Grounding sound change in ideal observer models of perception

Zachary J. Burchill
Dept. of Brain & Cognitive Sciences
University of Rochester
`zburchil@ur.rochester.edu`

T. Florian Jaeger
Dept. of Brain & Cognitive Sciences
University of Rochester
`fjaeger@ur.rochester.edu`

Abstract

An important predictor of historical sound change, functional load, fails to capture insights from speech perception. Building on ideal observer models of word recognition, we devise a new definition of functional load that incorporates both *a priori* expectedness and perceptual information. We explore this new measure with a simple model and find that it outperforms traditional measures.

1 Introduction

Whether a phonemic contrast (e.g., /t/ vs /d/) merges over historical time has been taken to depend on its **functional load**—essentially its utility in distinguishing different meanings (Gilliéron, 1918; Jakobson, 1931; Malthesius, 1931). Traditionally, functional load is operationalized as the number of minimal pair words distinguished only by the contrast (e.g., for the /t/-/d/ contrast, the words *ten* and *den*). Studies based on this measure have found higher functional load to be associated with lowered likelihood of merging (Wedel et al., 2013). However, traditional measures of functional load fail to capture important properties from speech perception that are critical if one seeks to understand how a phonological contrast adds to word identification in actual language use. Specifically, research on speech perception has found that the functional contribution of phonemic contrasts to word identification depends on:

i. the *degree* to which the contrast is perceptually distinguishable,

ii. the *degree* to which the contrast gradiently adds to the distinction between *any* pair of words, not only minimal pairs,

iii. the *interaction* of (i) and (ii): how the perceptual distinguishability affects word recog-

nition depends on the word's specific phonological neighbors.

Building on ideal observer models of spoken word recognition such as Norris and McQueen (2008) and similarly, Luce and Pisoni (1998), we integrate (i)-(iii) into a revised definition of functional load. We then evaluate this novel definition empirically.

2 Theoretical model

In order to understand the amount of "work" a contrast does, we first seek to understand the implications of *losing* the contrast. Traditionally functional load is thought of as the number of meaning distinctions lost when a contrast disappears. Our revised definition measures functional load as the overall decrease in *probability* of correctly recognizing meaning. We measure functional load by first calculating the confusion in the system with the contrast in place, then, merging the contrast in question, re-calculate the confusion of the system without the contrast and measure the difference. For the present purpose, we follow previous work and operationalize meanings as lexical entries (or short, words). Thus:

$$FL_c = \mathrm{E}[\mathrm{p}(\mathrm{CorrectWordRecognition_{with\ c}})] - \mathrm{E}[\mathrm{p}(\mathrm{CorrectWordRecognition_{without\ c}})] \tag{1}$$

Equation (1) defines functional load as the expected value (i.e. the mean) of the probability of correct word recognition when the system has the contrast c minus the expected value for the same language without the contrast. In this paper we entertain two conceptualizations for the expected value: one for which each word/meaning type is equally important (which we shall refer to as *type-based* functional load) and one in which words are weighted by their relative frequencies (*token-based* functional load).

Taking context into account, the type-based expected value of the probability of recognizing a word (i.e.: $E[p(CorrectWordRecognition)]$) would be equivalent to taking the average of the probabilities of recognizing all words in the lexicon, i.e.:

$$\frac{\sum_{w_s}^{L} f(p(w_h|w_s, ctxt))}{|L|} \quad (2)$$

where L is the set of word types in the lexicon, w_s is the word spoken by the talker, w_h is the word heard by the interlocutor, and $ctxt$ is the context the word was spoken in. f is a function of the choice rule that determines $p(w_h = w_s|w_s, ctxt)$ from the distribution of all potentially heard words, $p(w_h|w_s, ctxt)$. For example, f could be the *criterion rule* (Green and Swets, 1966) whereby the probability of the listener correctly recognizing the word is calculated as the probability that the intended word has the highest chance of being heard out of all other candidates.

Token-based functional load can be formulated similarly, but instead of merely averaging the probabilities of recognizing each word, one takes the weighted average, where each probability is weighted by the word's frequency:

$$\frac{\sum_{w_s}^{L} p(w_s|ctxt) \cdot p(ctxt) \cdot f(p(w_h|w_s, ctxt))}{\sum_{w_s}^{L} p(w_s|ctxt) \cdot p(ctxt)} \quad (3)$$

These two approaches make different claims about the pressures on language change—token-based accounts imply that the average number of misrecognitions exerts the pressure, whereas type-based accounts would suggest the number of meanings confused is most important. Similar distinctions have been made previously, for example, Martinet (1952) and Hockett (1955) argue that lexical frequency should affect functional load, similar to how we weigh probabilities of recognition in our token-based account. Conceivably, either approach might better capture how languages change, or perhaps some higher-level utility function weighs words' relative importance (e.g., words that have greater utility in recognizing contribute more). One of the benefits of our modeling approach is that we can quantitatively compare the two accounts.

Our proposal shares with entropic accounts (Hockett, 1967; Surendran and Niyogi, 2003; Surendran and Niyogi, 2006) that it incorporates the *probability* of recognizing words (via contextual predictability). However, the present proposal also incorporates perceptual confusability. On the other hand, ideal observer models predict that word recognition depends on both perceptual distinguishability and contextual predictability (capturing both bottom-up and top-down influences on word recognition). We therefore estimate $f(p(w_h = w_i|w_i, ctxt))$, the probability of correctly recognizing word w_i, by comparing the *perceptual* confusability of w_i with all other words in the lexicon, weighted by frequency, following Luce and Pisoni (1998).

3 Approach

3.1 Estimating word recognition rates

Although contextual predictability could be included in future versions of our model, we ignore it for simplicity's sake, essentially approximating rates of isolated word recognition. We calculate the predicted probability of recognizing an isolated word w_i as:

$$\frac{PerceptualConfusability(w_i, w_i) \cdot \log Freq_i}{\sum_{j \in L} PerceptualConfusability(w_j, w_i) \cdot \log Freq_j} \quad (4)$$

where the L is the set of all word forms in the lexicon (i.e. pairwise calculations with w_i and every other word form, w_j, including itself). The general formulation of Equation (4) follows Luce and Pisoni (1998), using Luce's choice rule—i.e. hearing each word out of all the alternatives at a probability equal to the probability that particular word (Luce, 1959). In essence, the numerator is a measure of how likely the listener is to hear the phones in w_i correctly, weighted by frequency of w_i. The denominator is the frequency-weighted sum of the probabilities of mishearing the phones in w_i as all other possible words.

We estimate the perceptual confusability of two strings of phones, $PerceptualConfusability(x_j, x_i)$, as the product of the probabilities of confusing phones in the first string (x_i) with phones at the same locations in the second string (x_j), using probabilities from a phoneme-to-phoneme confusion matrix we built with perceptual data collected from Cutler et al. (2004).

$$\text{PerceptualConfusability}(x_j, x_i) = \prod_{a=1}^{n} \mathrm{P}(p_{aj}|p_{ai})$$
(5)

where n is the number of phones in x_i and $\mathrm{P}(p_{aj}|p_{ai})$ represents the conditional probability of hearing the ath phone in x_i (p_{ai}) as the ath phone in x_j (p_{aj}). The formulation of Equation (5) currently entails that we are only considering words of equal length to the target when calculating perceptual confusion. Future work can consider deletion and insertion confusions (Levy, 2008).

With the phonological data from the CMU Pronouncing Dictionary (Weide, 1998) and lexical frequencies collected from the SUBTLEX-US frequency database (Brysbaert and New, 2009), we can estimate the probability of correctly recognizing over 42,000 word forms of American English. The recognition rates predicted by this method have been shown to mirror properties of spoken word recognition (Luce and Pisoni, 1998). Combining these data, we can calculate $\mathrm{E}[\mathrm{p}(CorrectWordRecognition_{\text{with c}})]$ from (1).

3.2 Simulating contrast loss

The perceptual confusion matrix our model uses can be considered an estimate of the perceptual distinguishability of all phonemes in English as it exists currently. To simulate a hypothetical version of English that *does not* contain a certain contrast, we can artificially manipulate the confusion matrix our model uses to make the phonemes in the contrast perceptually indistinguishable.

For example, to simulate a variety in English in which there is no longer a distinction between /t/ and /d/, we can redistribute the probability mass of the confusion matrix so that $\mathrm{p}(/t/_{heard}|/d/_{spoken}) = \mathrm{p}(/t/_{heard}|/t/_{spoken})$ and vice versa (Figure 1). The probability that a listener would recognize a /t/ as a /t/ would then be equal to the probability they recognize the phoneme as a /d/. Combining this altered confusion matrix with (4), we can rerun our model, calculating the average probability of correctly recognizing words without the contrast, and estimate $\mathrm{E}[\mathrm{p}(CorrectWordRecognition_{\text{without c}})]$.
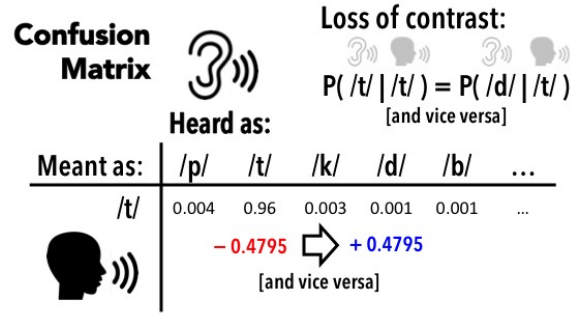

Figure 1: Manipulating the probability mass in the phoneme-to-phoneme confusion matrix to simulate losing the /t/-/d/ contrast.

4 Verifying our model's properties

Earlier, we stated that a definition of functional load that captures what is known about spoken word recognition should depend on (i)-(iii). In the follow sections, we verify that our model captures these properties (though under simplifying assumptions).

4.1 Revised functional load depends on contrasts' *a priori* perceptual distinguishability

Phonemes differ in their ease of recognition and differentiation acoustically and perceptually (Miller and Nicely, 1955; Wang and Bilger, 1973; Cutler et al., 2004). Some phonemes are harder to correctly recognize than others, and some contrasts are more readily confused. The loss of a phonemic contrast with high *a priori* perceptual confusability will affect word recognition less than losing a contrast between two very perceptually distinct phonemes, all other things being equal. For example, the /ɑ/-/ɔ/ contrast in American English has low perceptual distinguishability, even in non-merging dialects (Cutler et al., 2004). Losing this contrast increased predicted misrecognition rates by 6.6%. In comparison, removal of the same contrast in a simulated version of English where /ɑ/-/ɔ/ are highly distinct (i.e., the original confusability reduced by 80%) increased misrecognition rates by 10.1%.

4.2 Contrast loss affects more than just minimal pairs

Traditional functional load—including recent extensions (Surendran and Niyogi, 2006; Wedel et al., 2013)—fail to capture the fact that real-world
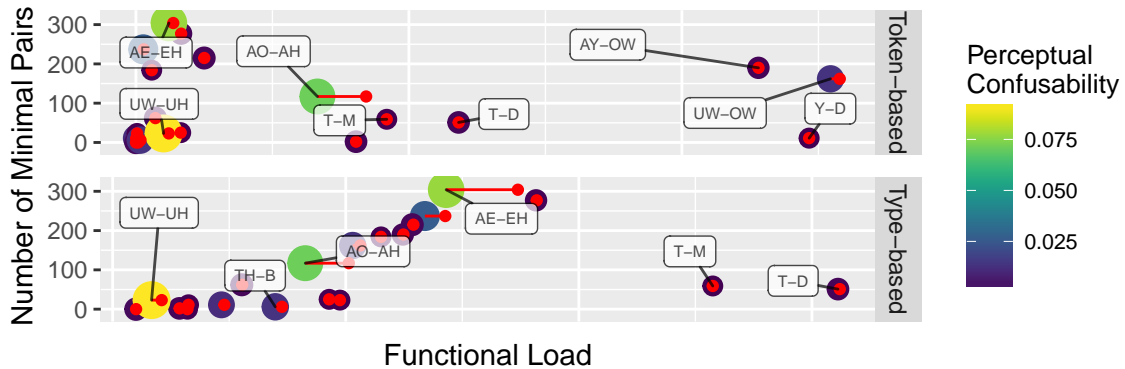
Figure 2: Twenty randomly chosen mergers demonstrating the complex relationship of perception, phonotactics, and the lexicon (with Arpabet labels). Size and color of dots indicate the *a priori* perceptual confusability of the contrast, as estimated by the original confusion matrix. Axes are scaled independently. The red lines show the increase in the contrast's functional load had the contrast originally been perfectly perceptually distinct—in essence, the sensitivity of the contrast's functional load to its *a priori* confusability.

spoken word recognition is influenced by word-to-word confusability beyond the confusability of minimal pairs. For example, the perceptual confusability of pairs like /gɑt/ and /cɔt/ will never be taken into account, yet these two words are likely to be relatively confusable, even in American dialects that preserve the /ɑ/-/ɔ/ distinction. Following a /g/-/k/ merger, /gɑt/ and /cɔt/ would not be completely indistinguishable, but the two words would be *highly* confusable.

Our approach compares the perceptual distinguishability of each word and *all* other words in the lexicon (the present implementation: all words with the same number of phonemes), avoiding this problem. To illustrate how this changes the operationalization of functional load, we compared token-based functional load of the /g/-/k/ contrast under two different assumptions: one in which only minimal pairs were considered neighbors and one in which all the words in the lexicon were considered. When our model only used minimal pairs as neighbors (i.e. for each w_i in (4), L was restricted to w_i and its /g/-/k/ minimal pairs), words containing /g/ or /k/ had a mean probability of misrecognition of 0.023; when the model used all non-minimal pair neighbors of the same length (as we do in the rest of this paper), misrecognition increased to a probability of 0.064, corresponding to a three-fold increase in odds.

4.3 Functional load depends on a complex interaction of elements

Our definition of functional load does predicts interactions between a contrast's perceptual distinguishability and its distribution across words and neighbors in the lexicon. In this section, we first highlight the differences in predictions between our type- and token-based accounts. We then demonstrate how perceptual confusability, distribution of the phones in the lexicon, and the interactions between the two are predicted to affect functional load. Figure (2) demonstrates the outcomes of these complex relationships for 20 randomly chosen phonemic contrasts. Both type- and token-based functional load are plotted against the number of minimal pairs formed by each contrast. The red lines represent how sensitive that contrast's functional load is to *a priori* perceptual confusability. The red dots represent a contrast's functional load, had its phones been completely perceptually distinguishable to begin with. In general, we would expect that contrasts with low *a priori* perceptual confusability would be less sensitive to eliminating this confusability, as they are already closer to perfect distinguishability.

We see first that type- and token-based accounts of functional load differ noticeably in their estimates of functional load. Figure (2) shows a fairly linear relationship between type-based functional load and the traditional measure, and although this relationship becomes less clear with
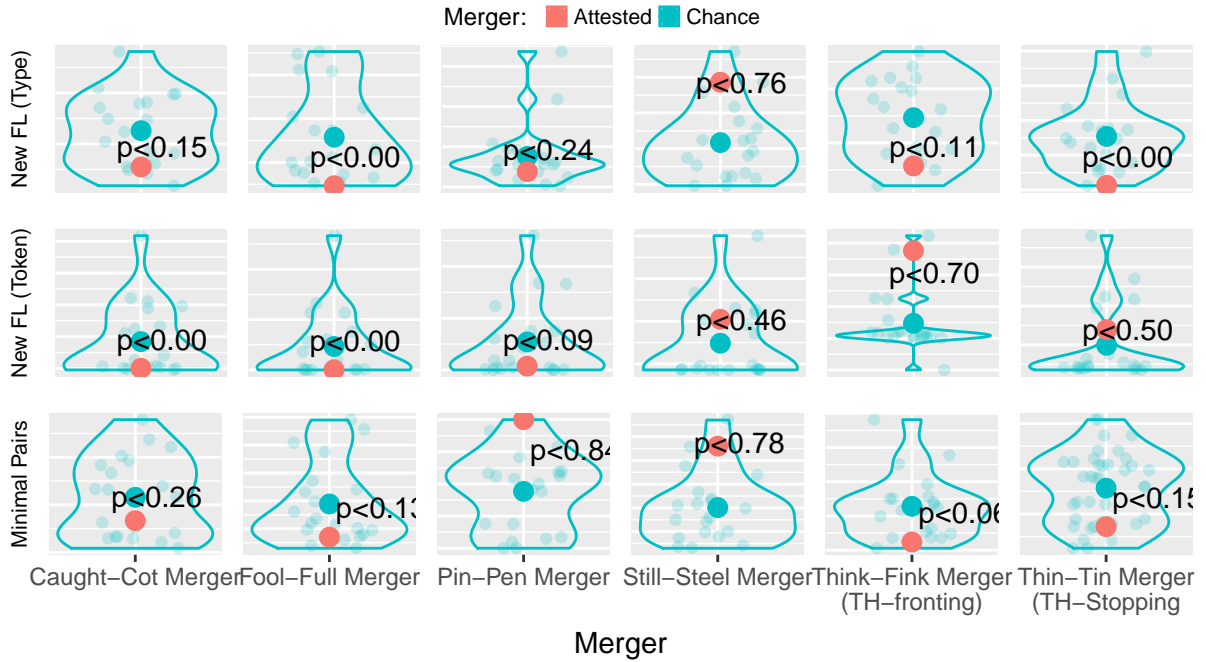
Figure 3: The functional loads of six attested mergers in North American English each compared to 20 randomly generated mergers in the same environments, measured by models implementing our new definition of functional load vs. the traditional measure of number of minimal pairs. The probability that a sample drawn from the distribution of random mergers has less than or equal functional load than the attested merger is indicated. Axes are scaled independently for each plot.

more contrasts (not pictured), the token-based account is noticeably less correlated with the number of minimal pairs. This difference makes sense: type-based functional load and the number of minimal pairs a contrast separates both emphasize word types. However, it also serves as a demonstration that lexical-phonological environments are not generally uniform across words of different frequencies–further evidence that choosing between type- vs. token- approaches should be a matter of careful consideration (Hockett, 1955; Martinet, 1952; Wedel et al., 2013).

Secondly, there are contrasts whose functional loads seem to be determined more by their initial perceptual distinguishability, or possibly by how they are distributed across non-minimal pair neighbors. Although the /t/-/m/ contrast does not separate many minimal pairs, the type-based account predicts that a merger would be unlikely, possibly due to the low *a priori* perceptual confusability of the phones. A benefit of our model is that it allows the researcher to ask questions about how the lexicon and phonological system interact; exactly how these contribute to a contrast's functional load can be teased apart.

Third, although the functional loads of naturally confusable contrasts are generally more sensitive to perceptual changes, there are exceptions to this pattern. For example, the functional load for the most perceptually confusable contrast, /u/-/ʊ/ ("UW-UH" in Figure 2), hardly increases at all when the original contrast is made less confusable. Even if this phonemic contrast were conveyed with perfect clarity, it would not drastically change how accurately words are recognized. This suggests that the functional load of the /u/-/ʊ/ contrast is more a product of its distribution in the lexicon than its perceptual confusability.

Finally, type- and token-based accounts also vary in how sensitive the functional load of their contrasts are to their *a priori* perceptual confusability. For example, the /æ/-/ɛ/ contrast ("AE-EH" in Figure 2), has high *a priori* perceptual confusability. The token-based functional load of this contrast increases only slightly when the contrast is made perfectly perceptually distinct, demonstrating a relative robustness to initial perceptual confusability. However, the type-based functional load of the same contrast shows the *most* sensitivity to initial perceptual confusability. The fact that

the weighting of lexical frequency can modulate how perceptual contributions affect our model's functional load suggests that in addition to capturing the impact of perceptual and lexical factors impact individually, our model also captures their interactions.

5 Evaluating our new definition

Although our definition of functional load has been built on a model of spoken word recognition that is known to work well in predicting human data (Luce and Pisoni, 1998), the question has been left open of whether this operationalization is actually a good predictor of sound change. If the previously observed effects of functional load on sound change are indeed relevant to a contrast's functional contribution to spoken word recognition, then we expect that models such as ours should be successful in predicting mergers.

In order to evaluate performance, we compare the functional load of actual attested mergers with randomly chosen phonological mergers in similar environments. The logic here is that the if high functional load prevents mergers[1], we would expect to see that attested mergers have significantly less functional load than random hypothetical mergers.

5.1 Methods

To get the most accurate sense of a merger's functional load, we need to calculate it in the environment of the language in which it is taking place. Because our model is based on American English, we limit ourselves to mergers that are currently taking place in varieties of American English. The attested vowel mergers here represent the four most attested vowel mergers in North America (Labov et al., 2006)[2]: the *caught-cot* merger (also known as the "low back merger") is a environment-independent (in this model) merger of /ɑ/ and /ɔ/. The *fool-full* merger merges /u/ and /ʊ/ before /l/, the *pin-pen* merger merges /ɪ/-/ɛ/ before nasals, and the *still-steel* merger merges /ɪ/-/i/ before /l/. Consonant mergers in English are much

more rare, and the consonant mergers here represent the two most straightforward attested mergers in varieties of American English we could find. The "think-fink" merger (*th-fronting*) merges dental fricatives and labiodental fricatives, and the "thin-tin" merger (*th-stopping*) merges dental fricatives and dental plosives.

To compare the functional load of attested mergers with random mergers, we compare each attested merger with 20 randomly chosen phonological mergers that take place in the same phonetic environments as the attested mergers. For example, the *pin-pen merger* merges /ɪ/-/ɛ/ before nasals; to compare, we chose 20 random vowel contrasts and merged them before nasals as well. The contrasts in the hypothetical mergers were chosen by randomly sampling all possible pairings of phonemes, the only constraint being that only vowels were considered for comparison to attested vowel mergers and only consononants for the attested consonant mergers[3].

Using a kernel density estimate calculated from the distribution of random mergers, we calculate the probability that a sample drawn from the distribution of theoretically possible similar mergers would have a functional load less than or equal to that of the attested merger. Attested mergers should have lower functional load than theoretical mergers, as they have actually happened in various dialects of American English. Figure (3) compares our models to a traditional measure, the number of minimal pairs erased by a contrast.

5.2 Results

As shown in Figure (3), both of our models outperform the traditional measure for at least four of the six mergers. The type-based model predicts each attested merger relatively well: with the exception of the *still-steel* merger (which no measure was able to predict), the type-based model predicts each attested merger with at least $p < 0.3$. Although the token-based model performs relatively well for the rest of the six, *TH-stopping* and *TH-fronting* have higher-than-average token-based functional loads compared to the random mergers. The only merger the number of minimal pairs predicts better than both of our mod-

---

[1] Low functional load *driving* mergers and high functional load *preventing* mergers are two separate hypotheses. In practice, they are hard to distinguish, and doing so is not the purpose of the current paper. Rather, we seek to assess whether our functional load measure is a good predictor of mergers compared to the standard notion.

[2] Labov et al. (2006) also mention that there is preliminary evidence for five more possible mergers in North American English before /l/, but note that these require further study.

---

[3] Limiting the comparison to, say, only mergers that are attested in other languages, would have introduced elements of cirularity into the logic of our test. It is possible that contrasts have merged *because* they have low functional load to begin with.

els is for *TH-fronting*, and its performance over the type-based model is relatively small (p<0.06 vs. p<0.11). Both of our models predict the attested mergers better numerically than the traditional measure for all four of the vowel mergers, though for the *still-steel* merger the type-based model and the traditional measure are practically identical (p<0.76 vs. p<0.78).

The difference is less clear when comparing the type- and token-based models. Numerically, the token-based model outperforms the type-based model on the *caught-cot*, *pin-pen*, and *still-steel* mergers, but for these mergers, the type-based model still does relatively well (p<0.15 for the *caught-cot* merger and p<0.24 for the *pin-pen* merger) or both models do poorly (for the *still-steel* merger, token: p<0.46 vs type: p<0.78). For the two mergers where the type-based model predicts the attested mergers better than the token-based model, it makes relatively accurate predictions while the token-based model does not predict the merger (for *TH-fronting*: p<0.11 vs. p<0.70, and for *TH-stopping*: p<0.00 vs. p<0.50). This could suggest that the type-based account of functional load better captures the pressures of sound change, but further testing is be required to make a more definitive comparison.

6 Discussion

As more studies highlight the relationships between the lexicon, perceptual information, and language change (Ohala, 1993b; Ohala, 1993a; Hall, 2009; Hall et al., submitted; Kang and Cohen, 2016), some have found that incorporating perceptual data into accounts of functional load can succeed where traditional measures cannot (Tsui, 2012). In this paper, we have incorporated key insights from speech perception with ideal observer models of word recognition to form a revised definition of functional load. Implementing this definition with a simple model, we find that it outperforms the traditional measure of functional load, and gives us new insight into the process of language change.

For example, the type-based model's better confidence in predicting attested mergers compared to the token-based model tentatively suggests that misrecognizing meaning *types* exerts more pressure on sound change than the average number of times one misrecognizes words (i.e. according to the frequency with which one encounters them).

These results are in line with Wedel et al. (2013), insofar both argue against the claim from Martinet (1952) that functional load should be determined by weighing words' functional contribution with their frequencies.

6.1 Future directions

We close by briefly discussing the simplifying assumptions of the present model, and how they can be addressed in future work. In the current study, perceptual confusability between two phonemes is calculated from a phoneme-to-phoneme confusion matrix, which simplifies phoneme recognition considerably. Unlike the smaller data set in (Luce and Pisoni, 1998), the data from (Cutler et al., 2004) was collected through a forced-choice task, with participants hearing only VC and CV syllables, and being only able to respond in kind. Although less constrained studies could offer more sophisticated behavioral data (e.g. including consonant clusters, etc.), experimentally collecting so much data can quickly become onerous.

The benefits of using acoustic/perceptual information also go beyond what has been covered in the current paper. For instance, it is worth noting that in all previous conceptions of functional load, phonological mergers have always been viewed as detrimental to recognition or to separating meanings, etc. However, when considering how sound change could affect actual spoken message transmission, we realize that mergers could in theory *benefit* communication. Imagine for example, a phone that is important to communication–but easily confused with another important phone–merges with a relatively unimportant phone. If the process of merging moves the phone in question acoustically (and perceptually) away from its confusable competitor, then the gain in recognition could in theory outweigh the loss from merging. In short, not all mergers of a phonemic contrast are equal: the direction in which a contrast merges should impact the system differently. Our model can thus in principle be used to make predictions of about the specific acoustic outcome of merges, and these predictions can be tested against data. For example, in the *pin-pen* merger, the result of the merger is generally closer to /ɪ/ than /ɛ/. Does this direction have a lower functional load than a merger in the opposite direction?

Additionally, ideal observer models predict word recognition based on contextual

predictability—although we ignore contextual predictability in this paper, it would be relatively straight forward to incorporate it in future versions of the model, for example, via measures of average predictability. Additionally, the model in the current paper analyzes *isolated* spoken word recognition. In the future, we could estimate word recognition in the context of sentences by incorporating *n*-gram frequencies from conversational corpora into the model.

7 Acknowledgments

References

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A. critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–990.

Anne Cutler, Andrea Weber, Roel Smits, and Nicole Cooper. 2004. Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, 116(6):3668–3678.

Jules Gilliéron. 1918. *G.énéalogie des mots qui désignent l'abeille d'après l'Atlas linguistique de la France*, volume 225. É.. Champion, Paris.

D. M. Green and J. A. Swets. 1966. *Signal detection theory and psychophysics*. Robert E. Krieger Publishing Co., Huntington, NY.

Kathleen Currie Hall, Elizabeth Hume, T. Florian Jaeger, and Andrew Wedel. submitted. The message shapes phonology.

Kathleen Currie Hall. 2009. *A probabilistic model of phonological relationships from contrast to allophony*. Ph.D. thesis, The Ohio State University.

Charles Francis Hockett. 1955. *A manual of phonology*. Waverly Press.

Charles Francis Hockett. 1967. The quantification of functional load. *Word*, 23(1-3):300–320.

Roman Jakobson. 1931. Prinzipien der historischen phonologie. *Travaux du cercle linguistique de Prague*, 4:246–267.

Shinae Kang and Clara Cohen. 2016. Relationships between functional load and auditory confusability under different speech environments. In *Interspeech 2016*. International Speech Communication Association.

William Labov, Sharon Ash, and Charles Boberg. 2006. *Atlas of North American English: Phonology and phonetics*. Mouton de Gruyter, Berlin.

Roger Levy. 2008. A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the conference on empirical methods in natural language processing*, pages 234–243. Association for Computational Linguistics.

Paul A. Luce and David B. Pisoni. 1998. Recognizing spoken words. *Ear and Hearing*, 19(1):1–36.

R. Duncan Luce. 1959. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley and Sons.

V. Malthesius. 1931. Zum Problem der Belastungs- und Kombinationsfähigkeit der Phoneme. *Travaux du cercle linguistique de Prague*, 4:148–152.

Andre Martinet. 1952. Function, structure, and sound change. *Word*, 8:1–32.

George A. Miller and Patricia E. Nicely. 1955. An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Dennis Norris and James M. McQueen. 2008. Shortlist B.: a Bayesian model of continuous speech recognition. *Psychological review*, 115(2):357.

John J. Ohala. 1993a. The phonetics of sound change. In Charles Jones, editor, *Historical linguistics: Problems and perspectives*, pages 237–278. Longman Publishing.

John J. Ohala. 1993b. Sound change as nature's speech perception experiment. *Speech Communication*, 13(1-2):155–161.

Dinoj Surendran and Partha Niyogi. 2003. Measuring the usefulness (functional load) of phonological contrasts. Technical report, Technical Report TR-2003.

Dinoj Surendran and Partha Niyogi. 2006. Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In O. N. Thomson, editor, *Competing models of linguistic change: Evolution and beyond*, pages 43–58. John Benjamins, Amsterdam.

Tsz-Him Tsui. 2012. Tonal variation in Hong Kong Cantonese acoustic distance & functional load. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 48, pages 579–588. Chicago Linguistic Society.

Marilyn D. Wang and Robert C. Bilger. 1973. Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, 54(5):1248–1266.

Andrew Wedel, Abby Kaplan, and Scott Jackson. 2013. High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2):179–186.

Robert L. Weide. 1998. The CMU pronouncing dictionary [cmudict 0.7].