Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 220–228
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

There goes the neighborhood: Lipreading and the structure of the mental lexicon

Julia Feld *, Mitchell Sommers

Washington University in St. Louis, Department of Psychology, Campus Box 1125, St. Louis, MO 63103, United States

Received 24 June 2010; received in revised form 10 September 2010; accepted 14 September 2010

Abstract

A central question in spoken word recognition research is whether words are recognized relationally, in the context of other words in the mental lexicon (McClelland and Elman, 1986; Luce and Pisoni, 1998). The current research evaluated metrics for measuring the influence of the mental lexicon on visually perceived (lipread) spoken word recognition. 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 novel statistical method for calculating perceptual confusability, based on the Phi-square statistic.

The Phi-square statistic proved an effective measure for assessing lexical competition and explained significant variance in visual spoken word recognition beyond that accounted for by traditional metrics. Because these values include the influence of a large subset of the lexicon (rather than only perceptually similar words), it suggests that even perceptually distant words may receive some activation, and therefore provide competition, during spoken word recognition. This work supports and extends earlier research (Auer, 2002; Mattys et al., 2002) that proposed a common recognition system underlying auditory and visual spoken word recognition and provides support for the use of the Phi-square statistic for quantifying lexical competition.

Keywords: Lipreading; Spoken word recognition; Lexical competition; Lexical neighborhoods

1. Introduction

The speed and accuracy of spoken word recognition demands an efficient method of searching a highly organized mental lexicon. There is a growing consensus that the way words are organized in memory influences our ability to recognize them (see Jusczyk and Luce, 2002 for a review). Although the majority of work on the organization of the mental lexicon has been done within the realm of auditory (A-only) word recognition, there is some evidence that visually perceived (V-only) speech is also affected by lexical organization[1] (Auer, 2002; Kaiser et al., 2003; Mattys et al., 2002; Tye-Murray et al., 2007). Most current models of A-only spoken word recognition propose that stimulus input activates perceptually similar lexical candidates in memory, and that these lexical candidates compete for recognition (McClelland and Elman, 1986; Luce and Pisoni, 1998; Morton, 1979; Norris, 1994). In general, these *Activation-Competition* models propose that the acoustic–phonetic stimulus information activates representations of words in the mental lexicon, with the degree of activation depending on the degree of

---

[1] Here, the terms ‘lipreading’ and ‘visual (V-only) speech perception’ refer to tasks in which no auditory information is available. The term ‘speechreading’ is used to refer to processing visual speech information in the presence of congruent auditory signals (Tye-Murray, 2008).

* Corresponding author.
E-mail address: juliafeld5@gmail.com (J. Feld).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.09.003

similarity between the input and the memory representation (Luce and Pisoni, 1998). Because these representations then compete for recognition, words that have many perceptually similar competitors will be more difficult to identify than words with few perceptually similar competitors (Luce and Pisoni, 1998). Studies of using perceptual identification, lexical decision, and auditory naming have provided considerable support for Activation-Competition models by demonstrating that words in large neighborhoods (those with many neighbors, or perceptually similar words) are recognized more slowly and less accurately than words in small neighborhoods (Luce and Pisoni, 1998; Goldinger et al., 1989; Vitevitch and Luce, 1998).

Following work in the A-only domain showing that the speed and accuracy of word recognition depends on the lexical properties of the stimulus (e.g., neighborhood density and the frequency with which the word occurs in the language), several studies have explored whether the lexical properties of a stimulus word also influence lipread word recognition accuracy (Auer, 2002; Mattys et al., 2002; Tye-Murray et al., 2007). If spoken word recognition depends upon similar underlying processes, regardless of whether the stimulus information is perceived visually or aurally, we would expect that V-only word recognition would also be influenced by the frequency of the stimulus word and its perceptual similarity to other words in the lexicon. Indeed, stimulus words that are high in frequency are lipread more accurately than are low-frequency words (Mattys et al., 2002), and words that are perceptually similar to few other words in the lexicon are lipread more easily than those that are visually similar to many other words (Auer, 2002; Mattys et al., 2002; Tye-Murray et al., 2007).

Within the A-only domain, the amount of competition a stimulus word encounters has usually been operationalized as a measure of its acoustic–phonetic similarity to other words. Words with many similar-sounding words are harder to recognize than words with few similar-sounding words. Therefore, within the V-only domain, the amount of competition should depend upon the *visual* similarity of a stimulus word and its competitor word (i.e., the similarity of the facial and articulator movements required to produce the words). Importantly, there is reason to expect that the extent to which two words are perceptually similar may differ depending on the modality of presentation. Because of the nature of the A-only and V-only signals, phonemic contrasts that are difficult to discriminate in one modality may be easily differentiated in the other (Iverson et al., 1998). For example, acoustic cues to place of articulation for consonants are often difficult to perceive in noisy environments, but the shape of the mouth and articulatory movements that correspond to those consonants are often visually clear (Summerfield, 1992). Therefore, words that are perceptually similar in one modality may not be in the other. As a result, any given word may be expected to have different competitors in A-only and V-only domains. There is some evidence (Tye-Murray et al., 2007) that the simultaneous activation of a word's A-only and V-only competitors may reduce the number of possible word candidates in audiovisual speech, facilitating recognition performance.

In the V-only domain, the perceptual similarity of phonemes has been indexed by categorizing phonemes into visually similar groupings called *visemes*[2] (Walden et al., 1977; Auer and Bernstein, 1997). From viseme groupings, clusters of visually similar words, called *homophenes*, may be derived (Mattys et al., 2002; Tye-Murray et al., 2007). Words that differ only by position-specific phonemes within the same viseme groups are members of the same homophene group. For example, if /b/, /m/, and /p/ are categorized within the same viseme group, then "bat," "mat," and "pat" are homophenes. Although within-viseme group substitutions are possible for any word, not all substitutions result in lexically valid outcomes. For instance, "bought" will have fewer homophenes than "bat," because "mought" and "pought" are nonwords. Homophene group size has been demonstrated to influence V-only word recognition: words in small homophene groups are identified more accurately than words in large homophene groups (Mattys et al., 2002; Tye-Murray et al., 2007).

A critical question that arises in assessing lexical competition effects within any modality is how to quantify perceptual similarity and hence, competition. Most previous studies (Mattys et al., 2002; Tye-Murray et al., 2007) have elected to use discrete measures of perceptual similarity, largely because such measures are computationally more tractable than comparing a target item to a large subset of words in the lexicon using a continuous method. For example, research on auditory neighborhood density effects have defined neighbors or competitors as any word that can differ from a target item by the addition, deletion, or substitution of a single phoneme. Similarly, the few studies of visual neighborhood effects have usually defined competitors as homophenes of a target item. Although some research has described visemes and homophenes as groups of indistinguishable units (Jackson et al., 1988), recent work suggests that lipreaders may be sensitive to within viseme distinctions, such as /b/ and /m/ (Bernstein et al., 1997), suggesting that viseme groups may underestimate the perceptual information available to a lipreader.

An alternative to discrete measures of perceptual similarity that has been used is to assess similarity empirically on a continuous scale: Auer (2002), for instance, adapted Luce and Pisoni's (1998) method for assessing A-only neighborhood density to quantify V-only competition. This method uses phoneme confusions (obtained in a forced-choice identification task) as a proxy for perceptual similarity. In order to adapt the procedure to V-only speech, Auer used confusion matrices generated from a V-only phoneme

---

[2] The terms viseme group and Phonemic Equivalence Class (PEC) are synonymous, as are homophene group, Lexical Equivalence Class (LEC), and visual neighborhood. We use the terms viseme and homophene because of their established place in the literature.

identification task. From these phoneme confusion probabilities, the perceptual confusability of two words is calculated through the conditional probability of confusing the words' phoneme segments. For example, the probability of mistakenly identifying the word "bet" as the word "mad" is calculated as p(mæd|bɛt) = p(m|b) ∗ p(æ|ɛ) ∗ p(d|t). Using this method, Auer calculated the perceptual similarity between a given stimulus word and every other word in a 1500 word lexicon, derived from the PhLex online lexical database (Seitz et al., 1998). The sum of these 1500 similarity values (a prediction of the amount of competition a stimulus word encounters) correlated negatively with word recognition accuracy, indicating that words in sparse regions of the lexicon were identified more accurately than words in dense regions (Auer, 2002). Based on the similarity of these findings to earlier work in A-only speech perception (Luce and Pisoni, 1998), Auer interpreted these results as evidence for a common spoken word recognition system for both A-only and V-only speech.

Although both categorical measures (homophene groupings) and continuous measures (based on the probability of phoneme confusability) of lexical competition have had success at predicting spoken word recognition accuracy, each method has potential limitations. A disadvantage with categorical measures of density is that a continuous variable (perceptual similarity) is reduced to categorical, resulting in information loss. Phonemes within a viseme group (and, therefore, words within a homophene group) are interpreted as perceptually identical, whereas units that don't share a group are treated as completely distinct. The rigidity of this system has been questioned previously, "... the concept of [homophene group size] is a convenient simplification of the problem of spoken words' visual similarity" (Mattys et al., 2002, p. 668).

In addition to the possible methodological limitation of the homophene method, there is also a theoretical issue that warrants further examination. Although homophene groupings may have been derived as a convenient method to assess lexical competition, the underlying assumption of the method is that only perceptually similar words provide competition for a stimulus word. That is, words that fall outside of the homophene group are not assumed to exert any influence on the recognition of the stimulus word. This is a theoretical divergence from continuous measures of similarity, which include the competitive influence of a larger subset of words in the lexicon.

There may also be potential disadvantages to using probability of phoneme confusion as an estimate of word similarity. First, the number of perceptually similar alternatives may interact with the response accuracy (Iverson et al., 1998). For example, the phonemes /f/ and /v/ are visually very similar, and will, in general, be confused on approximately 50% of trials. The phonemes /tʃ/, /dʒ/, /ʃ/, and /ʒ/ are also very visually similar, so any of the four will be confused with another on approximately 25% of trials. In this latter case, the response accuracy gives the erroneous impression that /f/ and /v/ are more confusable than, for instance /tʃ/ and /dʒ/, despite equivalent visual similarity.

A second potential limitation of using probability of confusion values is that response biases and asymmetries in the data set may influence patterns of confusion (see Luce and Pisoni, 1998). For example, a participant in a V-only phoneme identification task may select response /m/ at a disproportionate rate for a reason that is unrelated to signal information (e.g., the response button for /m/ appears at the center of the screen). In this case, the probability of confusion will result in artificially deflated relationships between /b/ and /p/ (which are visually very similar to /m/).

To overcome confounds of using probability of confusion as a similarity estimate, Iverson et al. (1998) introduced the Phi-square statistic to the speech perception literature. The Phi-square statistic, a normalized version of the chi-squared test, quantifies the similarity of two response distributions and is mathematically expressed as:

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

with x_i and y_i being the frequencies with which phonemes x and y were identified as category i, $E(x_i)$ and $E(y_i)$ being the expected frequencies of response for x_i and y_i if the two phonemes are equal, and N being the total number of responses to phonemes x_i and y_i. The output reaches a value of one when the distributions of responses for two phonemes are identical (participants show the same frequency of selecting each response alternative 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 stimuli).[3] Because the statistic compares the response distributions across all categories, the magnitude of the output is independent of the number of similar alternatives. It also overcomes the problems associated with response biases because it compares overall response distributions without taking into account which response options are selected.

Fig. 1 shows a graphical representation of response distributions. The horizontal axis shows all possible response alternatives, and the vertical axis shows the frequency with which these responses occur to stimuli /b/, /m/, and /s/. From this figure, it is clear that the response distributions of /b/ and /m/ are much more similar to one another than they are to /s/. That is, participants show more similar patterns of responding to /b/ and /m/ than to /b/ and /s/ or /m/ and /s/. The similarity between /b/ and /m/ is quantified

---

[3] In Iverson et al. (1998), Phi-square values were not subtracted from one. The change is made here for two reasons. First, if Phi-square values are not subtracted from one, the value of any phoneme, given itself, is 0. This confounds analyses that involve the calculation of conditional probabilities (described below). A second reason for this transformation is ease of interpretation: it makes the scale of Phi-square values the same direction as probability of confusion (higher numbers represent greater similarity).

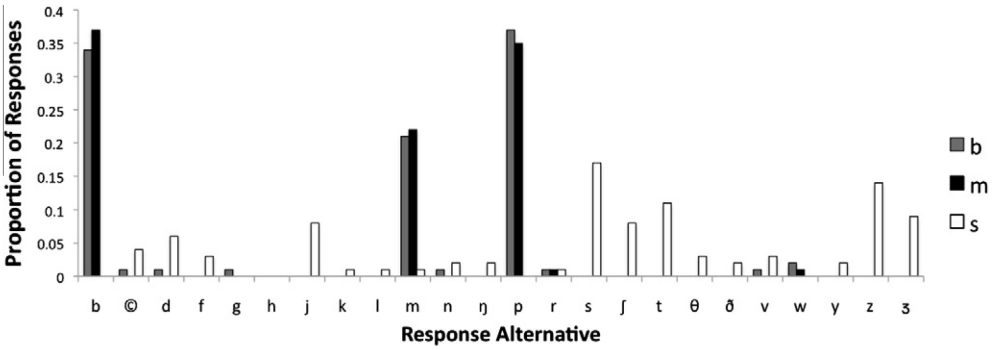

Fig. 1. Graphical representation of responses to visually presented /b/, /m/, and /s/. The horizontal axis shows all possible response alternatives, and the vertical axis shows the frequency with which these responses occur to stimuli /b/, /m/, and /s/.

as $\Phi^2 = .87$, while the similarity of /b/ and /s/ and of /m/ and /s/ are $\Phi^2 = .06$, and $\Phi^2 = .05$, respectively.

Iverson et al. (1998) used the Phi-square statistic to provide a mathematical basis for categorizing sounds into viseme groups. However, because the statistic renders a value quantifying the similarity between every pair of phonemes, it is also possible to use these values directly as a measure of perceptual similarity. This would eliminate the problem of information loss in categorical groupings as well as the confounds of using percent correct as a metric for similarity. The purpose of the present study was to use metrics based on the Phi-square statistic to overcome methodological difficulties in quantifying lexical competition in V-only speech and to determine if these similarity metrics support previous conclusions regarding lexical competition and spoken word identification. Support for this proposal would provide converging evidence that similar mechanisms mediate lexical selection in spoken word recognition for both A-only and V-only presentation.

2. Methods

2.1. Participants

Fifty participants (31 female, 19 male, mean age: 20.6, SD = 1.5) were recruited from Washington University's undergraduate participant pool. All participants were native English speakers and reported normal hearing and normal or corrected-to-normal vision. Participants received course credit or $10/h for participating, and all procedures were approved by Washington University's institutional review board.

2.2. Stimuli

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,) and 14 vowels (i, ɪ, ɛ, eɪ, æ, a, aʊ, aɪ, ʌ, ɔɪ, oʊ, ʊ, u, ɝ). Phonemic contexts (/hVd/ for vowels and /aCa/ for consonants) were selected to minimize co-articulation. All stimuli were recorded with a Cannon Elura 85 digital video camera connected to a Dell Precision PC. Digital capture and editing was done in Adobe Premiere Elements 1.0. The talkers (three male, three female were used) sat in front of a neutral grey background at a distance so that their head and neck filled the display and spoke the stimuli into the camera as they appeared on a teleprompter. Auditory and visual information was recorded for all stimuli, but only the visual signal was presented for the identification tasks.

2.3. Procedures

Participants were seated approximately .5 m from a 17-in. Touchsystems monitor (ELO-170C) running Superlab presentation software (Version 4.0.7b, Cedrus Corporation, 2009). They were presented with a video clip of a syllable being spoken, followed by a response screen listing each phoneme and an example word that contained it. Participants responded to each presentation by touching the button with the appropriate phoneme. Participants were given as much time as they wanted to make the response and no feedback was provided.

Each token by each speaker was repeated twice, resulting in 288 consonant trials and 168 vowel trials per participant. Consonant and vowel tokens were identified in separate blocks and presented in a pseudo-randomized order, blocked by speaker. Participants completed practice trials for both consonant and vowel identification that consisted of one presentation of each token by a different actor than was used in the test trials.

2.4. Analyses

To examine how lexical variables, including frequency and similarity, contribute to visual word recognition, we conducted a series of regression analyses. The criterion variable for all analyses was lipread word recognition accuracy from the Hoosier Audiovisual Multitalker Database [HAMD; Lachs and Hernández, 1998], an existing database that includes recognition accuracy data for visual-only presentation of 300 CVC words, spoken by six talkers. From this database, 260 words (referred to here as the

HAMD stimulus words) were selected that met the following criteria: (1) they had been presented to at least 80 participants; (2) they were identified accurately at least once; (3) they had entries in the English Lexicon Project (ELP, see below) (Balota et al., 2007). Scoring was based on an exact match criterion with the target items (i.e., morphological variants were counted as incorrect).

In order to estimate neighborhood structure, a phonetically coded lexicon was obtained from the ELP (Balota et al., 2007). The ELP is an online, searchable database that contains 40,000 phonetically-coded English words and corresponding lexical variables, including frequency of occurrence, phonological and orthographic neighborhood, and mean speeded naming and lexical decision reaction times, collected from 1200 subjects at six universities. From the ELP, a list of all consonant–vowel–consonant (CVC) words in English was created. This list was trimmed to exclude proper nouns and count homophones only once, rendering a lexicon of 1346 words (referred to here as the ELP-CVC Lexicon). For each of the HAMD stimulus words, the HAL_{log} frequency (Lund and Burgess, 1996) of occurrence was obtained from the ELP (Balota et al., 2007). These frequency counts are based on 131 million words gathered across 3000 usenet newsgroups in 1995, and values represent the log-transformed number of times a given word appeared in the corpus (Lund and Burgess, 1996).

2.4.1. Homophene analysis

The 260 HAMD stimulus words and 1346 ELP-CVC words were phonetically transcribed, and each phoneme was coded by viseme group using the categories determined by Iverson et al. (1998) (see Table 1). For example, the word "bat" was phonetically transcribed as /bæt/, and then viseme coded as /C1, V1, C7/. To quantify lexical competition, the ELP-CVC was searched to find the number of words with identical viseme strings to each of the HAMD stimulus words. For instance, "back" and "peg" constitute homophenes, because they are both coded as /C1, V1, C8/.

Table 1
Viseme groupings, determined by Iverson (1998).

Phonemes	Grouping
b, m, p	C1
f, v	C2
θ, ð	C3
w	C4
r	C5
tʃ, dʒ, ʒ, ʃ, d	C6
t, s, z	C7
k, g, h, ŋ, j	C8
n	C9
l	C10
i, ɪ, eɪ, aɪ, ɛ, æ, ʌ	V1
ɝ, oʊ, ɔɪ, ʊ, u	V2
a	V3
aʊ	V4

Table 2
Descriptive data for accuracy and measures of competition and frequency.

Variable	Range	Mean	SD
CVC identification accuracy	0.01–0.78	0.16	0.16
HAL_{log} frequency	5.55–14.77	9.9	1.68
Homophene group size	1–42	15.46	10.27
Probability density	0.02–0.53	0.26	0.12
Phi-square density	2.95–51.46	21.98	11.06

2.4.2. Probability analysis

Responses for the phoneme identification tasks were collapsed across the 50 participants, and confusion matrices were generated for consonant and vowels identification (see Appendix A). Following the procedure described in Section 1, perceptual similarity at the word level was determined by multiplying the conditional probabilities that position-specific stimulus phonemes would be confused with the competitor phonemes. This calculation was computed to determine the perceptual similarity of each of the HAMD stimulus words to every word in the ELP-CVC lexicon. For each stimulus word, the competitor probabilities were summed to obtain the predicted amount of competition, referred to here as probability density. Descriptive statistics for probability density are listed in Table 2.

2.4.3. Phi-square analysis

To calculate Phi-square values, the raw frequency syllable confusion matrices obtained from the 50 participants tested were converted to Phi-square values using SPSS (SPSS for Windows, version 18.0). These matrices are shown in Appendix A. To calculate the perceptual similarity of word pairs based on the Phi-square statistic, each stimulus word was compared to the 1346 words in the ELP-CVC lexicon, following the procedure of calculating conditional probability described above, but using Phi-square values in place of probability of confusion. For instance, the Phi-square similarity of $\text{mæd}|\text{bɛt} = \Phi^2(\text{m}|\text{b}) * \Phi^2(\text{æ}|\text{ɛ}) * \Phi^2(\text{d}|\text{t})$. Using this method, the Phi-square perceptual similarity of each HAMD stimulus word and every other word in the ELP-CVC lexicon was calculated. For each HAMD stimulus word, the perceptual similarities of all other words in the ELP-CVC lexicon were summed to quantify amount of competition (Phi-square density).

3. Results

Average identification accuracy of the 260 HAMD stimulus words (Lachs and Hernández, 1998) was 16% (SD = 16%) and ranged from 1% to 78%. Pearson product–moment correlations were calculated between recognition accuracy, word frequency, and all measures of competition (homophene group size, probability density, and Phi-square density). As is apparent from Table 3, all measures of lexical competition were significantly correlated with recognition accuracy. Regardless of how competition

Table 3
Pearson correlations between accuracy and measures of competition and frequency.

	HAL frequency	Homophene group size	Probability density	Phi-square density
Word identification accuracy	.31**	−.46**	−.46**	−.53**
HAL frequency		−.14*	−.14*	−.04
Homophene group size			.51**	.55**
Probability density				.52**

[*] $p < 0.05$.
[**] $p < .01$.

Table 5
Summary of multiple regression for predicting word identification accuracy with frequency, probability density, and Phi-square density.

	Variable	ß	R^2	ΔR^2
Step 1			.10	.10**
	Frequency	.31		
Step 2			.27	.17**
	Frequency	.25		
	Probability density	−.43		
Step 3			.40	.13**
	Frequency	.27		
	Probability density	−.21		
	Phi-square density	.42		

[**] $p < .01$

was quantified, words in perceptually sparse regions of the lexicon were accurately recognized more often than words in dense regions. In addition, recognition accuracy was higher for more frequent words. Measures of competition were also correlated with one another, so a series of regression analyses was conducted to assess the independent contributions of each measure of competition.

Two hierarchical regression analyses were conducted to determine whether the Phi-square statistic accounts for additional variance in word recognition accuracy beyond that explained by homophene group size and probability density (see Tables 4 and 5). For all analyses, word frequency was entered as the first step, either homophene group size or probability density was entered second, and Phi-square density was entered in the third step. These analyses revealed that, although stimulus frequency, homophene group size, and probability density account for significant variance in lipread word recognition, Phi-square density explains significant variance above and beyond these measures.

As discussed in the Introduction, there is a theoretical difference (in addition to the methodological difference) between the homophene group method and continuous measures. The underlying theoretical assumption of the homophene group method is that stimulus words are evaluated in the context of perceptually similar words in the lexicon. The assumption of the continuous measures, however, is that stimulus words are evaluated in the context a larger subset of the lexicon. That is, many words provide some competition (albeit very little in some cases) for the stimulus word. This study is not able to test the prediction that *all* words in the lexicon influence recognition, because the HAMD stimulus words and ELP-CVC lexicon were all CVC words (see Section 4). However, it is possible to test whether the influence of perceptually more distant competitors influence recognition, using an additional analysis.

Table 4
Summary of multiple regression for predicting word identification accuracy with frequency, homophene group size, and Phi-square density.

	Variable	ß	R^2	ΔR^2
Step 1			.10	.10**
	Frequency	.31		
Step 2			.27	.17**
	Frequency	.25		
	Homophene group size	−.42		
Step 3			.39	.12**
	Frequency	.27		
	Homophene group size	−.19		
	Phi-square density	−.42		

[**] $p < .01$

For each stimulus word, Phi-square values were computed for words in the same homophene group, and these values were summed. For example, if stimulus word "bar" has two homophenes, "par" and "mar" the Phi-square values of "par" | "bar" and "mar" | "bar" were summed. This analysis maintains the method of comparing a stimulus word to a small subset of perceptually similar competitors, but allows the similarity of the competitors to be measured continuously. If Phi-square density (which includes the influence of all CVCs) accounts for significant variance beyond that accounted for by the homophene group's Phi-square density, it suggests that words outside the homophene group influence target recognition. This is exactly the pattern that emerges from a hierarchical regression: Word frequency accounts for 10% of the variance in word recognition, homophene group Phi-square density explains a further 12%, and Phi-square density accounts for an additional 16% of the variance ($p < .01$ for all). This suggests that a larger subset of the lexicon is activated during word recognition than is represented by the homophene-based system.

4. Discussion

The present study assessed the influence of the mental lexicon on V-only spoken word recognition using both existing metrics of competition as well as novel methods based on the Phi-square statistic. The findings support and extend previous work (Auer, 2002; Mattys et al., 2002; Tye-Murray et al., 2007) suggesting that V-only recognition of spoken words is sensitive to the lexical properties of the input and suggests that both A-only and V-only word recognition rely upon bottom-up activation of multiple word candidates and competition among these candidates.

A measurement question posed here was whether the measures of lexical density that are based on the Phi-square statistic would have greater power at predicting spoken word recognition than would measures of lexical density based on confusion probabilities. Although this is the first investigation to use the Phi-square statistic as a continuous measure of perceptual similarity, it might be expected to be a more predictive measure than a parallel analysis that uses confusion probabilities, based on several confounds associated with confusion probability methods (described in Section 1; see Iverson et al., 1998). Overcoming these confounds may result in a more accurate measure of perceptual similarity, and therefore, lexical density. The results of the current investigation demonstrate that quantifying lexical competition using metrics based on the Phi-square statistic are more effective at predicting V-only spoken word recognition than are metrics based on probability density or homophene groupings.

The majority of work on lexical effects in lipreading (see Auer, 2002 for an exception) has quantified competition in the V-only domain by comparing a stimulus word to perceptually similar competitors (homophenes). A model of word recognition that makes use of such a method might hypothesize that stimulus input activates a large set of lexical items, but only items that meet some threshold of similarity to the stimulus representation receive enough activation to provide competition to the stimulus. This work does not support such a model, and rather, suggests that many words are initially activated and provide competition for a stimulus word.

A possible limitation of the method used here is that syllable and word stimuli were produced by different speakers and identified by different participants. It has been well established that patterns of phoneme similarity depend on speaker idiosyncrasies and the phonemic characteristics of the stimulus materials (e.g., the phonemic context in which phonemes are identified (Jackson et al., 1988). Given that the phoneme confusion matrices from which both probability density and Phi-square density were derived were obtained from different speakers than those who produced the word recognition stimuli, it is particularly notable that they have such predictive power. In addition, it would be preferable to use viseme groupings specific to the speakers and listeners being tested. In some investigations (as was the case in the current study), however, it is not possible to classify all phonemes into viseme groups based on traditional 75% within-group identification rates (Owens and Blazek, 1985). A final way in which the predictive power of these calculations may be improved would be to devise a metric for comparing the similarity of word strings of different lengths, (eg. “best” | “bet”). It seems likely that it would be difficult to discriminate between sequential segments with high Phi-square values (/s/ and /t/), and therefore “best” would be expected to provide competition for “bet.” Inclusion of such a metric could further test the extent of lexical competition.

In conclusion, this work demonstrates that V-only, like A-only spoken word recognition, is sensitive to the

Table A1
Confusion matrix showing the percent of trials on which each consonant was identified as every other consonant.

	b	tʃ	d	f	g	h	dʒ	k	l	m	n	ŋ	p	r	s	ʃ	t	θ	ð	v	w	y	z	ʒ
b	0.34	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.21	0.01	0.00	0.37	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00
tʃ	0.00	0.21	0.01	0.00	0.00	0.00	0.25	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.18	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.27
d	0.00	0.02	0.13	0.03	0.04	0.01	0.06	0.06	0.04	0.00	0.13	0.04	0.00	0.02	0.05	0.03	0.11	0.06	0.04	0.01	0.00	0.04	0.05	0.03
f	0.00	0.00	0.00	0.59	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.35	0.00	0.00	0.00	0.00
g	0.00	0.01	0.05	0.00	0.14	0.02	0.01	0.14	0.23	0.01	0.14	0.08	0.00	0.01	0.01	0.00	0.02	0.01	0.01	0.01	0.00	0.09	0.00	0.00
h	0.01	0.00	0.02	0.01	0.13	0.48	0.01	0.08	0.04	0.01	0.02	0.08	0.01	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01
dʒ	0.00	0.17	0.01	0.03	0.01	0.00	0.27	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.18	0.01	0.01	0.01	0.01	0.00	0.01	0.02	0.24
k	0.00	0.00	0.05	0.00	0.17	0.04	0.02	0.19	0.15	0.00	0.16	0.08	0.00	0.01	0.00	0.00	0.03	0.01	0.01	0.00	0.00	0.07	0.00	0.00
l	0.00	0.00	0.04	0.00	0.01	0.01	0.01	0.01	0.54	0.00	0.04	0.01	0.00	0.01	0.01	0.01	0.03	0.15	0.11	0.01	0.00	0.01	0.00	0.00
m	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.00	0.35	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
n	0.00	0.01	0.08	0.00	0.07	0.00	0.00	0.07	0.20	0.01	0.20	0.04	0.01	0.01	0.02	0.00	0.05	0.09	0.06	0.01	0.01	0.05	0.02	0.01
ŋ	0.00	0.00	0.03	0.00	0.14	0.06	0.01	0.18	0.20	0.00	0.12	0.10	0.00	0.04	0.00	0.00	0.03	0.01	0.00	0.01	0.00	0.05	0.00	0.00
p	0.31	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.01	0.00	0.41	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
r	0.00	0.00	0.01	0.09	0.01	0.00	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.52	0.00	0.00	0.01	0.01	0.00	0.11	0.12	0.02	0.00	0.01
s	0.00	0.04	0.06	0.03	0.00	0.00	0.08	0.01	0.01	0.01	0.02	0.02	0.00	0.01	0.17	0.08	0.11	0.03	0.02	0.03	0.00	0.02	0.14	0.09
ʃ	0.00	0.15	0.02	0.01	0.00	0.00	0.22	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.02	0.21	0.01	0.01	0.01	0.00	0.00	0.01	0.05	0.25
t	0.00	0.05	0.13	0.01	0.04	0.01	0.06	0.04	0.03	0.01	0.06	0.05	0.01	0.02	0.10	0.04	0.12	0.02	0.03	0.01	0.00	0.03	0.10	0.05
θ	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.49	0.44	0.00	0.00	0.00	0.00	0.00
ð	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.07	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.46	0.40	0.00	0.00	0.00	0.00	0.00
v	0.00	0.00	0.01	0.56	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.36	0.00	0.01	0.00	0.00
w	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.90	0.01	0.00	0.00
y	0.00	0.01	0.09	0.00	0.06	0.01	0.03	0.07	0.16	0.00	0.18	0.05	0.01	0.02	0.03	0.02	0.07	0.05	0.03	0.01	0.00	0.07	0.02	0.01
z	0.00	0.04	0.07	0.04	0.01	0.00	0.08	0.02	0.02	0.00	0.02	0.01	0.00	0.02	0.17	0.07	0.13	0.02	0.01	0.02	0.00	0.02	0.16	0.05
ʒ	0.00	0.16	0.01	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.02	0.21	0.01	0.01	0.01	0.01	0.00	0.01	0.03	0.24

frequency of the stimulus word and the extent to which perceptually similar words exist and compete for recognition. By using the Phi-square statistic to overcome the previous methodological limitations in quantifying lexical competition, the strength of the relationship between lexical density and word recognition increased. Future work should address the extent to which similarities in A-only and V-only word recognition persist across recognition contexts. For example, recent work provides initial evidence that the observed similarities are not specific to single word recognition, and rather, that they extend to sentence recognition, where additional top-down influences, such as available semantic information is available. Auer and Reed (2008) found evidence that the influence of the lexicon on V-only word recognition is reduced when words are presented in a sentence context. In the A-only domain, Sommers and Danielson (1999) found that, when available, semantic information can serve as an additional source of information for lexical discriminations and reduce the likelihood of within neighborhood recognition errors. Additional parallel findings of this nature would further strengthen the hypothesis (Auer, 2002) that spoken word recognition is achieved by similar processes in both A-only and V-only speech.

Table A2
Matrix showing the Phi-square values for every pair of consonants.

	b	tʃ	d	f	g	h	dʒ	k	l	m	n	ŋ	p	r	s	ʃ	t	θ	ð	v	w	y	z	ʒ
b	1	0.03	0.06	0.04	0.05	0.05	0.04	0.04	0.04	0.87	0.06	0.05	0.85	0.09	0.06	0.04	0.06	0.01	0.03	0.04	0.07	0.06	0.06	0.04
tʃ	0.03	1	0.21	0.04	0.08	0.08	0.82	0.08	0.06	0.03	0.09	0.07	0.03	0.07	0.34	0.81	0.27	0.04	0.05	0.03	0.03	0.14	0.30	0.82
d	0.06	0.21	1	0.10	0.45	0.26	0.25	0.44	0.31	0.05	0.58	0.42	0.05	0.20	0.56	0.28	0.74	0.15	0.18	0.10	0.06	0.68	0.57	0.27
f	0.04	0.04	0.10	1	0.06	0.05	0.07	0.05	0.06	0.03	0.06	0.06	0.03	0.23	0.11	0.05	0.08	0.04	0.04	0.86	0.03	0.07	0.12	0.05
g	0.05	0.08	0.45	0.06	1	0.38	0.10	0.82	0.39	0.04	0.66	0.78	0.05	0.15	0.22	0.11	0.38	0.08	0.14	0.06	0.05	0.68	0.23	0.11
h	0.05	0.08	0.26	0.05	0.38	1	0.09	0.42	0.13	0.05	0.27	0.46	0.05	0.16	0.14	0.08	0.24	0.05	0.07	0.05	0.06	0.31	0.16	0.09
dʒ	0.04	0.82	0.25	0.07	0.10	0.09	1	0.09	0.08	0.04	0.11	0.09	0.04	0.11	0.37	0.81	0.29	0.05	0.05	0.07	0.04	0.16	0.34	0.81
k	0.04	0.08	0.44	0.05	0.82	0.42	0.09	1	0.30	0.04	0.59	0.82	0.04	0.14	0.20	0.10	0.36	0.07	0.12	0.05	0.05	0.62	0.21	0.10
l	0.04	0.06	0.31	0.06	0.39	0.13	0.08	0.30	1	0.03	0.51	0.33	0.04	0.11	0.20	0.09	0.24	0.30	0.37	0.05	0.04	0.43	0.19	0.09
m	0.87	0.03	0.05	0.03	0.04	0.05	0.04	0.04	0.03	1	0.05	0.04	0.86	0.07	0.05	0.03	0.05	0.02	0.03	0.04	0.05	0.05	0.05	0.04
n	0.06	0.09	0.58	0.06	0.66	0.27	0.11	0.59	0.51	0.05	1	0.56	0.05	0.15	0.29	0.13	0.45	0.21	0.27	0.06	0.05	0.79	0.29	0.13
ŋ	0.05	0.07	0.42	0.06	0.78	0.46	0.09	0.82	0.33	0.04	0.56	1	0.04	0.18	0.20	0.09	0.35	0.07	0.12	0.06	0.06	0.59	0.21	0.10
p	0.85	0.03	0.05	0.03	0.05	0.05	0.04	0.04	0.04	0.86	0.05	0.04	1	0.07	0.05	0.04	0.06	0.02	0.04	0.03	0.06	0.05	0.05	0.04
r	0.09	0.07	0.20	0.23	0.15	0.16	0.11	0.14	0.11	0.07	0.15	0.18	0.07	1	0.17	0.09	0.18	0.05	0.07	0.23	0.19	0.17	0.19	0.09
s	0.06	0.34	0.56	0.11	0.22	0.14	0.37	0.20	0.20	0.05	0.29	0.20	0.05	0.17	1	0.43	0.70	0.10	0.12	0.11	0.05	0.37	0.85	0.41
ʃ	0.04	0.81	0.28	0.05	0.11	0.08	0.81	0.10	0.09	0.03	0.13	0.09	0.04	0.09	0.43	1	0.35	0.05	0.06	0.05	0.04	0.18	0.39	0.86
t	0.06	0.27	0.74	0.08	0.38	0.24	0.29	0.36	0.24	0.05	0.45	0.35	0.06	0.18	0.70	0.35	1	0.10	0.13	0.08	0.06	0.55	0.71	0.33
θ	0.01	0.04	0.15	0.04	0.08	0.05	0.05	0.07	0.30	0.02	0.21	0.07	0.02	0.05	0.10	0.05	0.10	1	0.84	0.03	0.02	0.15	0.08	0.05
ð	0.03	0.05	0.18	0.04	0.14	0.07	0.05	0.12	0.37	0.03	0.27	0.12	0.04	0.07	0.12	0.06	0.13	0.84	1	0.04	0.03	0.20	0.10	0.06
v	0.04	0.03	0.10	0.86	0.06	0.05	0.07	0.05	0.05	0.04	0.06	0.06	0.03	0.23	0.11	0.05	0.08	0.03	0.04	1	0.03	0.07	0.12	0.05
w	0.07	0.03	0.06	0.03	0.05	0.06	0.04	0.05	0.04	0.05	0.05	0.06	0.06	0.19	0.05	0.04	0.06	0.02	0.03	0.03	1	0.06	0.06	0.04
y	0.06	0.14	0.68	0.07	0.68	0.31	0.16	0.62	0.43	0.05	0.79	0.59	0.05	0.17	0.37	0.18	0.55	0.15	0.20	0.07	0.06	1	0.38	0.18
z	0.06	0.30	0.57	0.12	0.23	0.16	0.34	0.21	0.19	0.05	0.29	0.21	0.05	0.19	0.85	0.39	0.71	0.08	0.10	0.12	0.06	0.38	1	0.37
ʒ	0.04	0.82	0.27	0.05	0.11	0.09	0.81	0.10	0.09	0.04	0.13	0.10	0.04	0.09	0.41	0.86	0.33	0.05	0.06	0.05	0.04	0.18	0.37	1

Table A3
Confusion matrix showing the percent of trials on which each vowel was identified as every other vowel.

	i	ɪ	ɛ	eɪ	æ	ɑ	aʊ	aɪ	ʌ	oɪ	o	ʊ	u	ɝ
i	0.65	0.21	0.03	0.02	0.00	0.00	0.01	0.01	0.00	0.03	0.00	0.01	0.03	0.01
ɪ	0.37	0.48	0.07	0.03	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01
ɛ	0.01	0.05	0.23	0.19	0.29	0.04	0.01	0.16	0.02	0.00	0.01	0.00	0.00	0.01
eɪ	0.04	0.09	0.22	0.39	0.11	0.01	0.01	0.11	0.00	0.01	0.00	0.00	0.00	0.01
æ	0.01	0.02	0.11	0.08	0.33	0.04	0.04	0.33	0.02	0.00	0.00	0.01	0.00	0.00
ɑ	0.00	0.01	0.01	0.02	0.01	0.71	0.03	0.03	0.08	0.03	0.03	0.02	0.00	0.02
aʊ	0.00	0.00	0.00	0.00	0.00	0.03	0.94	0.00	0.00	0.01	0.01	0.00	0.01	0.00
aɪ	0.00	0.01	0.04	0.07	0.14	0.07	0.01	0.61	0.03	0.01	0.00	0.00	0.00	0.00
ʌ	0.00	0.01	0.03	0.03	0.04	0.39	0.02	0.12	0.26	0.02	0.01	0.04	0.00	0.04
oɪ	0.00	0.00	0.00	0.01	0.00	0.02	0.01	0.00	0.01	0.70	0.12	0.03	0.10	0.00
o	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.01	0.10	0.71	0.02	0.13	0.00
ʊ	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.05	0.06	0.02	0.47	0.08	0.28
u	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.04	0.03	0.06	0.81	0.01
ɝ	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.05	0.03	0.02	0.38	0.06	0.42

Table A4
Confusion matrix showing the Phi-square values for every pair of vowels.

	i	ɪ	ɛ	eɪ	æ	ɑ	aʊ	aɪ	ʌ	oɪ	o	ʊ	u	ɝ
i	1	0.64	0.13	0.20	0.10	0.08	0.03	0.09	0.09	0.07	0.07	0.08	0.07	0.08
ɪ	0.64	1	0.17	0.24	0.13	0.07	0.02	0.10	0.09	0.03	0.02	0.04	0.02	0.05
ɛ	0.13	0.17	1	0.66	0.69	0.15	0.04	0.47	0.27	0.05	0.03	0.06	0.03	0.06
eɪ	0.20	0.24	0.66	1	0.45	0.11	0.04	0.34	0.20	0.05	0.04	0.05	0.03	0.06
æ	0.10	0.13	0.69	0.45	1	0.17	0.07	0.65	0.30	0.05	0.04	0.06	0.04	0.07
ɑ	0.08	0.07	0.15	0.11	0.17	1	0.08	0.20	0.61	0.12	0.09	0.14	0.09	0.14
aʊ	0.03	0.02	0.04	0.04	0.07	0.08	1	0.05	0.06	0.05	0.05	0.04	0.04	0.04
aɪ	0.09	0.10	0.47	0.34	0.65	0.20	0.05	1	0.34	0.06	0.03	0.06	0.04	0.07
ʌ	0.09	0.09	0.27	0.20	0.30	0.61	0.06	0.34	1	0.09	0.06	0.17	0.08	0.18
oɪ	0.07	0.03	0.05	0.05	0.05	0.12	0.05	0.06	0.09	1	0.33	0.19	0.21	0.15
o	0.07	0.02	0.03	0.04	0.04	0.09	0.05	0.03	0.06	0.33	1	0.16	0.23	0.13
ʊ	0.08	0.04	0.06	0.05	0.06	0.14	0.04	0.06	0.17	0.19	0.16	1	0.22	0.82
u	0.07	0.02	0.03	0.03	0.04	0.09	0.04	0.04	0.08	0.21	0.23	0.22	1	0.18
ɝ	0.08	0.05	0.06	0.06	0.07	0.14	0.04	0.07	0.18	0.15	0.13	0.82	0.18	1

Acknowledgments

We thank Lorin Lachs, Luis Hernandez, and David Pisoni for generously sharing with us their identification data from the Hoosier Audiovisual Multitalker Database. This work was supported in part by grant award number RO1 AG 18029-4 from the National Institute on Aging.

Appendix A

See Tables A1–A4.

References

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

Auer, E., Bernstein, L., 1997. Speechreading and the structure of the lexicon: computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. J. Acoust. Soc. Amer. 102 (6), 3704–3710.

Auer, E., Reed, R., 2008. Investigating lexical influences on the accuracy of speechreading words presented in isolation and in sentence context. J. Acoust. Soc. Amer. 124 (4), 2459.

Balota, D., Yap, M., Cortese, M., Hutchison, K., Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G., Treiman, R., 2007. The English Lexicon Project. *Behav. Res. Methods, 39*, 445–459. Available online at: <http:/elexicon.wustl.edu/>.

Bernstein, L.E., Iverson, P., Auer, E.T, Jr., 1997. Elucidating the complex relationships between phonetic perception and word recognition in audiovisual speech perception. In: Benoit, C., Campbell, R., (Eds.), Proc. ESCA/ESCOP Workshop on Audio–Visual Speech Processing, Rhodes, Greece, pp. 21–24.

Goldinger, S.D., Luce, P.A., Pisoni, D.B., 1989. Priming lexical neighbors of spoken words: effects of competition and inhibition. J. Memory Lang. 28, 501–518.

Iverson, P., Bernstein, L., Auer, E., 1998. Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. Speech Commun. 26, 45–63.

Jackson, P.L., 1988. The theoretical minimal unit for visual speech perception: visemes and coarticulation. In: DeFillipo, C., Sims, D. (Eds.), New Reflections on Speechreading. A.G. Bell Association for the Deaf, Washington, DC, pp. 99–115.

Jusczyk, P.W., Luce, P.A., 2002. Speech perception and spoken word recognition: past and present. Ear Hear. 23, 2–40.

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

Lachs, L., Hernández, L.R., 1998. Update: the hoosier audiovisual multitalker database. In: Research on Spoken Language Processing Progress Report 22. Speech Research Laboratory, Indiana University, Bloomington, IN, pp. 377–388.

Luce, P., Pisoni, D., 1998. Recognizing spoken words: the neighborhood activation model. Ear Hear. 19, 1–36.

Lund, K., Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods – Instrum. Comput. 28, 203–208.

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

McClelland, J., Elman, J., 1986. The TRACE model of speech perception. Cognit. Psychol. 18, 1–86.

Morton, J., 1979. Word recognition. In: Morton, J., Marshall, J.C. (Eds.), Psycholinguistics Structures and Processes, Vol. 2. Paul Elek, London.

Norris, D., 1994. Shortlist: a connectionist model of continuous speech recognition. Cognition 52, 189–234.

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

Seitz, P.F., Bernstein, L.E., Auer Jr., E.T., MacEachern, M.E., 1998. PhLex (Phonologically Transformable Lexicon): a 35,000-Word Computer Readable Pronouncing American English Lexicon on Structural Principles, with Accompanying Phonological Transformations, and Word Frequencies. House Ear Institute, Los Angeles.

Sommers, M., Danielson, S., 1999. Inhibitory processes and spoken word recognition in younger and older adults: the interaction of lexical competition and semantic context. Psychol. Aging 14, 458–472.

Summerfield, Q., 1992. Lipreading and audio–visual speech perception. Philos. Trans. Biol. Sci. (1273), 71–78.

Tye-Murray, N., 2008. Foundations of Aural Rehabilitation: Children, Adults, and their Family Members. Delmar Cengage Learning, Clifton Park, NY.

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

Vitevitch, M., Luce, P., 1998. When words compete: levels of processing in spoken word perception. Psychol. Sci. 9, 325–329.

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