Auditory and Visual Lexical Neighborhoods in Audiovisual Speech Perception

Nancy Tye-Murray, PhD, Mitchell Sommers, PhD,
and Brent Spehar, PhD

Much evidence suggests that the mental lexicon is organized into auditory neighborhoods, with words that are phonologically similar belonging to the same neighborhood. In this investigation, we considered the existence of visual neighborhoods. When a receiver watches someone speak a word, a neighborhood of homophenes (ie, words that look alike on the face, such as *pat* and *bat*) is activated. The simultaneous activation of a word's auditory and visual neighborhoods may, in part, account for why individuals recognize speech better in an auditory–visual condition than what would be predicted by their performance in audition-only and vision-only conditions. A word test was administered to 3 groups of participants in audition-only, vision-only, and auditory–visual conditions, in the presence of 6-talker

babble. Test words with sparse visual neighborhoods were recognized more accurately than words with dense neighborhoods in a vision-only condition. Densities of both the acoustic and visual neighborhoods as well as their intersection overlap were predictive of how well the test words were recognized in the auditory–visual condition. These results suggest that visual neighborhoods exist and that they affect auditory–visual speech perception. One implication is that in the presence of dual sensory impairment, the boundaries of both acoustic and visual neighborhoods may shift, adversely affecting speech recognition.

Keywords: Lexical neighborhood; auditory–visual speech perception; integration

Most people, whether they have normal or impaired hearing, can recognize speech better when they can both see and hear the talker than when listening alone.^{1,2} Often, the advantage of supplementing listening with watching is more than additive. For instance, Sommers et al³ tested 38 young adults (between the ages of 18 and 25 years) in an audition-only (A), vision-only (V), and audition-plus-vision (AV) condition with 3 stimuli types—consonants, words, and sentences. The per-

formance in the AV condition for all 3 stimulus types surpassed what would be predicted by simply adding the scores obtained in the A and V conditions. On average, for example, the subjects recognized about 10% of the words in the Iowa Sentence Test⁴ in a V condition, about 40% of the words in an A condition with a background of 6-talker babble, and nearly 80% of the words in an AV condition, also with a background of 6-talker babble. A similar superadditive effect was observed for the group of 44 older adults (ages 65 years and older) who were also included in the study and for the consonant and word tests.

One reason for this superadditive effect is the complementary nature of the auditory and visual speech signals.^{1,5} For example, cues about nasality and voicing are typically conveyed very well by the auditory signal, even in adverse listening situations, whereas the visual signal does not convey them at all, even in the best of viewing conditions. On the other hand, cues about place of articulation are conveyed by the visual signal but not very well by a degraded

From CID at Washington University School of Medicine (NTM, BS) and Department of Psychology, Washington University in St Louis (MS), St Louis, Missouri.

The contents of this article were presented orally at the State of the Science Conference on Hearing Enhancement, September 20, 2006, Gallaudet University, Washington, DC.

Address correspondence to: Nancy Tye-Murray, PhD, Department of Otolaryngology, Campus Box 8115, Washington University School of Medicine, 660 South Euclid Avenue, St. Louis, MO 63110; e-mail: murrayn@ent.wustl.edu.

auditory signal, as when listening with a hearing loss or listening in the presence of background noise.

To understand how decreased vision and decreased audition might affect speech recognition, we must examine this superadditive effect and explain how persons with normal vision and/or normal audition combine what they see with what they hear. We can then seek to understand the challenges that persons with impaired hearing and vision confront and perhaps develop remedial procedures and counseling recommendations for alleviating their difficulties. For instance, it may be that the integrative skills of persons with dual sensory impairment are better than individuals with a unimodal impairment or individuals with intact sensory systems, owing to a greater need to exploit the integrative skill to compensate for the dual impairment. It also is possible that a dual sensory impairment might result in less well-developed integrative skills because of degraded unisensory inputs. Finally, even if integrative abilities remain intact, persons with dual sensory impairments will be at a considerable disadvantage encoding sensory information within each modality. Consequently, they will not enjoy the same degree of superadditive benefit for AV presentations as will individuals who have either a single or no sensory loss.

At least 2 groups of investigators have proposed a conceptual model for AV speech recognition that entails 3 stages.^{1,6,7} These 3 stages include an initial stage of perceiving the auditory and visual signals, a second stage of integrating the 2 signals, and a third stage of making discrete phonetic and lexical decisions (see Figure 1). For example, the model by Grant et al¹ suggests that the 3 stages proceed from left to right in time. The first stage entails perceiving the auditory and visual cues associated with a spoken word. These cues are integrated in a second, distinct stage. In the final stages, top-down semantic, syntactic, and contextual constraints factor in, and recognition occurs.

The possible existence of a distinct stage of integration has motivated us and other investigators to evaluate whether integration ability might be a quantifiable skill. That is, if there exists a distinct stage at which the 2 signals are combined, it may be that some persons are more skilled at integrating the 2 signals than are others. This has led to attempts to quantify integration ability and to experiments comparing the integration skills of different populations.^{8,9}

Perhaps the clearest attempt to quantify integration, independent of either A or V performance, is the Prelabeling (PRE) model developed by Braida.¹⁰

Figure 1. A model of auditory-visual speech perception, adapted from Grant et al.¹ The model entails a stage of perceiving the auditory and visual cues, a second stage of integrating the 2 kinds of cues, and a third stage of accessing the mental lexicon. The words in the lexicon, in some cases, might affect how one interprets the auditory and visual cues, as indicated by the dotted lines.

The PRE model predicts AV performance from consonant confusion error patterns obtained with A and V presentations. Specifically, the model uses multi-dimensional scaling of A and V consonant confusion matrices to predict optimal AV integration (ie, integration that would be predicted using an optimum observer model). The optimum observer in this case is assumed to combine the individual phonetic cues (eg, place of articulation) optimally and to be free from masking or other forms of interference in both the A and V modalities. If obtained AV scores are below the predicted performance, then an individual is said to have demonstrated less than optimal integration skills, and the ratio of predicted to obtained scores provides an index of integration abilities that is independent of unimodal encoding. Consistent with the ability to predict optimum AV performance from unimodal confusion matrices, past studies using the PRE model have found that predicted AV performance is consistently higher than obtained scores.^{1,11}

The primary limitation of the PRE model as a measure of integration is that because it uses multi-dimensional scaling of unimodal (A and V) performance to predict optimal AV scores, it requires highly stable estimates of confusion patterns obtained from closed-set procedures. Consequently, not only does it require extensive testing of individual participants, but its use has been limited exclusively to consonant identification.

Recently, Tye-Murray et al⁹ (see also Sommers et al¹²) used a measure of integration, termed *integration enhancement*, to examine integration for words and sentences. Integration enhancement is based on the work of Blamey et al¹³ and predicts performance in the AV condition based on independent identification in the 2 unimodal conditions, in that it assumes errors in the AV condition are made only when identification is incorrect in *both* modalities. Unlike the

PRE model, participants generally perform better than predicted by the model, and this difference between predicted and obtained scores serves as an index of integration. That is, individual differences between obtained and predicted performance are proposed to reflect individual differences in integration ability. The obvious advantage of using integration enhancement as a measure of AV integration is that, theoretically, it can be used for any type of material ranging from single consonants to extended discourse.

Recent experiments from our laboratory have used both the PRE and integration enhancement measures to determine whether younger adults integrate consonants, words, and sentences more effectively than do older adults^{8,3} and whether older persons with acquired hearing loss integrate more effectively than do older persons with normal hearing.⁹ We concluded, as did Cienkowski and Carney⁸ in an earlier study, that age affects one's ability to lip-read (ie, to recognize speech using only the visual signal) but does not affect one's ability to integrate. Similarly, Tye-Murray et al⁹ compared integration enhancement of a group of older participants with age-appropriate hearing loss to a group of older participants with normal hearing and found no difference in their integration abilities.

The findings from both Sommers et al³ and Tye-Murray et al⁹ have led us to reconceptualize how receivers might combine the auditory and visual speech signals. Perhaps there are no differences between older and younger persons' abilities to combine A and V information and no difference between older normal-hearing and hearing-impaired persons' abilities because there is no distinct integration stage as depicted in Figure 1. In addition or alternatively, the properties of the speech test stimuli may determine in large degree how much integration of the auditory and visual signals occurs. To advance these arguments, we draw on the literature concerning the Neighborhood Activation Model (NAM)¹⁴ and illustrate how the second stage of processing illustrated in Figure 1 might be reconceptualized.

The NAM is 1 example of a broader class of speech perception models that are often referred to as activation-competition models. In these models, presentation of a word activates a set of lexical candidates that "compete" for a best match with the incoming stimulus. In the NAM, the set of lexical candidates that are initially activated is referred to as a neighborhood, and neighborhoods are defined operationally as containing any word that can be

created from a target word by adding, deleting, or substituting a single phoneme (eg, *kit*, *cab*, and *scat* are all neighbors of the word *cat*). Considerable evidence is now available to support the importance of lexical neighborhoods in determining both the speed and accuracy of spoken word recognition (see Luce and Pisoni¹⁴ for a review of these findings). Bottom-up processing is initiated with the spoken word. Members of a neighborhood that sound similar to the speech waveform receive a high level of activation, whereas items that are dissimilar receive less activation. As more of the waveform is received, items in the selected neighborhood are eliminated until 1 item reaches a criterion value. So-called top-down information affects recognition as well. For instance, all other things being equal, words that have a high frequency of occurrence in the language receive a greater level of activation than words that have a low frequency of occurrence. A lexical neighborhood might be characterized as either dense or sparse. Dense neighborhoods include many words that sound alike, whereas sparse neighborhoods contain few words. For example, the words *cat* and *song* are both highly familiar monosyllabic words, but *cat* has a high neighborhood density (35), whereas *song* has a relatively low neighborhood density (11). In a noisy condition, a listener is more likely to recognize a word that belongs to a sparse lexical neighborhood than to recognize a word that has a dense lexical neighborhood. For instance, in the presence of background noise, a listener with normal hearing might recognize the word *song* more readily than the word *cat* because *song* has fewer lexical neighbors that can serve as competitors.

In Figure 2, we show how the NAM might be applied to audiovisual speech recognition. Figure 2a shows a lexical neighborhood schematic for the word *fish*. The left circle represents the acoustic lexical neighborhood, as determined by the single addition, deletion, or substitution rule described above. As the figure illustrates, *fish* belongs to a dense neighborhood, with such words as *fizz*, *fear*, and *fit* (total words = 13).

We propose not only that individual words belong to an acoustic neighborhood but also that each belongs to a visually defined neighborhood, consisting of words that look similar on the face. This proposal is consonant with findings from a previous investigation demonstrating that visually similar words appear to comprise "lexical equivalence classes." Mattys et al¹⁵ found that a word that looked

Figure 2. The upper half of the figure illustrates the auditory and visual lexical neighborhoods for the words *fork* and *fish*. Both words have similar densities for their auditory and their visual neighborhoods. The lower half of the figure illustrates what may happen when the auditory and visual neighborhoods for the 2 words are activated simultaneously. A larger number of words are activated in the intersection density for *fork* (a) than for *fish* (b). Words outside of the intersection receive minimal or no activation. In less than ideal listening or viewing conditions, an individual may be more likely to recognize *fish* than *fork*. A = audition only; V = vision only; AV = audition plus vision.

like few other words on the face was more likely to be recognized in a vision-only condition, by participants with either normal or impaired hearing, than a word that looked like many words. They suggested that similar lexical neighborhood effects may influence both auditory and visual word recognition.

The adjacent circle in Figure 2a illustrates a visual lexical neighborhood. Words that appear visually similar to *fork* are listed. Selection of these visually similar words was based on an extrapolation of viseme groups reported by Lesner et al.¹⁶ and Berger¹⁷ and is described in more detail in the Methods section of this report. Visemes are groups of sounds that look identical on the lips, such as /p, b, m/ and /t, d, n, s/. Although the visual neighborhood for the word *fork* is less dense than the acoustically defined neighborhood, 5 candidate alternatives are available. The bottom half of Figure 2a demonstrates what happens if an individual receives both the acoustic and visual speech signals simultaneously. The candidate choices are fewer than for either the

A or the V conditions alone, but still ambiguity exists. If either listening or viewing conditions are less than ideal, an individual might well misperceive the word.

Figure 2b shows a lexical neighborhood for the word *fish*. This word belongs to a similarly dense acoustic and similarly dense visual neighborhood as the word *fork*. However, if an individual is asked to recognize the word in an AV condition, and both the acoustic and visual neighborhoods are activated efficiently, then response alternatives are narrowed to only 1 activated word. As such, in difficult viewing or listening conditions, an individual is much more likely to recognize the word *fish* than the word *fork*, if factors such as semantic and syntactic contextual support are equal.

The principles of the NAM, with the assumption that visual as well as acoustic lexical neighborhoods influence AV performance, lead to 3 testable hypotheses. First, a word that has few words in the overlapping regions of the auditory and visual neighborhoods will have a greater likelihood of being

recognized correctly (Figure 2b) in the AV condition than a word that has many words (Figure 2a) in adverse listening conditions (ie, in the presence of 6-talker babble). Second, the density of the acoustic and visual lexical neighborhoods will predict how well words will be recognized in an AV condition. Third, acoustic lexical density will be predictive of A performance but not of V, whereas visual lexical density will be predictive of V performance but not of A.

Current Study

The purpose of the present investigation was to address these 3 hypotheses. The Children's Audiovisual Enhancement Test¹⁸ was administered to a group of young adults, older adults who have normal hearing, and older adults who have impaired hearing. The Children's Audiovisual Enhancement Test consists of words that have been shown to be highly recognizable in a V condition and words that have been shown to be poorly recognizable in a V condition. In this experiment, we determined the density of the acoustic and visual lexical neighborhoods for each of the test items. We used these density measures and their intersection densities (ie, the number of words that are both acoustic and visual neighbors of the target) to predict A, V, and AV performance. If distinct acoustic and visual lexical neighborhoods exist, then both should account for some of the variance seen in AV performance. Moreover, the density levels of the acoustic lexical neighborhoods should be predictive of A performance whereas the density levels of the visual lexical neighborhoods should be predictive of V performance.

Kaiser et al¹⁹ examined lexical effects on audiovisual word recognition by adults who have normal hearing and adults who use cochlear implants with a test that included words from lexically dense neighborhoods and lexically sparse neighborhoods. In this study, neighborhood density was based on acoustic measures only. The investigators examined visual enhancement for the 2 types of words. Visual enhancement was based on the equation presented by Sumby and Pollack² and represents the amount of improvement obtained in an AV condition compared with an A condition, normalized to the A performance. The investigators did not find a significant difference in visual enhancement between the 2 word types, although the comparison just missed significance. The present work differs from this experiment because lexical neighborhoods were defined independently for visual and auditory presentations, whereas Kaiser

et al¹⁹ used neighborhoods based solely on the acoustic/ phonetic characteristics of words.

Methods

Participants

One hundred and thirty-one adults were tested. All participants were tested as part of a larger study investigating audiovisual integration in young and older adults (see Tye-Murray et al⁹ and Sommers et al³). Participants were divided into 3 groups: young normally hearing persons (YNH, $n = 52$), older normally hearing persons (ONH, $n = 53$), and older adults having a mild to moderate sensorineural hearing loss (OHI, $n = 26$). The YNH participants (32 females and 20 males, mean age = 21.3 years, $SD = 2.1$, minimum = 18, maximum = 26) were recruited through the participant pool maintained by the Department of Psychology, Washington University. The ONH (36 females and 17 males, mean age = 73.3 years, $SD = 4.5$, minimum = 65.6, maximum = 85.2) and OHI (18 females and 8 males, mean age = 74.1 years, $SD = 7.6$, minimum = 65.5, maximum = 91.8) participants were community-dwelling individuals who were recruited from the participant pool maintained by the Aging and Development program at Washington University. No age difference between the ONH and the OHI groupings was found ($t_{77} = .552$, $P = .583$). All participants were paid \$10 per hour or assigned class credit for participating in the study. All participants were screened to include only those with visual acuity or corrected acuity equal to or better than 20/40 using the standard Snellen eye chart and contrast sensitivity better than 1.8 as assessed with the Pelli-Robison Contrast Sensitivity Chart. Older participants were also screened for dementia (Mini-Mental Status Exam) and normal intelligence (Wechsler Adult Intelligence Scale–V). Participants taking medications that might affect the central nervous system were excluded.

All young adults indicated good or excellent hearing in both ears via questionnaire. Older participants were screened for hearing acuity via pure-tone audiometry. Only older participants with a pure-tone average (PTA; average of pure tone threshold values at 500, 1000, and 2000 Hz) between 30 and 55 dB HL were considered for inclusion in the OHI group (mean PTA across all ears = 43.2, $SD = 6.1$). OHI participants did not use

amplification during testing. ONH participants were required to have PTAs better than 25 dB HL in both ears (mean across all ears = 14.0, SD = 6.5). Volunteers with interoctave slopes greater than 15 dB at frequencies between 500 and 4000 Hz were screened out.

The Children's Audiovisual Enhancement Test (CAVET)

The CAVET¹⁸ consists of 3 word lists of 20 words each along with 10 practice items. The test words are embedded in the carrier phrase, "Say the word . . ." Each test word has between 1 and 3 syllables. The test items are head-and-shoulder recordings of a single female talker with general American dialect. All of the test words have been shown to be familiar vocabulary to fourth-grade children who have significant hearing loss. Test words for the CAVET were selected on the basis of their visibility. During pilot testing, a large corpus of words were videotaped and shown to a group of 50 undergraduate college students. Based on the students' ability to lip-read the words, a subset of words was extracted from the larger corpus and labeled as "difficult" because between 10% and 35% of the students recognized them. Another subset was labeled as "easy" because between 40% and 90% of the students recognized them. Although this division of items into visually easy and hard words resulted in a somewhat larger range of performance for easy than for hard words, it enabled the CAVET to have equivalent numbers of words in each category that would be highly familiar to both normal-hearing and hearing-impaired listeners. The words were divided into 3 test lists such that each list contains 10 difficult test words and 10 easy test words. Examples of easy words include *family*, *bath*, and *butterfly*. Examples of difficult words include *line*, *kiss*, and *vegetable*. Subsequent research with the CAVET has confirmed that participants lip-read the easy words significantly better than the difficult words and that the 3 lists yield equivalent performance in a V condition.²⁰

Procedures

Participants were tested in a sound-treated room. They sat approximately 0.5 m from a 17-inch Touchsystems monitor (ELO-17OC) and responded to the test stimuli by repeating the target words. Stimuli were presented in a full screen mode. Presentation, scoring, and

Table 1. Conversion Values Used to Convert the Phonetically Transcribed Database to a Visemically Transcribed Database

Phoneme Group	Viseme Group
/p, b, m/	[P1]
/f, v/	[P2]
/θ, ð/	[P3]
/j, ʒ, dʒ, tʃ/	[P4]
/w, r/	[P5]
/l/	[P6]
/t, d, s, z, n, k, g, j, ŋ/	[P7]
/h/	[P8]
All vowels	[V]

Note: For example, /p, b, m/ are all indistinguishable when produced on the lips and therefore become transcribed as a single viseme (P = phoneme group).

experiment flow were all conducted using software written in LabView (National Instruments) for audio-visual speech perception studies. Only words repeated verbatim were counted as correct. The CAVET test lists were counterbalanced across the 3 testing conditions of A, V, and AV. All conditions were presented along with 6-talker babble (the V condition also included background babble for consistency across the conditions). Background noise was used to avoid ceiling effects in the audiovisual condition and to help equate performance in the A condition to between 40% and 50% across participants. Stimulus level was held constant at 60 dB SPL. The background babble was adjusted for each individual during a pretesting phase using a modified American Speech-Language-Hearing Association speech-reception-threshold procedure²¹ to obtain approximately 50% correct identification in the A condition. The same babble level was then used for all 3 conditions (A, V, and AV). Signal-to-noise ratios averaged -8.6 dB (SD = 1.2) for the YNH group, -6.5 dB (SD = 1.7) for the ONH group, and -2.4 dB (SD = 2.4) for the OHI group.

Quantification of the Acoustic and Visual Lexical Neighborhoods

Auditory neighborhood density was determined from the Lexical Neighborhood Database maintained at the Psychology Department at Washington University (<http://neighborhoodsearch.wustl.edu>). This database indicates the neighborhood density for a particular word as determined by a single addition, deletion, or substitution rule (ie, any word that can

be created from a target word by adding, deleting, or substituting a single phoneme is operationally defined as belonging to the neighborhood of the target item). The entire database (19319 words) was recoded to convert the available phonetic transcription of each word into a visemic transcription using the coding strategy adapted from Lesner et al,¹⁶ Table 1). No attempt was made to code vowel visemes, because vowels tend to have poor visibility on the face.¹⁷ This allowed us to determine the visual lexical neighborhoods of the test words included in the CAVET. The visual neighborhoods that we defined here might be conservative because we only included homophonous words in a neighborhood and not words that began with visually similar mouth gestures but ended differently than the target word. We identified the homophenes for each word and took the total number of each to define the size of the visual lexical neighborhood. For example, the words *sit* and *cat* were recoded to be [P7][V][P7] and would be considered part of the same visual lexical neighborhood. Six of the 60 words on the CAVET were not available for this analysis because they were compound words (ie, *newspaper*, *policeman*) or contained multiple words (ie, *ice cream cone*, *birthday cake*). Thus, the visual lexical neighborhood analysis yielded data for 54 of the CAVET words.

For each test word, we also determined intersection density. This was the number of words in the overlapping regions of its auditory and visual neighborhood and was computed by counting the number of words that were found in both neighborhoods.

Results

We first computed the number of homophenes per word in the CAVET test lists. The average number of homophenes for the easy words (as determined in developing the CAVET¹⁸) was 12.6 (SD = 18.8), whereas the average for the hard words was 110.3 (SD = 95.3). A *t* test indicated that this difference was significant ($t_{52} = 5.13$, $P < .0001$). Thus, as predicted, the words that have been labeled as easy to identify in a V condition have sparse visual neighborhoods, whereas the words that have been labeled as difficult have dense visual neighborhoods.

The average percent correct scores for the 3 groups of participants appear in Figure 3 as a function of word type and condition for each group. For each of the 3 conditions (A, V, AV), a 2-way analysis of variance (ANOVA) was conducted to examine

Figure 3. The percent words correct for the Children's Audiovisual Enhancement Test easy and hard words for the young normally hearing persons (YNH), older normally hearing persons (ONH), and older adults having a mild to moderate sensorineural hearing loss (OHI) in the 3 test conditions. Error bars indicate standard error. A = audition only; V = vision only; AV = audition plus vision.

differences as a function of participant group and word difficulty. The 3 groups did not differ significantly in

the A condition ($F_{2,256} = 2.41, P > .05$), indicating that the procedure to equate performance across the groups in the A condition was successful. As predicted, words from visually dense neighborhoods (hard words) were significantly more difficult to identify than words from visually sparse neighborhoods (easy words) in all 3 conditions (A, $F_{1,256} = 208.4, P < .01$; V, $F_{1,256} = 168.1, P < .01$; AV, $F_{1,256} = 208.9, P < .01$), and no interactions between group and word difficulty were found for any condition. For the AV condition, a group difference in overall performance was revealed ($F_{2,256} = 3.08, P = .05$). Bonferroni-Dunn corrected post hoc testing, however, did not find any differences in group comparisons that reached a conservative level of significance (ie, all P values $> .017$). The ANOVA for the V condition indicated differences between the groups ($F_{2,256} = 17.33, P < .01$). Bonferroni-Dunn corrected post hoc testing indicated that the group differences in V scores were attributable to the difference between the ONH and the YNH groups ($P < .017$) and the difference between the OHI and ONH groups ($P < .017$).

To test the first hypothesis, we divided the entire list of 54 CAVET words into 2 subgroups, creating a division at the median value for intersection density (median = 5 words). The 26 high-intersection density words had between 6 and 16 items (mean = 10.7, SD = 3.0) in the overlap between their visual and auditory neighborhoods. The 28 low-intersection density words had between 1 and 5 words (mean = 2.6, SD = 1.7) in the intersection of their visual and auditory neighborhoods. A t test was performed to compare performance in the AV condition between the 2 subsets. As predicted, words that had few items in the overlapping regions of their auditory and visual neighborhoods were recognized significantly more often than words that had many items ($t_{52} = 3.3, P < .01$).

We next determined whether auditory and visual neighborhood densities were independently predictive of AV performance. The acoustic neighborhoods ranged in size from 0 words to 36 words with a mean of 15.8 words (SD = 11.2). The visual neighborhoods ranged in size from 0 words to 224 words with a mean of 63.2 words (SD = 85.0). A hierarchical regression analysis was performed to predict AV performance with the predictor variables of both auditory and visual neighborhood density. The analysis revealed that auditory neighborhood density accounted for 19% of the variance ($F_{1,52} = 12.43, P < .01$), whereas visual neighborhood density accounted for an additional 11% (F change $_{1,51} = 7.98, P < .01$). These results are consistent with our second hypothesis, namely, that the 2 kinds of neighborhood density would be predictive of AV performance.

The last hypothesis, that auditory neighborhood density would predict A performance and visual neighborhood density would predict V performance, was addressed by performing 2 additional hierarchical regression analyses. Visual neighborhood density explained 16% ($F_{1,52} = 9.767, P < .01$) of the variability in the V scores, whereas auditory neighborhood density did not provide any addition to the model (F change $_{1,51} = .003, P = .648$). Similarly, auditory neighborhood density explained 15% ($F_{1,52} = 9.13, P < .01$) of the variance in the A scores, and visual neighborhood density did not explain any additional variance (F change $_{1,51} = 1.96, P = .167$). In summary, the variability in V scores across the 3 groups, beyond that accounted for by visual neighborhood density, could not be explained by auditory neighborhood density. Additionally, the variability in A scores, beyond that attributed to auditory neighborhood density, was not explained by visual neighborhood density.

Discussion

Persons who have both hearing and vision deficits are at a double disadvantage when it comes to recognizing speech. They receive neither a distinct auditory nor visual signal, so both listening and speech reading suffer as a result. For these individuals, the superadditive effects to be gained by combining what they hear with what they see might be especially beneficial, yet their sensory impairments may make such benefits particularly difficult for this group to attain. In this investigation, we took a step toward explaining the superadditive effect by demonstrating that superadditivity may be, at least in part, an epiphenomenon of the interaction between acoustic and visual lexical neighborhoods during ongoing speech recognition.

Acoustic lexical neighborhoods have received significant attention in the literature and have garnered much experimental support,¹⁶ but the possible existence of visual lexical neighborhoods has received much less attention.¹⁵ The present investigation provides additional evidence in support of visual lexical neighborhoods. The analyses presented here showed that the visual neighborhood density of words predicts performance in a V condition, whereas auditory neighborhood density does not. Conversely, auditory neighborhood density predicts performance in an A condition, whereas visual density does not. Both types are predictive of AV performance. Moreover, words that have few items in the overlapping regions of their intersections are more likely to be recognized correctly in an AV condition than words that have many items. These

results suggest that when an individual is asked to recognize a word in an AV condition, there may be a simultaneous activation of the acoustic and visual lexical neighborhoods, leading to a winnowing away of candidate word alternatives as the speech signal unfolds.

It remains to be determined whether AV speech perception includes a distinct integrative stage and whether some persons have better integrative skills than do others. By understanding how the properties of the test stimuli influence AV and V performance, however, as we attempted to do in this experiment, we can better describe what remains to be performed by an integrative mechanism above and beyond that which can be explained by the properties of the stimuli themselves.

Persons with dual sensory loss may experience shifts in the boundaries of both their acoustic and visual lexical neighborhoods. One consequence of this shift in both types of neighborhood characteristics is that compared with individuals who have deficits in only a single perceptual system (eg, only visual or only auditory impairments), AV enhancement may be diminished because the intersection of A and V neighborhoods may include many more candidate items. To address this directly, future investigations will assess how simulated auditory and visual impairments affect AV performance and how they affect lexical neighborhoods. Additionally, because auditory and visual neighborhoods combined accounted for less than 50% of the variance in AV performance, we will examine how stimulus factors other than neighborhood characteristics (eg, frequency) also contribute to combined auditory–visual speech perception. The eventual goal of these investigations will be to identify how the sensory and cognitive abilities mediating AV speech perception interact with stimulus characteristics, so we can design rehabilitation programs that target specific impairments for those individuals with impaired visual and auditory processing capabilities.

Acknowledgment

This work was supported by grant RO1 AG-180291 from the US National Institutes of Health, National Institute on Aging, Bethesda, Maryland.

References

1. Grant KW, Walden BE, Seitz PF. Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *J Acoust Soc Am*. 1198;103:2677-2690.
2. Sumby WH, Pollack I. Visual contributions to speech intelligibility in noise. *J Acoust Soc Am*. 1954;26:212-215.
3. Sommers MS, Tye-Murray N, Spehar B. Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear Hear*. 2005; 26:263-275.
4. Tyler RD, Preece J, Tye-Murray N. *The Iowa Laser Videodisk Tests*. Iowa City, Iowa: University of Iowa Hospitals; 1986.
5. Summerfield Q. Visual perception of phonetic gestures. In: Mattingly IG, ed. *Modularity and the Motor Theory of Speech Perception*. Hillsdale, NJ: Laurence Erlbaum; 1989:117-137.
6. Massaro D. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press; 1998.
7. Ouni S, Cohen M, Ishak H, Massaro D. Visual contribution to speech perception: measuring the intelligibility of animated talking heads. *Eurasip J Audio Speech Music Process*. In press.
8. Cienkowski KM, Carney AE. Auditory–visual speech perception and aging. *Ear Hear*. 2002;23:439-449.
9. Tye-Murray N, Sommers M, Spehar B. Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear Hear*. In press.
10. Braida L. Crossmodal integration in the identification of consonant segments. *Psych Quart J Exp Psych*. 1991;43: 647-677.
11. Grant KW. Measures of auditory-visual integration for speech understanding: a theoretical perspective. *J Acoust Soc Am*. 2002;112:30-33.
12. Sommers MS, Spehar B, Tye-Murray N. The effects of signal-to-noise ratio on auditory-visual integration: integration and encoding are not independent. *J Acoust Soc Am*. 2005;117:2574.
13. Blamey PJ, Cowan RSC, Alcantara JI, Whitford LA, Clark GM. Speech perception using combinations of auditory, visual, and tactile information. *J Rehab Res Devel*. 1989;26:15-24.
14. Luce PA, Pisoni DB. Recognizing spoken words: the neighborhood activation model. *Ear Hear*. 1998; 19:1-36.
15. Mattys SL, Bernstein LE, Auer ET. Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Percept Psychophys*. 2002;64:667-679.
16. Lesner SA, Sandridge SA, Kricos PB. Training influences on visual consonant and sentence recognition. *Ear Hear*. 1987;8:283-287.
17. Berger KW. *Speechreading: Principles and methods*. Kent, Ohio: Herald; 1972.
18. Tye-Murray N, Geers A. *Children's Audiovisual Enhancement Test*. St. Louis, MO: Central Institute for the Deaf; 2001.
19. Kaiser AR, Kirk KI, Lachs L, Pisoni DB. Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *J Speech Lang Hear Res*. 2003; 46:390-404.
20. Spehar B, Tye-Murray N, Sommers M. Time-compressed visual speech and age: a first report. *Ear Hear*. 2004;25:565-572.
21. American Speech-Language-Hearing Association. Guidelines for determining threshold level for speech. *ASHA*. 1988; 30:85-89.