ELSEVIER

Speech Communication 26 (1998) 45–63

SPEECH COMMUNICATION

Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition

Paul Iverson *, Lynne E. Bernstein, Edward T. Auer Jr.

Spoken Language Processes Laboratory, House Ear Institute, 2100 West Third St., Los Angeles, CA 90057, USA

Received 30 January 1998; received in revised form 14 June 1998; accepted 16 July 1998

Abstract

Studies of audiovisual perception of spoken language have mostly modeled phoneme identification in nonsense syllables, but it is doubtful that models or theories of phonetic processing can adequately account for audiovisual word recognition. The present study took a computational approach to examine how lexical structure may additionally constrain word recognition, given the phonetic information available under vocoded audio, visual and audiovisual stimulus conditions. Subjects made phonemic identification judgments on recordings of spoken nonsense syllables. Hierarchical cluster analysis was used first to select classes of perceptually equivalent phonemes for each of the stimulus conditions, and then a machine-readable phonemically transcribed lexicon was retranscribed in terms of these phonemic equivalence classes. Several statistics were computed for each of the transcriptions, including percent information extracted, percent words unique and expected class size. The findings suggest that superadditive levels of audiovisual enhancement are more likely for monosyllabic than for multisyllabic words. That is, impoverished phonetic information may be sufficient to recognize multisyllabic words, but the recognition of monosyllabic words seems to require additional phonetic information.

Keywords: Audiovisual perception; Speech perception; Word recognition

1. Introduction

Audiovisual speech perception research has focused mainly on phonetic processing. The principal research problem has been to account for the phonemic categorization of audiovisual nonsense syllables, given the categorization of these syllables under unimodal audio and visual conditions. A main methodology of this work has involved examining the phonemic categorization of audiovisual nonsense syllables that have altered or mismatched audio and visual components (e.g., Green and Kuhl, 1989; Massaro, 1987; Sekiyama and Tohkura, 1991). Computational models have been formulated to attempt to account for how the integration of audio and visual phonetic information leads to the obtained patterns of audiovisual phonemic categorizations (e.g., Braida, 1991; Massaro, 1987). However, the practical significance of audiovisual integration lies with its ability to enhance the intelligibility of words in isolation and in sentences (e.g., Reisberg et al., 1987; Sumby and Pollack, 1954), particularly under conditions of acoustic noise or distortion. Furthermore, it is doubtful that audiovisual word recognition can be adequately understood in terms of phonetic

* Corresponding author. Tel.: +1 213 353 7029; fax: +1 213 413 0950; e-mail: piverson@hei.org.

0167-6393/98/$ – see front matter
PII: S0167-6393(98)00049-1

processing alone, considering that word recognition probably involves additional sources of information, such as the distribution of words in the mental lexicon.

The aim of the present investigation was to examine the interactions of phonetic information and lexical structure in audiovisual word recognition. The process of phonetic integration per se was not examined, but instead computational techniques were used to assess the effects of audiovisual phonetic information on the distinctiveness of words in the lexicon. The approach followed that of Auer and Bernstein (1996, 1997), who examined computationally how the phonetic information available during lipreading affects word intelligibility (see also Altmann, 1990; Carter, 1987; Huttenlocher and Zue, 1984).

Within current audiovisual integration models of phoneme perception (e.g., Braida, 1991; Massaro, 1987), the enhanced intelligibility of audiovisual phonemes, frequently obtained when the acoustic signal is degraded in some way, is determined by the redundancy of audio and visual phonetic structures. For example, Grant and Walden (1996) compared audio, visual and audiovisual intelligibility for consonants under several conditions in which the audio was band-pass filtered. Audiovisual intelligibility was superior to audio intelligibility in all cases, but the enhancement in intelligibility was not monotonic; conditions with similar levels of audio intelligibility were not equally intelligible in the corresponding audiovisual conditions. Enhancement was low when the information in the audio and visual modalities was highly correlated (i.e., the visual modality provided redundant information), and enhancement was high when each modality provided different phonetic information.

However, the perceptual integration of phonetic information from vision and hearing is not the only relevant factor in audiovisual perception of spoken language; the available phonetic information during word recognition is processed in relation to the structure of the mental lexicon. For example, the lexical structure can help to identify accurately the word *sandwich* when /s/ and /ʃ/ are perceptually ambiguous (i.e., the lexicon resolves this ambiguity, because *shandwich* is not a word), but this lexical information is redundant when /s/ and /ʃ/ are perceptually distinct. [1] However, the lexical structure cannot help identify accurately the word *sack* when /s/ and /ʃ/ are perceptually ambiguous (i.e., the lexicon cannot resolve this ambiguity, because *shack* is also a word), and this lexical information is not redundant when /s/ and /ʃ/ are perceptually distinct.

The lexicon thus has the potential to provide information during word recognition because the set of words in the lexicon is only a subset of the combinatorially possible strings of phonemes in the language. Less perceptual resolution is required to recognize words that reside in relatively sparse regions of the lexicon (i.e., few other words in the lexicon are phonemically similar), and more perceptual resolution is required to recognize words that reside in relatively dense regions of the lexicon (i.e., many other words in the lexicon are phonemically similar). For the purpose of predicting the magnitude of gains in audiovisual intelligibility based on unimodal audio and visual information, phonetic integration models (e.g., Braida, 1991; Massaro, 1987) may be sufficient to account for enhancement levels for phonemes, but enhancement levels for words may additionally depend on the lexical structure.

1.1. Lexical structure in human and machine word recognition

Research on spoken word recognition by humans suggests that the accuracy and speed of word identification are influenced by both lexical structure and phoneme intelligibility (Lahiri and Marslen-Wilson, 1991; Luce et al., 1990; McClelland and Elman, 1986; Norris, 1994). Words that are in the sparse regions of the lexicon are easier to identify than words that are in the dense regions (Luce, 1986; Luce et al., 1990), and this supports the notion that human word recognition makes use of lexical constraints. In addition, identifica-

[1] Although effects of redundant lexical information might not be detectable in measures of recognition accuracy, lexical information can influence the speed of word recognition even under conditions with no phonemic ambiguity (Luce, 1986).

tion accuracy is influenced by how frequently words are used in the language. Words are easier to identify when they are frequently used (Savin, 1962), but words are harder to identify when their lexical regions contain other high-frequency words (Luce, 1986; Luce et al., 1990).

The potential constraints afforded by the lexical structure have been examined in the work related to automatic recognition of acoustic speech (Altman and Carter, 1989; Aull and Zue, 1985; Carter, 1987; Huttenlocher and Zue, 1984). Specifically, studies have assessed the feasibility of using broad phonemic transcriptions and/or lexical stress patterns for the initial stages of lexical access (Aull and Zue, 1985; Carter, 1987; Huttenlocher and Zue, 1984). For example, Huttenlocher and Zue (1984) examined what happens to the information in the lexicon when phonemes are transcribed with broad phonetic categories comprising only the manner of articulation (stop, nasal, liquid or glide, strong fricative, weak fricative, vowel). For example, the word *tin* would be transcribed [STOP][VOWEL][NASAL]. At this level of transcription, they found that approximately 33% of a 20,000-word lexicon was uniquely specified (i.e., not transcribed the same as any other word). This was taken as strong evidence that the structure afforded by the distribution of words in the lexicon could effectively constrain the phonetic information required by large vocabulary automatic speech recognition systems.

1.2. The current study

The approach of the current study follows that of Auer and Bernstein (1997), who adapted measures developed in the automatic speech recognition literature (Altman and Carter, 1989; Aull and Zue, 1985; Carter, 1987; Huttenlocher and Zue, 1984) to examine computationally how the phonetic information available through vision alone (i.e., lipreading) affects lexical distinctiveness. The current study differs from Auer and Bernstein in that phonetic information is assessed here under a wider range of stimulus conditions, with the aim of examining how lexical distinctiveness is affected by enhancements in phonetic information due to audiovisual integration.

The specific stimulus conditions of the present study were visual-only lipreading (V), two audio-only conditions in which the audio was processed by a vocoder using two different algorithms (F1A and F2A), and two audiovisual conditions in which the vocoded audio was presented with visual information (F1AV and F2AV). The audio vocoder consisted of an input stage in which a bank of band-pass filters analyzed the speech signal, and an output stage consisting of a set of amplitude modulated phase-locked sinusoids. The frequencies of the output sinusoids matched the center frequencies of the band-pass filters, and the energy within each band-pass filter was used to amplitude-modulate the sinusoid with corresponding frequency. The output retained the gross spectral-temporal amplitude information of the original speech signal, but eliminated fine-grained source characteristics such as F0 variations. In addition, the vocoder algorithms were band-limited; the F1 algorithm had filters with center frequencies ranging between 75 and 900 Hz, and the F2 algorithm had filters with center frequencies ranging between 825 and 2625 Hz.

F2 has long been considered to be more critical to speech intelligibility than is F1 (e.g., Liberman et al., 1967). However, the addition of lipreading information enhances intelligibility more for F1 than for F2, because lipreading errors are less redundant with errors made when listening to F1 (Grant and Walden, 1996). In the present study, the two vocoder algorithms were selected with the expectation that they would produce degraded stimuli with different levels of intelligibility (F1 being less intelligible than F2), and different patterns of phoneme identifications (F2 preserving place information and F1 preserving manner), with the aim of examining how different types of speech signals affect lexical distinctiveness.

Under each stimulus condition, phoneme identifications were collected for audiovisual recordings of spoken nonsense syllables. Following the computational method of Auer and Bernstein (1997), hierarchical cluster analysis (Aldenderfer and Blashfield, 1984; Norusis, 1993) was then applied to these data to find *phonemic equivalence classes* (i.e., groups of phonemes that are

putatively perceptually equivalent) [2] for each stimulus condition. The phonemic equivalence class concept is a generalization (i.e., applicable to any sensory modality) from the lipreading literature, in which phonemes with high visual perceptual ambiguity (e.g., /b/, /p/ and /m/) are referred to as *visemes* (Fisher, 1968; Woodward and Barber, 1960). The phonemic equivalence classes were then used to define rules to retranscribe a phonemically transcribed lexical database of English (PhLex; Seitz et al., 1995). For example, if the phonemes /b/, /p/ and /m/ formed a phonemic equivalence class, then the words *bark*, *park* and *mark* were predicted to be perceptually equivalent (i.e., these words were predicted to form a *lexical equivalence class*). Finally, statistics were calculated after the application of these transcription rules to predict quantitatively how the phonemic distinctiveness under each of the stimulus conditions may affect lexical distinctiveness. In the present study, the analysis of lexical distinctiveness was conducted (1) for all words in the lexicon, (2) for the subset of monosyllabic words, and (3) for the subset of multisyllabic words.

This computational approach to examining lexical distinctiveness contrasts with modeling methods such as those developed by Boothroyd and Nittrouer (1988), in which lexical constraints are represented by the factor j (i.e., $P_W = P_P^j$, where the probability of identifying a word, P_W, is equal to the probability of identifying a phoneme, P_P, raised to a power, j, equal to the number of statistically independent phonemes within the word). In Boothroyd and Nittrouer's method, the j factor is a free parameter that is fit to experimental measures of phoneme and word identification accuracy, and the j factor must be re-fit whenever a new set of words are tested (i.e., different experimental lists of words can have different j factors). The computational method of the current study contrasts with parameter-fitting models in that the statistical constraints of the lexicon are determined directly from analyses of the lexical database (i.e., no free parameters are required), and make a priori estimates of word intelligibility.

[2] A caveat here is that phonemic equivalence classes simplify the fine-grained structure of the phoneme identification responses. That is, phonemes within a phonemic equivalence class become operationally defined to be identical, even though there may have been subtle differences in the identification responses for these phonemes.

2. Method

2.1. Subjects

All subjects were adult native speakers of English between 18 and 45 years of age. All reported having normal hearing, and normal or corrected-to-normal vision. Six subjects at Gallaudet University participated in the consonant identification task. Six subjects at House Ear Institute participated in the vowel identification task.

2.2. Stimuli and apparatus

2.2.1. Videodisc recordings

All stimuli were audiovisual recordings stored on laser videodiscs ((Bernstein and Eberhardt, 1986); for details on the recording procedure, see (Bernstein et al., 1989)). The stimuli were recorded by a male talker of American English. This individual was recorded seated before a dark background, and his face filled most of the screen area. He was illuminated by two floodlights directed at his face and by two fill lights positioned above his head.

For the consonant identification task, the stimuli were recordings of C-/ɑ/ syllables with the consonants /p/, /b/, /m/, /f/, /v/, /θ/, /ð/, /w/, /r/, /tʃ/, /dʒ/, /ʃ/, /ʒ/, /t/, /d/, /s/, /z/, /k/, /g/, /n/, /l/ and /h/, along with an isolated /ɑ/ syllable. For the vowel identification task, the stimuli were recordings of /h/-V-/d/ syllables with the vowels /i/, /ɪ/, /eɪ/, /aɪ/, /ɛ/, /ɚ/, /æ/, /ɑ/, /ɔ/, /oʊ/, /aʊ/, /ɔɪ/, /ʌ/, /ʊ/ and /u/. At least for lip reading, the intelligibility of the target phonemes in both these syllable contexts is unlikely to be impaired by coarticulation (e.g., Montgomery et al., 1987). For example, if the consonants had been presented in a C-/u/ context, anticipatory lip rounding would have been expected to impair lip reading performance. There were two recordings made of each syllable.

2.2.2. Stimulus conditions

In the V condition, the video track from the videodisc recordings was displayed with no audio. For the vowel identification task, video was displayed on an 18-in color monitor, placed approximately 1.75 m from the viewer's face. For the consonant identification task, video was displayed on a 14-in color monitor, placed approximately 0.50 m from the viewer's face.

In the F1A and F2A conditions, two acoustic vocoder algorithms were used with band-pass filters in the typical ranges of F1 and F2. The audio track from the videodisc recordings was processed in real-time and played to subjects binaurally via headphones. The basic hardware design and software algorithms for the vocoder are described in (Engebretson and O'Connell, 1986). However, the current vocoder algorithms were custom-designed in our laboratory (Bernstein et al., 1993). Each of the vocoder algorithms comprised a set of sixth-order band-pass filters. The F1A filters were spaced 75 Hz apart (from 75 to 900 Hz), and the F2A filters were spaced 150 Hz apart (from 825 to 2625 Hz). The energy passed by each band modulated the amplitude of a fixed-frequency sinusoid at the corresponding center frequency. These sinusoids were phase-locked. Although the vocoder processor can be considered to be real-time, the filters introduced small latencies to the output, ranging from 17 ms for the lowest frequency band of F1 to 2 ms for the highest frequency band of F2. Fig. 1 displays the center frequencies and gains for these algorithms.

The vocoder thus reproduced time-varying amplitude information within the range of each filter bank. However, the gains of the channels were the complement of the energy within those channels for typical speech stimuli, so the gain structures tended to flatten the output spectrums. Furthermore, the fine-grained structure of the original audio recording was lost; the output of each vocoder had a flat F0 (75 Hz for F1 and 150 Hz for F2), and the output was the same regardless of whether the input signal contained voiced, aspirated or fricative energy. The output was intelligible without training, but the timbre of the output voice had a synthetic, robot-like, quality.

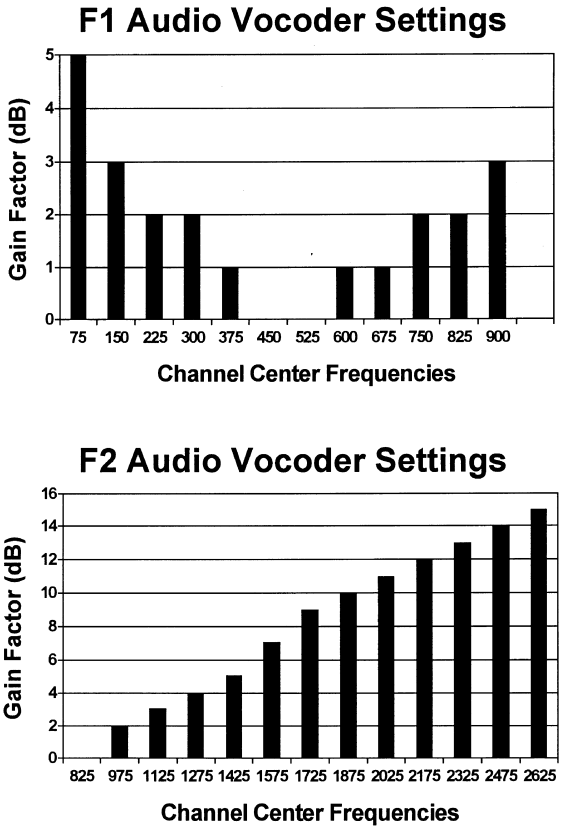

Fig. 1. Amount of gain within each frequency band for the audio vocoder using the F1 and F2 settings. The gain reflects the amount of amplification applied to a frequency band, not the absolute level of the output. For example, a gain of zero indicates that the amplitude of the input matched the amplitude of the output (i.e., no amplification).

In the F1AV and F2AV conditions, the vocoded audio and video recordings were both presented. Specifically, the videodisc player output the recorded audio and visual signals simultaneously, and the vocoder processed the audio signals in real-time.

2.3. Procedure

On each trial, participants gave a forced-choice phoneme identification of each stimulus. For consonant identification, responses were restricted to the 22 stimulus consonants and the isolated /ɑ/ vowel. For vowel identification, responses were restricted to the 15 stimulus vowels. Responses

were made by pressing marked keys on a computer keyboard, and examples of these phonemes in printed words were provided. After each response, subjects received feedback that gave them the correct answer for that trial. [3]

Each block of trials began with a practice phase consisting of one presentation of each stimulus in random order (i.e., 46 trials for consonants and 30 trials for vowels). This was followed by a test phase consisting of five repetitions of each stimulus in random order (i.e., 230 trials for consonants and 150 trials for vowels). For consonants, each subject completed two blocks for each of the five stimulus conditions. For vowels, each subject completed one block for each of these conditions. The order of conditions was randomized for each subject, with the constraint that subjects needed to be tested on an audio condition prior to being tested on the audiovisual condition that used the same vocoder algorithm. [4]

3. Results

3.1. Phoneme identification matrices

For each condition, the responses were compiled into stimulus-response confusion matrices that listed the percentage of trials for which each response label was chosen for each of the stimulus phonemes; this averaged the responses across different subjects and across the two different recordings of each phoneme. The confusion matrices are displayed in Appendix A. The levels of accuracy, listed in Table 1, varied across conditions from 29% for consonants in F1A and 51% for vowels in V, to 80% for consonants and 86% for vowels in F2AV. All scores were substantially above chance performance levels (4.3% for consonants and 6.7% for vowels).

3.2. Phonemic similarity matrices

In order to generate phonemic equivalence classes, it was necessary to transform confusion data into similarity estimates. For every pair of stimulus phonemes, similarity was estimated by calculating the phi-square statistic [5] (Norusis, 1993) on the distributions of responses given to these stimuli. This statistic is a normalized version of the chi-square test of equality for the two response distributions. It reaches a value of zero when the distributions of phoneme identification responses are identical for a pair of stimulus phonemes; the phonemes are then considered to be maximally similar. The statistic reaches a value of one when the distributions have no overlap (i.e., if subjects did not use any of the same response categories for these two phonemes); the

[3] The feedback was primarily intended to encourage participants to attempt to give responses that matched the stimulus (i.e., to discourage random guessing), even in stimulus conditions with low intelligibility. The feedback may also have had a secondary effect of providing training, and thus may have somewhat elevated the obtained performance levels.

[4] An inherent problem with using recordings of natural speech is that phonetically irrelevant visual details of the recordings (e.g., subtle differences in eye movements or head position) cannot be completely controlled. The randomization of V, F1AV and F2AV was chosen to limit the potential effect of learning these details on the experimental results; the potential accuracy gain from training on these stimuli is probably 3–5% points (see Bernstein et al., 1991). It is possible, given the design of the experiment, that the scores in audiovisual conditions were enhanced somewhat due to learning that had occurred in the audio-only conditions. Within each individual audio-only and audiovisual condition, we have examined the main effects of the first half of testing compared to the second half of testing, and have estimated that enhancements due to auditory learning were not likely to have been greater than three percentage points. This potential magnitude of error in the estimates of audiovisual phonetic perception is unlikely to have had a substantial effect on the pattern of results or even on the absolute predicted levels for the lexical statistics in this study.

[5] In mathematical notation, the phi-square statistic is

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

with x_i and y_i equaling the frequencies that phonemes x and y were identified as response category i, $E(x_i)$ and $E(y_i)$ equaling the expected frequencies of responses for x_i and y_i if phonemes x and y are equivalent (i.e., the expected frequencies in a Pearson chi-square test), and N equaling the total number of responses to phonemes x and y.

Table 1
Phoneme identification results

Condition	Phonemes percent correct		Number of phonemic equivalence classes		Phonemic equivalence classes
	Consonants	Vowels	Consonants	Vowels	
V	47.8%	51.3%	9	4	{p, b, m} {f, v} {ð, θ} {w} {r} {tʃ, dʒ, ʃ, ʒ, d}{t, s, z}{k, g, h, ŋ, j}{n, l} {i, ɪ, eɪ, aɪ, ɛ, a, ʌ, ə}{ɚ, o, ɔɪ, ʊ, u} {ɑ, ɔ}{aʊ}
F1A	28.9%	67.9%	4	9	{p, t, k, l, h} {b, f, v, ð, θ, w, tʃ, dʒ, ʃ, ʒ, d, s, z, g, j} {m, n, ŋ} {r} {i} {ɪ} {eɪ} {aɪ} {ɛ, ɚ, ɔɪ, ʌ, ə} {a, ɑ, ɔ, aʊ} {o} {ʊ} {u}
F2A	56.5%	81.3%	12	10	{p, b, f, v, ð, θ, s, z, h} {m} {w} {r, l} {tʃ, dʒ, ʃ, ʒ} {t, d} {k, g} {n} {ŋ} {j} {i} {ɪ} {eɪ} {aɪ} {ɛ, a} {ɛ, a} {ɚ} {ɑ, ɔ, ʌ} {o, u}{ɔɪ} {ʊ} {ə }
F1AV	66.4%	80.2%	14	12	{p} {b} {m} {f, v} {ð, θ} {w} {r} {tʃ, dʒ, ʃ, ʒ, d, g} {t, s, z} {k, h} {n} {l} {j} {ŋ} {i} {ɪ} {eɪ} {aɪ} {ɛ, ʌ, ə} {ɚ, ɑ, ɔ} {a} {o} {aʊ} {ɔɪ} {ʊ} {u}
F2AV	80.1%	86.3%	18	12	{p} {b} {m} {f, v} {ð, θ} {w} {r} {tʃ, dʒ} {ʃ, ʒ} {t, d} {s, z} {k} {g} {n} {l} {h} {j} {ŋ} {i} {ɪ} {eɪ} {aɪ} {ɛ, a} {ɚ} {ɑ, ɔ, ʌ} {o, u} {aʊ} {ɔɪ} {ʊ} {ə}

phonemes then are considered to be maximally dissimilar.

It is more common to evaluate similarity within phoneme identification experiments by examining only the response categories associated with the individual phoneme pairs. For example, the similarity between any two stimulus phonemes can be estimated by averaging the percentage of trials in which the first phoneme was identified as the second phoneme and the percentage of trials in which the second phoneme was identified as the first phoneme. However this similarity estimate has disadvantages. First, the response percentages for a particular pair interact with the similarity of other phonemes in the experiment. For example, if two phonemes are identical under a stimulus condition but are distinguishable from all other phonemes, then these phonemes are likely to be confused on 50% of the trials. If four other phonemes are identical under that stimulus condition but are distinguishable from all other phonemes, then any pair of these phonemes are likely to be confused on 25% of the trials. Response percentages would therefore imply that phonemes in the group of two have twice the similarity to each other as phonemes in the group of four, despite the fact that the magnitude of similarity is actually equivalent.

The phi-square coefficient is used here to correct this problem, because it compares the response distributions across all available response categories. In the example above, phonemes in the group of two should have a phi-square coefficient near zero, because the responses for those phonemes are divided among the same two categories. Pairs of phonemes in the group of four should also have a phi-square coefficient near zero because the responses for those phonemes are divided among the same four categories. The magnitude of the phi-square coefficient for an individual phoneme pair is thus independent of the similarity of other phonemes in the experiment.

The phi-square coefficient also has advantages when there are response biases and asymmetries. For example, given three phonemes, /p/, /b/ and /m/, that are indistinguishable in a hypothetical stimulus condition, a problem can arise if subjects are biased to choose /p/ for their identification response for all three phonemes. If only individual pairs are considered, the similarity between /b/ and /m/ would be erroneously low because /b/ would be rarely identified as /m/, and /m/ would be rarely identified as /b/; the bias to identify tokens as /p/ would decrease the use of these two other categories. The phi-square coefficient solves this, because it measures the similarity between response distributions for phoneme pairs, but is not dependent on what specific response categories are used.

3.3. Phonemic equivalence classes

Following the phi-square transformations, phonemic equivalence classes were found following the procedures used by Walden et al. (1977) to define visemes. First, the phonemic similarity matrices (i.e., the matrices of phi-square statistics) were analyzed using hierarchical cluster analysis (Norusis, 1993). This procedure generates an inverted tree structure in which phonemes join classes based on similarity. At the lowest level of the structure, no phonemes are joined together (i.e., each phoneme belongs to its own equivalence class). At each succeeding level, the most similar pair of classes is joined together. This continues until, at the highest level, all phonemes join a single equivalence class. An average-linkage-within-groups method was used to determine which classes to join at each level in the hierarchy; two classes were joined if they had the minimum average within-class distance at that level. As the level of the hierarchy goes from low to high, the average within-class distances become larger (i.e., phonemes within a class become more distantly related) and the number of classes become fewer.

In order to perform the computational analysis of the lexicon, a single level of the tree structure was chosen to represent the phonemic equivalence classes for each stimulus condition. In accord with Walden et al.'s definition of a viseme (Walden et al., 1977) the phonemic equivalence classes were chosen by finding the first level in which at least 75% of all the responses were within-class. [6] For example, if the phonemes /b/ and /m/ were in the same class, then a /b/ response to a /mɑ/ stimulus would be considered to be a within-class response. Under stimulus conditions with excellent intelligibility, this criterion should be met at a relatively low level of the hierarchy, indicating that there are many distinct classes of phonemes (i.e., few equivalent phonemes). Under stimulus conditions with poor intelligibility, this criterion should be met at a relatively high level of the hierarchy, indicating that there are few distinct classes of phonemes (i.e., many equivalent phonemes).

The phonemic equivalence classes are listed in Table 1. [7] As predicted, the number of classes was higher under more intelligible conditions (F1AV and F2AV), indicating better phonemic distinctiveness. In addition, the membership of individual classes varied substantially between conditions. The classes for V were in accord with traditional descriptions of visemes (e.g., Kricos and Lesner, 1982; Owens and Blazek, 1985). The consonant classes were primarily differentiated by place and secondarily by manner (e.g., {t, s, z} versus. {n, l}). The vowels were relatively indistinct, with the classes being differentiated by degree of lip rounding and jaw height. For F1A, the consonants were poorly intelligible, and they roughly fell into classes differentiated by whether an initial F1 transition was present or absent (i.e., {p, k, θ, l}) and by whether there was nasaliza-

[6] Under the present interpretation of this criterion, every phoneme is assigned to a phonemic equivalence class; no phonemes are left unassigned.

[7] For the purposes of examining information in the lexicon, the phonemes /y/, /ŋ/ and /ə/ were included into the phonemic equivalence classes, despite the fact that perceptual data were not collected for these phonemes. These phonemes were placed into phonemic equivalence classes that seemed most plausible given the classification of other phonemes based on the perceptual data. In addition, the phonemic equivalence class for the /ɑ/ stimulus was based on the vowel identification data, not on the identification of this phoneme within the consonant identification task.

tion or a low F3. The vowels roughly fell into clusters of low, mid and a number of high vowels that were uniquely specified. For F2A, the consonants were mostly specified by place, but the classifications were different from those based on place in V. In particular, place distinctions near the front of the mouth (e.g., /v/ versus /θ/) were less distinctive in the F2A condition. Vowels were mostly uniquely specified, although there were difficulties in distinguishing some mid and low vowels (i.e., {ɛ, æ} and {ɑ, ɔ, ʌ}). For the AV conditions, the higher levels of intelligibility led to more phonemes that were uniquely specified. For consonants, both conditions seemed to make manner distinctions among fricatives difficult, although phonemic distinctiveness was somewhat higher for F2AV. For vowels, both conditions had a high degree of phonemic distinctiveness, although there were small differences between the two conditions. The five conditions thus produced substantial variation in the membership of phonemic equivalence classes, affording a good data set for investigating how the presence of different phonetic information affects lexical distinctiveness.

3.4. Lexical modeling

The phonemic equivalence classes (see Table 1) were used to retranscribe the PhLex lexical database (Seitz et al., 1995), following the procedures of Auer and Bernstein (1997). For example, when the word *hop* was retranscribed using the phonemic equivalence classes of the V condition, the /h/ was replaced with a new symbol standing for the entire equivalence class {h, g, k}, the /ɑ/ was replaced with a new symbol standing for the entire equivalence class {ɑ, ɔ}, and the /p/ was replaced with a new symbol standing for the entire equivalence class {b, p, m}. Therefore, the words *hop*, *cop* and *cob* were retranscribed with the same symbols under the transcription rules for the V condition.

Lexical statistics, listed in Table 2, were calculated to quantify the effects of the loss of phonemic distinctiveness on lexical uniqueness. In separate analyses, statistics were calculated on the full corpus of 31,075 words in PhLex, on the 5,073 monosyllabic words in PhLex, and on the 26,002 multisyllabic words in PhLex. Monosyllabic words are particularly interesting to examine because

Table 2
Lexical statistics

	Condition	Percent information extracted	Percent words unique [a]	Expected class size [a]	Expected percent correct [a]	Expected enhancement [a]
All words	V	92.4%	67.6% (59.2%)	2.9 (3.7)	76.8% (70.1%)	– –
	F1A	91.4%	60.8% (51.5%)	4.7 (5.9)	70.5% (62.9)%	– –
	F2A	95.4%	77.4% (70.5%)	1.8 (2.2)	84.9% (79.9%)	– –
	F1AV	98.3%	87.5% (82.6%)	1.2 (1.3)	92.7% (89.7%)	15.9% (19.7%)
	F2AV	99.1%	91.9% (88.3%)	1.1 (1.2)	95.6% (93.6%)	18.9% (23.6%)
Monosyllabic words	V	86.3%	15.1% (12.5%)	9.2 (9.8)	31.4% (28.7%)	– –
	F1A	85.2%	7.8% (7.1%)	16.8 (16.7)	21.0% (20.4%)	– –
	F2A	91.6%	27.9% (26.0%)	4.9 (5.0)	46.9% (45.3%)	– –
	F1AV	97.0%	53.3% (50.4%)	2.1 (2.1)	71.1% (69.4%)	39.7% (40.7%)
	F2AV	98.2%	66.7% (63.3%)	1.5 (1.5)	81.5% (79.7%)	50.1% (51.0%)
Multisyllabic words	V	97.9%	77.8% (73.8%)	1.7 (1.8)	85.6% (83.0%)	– –
	F1A	96.7%	71.2% (65.4%)	2.3 (2.5)	80.1% (76.2%)	– –
	F2A	99.0%	87.0% (84.4%)	1.3 (1.3)	92.3% (90.7%)	– –
	F1AV	99.6%	94.2% (92.7%)	1.1 (1.1)	96.9% (96.1%)	11.2% (13.1%)
	F2AV	99.8%	96.9% (96.1%)	1.0 (1.0)	98.4% (98.0%)	12.8% (15.0%)

[a] Frequency-weighted statistics are in parentheses.

they are more frequent, in spoken English as well as in perceptual experiments, than are multisyllabic words. Furthermore, the lexicon probably provides weaker constraints for the recognition of monosyllabic words than for multisyllabic words, considering that monosyllabic words have fewer phonemes and therefore have less potential to be phonemically distinct from other monosyllabic words in the lexicon.

The percent information extracted statistic (PIE; Carter, 1987) was calculated to estimate the amount of information reduction in the lexicon resulting from the phonemic ambiguity within each stimulus condition. It was calculated according to the following formula:

$$\mathrm{PIE} = \frac{\sum_{a=1}^{n_E} p_a \log_2 p_a}{\sum_{i=1}^{n_L} p_i \log_2 p_i} \times 100,$$

where n_L is the number of words in the lexicon, n_E is the number of lexical equivalence classes after retranscription, p_i is the occurrence probability for word i, [8] and p_a is the occurrence probability for equivalence class a. [9] PIE is the quantity of information in the retranscribed lexicon (i.e., the number of bits required to differentiate the lexical equivalence classes) expressed as a percentage of the quantity of information in the original lexicon (i.e., the number of bits required to differentiate all the words).

The PIE statistics, listed in Table 2, demonstrate that a high percentage of information remains after every retranscription. In the condition with the lowest PIE scores (i.e., F1A), 91.4% of the information remains when all words in the lexicon are considered, 85.2% remains when only monosyllabic words are considered, and 96.7% remains when only multisyllabic words are considered. This occurred despite the fact that the 40 phonemes in the lexicon were reduced to 13 phonemic equivalence classes. PIE scores were higher for conditions with a larger number of phonemic equivalence classes. From the point of view of PIE, poor phonemic intelligibility does little to reduce the lexical structure.

However, note that the PIE statistic measures the information content of the lexicon, but does not predict word intelligibility. For example, consider a hypothetical lexicon composed of 1024 words that each have the same occurrence probability. This lexicon would have 10 bits of information (i.e., $2^{10} = 1024$). If retranscription formed 512 lexical equivalence classes, each composed of two words, then it would have 9 bits (i.e., $2^9 = 512$). PIE would thus be 90%, despite the fact that none of the words in the retranscribed lexicon would be unique (all words could be confused with one other word).

The percentages of words that are unique after retranscription (i.e., belong in their own lexical equivalence class) are listed in Table 2, along with a version of this statistic weighted by the frequency with which each word occurs in the language. [10] Even in the poorest condition (i.e., F1A), a majority of words (60.8% unweighted, 51.5% weighted) in the full lexicon are predicted to be unique. However, uniqueness is predicted to be much poorer for monosyllabic than for multisyllabic words; in F1A, only 7.8% (7.1% weighted) of monosyllabic words are unique, but 71.2% (65.4% weighted) of multisyllabic words are unique. Monosyllabic words thus seem to require more phonetic information to achieve uniqueness than do multisyllabic words. In addition, the weighted versions of these statistics are lower, suggesting that there is a tendency for uniqueness to decline for higher-frequency words.

[8] This was calculated by dividing the frequency with which the word i occurs in the Brown corpus (Kucera and Francis, 1967) by the sum of all word frequencies. It is used as an estimate of the a priori probability that an individual word encountered when speaking English will be the word i. Note that these frequencies are based on written, not spoken, English.

[9] This was calculated by dividing the sum of the word frequencies within equivalence class a by the sum of all word frequencies in the lexicon. It is used as an estimate of the a priori probability that an individual word encountered when speaking English will be a member of equivalence class a.

[10] The frequency weighted average was calculated by dividing the sum of the frequencies of unique words by the sum of the frequencies of all words. The word frequency values for this and the other weighted statistics reported in this paper were found by calculating the log of the occurrence frequencies of words in the Brown corpus (Kucera and Francis, 1967).

If subjects were asked to identify words in an experiment, words that are unique after retranscription would be predicted to be identified correctly on nearly 100% of the trials, because no words in the lexicon should be confused with them. But accuracy would not be expected to fall to 0% for words that have larger lexical equivalence classes. Instead, identification percentages would be predicted to reflect the probability that the correct word will be chosen from among the other words in its equivalence class. The choice of an individual word from a lexical equivalence class is likely to be influenced by factors such as word frequency and familiarity (i.e., people are more likely to choose frequent and familiar words within the class). In the current study, we have adopted a working assumption that the probability of identifying the correct word is the inverse of the lexical equivalence class size. For example, if a word was perceptually identical to four other words in the lexicon (i.e., lexical equivalence class size of four), then subjects should identify the correct word on 25% of the trials. Therefore, the percentage of unique words is not a good estimate for the average level of identification accuracy for all of the words; such an estimate also requires examining the lexical equivalence class sizes for non-unique words.

The expected lexical equivalence class size was calculated by averaging the lexical equivalence class sizes across all words. The weighted expected lexical equivalence class size was calculated by multiplying the class size for each word by the word's frequency, and then calculating the sum of this statistic for all words, divided by the total word frequency. These statistics (see Table 2) reveal that expected lexical equivalence classes are indeed larger for monosyllabic words; these words tend to be located in dense regions of the lexicon that are less capable of resolving perceptual ambiguities. In addition, expected lexical equivalence class sizes for each set of words become somewhat larger when words are weighted by frequency.

Table 3
Consonant identification percentages: V condition

	Response																						
Stimulus	p	b	m	f	v	ð	θ	w	r	tʃ	dʒ	ʃ	ʒ	t	d	s	z	k	g	n	l	h	ɑ
p	66	28	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	30	44	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	28	40	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	73	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
v	0	0	0	49	49	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
ð	0	0	0	0	0	66	31	0	0	0	0	0	3	0	0	0	0	0	0	0	1	0	
θ	0	0	0	0	0	48	52	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
w	1	0	0	0	0	0	0	94	5	0	0	0	0	0	0	0	0	0	0	0	0	0	
r	0	0	0	0	0	0	0	6	94	0	0	0	0	0	0	0	0	0	0	0	0	0	
tʃ	0	0	0	0	0	0	0	0	0	30	40	13	17	0	0	0	0	1	0	0	0	0	
dʒ	0	0	0	0	0	0	0	0	0	36	33	18	11	0	0	1	3	0	0	0	0	0	
ʃ	0	0	0	0	0	0	1	1	0	42	18	23	13	0	0	1	1	0	0	0	0	0	
ʒ	0	0	0	0	0	0	0	0	0	46	25	13	13	0	0	0	2	0	0	1	0	0	
t	0	0	0	0	0	0	0	0	0	0	0	0	0	27	23	22	16	1	12	1	0	0	
d	0	0	0	0	0	0	2	0	0	0	0	0	0	22	20	2	2	6	13	23	8	1	3
s	0	0	0	0	0	0	0	0	0	0	0	0	3	9	5	44	37	0	1	0	0	0	
z	0	0	0	0	0	0	0	0	0	0	1	0	1	13	8	40	38	0	0	0	0	0	
k	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	49	20	7	3	8	12
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	24	23	8	1	28	17
n	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	3	4	53	35	2	0
l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	12	87	0	0
h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	4	12	14	4	28	36
ɑ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	2	0	0	32	62

Table 4
Consonant identification percentages: F1A condition

	Response																					
Stimulus	p	b	m	f	v	ð	θ	w	r	tʃ	dʒ	ʃ	ʒ	t	d	s	z	k	g	n	l	h	ɑ
p	14	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	3	13	67
b	1	87	2	1	8	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
m	0	0	84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0
f	3	48	0	7	10	1	1	2	0	0	1	0	0	0	7	0	3	1	1	0	4	3	10
v	2	4	1	4	19	3	3	15	10	0	0	0	0	0	8	1	3	2	9	0	6	4	7
ð	4	26	0	3	14	0	3	12	0	2	1	0	1	1	10	2	1	2	3	2	4	3	8
θ	0	7	4	0	13	2	2	45	13	1	2	0	1	2	3	2	1	1	1	0	3	0	0
w	1	0	1	0	0	2	1	38	31	1	0	1	3	0	0	0	3	0	0	0	21	0	0
r	0	1	0	0	0	0	0	1	93	0	1	0	0	0	0	0	0	0	4	0	1	0	0
tʃ	4	8	0	3	14	3	4	18	3	1	0	1	1	3	10	3	4	4	9	1	3	3	3
dʒ	0	1	2	2	4	0	2	45	0	0	1	0	1	0	13	4	7	1	9	3	3	3	0
ʃ	2	11	1	3	4	0	0	34	1	0	1	1	1	1	15	7	9	2	3	0	5	0	1
ʒ	0	0	0	0	5	0	1	68	2	0	3	1	1	1	0	1	10	0	6	0	3	0	0
t	12	1	0	0	0	0	0	2	1	3	1	0	1	3	1	1	1	8	2	0	3	13	49
d	3	13	0	3	9	4	2	27	2	3	1	1	1	0	15	3	3	1	10	0	1	0	1
s	2	14	2	5	10	2	2	14	0	0	1	2	2	0	8	4	5	2	13	0	5	4	5
z	0	4	2	0	5	3	2	45	5	0	0	0	3	1	5	3	13	2	5	2	3	0	0
k	1	0	0	0	0	0	1	2	1	1	1	1	0	2	0	0	2	13	2	2	6	17	52
g	0	0	1	1	3	3	0	25	1	1	1	1	3	0	9	1	2	4	38	1	3	3	3
n	0	0	48	0	2	0	0	8	2	0	2	0	2	0	0	0	0	0	0	31	7	0	0
l	0	0	6	0	0	1	1	3	3	1	0	0	2	1	0	0	0	0	2	0	83	0	0
h	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	33	63
ɑ	0	0	0	1	0	0	0	0	1	0	0	0	0	1	2	0	1	0	0	0	3	3	89

The expected percent correct was calculated by averaging the inverse of the lexical equivalence class sizes across all words. The weighted expected percent correct was calculated by multiplying the inverse of the expected class size for each word by its frequency, and then calculating the sum of this statistic for all words divided by the total word frequency. The expected percent correct for words in the full lexicon is quite high for all conditions (see Table 2); words should be identified correctly 70.5% (62.9% weighted) of the time even in the F1A condition (i.e., the condition with the poorest intelligibility). The expected percent correct is much lower for monosyllabic words (e.g., 21.0% unweighted and 20.4% weighted for F1A) than for multisyllabic words (e.g., 80.1% unweighted and 76.2% weighted for F1A). The expected percent correct statistics are somewhat lower when weighted by frequency, but this result is insufficient for making strong predictions about whether high-frequency words would be identified less accurately than low-frequency words under conditions with reduced phonemic intelligibility. [11]

For the audiovisual conditions, the expected enhancement in intelligibility due to adding auditory information to lip reading was calculated by subtracting the expected percent correct values for V from that of F1AV and F2AV (see Table 2). Considering all words in the lexicon, the expected enhancement is moderate (15.9–23.6%) for both conditions. However, the magnitude of expected enhancement is much larger for monosyllabic (39.7–51.0%) than for multisyllabic (11.2–15.0%)

[11] This computational result occurred because high-frequency words are more likely to have larger lexical equivalence classes, but this effect may be reversed in practice by the fact that high-frequency words within a lexical neighborhood are more likely to be given as a response (Luce, 1986; Luce et al., 1990). The bias to give a high-frequency word response is not incorporated into the weighted expected percent correct statistics of the current experiment.

Table 5
Consonant; identification percentages: F2A condition

	Response																					
Stimulus	p	b	m	f	v	ð	θ	w	r	tʃ	dʒ	ʃ	ʒ	t	d	s	z	k	g	n	l	h	ɑ
p	53	18	0	0	0	1	1	8	0	0	0	0	0	7	13	0	0	1	0	0	0	0	0
b	1	82	8	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	1
m	0	0	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2
f	16	6	3	38	18	6	2	0	0	0	0	0	0	0	2	0	0	0	0	2	0	6	3
v	40	8	0	9	19	3	3	1	0	0	0	0	2	0	1	0	0	1	1	0	10	2	1
ð	2	8	1	12	10	13	13	0	0	2	0	0	0	1	3	18	8	1	3	3	1	3	2
θ	11	0	0	3	0	8	10	0	0	0	0	0	1	6	3	6	8	12	8	16	2	5	3
w	0	0	0	0	0	0	0	93	3	0	0	0	0	0	0	0	0	0	0	0	3	0	0
r	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	1	0	0
tʃ	0	0	0	0	0	0	0	0	0	78	20	1	1	0	0	0	0	0	0	0	0	0	0
dʒ	0	0	0	0	0	0	1	0	0	45	49	0	3	0	0	0	0	0	3	0	0	0	0
ʃ	0	0	0	0	0	0	1	0	0	4	1	68	25	0	0	0	2	0	0	0	0	0	0
ʒ	0	0	0	0	0	0	1	0	0	12	16	28	43	0	0	0	1	0	0	0	0	0	0
t	0	0	0	0	0	1	3	0	0	0	0	0	0	48	37	0	0	9	3	0	0	0	0
d	0	0	0	0	0	0	1	0	0	0	0	0	0	1	93	0	0	0	4	0	1	0	0
s	0	0	0	12	4	8	8	0	0	0	0	2	3	0	0	41	8	1	0	13	0	0	0
z	1	0	0	8	6	13	5	0	0	5	2	0	4	2	0	10	4	7	1	12	1	14	6
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	77	23	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	83	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
l	0	0	0	0	0	0	0	0	61	0	0	1	0	0	0	0	0	0	0	0	38	0	0
h	7	0	2	8	26	0	1	11	1	0	0	0	0	2	0	0	0	5	0	3	3	29	4
ɑ	2	3	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	7	6	43

words. In fact, for monosyllabic words, the magnitude of expected enhancement is large enough to predict that audiovisual integration for these words should be superadditive (i.e., the level of audiovisual intelligibility should exceed the sum of the levels of audio and visual intelligibility).

4. General discussion

The main result of this study is that the effects of audiovisual phonetic information on lexical distinctiveness are dependent on the specific words that are examined. The computational models predict that monosyllabic words are less frequently unique, have larger lexical equivalence classes, and have lower levels of intelligibility than do multisyllabic words under the same perceptual conditions. Furthermore, this prediction is the same across the range of phonemic intelligibility levels and phonetic structures examined in the present experiment. Monosyllabic words are predicted to show superadditive levels of enhancement in audiovisual conditions, but the enhancement is predicted to be much lower for multisyllabic words. The results of this study suggest that models of phonetic processing are inadequate in themselves to account for audiovisual recognition of words under conditions in which the phonetic information is degraded or impoverished.

Superadditive levels of enhancement probably arise because small gains in phonetic information cause larger increases in word intelligibility as the number of ambiguous words in the lexicon becomes smaller. For example, if an increase in phonetic information causes the lexical equivalence class size to decrease from 2 to 1, word identification accuracy would be expected to rise by 50 percentage points (from 50% to 100%). However, if an increase in phonetic information causes the lexical equivalence class size to decrease from 10 to 5, word identification accuracy would be expected to increase by only 10 percentage points (from 10% to 20%). Therefore, superadditivity is likely to

Table 6
Consonant identification percentages: F1AV condition

	Response																					
Stimulus	p	b	m	f	v	ð	θ	w	r	tʃ	dʒ	ʃ	ʒ	t	d	s	z	k	g	n	l	h	ɑ
p	97	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1
b	1	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f	1	3	0	44	52	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
v	0	0	0	27	73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ð	0	0	0	0	1	73	24	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
θ	0	0	0	0	0	57	42	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
w	0	0	3	0	0	0	0	95	3	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	0	1	98	0	0	0	0	0	0	0	0	0	1	0	0	0
tʃ	0	1	0	0	1	0	0	0	0	31	13	8	26	3	1	3	13	0	0	0	2	0
dʒ	0	0	0	0	0	0	0	0	1	10	42	5	28	0	2	0	9	0	1	0	3	0
ʃ	0	0	0	0	0	0	0	0	0	18	16	17	32	0	0	3	16	0	0	0	0	0
ʒ	0	0	0	0	0	0	0	0	0	8	18	0	50	0	1	2	21	0	2	0	0	0
t	0	0	0	0	0	0	0	0	0	0	0	0	1	51	4	9	8	13	1	0	3	11
d	0	0	0	0	0	5	3	0	0	0	0	0	2	1	36	5	15	6	28	0	1	0
s	0	0	0	0	0	0	0	0	0	0	0	0	3	5	13	40	38	1	0	0	0	0
z	0	0	0	1	0	0	0	0	0	0	0	0	3	2	7	13	75	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	45	7	2	2	25	18
g	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	2	3	92	0	0	0	0
n	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	91	8	0	0
l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	99	0	0
h	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	48	48
ɑ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	8	91

occur whenever the gain in phonetic information due to audiovisual integration causes the lexical equivalence class size to approach 1.

Although the perceptual integration models of Massaro (1987) and Braida (1991) can account for superadditivity effects at the phonetic level, information about lexical structure is required to account for word superadditivity. In the present experiment, superadditivity effects were predicted for monosyllabic words but not for multisyllabic words, because monosyllabic words tend to reside in denser regions of the lexicon than do multisyllabic words. The structure of the lexicon provides more information for the recognition of words located in sparse regions (i.e., lexical structure can resolve phonemic ambiguities for most multisyllabic words), so the limited phonetic information available in the unimodal conditions was sufficient for many of these words to remain unique. Additional phonetic information under audiovisual conditions provided little improvement in expected performance because there was little remaining lexical ambiguity. Words that reside in dense regions of the lexicon (i.e., most monosyllabic words) require a greater degree of phonetic information to resolve word ambiguity, so there is more potential benefit of audiovisual integration. Superadditivity effects therefore depend both on the level of phonetic information that is required to disambiguate words in specific regions of the lexicon, and on the structures of phonetic information that are available in unimodal and bimodal conditions. Large gains in audiovisual performance should not occur either when unimodal phonetic information is sufficient to eliminate lexical ambiguity or when bimodal phonetic information is insufficient to eliminate lexical ambiguity (i.e., lexical equivalence classes remain large). Large gains in audiovisual performance are predicted when the phonetic information required by the lexicon and the phonetic information available through perception coincide, so that lexical ambiguity remains high in unimodal conditions but is eliminated in bimodal conditions.

Table 7
Consonant identification percentages: F2AV condition

	Response																						
Stimulus	p	b	m	f	v	ð	θ	w	r	tʃ	dʒ	ʃ	ʒ	t	d	s	z	k	g	n	l	h	ɑ
p	90	9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	2	96	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	72	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
v	2	0	0	31	68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ð	0	0	0	0	0	69	27	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0
θ	0	0	0	1	0	64	31	0	0	0	0	0	1	1	0	1	0	0	1	0	0	0	0
w	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
tʃ	0	0	0	0	0	0	1	0	0	79	20	0	0	0	0	0	0	0	0	0	0	0	0
dʒ	0	0	0	0	0	0	0	0	0	32	68	0	0	0	0	0	0	0	0	0	0	0	0
ʃ	0	0	0	0	0	1	0	0	0	2	5	60	29	0	0	2	1	0	0	0	0	0	0
ʒ	0	0	0	0	0	0	0	0	0	11	12	20	57	0	0	0	0	0	0	0	0	0	0
t	0	0	0	0	0	0	0	0	0	0	0	0	0	64	29	0	0	6	1	0	0	0	0
d	0	0	0	0	0	0	0	0	0	0	0	0	0	3	97	0	0	0	0	0	0	0	0
s	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	88	10	0	0	1	0	0	0
z	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	46	46	0	0	3	0	2	1
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	89	11	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	86	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
l	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	99	0	0
h	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	3	3	87	4
ɑ	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	98

Table 8
Vowel identification percentages: V condition

	Response													
Stimulus	i	ɪ	eɪ	aɪ	ɛ	ɚ	æ	ɑ	ɔ	o	aʊ	ɔɪ	ʌ	ʊ
i	42	48	2	0	8	0	0	0	0	0	0	0	0	0
ɪ	18	52	7	2	20	0	0	0	0	0	0	0	2	0
eɪ	0	3	25	10	20	0	38	0	2	0	0	0	2	0
aɪ	0	0	8	25	17	0	22	18	3	0	0	0	7	0
ɛ	0	2	27	8	35	0	20	7	0	0	0	0	2	0
ɚ	0	0	0	0	0	47	0	0	8	8	2	7	2	18
æ	0	0	28	7	3	0	60	2	0	0	0	0	0	0
ɑ	0	0	0	2	0	7	0	58	25	0	2	2	5	0
ɔ	0	0	0	0	0	8	0	28	60	0	0	0	3	0
o	0	0	0	0	0	3	0	0	2	70	8	3	0	2
aʊ	0	0	0	0	0	0	0	2	7	0	92	0	0	0
ɔɪ	0	0	0	0	0	3	0	0	8	10	0	57	0	15
ʌ	0	0	2	3	28	0	5	25	5	0	0	0	32	0
ʊ	0	0	0	0	0	3	0	0	2	2	0	0	2	58
u	0	0	0	0	0	8	0	0	0	3	0	0	0	30

A caveat here is that the results of this experiment are based on an English lexical database. Lexicons of other languages are known to differ structurally from the English lexicon (Carlson et al., 1985). However, it is unknown whether the lexicons of other languages have distributional characteristics analogous to those in English. Specifically, in English, monosyllabic words tend

Table 9
Vowel identification percentages: F1A condition

	Response														
Stimulus	i	ɪ	eɪ	aɪ	ɛ	ɚ	æ	ɑ	ɔ	o	aʊ	ɔɪ	ʌ	ʊ	u
i	98	0	0	0	2	0	0	0	0	0	0	0	0	0	0
ɪ	3	93	0	2	0	0	0	0	0	0	0	0	0	0	2
eɪ	2	0	92	0	2	0	2	0	2	0	0	0	0	2	0
aɪ	0	2	0	97	0	0	2	0	0	0	0	0	0	0	0
ɛ	3	12	0	0	57	15	7	0	3	0	0	0	3	0	0
ɚ	0	0	2	0	32	23	5	5	2	2	2	17	8	2	2
æ	0	0	0	0	2	0	73	23	2	0	0	0	0	0	0
ɑ	0	0	2	0	0	0	20	47	27	0	3	0	2	0	0
ɔ	0	0	2	0	0	0	18	50	28	0	0	0	2	0	0
o	0	0	0	0	0	0	0	0	2	85	13	0	0	0	0
aʊ	0	0	0	5	0	0	2	3	15	0	75	0	0	0	0
ɔɪ	0	0	2	3	3	0	3	0	2	0	3	82	2	0	0
ʌ	0	0	0	0	43	5	18	12	5	0	0	0	15	0	2
ʊ	2	8	0	0	3	0	0	0	0	0	0	0	0	78	8
u	10	0	0	0	0	0	0	0	0	2	3	0	0	10	75

Table 10
Vowel identification percentages: F2A condition

	Response														
Stimulus	i	ɪ	eɪ	aɪ	ɛ	ɚ	æ	ɑ	ɔ	o	aʊ	ɔɪ	ʌ	ʊ	u
i	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ɪ	0	98	0	0	2	0	0	0	0	0	0	0	0	0	0
eɪ	2	0	98	0	0	0	0	0	0	0	0	0	0	0	0
aɪ	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
ɛ	0	18	0	2	75	0	5	0	0	0	0	0	0	0	0
ɚ	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
æ	0	0	0	0	43	0	57	0	0	0	0	0	0	0	0
ɑ	0	0	0	0	0	0	0	60	37	0	0	0	3	0	0
ɔ	0	0	0	0	0	0	0	18	72	2	2	0	5	0	2
o	0	0	0	0	0	0	0	0	0	78	3	0	0	7	12
aʊ	0	0	0	0	0	0	0	0	2	0	97	2	0	0	0
ɔɪ	0	0	0	0	0	0	0	0	0	0	2	98	0	0	0
ʌ	0	0	0	0	0	0	0	32	5	0	0	0	63	0	0
ʊ	0	0	0	0	0	0	0	0	0	0	0	0	8	77	15
u	0	0	0	0	0	0	0	0	0	35	8	0	0	10	47

to reside in denser regions of the lexicon than do multisyllabic words, and high-frequency words tend to reside in denser regions of the lexicon than do low-frequency words ((Landauer and Streeter, 1973); but see (Pisoni et al., 1985)). These characteristics may not be true for the lexicons of all languages, or for the lexicons of limited-vocabulary automatic speech recognition systems. The results probably do extend to other lexicons with regard to lexical density; less phonetic information is required to recognize words in sparse regions of the lexicon, and increases in phonetic information can have different implications for words located in dense and sparse regions.

Preliminary experimental results from our laboratory (Bernstein et al., 1997; Iverson et al., 1997) suggest that the current computational results are predictive of human word recognition performance. We have collected, under the five stimulus conditions used for the current models, open-set

Table 11
Vowel identification percentages: F1AV condition

	Response														
Stimulus	i	ɪ	eɪ	aɪ	ɛ	ɚ	æ	ɑ	ɔ	o	aʊ	ɔɪ	ʌ	ʊ	u
i	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ɪ	0	97	0	3	0	0	0	0	0	0	0	0	0	0	0
eɪ	0	0	98	0	0	0	2	0	0	0	0	0	0	0	0
aɪ	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
ɛ	0	2	0	0	88	0	8	2	0	0	0	0	0	0	0
ɚ	0	0	0	0	3	58	5	7	5	3	2	3	7	7	0
æ	0	0	2	0	7	2	82	8	0	0	0	0	0	0	0
ɑ	0	0	0	0	0	0	10	63	27	0	0	0	0	0	0
ɔ	0	0	2	0	0	0	0	45	52	2	0	0	0	0	0
o	0	0	0	0	0	0	0	0	0	82	8	0	0	5	5
aʊ	0	0	0	0	0	0	0	0	12	2	87	0	0	0	0
ɔɪ	0	0	0	0	0	0	0	0	0	0	2	97	0	0	2
ʌ	0	0	0	0	32	2	12	17	2	0	0	0	37	0	0
ʊ	0	0	2	0	0	0	0	0	0	3	2	0	0	83	10
u	0	0	0	0	0	0	0	0	0	0	2	0	0	18	80

Table 12
Vowel identification percentages: F2AV condition

	Response														
Stimulus	i	ɪ	eɪ	aɪ	ɛ	ɚ	æ	ɑ	ɔ	o	aʊ	ɔɪ	ʌ	ʊ	u
i	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ɪ	0	93	0	2	5	0	0	0	0	0	0	0	0	0	0
eɪ	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
aɪ	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
ɛ	0	3	0	0	92	0	5	0	0	0	0	0	0	0	0
ɚ	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
æ	0	0	0	0	30	0	70	0	0	0	0	0	0	0	0
ɑ	0	0	0	0	0	0	0	67	30	2	0	0	2	0	0
ɔ	0	0	0	0	0	0	0	37	62	0	2	0	0	0	0
o	0	0	0	0	0	0	0	0	0	93	7	0	0	0	0
aʊ	0	0	0	0	0	0	0	5	5	0	90	0	0	0	0
ɔɪ	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
ʌ	0	0	0	0	0	0	0	22	0	0	0	0	73	3	2
ʊ	0	0	0	0	0	0	0	0	0	0	0	0	5	82	13
u	0	0	0	0	0	0	0	0	0	22	2	0	0	3	73

word identifications for monosyllabic words that reside in relatively dense regions of the lexicon. The absolute levels of obtained performance were lower than predicted by the expected percent correct computations for monosyllabic words, but the relative performance across conditions matched predictions and audiovisual performance was superadditive; subjects accurately identified the stimulus words at a rate of 11.1% for V, 6.1% for F1A, 32.9% for F2A, 43.7% for F1AV, and 62.5% for F2AV. In addition, expected percent correct correlated with these word identification results on an individual-word basis; within each condition, words in small equivalence classes tended to be identified more accurately than words in large equivalence classes. Our preliminary experimental evidence thus suggests that this computational method is valid, and that, despite the fact that there can be individual differences in phoneme identification, the data collected in the present

study can be extended to make predictions for other participants.

To summarize, the findings suggest that audiovisual enhancement levels in word recognition result from an interaction of the phonetic information required for lexical uniqueness and the phonetic information available under audio, visual and audiovisual conditions. Superadditive levels of enhancement appear more likely for monosyllabic than for multisyllabic words, because even small quantities of unimodal phonetic information may be sufficient for accurate recognition of multisyllabic words. Considering that the practical significance of audiovisual integration lies in the ability of visual information to aid in word recognition when the available audio information is impoverished, it seems crucial to extend investigation beyond the level of phonetic perception.

Acknowledgements

We thank Brian K. Chaney for computer programming, John Jordan for vocoder maintenance, and Paula E. Tucker for assistance in the preparation of this article. This research was funded by National Institutes of Health grants (DC00695, DC02107) to L.E. Bernstein. Data collection for consonants took place while L.E. Bernstein and E.T. Auer Jr. were at Gallaudet University; all other work took place at House Ear Institute. Finally, we are grateful to Christian Benoît, both for actively promoting the audiovisual speech research field, and for inviting us to contribute to this volume.

Appendix A. Phoneme identification confusion matrices

Tables 3–12 contain the stimulus-response confusion matrices for each of the five experimental conditions. The number in each cell reflects the percentage of trials in which the phoneme stimuli for that row were identified as the phoneme of that column. The percentages were calculated based on the total number of trials for each stimulus phoneme (i.e., the sum of the percentages in a row equals 100).

References

Aldenderfer, M.S., Blashfield, R.K., 1984. Cluster Analysis. Sage, Beverly Hills, CA.

Altmann, G.T.M. (Ed.), 1990. Cognitive Models of Speech Processing, MIT Press, Cambridge, MA.

Altman, G., Carter, D., 1989. Lexical stress and lexical discriminability: Stressed syllables are more informative, but why? Computer Speech and Language 3, 265–275.

Auer Jr., E.T., Bernstein, L.E., 1996. Lipreading supplemented by voice fundamental frequency: To what extent does the addition of voicing increase lexical uniqueness for the lipreader? In: ICSLP '96 Proc., Philadelphia, PA, 3–6 October 1996, pp. 86–93.

Auer Jr., E.T., Bernstein, L.E., 1997. Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. Journal of the Acoustical Society of America 102, 3704–3710.

Aull, A.M., Zue, V.W., 1985. Lexical stress determination and its application to large vocabulary speech recognition. In: ICASSP Proc., Tampa, FL, 26–29 March 1985, pp. 1549–1552.

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

Bernstein, L.E., Eberhardt, S.P., Demorest, M.E., 1989. Single-channel vibrotactile supplements to visual perception of intonation and stress. Journal of the Acoustical Society of America 85, 397–405.

Bernstein, L.E., Demorest, M.E., Coulter, D.C., O'Connell, M.P., 1991. Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. Journal of the Acoustical Society of America 90, 2971–2984.

Bernstein, L.E., Coulter, D.C., O'Connell, M.P., Eberhardt, S.P., Demorest, M.E., 1993. Vibrotactile and haptic speech codes. In: Risberg, A., Felicetti, S., Plant, G., Spens, K.-E. (Eds.), Proc. Second International Conference on Tactile Aids, Hearing Aids, and Cochlear Implants, Stockholm, 7–11 June 1992, pp. 57–70.

Bernstein, L.E., Iverson, P., Auer Jr., E.T., 1997. Elucidating the complex relationships between phonetic perception and word recognition in audiovisual speech perception. In: Benoît, C., Campbell, R. (Eds.), Proc. ESCA/ESCOP Workshop on Audio-Visual Speech Processing: Cognitive and Computational Approaches, Rhodes, Greece, 26–27 September 1997, pp. 89–92.

Braida, L.D., 1991. Crossmodal integration in the identification of consonant segments. Quarterly Journal of Experimental Psychology 43A, 647–677.

Carlson, R., Elenius, K., Granstrom, B., Hunnicutt, S., 1985. Phonetic and orthographic properties of the basic vocabulary of five European languages. Speech Transmission Laboratory-Quarterly Progress and Status Report 1/1985, 63–94.

Carter, D.M., 1987. An information-theoretic analysis of phonetic dictionary access. Computer Speech and Language 2, 1–11.

Engebretson, A.M., O'Connell, M.P., 1986. Implementation of a microprocessor-based tactile hearing prosthesis. IEEE Transactions on Biomedical Engineering BME-33, 712–716.

Fisher, C.G., 1968. Confusions among visually perceived consonants. Journal of Speech and Hearing Research 11, 796–804.

Grant, K.W., Walden, B.E., 1996. Evaluating the articulation index for auditory-visual consonant recognition. Journal of the Acoustical Society of America 100, 2415–2424.

Green, K.P., Kuhl, P.K., 1989. The role of visual information in the processing of place and manner features in speech perception. Perception and Psychophysics 45, 34–42.

Huttenlocher, D.P., Zue, V.W., 1984. A model of lexical access from partial phonetic information. In: ICASSP Proc., San Diego, CA, 19–21 March 1984, pp. 26.4.1–26.4.4.

Iverson, P., Bernstein, L.E., Auer Jr., E.T., 1997. A comparison of perceptual word similarity metrics. Journal of the Acoustical Society of America 102, 3189.

Kricos, P.B., Lesner, S.A., 1982. Differences in visual intelligibility across talkers. Volta Review 84, 219–225.

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

Lahiri, A., Marslen-Wilson, W., 1991. The mental representation of lexical form: A phonological approach to the recognition lexicon. Cognition 38, 245–294.

Landauer, T.K., Streeter, L.A., 1973. Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. Journal of Verbal Learning and Verbal Behavior 12, 119–131.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M., 1967. Perception of the speech code. Psychological Review 74, 431–461.

Luce, P.A., 1986. Neighborhoods of words in the mental lexicon. Indiana University Ph.D. Dissertation.

Luce, P.A., Pisoni, D.B., Goldinger, S.D., 1990. Similarity neighborhoods of spoken words. In: Altmann, G.T.M. (Ed.), Cognitive Models of Speech Processing. MIT Press, Cambridge, MA, pp. 122–147.

Massaro, D.W., 1987. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Erlbaum, Hillsdale, NJ.

McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cognitive Psychology 18, 1–86.

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

Norusis, M.J., 1993. SPSS Professional Statistics 6.1. SPSS, Chicago.

Owens, E., Blazek, B., 1985. Visemes observed by hearing-impaired and normal-hearing adult viewers. Journal of Speech and Hearing Research 28, 381–393.

Pisoni, D.B., Nusbaum, H.C., Luce, P.A., Slowiaczek, L.M., 1985. Speech perception, word recognition and the structure of the lexicon. Speech Communication 4, 75–95.

Reisberg, D., McLean, J., Goldfield, A., 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In: Dodd, B., Campbell, R. (Eds.), Hearing by Eye: The Psychology of Lip-Reading. Erlbaum, London, pp. 97–113.

Savin, H.B., 1962. Word-frequency effect and errors in the perception of speech. Journal of the Acoustical Society of America 35, 200–206.

Seitz, P.F., Bernstein, L.E., Auer Jr., E.T., 1995. PhLex (Phonologically Transformable Lexicon), A 35,000-word computer readable pronouncing American English lexicon on structural principles, with accompanying phonological rules and word frequencies. Gallaudet Research Institute, Washington, DC.

Sekiyama, K., Tohkura, Y., 1991. McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. Journal of the Acoustical Society of America 90, 1797–1805.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 26, 212–215.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., Jones, C.J., 1977. Effects of training on the visual recognition of consonants. Journal of Speech and Hearing Research 20, 130–145.

Woodward, M.F., Barber, C.G., 1960. Phoneme perception in lipreading. Journal of Speech and Hearing Research 3, 212–222.