Perception & Psychophysics
2007, 69 (7), 1070-1083

Similarity structure in visual speech perception and optical phonetic signals

Jintao Jiang
House Ear Institute, Los Angeles, California
and University of California, Los Angeles, California

Edward T. Auer, Jr.
House Ear Institute, Los Angeles, California
and University of Kansas, Lawrence, Kansas

Abeer Alwan and Patricia A. Keating
University of California, Los Angeles, California

and

Lynne E. Bernstein
House Ear Institute, Los Angeles, California

A complete understanding of visual phonetic perception (lipreading) requires linking perceptual effects to physical stimulus properties. However, the talking face is a highly complex stimulus, affording innumerable possible physical measurements. In the search for isomorphism between stimulus properties and phonetic effects, second-order isomorphism was examined between the perceptual similarities of video-recorded perceptually identified speech syllables and the physical similarities among the stimuli. Four talkers produced the stimulus syllables comprising 23 initial consonants followed by one of three vowels. Six normal-hearing participants identified the syllables in a visual-only condition. Perceptual stimulus dissimilarity was quantified using the Euclidean distances between stimuli in perceptual spaces obtained via multidimensional scaling. Physical stimulus dissimilarity was quantified using face points recorded in three dimensions by an optical motion capture system. The variance accounted for in the relationship between the perceptual and the physical dissimilarities was evaluated using both the raw dissimilarities and the weighted dissimilarities. With weighting and the full set of 3-D optical data, the variance accounted for ranged between 46% and 66% across talkers and between 49% and 64% across vowels. The robust second-order relationship between the sparse 3-D point representation of visible speech and the perceptual effects suggests that the 3-D point representation is a viable basis for controlled studies of first-order relationships between visual phonetic perception and physical stimulus attributes.

Speech production biomechanics generate optical phonetic as well as acoustic phonetic signals, and humans typically integrate the information afforded by both. A growing list of audiovisual phenomena demonstrates the influence of visual speech stimuli on speech perception. The well-known McGurk (McGurk & MacDonald, 1976) and ventriloquist (De Gelder & Bertelson, 2003) effects demonstrate audiovisual integration. Being able to see a talker produces substantial gains to comprehending acoustic speech in noise (MacLeod & Summerfield, 1987; Sumby & Pollack, 1954), improvements in comprehending difficult messages even under good listening conditions (Arnold & Hill, 2001; Reisberg, McLean, & Goldfield, 1987), speech detection under adverse acoustic signal-to-noise conditions (Bernstein, Auer, & Takayanagi, 2004; Grant, 2001; Grant & Seitz, 2000), compensation for auditory speech information that is reduced by filtering out various frequency bands (Grant & Walden, 1996) or by hearing loss (Erber, 1975; Grant, Walden, & Seitz, 1998), and superadditive levels of speech perception from combinations of extremely minimal auditory speech information and visible speech (Breeuwer & Plomp, 1986; Iverson, Bernstein, & Auer, 1998; Kishon-Rabin, Boothroyd, & Hanin, 1996; Moody-Antonio et al., 2005).

Visual speech stimuli alone afford reduced phonetic information, relative to auditory speech stimuli that are presented under good listening conditions. Speech production activities are partially occluded from view by the lips, cheeks, and neck (e.g., hidden from view is vocal fold vibration, related to the phonological voicing distinction; partially hidden is the type of vocal tract closure made by the tongue, related to the phonological manner distinctions;

J. Jiang, jjiang@hei.org

and hidden is the state of the velum, related to nasality). As a result, fairly systematic, although far from invariant, clusters of confusions (e.g., /m, b, p/) among visual speech segments are regularly observed (cf. Kricos & Lesner, 1982; Owens & Blazek, 1985; Walden, Prosek, Montgomery, Scherr, & Jones, 1977). Fisher (1968) coined the term *viseme* to capture this sort of perceptual similarity among groups of phonemes. Visemes are sometimes regarded as unitary perceptual categories, having no internal perceptual structure that conveys additional phonemic information (e.g., Massaro, 1998). Nevertheless, lip-readers are able to discriminate and identify words whose segments are within the same viseme groups (Bernstein, 2006).

Despite the reduction in phonetic information afforded by visible speech stimuli, speech can be perceived via lipreading alone (also referred to as *speechreading*; see, e.g., Bernstein, Demorest, & Tucker, 2000; Dodd, McIntosh, & Woodhouse, 1998). Estimates of the information available via lipreading vary. On the low side, lipreading accuracy has been reported to be approximately 10%–30% words correct in isolated sentences (e.g., Breeuwer & Plomp, 1986; Demorest & Bernstein, 1992; Rönnberg, 1995; Rönnberg, Samuelsson, & Lyxell, 1998) and approximately 30%–60% phonemes correct in syllables (e.g., Fisher, 1968; Montgomery & Jackson, 1983; Owens & Blazek, 1985). On the other hand, congenitally deaf adults frequently demonstrate very good accuracy levels (Andersson & Lidestam, 2005; Auer & Bernstein, 2007; Auer, Bernstein, & Tucker, 2000; Bernstein, Auer, & Tucker, 2001; Bernstein, Demorest, & Tucker, 1998, 2000; Mohammed et al., 2005). In a group of 72 deaf adults, the top quartile lip-read in the range of 65%–85% words correct on a set of isolated prerecorded sentences (Bernstein et al., 2000). Although adults with normal hearing are generally less accurate lip-readers, their accuracy levels can also be moderately high (Auer & Bernstein, 2007; Bernstein et al., 2001; Bernstein et al., 2000).

The variance in individual lipreading performance can be accounted for, to a large extent, with measures of visual phonetic perception (e.g., nonsense syllable identification) and isolated spoken word recognition (Andersson & Lidestam, 2005; Bernstein et al., 2000). Visual spoken word recognition and discrimination are sensitive to and can be predicted on the basis of visual perceptual phonetic distinctiveness (Auer, 2002; Auer & Bernstein, 1997; Bernstein, 2006; Iverson et al., 1998; MacEachern, 2000; Mattys, Bernstein, & Auer, 2002).

Relatively little is known about the segmental phonetic (subphonemic) level of visual speech processing. Researchers who use visual speech stimuli typically record natural talkers and describe the stimuli primarily in terms of the recording conditions, the gender and language of the talker, and the linguistic content of the utterances (phonemes, words, sentences, etc.; see Munhall & Vatikiotis-Bateson, 1998). Some effects of global physical stimulus factors have been examined, such as overall spatial resolution (Erber, 1979; Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004), viewing angle (Jordan & Thomas, 2001), presence or absence of dynamic information (Campbell, 1996; Rosenblum & Saldaña, 1998), and color and luminance (McCotter & Jordan, 2003).

Another global approach has been to present isolated face parts or faces with parts removed (e.g., Benoît, Guiard-Marigny, Le Goff, & Adjoudani, 1996; IJsseldijk, 1992; Marassa & Lansing, 1995; Scheinberg, 1980; Thomas & Jordan, 2004). This research has been criticized on the grounds that face parts were not necessarily isolated adequately, that the displays encouraged unnatural attentional strategies, that configural properties of the stimuli were disrupted, and that the isolation techniques introduced potentially unnatural elements, such as edges around the relevant stimulus (Thomas & Jordan, 2004). Thomas and Jordan isolated the lip area within the face and produced stimuli without edges around the visible moving area of the face. They reported comparisons among regions of the face showing the lip region to be most informative. But studies that isolate or freeze parts of the face beg the question of how to characterize the phonetically relevant physical information, as is the case also for studies that manipulate global optical signal properties. In addition, as is the case with acoustic speech signals (Repp, 1981), optical phonetic signals are likely to involve complex relationships among cues.

Physical Signals and Speech Perception

Decades of research on acoustic speech signals have resulted in a detailed understanding of the physical acoustic characteristics of speech (Repp, 1981; Stevens, 1998). For example, relationships among formant frequency patterns that support consonant perception are quite well understood. Acoustic phonetic characteristics have been manipulated in scores of auditory speech perception experiments designed to test their perceptual relevance. As a result, rules are known for acoustic synthesis (Klatt, 1987). The talking face is a complex spatiotemporal stimulus. Knowledge about its phonetically relevant descriptors lags far behind knowledge about acoustic phonetic cues. A nonarbitrary approach is needed for learning about phonetically relevant optical descriptors.

Shepard and Chipman (1970) considered the problem of establishing the isomorphism between physical stimuli and internal (perceptual) representations. They noted that internal representations are unlikely to be structurally isomorphic with stimuli, in the sense that the internal representation of a square is not likely to be square. In order to approach the problem of establishing relationships between complex stimuli and internal representations, they argued that an "isomorphism should be sought—not in the first-order relation between (a) an individual object, and (b) its corresponding internal representation—but in the second-order relation between (a) the relations among alternative external objects, and (b) the relations among their corresponding internal representations. Thus, although the internal representation for a square need not itself be square, it should (whatever it is) at least have a closer functional relation to the internal representation for a rectangle than to that, say, for a green flash or the taste of a persimmon" (p. 2).

A few previous studies have attempted to relate optical speech measures to perception, but, to our knowledge, all of them incorporated attributes of first-order isomorphism, and all reported limited success. Plant (1980) studied 20 Australian English vowels and diphthongs (in /b/–vowel–/b/ context). Participants viewed recordings in an identification task. Measurements of inner and outer lip vertical and horizontal extent and of chin excursion were performed on individual stimulus frames at the midpoint of each vowel. Evaluation of the phonetic-to-perceptual relations in the data was qualitative. Selection of measures interpreted as likely to be perceptually important was influenced by their absolute magnitudes. For example, lower lip excursions, which are larger than upper lip excursions, were thought to be related to perceptual identification accuracy and patterns of mutual confusions among vowels.

Jackson, Montgomery, and Binnie (1976) used physical measures of the lips to account directly for perceptual similarity ratings among pairs of vowels from a set of 15 /h/–vowel–/g/ syllables spoken by four female talkers. The similarity ratings were submitted to multidimensional scaling (MDS; Kruskal & Wish, 1978), and a 5-D perceptual space was generated. Correlations were computed between the component for each vowel on a perceptual dimension and physical measures of its horizontal lip-spreading, lip-rounding, and lip-opening area. That is, each of the derived dimensions of perceptual similarity was associated directly to each of the face measurements. Overall, the method was a hybrid account comprising elements of both first- and second-order isomorphism.

Montgomery and Jackson (1983) conducted additional analyses on 10 vowels from Jackson et al.'s (1976) study. MDS of perceptual identifications resulted in a 2-D space that they labeled as lip spreading/rounding and tongue height. Tracings were made on a single frame of the lips during their maximum opening or constriction, and measures were performed on the tracings: They were lip height, lip width, lip aperture area, acoustic duration, and visual duration. Physical difference scores computed between pairs of stimuli expressed as absolute values were used as predictors in multiple regression models. The perceptual judgments informed the selection of stimulus features for use in the regression models. Preliminary analyses showed that the physical measures were "less than perfect" in reproducing the distinctions that were perceived. Variance accounted for ranged between 24% and 68% across talkers, with the former not significant. The significant variables in the regression models varied across talkers. The authors concluded that perceivers must have means to compensate for stimulus noninvariances, that static measures are necessarily exploratory, and that much of lipreading behavior remained to be explained. The approach appears to exemplify second-order isomorphism, but the use of the perceptual identifications to select specific physical measures of the lips seems appropriately characterized as a first-order approach. Had the results been more successful, the likely interpretation would have been that perceivers internally represent the vowels using the quantities that were directly measured. The mixed results across talkers demonstrate the hazards of commitment to specific stimulus features among the plethora available even with a single video frame. In addition, restricting analysis to a single video frame precluded measures that would reflect speech motion.

In summary, previous accounts of visual segmental phonetic perception based on physical stimulus measures have been modestly successful. The talking face is a highly complex stimulus, affording innumerable physical measurements. In order to gain traction on the relationship between optical quantities and phonetic perceptual effects, it may be necessary first to establish a reliable second-order isomorphism between perception and physical signals. Ideally, the physical signals should be reduced in complexity, relative to natural visible speech, so that subsequent studies can focus narrowly on signal attributes. For example, the early pattern playback acoustic speech synthesizer produced a reduced and highly schematized speech signal that led to productive research on the first-order relationships between acoustic cues and speech perception. The rule for auditory phonetic perception is that multiple complex acoustic quantities (cues) contribute to phonetic percepts (Repp, 1981). The same is likely true for visual phonetic perception. Shepard and Chipman (1970) argued for the validity of second-order isomorphism when knowledge about the descriptors of a first-order isomorphism is not available. They also noted that physical properties of a stimulus could be internally represented with different weightings but that the overall similarity of objects has functional significance to organisms. The present study had the goal of demonstrating that visual phonetic similarity (dissimilarity) is directly related to physical stimulus similarity (dissimilarity). An alternative possible hypothesis is that visual phonetic stimuli are mapped to abstract linguistic categories, so that a second-order isomorphism relationship cannot be established.

The Present Study

In this study, physical quantities were used that represented the movements in three dimensions of points on the faces of talkers saying a large inventory of syllables. No derived visual features were measured. In a perceptual experiment (Experiment 1), natural speech stimuli (silent video) were presented for perceptual identification. In an experiment designed to demonstrate second-order isomorphism (Experiment 2), (1) physical distances (dissimilarities) were computed between stimulus tokens, using the channels of 3-D data for each stimulus across an interval of 280 msec; (2) perceptual identifications were transformed into perceptual dissimilarities; and (3) the correspondence between physical and perceptual dissimilarities was evaluated.

Well-defined computational procedures exist for representations in terms of similarities/dissimilarities (Edelman, 1998), particularly in terms of either feature or geometric models (Goldstone, 1999; Nosofsky & Stanton, 2005). Feature-based similarity can be computed by tallying the number of shared (redundant) and unshared (contrastive) features (e.g., Frisch, Pierrehumbert, & Broe, 2004). But as has already been suggested, optical phonetic features have not yet been defined (Bernstein,

2006). A variety of methods exist for quantifying perceptual similarity structure without recourse to features. For example, direct perceptual similarity judgments on stimulus pairs can be obtained, followed by an analysis of the obtained response matrix (Jackson et al., 1976). However, the method of similarity judgment is prohibitively time consuming whenever there are many stimuli, as was the case here, and similarity judgments need not engage the same perceptual processes as perceptual identification (Jackson et al., 1976), which is arguably closer to everyday speech perception. In Experiment 1, perceptual identifications of nonsense syllables were obtained. The stimulus set comprised variations across talker, vowel, and consonant and was considered to be a challenge, due to its diversity, for achieving a statistically reliable second-order isomorphism. The perceptual identifications were subsequently (Experiment 2) submitted to MDS to locate the stimuli in perceptual spaces and to facilitate computation of Euclidean distances (dissimilarities) among stimulus pairs. Three-dimensional optical data that were obtained simultaneously with the video recordings of the stimuli were used to estimate physical stimulus dissimilarities. Small retroreflectors were glued on the talkers' faces and were optically tracked in real time. The optical data were an accurate record of the speech movements at a set of face locations.

The signal-processing technique known as *least-squares linear estimation* (Kailath, Sayed, & Hassibi, 2000) was used to transform the high-dimensionality physical distance matrices of stimulus pairs into the same dimensionality as the perceptual dissimilarities and to warp the physical dissimilarity data. In order to achieve independence between the calculation of the least-squares linear estimation weights and the evaluation of the second-order isomorphism, different perceptual data were used for the two operations. The variance accounted for in the relationship between the physical and the perceptual similarities was evaluated with Pearson correlations and multiple regression.

EXPERIMENT 1

Method

Participants. Six participants (labeled as P1 to P6; mean age, 32 years; range, 22–43 years; 2 males), with normal hearing, normal or corrected-to-normal vision, English as a native language, and average or better lipreading (as determined using a screening procedure; Auer & Bernstein, 2007) were recruited. They gave informed consent and were paid $10/h. Five participants completed testing within 3 weeks, and the 6th within 8 weeks.

Talkers and stimulus materials. The materials were spoken by four talkers (two males [M1 and M2] and two females [F1 and F2]) with English as a native language. The talkers had been selected from a larger pool that had initially been screened for their visual intelligibility by presenting video-recorded sentences to five deaf lip-readers. Screening was conducted to obtain talkers who varied in their visual intelligibility. Subsequently, extensive visual-only speech perception testing (deaf perceivers, $N = 8$) of 320 IEEE sentences (IEEE, 1969) produced by each of these talkers showed that the order of talker intelligibility in terms of percentage of words correct from most to least intelligible was Talker F2, Talker M2, Talker M1, and Talker F1.

The speech stimuli spoken by each talker comprised two tokens each of 69 consonant–vowel (CV) nonsense syllables in American English, for which the vowel was one of /ɑ, i, u/, and the consonant was one of the 23 consonants: /y, w, r, l, m, n, p, t, k, b, d, g, h, θ, ð, s, z, f, v, ʃ, ʒ, tʃ, dʒ/. The total stimulus set comprised 552 video-recorded utterances (69 CVs × 2 tokens × 4 talkers).

Prior to being recorded, the talkers practiced saying each syllable with a falling intonation at a normal speaking rate. During recording, syllables were pseudorandomly presented on a teleprompter. The talker's face filled the video screen. Lighting on the talker was from spotlights at both sides and slightly below the talker's head. Multiple tokens were recorded, and only tokens that began and ended with a closed mouth were selected for the experiment.

The video-recording equipment was a production quality camera (Sony DXC-D30 digital) and video recorder (Sony UVW 1800). The optical recording equipment was a three-camera, 3-D optical recording system (Qualisys MCU120/240 Hz CCD Imager), which digitally recorded the positions of passive retroreflectors during infrared flashes (not perceptible to the talker).

The stimuli were presented from BETACAM SP videotapes. Stimulus tapes were blocked by talker, and tokens were pseudorandomized across consonants and vowels. Two stimulus tapes were generated. On Tape 1, the talker order was M1–F2–M2–F1. On Tape 2, the order was F2–M1–F1–M2.

Procedure. The participants viewed the silent video and made 23-alternative forced choice consonant identifications using a computer mouse and a PC graphical display monitor located adjacent to the video display monitor. At the beginning of each testing session, instructions were displayed on the PC monitor. The stimuli were presented on a 14-in. high-resolution color monitor (Sony Trinitron) at a distance of 1 m from the participant. After the participants had acknowledged reading the instructions, a computer program presented each of the stimuli. After each stimulus presentation, the video monitor became black. At the same time, the graphical display on the PC monitor was activated, showing 23 consonant labels with corresponding sample words to exemplify pronunciation. The participant then chose a response. Following the response, the computer program presented the next syllable. No feedback was given at any time. A practice set of 10 trials was given on Day 1 only. Testing took place in a sound-treated IAC booth.

Testing was administered in blocks of 138 pseudorandomized items (2 tokens × 69 CVs) for each of the four talkers. Typically, the participants responded to one list for each talker on each day of testing. Each list required approximately 16 min to complete, and a 5-min break was given between lists. The order of viewing the tapes was counterbalanced across participants. Occasionally, the participants received more than four lists of trials, but not more than eight per day of testing. A half-hour rest was always given if the number of lists exceeded four. Each participant contributed a total of 10 responses for each stimulus token.

Analyses. For each CV syllable type (i.e., 23 consonants × 3 vowels), 480 identification responses (i.e., 2 tokens × 10 trials × 4 talkers × 6 participants) were obtained. The data were pooled into 12 (4 talkers × 3 vowels) stimulus–response confusion matrices (23 × 23). Each confusion matrix contained 120 responses (2 tokens × 10 trials × 6 participants) per CV syllable type.

Results and Discussion

Figure 1 displays the percent correct consonant identification scores for each talker, vowel, and participant. P1 was the most and P6 the least accurate participant.

An omnibus repeated measures ANOVA was carried out with talker (4) and vowel (3) as the repeated factors. The main effect of talker was significant [$F(3,3) = 23.64$, $p = .014$]. The mean proportion correct for each talker was (from high to low) the following: F2, .351 ($SD = .040$); M2, .347 ($SD = .059$); M1, .329 ($SD = .051$); and F1,

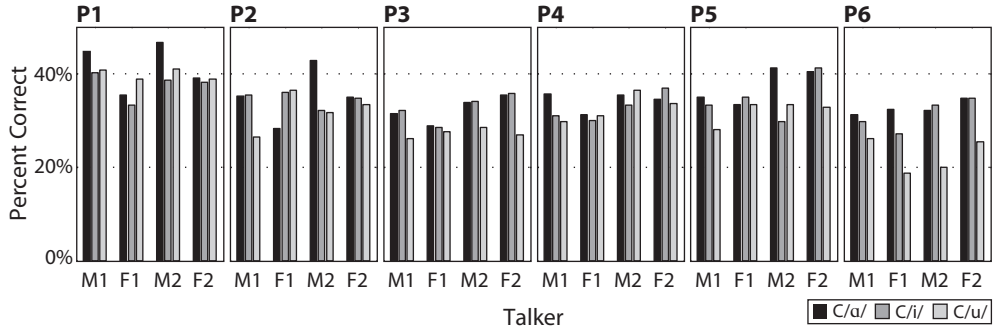


Figure 1. Mean percent correct scores as a function of talker and vowel context for each participant (P1–P6).

.315 (*SD* = .045). Contrast analyses showed that percent correct scores for Talker F2 were not significantly different from those for Talker M2 [$F(1,5) = 0.16$, $p = .706$] or M1 [$F(1,5) = 3.24$, $p = .132$] but were significantly higher than those for Talker F1 [$F(1,5) = 29.92$, $p = .003$]. Percent correct scores for Talker M2 were significantly higher than those for Talkers M1 [$F(1,5) = 7.73$, $p = .039$] and F1 [$F(1,5) = 17.05$, $p = .009$]. Percent correct scores for Talkers M1 and F1 were not significantly different [$F(1,5) = 1.57$, $p = .266$].

The main effect of vowel was significant [$F(2,4) = 10.41$, $p = .026$]. The mean proportion correct for each vowel was the following: C/ɑ/, .356 (*SD* = .047); C/i/, .340 (*SD* = .036); and C/u/, .311 (*SD* = .060). These results are commensurate with others in the literature (Iverson et al., 1998; Owens & Blazek, 1985; Walden et al., 1977). There were no interactions between talker and vowel [$F(6,30) = 2.41$, $p = .051$].

Contrast analyses showed that percent correct scores for C/ɑ/ syllables were significantly higher than those for C/i/ [$F(1,5) = 7.67$, $p = .039$] and C/u/ [$F(1,5) = 11.74$, $p = .019$] syllables, and that scores for C/i/ and C/u/ syllables were not significantly different [$F(1,5) = 3.28$, $p = .130$].

Figure 2 shows that individual consonant scores averaged across talkers and vowels ranged from .04 (/ʒ/) to .89 (/w/). Previously, patterns of consonant confusions (viseme groupings) were shown to vary as a function of vowel (Owens & Blazek, 1985), suggesting that consonant dissimilarity is sensitive to vowel context. In the present study, the pattern of correct scores were similar between C/u/ and C/ɑ/ ($r = .74$), between C/u/ and C/i/ ($r = .78$), and between C/ɑ/ and C/i/ ($r = .90$) ($df = 21$, $p < .001$). Correlations were not significantly different from each other.

Percent correct scores represent only the diagonal in the stimulus–response confusion matrices. To examine response patterns in the overall matrices, a *phoneme equivalence class* (PEC) analysis (Auer & Bernstein, 1997; Iverson et al., 1998) was applied. The analysis derives groups of perceptually similar consonants, using hierarchical cluster analysis. PECs were chosen by finding the first level in the cluster hierarchy at which at least 75% of all the responses were within the cluster, similar to Walden et al.

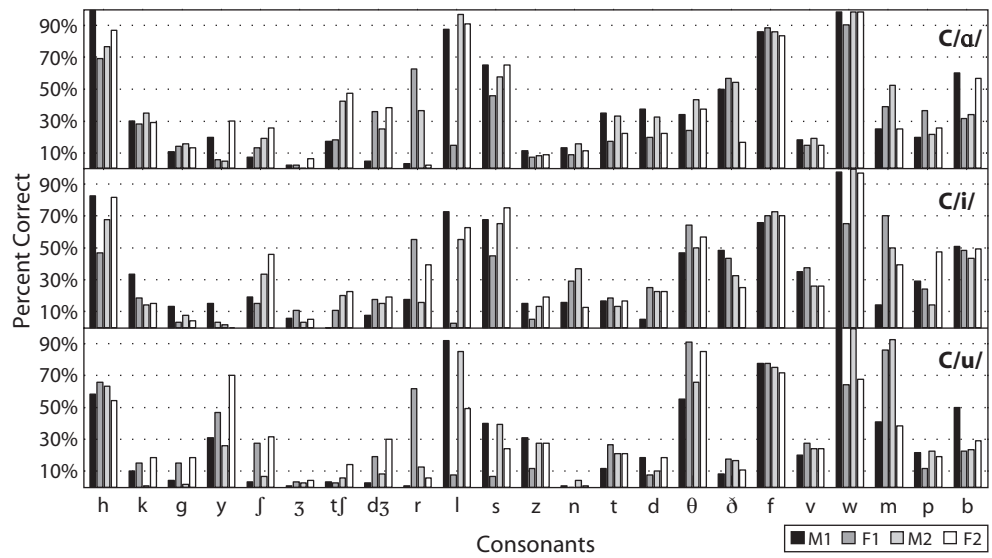


Figure 2. Identification accuracy for 23 consonants by talkers and vowels.

(1977). Table 1 lists the PECs for each talker and vowel context. The analyses showed that the similarity among consonants varied across talkers and vowels and that consonants tended to be grouped by place of articulation. For example, there were fewer PECs for Talker F1 than for Talker M2. Talker F1 was least intelligible, and the PEC analysis showed that her nonfront consonants produced large undifferentiated PECs for the three vowel contexts (including the consonant /l/). The PEC analyses showed that the more difficult C/u/ syllables were generally associated with fewer PECs. In other words, when the consonant was followed by /u/, it was perceived as more similar to the other consonants. PEC analysis is useful in showing broad similarity groupings; however, previous research has shown good to excellent within-PEC discrimination and identification with word stimuli (Bernstein, 2006).

In summary, the perceptual results showed that consonant intelligibility varied greatly across the 23 consonants and three vowels. Individual talkers differed in the information they afforded the participants, and the participants varied in their perception of the stimuli. Previously, Montgomery and Jackson (1983) were unable to show consistent results across talkers in their study of vowels, which are arguably phonetically simpler. The extent of variation in Experiment 1 was considered a serious challenge to demonstrating second-order isomorphism in Experiment 2.

EXPERIMENT 2

In the previous attempts to relate physical to perceptual measures, perceptual processing was taken into account by making explicit measures of the stimuli—that is, by assuming that certain stimulus attributes were relevant to perception. But the goal here was to impose as few constraints or presuppositions as possible in constructing the second-order isomorphism. For example, no attempt was made to realize a sensory–perceptual front-end vision processor to warp perceptually the physical data, an undertaking far beyond the scope of this study. Instead, the approach was to compute a set of linear weights that minimized the second-order distance between physical and perceptual dissimilarities. This was accomplished using least-squares linear estimation (Kailath et al., 2000). Specifically, least-squares linear estimation is used to obtain a weights vector that is multiplied by a higher dimension matrix (Euclidean distances for multiple channels of data between stimuli), reducing its dimensionality and minimizing the distance between it and the lower dimension matrix (perceptual dissimilarities). The minimization equation (Equation 4; see below) for this operation has a standard solution (Equation 5; see below) (Kailath et al., 2000).

Movements of individual points, particularly ones that were near each other, could not be considered to contribute independent information; neither could individual dimensions in x, y, and z. The multicollinearity among points and dimensions precluded applying statistical testing of the variance accounted for by each weight. All weights were retained. Subsequent statistical analyses were used to evaluate the variance accounted for in the relationship between the warped physical similarities and the perceptual similarities. Pearson correlations were computed across all the pairs of physical versus perceptual dissimilarities. Multiple regression was also applied conventionally to obtain statistical evaluation of beta weights for physical data partitioned into sets for the chin, cheeks, and lips. This analysis assumed independence among the subparts of the face, although some collinearity was surely captured in the analysis.

Method

Figure 3 is a flow diagram of the methods in Experiment 2. They will be described in detail below.

Perceptual dissimilarities. The data from Experiment 1 were pooled into several different matrices—for example, as a function of talker and vowel. For every matrix that resulted from pooling, the following methods were applied.

First, distributions of responses to individual pairs of stimuli were submitted to a *phi-square* transformation (SPSS, 2003), which expresses the *substitutability* of the phonemes within the pair and is a normalized version of the chi-square coefficient. The phi-square transformation compensates for response biases and asymmetries in the confusion matrix of stimulus–response counts (Iverson et al., 1998; Siegel & Castellan, 1988). Phi-square coefficients replace the count data and are computed as

$$\phi^2(x,y) = \frac{1}{\sqrt{N}} \cdot \sqrt{\sum_i \frac{\left[x_i - \mathrm{E}_{xy}(x_i)\right]^2}{\mathrm{E}_{xy}(x_i)} + \sum_i \frac{\left[y_i - \mathrm{E}_{xy}(y_i)\right]^2}{\mathrm{E}_{xy}(y_i)}}, \quad (1)$$

where x_i and y_i are the frequencies with which phonemes x and y are identified as response category i. N equals the total number of

Table 1
Phoneme Equivalence Classes Obtained by Cluster Analyses for Confusion Matrices of Each Talker and Vowel Context

Talker	Vowel	Phoneme Equivalence Classes
M1	ɑ	{w} {m p b} {r f v} {h} {θ ð y l n k g} {t d s z ʃ ʒ tʃ dʒ}
	i	{w} {m p b} {r f v} {y k g h} {θ ð l n t d} {s z ʃ ʒ tʃ dʒ}
	u	{w y k g h} {m p b} {r f v} {θ ð l n t d} { s z ʃ ʒ tʃ dʒ }
F1	ɑ	{w r} {m p b} {f v} {θ ð} {y l n k g h t d s z ʃ ʒ tʃ dʒ }
	i	{w r} {m p b} {f v} {θ ð} {y l n k g h t d s z ʃ ʒ tʃ dʒ }
	u	{w r m p b} {f v} {θ ð} {y l n k g h t d s z ʃ ʒ tʃ dʒ}
M2	ɑ	{w r} {m p b} {f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
	i	{w r} {m p b} {f v} {θ ð} {y l n t k d g h} {s z ʃ ʒ tʃ dʒ}
	u	{w r} {m p b} {f v} {θ ð} {y l n k g h} {t d s z ʃ ʒ tʃ dʒ}
F2	ɑ	{w} {m p b} {r f v} {y l n k g h θ ð t d s z ʃ ʒ tʃ dʒ}
	i	{w} {m p b} {r f v} {θ ð} {y l n k g h} {t d s z} {ʃ ʒ tʃ dʒ}
	u	{m p b} {r f v} {θ ð} {w y l n k g h t d s z ʃ ʒ tʃ dʒ}

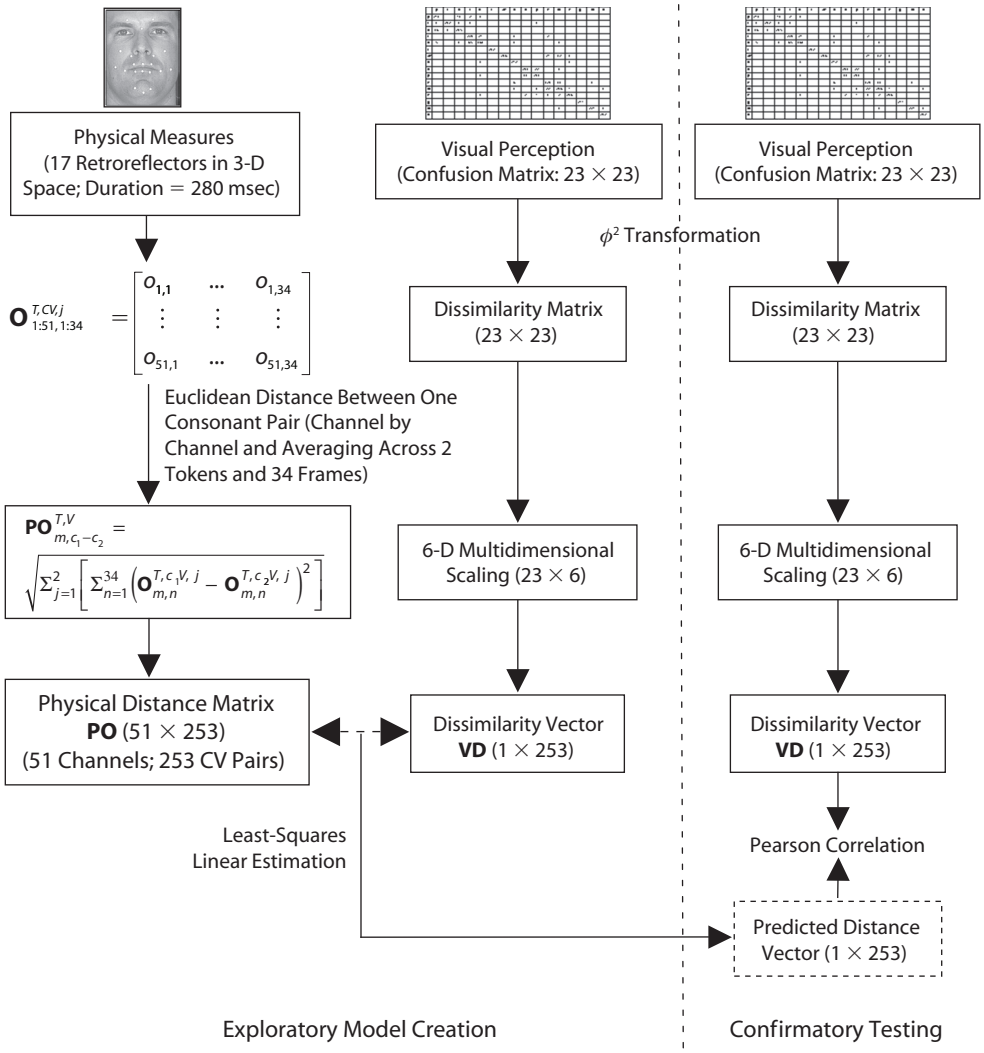


Figure 3. A diagram that shows how the analysis of the relationship between visual consonant perception and physical measures was carried out. The left side, labeled "exploratory model creation," refers to the calculation of least-squares linear estimation weights, and the right side, labeled "confirmatory testing," refers to the evaluation of the second-order isomorphism.

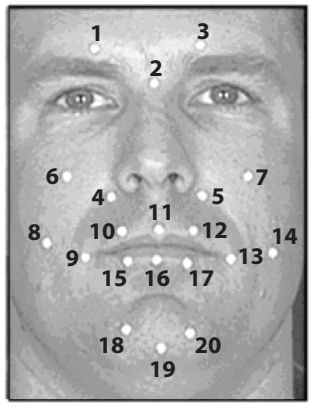
responses to phonemes x and y. $E_{xy}(x_i)$ and $E_{xy}(y_i)$ equal the expected frequencies of responses for x_i and y_i, if phonemes x and y are equivalent. The resulting matrices were symmetric, with distances on the diagonal equal to 0.

Second, the ϕ^2-transformed matrices were submitted to MDS (Kruskal & Wish, 1978; SPSS, 2003). MDS provides mutual geometric constraints among all the stimuli, whereas phi-square coefficients only represent substitutability between stimulus pairs. MDS is an optimization procedure that positions objects in space, so that relative distances between these objects best account for the observed data. The solutions that were chosen for all of the analyses had six dimensions, for which more than 95% of the variance of the ϕ^2-transformed matrices was accounted for. In contrast with many applications of MDS, the goal here was not to reduce dimensionality for the purpose of visualization but to obtain perceptual dissimilarities. Therefore, acceptance of a high-dimensionality set of solutions was appropriate.

Third, Euclidean distances between stimulus pairs in a perceptual space were computed. Two hundred fifty-three stimulus–response Euclidean distance coefficients were obtained for every set of 23 consonants. The notation, $\mathbf{VD}_{T,V}$, represents the coefficients for talker T in vowel context V, which is a 253-component, 1-D vector.

Physical stimulus measures. Physical measures were obtained using 3-D optical motion data. Retroreflector positions were chosen in an attempt to record the speech motion comprehensively. It was not possible a priori to guarantee that for each talker, the number and location of the retroreflectors were optimal for representing second-order isomorphism. The number and density of the retroreflectors were limited by the accuracy of the optical recording system to resolve and consistently track neighboring retroreflectors. Figure 4 shows the placement of the 20 retroreflectors on one of the talkers. The retroreflector locations were the bridge of the nose (1), eyebrows (2), lip contour (8), chin (3), and cheeks (6). Retroreflectors 1, 2, and 3 were used for head motion compensation only and were not used in the analyses (see below). The 17 retroreflectors were also subdivided into sets for the lip (9, 10, 11, 12, 13, 15, 16, and 17), chin (18, 19, and 20), and cheeks (4, 5, 6, 7, 8, and 14).

The motion capture system recorded 2-D coordinates of the retroreflectors, and the 2-D recordings were used to reconstruct 3-D motion at a sampling rate of 120 Hz. A DAT recording was made



Qualisys Retroreflectors:

1. Brow left	2. Nose bridge
3. Brow right	4. Nose left
5. Nose right	6. Cheek left
7. Cheek right	8. Middle left face
9. Middle left center	10. Upper lip left
11. Upper lip center	12. Upper lip right
13. Middle right center	14. Middle right face
15. Lower lip left	16. Lower lip center
17. Lower lip right	18. Chin left
19. Chin center	20. Chin right

Figure 4. Placement of optical face retroreflectors.

simultaneously with the video and motion capture recordings, and all data types were synchronized (Jiang, Alwan, Keating, Auer, & Bernstein, 2002). The DAT recordings were used in selecting the starting and ending points for the optical signal analyses.

Despite the instructions to sit quietly and focus on the teleprompter, talkers made small head movements that had to be removed from the data before quantifying movements during speech; otherwise, the overall head movements would sum with the speech movements. Also, very occasionally, the optical recording system failed to capture a data point. Methods described in Jiang et al. (2002) were used to remove nonspeech movements and replace occasional missing data.

Physical data analyses. Analyses of 3-D motion data were restricted to the initial part of the CV syllables, during the consonant onset, transition, and the initial part of the vowel. The timing relationship between syllable acoustic onset and speech motion varied across different consonants. Nevertheless, a fixed starting point for analyses was set, following visual examination of the acoustic spectra and visual trajectories. The analysis window easily captured the important consonant information in each syllable. The acoustic syllable onset was identified, and then, because speech movements often (not always) were initiated prior to acoustic signal onsets, the beginning point for optical signal analysis was set 30 msec prior to the acoustic onset. Analyses were then applied for 280 msec. At 120 frames/sec, the 280-msec analysis window was equivalent to 34 optical frames.

The data for each stimulus token were organized into a matrix $\mathbf{O}^{T,CV,j}$,

$$\mathbf{O}^{T,CV,j}_{1:51,1:34} = \begin{bmatrix} o_{1,1} & \cdots & o_{1,34} \\ \vdots & \vdots & \vdots \\ o_{51,1} & \cdots & o_{51,34} \end{bmatrix}, \tag{2}$$

where T is the talker, CV is the syllable, and j is the token number. For example, $\mathbf{O}^{\mathrm{M1,ba,1}}$ represented data for the first token of syllable /bɑ/ for Talker M1. Each matrix had 34 columns corresponding to the 34 motion capture frames and 51 rows corresponding to the 51 optical channels (17 retroreflectors in a 3-D space).

The physical Euclidean distance between a pair of consonants (c_1, c_2) represented by two tokens, with vowel context V for talker T, was computed as follows:

$$\mathbf{PO}^{T,V}_{m,c_1-c_2} = \left\| \begin{bmatrix} \mathbf{O}^{T,c_1V,1}_{m} \\ \mathbf{O}^{T,c_1V,2}_{m} \end{bmatrix} - \begin{bmatrix} \mathbf{O}^{T,c_2V,1}_{m} \\ \mathbf{O}^{T,c_2V,2}_{m} \end{bmatrix} \right\| = \sqrt{\sum\nolimits_{j=1}^{2} \left[\sum\nolimits_{n=1}^{34} \left(\mathbf{O}^{T,c_1V,j}_{m,n} - \mathbf{O}^{T,c_2V,j}_{m,n} \right)^2 \right]}, \tag{3}$$

where m is the optical channel number (1–51), n is the frame number (1–34), and j is the token number (1–2). The Euclidean distances among the 23 consonants in a vowel context V for talker T resulted in a 51 (row) × 253 (column) matrix, $\mathbf{PO}^{T,V}$, where the rows were optical channels and the columns were consonant pairs. The distance between consonants in a pair for each channel m was calculated first frame by frame, and then a Euclidean distance was taken across the 34 frames. Therefore, both dynamical and geometric information was captured to some degree in the distance measures, but information was reduced by combining measures across frames.

Three subsets of distances were also derived on the basis of the subsets of data for the lip, cheek, and chin retroreflectors. The distance calculations, thus, resulted in estimates of the dissimilarities as a function of talkers, vowels, and subregions of the talkers' faces. These measures are referred to as $\mathbf{PO}^{T,V}$ (51 × 253, 17 retroreflectors on the face), $\mathbf{PO}^{T,V}_{\mathrm{lips}}$ (24 × 253, 8 retroreflectors on the lips), $\mathbf{PO}^{T,V}_{\mathrm{cheeks}}$ (18 × 253, 6 retroreflectors on the cheeks), and $\mathbf{PO}^{T,V}_{\mathrm{chin}}$ (9 × 253, 3 retroreflectors on the chin).

Analyses applied to optical and perceptual distances. Euclidean optical distances were used to investigate the simple relationship between the raw physical and the perceptual distances. That is, the matrix $\mathbf{PO}^{T,V}$ (51 × 253) was converted to a (1 × 253) vector by computing Euclidean distances directly from the 51-channel distances. Then this vector was correlated with the corresponding vector of perceptual dissimilarities.

Perceptual weighting was also computed using least-squares linear estimation and applied to the $\mathbf{PO}^{T,V}$ matrix. Least-squares linear estimation was used to fit a linear combination of the components of a multichannel signal **PO** and a constant vector **c** to a single-channel signal **VD**. This minimization problem,

$$\mathbf{weights} = \arg\min \left\{ \left\| \begin{bmatrix} \mathbf{c} \\ \mathbf{PO} \end{bmatrix}^T \cdot \mathbf{weights} - \mathbf{VD}^T \right\|_2 \right\}, \tag{4}$$

has a standard solution (Kailath et al., 2000), which is

$$\mathbf{weights} = \left(\begin{bmatrix} \mathbf{c} \\ \mathbf{PO} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{c} \\ \mathbf{PO} \end{bmatrix}^T \right)^{-1} \cdot \begin{bmatrix} \mathbf{c} \\ \mathbf{PO} \end{bmatrix} \cdot \mathbf{VD}^T. \tag{5}$$

The obtained least-squares linear estimation solution provides a set of optimal weights (scalars) for transforming the physical measures (51 dimensions—i.e., 3 dimensions of 17 retroreflectors) into the same dimensionality as that for the perceptual measures by weighting and summation. The procedure yielded the optimal weights for a $\mathbf{PO}^{T,V}$ and $\mathbf{VD}_{T,V}$ pair.

Each participant had viewed 10 lists of stimuli. These 10 lists were randomly divided into two 5-list sets that were balanced in terms

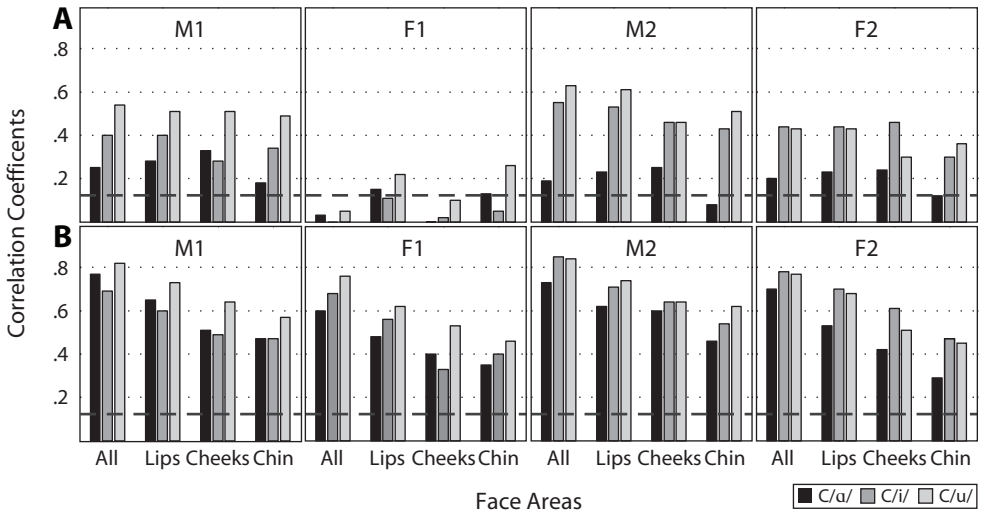


Figure 5. Pearson correlations between physical and perceptual distances for pairs of stimuli. Physical distances were computed without weights (A) and with weights (B). Correlations are listed in terms of talkers (M1, F1, M2, and F2), face areas (lips, cheeks, and chin), and vowel context (C/ɑ/, C/i/, and C/u/). Correlations above the dashed lines are significant ($ps < .05$).

of talkers and participants. Only one set was used to compute the weights on the physical data. That is, one set of perceptual data was used to calculate the weights for physical data. Then those weights were applied to the physical data to obtain the warped physical similarities. The other set of perceptual data was used for evaluating the second-order isomorphism relationship. Pearson correlation coefficients were used to estimate the variance accounted for by the relationship between the perceptual and the physical dissimilarities.

Results and Discussion

Pooled data analyses. Figure 5 shows the Pearson correlations that were obtained between the perceptual distances and the physical distances. The correlations in Figure 5A used unweighted physical distances and were mostly below .65. Correlations for Talker F1 were below .30. That is, the variance accounted for (i.e., the square of the correlation) was modest at best. In addition, the correlations using data from the full face were not consistently higher than those using a subset of the data.

Correlations were higher when the weighted physical distances were used, as is shown in Figure 5B and confirmed by a paired t test [$t(47) = 13.45, p = .000$]. Correlations exceeded .60 when all of the face points were used. Across talkers and vowels, the variance accounted for was 56%. Partitioning of the data into subsets based on face parts resulted in reduced correlations. The lips accounted overall for more variance (41%) than did the cheeks (28%) or the chin (21%).

The mean Pearson correlations across vowels with the complete set of face points were .76 for M1, .68 for F1, .81 for M2, and .75 for F2 ($p < .001$). When the data were subdivided as a function of vowel, the mean correlations were .70 for C/ɑ/, .75 for C/i/, and .80 for C/u/ ($p < .001$). When subdivided into face parts, the mean correlations were .75 for the whole face, .64 for lips only, .53 for chin only, and .46 for cheek areas only ($p < .001$).

In order to estimate the relative perceptual importance of the face parts, the weighted physical dissimilarity vectors were calculated separately for the lips, cheeks, and chin. The sets were used in a multiple regression analysis to predict perceptual dissimilarities. The results in Figure 6 show that individual talkers were associated with individual differences in the contributions of the various parts of the face. The standardized betas for the lips were the largest and were consistently significant. The chin betas were not significant, except for Talker M2's C/u/ syllables. The standardized betas for the cheeks were significant, except for Talker M1. Talker M2 demonstrated the highest R^2, in addition to the sole example of a reliable beta for the chin. This talker also yielded the relatively high perceptual scores in Experiment 1. Although the chin accounted for 21% of the variance (see above), the multiple regression results suggest that it provides primarily redundant information.

Individual-participant analyses. A potential pitfall in using MDS is that averaging data, particularly across participants, can fundamentally change the structure of the confusion matrices (Ashby, Maddox, & Lee, 1994). In order to show that the results were not dependent on pooling data across individuals, the group analysis procedures used to obtain perceptual distances were replicated using individual-participant data. Analyses were then conducted on the full set of face points. The individuals' Pearson correlation coefficients (see Table 2) were generally of the same magnitude as the group's Pearson correlation coefficients in Figure 5B. Figure 7 shows example scatterplots of perceptual distances versus weighted physical distances with C/ɑ/ from Talker F1 and C/u/ from Talker M2 for the most (P1) and least (P6) accurate lip-readers in Experiment 1. Figure 7 shows that the perceptual distances were distributed across the perceptual scale but that many were clustered

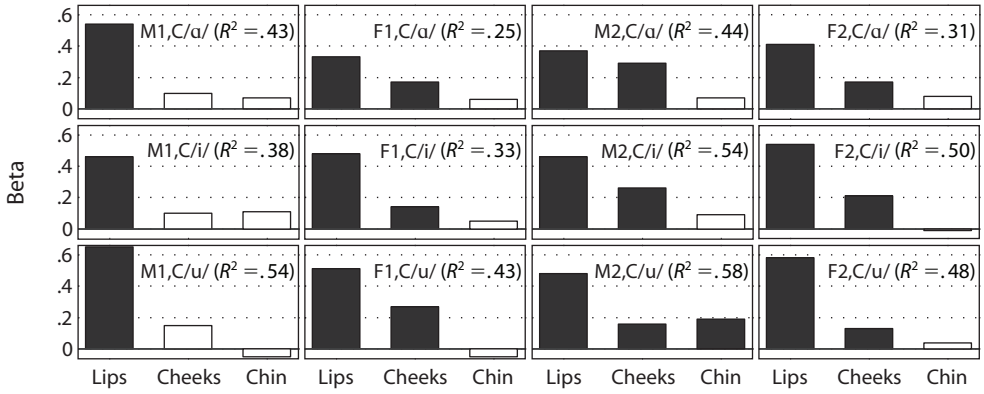


Figure 6. Standardized beta values obtained when predicting visual perceptual distances from the estimated visual distances from lips, cheeks, and chin using the 3-D motion data. Talker and vowel identities and the R^2 statistic are listed inside each panel. Filled bins indicate significant beta values ($ps < .05$).

at higher distances. The figure also shows that the weighting method tended to result in greater estimated distances for stimulus pairs with small obtained perceptual distances. That is, the method tended to amplify (or failed to reduce) distances that were perceptually closer together.

An omnibus repeated measures ANOVA with talker (4) and vowel (3) as the repeated factors was carried out using the individual-participant Pearson correlation coefficients. Overall, the results were consistent with the results from pooled data analyses. The effects of talker [$F(3,3) = 13.72$, $p = .029$] and vowel [$F(2,4) = 33.88$, $p = .003$] were significant. However, the talker × vowel interaction was also significant [$F(6,30) = 11.35$, $p = .000$]. Table 3 shows the mean Pearson correlation coefficients for the main effects of talker and vowel. The table shows that the correlations for C/u/ varied less across talkers than did those for C/ɑ/ and C/i/. The significant interaction is due partly to the interaction involving C/i/ versus C/u/ and Talkers M1 versus F2 [$F(1,5) = 26.41$, $p = .004$] and partly to the interaction involving C/ɑ/ versus C/i/ and Talkers M1 versus F2 [$F(1,5) = 114.59$, $p = .000$]. An explanation for these interactions is not obvious.

Table 2 shows that the Pearson correlations for P1 were smaller than those for P6 [paired $t(11) = 4.97$, $p = .000$], yet P1 was the most accurate lip-reader and P6 the least accurate. This result could be due to the sparse 3-D representation of perceptually relevant physical data. For example, the retroreflectors were pasted near the outer lip margins, but not on the inner margins as well. Nor were there measurements of tongue movement. Lip-readers report using glimpses of the tongue inside the mouth. The inner lip margins likely provide somewhat different information than do the outer lip margins (Montgomery & Jackson, 1983; Plant, 1980). If the retroreflector data had included the information used by the most accurate lip-reader, the correlations for that perceiver would be highest, as was expected. That is, the most accurate lip-reader might use information measurable with inner lip margin markers.

Table 2
Pearson Correlations Between Physical and Perceptual Dissimilarities for Pairs of Stimuli for Participants P1 and P6

	Participant P1			Participant P6		
Talker	ɑ	i	u	ɑ	i	u
M1	.66	.59	.73	.75	.70	.81
F1	.52	.61	.67	.51	.65	.70
M2	.63	.76	.76	.79	.81	.83
F2	.61	.70	.61	.64	.72	.76

Note—Physical dissimilarities were weighted distances of all face points. Correlations are listed in terms of talkers (M1, F1, M2, and F2) and vowel context (C/ɑ/, C/i/, and C/u/).

When the optical–perceptual correlations (using all face retroreflectors) in Figure 5B were averaged across vowels, they were somewhat related to the four talkers' visual sentence intelligibilities. That is, Talker M2 had relatively high visual sentence intelligibility and the highest optical–perceptual correlation ($r = .81$); Talker F1 had the lowest visual sentence intelligibility and the lowest optical–perceptual correlation ($r = .68$); Talker M1, with intermediate visual sentence intelligibility, yielded intermediate optical–perceptual correlations; however, Talker F2, with the highest visual sentence intelligibility, yielded intermediate optical–perceptual correlations.

By informally examining the video, we saw that Talker M2 had extensive movements, whereas Talker F1 did not move her cheeks and chin much. This was confirmed with a quantitative examination of mouth area/opening, which was approximated using the eight lip retroreflectors across 4,692 frames (23 consonants × 3 vowels × 2 tokens × 34 frames) for Talkers F1 and M2. The variances in mouth area/opening were 2.40 cm^4 for F1 and 4.59 cm^4 for M2. The talker difference in face movements could affect the internal structure of the second-order isomorphism (see Figure 7). For example, for Talker F1, the prediction error in the least-squares linear estimation of perceptual dissimilarities was large for the /du/–/gu/ pair (1.05; back place of articulation) but low for the /wu/–/pu/ pair (.10; front place of articulation); and for Talker M2, the prediction error was low for the /du/–/gu/ pair (.07) but large for the /wu/–/pu/ pair (.44).

In our previous study (Jiang et al., 2002), in which predictions were made across acoustic and optical signals

Table 3
Mean Pearson Correlations and the Corresponding Standard Errors of Means in Terms of Talkers (M1, F1, M2, and F2) and Vowel Context (C/ɑ/, C/i/, and C/u/) From Individual-Participant Analyses

	ɑ		i		u	
Talker	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
M1	.70	.02	.63	.02	.75	.02
F1	.53	.02	.61	.02	.65	.02
M2	.66	.03	.77	.01	.76	.02
F2	.64	.01	.73	.01	.69	.02

Note—Physical dissimilarities were weighted distances of all face points.

for these talkers, speech acoustics were most accurately predicted from face movements for Talker M2 and least accurately predicted for Talker F1. Converging evidence from this and the prior study suggests that Talker F1 provided the least linguistically relevant information to the lip-readers and to the motion capture system.

Overall, intelligibility is not straightforwardly related to the strength of the association between the physical and the perceptual measures. Another example of the complexity among these relationships arises between the optical–perceptual correlations in Figure 5B when averaged across talkers versus lipreading scores in Experiment 1. For example, the C/u/ syllables resulted in relatively low perceptual identification scores (.311) but the highest second-order association ($r = .80$). This might be because with C/ɑ/ syllables, the tongue was visible to lip-readers, but the motion capture system did not record the tongue motion. Thus, the linguistically relevant speech information is likely too sparsely sampled by the 3-D optical data in this study, and this could be improved by also representing data points on the tongue (Jiang et al., 2002). Although we obtained such data, the apparatus to record tongue movement interferes with viewing the talker, and therefore the video taken at the same time as the magnetometer recordings was not used here for the stimuli.

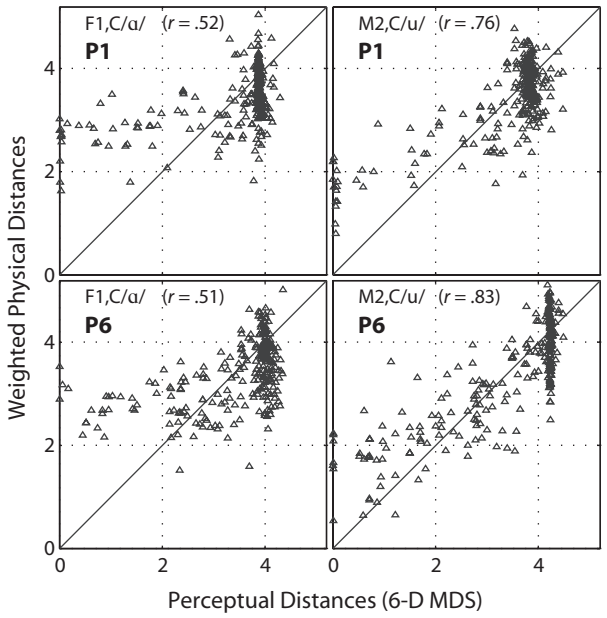


Figure 7. Example scatterplots of weighted physical distances versus perceptual distances (modeled with 6-D multidimensional scaling; *x*-axis) for Participants P1 and P6 with C/ɑ/ and C/u/ from Talkers F1 and M2, respectively. The correlation (*r*) is also displayed.

GENERAL DISCUSSION

A main empirical goal of this study was to demonstrate the degree of second-order isomorphism between physical speech signals and visual speech perception. Perceptual identifications of nonsense syllables were transformed into perceptual dissimilarities between pairs of stimuli. The 3-D optical data were also transformed into dissimilarities that were linearly weighted to correspond optimally with perceptual dissimilarities. The methods resulted in a robust demonstration of second-order isomorphism, both for perceptual data pooled across individual participants and for data at the level of individual participants. When the data were pooled across participants, the variance accounted for in the physical–perceptual relationship ranged between approximately 36% and 72% across talkers and vowels, when the data comprised all of the face points. The variance accounted for was slightly lower when the individual-participant data were used separately to estimate dissimilarities. Overall, the method was successful across talkers, vowels, and perceivers. The success of the method is considered impressive, given the number of factors that contributed variation and the degree to which the 3-D optical data are a sparse representation of the natural visual stimuli.

The demonstration of a second-order isomorphism relationship without recourse to a feature extraction stage in the processing of the physical signals suggests that visual phonetic distinctions could be represented during perceptual processing without explicit extraction of features such as mouth shape, jaw position, and so forth. Alternatively, the physical distinctiveness that was demonstrated among the signals could be the basis for a perceptual feature representation of visible speech. Phonological features are theorized to take advantage of similarities and differences that are output by a perceptual system (Bernstein, 2006; Stevens, 1998; Stevens, Manuel, Shattuck-Hufnagel, & Liu, 1992). Of course, the question of whether there is a feature level of visual phonetic perceptual processing is an empirical one and beyond the scope of this study. However, the advantage of having established a second-order isomorphism is that we do not need to know what the representation is in order to investigate some of its implications. Furthermore, the advantage of representing visual speech stimuli with such a high degree of reduction (e.g., the 3-D) in the data leads to the possibility of systematically applying transformations to the data, analogous perhaps to the use of the acoustic speech synthesizer for research in acoustic phonetics. In our current studies, we are using stimuli at predicted perceptual distances and obtaining perceptual and neural measures as a function of the distances. In other work, we are using the 3-D face points to drive a synthetic talking face, thus achieving more complete stimulus control.

Additional modeling approaches can be applied to our data. For example, Shepard (1965) suggested that confusions tend to decay exponentially with increasing interstimulus distances and that improvements in accounting for variance could take place by transformation into a warped perceptual space. A desirable approach could be to use a psychophysically motivated transformation on the perceptual data (Nosofsky, 1985). As a preliminary investigation, we further processed the ϕ^2-transformed perceptual dissimilarities in this study, so that

$$d = \sqrt{-\log\left(1 - \phi^2\right)} \quad (6)$$

prior to the MDS transformation (Nosofsky, 1985). Then the same methods as were used in Experiment 2 were applied to evaluate the physical–perceptual relationships. Table 4 shows the results with and without the nonlinear transformation (Equation 6; see above). The mean difference in the Pearson correlation coefficients was .035, which was a small but nevertheless significant difference [paired $t(11) = 2.77$, $p = .018$]. Future studies might compare alternative psychophysical or neurobiologically motivated similarity models for their ability to improve the correspondence between physical and perceptual similarities.

Limitations of the Present Study

In this study, the participants did not make similarity judgments. Perceptual similarity structure was derived. The forced choice perceptual identification was a metalinguistic task, whereas the physical dissimilarities/similarities were based on optical measurements. A valid question is, what type of stimulus processing could have resulted in the physical dissimilarities being shown to be unrelated to the perceptual similarities? This could arise if visual speech perception engaged, for example, highly nonlinear processes that transformed information into abstract feature representations. The results here seem to argue against that possibility, suggesting that phonetic information in visual stimuli is, to a first approximation, well represented by linearly related dissimilarities.

Several potential sources of visual information were not represented in the physical measures in the present study, as was previously discussed. Several studies have shown the importance of the oral cavity (Smeele, Hahnlen, Stevens, Kuhl, & Meltzoff, 1995) and being able to see the teeth (McGrath, 1985; Summerfield, MacLeod, McGrath, & Brooke, 1989).

Table 4
Pearson Correlations Obtained in Experiment 2 and Those Obtained Using the Same Methods Except That the Phi-Square Dissimilarities Were Submitted to a Nonlinear Transformation Prior to Multidimensional Scaling

	Experiment 2			Nonlinear Transformation		
Talker	ɑ	i	u	ɑ	i	u
M1	.77	.69	.82	.77	.77	.83
F1	.60	.68	.76	.70	.71	.72
M2	.73	.85	.84	.83	.86	.83
F2	.70	.78	.77	.75	.83	.81

Note—Physical dissimilarities were weighted distances of all face points.

The dynamic information in the physical data was not analyzed separately from the configural information. Both types of information were implicit in the distance measures, because distance on a particular face point and dimension increases whenever movement over time differs across syllables. For example, the timing difference in bilabial release between /ba/ and /wa/ produces a large physical distance. Dissimilarities between syllables could be computed in terms of differences in velocities and/or accelerations. The extent to which kinematic information alone is sufficient for visual speech perception is an ongoing topic (Campbell, 1996; Rosenblum & Saldaña, 1998). The methods of the present study could be extended to determine the extent to which perceptual dissimilarity is accounted for by kinematic dissimilarity.

Individual perceiver and talker differences were obtained in this study. A hypothesis is that individual perceivers differ in their sensitivity to the stimulus information that supports differentiating among consonants. Interestingly, in the individual-participant analyses, higher correlations were obtained for the least proficient lip-reader (P6) relative to the most proficient lip-reader (P1), a seeming contradiction to that hypothesis. However, an explanation could be that the sparse physical data were more adequate to represent the information perceived by P6 than that perceived by P1, who might have relied more on tongue movement and inner lip borders, information that was not captured by the 3-D optical recordings.

A hypothesis about individual-talker differences is that they vary in producing phonetic differences. Talker F1 was least intelligible. Physical dissimilarities in her data produced the lowest correlations with perceptual dissimilarities. This result would be expected whenever the range of dissimilarities is compressed in the data, which Table 2 suggests was the case for Talker F1.

Conclusions

Phonetic categories are sometimes considered mental concepts with arbitrary relationships to physical stimulus properties. The physical measures on 3-D optical recordings in this study represent a drastic reduction in stimulus data, relative to full video, yet a high level of correlation was demonstrated between perceptual and physical dissimilarity without recourse to phonetic categories. This suggests that visual phonetic perception is mediated by a nonarbitrary relationship among the physical stimuli. Furthermore, it suggests that manipulations and transformations on sparse 3-D optical data are relevant to acquiring further understanding of visual phonetic perception.

This study shows that phonetic distinctiveness can be quantified with perceptual and physical measures that show a high level of second-order isomorphism with each other. We think that these results support the view that physical signals that were measured are linguistically relevant and, therefore, are relevant to the underlying neural processes that support speech perception. Understanding and control of the physical optical speech signals via 3-D

motion representations will facilitate future studies of perception and neural processing in terms of first- and second-order isomorphism.

AUTHOR NOTE

Some of the results were presented at EUROSPEECH (Aalborg, Denmark, September 2001), AVSP (Scheelsminde, Denmark, September 2001), and ICASSP (Orlando, Florida, May 2002). This research was supported by Awards NSF IIS 9996088, IIS 0312434, and NIH/NIDCD DC006035 (L.E.B., PI). We acknowledge the help of Brian Chaney, Patrick Barjam, Jennifer Yarbrough, Sven Mattys, Sumiko Takayanagi, and Taehong Cho in carrying out the study. The authors also thank Laurel Fisher for helpful discussions and José Benkí and two anonymous reviewers for their comments on the manuscript. Correspondence concerning this article should be addressed to J. Jiang, Department of Communication Neuroscience, House Ear Institute, 2100 West Third Street, Los Angeles, CA 90057 (e-mail: jjiang@hei.org).

REFERENCES

ANDERSSON, U., & LIDESTAM, B. (2005). Bottom-up driven speechreading in a speechreading expert: The case of AA (JK023). *Ear & Hearing*, **26**, 214-224.

ARNOLD, P., & HILL, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, **92**, 339-355.

ASHBY, F. G., MADDOX, W. T., & LEE, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, **5**, 144-151.

AUER, E. T., JR. (2002). The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness. *Psychonomic Bulletin & Review*, **9**, 341-347.

AUER, E. T., JR., & 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.

AUER, E. T., JR., & BERNSTEIN, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech, Language, & Hearing Research*, **50**, 1-9.

AUER, E. T., JR., BERNSTEIN, L. E., & TUCKER, P. E. (2000). Is subjective word familiarity a meter of ambient language? A natural experiment on effects of perceptual experience. *Memory & Cognition*, **28**, 789-797.

BENOÎT, C., GUIARD-MARIGNY, T., LE GOFF, B., & ADJOUDANI, A. (1996). Which components of the face do humans and machines best speechread? In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (pp. 315-328). Berlin: Springer.

BERNSTEIN, L. E. (2006). Visual speech perception. In E. Vatikiotis-Bateson, G. Bailly, & P. Perrier (Eds.), *Audio–visual speech processing*. Cambridge, MA: MIT Press.

BERNSTEIN, L. E., AUER, E. T., JR., & TAKAYANAGI, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, **44**, 5-18.

BERNSTEIN, L. E., AUER, E. T., JR., & TUCKER, P. E. (2001). Enhanced speechreading in deaf adults: Can short-term training/practice close the gap for hearing adults? *Journal of Speech, Language, & Hearing Research*, **44**, 5-18.

BERNSTEIN, L. E., DEMOREST, M. E., & TUCKER, P. E. (1998). What makes a good speechreader? First you have to find one. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 211-228). Hove, U.K.: Psychology Press.

BERNSTEIN, L. E., DEMOREST, M. E., & TUCKER, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, **62**, 233-252.

BREEUWER, M., & PLOMP, R. (1986). Speechreading supplemented with auditorily presented speech parameters. *Journal of the Acoustical Society of America*, **79**, 481-499.

CAMPBELL, R. (1996, October). *Seeing speech in space and time: Psychological and neurological findings*. Paper presented at the ICSLP 1996, Philadelphia.

DE GELDER, B., & BERTELSON, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences*, **7**, 460-467.

DEMOREST, M. E., & BERNSTEIN, L. E. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech & Hearing Research*, **35**, 876-891.

DODD, B., MCINTOSH, B., & WOODHOUSE, L. (1998). Early lipreading ability and speech and language development of hearing-impaired pre-schoolers. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 229-242). Hove, U.K.: Psychology Press.

EDELMAN, S. (1998). Representation is representation of similarities. *Behavioral & Brain Sciences*, **21**, 449-498.

ERBER, N. P. (1975). Auditory–visual perception of speech. *Journal of Speech & Hearing Disorders*, **40**, 481-492.

ERBER, N. P. (1979). Auditory–visual perception of speech with reduced optical clarity. *Journal of Speech & Hearing Research*, **22**, 212-223.

FISHER, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech & Hearing Research*, **11**, 796-804.

FRISCH, S. A., PIERREHUMBERT, J. B., & BROE, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, **22**, 179-228.

GOLDSTONE, R. L. (1999). Similarity. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 763-765). Cambridge, MA: MIT Press.

GRANT, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America*, **109**, 2272-2275.

GRANT, K. W., & SEITZ, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, **108**, 1197-1208.

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.

GRANT, K. W., WALDEN, B. E., & SEITZ, P. F. (1998). Auditory–visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory–visual integration. *Journal of the Acoustical Society of America*, **103**, 2677-2690.

IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio & Electroacoustics*, **17**, 225-246.

IJSSELDIJK, F. J. (1992). Speechreading performance under different conditions of video image, repetition, and speech rate. *Journal of Speech & Hearing Research*, **35**, 466-471.

IVERSON, P., BERNSTEIN, L. E., & AUER, E. T., JR. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Communication*, **26**, 45-63.

JACKSON, P. L., MONTGOMERY, A. A., & BINNIE, C. A. (1976). Perceptual dimensions underlying vowel lipreading performance. *Journal of Speech & Hearing Research*, **19**, 796-812.

JIANG, J., ALWAN, A., KEATING, P., AUER, E. T., JR., & BERNSTEIN, L. E. (2002). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Applied Signal Processing*, **2002**, 1174-1188.

JORDAN, T. R., & THOMAS, S. M. (2001). Effects of horizontal viewing angle on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **27**, 1386-1403.

KAILATH, T., SAYED, A. H., & HASSIBI, B. (2000). *Linear estimation*. Upper Saddle River, NJ: Prentice Hall.

KISHON-RABIN, L., BOOTHROYD, A., & HANIN, L. (1996). Speechreading enhancement: A comparison of spatial–tactile display of voice fundamental frequency (F_0) with auditory F_0. *Journal of the Acoustical Society of America*, **100**, 593-602.

KLATT, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, **82**, 737-793.

KRICOS, P. B., & LESNER, S. A. (1982). Differences in visual intelligibility across talkers. *Volta Review*, **84**, 219-225.

KRUSKAL, J. B., & WISH, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

MACEACHERN, E. (2000). On the visual distinctiveness of words in the English lexicon. *Journal of Phonetics*, **28**, 367-376.

MACLEOD, A., & SUMMERFIELD, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, **21**, 131-141.

MARASSA, L. K., & LANSING, C. R. (1995). Visual word recognition in two facial motion conditions: Full-face versus lips-plus-mandible. *Journal of Speech & Hearing Research*, **38**, 1387-1394.

MASSARO, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

MATTYS, S. L., BERNSTEIN, L. E., & AUER, E. T., JR. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics*, **64**, 667-679.

MCCOTTER, M. V., & JORDAN, T. R. (2003). The role of facial colour and luminance in visual and audiovisual speech perception. *Perception*, **32**, 921-936.

MCGRATH, M. (1985). *An examination of cues for visual and audio–visual speech perception using natural and computer-generated faces*. Unpublished thesis, University of Nottingham.

MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

MOHAMMED, T., CAMPBELL, R., MACSWEENEY, M., MILNE, E., HANSEN, P., & COLEMAN, M. (2005). Speechreading skill and visual movement sensitivity are related in deaf speechreaders. *Perception*, **34**, 205-216.

MONTGOMERY, A. A., & JACKSON, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America*, **73**, 2134-2144.

MOODY-ANTONIO, S., TAKAYANAGI, S., MASUDA, A., AUER, E. T., JR., FISHER, L., & BERNSTEIN, L. E. (2005). Improved speech perception in adult congenitally deafened cochlear implant recipients. *Otology & Neurotology*, **26**, 649-654.

MUNHALL, K. G., KROOS, C., JOZAN, G., & VATIKIOTIS-BATESON, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics*, **66**, 574-583.

MUNHALL, K. G., & VATIKIOTIS-BATESON, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 123-139). Hove, U.K.: Psychology Press.

NOSOFSKY, R. [M.] (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics*, **38**, 415-432.

NOSOFSKY, R. M., & STANTON, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception & Performance*, **31**, 608-629.

OWENS, E., & BLAZEK, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech & Hearing Research*, **28**, 381-393.

PLANT, G. L. (1980). Visual identification of Australian vowels and diphthongs. *Australian Journal of Audiology*, **2**, 83-91.

REISBERG, D., MCLEAN, J., & GOLDFIELD, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-113). London: Erlbaum.

REPP, B. H. (1981). On levels of description in speech research. *Journal of the Acoustical Society of America*, **69**, 1462-1464.

RÖNNBERG, J. (1995). Perceptual compensation in the deaf and blind: Myth or reality? In R. A. Dixon & L. Bäckman (Eds.), *Compensating for psychological deficits and declines: Managing losses and promoting gains* (pp. 251-274). Mahwah, NJ: Erlbaum.

RÖNNBERG, J., SAMUELSSON, S., & LYXELL, B. (1998). Conceptual constraints in sentence-based lipreading in the hearing impaired. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 143-153). Hove, U.K.: Psychology Press.

ROSENBLUM, L. D., & SALDAÑA, H. M. (1998). Time-varying information for visual speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 61-81). Hove, U.K.: Psychology Press.

SCHEINBERG, J. S. (1980). Analysis of speechreading cues using an interleaved technique. *Journal of Communication Disorders*, **13**, 489-492.

SHEPARD, R. N. (1965). Approximation to uniform gradients of generalization by monotone transformations of scale. In D. I. Mostofsky (Ed.), *Stimulus generalization* (pp. 94-110). Stanford, CA: Stanford University Press.

SHEPARD, R. N., & CHIPMAN, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, **1**, 1-17.

SIEGEL, S., & CASTELLAN, N. J., JR. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

SMEELE, P. M. T., HAHNLEN, L. D., STEVENS, E. B., KUHL, P. K., & MELTZOFF, A. N. (1995). Investigating the role of specific facial information in audio–visual speech perception [Abstract]. *Journal of the Acoustical Society of America*, **98**, 2983.

SPSS (2003). *SPSS base 12.0 user's guide*. Chicago: Author.

STEVENS, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

STEVENS, K. N., MANUEL, S. Y., SHATTUCK-HUFNAGEL, S., & LIU, S. (1992, October). *Implementation of a model for lexical access based on features*. Paper presented at the ICSLP 1992, Banff, Canada.

SUMBY, W. H., & POLLACK, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215.

SUMMERFIELD, Q., MACLEOD, A., MCGRATH, M., & BROOKE, M. (1989). Lips, teeth, and the benefits of lipreading. In A. W. Young & H. D. Ellis (Eds.), *Handbook of research on face processing* (pp. 223-233). Amsterdam: North-Holland.

THOMAS, S. M., & JORDAN, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **30**, 873-888.

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 & Hearing Research*, **20**, 130-145.

(Manuscript received February 16, 2005;
revision accepted for publication February 15, 2007.)