Making Long-Distance Relationships Work: Quantifying Lexical Competition with Hidden Markov Models[*]

Julia Strand
Department of Psychology
Carleton College
jstrand@carleton.edu

David Liben-Nowell
Department of Computer Science
Carleton College
dlibenno@carleton.edu

April 15, 2016

Abstract

A listener recognizes a stimulus word from acoustic–phonetic input by discriminating that word's representation from those of other words. The Neighborhood Activation Model (NAM) [28] is a long-standing and deeply influential model quantifying how properties of the stimulus word and its competitors influence recognition. The current project incorporates Hidden Markov Models (HMMs) into the NAM's framework to more flexibly evaluate the influence of multiple lexical properties, thereby allowing us to pose novel questions about the process of spoken-word recognition. Analyses using HMMs' power to evaluate the stimulus's "distance" even to very distant words suggest that faraway words still act as competitors, suggesting that a larger subset of the lexicon is activated during recognition than has been previously assumed. Another analysis reveals that the way competition is distributed among other words significantly influences word recognition. HMMs have been widely applied in other domains, and our results demonstrate that they may be similarly suited to quantifying the processes underlying spoken-word recognition.

Keywords: Lexical competition, spoken-word recognition, Neighborhood Activation Model, Hidden Markov Models.

[*] This is a prepublication version of the paper. The final version appears in *Journal of Memory and Language,* 2016. We thank Emily Massell and Violet Brown for assistance with data collection, and Anna Rafferty for helpful discussions. This work was supported by Carleton College. Portions of this work were performed while the second author was visiting Churchill College and the Computer Laboratory at the University of Cambridge; thanks to both for their support and hospitality.

Introduction

Recognizing spoken words is an impressive perceptual and cognitive accomplishment; humans can process 250 words per minute [16] and distinguish between timing differences in speech stimuli as short as 20ms [13]. In the last 60 years, numerous models have been developed to quantify the processes underlying spoken-word recognition [27,28,30–32]. The models are generally in agreement about several features of the word-recognition process. First, incoming acoustical input activates multiple lexical representations in memory. For example, when a listener hears a stimulus word "bird," perceptually similar lexical entries (often called *competitors* or *neighbors*) such as "burn," "heard," and "bud" are also partially activated. In addition, the degree of activation is relative to the perceptual similarity between the stimulus word and the competitor [28, 32]. When hearing "bird," the highly confusable competitor "burn" is assumed to be more highly activated than the less confusable "bin." The activation of the competitors influences recognition of the stimulus word, with greater activation from competitors hindering recognition. Therefore, words that are more perceptually distinctive, and therefore have less lexical competition, are recognized more quickly and accurately than those with more competition [28, 46].

In addition to the dynamics of lexical activation and competition, current models of spoken-word recognition also include mechanisms to account for the well-established effects of frequency of occurrence. Words that occur frequently in the language are identified more quickly and accurately than those that are rare [40]. Word frequency also appears to modulate lexical competition effects; words with low-frequency competitors are identified more quickly and accurately than those with high-frequency competitors [28]. Most models assume that frequency acts by weighting activation levels of words prior to recognition or by biasing responses toward more frequent words following identification (see [11]).

To simultaneously represent the influence of frequency effects and lexical competition on spoken-word recognition, Luce & Pisoni [26, 28] proposed the *Neighborhood Activation Model (NAM).* The NAM has since become the most influential mathematical model of spoken-word recognition (see [29] for a detailed discussion of the difference between mathematical models that seek to quantify lexical competition and other forms of computational models that simulate the process). Implementations of the NAM can readily generate predictions about how difficult specific words or classes of words will be to recognize, and these predictions may then easily be tested against human accuracy in spoken-word recognition. The flexibility of the the NAM framework makes it an attractive candidate for testing novel predictions about the processes underlying spoken-word recognition. For example, the NAM predicts word recognition in cochlear implant users [22], older adults [43], and in visually perceived (lipread) speech [2] when the input to the model is changed to make it appropriate for the population or modality. Below, we outline the architecture of the NAM, identify issues in spoken-word recognition that the NAM has difficulty handling in its current form, and propose a novel method for quantifying the processes of lexical activation, and, therefore, competition.

Architecture of the Neighborhood Activation Model

The NAM assumes that word recognition involves discriminating among lexical representations that are activated in parallel. This process occurs using a collection of *word decision units* that simultaneously monitor three sources of information: the acoustic–phonetic input (bottom-up support for the lexical candidate), higher-level lexical information (word frequency), and the overall level of activity of other word decision units (lexical competition). Word recognition occurs when a specific word decision unit reaches a criterion and is discriminated from other activated lexical representations.

Given the high-level nature of the description that we have just given, there are many different operationalizations that are consistent with this model. As an anonymous reviewer of a previous version of this paper noted, it is important to distinguish the abstract verbal description of the process of word recognition from the numerical output of any particular formula motivated by that description. Some claims of the NAM, such as the beneficial effects of word frequency, are straightforward to turn into quantified predictions about word recognition. However, others, such as precisely what lexical competition means, are subtler. The original NAM paper quantifies competition in a particular way and those quantifications are tested in the paper. Here, we propose other methods of quantification that, while in keeping with the spirit of NAM [28], make different predictions from the formulas contained therein. Comparing multiple ways of operationalizing lexical competition, each of which has implicit assumptions about the underlying process, allows us to pose questions about the process of spoken-word recognition.

As described by Luce & Pisoni [28], the process within a word decision unit is quantified using *Frequency-Weighted Neighborhood Probability (FWNP)*, which combines the influences of all three sources of information: acoustic-phonetic support, word frequency, and lexical competition. For a word w with phonemes $w_1, w_2, \ldots, w_n$, FWNP is

$$FWNP(w) := \frac{\overbrace{\left[\left(\prod_i p(w_i|w_i)\right)\cdot \mathit{freq}(w)\right]}^{\boxed{\text{frequency-weighted } SWP(w)}}}{\underbrace{\left[\left(\prod_i p(w_i|w_i)\right)\cdot \mathit{freq}(w)\right]}_{\boxed{\text{frequency-weighted } SWP(w)}} + \underbrace{\left[\sum_{\text{competitor words } w'}\left[\left(\prod_i p(w'_i|w_i)\right)\cdot \mathit{freq}(w')\right]\right]}_{\boxed{\text{frequency-weighted } NWP(w)}}}$$

Here $\mathit{freq}(x)$ quantifies the frequency of occurrence of a word x (typically represented as the log of the number of occurrences per million words), and $p(x|y)$ denotes the conditional probability of identifying a presented phoneme y as the phoneme x in a forced-choice phoneme identification task.

The numerator of $FWNP(w)$ is the *Stimulus Word Probability (SWP),* weighted by frequency. The SWP represents the bottom-up support for the stimulus word and can be thought of as a measure of intelligibility, because it quantifies the probability of perceiving the phonemes of the stimulus word given that those phonemes were presented. For example, the SWP of the word "bat" /bæt/ is

$$SWP(\text{bat}) = p(\text{b}|\text{b}) \cdot p(\text{æ}|\text{æ}) \cdot p(\text{t}|\text{t}).$$

The NAM posits that the decision units of words containing easily identified segments receive more support from the acoustic–phonetic input than words containing segments that are difficult to identify. For example, if participants correctly identify /w/ more often than /θ/ on the forced-choice phoneme identification task, then the word "win" will have a higher SWP than the word "thin," and therefore, *ceteris paribus,* will be more likely to be correctly identified. Thus, the SWP reflects the likelihood of correctly identifying a stimulus word, based on the word's segments themselves.

The denominator of $FWNP(w)$ contains the summed frequency-weighted support for each lexical competitor w' (NWP, *Neighbor Word Probability*) from the acoustic–phonetic input for w, along with the frequency-weighted SWP (the support for the stimulus word itself). Competitors that are highly perceptually confusable with the stimulus word generate high NWPs, representing strong activation from the acoustic–phonetic input. Similarity is quantified in a similar manner to SWPs, again using the conditional probability of confusing the stimulus word's phonemes with the competitor's position-specific phonemes. For example, the probability of responding "meet" /mit/ given the stimulus "bead" /bid/ is calculated as

$$NWP(\text{meet}|\text{bead}) = p(\text{m}|\text{b}) \cdot p(\text{i}|\text{i}) \cdot p(\text{t}|\text{d}).$$

Using this method, it is possible to calculate the NWP of any competitor word w' for a stimulus word w of the same length, even if the stimulus word and the competitor share no phonemes. But comparing words that differ in length (e.g., "meets" and "bead," or "lease" and "least") requires an additional step. To achieve this, Luce & Pisoni [28] included a "null response" category in the forced-choice consonant identification task. On some trials, no consonant was presented, and participants had "nothing presented" as a response category on all trials. Therefore, it is possible to determine the likelihood that participants would falsely report hearing a specific phoneme when nothing was presented (i.e., "hallucinating" that phoneme), or the likelihood that participants would respond that they had not heard anything when a specific phoneme was presented (i.e., missing the phoneme). Using the null-response data, Luce & Pisoni [28] are able to compute a quantity corresponding to the chances of perceiving, say, "meets" /mits/ when the actual stimulus presentation was "bead" /bid/: they align the vowels of the two words, and, working outwards from the vowel, evaluate the phoneme-by-phoneme similarities, using the null-response category where the words do not overlap. That is, they compute the probability of perceiving "meets" given "bead" as

$$p(\text{m}|\text{b}) \cdot p(\text{i}|\text{i}) \cdot p(\text{t}|\text{d}) \cdot p(\text{s}|\varnothing).$$

Conversely, the likelihood of perceiving "me" /mi/ given "bead" /bid/— missing the /d/—is computed as

$$p(\text{m}|\text{b}) \cdot p(\text{i}|\text{i}) \cdot p(\varnothing|\text{d}).$$

In general, if we renumber the phonemes of a univocalic stimulus and competitor so that the vowel is the 0th phoneme of each (so that phonemes in the onset have negative indices, and those in the coda have positive indices), then we can express NWP as

$$NWP(w'|w) := \prod_{i=\ldots,-2,-1,0,1,2,\ldots} p(w'_i|w_i),$$

where $p(w'_i|\varnothing)$ and $p(\varnothing|w_i)$, respectively, correspond to the probabilities of hallucinating w'_i and missing w_i.

(Note that, under this description, perceiving "scat" /skæt/ when presented with "cat" /kæt/ is the result of hallucinating an /s/ before the /k/. An alternative description of this misperception is that we have mistaken the single phoneme /k/ for the diphone /sk/: in either case, the effect is to have perceived /skæt/ when the stimulus was /kæt/. The key difference between these ways of stating the process is that the /sk/-for-/k/ confusion implicitly includes the context in which the hallucination occurred: "the chance of hallucinating /s/ before a truly presented word-initial /k/" is identical to "the chance of hearing /sk/ when presented with a word-initial /k/." But they may not be identical to "the chance of hallucinating /s/," which omits the context. For more, see the paragraphs marked "Potential limitations in quantifying competition" in the Discussion.)

To evaluate the total amount of lexical competition that a stimulus word encounters, Luce & Pisoni [26, 28] calculated the NWPs of every English monosyllabic word, given any one of their (monosyllabic)[1] stimulus words, and summed these values. Overall, then, the computed likelihood of correctly recognizing the stimulus word w is the value $FWNP(w)$, the proportion of the frequency-weighted support for words in the lexicon that the stimulus word contributes. The model predicts that recognition should be most difficult for words that have low SWP (that is, low predicted intelligibility), low frequency, and are perceptually similar to high-frequency competitors. Although the computations within the word decision units are relatively simple, the model has good predictive power; correlations between FWNP and word-recognition scores range from $r = .23$ to $r = .47$ [26, 28].

By simultaneously combining the effects of predicted intelligibility, frequency, and lexical competition, the FWNP makes specific, testable predictions about human word-recognition performance. For example, if a target word is preceded by a phonetically related prime, the residual activation from the prime should increase the NWP for that prime as a competitor, thereby reducing the target word FWNP. Indeed, priming a target word (e.g., “bull”) with the competitor with the highest NWP that does not share phonemes (e.g., “veer”) reduced identification accuracy [20] and increased shadowing latencies [27] for the target. In addition, FWNPs can be predictive in situations that cause listeners to activate a particular set of words; FWNPs calculated for the names of drugs, using a lexicon of drug names rather than all English words, significantly predict clinicians’ identification accuracy in noise [23, 24]. The mechanisms of activation and competition described by FWNPs are not limited to auditory speech; when the perceptual input is visual (lipread), FWNPs derived from visual confusion matrices predict lipread word-identification accuracy [2, 14].

Although FWNPs successfully predict human word-recognition performance, the simplicity of these computations makes some implicit assumptions about the quantification of lexical activation that may or may not have been intended by the creators of NAM. That is, some components of the FWNPs may reflect computational challenges rather than theoretical claims. As one example, the NWP for “cast” /kæst/ given the stimulus word “cat” /kæt/ is

$$p(\text{k}|\text{k}) \cdot p(\text{æ}|\text{æ}) \cdot p(\text{s}|\text{t}) \cdot p(\text{t}|\varnothing),$$

representing the participant correctly hearing /k/ and /æ/, then mistaking /s/ for /t/, and finally hallucinating a /t/ when nothing was presented. Although this sequence certainly is one way that the two words could have been mistaken, it is also possible that the error was a single hallucination of an /s/ (while perceiving every other phoneme correctly), quantified as

$$p(\text{k}|\text{k}) \cdot p(\text{æ}|\text{æ}) \cdot p(\text{s}|\varnothing) \cdot p(\text{t}|\text{t}).$$

The overall confusability of word pairs (and therefore degree of lexical competition for a particular stimulus word) may be more accurately represented by a metric that considers other possible ways of confusing the words.

The NAM successfully predicts many phenomena in human word recognition and has been very influential in the field. Incorporating more flexible methods for computing perceptual similarity and

[1] The reference lexicon used by Luce & Pisoni [28] included only monosyllabic words. However, the authors specify that this restriction was imposed to simplify the computational analysis, rather than as a theoretical claim about which words are simultaneously activated. Although the calculations for the FWNP apply directly to polysyllabic univocalic words (“cattle” as a competitor for “cat”), as implemented by Luce & Pisoni [28], FWNP cannot be computed for words with multiple vowels without modification.

lexical competition will allow us to extend the NAM framework to ask additional questions about the processes underlying spoken-word recognition. Indeed, Magnuson, Mirman, & Harris [29, p. 79] suggest that "using other similarity metrics in the NAM framework would be an excellent strategy for making further progress on identifying general constraints on spoken-word recognition."

A More Flexible Framework for Measuring Competition: Hidden Markov Models

The NAM framework allows us to focus our attention on one concrete computational task: given a stimulus word w and any particular competitor word w', we must compute the probability that a listener would hear w' instead of w. There are two distinct difficulties for doing this computation in the basic NAM framework. First, we seem to get the wrong probabilities for some words: the NWP assumes a particular structure in the perceptions and misperceptions that caused w' to be heard instead of w. Namely, hallucinated and missing phonemes can occur only at the periphery of the word (as in the "cast|cat" example above), while the phonemes adjacent to the aligned vowel may only be mistaken for each other. That assumed structure may not be the most probable way of confusing w and w'. Second, we cannot even perform this computation for every potential competitor: for those competitors that contain multiple vowels, the NWP does not allow us to compute a confusion probability at all because "aligning the vowels of w and w'" is not well defined.

Here, we suggest a generalization in the NAM framework to simultaneously address these issues by adapting to our setting a probabilistic modeling approach called *Hidden Markov Models* (HMMs; see [21,39] for an introduction to this technique). HMMs have been used successfully in a wide range of domains, including computational biology [12,49], visual perception [7], and automatic speech recognition [18,37]; see also [41] for links between automatic and human speech recognition), but have not yet been applied to quantifying lexical competition. Intuitively, a Hidden Markov Model is an abstract mathematical "machine" that describes a probabilistic process that generates an "output" (in our setting, a sequence of phonemes). We will construct an HMM for every word in the lexicon, with the goal that the machine for w will generate w' as output with higher probability the more perceptually similar w and w' are.

To be concrete, we will describe the HMM for a particular word, "cat" /kæt/. The HMM M_{cat} consists of a collection of *states*: one state for each of the three phonemes in "cat," and four "hallucination states" before and after each of these phonemes. We also add a *start state* and an *end state* (see Figure 1).

Each of these states is associated with two probability distributions: the *emission probabilities* (for example, when the machine is in the /æ/ state, what is the probability that the phoneme /æ/ will be generated? Or /o/? Or no phoneme at all?) and *transition probabilities* (if the machine is currently in the /æ/ state, what is the probability that it will be in the /t/ state in the next step? Or hallucination state 3?). Our probability distributions are derived experimentally, as we discuss below. "Running" an HMM M_{cat} means visiting a sequence of states of the machine, generating a phoneme (or $\varnothing$) in each state that we visit. More concretely, we begin in the start state, and repeatedly (a) append to the output sequence a phoneme chosen according to the current state's emission probabilities;[2] and (b) move to a new state, chosen according to the current state's

[2] Given the data we have, the emission probability for phoneme x in a particular state y is the confusion probability $p(x|y)$. Phonemic context could be incorporated into an HMM by modifying the confusion probabilities for each particular state based on one or more preceding phonemes. To add one phoneme of context, for example, the

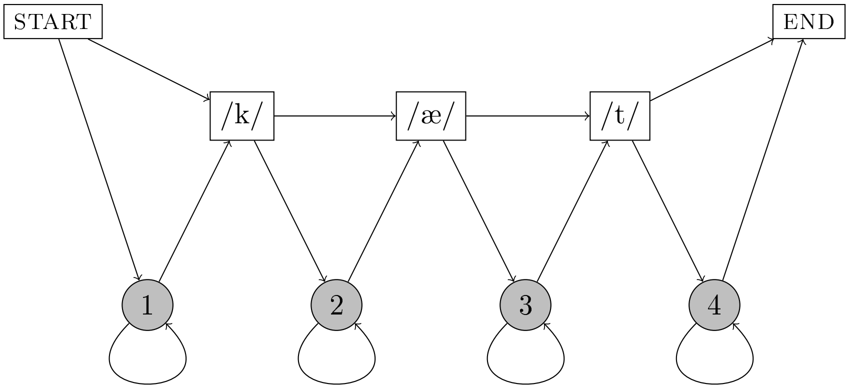

Figure 1: The Hidden Markov Model for the word "cat."

transition probabilities. We continue to run the machine until it arrives in the final state. This process results in the machine *generating* a particular output sequence of phonemes.[3]

Note that the HMM M_{cat} can generate many different outputs, depending on the probabilistic choices that happen to be made during the run. For example, M_{cat} can generate "cat" by following the state sequence

$$\textsc{start} \to /\text{k}/ \to /\text{æ}/ \to /\text{t}/ \to \textsc{end},$$

and successively generating $/\varnothing, \text{k}, \text{æ}, \text{t}, \varnothing/$ in those five states. Or M_{cat} can generate "claps" by following the state sequence

$$\textsc{start} \to /\text{k}/ \to 2 \to /\text{æ}/ \to /\text{t}/ \to 4 \to \textsc{end},$$

and successively generating $/\varnothing, \text{k}, \text{l}, \text{æ}, \text{p}, \text{s}, \varnothing/$ in those seven states. (Note that in the latter run the state /t/ generated /p/: a less probable phoneme for the /t/ state to generate than /t/, but a possibility.) A crucial mathematical fact about an HMM is that

$$\sum_{\textit{phonemic sequences } w'} p(w'|M_w) = 1,$$

i.e., that the quantities $p(w'|M_w)$, the probabilities of the machine generating any particular output sequence w', in fact form a probability function. This fact allows us to reason about that probability distribution rigorously. Intuitively, a competitor w' that is generated with a higher probability by the HMM M_w is more similar to w.

Importantly, an HMM can generate the same phonemic sequence in multiple ways—that is, via multiple paths (sequences of hidden states) through the machine. For example, the word "cast" could be generated by M_{cat} along the path

$$\textsc{start} \to /\text{k}/ \to /\text{æ}/ \to /\text{t}/ \to 4 \to \textsc{end},$$

probability of hallucinating /s/ in State 3 in Figure 1 (to generate the word "cast") would be set to the (experimentally derived) probability of hallucinating an /s/ immediately after /æ/; the probability of hallucinating an /s/ in State 4 (to generate the word "cats") would be the (experimentally derived) probability of hallucinating an /s/ immediately after /t/. Similarly, the emission probabilities in the /æ/ state of M_{cat} would differ from the emission probabilities in the corresponding /æ/ state of M_{pat}.

[3] We can observe the phonemes in the output, but which state the HMM is in at any particular time is unobservable; the "hidden" in the name "Hidden Markov Model" derives from this fact.

where state /t/ generated /s/ and state 4 (the last hallucination state) generated /t/. That path corresponds to what the original NAM computes. Or "cast" could be generated along a more probable path, following the state sequence

$$\textsc{start} \to /\text{k}/ \to /\text{æ}/ \to 3 \to /\text{t}/ \to \textsc{end},$$

where state 3 generated /s/. Or "cast" could be generated along a much less probable path,

$$\textsc{start} \to /\text{k}/ \to /\text{æ}/ \to /\text{t}/ \to 4 \to 4 \to 4 \to \textsc{end},$$

where state /k/ generated ∅, state /æ/ generated ∅, state /t/ generated /k/, and state 4 successively generated /æ/, /s/, and /t/. The probabilities of these respective sequences are

$$p(\text{non-hall}) \cdot p(\text{k}|\text{k}) \cdot p(\text{non-hall}) \cdot p(\text{æ}|\text{æ}) \cdot p(\text{non-hall}) \cdot p(\text{s}|\text{t}) \cdot p(\text{hall}) \cdot p(\text{t}|\varnothing) \cdot p(\text{non-hall})$$

$$p(\text{non-hall}) \cdot p(\text{k}|\text{k}) \cdot p(\text{non-hall}) \cdot p(\text{æ}|\text{æ}) \cdot p(\text{hall}) \cdot p(\text{t}|\varnothing) \cdot p(\text{non-hall}) \cdot p(\text{t}|\text{t}) \cdot p(\text{non-hall})$$

$$\begin{aligned}&p(\text{non-hall}) \cdot p(\varnothing|\text{k}) \cdot p(\text{non-hall}) \cdot p(\varnothing|\text{æ}) \cdot p(\text{non-hall}) \cdot p(\text{k}|\text{t}) \cdot p(\text{hall}) \cdot p(\text{æ}|\varnothing)\\ &\quad \cdot p(\text{hall}) \cdot p(\text{s}|\varnothing) \cdot p(\text{hall}) \cdot p(\text{t}|\varnothing) \cdot p(\text{non-hall})\end{aligned}$$

where p(hall) and p(non-hall) denote the probability of hallucinating and not hallucinating, respectively. Unlike the original NAM formulation, the HMM imposes no assumption on the "alignment" of a competitor word to the baseline stimulus word from which the machine was constructed. Instead, we can simultaneously consider all possible ways that a stimulus word w can be mistaken for a competitor w'. Using versions of two well-known algorithms for HMMs, called the Viterbi Algorithm and the Forward Algorithm (see [21, 39]), we can efficiently compute, for any particular HMM M_w and any particular word w', two key quantities: (a) the *maximum path probability* of w' (out of all paths through M_w that generate w', what is the probability of the most probable path?); and (b) the *total probability* of w' (out of all paths through M_w that generate w', what is the sum of the probabilities of these paths?). We can use both of these quantities as measures of the confusability of words w and w'.

The Current Study

A key principle of the NAM is that word identification is influenced by "the number and nature of lexical items activated by the stimulus input" [28, p. 12]. Using HMMs in the NAM framework will allow us to further explore the processes underlying spoken-word recognition by more flexibly evaluating the number and nature of activated representations. Below, we describe four novel research questions that may be addressed using HMMs.

Aim 1. Measuring perceptual intelligibility.

The NAM and other models of word recognition posit that lexical activation is a function of the perceptual similarity between the acoustic–phonetic input and the lexical item in memory. In the NAM, intelligibility is calculated as the perceptual support for the stimulus word from the stimulus input, e.g., the SWP of "bat" /bæt/ is

$$SWP(\text{bat}) = p(\text{b}|\text{b}) \cdot p(\text{æ}|\text{æ}) \cdot p(\text{t}|\text{t}).$$

However, it is possible for a word to be identified correctly despite misperceptions (to be "right for the wrong reasons"): e.g., to identify "bat" as "bat" but doing so by *both* missing the /b/ *and* then immediately after hallucinating a /b/. HMMs enable us to evaluate multiple paths to correct recognition. This capability will enable us to evaluate whether having other paths to "correct" recognition facilitates recognition because there are more ways to hear the right word, or hinders recognition because other paths represent a kind of "self-confusability."

Aim 2. Measuring perceptual confusability.

Quantifying perceptual similarity effectively is a critical component to modeling lexical competition. To the best of our knowledge, no studies to date have included multiple routes for confusing pairs of words, as in the example above. This omission is likely due to computational difficulty, rather than theoretical motivation. Therefore, we will use HMMs to more flexibly represent the perceptual similarity of word pairs, computing all of the ways in which a pair of words can be confused for each other rather than just one.

Aim 3. Quantifying the spread of lexical activation.

This project also aims to make progress in evaluating the extent to which activation spreads throughout the lexicon. Although models of word recognition agree that highly perceptually similar competitors receive the most activation, it is unclear whether even relatively dissimilar words are also activated to some degree. In other words, the extent to which lexical activation diffuses throughout the lexicon remains unknown. In the original presentation of NAM, all stimulus words were consonant–vowel–consonant (CVC) words, and the reference lexicon included all monosyllabic words. Therefore, for the stimulus word "beat" /bit/, "be" /bi/ could act as a competitor but "beetle" /bidl/, a disyllabic word containing a syllabic /l/, could not. Simplified implementations of NAM have quantified lexical competition using the Deletion–Addition–Subtraction (DAS) rule, considering only words that differ by a single phoneme from the stimulus word: competitors of "beat" /bit/ include "be" /bi/ and "bleat" /blit/ and "meat" /mit/ but not "need" /nid/.

Given that HMMs allow distance computation between any two words, we can consider any subset of the lexicon as possible competitors for the stimulus word. To evaluate the extent to which activation spreads through the lexicon, we will include as the set of competitors all monosyllabic words (as the FWNPs in Luce & Pisoni's 1998 implementation of the NAM [28] did); all DAS competitors (as simplifications of the NAM do); all words with the same consonant/vowel pattern as the stimulus (e.g., all CVC competitors for a CVC stimulus); or the full lexicon. Though we can use HMMs to compute distances to any competitor in principle, we are constrained by the available data; given the confusion data that we are using, HMMs report zero confusability for much of the full lexicon. (Namely, we do not have data for vowel hallucinations, so we can compute nonzero confusability from a CVC stimulus only to other univocalic words, including polysyllabic words which contain syllabic consonants—like comparing "beat" /bit/ to "beatle" /bidl/ or "buttons" /bʌtnz/. Given our confusion data, HMMs report a zero distance from any stimulus to any word with more vowels than the stimulus.) See Figure 2.

The words that are included as competitors make implicit assumptions about the extent to which lexical activation spreads through the lexicon, but these assumptions have not been tested. If the spread of lexical activation is relatively limited, measures that include a small subset of the lexicon as possible competitors will be expected to perform just about as well as measures that use a larger

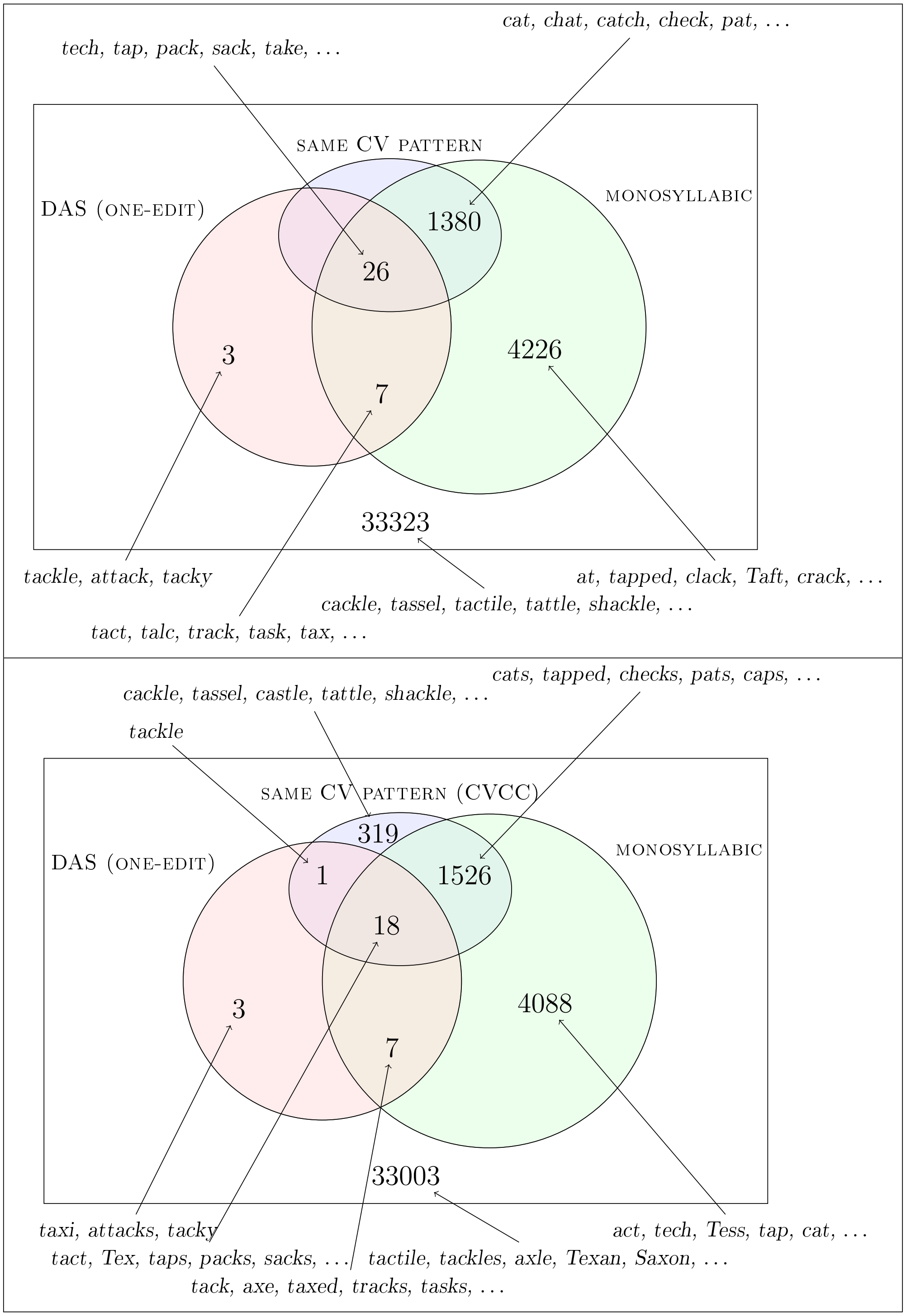
cat, chat, catch, check, pat, . . .
tech, tap, pack, sack, take, . . .
same CV pattern
DAS (one-edit)
monosyllabic
1380
26
3
4226
7
33323
tackle, attack, tacky
at, tapped, clack, Taft, crack, . . .
cackle, tassel, tactile, tattle, shackle, . . .
tact, talc, track, task, tax, . . .
cats, tapped, checks, pats, caps, . . .
cackle, tassel, castle, tattle, shackle, . . .
tackle
same CV pattern (CVCC)
319
DAS (one-edit)
1
1526
monosyllabic
18
3
4088
7
33003
taxi, attacks, tacky
act, tech, Tess, tap, cat, . . .
tact, Tex, taps, packs, sacks, . . .
tactile, tackles, axle, Texan, Saxon, . . .
tack, axe, taxed, tracks, tasks, . . .

Figure 2: Top panel, the closest competitors for the word "tack" divided into those competitors that are within a single deletion/addition/substitution, those competitors that are monosyllabic, those that have the same consonant/vowel pattern as the stimulus (i.e., CVC), and the other 33,323 competitors not in any of these categories. In each category, the number of competitors and the most similar words to "tack" are shown (according to our HMM-based p_{all} measure of similarity; see below). Bottom panel, the corresponding competitors for "tacks."

subset or the full lexicon. However, there is some reason to believe that the DAS rule is overly restrictive in designating which words are activated during word recognition; 60% of recognition errors for monosyllabic words differ from the stimulus word by more than one phoneme [15]. In addition, words with no DAS neighbors can still show effects of lexical competition; high phonological Levenshtein distance (the mean number of phoneme changes necessary to transform the word into its 20 nearest neighbors) impairs recognition [45] (see Discussion for more on Levenshtein distance). Because HMMs allow us to compute the similarity of any two words, even those that differ by a large number of phonemes, using them will allow us to help to evaluate whether a larger region of the lexicon is activated by stimulus words than is represented by the current computations of the NAM or the DAS rule.

Aim 4. Evaluating the source of lexical competition.

Recall that $FWNP(w)$ is the ratio between the (frequency-weighted) support for w and the (frequency-weighted) support for all words, including w and all competitors w'. The competition component of the FWNP's denominator simply sums the competition from all competitors. In doing so, it loses information about whether the majority of the competition came from a few highly similar competitors or a larger number of less similar competitors. For example, “bone” and “doom” have reasonably similar frequencies of occurrence, SWPs, and FWNP values, but they differ in the source of the competition. (See Figure 3.)

Strand [44] demonstrated that the dispersion of competition had a small but significant effect on word-recognition accuracy after controlling for the amount of overall lexical competition; stimulus words like “bone” whose competitors are more tightly clustered around the mean level of the stimulus's competition were recognized more accurately than words like “doom” that have more high-competition competitors (and thus more low-competition competitors too). In the current study, we focus on the extremes of these distributions of competition by evaluating the influence on recognition of the proportion of the total competition a word encounters that comes from its closest competitor.

Methods

The NAM and HMM calculations require measures of phoneme confusability as input. To test the predictions of the models, we also require measures of word-recognition accuracy.

Phoneme-identification data

Phoneme-identification data were obtained from an existing dataset [26, 28]. The stimuli consisted of 25 consonants (b, tʃ, d, f, ɡ, h, dʒ, k, l, m, n, ŋ, p, r, s, ʃ, t, θ, ð, v, w, j, z, ʒ) and 15 vowels (i, ɪ, ɛ, eɪ, æ, a, ɔ, aʊ, aɪ, ʌ, ɔɪ, ɔʊ, ʊ, u, ɛ˞). The consonant task also included a “null” condition ([28], described above) in which a vowel was presented in isolation but participants were given the opportunity to report that a consonant had been presented in addition to the vowel. The null condition renders values for perceptual hallucinations and omissions that enable us to quantify the similarity of words with differing lengths. The vowel data set included the rates at which participants failed to report the vowel they heard (representing omissions), but did not include a null stimulus category, so it is not possible to estimate the likelihood of vowel hallucinations. Thus, vowel-hallucination probabilities are set to zero in the present work. (For the implications of this

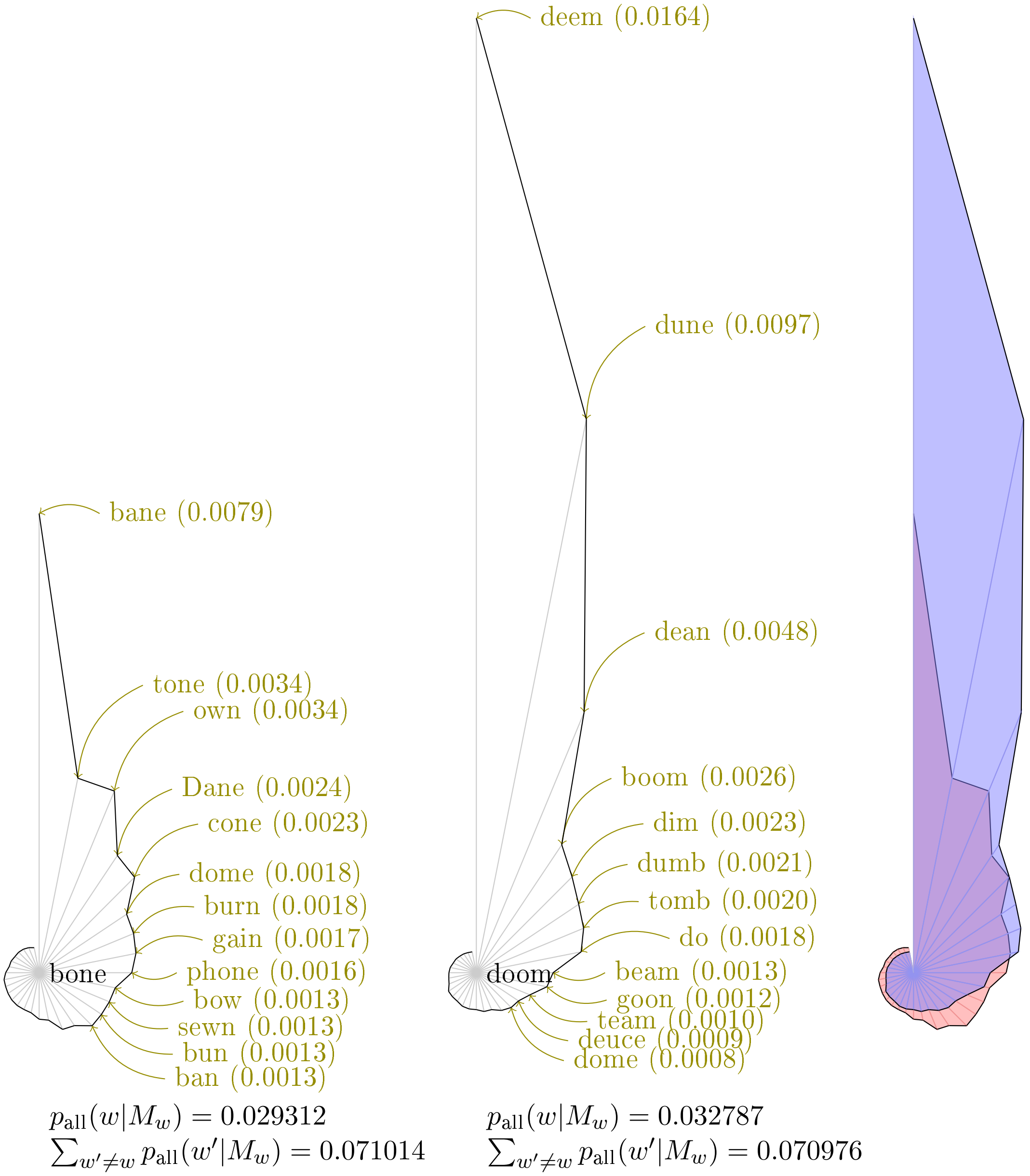

Figure 3: The distribution of competitors for the words "bone" and "doom." In the left and center panels, the 32 closest competitors for both words are shown in descending order of similarity, moving clockwise from the top; the length of the line for competitor w' is $p(w'|w)$. The 13 closest competitors for each word are labeled, along with their confusion probabilities (under the HMM-based p_{all} measure); for context, the p_{all} probability of getting the word right is also shown. In the right panel, the two distributions are superimposed, illustrating that substantially more of the competition for "doom" comes from the top 8 competitors, while "bone" has competition that is much more evenly distributed throughout the 32 competitors.

zero probability, see the paragraphs marked “Potential limitations in quantifying competition” in the Discussion.)

Luce & Pisoni [28] presented participants with phonemes to identify at three different signal-to-noise (SNR) ratios, −5, 5, and 15. Prior work has shown that although SNR affects the accuracy with which phonemes are identified, it does not systematically change the types of confusions that are made (see [33] for evidence that consonant confusions are consistent across SNRs). Therefore, in the current study, confusion matrices were collapsed across the three matrices to increase the number of data points in each cell and reduce the effects of small idiosyncrasies in an individual confusion matrix. Luce & Pisoni [28] also differentiated between consonant-initial and consonant-final confusions. However, this classification becomes problematic when comparing words of longer lengths; for example, should the “r” of “carry” be treated as a final consonant (because it follows “a”) or an initial consonant (because it precedes “y”)? In the current study, we collapsed across these two categories to create a single context-independent quantification of the confusability of consonant pairs.

Word recognition methods

Participants

53 native English speakers with self-reported normal hearing and normal or corrected-to-normal vision were recruited from the Carleton College community. Carleton College’s Institutional Review Board approved the research procedures.

Stimuli & Procedures

Stimuli included 400 CVC words selected from the English Lexicon Project (ELP) [5]. The talker was a Midwestern female, and recording was done at 16 bit, 44100Hz using a Shure KSM-32 microphone with a pop filter. Stimuli were equated for RMS amplitude using Adobe Audition, version 5.0.2, and presented in isolation through Seinheisser HD-280 headphones in background noise (six-talker babble), set at 65dB SP at SNR of 0. Participants were seated in a quiet room at a comfortable distance from a 21.5” monitor. Stimulus presentation was controlled with Superlab, version 5. Participants identified stimuli by typing their responses on a keyboard. They were encouraged to guess when unsure. Prior to analysis, recognition responses were hand-checked for obvious entry errors, such as a superfluous punctuation mark (e.g., “soup[”). Entry corrections accounted for approximately 1% of responses. No other deviations from the stimulus word (plurals, inflected forms) were counted as correct. This dataset has been reported on previously by [42].

Quantifying lexical competition

A lexicon of potential competitors was constructed from a list of 40,411 English words from the ELP [5], with two modifications. First, we replaced each /ɜ/ by /ʌ/ to make pronunciations consistent with the phoneme-identification data [26, 28]. Frequency counts were derived from the Subtlex norms [8]. Second, we combined entries for homophones, where the combined entry’s frequency count was the sum of the relevant raw frequency counts. We denote by $\mathit{freq}(w)$ the log (base 10) of the raw frequency counts with add-one smoothing for each of the resulting 38,967 words.

For each stimulus word w, we constructed an HMM M_w, as described previously. Then, for each potential competitor w' from the lexicon (including for $w' = w$ itself), we compute two probabilities:

Variable	Mathematical definition	Description
intel_all	$p_{\text{all}}(w\|M_w)$	probability that M_w generates w, along any path
intel_max	$p_{\max}(w\|M_w)$	probability that M_w generates w, along the most probable (correct) path
conf_all	$\sum_{w' \neq w} \left[p_{all}(w'\|M_w) \cdot \mathit{freq}(w')\right]$	probability that M_w generates any other word, weighted by the (log) frequency of the competitor.
NN_ratio	$\frac{p_{\text{all}}(NN(w)\|M_w) \cdot \mathit{freq}(NN(w))}{\texttt{conf\_all}}$	fraction of (log frequency–weighted) confusability conf_all coming from the largest term (that is, the competitor w' with the largest value of $p_{\text{all}}(w'\|M_w) \cdot \mathit{freq}(w')$.

Table 1: Variables derived from HMMs.

	$\sum_{w' \in S} \left[p_{all}(w'\|M_w) \cdot \mathit{freq}(w')\right]$
conf_all	S = entire lexicon (except w itself)
conf_mono	S = all monosyllabic words (except w itself)
conf_DAS	S = all words that are one deletion/addition/substitution away from w
conf_CV	S = all words with the same consonant/vowel pattern as w (except w itself)

Table 2: Variations on the conf_all measure, based on the subset of the lexicon included.

1. the probability that the HMM M_w generates w' as output, summed across all possible ways of M_w generating w' (that is, summed across all possible paths through the machine). We denote this quantity by $p_{\text{all}}(w'|M_w)$.

2. the probability that the HMM M_w generates w' as output, along only the single most probable path. We denote this quantity by $p_{\max}(w'|M_w)$.

We also define the *nearest neighbor (NN)* of w based on the p_{all} values:

$$NN(w) := \arg\max_{w' \neq w} \left[p_{\text{all}}(w'|M_w) \cdot \mathit{freq}(w')\right],$$

i.e., $NN(w)$ is the competitor w' of w that is most likely to be generated by the HMM M_w, weighted by frequency. Using these raw probabilities, we compute several measures for each stimulus word w. (See Table 1.)

We also consider restrictions to the set of competitors for a stimulus word w, giving us three variations of the conf_all measure. (See Table 2.)

Finally, we also compute two measures using the original NAM model of distance (SWP and NWP), instead of the HMM-based p_{all}. (See Table 3.) The code (written in Python) used to generate all of these measures, along with values for each variable described here for all English CVCs, are available at <http://go.carleton.edu/StrandLab>.

Variable	**Mathematical definition**
intel_NAM	$SWP(w) := \prod_{i=1}^{n} p(w_i \vert w_i)$
conf_NAM	$\sum_{w' \in S} NWP(w' \vert w) \cdot \mathit{freq}(w')$, where $NWP(w' \vert w) := \prod_{i=1}^{n} p(w'_i \vert w_i)$ and $S =$ all monosyllabic words (except w).

Table 3: Variables calculated using the original NAM methods.

Results

Mixed-effect models with a binomial distribution were created in R [36] and the R packages *lme4* and *languageR* (see [3]) were used to evaluate the influence of multiple lexical predictors on the criterion variable, word-recognition accuracy ("correct" vs "incorrect"). For each lexical variable, by-participant random slopes were included when they improved the fit of the model. Variables were centered around their means, and likelihood ratio tests [4] were used to evaluate whether models of increasing complexity provided better fits for the data. Excluding words that were skipped by participants and those that were not presented due to experimental error (6% of trials) resulted in 19,860 observations. An equivalent statistic to R^2 does not currently exist for evaluating the fit of a logistic model, and values derived by the multiple measures of proposed pseudo-R^2 can vary significantly [25]. Thus, assessing model fit for these type of data is notoriously difficult [19]. To evaluate the relative improvements caused by adding new variables to the model, we report the reduction in Akaike's Information Criterion (AIC). AIC reductions caused by novel variables can be contextualized by comparing them to AIC reductions from well-established variables, such as word frequency.

Baseline model

A model that contained only subjects and items as random effects was first constructed to serve as a comparison point for other models. As would be expected, adding centered, log-transformed frequency values [8] as a fixed factor significantly improved the fit of the model, $\chi^2(1) = 32.97, p < .001$, AIC reduction = 30.9; higher word frequency was associated with greater likelihood of word-recognition accuracy.

Given that frequency was entered as a control predictor, by-subject random slopes for frequency were not included (see [6] for more on this issue). See Table 4.

The multiple measures of predicted intelligibility and lexical competition are derived from the same confusion matrix with small modifications, resulting in high correlations among some of the measures. When correlated variables are simultaneously entered in the same model, the high degree of collinearity can complicate evaluating the unique contribution of each measure [17]. Therefore, in this case, a model that includes fixed factors X1 & X2 combined is compared to a model that includes only fixed factor X1, and one that includes only fixed factor X2. Given the collinearity between the measures, the coefficient estimates of such a model are not interpretable, but the likelihood ratio tests can reveal whether including multiple fixed factors explains additional variance beyond the single factors. The analyses below evaluate the influence of multiple lexical variables, and are

```
Fixed effects     Estimate       SE          z value           p
(Intercept)       .24            .09         2.65              .008
frequency         .45            .08         5.85              <.001

AIC          BIC          logLik           deviance         df.resid
22253.1      22284.7      -11122.5          22245.1         20174
```

Table 4: Summary of baseline model with frequency.

compared to the frequency-only model.

Analysis 1: Measuring perceptual intelligibility

Our first aim was to evaluate the predictive power of multiple measures of word intelligibility. These included `intel_NAM`, `intel_all`, and `intel_max` centered around their means. First, the effect of each measure was evaluated individually, by adding it as a fixed factor to the baseline model that included frequency as a fixed effect and subjects and items as random effects.

When entered individually, all measures of predicted intelligibility improved the fit of frequency-only model, `intel_NAM`: $\chi^2(1) = 61.88, p < .001$, AIC reduction $= 55.9$; `intel_all`: $\chi^2(1) = 61.34, p < .001$, AIC reduction $= 55.9$; and `intel_max`: $\chi^2(1) = 61.88, p < .001$, AIC reduction $= 55.9$. Including by-subject random slopes for each measure of intelligibility improved the fit of the models, indicating that participants differ in the extent to which they are influenced by effects of lexical intelligibility, `intel_NAM`: $\chi^2(1) = 11.6, p < .001$, AIC reduction $= 7.6$; `intel_all`: $\chi^2(1) = 11.57, p < .001$, AIC reduction $= 7.6$; `intel_max`: $\chi^2(1) = 11.6, p < .001$, AIC reduction $=$ 7.6. Note that the magnitude of the AIC reductions caused by adding intelligibility measures were numerically larger than the well-established effect of frequency, indicating strong effects of word intelligibility. The values of `intel_NAM` and `intel_max` are perfectly correlated: for stimuli of the same length, they only differ by a linear transformation (the probability of failing to hallucinate phonemes anywhere in the word) to make the latter true probabilities. Therefore, the results are identical when `intel_NAM` is substituted for `intel_max`. Although `intel_NAM` and `intel_max` are perfectly correlated for CVC-only stimuli, a benefit of `intel_max` is that it yields scores "on the same scale" as the HMM-based confusion measures that can in principle be applied to polyvocalic words (where the align-the-vowels step of NAM is not well-defined). Therefore, `intel_max` is included in subsequent models.

To evaluate whether `intel_max` and `intel_all` contribute uniquely to the model, both predictors were simultaneously added to the frequency-only model as fixed factors. A model that included both `intel_max` and `intel_all` provided a better fit than one that included only `intel_max`: $\chi^2(1) = 7.96, p = .005$, AIC reduction $= 6.0$; or only `intel_all`: $\chi^2(1) = 8.45, p = .004$, AIC reduction $= 6.5$, indicating that `intel_max` and `intel_all` are both contributing uniquely to the model, although the size of the effect is somewhat modest compared to the more robust effect of frequency.

By-subject random slopes for both models of predicted intelligibility improved the fit to the data, `intel_all`: $\chi^2(1) = 11.54, p = .003$, AIC reduction $= 7.6$; `intel_max`: $\chi^2(1) = 11.58, p = .003$, AIC reduction $= 7.6$, but a model with by-subject random slopes for both failed to converge.

Given the high degree of collinearity between `intel_max` and `intel_all`, the estimates of the

Fixed effects	Estimate	SE	z value	p
(Intercept)	0.24	0.09	2.76	0.01
frequency	0.44	0.07	6.14	<.001
intel_max	34.04	4.70	7.24	<.001
intel_all_resid	-0.20	0.07	-2.84	.005

AIC	BIC	logLik	deviance	df.resid
22198.8	22246.3	-11093.4	22186.8	20172

Table 5: Evaluating the contributions of `intel_max` and `intel_all_resid`.

coefficients are not interpretable, so it is not clear whether the two predictors are affecting recognition in the same direction. That is, are higher values for `intel_all` and `intel_max` associated with higher or lower rates of word identification? To render the directions of the effects interpretable, we residualized `intel_all` by conducting a simple linear regression, predicting `intel_all` from `intel_max`, to generate `intel_all_resid`. This step will enable us to enter `intel_max` into the model, along with `intel_all_resid`, giving `intel_max` (the variable that has been tested previously in the literature, in the form of `intel_NAM`) first access to the shared variance. Although there are circumstances under which orthogonalizing predictor variables by residualizing one variable against another is not appropriate [47], the benefit of this approach is that it allows us to simultaneously evaluate the unique contribution of `intel_all` on explaining word recognition scores, beyond the variance explained by `intel_max`. Importantly, this analysis will generate the same results for the predictor that was not residualized (`intel_max`) as being entered in the model alone. In addition, the coefficient estimate for the residualized predictor (`intel_all_resid`) alone will be the same as when it is simultaneously included with the non-residualized predictor (`intel_max`). The residualization will not improve the overall explanatory power of the model nor any indices of model fit, but will enable interpretation of the coefficient estimates. For more detail on the consequences of residualization, see Wurm and Fisicaro [47]. The output of this model is shown in Table 5.

The sign of the estimate of `intel_max` is positive, indicating that higher intelligibility values are associated with higher accuracy, as expected. However, `intel_all_resid` is negative, indicating that higher values are associated with lower accuracy. These results suggest that the most intelligible words have high probabilities along the highest probability path, and low probability along other paths to correct identification (but via incorrect paths).

Analysis 2: Measuring perceptual confusability

Our second question was whether evaluating lexical competition between a given stimulus word and a competitor in multiple ways (e.g., the "cast|cat" example described above) would improve the predictive power of the model over the lining-up-the-vowel method. To evaluate this question, we compared the effects of `conf_mono` and `conf_NAM`. Both of these variables include the same subset of the lexicon (all monosyllabic words) but they differ in that `conf_mono` allows confusions between stimulus word and competitor in multiple ways. Models that contain only `conf_mono` or `conf_NAM` both provided a better fit than the frequency-only model, `conf_mono`: $\chi^2(1) = 51.36, p < .001$, AIC reduction = 49.4; `conf_NAM`: $\chi^2(1) = 49.29, p < .001$, AIC reduction = 47.3. A model that included both `conf_mono` and `conf_NAM` did not provide a better fit than `conf_mono` alone,

$\chi^2(1) = 0.82, p = .37$, AIC reduction = -1.2. The model with both conf_mono and conf_NAM performed slightly better than conf_NAM alone, but the difference was only marginally significant, $\chi^2(1) = 3.15, p = .08$, AIC reduction = 1.2. This result indicates that the method of quantifying lexical competition from the HMM accounts for only a marginal degree of unique variance in word-recognition accuracy beyond that explained by the original NAM method.

Analysis 3: Quantifying the spread of lexical activation

The third aim was to assess the extent to which lexical activation spreads through the lexicon. To evaluate this question, we calculated lexical competition using differing subsets of the lexicon as potential competitors: DAS neighbors only (conf_DAS), substitution-only neighbors (conf_CV), monosyllabic words only (conf_mono); and all words in the lexicon (conf_all). Each of the measures on its own added significant unique variance to the frequency-only model, conf_DAS: $\chi^2(1) = 12.89, p < .001$, AIC reduction = 10.9 , conf_CV: $\chi^2(1) = 44.83, p < .001$, AIC reduction = 42.9, conf_mono: $\chi^2(1) = 51.36, p < .001$, AIC reduction = 49.40, conf_all: $\chi^2(1) = 52.30, p < .001$, AIC reduction = 50.3.

Next, we included multiple measures of competition simultaneously. A model that included both conf_CV and conf_DAS as fixed effects provided a better fit than one that included only conf_CV, $\chi^2(1) = 12.36, p < .001$, AIC reduction = 10.3 or a model that included only conf_DAS, $\chi^2(1) = 44.29, p < .001$, AIC reduction = 42.3, suggesting that words that are an addition or deletion away from the target are providing competition. A model that included conf_mono in addition to conf_CV and conf_DAS accounted for additional variance beyond the model with conf_CV and conf_DAS, $\chi^2(1) = 19.77, p < .001$, AIC reduction = 17.8. Finally, including the full lexicon (conf_all) in addition to conf_CV, conf_DAS, and conf_mono provided a better fit than the model without it, $\chi^2(1) = 15.97, p < .001$, AIC reduction = 14.0, suggesting that multisyllabic words outside of the DAS neighborhood are providing competition for the stimulus words.

In order to render interpretable estimates of the coefficients, we again residualized variables. Given that the most research has been done on the effect of DAS neighbors, we did simple linear regressions on conf_CV (removing the variance explained by conf_DAS, conf_mono, and conf_all), conf_mono (removing the variance explained by conf_DAS, conf_CV, and conf_all), and conf_all (removing the variance explained by conf_DAS, conf_CV, and conf_mono). Then, we entered conf_DAS into a model, along with conf_CV_resid, conf_mono_resid, and conf_all_resid. This analysis will render an estimate of the coefficient for conf_DAS that is the same as conf_DAS being entered in alone. (See Table 6.)

All four measures have negative estimates, indicating that competition from all subsets of the lexicon is negatively correlated with word-recognition accuracy.

Analysis 4: Evaluating the source of lexical competition

The fourth aim was to better understand how the distribution of competition across competitors influences recognition. Do recognition rates differ between words whose competition comes primarily from a single, highly similar neighbor, versus coming from a larger number of less similar neighbors? To that end, we first built models that include only the total competition (conf_all) and the proportion of total competition that comes from the single competitor that provides the most competition (NN_ratio). Conf_all provided a better fit than the frequency-only model, $\chi^2(1) =$

```
Fixed effects     Estimate       SE          z value            p
(Intercept)       0.24           0.08        2.91               .004
frequency         0.54           0.07        7.84               <.001
conf_DAS          -5.70          1.41        -4.03              <.001
conf_CV_resid     -1.10          0.16        -7.07              <.001
conf_mono_resid   -14.55         1.75        -8.32              <.001
conf_all_resid    -14.70         1.74        -8.47              <.001

AIC          BIC          logLik           deviance         df.resid
22168.1      22231.4      -11076.1         22152.1          20170
```

Table 6: Evaluating the contributions of `conf_DAS`, `conf_CV_resid`, `conf_mono_resid`, and `conf_all_resid`.

```
Fixed effects     Estimate       SE          z value            p
(Intercept)       0.25           0.09        2.85               .004
frequency         0.53           0.07        7.30               <.001
conf_all          -8.41          1.09        -7.76              <.001
NN_ratio          2.59           1.00        2.58               .01

AIC          BIC          logLik           deviance         df.resid
22198.2      22245.6      -11093.1         22186.2          20172
```

Table 7: Evaluating the influence of `conf_all` and `NN_ratio`.

$52.30, p < .001$, AIC reduction = 50.3; but `NN_ratio` did not, $\chi^2(1) = 2.73, p = .10$, AIC reduction = .08.

To evaluate the independent contributions of the proportion of competition from the nearest neighbor and the competition from the other neighbors, we tested a model that included both `conf_all` and `NN_ratio`. This model provided a better fit than `conf_all` alone, $\chi^2(1) = 6.59, p = .01$, AIC reduction = 4.6; or `NN_ratio` alone, $\chi^2(1) = 56.16, p < .001$, AIC reduction = 54.1. Because `conf_all` and `NN_ratio` are not correlated (r = .08), it was not necessary to generate a residualized measure to test the direction of the effects as in the previous analyses. Thus, we report the summary of the model with `conf_all` and `NN_ratio` entered simultaneously in Table 7.

These results indicate that the amount of competition that comes from the nearest neighbor accounts for a small but significant amount of unique variance in word-recognition accuracy, beyond that explained by the total amount of competition. After controlling for the total amount of competition, words with a highly similar neighbor (those with a larger fraction of their competition coming from a single competitor) are easier to recognize than those with less competition coming from the closest competitor.

Discussion

In this study, we report a novel method for generating measures of lexical activation and competition using HMMs within the framework of the NAM. The measures account for significant unique variance in spoken-word recognition accuracy beyond that explained by existing methods, and allow us to address novel questions about the dynamics of lexical competition.

Measuring perceptual intelligibility.

First, measures of perceptual similarity of word pairs that include multiple possibilities for confusability are richer predictors than those that compute similarity in only one way. Our results indicate that words that have a single clear path to recognition are recognized most easily, and that having multiple routes to correct recognition actually hinders the listener. This metric of "multiple methods of correct recognition" is likely quantifying how perceptually unique a given phoneme string is and serves as a more flexible measure of a particular stimulus word's intelligibility.

Measuring perceptual confusability.

When the number of lexical entries that are allowed to compete is held constant (at all monosyllabic words), calculating similarity using HMMs accounted for only marginally significant unique variance beyond the original NAM method. Given that these measures are constructed from the same confusion matrices and are highly correlated, the only very small difference between HMM- and NAM- derived measures may not be surprising. The key difference between the HMM and NAM measures is that the HMM measures produce scores that satisfy the mathematical definition of a probability function; HMMs also allow for multiple ways for the words to be confused, rather than considering only the line-up-the-vowel path to confusion. However, particularly for CVCs, the line-up-the-vowel method is likely to render the largest values for most word pairs; the "cast|cat" situation is relatively rare in the lexicon. However, as words get more complex, they will provide more opportunities for complex confusions. Longer words tend to have fewer neighbors, making DAS measures generally less informative. However, Suarez et al. [45] demonstrated effects of lexical competition from nearby words even for targets with no direct DAS neighbors (see discussion of Levenshtein distance below). Future research should evaluate whether our HMM-based quantifications of lexical competition predict recognition accuracy for longer words more accurately than Levenshtein distance metrics in the style of Suarez et al. [45].

Quantifying the spread of lexical activation.

The third analysis revealed that a larger subset of the lexicon is activated during spoken-word recognition than the DAS shortcut method would predict. Although the original implementation of the NAM included all monosyllabic words as potential competitors for each stimulus word, it is common in the literature to quantify lexical competition using only words within a one-phoneme radius of the stimulus word. Given that adding the influence of `conf_all` significantly improved the model after controlling for `conf_CV`, `conf_mono`, and `conf_DAS`, the results indicate that even perceptually distant words are simultaneously activated and therefore provide competition for stimulus words. The finding that adding any of the other confusability measures to `conf_CV` improves fit also provides further empirical support that words of differing CV structures also provide competition, as would be predicted by the NAM.

The finding that words that are several phonemes removed from the target influence recognition complements work using phonological Levenshtein distance [45] that demonstrated that lexical competition effects can emerge even for words with no direct DAS neighbors. Phonological Levenshtein distance is typically calculated as the average number phoneme substitutions, additions, or deletions required to turn the target word into its 20 closest neighbors in a lexicon [45]. Measures of phonological Levenshtein distance assign an equivalent "cost" to each phoneme change while HMMs calculate the differences between a word and a competitor continuously. However, the predictive power of phonological Levenshtein distance and the current HMM work suggest that a larger subset of the lexicon is activated during spoken-word recognition than the DAS shortcut method implicitly assumes. Although measures of Levenshtein distance quantify lexical competition more sensitively than simply counting DAS neighbors, HMMs are able to quantify the effects of even more distant competitors, which the current results suggest may be informative. In addition, HMMs avoid specifying an arbitrary competitor count (20 words) that makes up the competitor set. Given that orthographic Levenshtein distance predicts visual word recognition accuracy [48], future studies should seek to evaluate whether HMMs could also be applied to model visual word recognition.

The finding that perceptually distant words affect recognition is at odds with the Shortlist model [34], which posits that the process of lexical competition occurs within a small ("shortlist") of lexical items that closely match the bottom-up input. TRACE [32] and PARSYN [27], however, propose that any word candidates that match a portion of the speech input are activated. Therefore, these models could, in principle, account for our analyses' suggestion that words that are perceptually distant from the target (e.g., those that share only one feature of one phoneme) may receive a small amount of activation, and therefore provide competition for the target word.

Evaluating the source of lexical competition.

The fourth analysis revealed that the source of the lexical competition—that is, whether competition comes primarily from one frequent, highly similar competitor or from a greater number of moderately similar competitors—influences spoken-word recognition. Words that have a higher proportion of their competition coming from a single neighbor (those with high values for `NN_ratio`) were recognized more accurately than those whose total competition is more evenly distributed across multiple neighbors.[4] This result suggests that recognition is facilitated when a target word may be easily discriminated from most other words in the lexicon and is left to compete primarily with one highly activated competitor.

In contrast, Strand [44] found that greater dispersion in the distribution of competitors impaired word-recognition accuracy after controlling for the amount of overall lexical competition. That is, words that have a greater proportion of their competition originating from highly similar words are recognized less accurately than those whose competition is more uniform across competitors. That finding is seemingly difficult to reconcile with the current result, but several methodological differences complicate direct comparisons between the two studies: Strand [44] used an alternate metric for calculating NWPs and assessed the dispersion of the distributions, rather than the proportion of competition from the nearest neighbor. Future studies should seek to elucidate the apparently complex relationship between the distribution of competitors and word-recognition accuracy.

Although the two results seem to pull in opposite directions, both Strand [44] and the findings

[4] Follow-up analyses confirm that this finding is not the result of collinearity between NN_ratio and frequency or predicted intelligibility.

reported here agree that the shape of the distribution of competitors in lexical space, and not just the total amount of competition, influences word-recognition accuracy.[5] Critically, metrics that combine the influence of all competitors by a simple sum (as FWNPs do) neglect the influence of the variation in competitor distance. The fact that the distribution of competition influences recognition is difficult to explain using existing models of word recognition. Spoken-word recognition models such as TRACE [32] and PARSYN [27] assume that the degree of competition or inhibition is based on all the activated lexical representations and produce patterns of inhibitory and facilitative effects among competitors based on how and when they are activated (cf. [9, 10]). However, they do not include a mechanism that is sensitive to the distribution of those inputs. That is, the input integration mechanism of the TRACE, which calculates the activation of the target based on the activation and inhibition from other words, does not include a mechanism for incorporating the shape of the distribution. Although it is possible that these effects may be represented by patterns of interactive activation among competitors that are also competitors of one another, it is an open question whether TRACE could account for the present findings.

Potential limitations in quantifying competition.

There are several limitations to the current computations that may lead to improved predictive power if they are addressed. Some of these limitations are purely the result of limitations of the phoneme-confusion data, while others are more structural.

Our first, and most pervasive, data-based limitation is that the phoneme-confusion data from Luce & Pisoni [26, 28] included null-response categories for consonants (and data for participants' reports of consonants even when none was presented), but not for vowels. Thus we have no estimate for the probability of perceiving a vowel when none was presented, and therefore our vowel hallucination probabilities are all set to zero. This zero-probability setting means that our HMM-based measures implausibly report zero probability of confusion of any competitor with strictly more vowels than the stimulus word (e.g., "polite" for "plight"). Rerunning Analysis 3 with a better estimate of `conf_all` to include this missing data could shed more light on the spread of activation throughout the lexicon.

A second data-based limitation is that, in the current study, phoneme-confusion rates were derived from phoneme-identification tasks consisting of identifying individual phonemes embedded in a consistent phonemic context. However, it is likely that the patterns of confusion observed would vary in a different phonemic context, due to coarticulation or variation of perceptual salience in different acoustic contexts. Thus a single phoneme-to-phoneme confusion probability may not accurately represent the confusability of phoneme strings that appear in different contexts in different real words. As an extreme example, it may be that a given pair of vowels is very confusable when they follow a stop consonant, but if they follow a fricative consonant then they are clearly distinguishable. A similar limitation occurs for hallucinations: the current HMMs model the cost of inserting a given phoneme as identical regardless of what phonemes are adjacent. An alternative phrasing is that single phoneme identifications do not allow us to assess the similarity of individual phonemes and phoneme clusters. For example, our HMMs would compute "cast" and "cats" as equally similar to "cat." However, it is certainly possible that "st" and "t" are more confusable than "ts" and "s."

[5] An anonymous reviewer also points out that the two small effects that pull in opposite directions are comparatively weak, and may be explained more parsimoniously by claiming that there, in fact, is no effect; therefore, future studies should attempt to replicate and expand these findings to assess whether and how the shape of the distribution influences word recognition.

It is an advantage over the NAM's original formulation that HMMs can very naturally incorporate these distinctions—the emission probabilities of any particular state in the HMM could simply be defined based on that particular phoneme's context—but the current matrix of phoneme confusions does not include them. Thus, the current phoneme confusion data for calculating competition does not enable us to test these hypotheses.

The HMMs presented here also do not capture the fact that words unfold over time. This feature of word recognition is incorporated into TRACE [32], Shortlist [34], and Cohort [31] models, such that words that differ at onset provide less competition than those that differ at offset, a prediction that has been empirically supported [1]. Adapting HMMs to include this feature would require more substantial modification to the model's infrastructure; it is not clear how to add these features to HMMs in a mathematically coherent way. Another feature of these models that cannot be readily incorporated into HMMs is phonemic variability from coarticulation with upcoming sounds; the natural way to include context is conditioned on the previous phoneme(s) and not on upcoming phoneme(s). Future studies should explore methods to include these features in the framework of the NAM.

Clinical applications.

Because the only input necessary to calculate lexical distances is phoneme-confusion matrices, it is possible to map the topology of lexical space for population groups with specific perceptual characteristics, such as cochlear implant users or people with impaired hearing. Population-specific measures of lexical distances may also inform research on word recognition in older adults. Evidence exists that older adults may be especially impaired at recognizing words from regions of the lexicon that are perceptually dense (as quantified by the DAS shortcut method [43]). However, it is possible that older adults are less able to make distinctions between phonemic contrasts than are younger adults, due to either sensory deficits (i.e., presbycusis) or age-related cognitive changes [35]. Therefore, for any given stimulus word, there may be more words that serve as potential competitors for older adults than for younger adults, assuming that older adults will show impaired phoneme discrimination compared with younger adults. It may be that the observed interaction between age and lexical difficulty is due to age-related changes in lexical competition instead of (or, more likely, in addition to) age-related cognitive changes (e.g., inhibitory deficits) (see [38] for a discussion of how the mental lexicon changes with age).

Conclusions

The concepts of activation and competition are well supported in research on spoken-word recognition, and the NAM has proven itself to be an influential method of formalizing these concepts. The current study suggests that more flexible approaches for quantifying how activation and competition unfold may help shed light on the mechanisms underlying spoken-word recognition.

References

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

[2] E. T. Auer. The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness. *Psychonomic Bulletin & Review*, 9(2):341–347, 2002. doi:10.3758/BF03196291.

[3] R. H. Baayen. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R.* Cambridge University Press, 2008.

[4] R. H. Baayen, D. J. Davidson, and D. M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008. doi:10.1016/j.jml.2007.12.005.

[5] D. A. Balota, M. J. Yap, M. J. Cortese, K. A. Hutchison, B. Kessler, B. Loftis, . . ., and R. Treiman. The English Lexicon Project. *Behavior Research Methods*, 39(3):445–59, 2007. doi:10.3758/BF03193014.

[6] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278, 2013. doi:10.1016/j.jml.2012.11.001.

[7] M. Brand, N. Oliver, and A. Pentland. Coupled Hidden Markov Models for complex action recognition. In *Proceedings of Computer Vision and Pattern Recognition*, pages 1–6, 1997.

[8] M. Brysbaert and B. New. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990, 2009. doi:10.3758/BRM.41.4.977.

[9] Q. Chen and D. Mirman. Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, 199:417–430, 2012.

[10] Q. Chen and D. Mirman. Interaction between phonological and semantic representations: Time matters. *Cognitive Science*, 39:538–558, 2015.

[11] D. Dahan, J. S. Magnuson, and M. K. Tanenhaus. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, 42(4):317–67, 2001. doi:10.1006/cogp.2001.0750.

[12] S. R. Eddy. What is a Hidden Markov Model? *Nature Biotechnology*, 22(10):1315–6, 2004. doi:101038/nbt1004-1315.

[13] P. D. Eimas and J. D. Corbit. Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4(1):99–109, 1973. doi:10.1016/0010-0285(73)90006-6.

[14] J. Feld and M. S. Sommers. There goes the neighborhood: Lipreading and the structure of the mental lexicon. *Speech Communication*, 53(2):220–228, 2011. doi:10.1016/j.specom.2010.09.003.

[15] R. A. Felty, A. Buchwald, T. M. Gruenenfelder, and D. B. Pisoni. Misperceptions of spoken words: Data from a random sample of American English words. *The Journal of the Acoustical Society of America*, 134(1):572–85, 2013. doi:10.1121/1.4809540.

[16] E. Foulke. Listening comprehension as a function of word rate. *Journal of Communication*, 18(3):198–206, 1968. doi:10.1111/j.1460-2466.1968.tb00070.x.

[17] L. Friedman and M. Wall. Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, May 2005. doi:10.1198/000313005X41337.

[18] M. Gales and S. Young. The application of Hidden Markov Models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2007. doi:10.1561/2000000004.

[19] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[20] S. D. Goldinger, P. A. Luce, and D. B. Pisoni. Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28:501–518, 1989. doi:10.1016/0749-596X(89)90009-0.

[21] D. Jurafsky and J. Martin. *Speech and Language Processing*. Prentice Hall, 2nd edition, 2008.

[22] K. I. Kirk, D. B. Pisoni, and M. J. Osberger. Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear and Hearing*, 16(5):470–481, 1995. doi:10.1097/00003446-199510000-00004.

[23] B. L. Lambert, L. W. Dickey, W. M. Fisher, R. D. Gibbons, S.-J. Lin, P. A. Luce, …, and C. T. Yu. Listen carefully: the risk of error in spoken medication orders. *Social Science & Medicine (1982)*, 70(10):1599–608, 2010. doi:10.1016/j.socscimed.2010.01.042.

[24] B. L. B. Lambert, S. S.-J. Lin, S. Toh, P. A. Luce, C. T. McLennan, R. La Vigne, …, and J. W. Senders. Frequency and neighborhood effects on auditory perception of drug names in noise. *In Noise-CON (Vol*, 2005. 118, p. 1955). doi:10.1121/1.4781378.

[25] J. S. Long and J. Freese. *Regression Models for Categorical Dependent Variables Using Stata*. Stata Press, 2nd edition, 2001.

[26] P. A. Luce. A computational analysis of uniqueness in auditory word recognition. *Perception & Psychophysics*, 39(3):155–158, 1986.

[27] P. A. Luce, S. D. Goldinger, E. T. Auer, and M. S. Vitevitch. Phonetic priming, neighborhood activation, and PARSYN. *Perception*, 62(3):615–625, 2000. doi:10.3758/BF03212113.

[28] P. A. Luce and D. B. Pisoni. Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, 19(1):1–36, 1998. doi:10.1097/00003446-199802000-00001.

[29] J. S. Magnuson, D. Mirman, and H. Harris. Computational models of spoken word recognition. In M. J. Spivey, M. Joanisse, and K. McRae, editors, *Cambridge Handbook of Psycholinguistics*, pages 76–103. Cambridge University Press, 2012.

[30] J. S. Magnuson, M. K. Tanenhaus, R. N. Aslin, and D. Dahan. The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General*, 132(2):202–227, 2003. doi:10.1037/0096-3445.132.2.202.

[31] W. Marslen-Wilson. Functional parallelism in spoken word-recognition. *Cognition*, 25(1–2):71–102, 1987. doi:10.1016/0010-0277(87)90005-9.

[32] J. L. McClelland and J. L. Elman. The TRACE model of speech perception. *Cognitive Psychology*, 18(1):1–86, 1986. doi:10.1016/0010-0285(86)90015-0.

[33] G. A. Miller and P. E. Nicely. An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America, 27(2), 338*, 1955. doi:10.1121/1.1907526.

[34] D. Norris. Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3):189–234, 1994. doi:10.1016/0010-0277(94)90043-4.

[35] K. Pichora-Fuller. Processing speed and timing in aging adults: psychoacoustics, speech perception, and comprehension. *International Journal of Audiology*, 42(s1):59–67, 2003. doi:10.3109/14992020309074625.

[36] R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[37] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. In *Proceedings of the IEEE*, 1989.

[38] Michael Ramscar, Peter Hendrix, Cyrus Shaoul, Petar Milin, and Harald Baayen. The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, 6(1):5–42, January 2014.

[39] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, 3rd edition, 2009.

[40] H. B. Savin. Word-frequency effect and errors in the perception of speech. *The Journal of the Acoustical Society of America*, 35(2):200, 1963. doi:10.1121/1.1918432.

[41] O. Scharenborg. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336, 2007.

[42] J. Slote and J. F. Strand. Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods*, 2015. http://doi.org/10.3758/s13428-015-0599-7.

[43] M. S. Sommers. The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. *Psychology & Aging*, 11(2):333–41, 1996. doi:10.1037/0882-7974.11.2.333.

[44] J. F. Strand. Phi-square lexical competition database (phi-lex): an online tool for quantifying auditory and visual lexical competition. *Behavior Research Methods*, 46(1):148–156, 2014. doi:10.3758/s13428-013-0356-8.

[45] L. Suárez, S. H. Tan, M. J. Yap, and W. D. Goh. Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review*, 18(3):605–611, 2011. doi:10.3758/s13423-011-0078-9.

[46] M. S. Vitevitch and M. S. Sommers. The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*, 31(4):491–504, 2003. doi:10.3758/BF03196091.

[47] L. H. Wurm and S. A. Fisicaro. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, 72:37–48, 2014. doi:10.1016/j.jml.2013.12.003.

[48] T. Yarkoni, D. A. Balota, and Yap M. J. Beyond Coltheart's N: a new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15:971–979, 2008.

[49] B.-J. Yoon. Hidden Markov Models and their applications in biological sequence analysis. *Current Genomics*, 10(6):402–415, 2009. doi:10.2174/138920209789177575.