Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition

Julia F. Strand[a] and Mitchell S. Sommers
Washington University in Saint Louis, Department of Psychology, Campus Box 1125, Saint Louis, Missouri 63108

(Received 28 October 2010; revised 27 June 2011; accepted 28 June 2011)

Much research has explored how spoken word recognition is influenced by the architecture and dynamics of the mental lexicon (e.g., Luce and Pisoni, 1998; McClelland and Elman, 1986). A more recent question is whether the processes underlying word recognition are unique to the auditory domain, or whether visually perceived (lipread) speech may also be sensitive to the structure of the mental lexicon (Auer, 2002; Mattys, Bernstein, and Auer, 2002). The current research was designed to test the hypothesis that both aurally and visually perceived spoken words are isolated in the mental lexicon as a function of their modality-specific perceptual similarity to other words. Lexical competition (the extent to which perceptually similar words influence recognition of a stimulus word) was quantified using metrics that are well-established in the literature, as well as a statistical method for calculating perceptual confusability based on the phi-square statistic. Both auditory and visual spoken word recognition were influenced by modality-specific lexical competition as well as stimulus word frequency. These findings extend the scope of activation-competition models of spoken word recognition and reinforce the hypothesis (Auer, 2002; Mattys et al., 2002) that perceptual and cognitive properties underlying spoken word recognition are not specific to the auditory domain. In addition, the results support the use of the phi-square statistic as a better predictor of lexical competition than metrics currently used in models of spoken word recognition.
 [DOI: 10.1121/1.3613930]

PACS number(s): 43.71.Sy, 43.71.Es [PEI] Pages: 1663–1672

I. INTRODUCTION

A long-standing question in research on spoken word recognition has been how humans are able to map stimulus information about spoken words onto meaningful lexical representations in memory. Given the enormity of the mental lexicon [minimum estimates suggest at least 40 000 words in the average adult lexicon (Aitchison, 2003)], discriminating between the appropriate lexical item and all other items in memory is a large and complex task that surprisingly seems relatively automatic and effortless for most listeners. Many current models of spoken word recognition [Neighborhood Activation Model (NAM) (Luce, 1986; Luce and Pisoni, 1998); TRACE (McClelland and Elman, 1986), Shortlist (Norris, 1994)] explain the process by which this is accomplished using mechanisms of activation and competition (but see Marslen-Wilson, 1987 and Norris and McQueen, 2008). According to these models, acoustic-phonetic input from a stimulus word activates a set of perceptually similar lexical candidates in memory, and these lexical candidates compete for recognition. Because each perceptually similar word provides competition for the stimulus word, words with more perceptually similar competitors should be more difficult to recognize than words with fewer competitors. Activation-competition models have received much empirical support: Experiments using perceptual identification, lexical decision, and auditory naming tasks have shown that words with many competitors are recognized more slowly and less accurately than words with few competitors (Goldinger, Luce, and Pisoni, 1989; Luce and Pisoni, 1998; Vitevitch and Luce, 1998).

In addition to the findings in auditory (A-only) speech recognition, there is a growing body of work suggesting that lipreading[1] (V-only speech perception) is also a dynamic process involving activation of and competition between lexical candidates. In parallel to A-only findings, these studies show that the number of perceptually similar words influences the likelihood that a stimulus word will be successfully lipread (Auer, 2009; Feld and Sommers, 2011; Mattys *et al.*, 2002; Tye-Murray, Sommers, and Spehar, 2007). In addition, words that occur frequently in a language are more likely to be identified accurately in both A-only and V-only domains (Savin, 1963, Mattys *et al.*, 2002). These similarities between A-only and V-only spoken word recognition prompted the proposal that the process by which words are isolated in the mental lexicon is a function of their form-based similarity to other words (Mattys *et al.*, 2002).

Despite initial findings supporting parallels in A-only and V-only spoken word processing, several methodological issues cloud the comparison. The most critical of these is that different methods have been used to quantify competition in A-only and V-only domains. Only one study (Auer, 2002) has applied a measure of competition from one domain (A-only) to the other (V-only). However, this study was restricted to measures of V-only performance. Indeed, few studies to date have collected accuracy measures on both A-only and V-only spoken word presentations using the

[a]Author to whom correspondence should be addressed. Electronic mail: julia.f.strand@gmail.com

same participants, materials, and speakers in both modalities (Kaiser *et al.*, 2003; Tye-Murray *et al.*, 2007), and neither of these used a consistent method of quantifying competition that could be tailored to each modality, to reflect modality-specific competition. Although these studies are suggestive, it is necessary to use consistent presentation conditions and a consistent method for quantifying competition in both modalities to fully assess whether stimulus-based lexical distinctiveness governs word recognition in A-only and V-only word recognition.

Lexical competition has commonly been computed by identifying clusters of perceptually similar words in a given modality (e.g., Newman, Sawusch, and Luce, 1997; Sommers, 1996; Tye-Murray, Sommers, and Spehar, 2007). In A-only presentations, competitors are commonly operationally defined as any word that can be formed by the addition, deletion, or substitution of one phoneme of the stimulus. For instance, competitors [called "neighbors" in the NAM (Luce and Pisoni, 1998)] of "cat" include "cot" (a substitution), "at" (a deletion), and "cast" (an addition). Neighborhood density (number of neighbors) has been demonstrated to predict word identification accuracy: words with few neighbors are identified more quickly and accurately than those with many neighbors[2] (Kaiser *et al.*, 2003; Luce and Pisoni, 1998; Vitevitch and Luce, 1998; see also Marslen-Wilson, 1987 for an alternative account).

To quantify competition in the V-only domain, phonemes are categorized into visually similar clusters called viseme groups[3] (Fisher, 1968; Owens and Blazek, 1985; Walden, Prosek, Montgomery, Scherr, and Jones, 1977). From these viseme groups, clusters of visually similar words, called homophenes, are derived (Mattys *et al.*, 2002; Nitchie, 1926; Tye-Murray *et al.*, 2007). Words that differ only by position-specific phonemes within the same viseme groups are categorized as homophenes. For example, if /b/, /m/, and /p/ are members of the same viseme group, "bat," "mat," and "pat" are homophenes. In parallel with neighborhood density findings in A-only, words in small homophene groups are lipread more accurately than words in large homophene groups (Auer, 2002; Feld and Sommers, 2011; Mattys *et al.*, 2002; Tye-Murray *et al.*, 2007).

Making inferences about the similarity of lexical access in A-only and V-only domains using these metrics is problematic for two reasons. First, the competitors are created using different criteria. In the A-only one-phoneme shortcut method (as well as other metrics of A-only confusability), neighbors are selected to include similar, but perceptually distinguishable variations from the target (e.g., "bat" and "cat"). Homophene groups, on the other hand, are made by identifying very similar or "indistinguishable" words. If homophenes truly represent perceptually indistinguishable units and therefore cannot be discriminated based on the physical properties of the input alone, then density effects could be statistical artifacts. That is, selecting from among a small set of identical options will result in more correct answers than will selecting from among a larger set of identical options, simply by statistical chance. Under this proposal, homophenes are more similar to A-only homophones ("fair" and "fare") than they are to A-only neighbors. For example, an A-only presentation of "fair" (which is indistinguishable from an A-only presentation of "fare") would be expected to yield higher recognition scores than A-only presentation of "rode" because the former has only one auditory homophone whereas the latter has two ("road" and "rowed").

A more general, statistical limitation of the one-phoneme shortcut method and homophene grouping method is they do not incorporate perceptual similarity within a neighborhood or cluster as a factor affecting word recognition. Although some neighbors may be more confusable (perceptually similar to the stimulus word) than others, modeling competition using categorically derived neighborhoods assumes that all neighbors contribute equally to measures of neighborhood density. For instance, in the auditory domain, two words may have the same number of neighbors, but one may have many neighbors that differ by place of articulation (a feature easily lost in noise or reverberation), while the other may have a majority of neighbors that differ by voicing [a feature that is relatively resistant to interference (Binnie, Montgomery, and Jackson, 1974)]. The one-phoneme shortcut method, however, would predict similar identification performance for these two words given that neighborhood size is equivalent. Comparison of A-only and V-only lexical competition therefore requires a common metric for perceptual similarity that modulates lexical competition based on both the number and confusability of alternative lexical candidates.

An alternative to homophene grouping and the one-phoneme shortcut method (which use categorical groupings to quantify lexical competition) is to compare perceptual similarity on a continuous scale. This method has been used within the NAM using a metric called neighborhood word probability (NWP). NWPs quantify competition by comparing the perceptual similarity of a stimulus word's segments to a competitor's segments. These similarities are approximated using the probability that phonemes will be confused with one another in a forced-choice identification task. To calculate the confusability of two words, the probabilities that a stimulus word's position-specific phonemes will be confused with its competitor's position-specific phonemes are multiplied. The NWP is mathematically expressed as

$$\mathrm{NWP} = \prod_{i=1}^{n} p(\mathrm{PN}_i|\mathrm{PS}_i), \tag{1}$$

where PN_i is the ith phoneme of the neighbor and PS_i is the ith phoneme of the stimulus word. For example, the probability of responding "mad" given that the actual stimulus presentation was "bet" is given by mad|bet = p(m|b) * p(a|e) * p(d|t). This method provides a method for assessing the perceptual similarity of two words on a continuous scale. To quantify the overall competition any given stimulus word encounters within the lexicon, the NWPs comparing the stimulus word to all other words in the lexicon are summed [e.g., $p(\mathrm{word}_1|\mathrm{stimulus\ word}) + p(\mathrm{word}_2|\mathrm{stimulus\ word}) \cdots + p\ (\mathrm{word}_N|\mathrm{stimulus\ word})$]. Using an existing set of V-only phoneme confusions, Auer (2002) calculated V-only NWPs that represent the perceptual similarity (and by extension, amount of competition) of a stimulus word and an individual competitor word. Following the protocol of Luce and Pisoni

(1998), all the individual V-only NWPs were summed to quantify the total amount of competition exerted by all competitor words on the stimulus word. In parallel to the A-only findings of Luce and Pisoni (1998), Auer found that this visually based lexical density metric predicted word recognition accuracy, such that words with less competition were accurately identified more often than words with more competition.

The summed NWP method has the advantage of using a common method of quantifying competition in A-only and V-only, but a potential limitation rests in its use of probability of confusion as an estimate of perceptual similarity. Although the likelihood that two phonemes will be confused seems a reasonable proxy for how perceptually similar they are, this has a limitation: response percentages depend upon the number of perceptually similar alternatives (Iverson, Bernstein, and Auer, 1998). For instance, the phonemes /f/ and /v/ look very similar on the face, and will be confused on roughly 50% of V-only trials. The phonemes /tʃ/, /dʒ/, /ʃ/, and /ʒ/ also look very similar, so any two will be confused on roughly 25% of trials. Therefore, using the probability of confusion gives the erroneous impression that /f/ and /v/ are twice as similar as /tʃ/ and /dʒ/, despite the fact that both pairs are nearly identical.

To overcome this confound, Iverson *et al.* (1998) employed the phi-square statistic, a normalized version of the χ^2 test, which quantifies the similarity of two response distributions. It is expressed mathematically as

$$p(\mathrm{ID}) = 1 - \sqrt{\frac{\sum \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum \frac{(y_i - E(y_i))^2}{E(y_i)}}{N}}. \quad (2)$$

Here, x_i and y_i are the frequencies with which phonemes x and y were identified as category i, E_{x_i} and E_{y_i} are the expected frequencies of response for x_i and y_i if the two phonemes are perceptually identical, and N is the total number of responses to phonemes x_i and y_i. The expected values (E_{x_i} and E_{y_i}) are determined by summing the frequency with which phoneme x was identified as category i and the frequency with which phoneme y was identified as category i, divided by 2. The rationale for this method is that if phonemes x and y are perceptually identical, they should be identified as members of a given category with equal frequency. The phi-square statistic reaches a value of one when the distributions of responses for two phonemes are identical (participants select each response alternative equally for both phonemes), and reaches a value of zero when the distributions have no overlap (that is, participants did not use any of the same response categories for the two stimulus words).[4] Because the statistic compares the distributions across all response options, the magnitude of the output is independent of the number of similar alternatives.

Another advantage outlined in Iverson *et al.* (1998) to using the phi-square statistic instead of probability of confusion values is that it minimizes the influence of response biases and asymmetries in the data set. For example, a participant in a visual-only phoneme identification task may select response /b/ at a disproportionate rate for a reason that is unrelated to signal information (e.g., their name begins with /b/). In this case, the probability of confusion will result in artificially deflated similarity between /m/ and /p/ (which are visually very similar to /b/). This occurs because when /m/ or /p/ are presented, the response bias of choosing/b/ reduces the frequency with which the other option is chosen. The phi-square statistic overcomes the problems associated with response biases because it compares overall response distributions without taking into account which response options are selected.

Although Iverson *et al.* (1998) used the Phi-square statistic to establish viseme groups based on empirically derived confusion patterns, its output can be used directly as a measure of perceptual similarity. Using the phi-square statistic in this manner overcomes the limitation of using categorical groupings and obviates the measurement issues of using probability of confusion as a proxy for similarity. The phi-square statistic also provides an elegant solution to the difficulties of comparing V-only homophene groups to A-only neighborhoods described above. Given that the only input necessary to derive phi-square values is phoneme confusion matrices, perceptual similarities may be readily calculated for any pair of phonemes in any modality. This allows the opportunity to directly compare processes of lexical competition across different modalities.

Feld and Sommers (2011) used NWPs derived from probability of confusion values as well as phi-square values to predict V-only word recognition. Measures of competition based on phi-square values accounted for significant variance beyond that explained by measures of competition based on probability of confusion or by homophene group size. However, in this study, the speakers used to establish phi-square confusability at the phoneme level were not the same ones that produced the stimuli in the word recognition experiments. Because viseme and homophene groupings may differ based on speaker idiosyncrasies (Jackson, 1988), the perceptual similarity of words should be based on speaker-specific phoneme confusions. In addition, Feld and Sommers (2011) only assessed V-only identification scores, making it impossible to compare A-only and V-only identification using the same participants and stimulus materials.

The current study tested the hypothesis that words are recognized as a function of their modality-specific perceptual similarity to other words in the lexicon. To accomplish this, we quantified lexical competition using the same computational algorithm in both A-only and V-only domains, but used "modality-specific input that reflects the perceptual" properties of the input modality. Three measures of lexical competition were calculated for each modality: a categorical measure of competitor density (neighborhood density for A-only and homophene group size for V-only), a continuous measure of competitor density based on probability of confusion values, and a continuous measure of competitor density based on phi-square values. This serves as the first investigation that has assessed spoken word recognition in both A-only and V-only domains using the same participants and speakers, as well as a consistent method for quantifying competition in both modalities. These methodological controls provide the first direct test of the hypothesis that the

dynamics of spoken word recognition is similar in both A-only and V-only modalities, using analogous measures of similarity and identical stimulus materials in both auditory and visual modalities.

II. METHODS

A. Participants

Seventy-two native English speakers with self-reported normal hearing and normal or corrected-to normal vision were recruited from Washington University's undergraduate participant pool. Participants (55 female) ranged in age from 18 to 22 ($M = 19.1$, $SD = 1.07$). Testing took approximately 3 h, which was split into two 1.5 h sessions. Participants were awarded course credit for their participation, and all procedures were approved by the Human Research Protection Office of Washington University in St. Louis.

B. Stimuli

To select the stimulus words, a corpus of all consonant-vowel-consonant (CVC) words in English (excluding proper nouns and taboo or profane words and including homophones only once) was compiled, using the English Lexicon Project (ELP) (Balota *et al.*, 2007). From this corpus, two lists of 180 words were selected and randomly assigned to be used in the A-only and V-only conditions. The two lists were matched on Hal_{log} frequency ($p = 0.90$), mean lexical decision reaction times on the ELP ($p = 0.67$), number of substitution-only phonological neighbors ($p = 0.54$), and number of addition, substitution, or deletion phonological neighbors ($p = 0.94$). Additionally, the lists were checked against one another to ensure that they had equivalent numbers of each part of speech and similar representations of each phoneme in each position. These analyses were conducted to ensure that the A-only and V-only stimuli were equivalent on critical measures that may influence the accuracy with which they are processed, that they are representative of English CVCs in general, and that they contain a large range of values on all variables of potential interest.

The stimuli were recordings of phonemes and words produced by six talkers (three male, three female). Each talker was instructed to read words from a teleprompter and repeat them aloud in a natural speaking voice. The speakers wore black shirts and appeared in front of a neutral gray background. They were well-lit with studio lighting to ensure good illumination of their articulators. Stimuli were recorded with a Cannon Elura 85 digital video camera connected to a Dell Precision PC and recorded at a 16-bit resolution and sampling rate of 48 000. Digital capture and editing was done in Adobe Premiere Elements 1.0. The stimuli consisted of 24 consonants (b, tʃ, d, f, g, h, dʒ, k, l, m, n, ŋ, p, r, s, ʃ, t, θ, ð, v, w, j, z, ʒ) presented in an aCa context, 14 vowels (i, ɪ, ɛ, eɪ, æ, a, aʊ, aɪ, ʌ, ɔɪ, oʊ, ʊ, u, ɝ) presented in an hVd context, and 360 CVC words.

Auditory and visual information was recorded for all stimuli, but during the identification task participants received only one modality at a time. For the V-only tasks, the speakers' head, neck, and top of the shoulders appeared in the frame, and the visual stimuli filled the 17-in. Touchsystems monitor (ELO-170C). Visual stimuli were uncompressed and interleaved in .avi format files that measured 720×480 pixels, presented at 30 frames/s. For the A-only tasks, stimuli were equated for RMS amplitude using Adobe Audition and presented in background noise (six-talker babble), set at 60dB sound pressure level. Audio stimuli were presented through a Maico MA42 audiometer over two loudspeakers orientated +/− 45° in front of the participant. Amplitude levels were checked daily to ensure calibration using a handheld sound meter (Quest Technologies model 2004 Sound Meter). Pilot testing determined that using identical SNRs for all stimuli resulted in ceiling-level performance for vowels. Therefore, consonant and word stimuli were presented at an SNR of −4 dB, whereas vowels were presented at an SNR of −12 dB (see discussion for additional consideration of these issues).

C. Procedures

Participants read an information sheet, gave verbal consent, and were seated in a double-walled sound-attenuating booth (IAC 120A) approximately 0.5 m from the computer running Superlab presentation software (Version 4.0.7b, Cedrus Corporation, 2009). Participants were presented with short audio or video clips of phonemes and words in A-only and V-only conditions. They responded to the stimuli via touchscreen button presses or keyboard input. Order of completion of the tasks was randomly determined for each participant.

1. Phoneme identification

Participants were presented with a series of audio or video clips of a speaker producing a phoneme, followed by a response screen listing each phoneme and an example word that contains it. Participants made their identification responses by touching the button with the appropriate phoneme. There were two presentations of each phoneme, spoken by each talker, resulting in 288 consonant trials and 168 vowel trials per participant, per modality. Consonant and vowel tokens were identified in separate blocks and presented in a pseudo-random order (without replacement), blocked by speaker. Before each participant performed the consonant or phoneme task the first time, an experimenter spoke aloud each of the phoneme sounds, and participants demonstrated familiarity with them by repeating them aloud in the presence of the experimenter. Participants completed practice trials that consisted of one presentation of each token by a different speaker than was used in the test trials.

2. Word recognition

Participants were presented with clips of speakers producing a CVC word presented in the carrier phrase "Say the word _____," which they identified by typing their responses on a keyboard. In each modality, 180 words consisting of six sets of 30 words were presented for identification, with each 30-item set spoken by a different talker. The word sets were counterbalanced across six participant groups

($N = 12$ in each), so that each of the words was identified for every speaker. Participants completed six practice trials in each condition spoken by a different talker than the test trials.

III. LEXICAL VARIABLES

To estimate neighborhood structure, a phonetically coded lexicon was obtained from the ELP (Balota *et al.*, 2007). This lexicon consists of 40 000 phonetically coded English words and provides a number of lexical properties for each word, including orthographic and phonological neighborhood size, word frequency, and lexical decision and naming latencies. From the 40 000 word ELP, a sublist of words was selected to serve as the lexicon for calculating all measures of competition described below. This sublist (referred to here as the ELP-CVC list) consisted of all English CVC words, excluding proper nouns and counting homophones only once.

A. Categorical measures of density

Using the ELP-CVC list, A-only neighborhood size was manually generated by identifying the number of words that can be formed by a one-phoneme substitution.[5] For example, a subset of neighbors for the word "cat" are "bat," "cot," and "cap." A more common method of defining neighbors is to include as neighbors all words that may be formed by a one-phoneme substitution, addition, or deletion. Substitution-only neighbors are reported here because they more closely parallel substitution-only homophenes in V-only. Importantly, however, all analyses described below were also conducted on the one-phoneme substitution, addition, or deletion neighbors and the pattern of results was not different.

To calculate homophene group size, viseme groupings for consonants and vowels were determined based on the procedures of Walden *et al.* (1977) and Iverson *et al.* (1998). Responses for the vowel and consonant identification tasks were collapsed across speakers and participants, rendering confusion matrices that display the frequency with which each phoneme was identified as every other phoneme in that modality. These raw frequency confusion matrices were converted to phi-square values using SPSS (SPSS for Windows, version 18.0). The V-only phi-square matrices were submitted to a hierarchical cluster analysis. This procedure generated a tree structure that grouped phonemes by confusability. At the lowest level of the structure, each phoneme is in a unique class, and at each successive level, the most similar phoneme pair is joined until, at the highest level, all phonemes belong to a single class. Viseme groupings were defined operationally as the lowest level at which 70% of responses are within viseme class. For example, when presented with /b/, /m/, and /p/, if 70% of responses are either are /b/, /m/, or /p/, they constitute a viseme group. This criteria proved too rigid for only one cluster (ɑʌ), which showed 67% within viseme response. Because this cluster was otherwise unclassifiable, the criteria were relaxed to include it. Overall, the majority of responses fell within viseme group (80% for the vowel task, 88% for the consonant task). Although viseme groupings have been published previously, (Iverson *et al.*, 1998; Owens and Blazek, 1985; Walden *et al.*, 1977), speaker idiosyncrasies can result in differing patterns of phoneme confusion (Jackson, 1988). Because of this concern, a quantitative method for building speaker-specific viseme groups has the advantage of more specifically measuring the speech patterns of a given set of talkers. The resulting viseme groups are presented in Table I. Although viseme groups vary across studies (presumably due to differences in speakers and materials), these groupings are similar to others in the literature (see Jackson, 1988 for other examples). To build homophene groups, all stimulus words and all words from the ELP-CVC list were coded into viseme groups. Next, homophenes for each stimulus word were identified by selecting all words in the ELP-CVC list that had identical viseme strings.

B. Continuous measures of density

Continuous measures of density were calculated for both A-only and V-only domains that rely on either probability of confusion values or phi-square values as the input. Probability NWPs were calculated for each stimulus word using the modality-specific probabilities from the phoneme confusion matrices following the procedure described in the Introduction [the probability of responding "mad" given "mat" $= p(\text{m}|\text{m}) * p(\text{a}|\text{a}) * p(\text{d}|\text{t})$]. For each stimulus word, the NWPs of all other words in the ELP-CVC list were summed to quantify amount of competition exerted by all other words in the lexicon [e.g., $p(\text{word}_1|$ "mad") $+ p(\text{word}_2|$ "mad"). $+ p(\text{word}_N|$ "mad")]. These sums are referred to as A-only probability density and V-only probability density (the word "probability" is included to differentiate these values from a parallel analysis using the phi-square values, described below).

To calculate phi-square density, a parallel set of calculations was conducted, following the procedure of summing NWPs, but using the transformed phi-square values as the input in place of raw probability of confusion values. For instance, the phi-square NWP of "mad" given "bet" $= \Phi^2(\text{m}|\text{b}) * \Phi^2(\text{a}|\text{e}) * \Phi^2(\text{d}|\text{t})$. For each stimulus word, the NWPs of all other words in the ELP-CVC list were summed to quantify the amount of competition exerted by all other words in the lexicon, phi-square density.

Although phi-square values have not been regularly used to quantify lexical competition previously, the metric similarities to probability of confusion values suggest that phi-square values are suitable for input to NWPs. For example, phi-square and probability of confusion phoneme matrices are highly correlated ($r = 0.88$ for A-only consonants, $r = 0.82$ for A-only vowels, $r = 0.85$ for V-only consonants, $r = 0.90$ for V-only vowels, $p < 0.01$ for all) and values are similarly bounded (range from 0–1). As a result, the distributions of phi-square density values and probability density

TABLE I. Viseme groupings.

Consonants:	{b, m, p} {f, r, v} {tʃ, dʒ, ʃ, ʒ} {g, h, k, l, n, ŋ, y}{d, s, t, z} {θ, ð} {w}
Vowels:	{i, ɪ} {e, ɛ, æ, aɪ} {ɜ˞, ʊ} {ɑ, ʌ} {oɪ, o} {u} {aʊ}

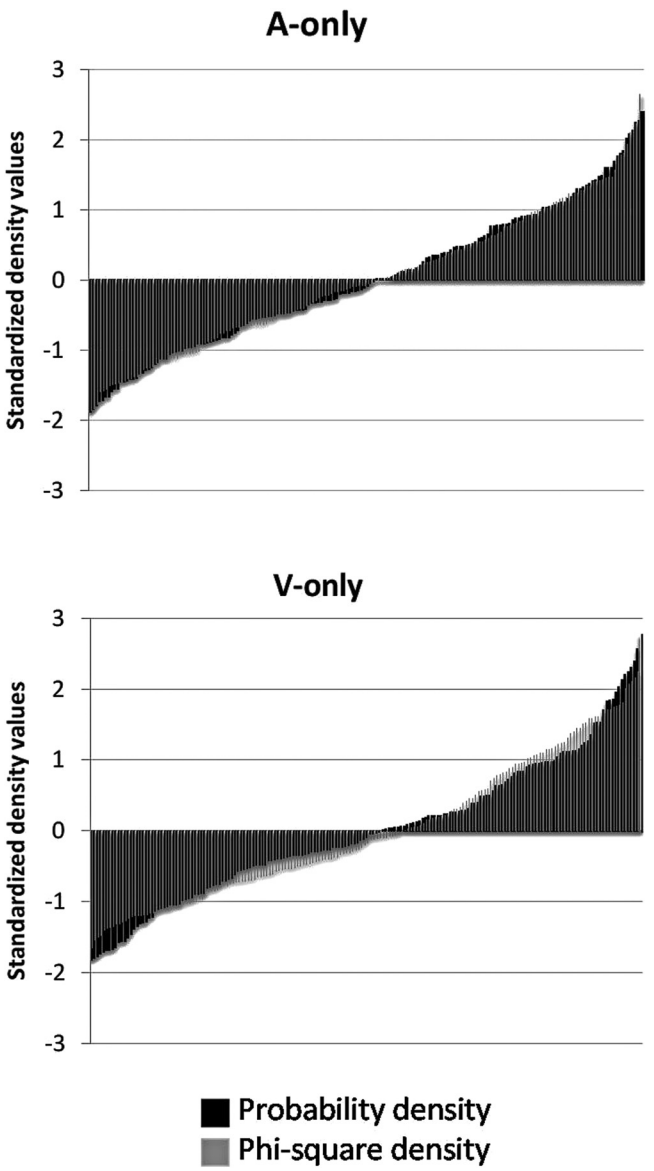

FIG. 1. Model outputs for A-only and V-only probability density and phi-square density. All values are displayed as standardized z-scores.

values are very similar (see results and Fig. 1). Therefore, the metric similarities to probability of confusion values justify the use of phi-square values as a reasonable alternative for NWP input.

IV. RESULTS

Prior to analysis, responses to the word recognition task were hand-checked for homophones and obvious entry errors. For homophonous stimulus words (e.g., "peace"), all alternate spellings were counted as correct ("piece"). For entry errors, only responses that formed nonwords were corrected, and these nonwords were only corrected in the following circumstances: the response contained a superfluous punctuation mark (e.g., "teeth]"), the response word had a letter pair reversed in a way that did not form a real word ("cheif"), the response word had a doubled letter that did not form a real word ("thiss"), or the response was misspelled in a phonetically probable way ("beed" for "bead"). These corrections accounted for approximately 1.5% of responses. No other deviations from the stimulus word (plurals, inflected forms) were counted as correct. Percent accuracy for each stimulus item was calculated and served as the criterion variable for all analyses described below. Words that were never identified accurately by any participant were excluded, to ensure exclusion of flawed or faulty stimuli. Analyses were conducted on the remaining 171 words for A-only (range: 0.01–0.86, mean accuracy = 0.30, SD = 0.19) and 149 words for V-only (range: 0.01–0.77, mean accuracy = 0.13, SD = 0.14). Regression analyses were conducted to examine the influence of lexical variables on word recognition accuracy. Descriptive statistics for the measures of competitor density are shown in Table II.

TABLE II. Descriptive statistics for measures of lexical competition.

	Range	Mean	SD
A-only neighborhood size	3–32	17.06	6.46
A-only probability density	0.06–0.49	0.25	0.10
A-only phi-square density	4.23–28.70	14.03	5.31
V-only homophene group size	0–38	12.60	10.13
V-only probability density	0.02–0.58	0.25	0.13
V-only phi-square density	1.20–39.10	16.53	9.62

A. A-only competitor density

Correlation analyses were conducted to determine the relationship between A-only word recognition and neighborhood size, A-only probability density, and A-only phi-square density. Because word frequency has been demonstrated to be a powerful predictor of spoken word recognition accuracy (Savin, 1963; Luce and Pisoni, 1998), partial correlations between measures of competitor density and accuracy, controlling for stimulus word frequency, are displayed in Table III. Stimulus word frequency values, based on the hyperspace analog to language (HAL) frequency counts reported by Lund and Burgess (1996) were taken from the ELP. All measures of competitor density were negatively correlated with recognition accuracy, indicating that words with less lexical competition were identified more accurately. Table III also reveals that all measures of competitor density were significantly correlated with one another.

In order to assess the efficacy of the measures of lexical competition at predicting A-only spoken word recognition, a series of hierarchical regressions was conducted (see Table IV). These regressions revealed that after controlling for stimulus word frequency and either neighborhood size or

TABLE III. Partial correlations between A-only word recognition accuracy and measures of lexical competition, controlling for stimulus word frequency.

	1	2	3	4
1. A-only recognition accuracy		−0.20[a]	−0.16[b]	−0.32[a]
2. Neighborhood size			0.65[a]	0.34[a]
3. A-only Probability density				0.49[a]
4. A-only Phi-square density				

[a] $p < 0.01$.
[b] $p < 0.05$.

TABLE IV. Hierarchical multiple regression predicting A-only word recognition with frequency and measures of lexical competition. Note: β represent values at the final step.

Neighborhood size and phi-square density				Probability density and phi-square density			
4A	β	R^2	ΔR^2	4C	β	R^2	ΔR^2
Step 1: Frequency	0.30	0.07	0.07[a]	Step 1: Frequency	0.30	0.06	0.06[a]
Step 2: Neighborhood size	−0.11	0.10	0.04[a]	Step 2: Probability density	−0.01	0.09	0.03[b]
Step 3: Phi-square density	−0.28	0.17	0.07[a]	Step 3: Phi-square density	−0.31	0.16	0.07[a]
4B	β	R^2	ΔR^2	4D	β	R^2	ΔR^2
Step 1: Frequency	0.30	0.07	0.07[a]	Step 1: Frequency	0.30	0.07	0.07[a]
Step 2: Phi-square density	−0.28	0.16	0.10[a]	Step 2: Phi-square density	−0.31	0.16	0.10[a]
Step 3: Neighborhood size	−0.11	0.17	0.01	Step 3: Probability density	−0.01	0.16	0.00

[a]$p < 0.01$.
[b]$p < 0.05$.

A-only probability density, A-only phi-square density accounted for significant additional variance in spoken word recognition accuracy. However, the inverse was not true: After accounting for stimulus word frequency and A-only phi-square density, both neighborhood size and A-only probability density failed to explain additional variance in A-only spoken word recognition.

B. V-only competitor density

A parallel set of analyses was conducted for V-only word recognition accuracy, using homophene group size, V-only probability density, and V-only phi-square density. Table V shows the partial correlations between V-only word recognition accuracy and measures of density, controlling for stimulus word frequency. Homophene group size, V-only probability density, and V-only phi-square density were negatively correlated with V-only spoken word recognition and were positively correlated with one another. Following the procedure of A-only analyses, a series of hierarchical regressions that include stimulus word frequency, homophene group size, V-only probability density, and V-only phi-square density is shown in Table VI. Taken together, these analyses demonstrate that Phi-square density accounts for more variance in spoken word recognition accuracy than existing measures and captures unique aspects of the lexical competition process in both A-only and V-only domains.

C. Cross-modality comparison

An important assumption of the NAM and other Activation Competition models (Luce and Pisoni, 1998) is that competitor effects are the result of the perceptual similarity of the stimulus word to its competitors. Therefore, because density depends on perceptually defined similarity within a given modality, confusions from one modality should not be expected to correlate with accuracy in another modality (see also Auer, 2002).[6] That is, the amount of competition a word encounters in the visual modality should not predict A-only recognition accuracy. Phi-square density reveals exactly this type of divergent validity: A-only Phi-square density values of words do not predict identification accuracy in V-only ($r = -0.08$, $p = 0.33$), nor do V-only density values predict A-only identification accuracy ($r = 0.11$, $p = 0.16$). Simply put, competitor density in one modality does not predict recognition accuracy in the other modality, supporting the NAM's prediction that the density effects depend on perceptually-derived similarity of the competitors.

Previous research (Auer, 2002; Mattys *et al.*, 2002) has demonstrated lexical competition effects in both A-only and V-only word recognition, but direct comparisons between the predictive power of lexical competition in the two modalities was not possible, owing to the methodological issues described previously. The use of identical stimulus materials and analogous competition measures for A-only and V-only presentations in the current investigations therefore serves as the first study capable of such direct comparisons.

The correlation between A-only phi density and A-only recognition accuracy was $r = -0.27$, and the correlation between V-only phi density and V-only recognition accuracy was $r = -0.48$ ($p < 0.01$ for both).[7] A Fisher r-to-z transformation revealed that the magnitude of these correlations were significantly different ($z = 2.19$, $p < 0.05$). Although A-only and V-only spoken word recognition are both significantly influenced by lexical competition, competition appears to have a greater influence in the visual than in the auditory modality.

It is possible that the differences in susceptibility to lexical competition observed in A-only and V-only domains are due to different accuracy distributions in the two domains (see Fig. 2). Indeed, the distributions of accuracy data differed significantly across modality (Kolmogorov–Smirnov $Z = 2.5$, $p < 0.001$). To assess whether the modality differences in distribution shape influenced the correlations with measures of lexical competition, an additional analysis was conducted. This assessed whether the modality differences in susceptibility to lexical competition persisted when the accuracy distributions were artificially equated. First, A-only and V-only

TABLE V. Partial correlations between V-only word recognition accuracy and measures of lexical competition, controlling for stimulus word frequency.

	1	2	3	4
1. V-only recognition accuracy		−0.41[a]	−0.37[a]	−0.57[a]
2. Homophene group size			0.60[a]	0.83[a]
3. V-only Probability density				0.59[a]
4. V-only Phi-square density				

[a]$p < .01$.

TABLE VI. Hierarchical multiple regression predicting V-only word recognition with frequency and measures of lexical competition. Note: β represent values at the final step.

Homophene group size and phi-square density				Probability density and phi-square density			
6A	β	R^2	ΔR^2	6C	β	R^2	ΔR^2
Step 1: Frequency	0.39	0.05	0.05[a]	Step 1: Frequency	0.37	0.05	0.05[a]
Step 2: Homophene group size	0.20	0.21	0.16[a]	Step 2: Probability density	−0.06	0.19	0.13[a]
Step 3: Phi-square density	−0.74	0.37	0.16[a]	Step 3: Phi-square density	−0.54	0.36	0.18[a]
6B	β	R^2	ΔR^2	6D	β	R^2	ΔR^2
Step 1: Frequency	0.39	0.05	0.05[a]	Step 1: Frequency	0.37	0.05	0.05[a]
Step 2: Phi-square density	−0.74	0.36	0.31[a]	Step 2: Phi-square density	−0.54	0.36	0.31[a]
Step 3: Homophene group size	0.20	0.37	0.01	Step 3: Probability density	−0.06	0.36	0.00

[a] $p < 0.01$.

accuracy data were separated into 1% accuracy bins (e.g., all words that were identified at 35% accuracy in A-only were grouped, words identified at 36% accuracy in A-only were grouped, etc). The average phi-square density value for each binned group was calculated (e.g., the mean A-only phi-square density for all words identified at 35% accuracy in A-only). Next, accuracy bins that had values for both A-only and V-only were selected (e.g., at least one word in A-only was identified at 35% accuracy and at least one word in V-only was identified at 35% accuracy). This resulted in a distribution of accuracy data that was identical for both A-only and V-only, based on a subset of the full data set. When the binned accuracy data were correlated with the mean phi-square density values of the bin, it rendered statistically significant correlations in both A-only ($r = -0.55$, $p < 0.01$) and V-only ($r = -0.77$, $p < 0.01$). Importantly, the magnitude of these correlations are still significantly different ($z = -1.42$, $p = 0.15$). This suggests that the accuracy distribution modality differences are not the cause of the modality differences in sensitivity to lexical competition.

It is also possible that the stronger correlation between V-only word recognition and phi-square density lies in the fact that different SNRs were used for the vowel identification task and the consonant and word tasks in the A-only condition. Given that pilot testing revealed very high accuracy rates for the vowel identification task at a SNR of −4 dB, it might be expected that word recognition at SNR = −4 dB would not be influenced by vowel errors. To assess this, we ran our analyses using the assumption of no vowel errors (i.e., only patterns of consonant errors were included in calculating density). Although these values predicted spoken word recognition accuracy, they were less strongly correlated with word identification accuracy than were the metrics that included vowel errors. After entering the vowel error-free measure in a multiple regression predicting word accuracy, the original metrics (that include vowel confusions) explained significant additional variance (6% additional variance in for both probability density and phi-square density, $p < 0.01$). This suggests that the vowel confusion data is informative to the model, despite the differences in presentation conditions. In addition, we found strong correlations ($r = 0.68$ to $r = 0.89$) between our vowel confusion matrix at −12 dB SNR and other confusion matrices at less demanding SNRs (Cutler *et al.*, 2004: SNR = 0 dB; Luce and Pisoni, 1986: SNR = −5 dB). This suggests that although the accuracy of responses certainly changes depending on the SNR, the overall patterns of confusion tend to remain relatively consistent.

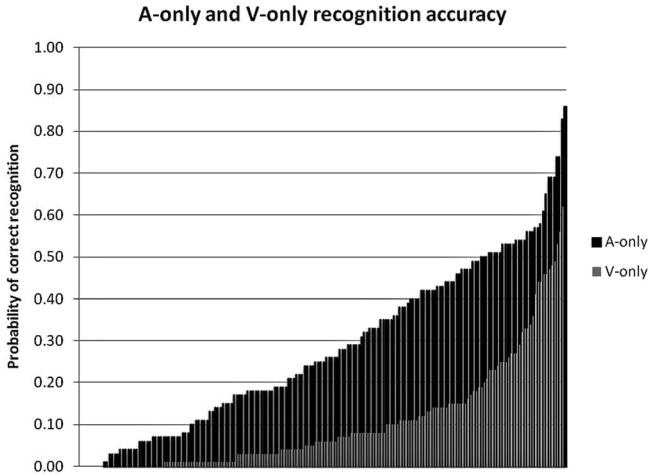

FIG. 2. Distribution of word identification accuracies in A-only and V-only domains.

V. DISCUSSION

The present study explored whether similar processes underlie spoken word recognition in A-only and V-only modalities. The results reveal that stimulus word frequency and competitor density account for significant unique variance in spoken word recognition in both A-only and V-only speech. These results support the hypothesis (Auer, 2002; Mattys *et al.*, 2002) that A-only and V-only lexical retrieval rely on similar processes. The present results are the first to use a consistent method for quantifying competition across modalities to directly assess the extent to which lexical competition operates similarly in A-only and V-only spoken word recognition. These consistent methods of quantifying competition also revealed that V-only spoken word recognition is more highly correlated with lexical density than is A-only word recognition.

Although all measures of density significantly predicted word recognition accuracy (and were all correlated with one another), measures based on the phi-square metric accounted for additional unique variance beyond that explained by other measures. This result is likely due to the fact that, unlike the one-phoneme shortcut method, the phi-square measure

quantifies competition on a continuous scale and, in contrast to the NWP (also a continuous measure of competition) measure, the phi-square measure overcomes the confounds associated with response biases and the interaction with number of response alternatives.

An alternative explanation for the predictive success of phi-square density is that it results in a distribution of output values that is more appropriate for correlation analyses than is the distribution of output values of probability density. If this were the case, the improved prediction of phi-square density could be a statistical artifact, rather than a result of better estimation of similarity at the segment level. To assess whether the distributions of probability density and phi-square density differ, the modality-specific values were standardized (see Fig. 1) and submitted to Kolmogorov–Smirnov tests for the equality of distributions. These revealed that the distributions of probability density and phi-square density were not significantly different for either A-only (Kolmogorov–Smirnov $Z = 0.42$, $p = 0.90$) or V-only (Kolmogorov–Smirnov $Z = 0.94$, $p = 0.32$) domains. This suggests that the difference in the predictive power of probability density and phi-square density is not an artifact of distribution type.

There are several issues pertaining to quantifying perceptual similarity that, if properly addressed, may further increase the predictive power of measures of lexical competition. In the current study, phoneme identification tasks consisted of identifying individual phonemes. A limitation of using single phoneme identifications as a metric for similarity is that it is not possible to assess the similarity of individual phonemes and phoneme clusters (e.g., how similar are /s/ and /st/). Some investigations (Auer, 2009) have included consonant clusters in viseme groups with single consonants (e.g., one viseme group consisted of /b/, /m/, /p/, and /pr/). If two perceptually similar consonants occur in succession, it could be that a CCVC word could be perceptually similar to, and therefore provide strong competition for, a CVC stimulus word. For example, in a visual presentation, "stop" and "top" may be easily confused. Because the current investigation only included CVC competitors (i.e., "top" is compared to "tip" and "hop", but not "stop" or "trip"), it may over or underestimate the average competition imposed on a stimulus word, based on whether it has perceptually similar words that are not CVCs. A more nuanced method that allows single phonemes to be compared to phoneme clusters might be expected to lead to greater predictive power in measures of competition.

This analysis reveals another area that should be addressed in future investigations: how to quantify the perceptual similarity of a stimulus word and a competitor word of different lengths (i.e., those that differ by the addition or subtraction of a phoneme). For example, to assess the similarity of the stimulus word "top" to a competitor such as "step," if the words are aligned at the vowel, the "t" of the stimulus word is aligned with the "t" of the competitor word. However, the /s/ of the competitor words does not align with any phoneme of the stimulus word, precluding any determination of the position-specific phoneme confusability. In this case, the probability of the phoneme /s/ is, conceptually, the probability of identifying /s/ when no phoneme was presented. Luce (1986) resolved this issue by including a "null" phoneme in the phoneme identification tasks. In some phoneme identification trials, no phoneme was presented, but participants were still forced to make a decision about what they heard. Participants also had the option of making the response that no phoneme had been presented. This enabled calculating conditional probabilities of identifying a specific phoneme when none was identified [e.g., $p(s|\emptyset)$] or the probability of failing to detect that a given phoneme had been presented [e.g., $p(\emptyset|s)$]. Using this method, Luce could calculate perceptual similarity for competitors that were longer or shorter than the stimulus words.

The method of including a null response works well for the A-only domain, when phonemes are masked and the background noise is perceptually similar to the signal. However, it is difficult to translate to the V-only modality where task difficulty stems not from similarity between signal and noise, but from an underspecified signal. The detection of a mouth movement is very salient, even if the identification of that mouth movement is difficult. Therefore, it seems extremely unlikely that a participant would ever fail to notice a speaker opening their mouth (choose the null response) or identify an unmoving face as a speaker producing a specific phoneme. Therefore, another method seems necessary for calculating the perceptual similarity of two words of differing length.

VI. CONCLUSIONS

The NAM and other activation-competition models were designed specifically to describe A-only spoken word recognition. However, the finding that V-only spoken word recognition is also achieved through lexical competition suggests that the scope of activation-competition models may be extended to other modalities of speech perception. The results suggest that similar processes of activation and competition occur for spoken word recognition whether the signal is seen or heard. It is especially interesting that measures of competition within a modality (e.g., A-only phi-square density) do not predict spoken word recognition in another modality (e.g., V-only spoken word recognition). This divergent validity suggests that competitor density is not an inherent property of a word, but rather, depends upon the nature of the perceptual signal through which it is perceived. Taken together, these results suggest that the dynamics of spoken word recognition are similar across auditory and visual input modalities, provided that modality-specific differences in perceptual similarity (and therefore competition) are taken into account.

ACKNOWLEDGMENTS

This work was supported in part by Grant No. R01 AG 18029 from the National Institute on Aging and T32 GM081739 from the National Institute of Health. Portions of this research are based on a dissertation submitted in partial fulfillment of the requirements for the Ph.D. degree by the first author.

[1]We use the term "lipreading" to refer to tasks in which no auditory information is available, whereas "speechreading" refers to processing visual speech information in the presence of auditory signals.

[2]The NAM proposes that all lexical items are potential competitors, but that activation is so low on those outside the one-phoneme alteration group that excluding them does not substantially alter predictions.

[3]The terms "viseme group" and "phonemic equivalence class" are synonymous, as are "homophene group," "lexical equivalence class," and "visual neighborhood." Here, the terms "viseme" and "homophone" are used because of their established place in the literature.

[4]In Iverson *et al*. (1998), Phi-square values were not subtracted from 1. The change is made here to allow non-zero values in calculating conditional probabilities (see below), as well as for ease of interpretation: higher numbers represent greater similarity.

[5]It is worth noting that manually generating neighborhood size resulted in values that are very similar to those available elsewhere (e.g., the ELP, Balota *et al*., 2007). The correlation between A-only density measures obtained from ELP and those generated using the above process was $r = 0.83$ ($p < 0.001$)

[6]Importantly, this rests on the assumption that perceptual similarity in the two modalities is not necessarily correlated. That is, words that have many similar competitors in A-only should not be expected to have many similar competitors in V-only. Indeed, A-only and V-only Phi-square density are not significantly correlated ($r = 0.08$, $p < 0.05$).

[7]These values are the correlation coefficients between density measures and word recognition accuracy, not controlling for frequency, whereas Tables III and IV display the partial correlations, controlling for word frequency.

Aitchison, J. (**2003**). *Words in the Mind: An Introduction to the Mental Lexicon* (Wiley-Blackwell Publishers, Oxford, UK), Chap. 1.

Auer, Jr., E. (**2002**). "The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness," Psychonomic Bull. Rev. **9**(2), 341–347.

Auer, Jr., E. (**2009**). "Spoken word recognition by eye," Scand. J. Psychol. **50**(5), 419–425.

Balota, D., Yap, M., Cortese, M., Hutchison, K., Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G., and Treiman, R. (**2007**). "The English Lexicon project," Behav. Res. Methods Instrum. Comput. **39**(3), 445–459.

Binnie, C., Montgomery, A., and Jackson, P. (**1974**). "Auditory and visual contributions to the perception of consonants," J. Speech Hear. Disord. **17**, 619–630.

Feld, J., and Sommers, M. (**2011**). "There goes the neighborhood: Lipreading and the structure of the mental lexicon," Speech Commun. **52**, 220–228.

Fisher, C. (**1968**). "Confusions among visually perceived consonants," J. Speech Hear. Res. **11**(4), 796–804.

Goldinger, S., Luce, P., and Pisoni, D. (**1989**). "Priming lexical neighbors of spoken words: Effects of competition and inhibition," J. Mem. Lang. **28**(5), 501–518.

Iverson, P., Bernstein, L., and Auer, Jr., E. (**1998**). "Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition," Speech Commun. **26**(1-2), 45–63.

Jackson, P. (**1988**). "The theoretical minimal unit for visual speech perception: Visemes and coarticulation," Volta Rev. **90**(5), 99–115.

Kaiser, A., Kirk, K., Lachs, L., and Pisoni, D. (**2003**). "Talker and lexical effects on audiovisual word recognition by adults with cochlear implants," J. Speech Lang. Hear. Res. **46**(2), 390–404.

Luce, P. A. (**1986**). "Neighborhoods of words in the mental lexicon," Research on Spoken Language Processing Technical Report No. 6, Department of Psychology, Indiana University, Bloomington, IN, pp. 1–151.

Luce, P., and Pisoni, D. (**1998**). "Recognizing spoken words: The neighborhood activation model," Ear Hear. **19**(1), 1–36.

Lund, K., and Burgess, C. (**1996**). "Producing high-dimensional semantic spaces from lexical cooccurrence," Behav. Res. Methods Instrum. Comput. **28**, 203–208.

Marslen-Wilson, W. (**1987**). "Functional parallelism in spoken word-recognition," Cognition **25**, 71–102.

Mattys, S., Bernstein, L., and Auer, E. (**2002**). "Stimulus-based lexical distinctiveness as a general word-recognition mechanism," Perception Psychophys. **64**(4), 667–679.

McClelland, J., and Elman, J. (**1986**). "The TRACE model of speech perception," Cog. Psych. **18**(1), 1–86.

Miller, G., and Nicely, P. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**(2), 338–352.

Newman, R., Sawusch, J., and Luce, P. (**1997**). "Lexical neighborhood effects in phonetic processing," J. Exp. Psychol.: Human Perception Perform. **23**(3), pp. 873–889.

Nitchie, E. (**1926**). *Lip-Reading Principles and Practise: A Hand-Book for Teachers and For Self Instruction* (Frederick A. Stokes Company, New York, NY), Chap. 1.

Norris, D. (**1994**). "Shortlist: A connectionist model of continuous speech recognition," Cognition **52**(3), 189–234.

Norris, D., McQueen, J., and Shortlist, B. (**2008**). "A Bayesian model of continuous speech recognition," Psychol. Rev. **115**(2), 357–395.

Owens, E., and Blazek, B. (**1985**). "Visemes observed by hearing-impaired and normal-hearing adult viewers," J. Speech Hear. Res. **28**(3), 381–393.

Savin, H. (**1963**). "Word frequency effect and errors in the perception of speech," J. Acoust. Soc. Am. **35**, 200–206.

Sommers, M. (**1996**). "The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition," Psychol. Aging **11**, 333–341.

Tye-Murray, N., Sommers, M., and Spehar, B. (**2007**). "Auditory and visual lexical neighborhoods in audiovisual speech perception," Trends Amplification **11**(4), 233–241.

Vitevitch, M., and Luce, P. (**1998**). "When words compete: Levels of processing in perception of spoken words," Psychol. Sci. **9**(4), 325–329.

Walden, B., Prosek, R., Montgomery, A., Scherr, C., and Jones, C. (**1977**). "Effects of training on the visual recognition of consonants," J. Speech Hear. Res. **20**(1), 130–145.