THE INFLUENCE OF THE LEXICON ON VISUAL SPOKEN WORD RECOGNITION

Edward T. Auer, Jr.[1], *Lynne E. Bernstein*[1], *& Sven Mattys*[2]

[1]House Ear Institute, 2100 W. Third St., Los Angeles, CA, [2]University of Bristol, 8 Woodland Rd., Bristol, England

ABSTRACT

In this paper, we report on experiments that investigated form-based similarity effects in visual spoken word recognition. Specifically, we tested whether accuracy of speechreading a word was related to the number of words (neighbors) perceptually similar to that stimulus word and to its frequency of occurrence. In the first Experiment, the Neighborhood Activation Model (NAM) [1,2] was adapted to generate predictions about the accuracy of visual spoken word identification. In the second Experiment, we used the concept of the Lexical Equivalence Class Size [3] to generate predictions regarding the accuracy of visual spoken word recognition. Both experiments provided evidence that words are identified more accurately if they have few neighbors and occur frequently in the language. Correlational analyses provided evidence that a word's neighbors, or close competitors, are based on perceptually defined similarity. The results of the current experiments are interpreted as evidence of a common spoken word recognition system for both auditory and visual speech information, which retains sensitivity to form-based stimulus similarity among words.

1. INTRODUCTION

Models of spoken word recognition posit a framework in which phonetic stimulus information activates multiple word candidates as a function of their form-based similarity to the stimulus. Word recognition is achieved via competition among the active candidates [1,2,4,5]. These models predict that a word's recognition accuracy will be determined in part by the number of words similar to it in the mental lexicon. Specifically, words similar to many other words will be harder to recognize than words similar to few other words. Empirical evidence supporting this prediction for auditory spoken word recognition has been obtained across several different recognition tasks [e.g., 2]. Although, this prediction should theoretically be equally applicable to visual spoken word recognition, the prediction has been tested only for the auditory modality.

At issue here is whether word similarity is determined by form-based stimulus similarity or by abstract linguistic relationships (e.g., feature communality). One current model of spoken word recognition, NAM [1,2], specifically predicts that a word's competitor environment is defined on the basis of form-based or perceptual similarity among words. Visual speech provides the opportunity, which we have taken in Experiment 1, for a strong test of whether word similarity is based on form-based similarity; because visual speech provides information complementary to that available in auditory speech degraded by noise. For example, acoustic speech presented in noise maintains voicing distinctions but reduces the cues for place of articulation. Thus, auditorily, *peek* is more similar to *teak* than it is to *beak*. In contrast, visual speech reduces the voicing contrast but transmits reliable cues for place of articulation. Thus, *peek* is visually more similar to *beak* than it is to *teak*.

2. EXPERIMENT 1

The NAM [1,2] for auditory spoken word recognition is a computationally explicit model for investigating the combined effects of the phonetic stimulus and lexical properties. In the NAM (see [1] and [2]; for a complete description), stimulus input activates a set of acoustic-phonetic patterns in memory. The activation levels of the memory patterns are a function of their similarity to the phonetic stimulus. Acoustic-phonetic input patterns then activate a set of word decision units tuned to the input patterns. Once activated by bottom-up phonetic input, these word decision units continuously compute a decision value. The decision values are computed with a frequency-biased, activation-based version of R. D. Luce's [6] choice rule,

$$p(ID) = \frac{p(S \mid S_c) * freq_s}{\{p(S \mid S_c) * freq_s\} + \sum_{j=1}^{n} \{p(N_j \mid S_c) * freq_{Nj}\}},$$

where *p(ID)* is the probability of correctly identifying the stimulus word, *p(S|Sc)* is the support for a stimulus word based on its constituent segments, $freq_s$ is the stimulus word's frequency of occurrence, *p(Nj|Sc)* is the support for the neighbor word *j* based on the stimulus word's constituent segments, and $freq_{Nj}$ is neighbor word *j*'s frequency of occurrence (see Equation 6 in [2]). The word decision value for a given acoustic-phonetic pattern

is hypothesized to be related to word recognition accuracy, such that high word decision values predict easier word recognition.

NAM's input representation can be applied to a variety of perceptual conditions for which phonetic information is impoverished, including speechreading. Experiment 1 used NAM to predict the influences of variation in optical-phonetic and lexical properties of spoken words on the accuracy of visual-only word identification. To adapt NAM to speechreading, previously collected speechread nonsense syllable identification data were used to define the input to the model (see below for details).

2.1. Method

2.1.1. Participants

All participants self-reported English as a native language, had vision 20/30 or better in each eye, as determined with a standard Snellen chart, and were paid for their participation.

Hearing participants. Twelve participants were recruited from the campuses of University of Southern California, California State University Northridge (CSUN) and the staff of St. Vincent's Hospital, LA, CA. Seven were female. The group mean age was 22.65 (range 18.4 to 33.9) years. The speechreading screening test resulted in mean percent words correct in sentences of 19 (range 3.11 to 49.42).

Deaf participants. Thirteen deaf participants were recruited from California State University Northridge (CSUN). All deaf participants reported use of English as their family's primary language and education in a mainstream and/or oral program for eight or more years. One was dropped due to a technical malfunction. Of the remaining twelve, half were female. The mean age across the group was 21 (range 18.0 to 24.7). All participants reported age at onset of hearing impairment was prior to 2 years, with the majority reporting congenital impairment. Ten participants had 90 dB HL or greater pure tone averages (profound hearing impairment) in both ears. Participants had average or better performance on a speechreading screening test. The mean percent words correct for speechread screening sentences was 53 (range 36.96 to 71.6).

2.1.2. Materials and Procedure

The stimuli were audiovisual recordings of monosyllabic spoken words spoken by a male talker of American English and stored on laserdisc [7]. The talker was seated before a dark background, with his face filling most of the recording frame. He was illuminated by two floodlights at oblique angles to his face, and by two fill lights positioned above his head.

NAM output values were computed for each of 153 consonant-vowel-consonant words. Each stimulus word was compared against 1506 (consonant-vowel-consonant) monosyllabic words drawn from the PhLex online lexical database [10].

The numerator (or segmental intelligibility) of the NAM output equation was computed as follows: The conditional probability of correctly identifying each of the word's segments was computed based on the nonsense syllable confusion matrices for visible speech segments. Assuming independent probabilities within cells, the conditional probabilities for each of the word's segments were multiplied to calculate an index of the word's overall segmental intelligibility. For example, the segmental intelligibility of the stimulus word "bit" is the product of the independent conditional probabilities of p(b|b), p(I|I), and p(t|t). Previously collected phoneme identification data were used to estimate phonetic similarity [8,9]. For the consonant identification data, the stimuli were recordings of C-/ɑ/ syllables. For the vowel identification data, the stimuli were recordings of /h/-V-/d/ syllables. The talker and stimulus recording conditions for the nonsense syllables were identical to those used in the current word identification study.

The denominator (or neighborhood density) was also computed using conditional probabilities obtained from nonsense syllable confusion matrices. The probability of identifying each of a stimulus word's neighbors was computed as follows: The conditional probability of identifying each of the neighbor word's segments, given the stimulus word's segments was obtained from the nonsense syllable confusion matrices. Again, assuming independent probabilities, the conditional probabilities for each of the neighbor word's segments were multiplied to calculate the neighbor word probability. For example, the probability of the visual neighbor word "pill," given the stimulus word "bit", p(pɪl|bɪt), is the product of the independent conditional probabilities p(p|b), p(ɪ|ɪ), and p(l|t). In addition, neighbor word probabilities were multiplied by their respective frequency of occurrence in English. Word frequencies were transformed (see [11]) using $f(p_i) = 40 + 10*(\ln p_i + 10)$, where p_i is the frequency of word *i* in [12], expressed as a proportion (i.e., p_i is the frequency of word *i* divided by the sum the frequencies of all words in PhLex [10]). For a given stimulus word, all of its neighbor word probabilities were then summed to arrive at a predicted neighborhood density.

To examine neighborhood density effects with segmental intelligibility controlled, the word list

contained two sets of 44 words each matched on segmental intelligibility [$t(86) = .043$, $p = .96$] and word frequency [$t(86) = 1.92$, $p = .058$], and contrasted on neighborhood density (high versus low) [$t(86) = 22.17$, $p < .001$]. Although the word frequency of the two word groups is marginally different, it is dense words that have the higher average word frequency. Because high frequency are easier to recognize, this marginal difference could make it more difficult to observe the density effect.

Procedure. Participants were tested individually in a quiet room. They were seated in front of a computer monitor and were given verbal and written instructions. A research assistant, skilled in sign language, administered verbal instructions to the deaf participants using simultaneous communication (i.e., signs in English word order produced in synchrony with speech). Participants were informed that they would be seeing a series of one-syllable English words presented one at a time. A trial consisted of the video presentation of a single monosyllabic word followed by the presentation of a prompt on the computer screen for the participants to type a response. Participants were instructed to enter a guess if they were not sure of what the word was. After they entered their response and pressed the "ENTER" key, a short pause preceded the presentation of the next trial.

2.3. Results and Discussion

Prior to analysis, the responses were corrected for obvious typographic errors and then all word responses were phonemically transcribed by software lookup. A response was considered correct only if its entire phonemic transcription matched the transcription of the stimulus word. An item analysis was performed to examine the level of identification accuracy over the entire set of stimulus words by each participant group. In each participant group, the percentage of participants correctly identifying each word was used for the item analyses. The deaf participants, who were screened to be skilled speechreaders, were more accurate than the randomly selected hearing participants (deaf: item mean percent correct 28%, range 0-100%; hearing: item mean percent correct 13%, range 0-83%) [$t(304) = 5.826$, $p < .001$]. Both participant groups had extremely wide ranges of performance, but identification accuracy at the item level was significantly correlated between the two participant groups [$r = .64$, $p < .01$]. In the deaf participant group, three of the 153 items were correctly identified by all of the participants, and 12 items were identified correctly by three quarters or more of the participants.

To assess NAM's predictions for visual spoken word identification accuracy, correlational analyses were performed between NAM output values, word frequency, and item identification accuracy within each group. Table 1 displays the Pearson correlation coefficients for the hearing and deaf groups, respectively. In both groups, the output of NAM was significantly correlated with the identification accuracy. NAM output values correlated with visual spoken word identification accuracy at levels comparable to correlations reported previously for auditory spoken word identification ($r = .23$ to $r = .47$) across a range of signal-to-noise ratios [4]. Word frequency was correlated with item percent correct for the deaf participant group only.

Participant Group	NAM Output	Word Frequency
Hearing	.44*	.14
Deaf	.48*	.28*

Table 1: Pearson correlation coefficients between item accuracy and lexical variables for both participant groups Note. * $p < .01$.

Figure 1 displays the results for the word sets that varied on predicted neighborhood density (high vs. low) but were controlled for segmental intelligibility. Mean percent words correct is displayed as a function of neighborhood density

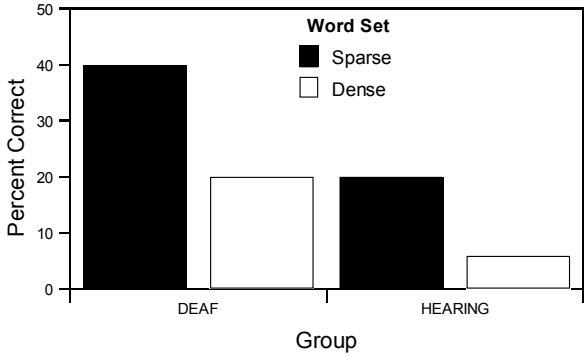

Figure 1: Mean percent words correct as a function of stimulus neighborhood density (sparse, dense) and participant group (deaf, hearing).

(high, low) and participant group (deaf, hearing). In the deaf group, words with many neighbors were more difficult to identify than words with few neighbors [$F(1, 11) = 68.54$, $p < .001$]. Likewise, in the hearing group, words with many neighbors were more difficult to identify than words with few neighbors [$F(1, 11) = 63.58$, $p < .001$]. A separate ANOVA was performed to examine participant group effects. The main effect of participant group was significant [$F(1, 22) = 26.87$, $p < .001$] and the participant group by neighborhood density factor interaction was significant [$F(1, 22) = 5.62$, $p = .027$].

These results demonstrated that the accuracy of visual spoken word identification is influenced by the number of perceptually similar words in the mental lexicon. Just as with auditory spoken word recognition, words in dense neighborhoods were more difficult to identify than words in sparse neighborhoods. Thus, our results are consistent with contemporary models of auditory spoken word recognition that posit stimulus-driven activation of multiple word candidates and competition among the active candidates during the recognition of a single candidate word. For an expanded analysis and discussion of the results of Experiment 1, see [13].

NAM specifically predicts that neighborhood effects obtained in the current study are the result of the **form-based similarity** of words [4]. Thus, a word's neighborhood can vary as a function of the stimulus presentation conditions. As noted above, visual speech provides the opportunity for a strong test of this prediction. It was hypothesized that if NAM output values depended critically on visual-phonetic similarity, output values computed using confusion data obtained for noise-degraded acoustic speech should not correlate with visual word identification accuracy.

To investigate this hypothesis, a previously collected set of confusions for acoustic speech presented in noise [1] were used as input to NAM. A set of confusions was chosen that roughly matched the visual confusion data on identification accuracy for consonants [visual = 48% correct; auditory = 51% correct; +5 S/N] and vowels [visual = 51% correct; auditory = 67% correct; + 5 S/N]. NAM_{audio} output values computed with the acoustic confusion data were not significantly correlated with visual word identification accuracy in the hearing group [r = .046], but were correlated for the deaf group [r = .28, $p < .05$]. However, the calculation of NAM output equation includes word frequency, and frequency was correlated with visual word identification scores for the deaf group but not the hearing group. Therefore, NAM output values were recomputed leaving word frequency out. In the new analysis, the correlations between identification accuracy and NAM_{audio} were no longer significant in either group [r = -.11, hearing group; r = .11 deaf group] but, identification accuracy and NAM_{video} remained significant for both participant groups [r = .33, hearing group; r = .24 deaf group]. This pattern of results supports NAM's prediction that a word's competitor environment depends on stimulus-based perceptual similarity.

3. EXPERIMENT 2

In Experiment 2 we sought to replicate and extend the results of Experiment 1. As in Experiment 1, words were presented that varied in the number of words perceptually similar to the stimulus word. However, in Experiment 2 the stimuli were spoken by a new talker and included familiar monosyllabic and disyllabic words with both high and low frequencies of occurrence. In addition, the new set of words was selected using lexical equivalence class (LEC) size and not NAM output values. Our previous computational research provided evidence that stimulus-based lexical dissimilarity can reduce the problem of phonetic impoverishment in visual spoken word recognition [3]. This research lead to the development of the concept of the lexical equivalence class size as an index of lexical form-based similarity. Differences exist between the computation of size of LEC and neighborhood density. For example, LEC computation sets an explicit threshold on those words considered to be competitors for recognition, whereas NAM does not. For the current purposes, LEC size provides a convenient method for estimating the number of words perceptually similar to a stimulus word. In Experiment 2, we investigated form-based similarity effects in visual word recognition by considering words from three classes of visual equivalence: words predicted to have no visual within-class competitors (LEC size of 1), words predicted to have a few competitors (LEC size of 2 to 6), and words predicted to have many competitors (LEC size of 10 to 60).

3.1. Method

3.1.1. Participants

A different set of participants who met the same requirements as did the participants in Experiment 1 were recruited and were paid for their participation.

Hearing participants. Eight hearing participants were recruited among undergraduate students at California State University, Northridge (CSUN). The group mean age was 23.5 years (range 21 to 28).

Deaf participants. Eight deaf participants were recruited from among undergraduate students at CSUN. The group mean age was 22.5 years (range 19 to 26). All participants reported age of hearing impairment onset prior to 2 years of age with the majority at birth. All participants had 95 dB HL or greater pure tone averages in both ears.

3.1.2. Materials and Procedure

The stimuli were monosyllabic words and initial-stress, disyllabic words chosen from the 35,000-word PhLex [10]. They were spoken by a female talker and recorded under conditions similar to those detailed in Experiment 1. They contrasted on their frequency of occurrence in English (low-frequency

vs. high-frequency) and on the number of words in their lexical equivalence class (unique, medium, and large LEC sizes).

Lexical confusability was estimated using the computational method in [3]. The method involves three steps: (1) Rules are developed to retranscribe words so that their transcriptions represent only the segmental distinctions that are estimated to be visually perceivable. The retranscription rules are in the form of *phoneme equivalence classes* (PECs). (2) Retranscription rules are applied to the words in a phonemically transcribed, computer-readable lexicon. (3) The retranscribed words are sorted so that words rendered identical (no longer notationally distinct) are placed in the same *lexical equivalence class* (LEC). PECs corresponded generally to conventionally obtained viseme classes.

The 282 test words were organized into three LEC-size categories, based on their retranscribed format: *Unique-LEC* words were from lexical equivalence classes of size one, that is, they did not have any visual competitors. *Medium-LEC* words were from LECs of size two to six. *Large-LEC* words were from LECs of size 10 to 60. Because of the limited number of available disyllabic words in large LECs, only 32 stimuli were used in the disyllabic low-frequency (16) and high-frequency (16) large-LEC categories. This limitation influenced the overall frequency of the stimuli in these categories, with a resulting lower mean frequency for the high-frequency stimuli of the large-LEC disyllable category than for the high-frequency stimuli in the two other categories. Another consequence was that the mean LEC size of the large-LEC words was 34.4 for the monosyllables and only 13.2 for the disyllables.

Procedure. All participants were tested individually in a quiet room. They were seated in front of a computer monitor and were given verbal instructions. A certified sign language interpreter or a deaf research assistant administered the instructions to the deaf participants using English signs in synchrony with speech. The 282 videorecorded words, presented one at a time, were spoken by a female talker, with her face filling most of the monitor frame. Words were presented in four blocks. Two blocks contained the monosyllables (75 words in each), and the other two contained the disyllables (66 words in each). Proportions of high- *versus* low-frequency words, and unique-, medium-, and large-LEC words were identical across all blocks. Block presentation order was rotated across participants. Within each block, word presentation was randomized for each participant. The experiment began with a practice block of 10 monosyllables and one of 10 disyllables. For both the practice and the experimental blocks, participants were asked to identify each word in an open-set format by typing it in on a computer keyboard. They were told that all of the stimuli were words and were therefore encouraged to provide a word response, but they were allowed to enter a non-word response, if they could not perceive a word that corresponded to the input. After entering a response, participants pressed a keyboard key to see the next word.

All responses were screened and corrected for misspelling or obvious typographical errors. Responses were then coded as "correct" or "incorrect." Incorrect responses included any departure from the target word such as another word, a nonsense word, an untranscribable response (e.g., "wqxa"), or no response. The percentage of correct responses was calculated for each cell of the design, examining group (hearing, deaf), word LEC size (unique, medium, large), word frequency (high, low), and word length (monosyllable, disyllable).

3.3. Results and Discussion

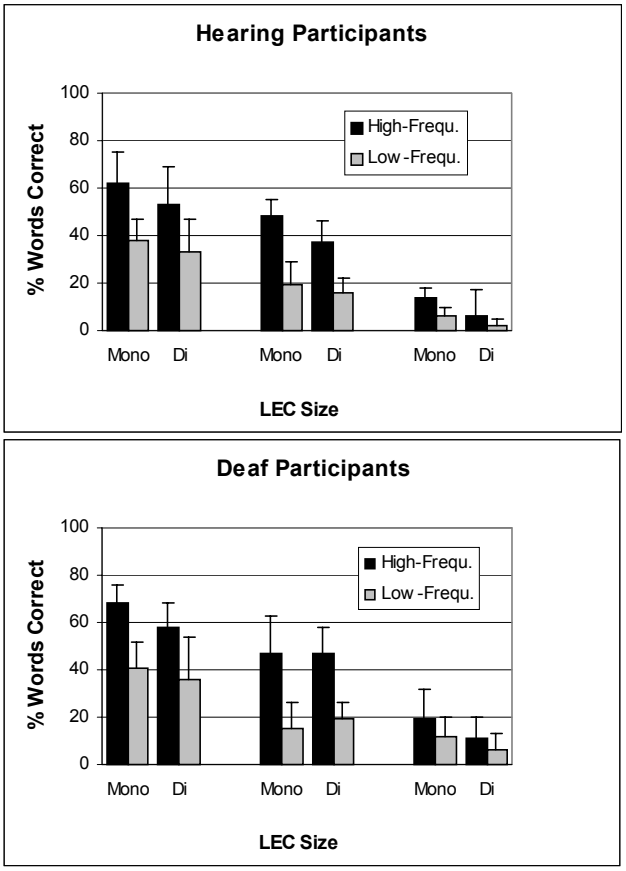

Figure 2: Percent words correct (and error bars) as a function of the LEC size (small to large), length, and frequency of occurrence of the test words. *Upper*: Normal hearing participants. *Lower*: Deaf participants.

The results are plotted in Figure 2. Analyses of variance were performed on the identification scores by subjects. Words were identified more accurately when the LEC size was low [$F(2, 28) = 239.45$, $p < .001$] and when the frequency of occurrence was high [$F(1, 14) = 463.16$, $p < .001$] There was also an advantage for monosyllables over disyllables [$F(1, 14) = 13.14$, $p < .005$]. LEC size interacted with word frequency [$F_1(2, 28) = 34.06$, $p < .001$]. This interaction indicated that the frequency effect was less pronounced among words with a large LEC size than in the two other LEC size categories. This interaction probably results from the combination of the design-induced relatively low frequency of the high-frequency, large-LEC disyllables and a potential floor effect for the words in the low-frequency, large-LEC category. These results clearly indicate that visual spoken word recognition is influenced by the number of visually similar words in the lexicon and by the frequency of occurrence of the test words. For an expanded analysis and discussion of Experiment 2, see [14].

4. CONCLUSION

The current results are consistent with contemporary models of auditory spoken word recognition that posit stimulus driven activation of multiple word candidates and competition among the active candidates to arrive at the recognition of a single candidate word. We interpret these results as evidence of common spoken word recognition processes for both auditory and visual speech information. However, modality-specific form-based similarity is what defines lexical similarity.

5. ACKNOWLEDGMENTS

This work was supported by research grant R01 DC02107 from the National Institutes on Deafness and Other Communication Disorders, National Institutes of Health. We thank Brian Chaney, Sheri Hithe, Jennifer Yarbrough, and Paula E. Tucker for their helpful advice and assistance.

6. REFERENCES

1. Luce, P. A. *Neighborhoods of words in the mental lexicon. (Research on Speech Perception, Technical Report No. 6*). Bloomington, IN: Speech Research Lab, Dept. of Psychol., Indiana University, 1986.

2. Luce, P. A. & Pisoni, D. B. Recognizing spoken words: The neighborhood activation model. Ear & Hearing, 19: pp. 1-36, 1998.

3. Auer, E.T., Jr., and Bernstein, L. E., Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. J Acoust. So. Amer., Vol. 102, 1997, p 3704-3641.

4. Marslen-Wilson, W. D. Access and integration: Projecting sound onto meaning. In W. D. Marslen-Wilson (Ed.), Lexical Representation and Process, MIT, Cambridge, MA, 1992, pp. 3-24.

5. Norris, D., Shortlist: A connectionist model of continuous word recognition. *Cognition, 52, 1994, pp. 189-234.*

6. Luce, R. D.,. *Individual choice behavior. New York: Wiley, 1959.*

7. Bernstein, L. E. & Eberhardt, S. *P., Johns Hopkins Lipreading Corpus I-II, Disc I*, The Johns Hopkins University, Baltimore, MD, 1986.

8. Bernstein, L. E., Coulter, D. C., O'Connell, M. P., Eberhardt, S. P., & Demorest, M. E., Vibrotactile and haptic speech codes. In A. Risberg, S. Felicetti, G. Plant, and K-E. Spens (Eds.), *Proc. Second International Conference on Tactile Aids, Hearing Aids, & Cochlear Implants. (Stockholm), 7-11 June 1992.*

9. Iverson, P., Bernstein, L. E., and Auer, Jr., E. T., Phonetic perception and word recognition. *Speech Comm., 26: pp. 45-63, 1998.*

10. Seitz, P. F., Bernstein, L. E., Auer, E. T., Jr., and MacEachern, M.E.. *PhLex (Phonologically Transformable Lexicon): A 35,000-word computer readable pronouncing American English lexicon on structural principles, with accompanying phonological transformations, and word frequencies.* Copyright 1998, House Ear Institute.

11. Carroll, J. B., An alternative to Julliand's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (SFI). Computer Studies in Humanities & Verbal Behavior, 3: pp. 61-65, 1970.

12. Kucera, H., and Francis, W., Computational Analysis of Present-Day American English. Brown University Press: Providence, RI., 1967.

13. Auer, E. T., Jr., The influence of the lexicon on speechread word recognition: Contrasting segmental and lexical distinctiveness. Submitted.

14. Mattys, S. Bernstein, L. E., and Auer, E. T., Jr., Stimulus-Based Lexical Distinctiveness as a General Word-Recognition Mechanism. Submitted.