rhot	Vcd	4	3	4	3	2	2	4	3	3	2	3	2	1	1	0	4	3	4	3	2	4	3	2	4
č	Pal	Obs	Aff	Vls	2	3	2	3	4	4	2	3	2	3	2	3	4	4	4	0	1	1	2	3	2	3	4	2
ǰ	Pal	Obs	Aff	Vcd	3	2	3	2	3	3	3	2	3	2	3	2	3	3	3	1	0	2	1	2	3	2	3	3
š	Pal	Obs	Fric	Vls	2	3	1	2	4	4	1	2	2	3	1	2	4	4	4	1	2	0	1	3	2	3	4	1
ž	Pal	Obs	Fric	Vcd	3	2	2	1	3	3	2	1	3	2	2	1	3	3	3	2	1	1	0	2	3	2	3	2
j	Pal	Son	Gld	Vcd	4	3	4	3	2	1	4	3	4	3	4	3	2	2	2	3	2	3	2	0	4	3	2	4
k	Vel	Obs	Stop	Vls	1	2	2	3	4	4	2	3	1	2	2	3	4	4	4	2	3	2	3	4	0	1	3	2
g	Vel	Obs	Stop	Vcd	2	1	3	2	3	3	3	2	2	1	3	2	3	3	3	3	2	3	2	3	1	0	2	3
ŋ	Vel	Son	Nas	Vcd	4	3	4	3	1	2	4	3	4	3	4	3	1	2	2	4	3	4	3	2	3	2	0	4
h	Glot	Obs	Fric	Vls	2	3	1	2	4	4	1	2	2	3	1	2	4	4	4	2	3	1	2	4	2	3	4	0

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition, 90*, 119–161.

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*, 144–151.

Baddeley, A. D. (1966). Short term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology, 18*, 362–365.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language, 44*, 568–591.

Bailey, T. M., & Plunkett, K. (2002). Phonological specificity in early words. *Cognitive Development, 17*(2), 1265–1282.

Benki, J. R. (2003). Analysis of english nonsense syllable recognition in noise. *Phonetica, 60*, 129–157.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*, 425–455.

Carter, D. M. (1987). An information-theoretic analysis of phonetic dictionary access. *Computer Speech and Language, 2*, 1–11.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition?. *Journal of Memory and Language, 32*, 193–210.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior, 20*, 611–629.

Denes, P. B. (1963). On the statistics of spoken English. *Journal of the Acoustical Society of America, 35*, 892–904.

Frisch, S. A. (1996). *Similarity and Frequency in Phonology*. PhD thesis, Dept. of Linguistics, Northwestern University, Evanston Illinois, US Available online as Rutgers Optimality Archive 198-0597: http://roa.rutgers.edu.

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

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

Garrett, M. F. (1975). The analysis of sentence production. In G. Bower (Ed.). *The psychology of learning and motivation* (Vol. 9, pp. 133–177). New York: Academic Press.

Gerken, L., Murphy, W. D., & Aslin, R. N. (1995). 3-year-olds' and 4-year-olds' perceptual confusions for spoken words. *Perception & Psychophysics, 57*, 475–486.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28*, 501–518.

Goldstein, L. M., & Fowler, C. (2003). Articulatory phonology: A phonology for public language use. In A. S. Meyer & N. O. Schiller (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 159–207). Berlin: Mouton de Gruyter.

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning. Memory, and Cognition, 20*, 3–28.

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

Hahn, U., & Bailey, T. M. (in press). What makes words sound similar? *Cognition*. Available online (http://www.sciencedirect.com/science/article/B6T24-4F8TKDH-1/2/40678ecdc7f35dd3e39134ab8afe9d6f).

Hahn, U., & Nakisa, R. C. (2000). German inflection: Single route or dual route? *Cognitive Psychology, 41*, 313–360.

Harley, T. A. (1993). Phonological activation of semantic competitors during lexical access in speech production. *Language and Cognitive Processes, 8*, 291–309.

Harley, T. A. (2001). *The psychology of language: From data to theory* (2nd ed.). Hove: Psychology Press.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*, 1–23.

Kenstowicz, M. (1994). *Phonology in generative grammar*. Cambridge, MA: Blackwell.

Levelt, W. J. M. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition, 42*, 1–22.

Luce, P. A. (1986). *Neighborhoods of Words in the Mental Lexicon*. PhD Dissertation, Dept. of Psychology, Indiana University, Bloomington Indiana.

Luce, P. A., Pisoni, D. B., & Goldinger, S. B. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and comutational perspectives* (pp. 122–147). Cambridge, MA: MIT Press.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Marslen-Wilson, W., Moss, H. E., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1376–1392.

McCarthy, J. J. (1988). Feature geometry and dependency: A review. *Phonetica, 43*, 84–108.

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

Mueller, S. T., Seymour, T. L., Kieras, D. E., & Meyer, D. E. (2003). Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29*, 1353–1380.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology, 23*, 94–140.

O'Grady, W., Dobrovolsky, M., & Aronoff, M. (1989). *Contemporary linguistics: An introduction*. New York: St. Martin's Press.

Padgett, J. (2002). Feature classes in phonology. *Language, 78*, 81–110.

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

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

Prasada, S., & Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes, 8*, 1–56.

Shattuck-Hufnagel, S., & Klatt, D. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior, 18*, 41–55.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.

Stemberger, J. P. (1990). Wordshape errors in language production. *Cognition, 35*, 123–157.

Stemberger, J. (1991a). Radical underspecification in language production. *Phonology, 8*, 73–112.

Stemberger, J. (1991b). Apparent anti-frequency effects in language production: the addition bias and phonological underspecification. *Journal of Memory and Language, 20*, 41–55.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

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

Weber, E. H. (1835/1978). Concerning touch (H.E. Ross, Trans.). New York: Academic Press.

Wicklegren, W. A. (1965). Distinctive features and errors in short-term memory for english vowels. *Journal of the Acoustical Society of America, 38*, 583–588.

Wicklegren, W. A. (1966). Distinctive features and errors in short-term memory for english consonants. *Journal of the Acoustical Society of America, 39*, 388–398.