ARTICLE IN PRESS

Journal of Memory and Language xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception

Susanne Gahl [a,*], Julia F. Strand [b]

[a] University of California at Berkeley, Department of Linguistics, 1203 Dwinelle Hall, Berkeley, CA 94720-2650, United States
[b] Carleton College, Department of Psychology, One North College Street, Northfield, MN 55057, United States

ARTICLE INFO

Article history:
Received 2 March 2015
revision received 9 December 2015
Available online xxxx

Keywords:
Phonological neighborhood density
Perceptual confusability
Lexical competition
Word identification
Spoken word production
Word duration

ABSTRACT

We examine the relationship of lexical representations, pronunciation variation, and word recognition, by investigating effects of two lexical variables: Phonological Neighborhood Density (the number of words that can be formed by a single phoneme substitution, addition, or deletion from the target word), as well as a measure of the perceptual similarity of a target word to other words in the lexicon. We show that perceptual similarity to other words affects recognition, but not production. Phonological Neighborhood Density, on the other hand, affects both word durations and recognition accuracy (words with many neighbors shorten and are difficult recognition targets). We interpret our results as indicating that effects of Phonological Neighborhood Density on pronunciation are not generally due to perceptual similarity of the target to other words. Our results are consistent with a more general line of research demonstrating effects of 'central' processes on 'peripheral' processes such as articulation, as well as effects of modality-specific properties, such as auditory similarity and motor movements, on measures thought to tap central processes.

Introduction

A central theme in the psychology of language concerns the role of perceptual, articulatory, and modality-neutral representations and processes in language processing. For example, verbal working memory has been argued to have both articulatory and auditory components (Gupta & MacWhinney, 1995; Wilson, 2001) and/or components not tied to any sensory modality (Jones, Beaman & Macken, 1996). Another example concerns the nature of percepts in speech perception, which have variously been characterized as auditory (Diehl & Kluender, 1989; Diehl, Lotto, & Holt, 2004) or as motor gestures (Galantucci, Fowler, & Turvey, 2006; Liberman & Mattingly, 1985). A third example addresses the question of what makes word forms similar to one another. Word form similarity affects many phenomena in speech production and perception, including phonological priming, speech errors, immediate verbal recall, and spoken word recognition difficulty (Conrad & Hull, 1964; Goldrick, Baker, Murphy, & Baese-Berk, 2011; Goldrick, Folk, & Rapp, 2010; Levelt et al., 1991; Page, Madge, Cumming, & Norris, 2007). What is less clear is the degree to which, or the circumstances under which, word form similarity is based on auditory or articulatory properties of words, or on some form of abstract representation shared across modalities.

A key concept in studies of effects of word form similarity is phonological neighborhood density. Informally, phonological neighborhood density is often defined as "the number of words that sound similar to a given word" (e.g. Vitevitch, 2007, p. 166). A common method for determining that number is to count the number of words ("neighbors") in a reference lexicon that differ from a target word through addition, deletion, or substitution of

* Corresponding author.
E-mail addresses: gahl@berkeley.edu (S. Gahl), jstrand@carleton.edu (J.F. Strand).

http://dx.doi.org/10.1016/j.jml.2015.12.006
0749-596X/

exactly one phonological segment, regardless of the degree of auditory similarity of target and neighbor. By that criterion, the neighbors of *cat* include *caught, pat,* and *can.* Convenience is doubtless one reason why these two characterizations of phonological neighborhood density – in perceptual/auditory terms and in terms of phonological segments – are each common, for conveying informal working definitions ("neighbors sound similar") and straightforwardly estimating phonological neighborhood density in electronically searchable dictionaries ("count anything as a neighbor that differs by one segment"). Convenience aside, these definitions and methods reflect the potential role of various sensorimotor properties vs. modality-neutral segmental representations in the phenomena that the literature on neighborhood density has brought to light.

Different definitions invite different inferences about the sources of effects of word form similarity generally, and phonological neighborhood density in particular. The definition of phonological neighbors as similar-sounding words implies that effects of phonological neighborhood density are due at least in part to auditory similarity; if that is so, then the degree to which words sound similar to one another should affect the strength of phonological neighborhood density effects. On the other hand, the one-segment-difference metric implies that words overlapping in segmental content without sounding particularly similar (e.g. *leap* and *lope*) should act as neighbors of one another. Conversely, words that do sound similar to one another, but do not form neighbors by the one-segment-difference metric (e.g. *this* and *fish*), are not expected to act as phonological neighbors under the segmental-overlap conception of word form similarity.

What makes word forms similar? Common sense suggests that the answer must depend on the nature of the task, such as identifying words over noise vs. producing tongue twisters. It follows that a measure of phonological neighborhoods that successfully predicts neighborhood effects in one domain needn't be relevant to some other domain. The fact that phonological neighborhood density as estimated by the one-segment difference criterion is predictive of a wide range of tasks is an empirical discovery that has had a major impact on models of speech production and comprehension. However, observations consistent with those models do not constitute proof of a causal role of one-segment-difference neighbors in the phenomena being modeled: Words that share segments in common require some of the same articulatory movements and often do sound similar to one another. Therefore, one and the same effect may be consistent with models based on 'amodal' segment overlap, articulatory, or perceptual similarity. As Vitevitch and Luce (2016) point out: "[M]etrics for computing similarity neighborhoods are not the same as theoretical statements about the proposed effects of similarity neighborhood activation on recognition." Yet, the predictiveness of phonological neighborhood density metrics is sometimes taken as the basis for inferred mechanisms. For example, effects of phonological neighborhood density on phonetic detail in speech production, which we discuss below, are sometimes explained in terms of the auditory similarity of words (e.g. Lindblom, 1990; Wright, 2004; Scarborough, 2004). At other times (including in our own previous work, e.g. Gahl, Yao, & Johnson, 2012), effects of phonological neighborhood density on speech production are linked to (amodal) segmental and articulatory similarity of words, without any assessment of auditory similarity.

The question of what makes words similar also figures in information-based models of spoken language processing (e.g. Aylett, 2000; Hale, 2003; Jaeger & Levy, 2006; Levy, 2008). The central idea in these models is that language processing is optimized for efficient communication, and that communication is most efficient when information is conveyed at a constant rate. The more an item (a word or sound) reduces the uncertainty about the message to be conveyed at any given point in an utterance, the heavier its informational load. A word that is highly similar to other words may generate high uncertainty. Segments of words in dense phonological neighborhoods may considerably narrow down the set of words matching the signal and hence reduce uncertainty. Applied to speech production, the idea is that speakers spend more time and/or articulatory effort, and/or produce signals that increase the probability of recognition for highly informative items than on less informative items (Fosler-Lussier & Morgan, 1999; Aylett, 2000; Jurafsky, 2001; Aylett & Turk, 2004, 2006; van Son & Pols, 2003; Pluymaekers, Ernestus & Baayen, 2005; Bell, Brenier, Girand & Jurafsky, 2009; Buz & Jaeger, 2015; Kuperman, Pluymaekers, Ernestus, & Baayen, 2007; Pate & Goldwater, 2015; Seyfarth, 2014; Tily & Kuperman, 2012). The question of what makes words similar can be stated as the question of how estimates of information load are to be fleshed out. In practice, information-based models of speech production and recognition have calculated the information carried by sounds and words in different ways. These differences depend in part on the unit being modeled, such as phonological segments (van Son & Pols, 2003), syllables (Aylett & Turk, 2006), or words (Buz & Jaeger, 2015), and on estimates of contextual predictability. In part, though, they depend on assumptions about what it takes to disambiguate a signal: Like metrics of phonological neighborhood density, some estimates of information density take auditory similarity into account, while others do not. Which estimates best predict variation in the signal or in recognition accuracy, and whether articulatory effort, intelligibility, and recognition probability all vary as a function of one measure of information load is an empirical question. One step towards answering that question was taken in Buz and Jaeger (2013), who described three different ways of estimating neighborhood density: The (log) count of the neighbors, the summed log frequency of the neighbors, and a measure taking into account both the segment-to-segment confusability of target and neighbors along with the frequency of neighbors (termed the frequency-weighted neighborhood probability in Luce & Pisoni, 1998, see below, but computed using a different confusability matrix). Buz and Jaeger (2013) found that the three measures behaved similarly as predictors of word duration in a picture naming experiment. (We return to the results of the picture naming experiment, which is described in greater detail in Buz & Jaeger, 2015, in the General Discussion.) Buz and Jaeger (2013 and

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

2015) did not attempt to assess whether the neighborhood density measures were equally good predictors of recognition accuracy of the recorded naming responses. More generally, the ability of different measures of phonological neighborhood density to predict recognition vs. production has typically been assessed in separate studies, using different sets of words. Tracking the effects of all aspects of word similarity and contextual predictability affecting production and recognition, in identical contexts, remains an unmet goal.

The present study takes a step in that direction. We track effects of variables indexing word-form similarity with and without taking perceptual similarity into account in two data sets representing auditory word recognition accuracy (the 'recognition' set) and word duration in a corpus of conversational speech (the 'production' set, taken from the Buckeye corpus; Pitt et al., 2007). We hypothesized that perceptual and segmental metrics of phonological similarity have different and, to some degree, separable effects in recognition accuracy vs. conversational speech, and that these differences would reveal themselves as asymmetries in the predictiveness of a perceptually-based vs. a segmentally-based estimate of phonological neighborhood density.

Several cautionary notes about our analyses of these two very different data sets are in order. Pronunciation and intelligibility of stimuli in the recognition task vs. the conversational speech differ enormously (see e.g. Johnson, 2004; Keune, Ernestus, Hout, & Baayen, 2005). Conversational speech is produced far more rapidly and contains far more contextual clues than laboratory speech. The differences in the time available for utterance planning and articulation undoubtedly affect what kinds of lexical information can be reflected in conversational speech vs. the recorded stimuli for the recognition task. Indeed, Gahl et al. (2012) speculated that some of the differences between effects of phonological neighborhood density in single-word production tasks vs. conversational speech may be due to the different temporal demands of the tasks. We return to these caveats and what we see as fruitful directions for future research in the Discussion.

Background: phonological neighborhood density in recognition and production

The core empirical fact that earned phonological neighborhood density its place among lexical variables of interest to psycholinguists is that words in dense phonological neighborhoods (those with many neighbors) are more difficult to recognize, other things being equal, than words in sparse neighborhoods (Goldinger, Luce, & Pisoni, 1989; Luce & Pisoni, 1998; Luce, Pisoni, & Goldinger, 1990; Vitevitch & Luce, 1998). That observation provided the empirical foundation of a highly influential model of spoken word recognition, the Neighborhood Activation Model (NAM, Luce & Pisoni, 1998) and for a large body of research on recognition and production (see e.g. Chen & Mirman, 2012, and Vitevitch & Luce, 2016 for overviews and discussion).

The observation that words in dense neighborhoods tend to be difficult to recognize fits a widely-held intuition: that neighborhood density effects arise because phonological neighbors tend to sound similar to one another. The more neighbors a word has, the more words it resembles, increasing the difficulty of the categorization task faced by the listener. However, from the start, measures of phonological neighborhood density were calculated in different ways, not all of which made reference to what words sounded like. The original instantiation of the NAM (Luce & Pisoni, 1998) employed the Frequency-Weighted Neighborhood Probability Rule (FWNP), which uses forced-choice phoneme confusions in noise to quantify the auditory confusability of a target word with all other words in the lexicon (Luce & Pisoni, 1998, Experiment 1). That is to say, the method of estimating the lexical "competition" faced by a target word took into account the fact that some segments are more confusable than others, as well as the fact that some words are more frequent than others. Furthermore, the FWNP took into account a target word's segment-by-segment confusability with all other words in the lexicon, not just those that differed from the target in one segment. However, when predicting auditory lexical decision and naming latencies in the absence of background noise, Luce and Pisoni (1998, Experiments 2 and 3) used the number of one-phoneme-difference neighbors as an estimate of phonological neighborhood density, rather than the segment confusion matrices (which rely on segments presented in background noise). By that criterion, all words differing from a target word by exactly one segment are target neighbors – and words that differ in more than one segment from the target are not. Numerous subsequent related studies, both in the recognition literature and the production literature, also use that shortcut estimate of phonological neighborhood density, sometimes weighted by lexical frequency, to quantify lexical competition (Cluff & Luce, 1990; Dell & Gordon, 2003; Gordon, 2014; Munson & Solomon, 2004; Scarborough, 2010; Scarborough, 2013; Vitevitch & Luce, 1998).

Alongside evidence for an inhibitory effect of phonological neighborhood density on spoken word recognition, there is evidence for a facilitative effect of high neighborhood density on spoken word production (Harley and Bown, 1998; Gordon, 2002; Marian & Blumenfeld, 2006; Peramunage, Blumstein, Myers, Goldrick, & Baese-Berk, 2010; Vitevitch, 1997, 2002; Vitevitch & Sommers, 2003). Dell and Gordon (2003) model this pattern of high PND facilitating word production, but inhibiting word recognition, as resulting from interactive feedback between lexical and segmental levels, consistent with Dell's interactive two-step model of lexical access and retrieval (Dell, 1986; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997): In production, feedback from phonological neighbors boosts target word activation, which is already high, due to the initial jolt of activation from the semantic level. Recognition, on the other hand, begins with activation of phonological segments, boosting the activation of target words, but also of other words containing those same phonological segments – leading to a net loss in target activation that is especially perilous when a target has many phonological neighbors. Chen and Mirman (2012) argue, on the basis of simulations in a domain-general interaction and competition model, that the reported pattern of

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

facilitation and inhibition is predicted in any model in which multiple representations are activated in parallel: Weak competition (such as that posed by phonological neighbors in spoken word production) yields a net benefit for the target, whereas strong competition (such as that posed by phonological neighbors in spoken word recognition) results in target inhibition. More recently, Chen and Mirman (2015) have shown that phonological neighbors produce a facilitative effect even in spoken word recognition when semantic context weakens the activation of phonological neighbors. Importantly for the current discussion, these explanations make no reference to auditory or articulatory similarity; the amount of competition among jointly activated phonological segments does not depend on how similar the segments are.

Complicating matters is the fact that phonological neighborhood density is highly correlated with several other variables, including lexical frequency (Frauenfelder, Baayen, Hellwig, & Schreuder, 1993), various measures of phonotactic probability (such as the probability of a given segment appearing in a given position, possibly conditioned on the segment preceding or following it; Vitevitch, Armbrüster, & Chu, 2004; Vitevitch & Luce, 2005), onset density (i.e. the proportion of neighbors with the same initial segment as a target word), and the 'spread' of the neighborhood (i.e. the number of segmental positions at which neighbors can be formed; Vitevitch, 2007). Each of these variables affects speech production, but not all of them do so in the same direction as phonological neighborhood density. While high lexical frequency and phonotactic probability are associated with shorter naming latencies, high onset density has been found to elicit longer naming latencies when phonological neighborhood density is controlled for (Vitevitch et al., 2004). Several related models similarly predict patterns that take into account the left-to-right nature of word recognition (Allopenna, Magnuson, & Tanenhaus, 1998; Magnuson, Dixon, Tanenhaus, & Aslin, 2007) and production (Sevald & Dell, 1994). For example, Sevald and Dell (1994) argue that phonological selection is a serial process: Following lexical selection and during phonological encoding, target segments are accessed in the order in which they are to be articulated. Sevald and Dell (1994) found that speakers were able to repeat pairs of words more quickly when the words differed in their initial consonants (e.g. PICK-TICK) than when the difference was in the final consonants (e.g. PICK-PIN). Sevald and Dell (1994) interpret that observation as an effect of shared segments producing lexical competition: Words that share initial segments act as strong competitors of one another. That interpretation is consistent with Chen and Mirman (2015)'s simulations: While high neighborhood density generally yields target facilitation in word production, that facilitation gives way to inhibition at a point when the only remaining competitors are strong. That is, at the point when the initial two segments have been selected, competition between *pick* and *pin* is strong. In summary, the presence of multiple correlated variables (such as neighbors overlapping in onsets in a specific stimulus set or in the lexicon generally, as well as phonotactic probability and the positions in a word at which neighbors can be formed among) considerably complicate the interpretation of observed effects.

Evidence for a facilitative effect of high phonological neighborhood density on spoken word production is still relatively sparse, compared to the copious literature on its inhibitory effects on spoken word recognition, and the idea remains somewhat controversial. More research is needed on individual effects of variables that are correlated with phonological neighborhood density. For example, Sadat, Martin, and Costa (2014) argue that high phonological neighborhood density is associated with longer, not shorter latencies in picture naming. However, Sadat et al. report that the correlation between PND and 'onset density', i.e. the number of words that share the initial segments with the target, was .97 in their data set of Spanish nouns. Sadat et al.'s observation of longer latencies with increasing phonological neighborhood density may therefore be due to onset density, shown in previous studies to yield an inhibitory effect on spoken word production (cf. Vitevitch et al., 2004).

Effects of phonological neighborhood density on pronunciation variation

The inhibitory effects of phonological neighborhood density on word recognition inspired a line of inquiry in studies of pronunciation variation, exploring the possibility that talkers might pronounce words in dense neighborhoods more clearly than words in sparse neighborhoods, to compensate for the recognition difficulty (Munson & Solomon, 2004; R. Wright, 2004). Consistent with this possibility, a number of studies reported that words in dense phonological neighborhoods are hyperarticulated and/or phonetically enhanced compared to words in sparse phonological neighborhoods, as evidenced by longer VOTs (Baese-Berk & Goldrick, 2009; Fox, Reilly, & Blumstein, 2015; Goldrick, Vaughn, & Murphy, 2013), increased nasal coarticulation (Scarborough, 2004, 2010, 2013), or increased vowel dispersion (Munson, 2007; Munson & Solomon, 2004; Wright, 2004; but see Flemming, 2010, and Gahl, 2015, for critiques and reanalyses of several of these studies).

However, not all studies of phonological neighborhood density effects on pronunciation report high density to be associated with hyperarticulation or lengthening. Gahl et al. (2012) examined the effects of phonological neighborhood density on word durations and on vowel dispersion (Euclidean distance from a talker's average first and second vowel formants) in the Buckeye corpus of spontaneous speech (Pitt et al., 2007). Gahl et al. (2012) found that CVC (consonant-vowel-consonant) words tended to be shorter with increasing phonological neighborhood density, when other factors affecting word duration were controlled in a mixed-effects regression model. In addition, vowels in high-density words tended to be more centralized in F1/F2 space, i.e. more schwa-like, than vowels in low-density words. Since shortening and vowel centralization are also often observed in high-frequency words, which are retrieved more quickly than low-frequency words, Gahl et al. (2012) interpreted these findings as part of a broader pattern of phonetic reduction of words whose

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

retrieval is facilitated at early stages of language production: In other words, reduction of words with high phonological neighborhood density is attributed by these authors to be consistent with, and a consequence of, the facilitation in lexical retrieval consistent with Chen & Mirman and Dell & Gordon's models.

Several variables correlated with phonological neighborhood density have been explored in research linking pronunciation variation to lexical retrieval. Yiu and Watson (2015) recently demonstrated that initial overlap of words was associated with a greater degree of lengthening of word durations compared to final overlap. Yiu and Watson (2015) interpret that observation to result from words with shared overlap (PICK-PIN) being strong competitors of one another, as proposed in Sevald and Dell (1994). The idea is that the phonological planning process is slowed down while that competition is resolved.

As mentioned earlier, high phonological neighborhood density has been found to be associated with phonetic enhancement in a number of studies of voice onset times (VOT). (Baese-Berk & Goldrick, 2009; Fox et al., 2015; Goldrick et al., 2013 found longer VOTs in high-density vs. low-density targets. The interpretation of those findings is complicated by the presence of other correlated variables. Fricke, Baese-Berk & Goldrick (in press) evaluated the relationship of minimal pair status (i.e. whether a stop-initial target word had a neighbor differing only in voicing of the initial stop, e.g. *pig/big* vs. *peel/*beel*), phonological neighborhood density, and position-specific phonological neighborhood density, i.e., the number of neighbors that can be formed by changes at each position, on voice onset times (VOT) in initial stop consonants. Although both minimal pair status and phonological neighborhood density affected VOT when entered individually in a model of VOTs, neither accounted for significant variance when added to a model that included position-specific phonological neighborhood density. This raises the possibility that other reported effects of phonological neighborhood density on VOTs may likewise be due to position-specific measures, rather than phonological neighborhood density.

The search for lexical factors in pronunciation constitutes a departure from a research tradition in which details of pronunciation were either considered to be a matter of late stages of the language production processes in serial psycholinguistic models (such as the phonetic encoding stage, Levelt & Wheeldon, 1994) or considered to be outside of the scope of psycholinguistic models altogether (see Hickok, 2012, for an overview). (An early exception to that strategy is Balota, Boland, & Shields, 1989, who observed an effect of semantic priming on word durations). The lion's share of research on pronunciation variation has focused on effects of syllable frequency, n-gram probability (of segments and/or words), and phonotactic probability (see e.g. Jurafsky, 2003, for an overview). Such effects are well established, and there can be no doubt that word duration is in part due to factors affecting late stages of articulatory planning and motor execution, perhaps due to the availability of pre-compiled motor plans for frequently-produced syllables (cf. Cholin, Levelt, & Schiller, 2006; Levelt, Roelofs, & Meyer, 1999; Levelt & Wheeldon, 1994).

Since syllable frequency and lexical frequency are highly correlated, particularly in the case of monosyllabic words, many effects of lexical frequency on pronunciation can in principle be explained as effects of articulatory routinization (though not all; see Gahl, 2008, for discussion). Since the words in our datasets are monosyllables, it is questionable whether effects of syllable frequency and lexical frequency can be disentangled in our data. We see no reason to doubt the effects of phonotactic probability, syllable frequency, or lexical frequency on word durations. As explained below, we included syllable frequency in our models of word duration; given the high correlation with lexical frequency, we did not attempt to disentangle effects of syllable frequency from effects of word frequency.

Limitations of phonological neighborhood density as a measure of lexical confusability

As successful as phonological neighborhood density has proven to be in studies of word recognition, it has its limitations. The first is that the most commonly-used neighborhood density metrics categorically divide words into target neighbors vs. "non-neighbors." That categorical division can have unexpected and undesirable consequences, depending on the research question at hand: Some words within a set of neighbors might be expected to be more perceptually similar to the target word than others – and words outside a target's neighborhood may be more perceptually similar to the target than some target neighbors. For example, both *seen* and *shun* are neighbors of *sun*, but *shun* may be expected to be more highly confusable with *sun* because [ʃ] and[s] are perceptually similar, whereas *seen* is likely to be less confusable with *sun* because [i] and [ʌ] are less similar. Another potential limitation is that words that differ from the target by more than one phoneme are not included in measures of lexical density. For instance, *fish* and *this* are not neighbors by a one-phoneme difference criterion, but it would be reasonable to expect some confusability between the two words. Indeed, some words that differ by multiple phonemes (*fish* and *this*) may be more confusable than words that differ by only one (*seen* and *sun*). In part, these are limitations of the one-phoneme-difference shortcut measure: The FWNP (Luce & Pisoni, 1998, Experiment 1) assigns weights to target competitors based on auditory confusability of segments and takes into account all words in the lexicon, not just those words that differ from the target in one phoneme.

Neighborhood density metrics using confusion probabilities, such as the FWNP, have another limitation, owing to the fact that these measures fail to take into account the number of perceptually similar alternatives for each target segment (Iverson, Bernstein, & Auer, 1998). For example, a confusion matrix may reveal that [z,ð] are perceptually similar to one another, and [f,s,ʃ,θ] are also perceptually similar to one another. If [z,ð]) are confused on 50% of trials, and [f,s] are confused on 25% of trials, then based on p (z|ð) vs. p(f|s), [z] and [ð] will be judged to be more "similar" than [f] and [s] (Iverson et al., 1998). When these values are weighted by word frequency and used to compute FWNP, individual target-competitor comparisons will be

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

distorted by the number of response alternatives and no longer be based solely on the frequency-weighted perceptual similarity of the two words.

To correct this problem, Iverson et al. (1998) introduced Phi-square, which can be used to quantify segment similarity while taking into account the number of perceptually similar alternatives. The Phi-square statistic is quantified as follows:

$$\Phi^2 = 1 - \sqrt{\frac{\sum \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum \frac{(y_i - E(y_i))^2}{E(y_i)}}{N}}$$

where x_i and y_i are the frequencies that phonemes x and y were identified as phoneme i in a forced choice identification task, $E(x_i)$ and $E(y_i)$ are the expected frequencies of x_i and y_i if x and y were perceptually identical, and N is the total number of responses to x_i and y_i. If x and y are perceptually identical, they should be expected to be identified as members of a phoneme category equally often. Therefore, the expected frequencies, $E(x_i)$ and $E(y_i)$, are the average of the frequencies with which phonemes x and y were each identified as category i, because hypothetically, if [z] and [ð] were perceptually identical, participants should choose evenly between them when making a phoneme identification. Confusion probabilities quantify how regularly two phonemes are confused for one another (i.e., a single cell within a confusion matrix); the Phi-square value quantifies how similar the pattern of responses to the two phonemes are (i.e., comparing two rows in a confusion matrix). Using the entire distribution of responses for two phonemes negates the problem that the number of likely alternatives interacts with response probabilities. Phi-square values thus provides a measure of perceptually-based target competition that avoids the undesirable conclusion that highly ambiguous phones are less confusable than less ambiguous ones.

A second strength of using Phi-square values rather than confusion probabilities (as FWNP does) is that it reduces the influence of response biases that are present in forced-choice phoneme confusion tasks. If a participant disproportionately chooses a phoneme response for reasons that are not related to the task (e.g., always guessing [g] when unsure), it generates artifacts in the probability data. Phi-square values avoid these artifacts by evaluating the similarity of two phoneme response distributions, rather than simply evaluating the likelihood that two phonemes will be confused (see Iverson et al., 1998 and Strand, 2014).

Once the similarity of phoneme pairs has been established, word-level similarities may be calculated using the position-specific Phi-square values for a target and competitor. For example, the predicted confusability of "cat" and "cup" is quantified as $\Phi^2(k|k) * \Phi^2(æ|ʌ) * \Phi^2(t|p)$. Following the method of calculating FWNPs (Luce & Pisoni, 1998), Strand and Sommers (2011) calculated the perceptual similarity of each target word with every other word in a reference lexicon, and summed these values to obtain a measure of density called "Phi-square density." Critically, similarity values can be calculated between word pairs that differ by multiple phonemes (e.g., *fish* and *this*, to return to the example above), thereby removing the distinction between "neighbors" and "not neighbors"; in Phi-square density (as in Luce & Pisoni's FWNP), all words are allowed to compete. As a consequence, a word can have a phonological neighborhood density (PND) of 0, but still have high Phi-square density, if its segments are confusable with other segment combinations that also form words in the lexicon.

Phi-square density predicts additional variance in spoken word recognition beyond that accounted for by PND or by continuous measures of lexical competition based on confusion probabilities (Strand, 2014; Strand & Sommers, 2011). These studies suggest that the success of PND at predicting spoken word recognition accuracy may be due in part to the fact that it is correlated with and approximates measures of auditory confusability. However, Strand and Sommers (2011) did not also evaluate the influence of other measures that correlate with PND, such as syllable frequency. This leaves open the possibility that the improvement in predicting variance in spoken word recognition was not actually due to Phi-square density, but to other syllable-level, lexical, or segmental properties.

Aims and predictions

The goal of the present study is to model the effects of a variables targeting auditory vs. segment-based phonological neighborhood density on word recognition and production. We do so by modeling variation in word duration and recognition accuracy in two datasets that have figured in discussion of effects of phonological neighborhood density, but that so far been analyzed only from the perspective of perception (Slote, Strand, & a new timing method. Behavior Research Methods, in press) or production (Gahl et al., 2012), but not both.

Methods

Data sets

Spoken word recognition data were obtained from an existing dataset (Slote & Strand, in press). These data included word recognition in noise scores of 400 consonant-vowel-consonant (CVC) words by 53 college-aged listeners with normal hearing. Six of those words were excluded from the present analysis because they did not appear in the SUBTLEXUS database (Brysbaert & New, 2009). Excluding trials on which participants failed to respond left 19,860 observations. Words were presented in a background of six-talker babble at a signal to noise ratio of 0 at approximately 65 dB. Each word was presented individually with no carrier phrase and participants were instructed to type what they heard.

Word duration data were obtained from the Buckeye Corpus of conversational speech (Pitt et al., 2007; Pitt, Johnson, Hume, Kiesling, & Raymond, 2005), which consists of one hour of spontaneous speech from each of 40 talkers (20 male, 20 female; 20 under 40 years of age, 20 over 40 years of age) from Columbus, Ohio. Target words were all monomorphemic CVC content words in the

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

corpus, with the following exclusion criteria: (1) Words which did not appear in the lexical databases used for estimating syllable frequency, PND, or auditory confusability,; (2) Word forms that are frequently used as function words or as discourse markers, such as *right* or *like*; (3) Orthographic forms with multiple phonemic representations, such as *read* and *lead;* (4) Utterance-initial and utterance-final word tokens, as well as word tokens immediately following or immediately preceding filled pauses such as *um* and *uh*; (5) Words all of whose tokens had bigram probabilities of 1, i.e. probabilities given the immediately preceding or following word, which often represent parts of fixed expressions and/or hapax legomena in the corpus. The final data set contained 477 word types, represented by 11,095 tokens.

Our dataset only contains monosyllables, raising the question whether the observed effects are informative about speech perception and production more generally. As an anonymous reviewer points out, the correlation between lexical frequency and PND is approximately 0.5 in the 40,000 word dictionary of the English Lexicon Project (ELP; Balota et al., 2007). In our sample, the correlation was far lower than that (r = .13 in the recognition set and .02 in the production set). The reason for the large difference is the relationship between word length and PND: The 40,000-word ELP lexicon includes multisyllabic words, whereas our sample is restricted to CVC monosyllables. Long words tend to be lower in frequency and generally have fewer neighbors than short ones – by necessity, since long words are less likely than short words to differ by exactly 1 phone. In the ELP, 1-syllable words have an average of 12.4 neighbors. For 2-syllable words, the average drops to 2.1, and 3-syllable words have only .3 neighbors on average; see also (Frauenfelder et al., 1993). Despite the large differences in neighborhood size for long vs. short words, monosyllables form an important subset of the words speakers and listeners typically encounter. For example, 81% of word tokens in the Switchboard corpus of conversational speech are monosyllables (Greenberg, 1998). We also note that, although the majority of work on lexical competition has been done on monosyllabic words, bisyllabic words show similar effects of lexical competition as monosyllabic words (Vitevitch, Stamer, & Sereno, 2008), with high density words being recognized less accurately on a word recognition task and more slowly on a lexical decision task. This suggests that, while the processing of multisyllabic words is certainly a topic awaiting much more research, the properties of monosyllabic words form a useful starting point.

Description of variables in the regression models

To assess the influence of Phi-square density and PND on the two outcome variables, we fitted models containing these variables along with other known predictors of, respectively, recognition accuracy and word durations.

Lexical frequency. Word frequency of occurrence values were obtained from the $SUBTLEX_{US}$ database (Brysbaert & New, 2009) and represent the log-transformed number of times a given word appeared per million words.

Baseline duration. Some segments are inherently longer than others. For example, tense vowels tend to be longer than lax vowels, and nasal stops tend to be longer than voiceless oral stops (Bent, Bradlow, & Smith, 2008; Crystal & House, 1988; Peterson & Lehiste, 1960; Umeda, 1977). In addition, segment durations vary with phonological context, for example word length, position within a word, or (in the case of vowels) voicing of a following consonant. To control for the 'inherent' duration of the target words, i.e. the duration they might have if factors such as lexical frequency, neighborhood density, speaking rate, and so forth, had no effect, we estimated their 'baseline' durations, as follows: We calculated the median duration of each consonant and vowel in the Buckeye target words (i.e. CVC content words in fluent speech, using the criteria for inclusion described above). For consonants, we calculated separate medians for tokens in initial vs. final position. For vowels, we calculated separate median durations for tokens preceding voiced vs. voiceless consonants. As expected, the by-segment medians differed substantially based on position and final voicing. The "baseline duration" of each target word was the summed median duration of its segments, conditioned on position (initial vs. final consonants) and final voicing. Baseline durations were log transformed.

This estimate of baseline duration differs from that in Gahl et al. (2012), who calculated the mean (not the median) duration of each segment type across the entire Buckeye corpus, i.e. including tokens before disfluencies. Means are unduly affected by outliers (particularly for duration measures, which cannot be negative). Moreover, tokens before disfluencies are often substantially longer than segments in the subcorpus of target word productions. The baseline measure used in the current work is more firmly grounded in research on phone durations, reducing the possibility that variability due to position and final voicing might yield spurious effects of PND or other lexical variables.

Bigram probability given the word preceding/following the target: The (log-transformed) probability of a word, given the immediately preceding or following word in an utterance, which is a known predictor of word durations in connected speech (Bell et al., 2003; Fosler-Lussier & Morgan, 1999). Bigram probabilities were estimated based on the entire Buckeye corpus. Word types with average bigram probabilities of 1 were excluded from further analysis.

Speech rate (before/after): The (log-transformed) speaking rate, measured as syllables per second, in the stretch of speech from the preceding utterance boundary up to the target (Speech rate before) and from the target up to the end of the utterance (Speech rate after).

Syllable frequency (type, token). Syllable frequency was estimated using the method described in Cholin et al. (2006): Syllable type frequency was estimated as the number of word types in the CELEX data base (Baayen, Piepenbrock, & van Rijn, 1993) containing a given syllable. Syllable token frequency was estimated as the summed lexical frequency (according to CELEX) of all words containing a given syllable.

Phonological Neighborhood Density (PND). We calculated the number of words in the reference lexicon (Balota et al.,

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

Table 1
Pairwise (Spearman) correlations among lexical variables in the recognition data set (n = 394).

	Lexical frequency	Syllable type frequency	Syllable token frequency	Phi-square density
Lexical frequency				
Syllable type frequency	0.54			
Syllable token frequency	0.83	0.68		
Phi-square density	0.04	0.16	0.18	
PND	0.13	0.35	0.21	0.25

2007) that could be made by a single phoneme substitution from the target word. Although the substitution-only method for calculating phonological neighborhood density is less common than the method also counting words that can be made by addition or deletion from the target word, we used the substitution-only method here so that the reference lexicon was the same for calculating PND as calculating Phi-square density (see below). The recognition task on which the confusability measures for the Phi-square density metric is based did not allow for the possibility of confusing a segment with the "null segment" (see Luce & Pisoni, 1998), as participants knew there was some segment in each position. Values from the substitution-only method are highly correlated with values from the substitution/addition/deletion method (r = .98 for all CVCs in the reference lexicon; Strand, 2014), so this change is not likely to substantially influence the results. In the discussion to follow, we will use the abbreviation PND to mean the number of words that differ from a target in exactly one segment.

Phi-square density. Phi-square density was calculated following the method described in the introduction (Strand, 2014; Strand & Sommers, 2011). The Phi-square density of a word is the sum of a quantity indexing the pairwise perceptual similarity between the target word and every other word in the reference lexicon.

Tables 1 and 2 show the pairwise correlations of the lexical variables in the sets of 394 and 469 word types, respectively. It will be observed that PND and Phi-square density are only moderately correlated (r = .25). On the other hand, lexical frequency and syllable frequency were strongly correlated (type frequency: .54; token frequency: .83). The strong correlation of lexical frequency and syllable frequency was expected in our word list of monosyllabic words.

Modeling strategy

We wished to understand the role of perceptual similarity (Phi-square density), segmental neighborhood structure (PND), and articulatory fluency (lexical and syllable frequency) in word durations and spoken word recognition. We fitted models of word duration and recognition accuracy containing variables intended to tap these three potential sources of variation in recognition difficulty and pronunciation. A total of 201 word types appeared in both data sets (10,070 tokens in the recognition task, 7044 tokens in the production set). For a side-by-side comparison of the predictors of interest on the same set of words, we fitted models of recognition accuracy and word duration to the data sets using the 201 words for which both recognition and pronunciation data were available. Each model contained the variables targeting segmental (PND), perceptual (Phi-square density) and articulatory (Feature similarity, syllable frequency) similarity to other words, along with 'baseline' predictors that are known to affect spoken word recognition accuracy and word duration. In addition, we tested whether the pattern of significant and non-significant effects observed in the set of 201 words generalized to the larger set, to reduce the possibility that patterns in smaller set were due to idiosyncrasies of the 201 words.

Statistical treatment of the data

We fitted mixed-effects regression models of recognition accuracy and word durations. Two related sets of issues in regression modeling that have received a great deal of attention among psycholinguists concern the specification of the random effects structure (Barr, Levy, Scheepers, & Tily, 2013; Gelman & Hill, 2006; Bates, Kliegl, Vasishth & Baayen, 2015) and the order of entry or removal of variables (both in the fixed effects structure and the random effects). In the models reported here, we entered all fixed effects simultaneously, as opposed to entering or removing variables in a stepwise fashion.[1]

For the random effects structure, we used forward entry of (by-target word and/or by-participant) random slopes corresponding to the variables in the fixed effects. In many cases, including random slopes resulted in problematic models, either due to zero variances or to perfect correlations among variance components, or else resulted in failure to converge. To satisfy our (and a reviewer's) curiosity about the effect forward entry vs. backward reduction in the random effects structure, we explored forward entry and backward reduction; in no case did the pattern of significant fixed effects of the critical variables change as a result of the choice of method of entry in the random effects structure. The models reported here are the models with the maximal random effects structure that appeared to be supported by the data, on the basis of the variances and correlations in the random effects. Observations with

[1] In an earlier version of this work, we fitted a "baseline" model containing non-lexical variables (e.g. local speaking rate and a baseline measure of expected word durations based on average segment duration) and compared that baseline to models containing one additional predictor at a time: E.g. the baseline was compared to the baseline plus Phi-square density, or the baseline plus syllable frequency. In other words, PND, Phi-square density, and syllable frequency did not compete with one another. The pattern of significant effects of lexical frequency, Phi-square density, and PND was the same as the results presented here.

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

Table 2
Pairwise (Spearman) correlations among the lexical variables in the production data set (n = 469).

	Baseline duration	Lexical frequency	Syllable type frequency	Syllable token frequency	CV biphone	VC biphone	PND
Baseline duration							
Lexical frequency	−0.14						
Syllable type frequency	−0.16	0.38					
Syllable token frequency	−0.21	0.74	0.57				
Initial (CV) biphone probability	−0.04	−0.05	0.19	0.01			
Final (VC) biphone probability	−0.27	0	0.24	0.07	0.32		
PND	−0.23	0.02	0.24	0.11	0.38	0.44	
Phi-square density	−0.31	0.08	0.1	0.19	0.04	0.04	0.26

large residuals (more than 2.5 SDs) were removed at each modeling step and the model refitted without those cases. Continuous variables were log-transformed where doing so resulted in more nearly normal distributions. All numerical variables were centered around their means. Treatment coding was used for all factors.

The criterion variable in the word recognition data was accuracy ("correct" vs "incorrect"). These analyses were completed using logit mixed-effect models with a binomial distribution. The same approach to the random effects specification was used in the models of recognition accuracy and the models of word duration. All statistical analyses were performed using R (R Development Core Team, 2008) and the R package lme4 for mixed-effects modelling (Bates & Maechler, 2010, version 1.1–7, 2014).

Results

Recognition

The model of word recognition accuracy for the set of 201 words for which both recognition and production data were available is summarized in Table 3 (left columns). The pattern of significant effects was similar to those observed in the larger set of 394 (right columns). Along with the expected facilitatory effects of lexical frequency, Phi-square density and PND emerged as significant predictors of recognition accuracy; words with more lexical competition by either measure showed lower accuracy in the set of 394 words. In the set of 201 words, the effect of PND was marginally significant (p = −.06). Syllable token frequency (residualized against lexical frequency) failed to produce a significant effect in either dataset. The strongest correlations among the fixed effects estimates in the model using 201 word types was that between the estimate of the effect of Phi-square density and PND (r = −.29). All other correlations among fixed effects had absolute values smaller than .16. The strongest correlations among the fixed effects estimates in the model using 394 word types were those between the estimate of the effect of Phi-square density and PND (r = −.24) and between Phi-square density and syllable token frequency (r = −.24). All other correlations among fixed effects had absolute values smaller than .12. It will be observed that the models do not include random slopes. In an earlier version of the current work, we did include by-participant random slopes for lexical frequency and PND (the maximal random effects structure supported by the data). The pattern of significant fixed effects was identical to the model reported here. The corresponding random slopes were not supported by the available data for the model of word duration, complicating the comparison of the significant predictors of recognition vs. production.[2] We report the models with identical random effects structure here, so as to avoid creating the impression that the behavior of the fixed effects (particularly for the critical variables Phi-square density and PND) was due to differences in power arising from differences in the random effects structure.

A follow-up model using (residualized) syllable type frequency (i.e. the number of word types containing a given syllable), rather than syllable token frequency (i.e. the summed frequency of all words containing a given syllable) also failed to reveal a significant effect of syllable frequency apart from lexical frequency. The significant effect of PND diverges from prior work, which showed that PND failed to account for unique variance in word recognition accuracy when Phi-square density was controlled for (Strand & Sommers, 2011). However, the data in Strand and Sommers (2011) included fewer target words. In an earlier version of the current work, using only 118 word types, we observed that PND ceased to be significant when Phi-square density was entered into the model, consistent with Strand and Sommers (2011). Therefore, effects of PND that are separate from Phi-square density may be subtle and require a large data set to obtain.

Word duration

The models of word duration for the 201 "shared" target words and for the total set of 469 words are summarized in Table 4. In both models, there were significant effects of baseline duration, forward and backward bigram probabilities, and speaking rate before and after the target, in the expected direction: Increased baseline duration was associated with longer word durations. High lexical frequency, high bigram probability and high contextual speaking rate were each associated with shorter word durations. Residual syllable token frequency failed to produce a significant effect. Phi-square density failed to give rise to a significant effect in either model. PND (counting only substitution-related neighbors) also failed to give rise to a significant effect in the model of 201 words, but did do

[2] We thank Florian Jaeger for raising this point.

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

Table 3
Summary of the models of word recognition accuracy, using 201 (left columns) and 394 (right columns) word types.

	201 Word types (10,070 observations)			394 Word types (19,860 observations)		
	β (SE)	z	p	β (SE)	z	p
Fixed effects						
(Intercept)	0.429 (0.108)	3.96	<.0001	0.244 (0.088)	2.77	0.01
Lexical frequency	0.301 (0.129)	2.327	0.02	0.525 (0.074)	7.09	<.0001
Syllable token frequency$_{res}$	−0.050 (0.094)	−0.534	0.59	0.052 (0.047)	1.12	0.26
PND	−0.033 (0.018)	−1.866	0.06	−0.041 (0.013)	−3.30	<.0001
Phi-square density	−0.640 (0.257)	−2.491	0.013	−0.852 (0.188)	−4.530	<.0001
	Variance		SD	Variance		SD
Random effects						
Target (intercept)	1.6930		1.301	1.8855		1.3731
Participant (intercept)	0.1369		0.370	0.1354		0.3679

Table 4
Summary of the models of word duration, using 201 (left columns) and (right columns) word types.

	201 Word types		477 Word types	
	β	t	β	t
Fixed effects				
(Intercept)	0.074 (0.018)	4.03	0.102 (0.016)	6.28
Baseline duration (log)	0.779 (0.062)	12.53	0.707 (0.044)	16.23
Backward bigram (log)	−0.024 (0.002)	−13.77	−0.025 (0.001)	−18.17
(Backward bigram, log)^2	0.005 (0.001)	6.19	0.003 (0.001)	5.6
Forward bigram (log)	−0.012 (0.002)	−6.16	−0.012 (0.002)	−7.73
Speech rate, after (log)	−0.147 (0.009)	−15.58	−0.136 (0.008)	−17.69
Speech rate, before (log)	−0.087 (0.009)	−10.1	−0.083 (0.007)	−11.33
(Speech rate, before (log))^2	−0.045 (0.013)	−3.5	−0.025 (0.011)	−2.31
Lexical frequency	−0.028 (0.006)	−4.99	−0.032 (0.004)	−7.68
Syllable token frequency$_{res}$	−0.007 (0.01)	−0.71	−0.009 (0.007)	−1.28
PND	−0.001 (0.002)	−0.97	−0.002 (0.001)	−2.05
Phi-square density	0.021 (0.024)	0.9	0.009 (0.017)	0.54
	Variance	SD	Variance	SD
Random effects				
Target (intercept)	0.007	0.085	0.0078	0.088
Speaker (intercept)	0.0089	0.0941	0.0079	0.089
Residual	0.0523	0.2287	0.0584	0.242

so in the larger data set. In an earlier version of the current work, we included the neighbors related to the target through substitution, addition, and deletion of segments in our models of word durations. That measure reached significance in the set of 201 words ($\beta = -0.002$, SE = 0.001, $t = -1.96$), as well as in the larger dataset ($\beta = -0.003$, SE = 0.001, $t = -2.97$). The pattern of significant effects, including the non-significance of Phi-square density, was otherwise identical in the models with substitution-only vs. substitution-deletion-addition neighbors. By either measure, higher neighborhood density was associated with shorter word durations. Phi-square density failed to give rise to a significant effect in all models of word duration, regardless of the size of the data set.

The strongest correlations among the fixed effects estimates for both data sets were between lexical frequency and the Intercept (−.29 for the smaller set and −.25 for the larger set) and between Phi-square density and the baseline estimate of word duration (.34 and .31, respectively). Follow-up models using (residualized) syllable type frequency, rather than syllable token frequency, failed to reveal a significant effect of syllable frequency apart from lexical frequency.

The baseline duration variable in the model of word duration is necessary, but raises a potential problem, pointed out by a reviewer. The variable is necessary because segments differ in their 'inherent' duration, as well as in the degree to which their duration varies. For example, nasal stops are much more variable in duration than taps. Unsurprisingly, the duration of a word that one might expect, given the segments it contains, is in fact a strong predictor of actual word duration. A model of word duration that ignored segmental content would strike us as misguided. However, the baseline duration variable is, by necessity, correlated with Phi-square density, PND, and any other variable that is partly predictable from segmental content. The correlation between Phi-square density and baseline duration arises because both are ultimately based on properties of segments. For example, sibilants are fairly confusable with one another. Words containing sibilants therefore tend to have higher Phi-square density than words that do not contain sibilants – although not always and not necessarily. The correlation between baseline duration and Phi-square density means that part of the variance potentially attributable to Phi-square density is accounted for by the baseline. In order to explore which

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

Table 5
Summary of effects of lexical variables on recognition accuracy and word duration in two word lists (n = 394 and n = 469) and their intersection (n = 201). Non-significant effects are marked "n.s.".

	Perception			Production	
	n = 394	n = 201		n = 469	
Lexical frequency	Increased accuracy	Increased accuracy		Shortening	Shortening
Phi-square density	Decreased accuracy	Decreased accuracy		n.s.	n.s.
PND	Decreased accuracy	Decreased accuracy		n.s./shortening	Shortening
Syllable frequency	n.s.	n.s.		n.s.	n.s.

of the two variables (baseline duration vs. Phi-square density) best explains the variability in word duration that could in principle be explained by either, we fitted a model containing one variable (baseline duration or Phi-square density) and then compared that model to a model containing the other variable, as well. The results indicated that baseline duration, i.e. the duration of a word that one might expect, given the segment it contains, was a robust predictor of actual word duration, whereas Phi-square density was not.

There is a sizable body of evidence showing that phonotactic probability affects segment duration – and therefore, potentially, word durations. For example, Kuperman, Ernestus, and Baayen (2008) show that there is a robust relationship between phonotactic probabilities (measured as n-phones, i.e. n-grams of phones) and segment duration in Dutch, English, German, and Italian spontaneous speech. Phonotactic probability is also known to be correlated with PND (Frauenfelder et al., 1993). This raises the possibility that the effects of PND in our models could be due to phonotactic probabilities. To explore this possibility, we fitted models using biphone probabilities in place of and alongside PND and/or syllable frequency and lexical frequency. We used the phonotactic probabilities from the Phonotactic Probability Calculator (Vitevitch & Luce, 2004). The effects of biphone probabilities were non-significant, except in models excluding lexical frequency and syllable frequency, i.e. when biphone probability was the only variable capturing the probability of target strings of segments. We interpret this pattern as indicating that phonotactic probability is predictive of word durations, and that PND has a significant effect on word durations beyond the effect of combinations of segments captured by biphone probabilities. One limitation of the biphone probabilities used here, and possibly the reason for the non-significance of biphone probabilities in the models containing syllable frequency and word frequency, was the fact that only within-word biphones were considered (the probabilities associated with CV and VC in each target), as opposed to biphones across word boundaries, i.e. taking into account the segments preceding and following the target in each utterance. We suspect that word-in-context biphones may very well yield an additional shortening effect, as observed in Kuperman et al. (2008).

We were also interested in seeing whether our data sets gave any indication of an effect of syllable frequency beyond the effect of lexical frequency. Therefore, we fitted simple linear regression models predicting lexical frequency from syllable frequency and *vice versa.* We then added the residuals of those models to our mixed-effects regression models. We found that residualized measures of syllable frequency never predicted variability beyond that attributable to lexical frequency.

Table 5 summarizes the pattern of significant effects of variables related to neighborhood density in the models of recognition and production.

Discussion

We assessed the ability of a segmental measure and a perceptually-based measure of word form similarity to predict two outcome variables – word durations in conversational speech and spoken word recognition accuracy. The phoneme-based measure (PND, i.e. phonological neighborhood density estimated as the number of words differing from the target in one segment) was a significant predictor of both spoken word recognition accuracy and word durations. The perceptually-based measure (Phi-square density) was a significant predictor of spoken word recognition accuracy, but not of spoken word durations. We interpret the significant effect of PND in both the production and the perception data sets as effects of lexical neighbors that are not necessarily perceptually similar to the target, but have segments in common with the target, consistent with numerous previous studies of phonological neighborhood density. We interpret the significant effect of Phi-square density on spoken word recognition, but not word durations, as reflecting an effect of 'perceptual neighbors', i.e. words that sound similar to the target, on recognition. Each of these conclusions has theoretical implications.

The first conclusion – that phonological neighbors in the lexicon affect articulatory detail – adds to the growing literature documenting effects of early stages of language production, such as lexical retrieval, on pronunciation (Arnold, Kahn, & Pancani, 2012; Gahl, 2008; Goldrick et al., 2013; Heller & Goldrick, 2014, 2015; Lam & Watson, 2010; Wright, 1979; Fink & Goldrick, 2015; Mousikou & Rastle, 2015). The idea that early stages of language production affect spoken word durations in connected speech is consistent with a more general line of research demonstrating the role of 'central' representations and mechanisms on 'peripheral' processes such as articulation and response execution generally. A similar line of research is being pursued in research on typing (Crump & Logan, 2010) and handwriting (Kandel, Peereman, & Ghimenton, 2013; Kandel, Peereman, Grosjacques, & Fayol, 2011; Roux, McKeeff, Grosjacques, Afonso, & Kandel, 2013). Effects of strictly lexical properties on fine phonetic detail are consistent with cascading

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

or fully interactive models of language production, in which articulation may proceed even as retrieval processes are still ongoing.

The second conclusion confirms and extends previous research showing that Phi-square density, a measure of perceptual neighborhood density, produces an effect on spoken word recognition over and above the effect of PND (phonological neighborhood density based on segment substitution). Prior work (Strand & Sommers, 2011) found that PND predicted word recognition accuracy when Phi-square density was not included in the model, but the effects of PND disappeared when Phi-square density was included. The current study, however, found that the significant effects of PND remained when Phi-square density was included. A possible explanation for this discrepancy is the larger dataset used in the current study. If the effects of PND beyond Phi-square density are small, the larger sample of words may be necessary to detect them. The current results suggest that the success of PND at predicting word recognition accuracy is not solely attributable to the fact that PND is approximating auditory similarity.

Once one accepts that lexical properties (as opposed to only 'peripheral' properties specific to the domain of motor movements) can be reflected in word durations, the question arises whether the direction of the effect – the fact that higher PND was associated with shorter, not longer, word durations, is expected or unexpected. If it is indeed the case that high PND has a facilitating effect on word form retrieval (Dell & Gordon, 2003; Marian & Blumenfeld, 2006; Vitevitch, 2002), then proposals claiming that pronunciation variation can reflect the speed of word form retrieval entail the prediction that the effects of PND should parallel those of lexical frequency. Frequent words shorten, and so should words from dense phonological neighborhoods. The current results, and the models of the same corpus of conversational speech reported in Gahl et al. (2012), are consistent with that prediction.

The idea that phonetic detail reflects early stages of language production, as opposed purely being a matter of motor execution also means that response latencies (in single-word production) and word durations (in connected speech) should reveal many of the same factors: Any factor known to facilitate lexical retrieval might potentially result in shorter word durations. Therefore, response latencies and word durations might be expected to correlate positively, a pattern that has been observed at times (e.g. Arnold et al., 2012; Mousikou & Rastle, 2015), but is far from being well established.

We have commented elsewhere (Gahl, 2008) that participants' tendency to pace themselves evenly in tasks involving the production of word lists and short phrases may get in the way of studying lexical effects on word duration. Articulatory and acoustic properties of word-initial segments are another factor that can complicate studying the relationship between speech onset latencies and word durations, as noted in Kawamoto, Liu, Mura, and Sanchez (2008). For example, Buz and Jaeger (2015) observed a positive correlation between latencies and word durations, but also found that effects of latencies (as a measure of lexical planning) and (frequency-weighted) phonological neighborhood density in a model of word durations were largely independent of one another, as evidenced by very low fixed-effect correlations in a mixed-effects regression model of word durations. Buz and Jaeger (2015) interpret this as evidence for the independence of planning and articulation. However, these findings are complicated by several methodological issues in the words used in Buz and Jaeger (2015), of which we mention one here. The low-density target words had a larger number of initial voiceless stops (7 out of 18) than the high-density words (2 out of 18). Voiceless stops begin with a complete closure, i.e. acoustically a period of silence that is indistinguishable from the latency to begin speaking. The duration of stop closures is variable, but in the order of 80 ms (Umeda, 1977). Therefore, the recorded "latencies" for the low-PND words (reported to be 52 ms longer than for the high-PND words) may be substantially increased by the initial stop closures and thus inflated relative to the high-PND words.

The model of word duration also has implications for the role of perception in pronunciation variation. We have argued that the word duration data modeled here reflect word-level information that is independent of the perceptual confusability of target words with other words. The effect of the number of phonological neighbors on target word duration does not appear to be due to perceptual target confusability. It may be important to point out that we restricted our attention to fluent multi-word utterances. We suspect that words in very short utterances, as well as words near pauses and disfluencies, all of which were excluded from our data, may be a better place to look for effects of perceptual target confusability with specific alternatives: For example, talkers respond to requests for clarification and disambiguation ("I said hyPERarticulated, not hyPOarticulated") and make up their minds about tricky word pairs ("Stalagm-, no wait, I mean stalacTITE!") – choices, in other words, in which target words are being contrasted with confusable alternatives.

Caveats and limitations

Several caveats are in order. One limitation of the current study concerns the continuous vs. categorical nature of our outcome variables. Comparing predictors of a continuous variable (word duration) and a categorical one (accuracy of word identification) may yield spurious apparent task-dependent differences (Tooley & Bock, 2014; we thank Florian Jaeger for pointing us to that work). One continuous measure tapping the recognition process that one might conceivably use to address this problem is the auditory lexical decision (ALD) task, i.e. a task in which participants are asked to make speeded judgments about whether phoneme strings form real words. However, ALD tasks are typically conducted in the absence of background noise (Goldinger, 1996). Presenting stimuli without masking noise can be an advantage, as it enables making inferences about lexical processing without degrading the stimuli. However, Phi-square density values are derived from measures of phoneme confusion in noise. Applying Phi-square density to measures of word identification in the absence of noise is therefore problematic. Indeed, although Luce and Pisoni (1998) calculated continuous

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

measures of perceptual similarity for predicting identification accuracy, they employed PND when predicting ALD responses. Therefore, given our use of Phi-square density, we did not include ALD data in the current study.

A second limitation arises because conversational speech and words spoken in citation form (with or without masking noise) sound quite different from one another (cf. Johnson, 2004). This difference has consequences for our ability to assess the role of auditory similarity in conversational speech production. This is a serious limitation not just of the current study, but, to our knowledge, prior research in this area more generally. The problem arises because available measures of perceptual similarity of words are based on segment confusability of segments produced in a citation context (e.g. [aCa]), not segments produced in conversational speech, in which segments undergo coarticulatory and other connected-speech processes that inevitably affects their acoustic and perceptual properties (cf. Farnetani & Recasens, 1997 for an overview). To our knowledge, sizable data sets on the perceptual confusability of either segments or words as produced in spontaneous speech are unavailable. Using tokens from the Buckeye Corpus in a recognition task, perhaps with a continuous measure of recognition difficulty, strikes us as a useful direction for future research. Doing so would necessitate different independent measures, however: measures of auditory confusability (like Phi-square density) are based on confusability of tokens produced in a citation context (e.g. vowel-target-vowel), which may sound quite different when produced in conversational speech. Therefore, Phi-square density may be a poor predictor of auditory confusability of words as spoken in conversational speech. This of course raises the possibility that Phi-square density (and other available segment-based measure of the auditory confusability of words) is a poor predictor of word duration not because perceptual confusability doesn't affect word durations, but because of the difference in segment confusability.

An alternative approach might be to investigate to what extent word durations in conversational speech are predictable from the confusability of phones in conversational speech. As a first step in that direction, we compared the confusability matrices that formed the basis for the Phi-square density measure (Strand & Sommers, 2011) to transcriber agreement data the Buckeye corpus (Raymond et al., 2002) to determine whether the types of feature confusions made were consistent across databases. In both, the vast majority of confusions were made within manner and place class (i.e., fricatives confused for other fricatives). However, the available data from the Buckeye transcribers contained too few instances of transcriber disagreement to enable meaningful comparisons with the types of confusions made in the recognition task.

In any case, transcriber agreement cannot replace a full analysis of the phone-by-phone confusability of phones: Segments in conversational speech frequently undergo various (and sometimes extreme) forms of phonetic reduction (Ernestus, 2014; Keune et al., 2005). In addition, the task of the transcribers was to listen veridically to the corpus data; the transcribers' task may favor different results than the forced-choice-over-noise tasks typically used in studies of perceptual similarity of segments.

A reviewer (Florian Jaeger) points out a third potential limitation, which is that comparisons of the predictive power of variables across models with different control variables is problematic. While that is certainly true, simply using identical predictors in models of different phenomena would create new problems: For example, inherent segment duration is an important predictor of word duration – longer segment durations add up to longer word durations, other things being equal. Our model of word duration takes this into account, by incorporating a baseline duration measure. However, segment duration does not straightforwardly predict spoken word recognition (or even segment recognition). Including that baseline duration measure in a model of spoken word recognition would therefore not be justified and might well prevent true predictors of spoken word recognition from revealing themselves.

Most of the caveats just discussed apply to many previous studies, as well as to the present work: After all, interest in PND as a lexical variable began with studies of recognition accuracy in single-word recognition tasks, which then inspired a huge amount of subsequent work on speech production, which set aside the questions just raised. We hope that the present research serves to inspire research that can address these questions.

Conclusion

Without additional assumptions, PND is a measure, not a mechanism. PND (the measure) indirectly reflects several distinct properties of words which are relevant at different stages of language production and recognition. Taking effects of PND or any other measure as direct evidence for a causal role of any particular measure in language production and comprehension runs the risk of shoehorning widely different phenomena into explanations that look appealingly uniform, but are ultimately lacking.

The interest that effects of word form similarity have held is due to the fact that such effects are thought to reflect the organization of the mental lexicon and the workings of the production and comprehension processes. Understanding the role of perceptual, articulatory, and modality-neutral representations and processes in language processing will help clarify the specific mechanisms by which humans perceive and produce spoken language.

Acknowledgments

We would like to thank the audiences at the 2013 CUNY Conference on Sentence Processing and the 2013 conferences of the Linguistic Society of America for helpful feedback on earlier versions of this work. We are also very grateful to Neal Fox, Florian Jaeger, Keith Johnson, Antje Meyer, Fabian Tomaschek, and two anonymous reviewers for their thoughtful comments.

Appendix A

See Tables A.1–A.4.

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

Table A.1
Correlation of Fixed Effects in the model of recognition, 201 word types.

	Intercept	Lexical frequency	Syllable token frequency$_{res}$	PND
Lexical frequency	−0.003			
Syllable token frequency$_{res}$	−0.006	0.017		
PND	−0.009	0.050	−0.128	
Phi-square density	−0.006	−0.123	−0.150	−0.287

Table A.2
Correlation of Fixed Effects in the model of recognition, 394 word types.

	Intercept	Lexical frequency	Syllable token frequency$_{res}$	PND
Lexical frequency	−0.001			
Syllable token frequency$_{res}$	0.000	0.023		
PND	−0.005	−0.118	−0.078	
Phi-square density	−0.002	0.030	−0.235	−0.239

Table A.3
Correlation of Fixed Effects in the model of word duration, 201 word types.

	Intercept	Baseline	BackBigr	BackBigr^2	ForwBigr	Rate, after	Rate, before	(Rate before)^2	Lex Freq	Syll. freq.	PND
Baseline duration (log)	0.01										
Backward bigram (log)	0.124	0.055									
(Backward bigram, log)^2	−0.224	−0.027	−0.188								
Forward bigram (log)	0.094	0.003	−0.028	0.044							
Speech rate, after (log)	0.011	−0.005	0.061	0.038	0.008						
Speech rate, before (log)	0.011	−0.001	−0.016	0.004	0.032	−0.095					
(Speech rate, before (log))^2	−0.076	0.005	−0.011	−0.017	0.013	0.008	0.017				
Lexical frequency	−0.286	0.184	−0.078	0.108	−0.113	−0.026	−0.016	−0.008			
Syllable token frequency$_{res}$	−0.038	−0.047	−0.002	0.021	−0.046	−0.012	−0.005	0.006	0.046		
PND	−0.002	0.085	0.003	0.00	0.017	−0.013	−0.002	0.00	0.092	−0.11	
Phi-square density	−0.003	0.337	−0.001	−0.001	0.011	0.017	−0.003	0.012	−0.031	−0.169	−0.205

Table A.4
Correlation of Fixed Effects, 487 word types.

	Intercept	Baseline	BackBigr	BackBigr^2	ForwBigr	Rate, after	Rate, before	(Rate before)^2	Lex Freq	Syll. freq.	PND
Baseline duration (log)	0.002										
Backward bigram (log)	0.112	0.05									
(Backward bigram, log)^2	−0.225	−0.012	−0.223								
Forward bigram (log)	0.086	−0.003	−0.004	0.04							
Speech rate, after (log)	0.019	−0.009	0.089	0.016	−0.002						
Speech rate, before (log)	0.011	0.002	−0.021	0.015	0.039	−0.079					
(Speech rate, before (log))^2	−0.071	0.007	−0.016	−0.004	0.009	0.022	0.035				
Lexical frequency (log)	−0.249	0.164	−0.078	0.135	−0.109	−0.028	−0.015	−0.008			
Syllable token frequency$_{res}$	−0.051	0.023	−0.025	0.013	−0.069	−0.005	−0.003	0.001	0.071		
PND	−0.008	0.122	−0.007	0.001	0.004	−0.003	0.001	−0.006	0.066	−0.026	
Phi-square density (log)	−0.003	0.311	0.002	−0.004	0.00	0.007	0.003	0.008	−0.016	−0.162	−0.232

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Arnold, J. E., Kahn, J. M., & Pancani, G. C. (2012). Audience design affects acoustic reduction via production facilitation. *Psychonomic Bulletin & Review, 19*, 505–512.

Aylett, M. (2000). Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and care of articulation in spontaneous speech (Ph.D.). Edinburgh. Retrieved from <http://www.cogsci.ed.ac.uk/matthewa/thesissum.html>.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence and duration in spontaneous speech. *Language and Speech, 47*, 31–56.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America, 119*, 3048–3058.

Baayen, H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX Lexical Database* (CD-ROM).

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes, 24*, 527–554.

Balota, D. A., Boland, J. E., & Shields, L. W. (1989). Priming in pronunciation: Beyond pattern recognition and onset latency. *Journal of Memory and Language, 28*, 14–36.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods, 39*, 445–459.

Barr, D., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-33. http://CRAN.R-project.org/package=lme4.

Bates, D., Kliegl, R., Vasishth, S., Baayen, H., (2015). Parsimonious mixed models. arXiv:1506.04967.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60*, 92–111.

Bent, T., Bradlow, A. R., & Smith, B. L. (2008). Production and perception of temporal patterns in native and non-native speech. *Phonetica, 65*, 131–147.

Brysbaert, M., & New, B. (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*, 977–990.

Buz, E., & Jaeger, T. F. (2013). Comparing measures of word confusability and their effect on speech production. *Poster presented at CUNY conference on sentence processing*.

Buz, E., & Jaeger, T. F. (2015). The (in) dependence of articulation and lexical planning during isolated word production. *Language, Cognition and Neuroscience*, 1–21.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review, 119*, 417–430.

Chen, Q., & Mirman, D. (2015). Interaction between phonological and semantic representations: Time matters. *Cognitive Science, 39*, 538–558. http://dx.doi.org/10.1111/cogs.12156.

Cholin, J., Levelt, W. J., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition, 99*, 205–235.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 551–563.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology, 55*, 429–432.

Crump, M. J., & Logan, G. D. (2010). Hierarchical control and skilled typing: Evidence for word-level control over the execution of individual keystrokes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1369–1380.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America, 83*, 1553–1573.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283–321.

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production* (pp. 9–47). New York: Mouton De Gruyter.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*, 801–838.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology, 1*, 121–144.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology, 55*, 149–179.

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua, 142*, 27–41.

Farnetani, E., & Recasens, D. (1997). Coarticulation and connected speech processes. *The Handbook of Phonetic Sciences*, 371–404.

Fink, A., & Goldrick, M. (2015). The influence of word retrieval and planning on phonetic variation: Implications for exemplar models. *Linguistics Vanguard*. http://dx.doi.org/10.1515/lingvan-2015-1003.

Flemming, E. (2010). Modeling listeners: Comments on Pluymaekers et al. and Scarborough. In C. Fougeron, B. Kühnert, M. D'Imperio & N. Vallée (Eds.), *Laboratory phonology* (Vol. 10, pp. 587–606). Berlin: Mouton De Gruyter.

Fosler-Lussier, E., & Morgan, N. (1999). Effects of speaking rate and word predictability on conversational pronunciations. *Speech Communication, 29*, 137–158.

Fox, N. P., Reilly, M., & Blumstein, S. E. (2015). Phonological neighborhood competition affects spoken word production irrespective of sentential context. *Journal of Memory and Language, 83*, 97–117.

Frauenfelder, U. H., Baayen, H., Hellwig, F. M., & Schreuder, R. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language, 32*, 781–804.

Fricke, Baese-Berk & Goldrick (in press). Dimensions of similarity in the mental lexicon. *Language, Cognition, and Neuroscience*.

Gahl, S. (2008). "Time" and "thyme" are not homophones: Word durations in spontaneous speech. *Language, 84*, 474–496.

Gahl, S. (2015). Lexical competition in vowel articulation revisited: Vowel dispersion in the Easy/Hard database. *Journal of Phonetics, 49*, 96–116.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language, 66*, 789–806.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*, 361–377.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Goldinger, S. D. (1996). Auditory lexical decision. *Language and Cognitive Processes, 11*, 559–568. http://dx.doi.org/10.1080/016909696386944.

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.

Goldrick, M., Baker, H. R., Murphy, A., & Baese-Berk, M. (2011). Interaction and representational integration: Evidence from speech errors. *Cognition, 121*, 58–72.

Goldrick, M., Folk, J. R., & Rapp, B. (2010). Mrs. Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language, 62*, 113–134.

Goldrick, M., Vaughn, C., & Murphy, A. (2013). The effects of lexical neighbors on stop consonant articulation. *Journal of the Acoustical Society of America, 134*, EL172–EL177.

Gordon, J. K. (2002). Phonological neighborhood effects in aphasic speech errors: Spontaneous and structured contexts. *Brain and Language, 82*, 113–145.

Gordon, J. (2014). The aging neighborhood: Phonological density in naming. *Language and Cognitive Processes, 29*, 326–344.

Greenberg, S. (1998). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. Paper presented at the ESCA workshop on modeling pronunciation variation for automatic speech recognition, Kekrade (The Netherlands).

Gupta, P., & MacWhinney, B. (1995). Is the articulatory loop articulatory or auditory? Reexamining the effects of concurrent articulation on immediate serial recall. *Journal of Memory and Language, 34*, 63–88.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research, 32*(2), 101–123.

Harley, T. A., & Bown, H. E. (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology, 89*, 151–174.

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

Heller, J. R., & Goldrick, M. (2015). Erratum to: Grammatical constraints on phonological encoding in speech production. *Psychonomic Bulletin & Review, 22*, 1475.

Heller, J. R., & Goldrick, M. (2014). Grammatical constraints on phonological encoding in speech production. *Psychonomic Bulletin & Review, 21*, 1576–1582.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*, 135–145.

Iverson, P., Bernstein, L. E., & Auer, E. T. Jr., (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Communication, 26*, 45–63.

Jaeger, T. F., & Levy, R. P. (2006). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium* (pp. 29–54). Tokyo: The National International Institute for Japanese Language.

Jones, D. M., Beaman, P., & Macken, W. J. (1996). The object-oriented episodic record model. In S. E. Gathercole (Ed.), *Models of short-term memory* (pp. 209–238). Hove, UK: Psychology Press.

Jurafsky, D. (2003). Probabilistic modeling in Psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic Linguistics* (pp. 39–95). Cambridge, MA: MIT Press.

Kandel, S., Peereman, R., & Ghimenton, A. (2013). Further evidence for the interaction of central and peripheral processes: The impact of double letters in writing English words. *Frontiers in Psychology, 4*.

Kandel, S., Peereman, R., Grosjacques, G., & Fayol, M. (2011). For a psycholinguistic model of handwriting production: Testing the syllable-bigram controversy. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 1310.

Kawamoto, A. H., Liu, Q., Mura, K., & Sanchez, A. (2008). Articulatory preparation in the delayed naming task. *Journal of Memory and Language, 58*, 347–365.

Keune, K., Ernestus, M., Hout, R. v., & Baayen, H. (2005). Variation in Dutch: From written MOGELIJK to spoken MOK. *Corpus Linguistics and Linguistic Theory, 1*, 183–223.

Kuperman, V., Ernestus, M., & Baayen, H. (2008). Frequency distributions of uniphones, diphones, and triphones in spontaneous speech. *The Journal of the Acoustical Society of America, 124*, 3897–3908.

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *Journal of the Acoustical Society of America, 121*, 2261–2271.

Lam, T. Q., & Watson, D. G. (2010). Repetition is easy: Why repeated referents have reduced prominence. *Memory & Cognition, 38*, 1137–1146.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–75.

Levelt, W. J., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review, 98*, 122.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition, 50*, 239–269.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 403–439). Dordrecht: Kluwer.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19*, 1–36.

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

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*, 133–156.

Marian, V., & Blumenfeld, H. K. (2006). Phonological neighborhood density guides: Lexical access in native and non-native language production. *Journal of Social and Ecological Boundaries, 2*, 3–35.

Mousikou, P., & Rastle, K. (2015). Lexical frequency effects on articulation: A comparison of picture naming and reading aloud. *Frontiers in Psychology, 6*.

Munson, B. (2007). Lexical access, lexical representation, and vowel production. In J. Cole & J. I. Hualde (Eds.), *Laboratory phonology 9: Phonology and phonetics* (pp. 201–227). Berlin: Mouton de Gruyter.

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Speech, Language, and Hearing Research, 47*, 1048–1058.

Page, M. P., Madge, A., Cumming, N., & Norris, D. G. (2007). Speech errors and the phonological similarity effect in short-term memory: Evidence suggesting a common locus. *Journal of Memory and Language, 56*, 49–64.

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language, 78*, 1–17.

Peramunage, D., Blumstein, S. E., Myers, E. B., Goldrick, M., & Baese-Berk, M. (2010). Phonological neighborhood effects in spoken word production: An fMRI study. *Journal of Cognitive Neuroscience, 23*, 593–603.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America, 32*, 693–703.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye corpus of conversational speech (2nd release) <www.buckeyecorpus.osu.edu>.

Pluymaekers, M., Ernestus, M., & Baayen, H. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica, 62*, 146–159.

Roux, S., McKeeff, T. J., Grosjacques, G., Afonso, O., & Kandel, S. (2013). The interaction between central and peripheral processes in handwriting production. *Cognition, 127*, 235–241.

Sadat, J., Martin, C. D., & Costa, A. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive Psychology, 68*, 33–58.

Scarborough, R. A. (2004). *Degree of coarticulation and lexical confusability*. Paper presented at the proceedings of the 29th meeting of the Berkeley Linguistics Society.

Scarborough, R. A. (2013). Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. *Journal of Phonetics, 41*, 491–508.

Scarborough, R. A. (2010). Lexical and contextual predictability: Confluent effects on the production of vowels. In L. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Laboratory phonology* (pp. 557–586). Berlin, New York: De Gruyter Mouton. 10.

Sevald, C., & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition, 58*, 91–127.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition, 133*, 140–155.

Slote, J., & Strand, J. (2015). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods*, in press. http://dx.doi.org/10.3758/s13428-015-0599-7.

Strand, J. (2014). Phi-square Lexical Competition Database (Phi-Lex): An online tool for quantifying auditory and visual lexical competition. *Behavior Research Methods, 46*, 148–158. http://dx.doi.org/10.3758/s13428-013-0356-8.

Strand, J., & Sommers, M. (2011). Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *Journal of the Acoustical Society of America, 130*, 1663–1672.

Tily, H., & Kuperman, V. (2012). Rational phonological lengthening in spoken Dutch. *Journal of the Acoustic Society of America, 132*, 3935–3940.

Umeda, N. (1977). Consonant duration in American English. *Journal of the Acoustical Society of America, 61*, 846–858.

van Son, R. J. J. H., & Pols, L. C. W. (2003). *Information structure and efficiency in speech production*. Paper presented at the 2003 Eurospeech conference, Geneva, Switzerland.

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech, 40*, 211–228.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 735–747.

Vitevitch, M. S. (2007). The spread of the phonological neighborhood influences spoken word recognition. *Memory & Cognition, 35*, 166–175.

Vitevitch, M. S., Armbrüster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 514.

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9*, 325–329.

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers, 36*, 481–487.

Vitevitch, M. S., & Luce, P. A. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language, 52*, 193–204.

Vitevitch, M. S., & Luce, P. A. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics, 2*. http://dx.doi.org/10.1146/annurev-linguist-030514-124832.

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition, 31*, 491–504.

Wilson, M. (2001). The case for sensorimotor coding in working memory. *Psychonomic Bulletin & Review, 8*, 44–57.

Wright, C. E. (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition, 7*, 411–419.

Yiu, L. K., & Watson, D. G. (2015). When overlap leads to competition: Effects of phonological encoding on word duration. *Psychonomic Bulletin & Review*, 1–8.

Please cite this article in press as: Gahl, S., & Strand, J. F. Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language* (2016), http://dx.doi.org/10.1016/j.jml.2015.12.006