Published in final edited form as:

J Exp Psychol Hum Percept Perform. 2004 April ; 30(2): 378–396. doi:10.1037/0096-1523.30.2.378.

Cross-Modal Source Information and Spoken Word Recognition

Lorin Lachs and

Department of Psychology, California State University, Fresno

David B. Pisoni

Speech Research Laboratory, Department of Psychology, Indiana University Bloomington, and
DeVault Otologic Research Laboratory, Department of Otolaryngology, Head and Neck Surgery,
Indiana University School of Medicine

Abstract

In a cross-modal matching task, participants were asked to match visual and auditory displays of speech based on the identity of the speaker. The present investigation used this task with acoustically transformed speech to examine the properties of sound that can convey cross-modal information. Word recognition performance was also measured under the same transformations. The authors found that cross-modal matching was only possible under transformations that preserved the relative spectral and temporal patterns of formant frequencies. In addition, cross-modal matching was only possible under the same conditions that yielded robust word recognition performance. The results are consistent with the hypothesis that acoustic and optical displays of speech simultaneously carry articulatory information about both the underlying linguistic message and indexical properties of the talker.

Visual information about the articulation of spoken words has been shown to produce large effects on speech perception. Conflicting information about speech in the visual and auditory modalities can lead to the illusory perception of speech sounds not displayed in either modality alone, a phenomenon known as the “McGurk effect” (McGurk & MacDonald, 1976). In combination, auditory and visual information about speech can produce improved speech recognition performance (Sumby & Pollack, 1954). Much of the work on audiovisual speech perception has been concerned with the effects of conflicting auditory and visual cues on phonetic identification (e.g., Green, 1996; Green & Gerdeman, 1995; Massaro, 1998). However, a growing body of literature has shown that the speech signal carries important information about the indexical properties of the talker: properties such as gender, dialectal variation, and idiosyncratic speaking style (elements commonly referred to as a talker’s “voice”) are present in the speech waveform (Hirahara & Kato, 1992; Pisoni, 1993). Are these sources of voice information also carried in the visual signal generated by speaking as well?

The acoustic signal that carries linguistic and extra-linguistic or “indexical” information provides a rich and very detailed source of information about the speaker, speaking environment, and the linguistic message. When engaged in the act of speaking, a talker produces a single acoustic pattern that simultaneously conveys both the semantic content of the intended utterance and the unique acoustic elements that define the talker’s voice. In other words, different analyses of the same acoustic signal can yield two distinct sources of information about its production (Pisoni, 1993, 1997). It is not that there are two different

sources of information buried in the signal for each of these sets of attributes. Rather, the two properties of speech—the linguistic and indexical—are carried simultaneously and in parallel by the same acoustic signal. In principle, then, linguistic and indexical information are fundamentally inseparable in their initial acoustic form in the speech waveform.

The inextricable bond between linguistic and indexical information in the speech signal has been shown to affect speech perception and spoken word recognition (Creelman, 1957; Fourcin, 1968; Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Verbrugge, Strange, Shankweiler, & Edman, 1976). It is important to note that this research has revealed that the perceptual aspects of a spoken word that are associated with phonetic information and those attributes that are associated with talker information are not analyzed independently but rather are perceived and processed in a mutually dependent fashion (see Goldinger, 1998; Lachs, Mc-Michael, & Pisoni, 2003, for reviews).

What kind of information about a talker's voice is encoded in the speech signal, and how does that information contribute to speech perception? Bradlow, Torretta, and Pisoni (1996) found that although global characteristics of speech such as fundamental frequency and speaking rate had little effect on speech intelligibility, detailed changes in the acoustic–phonetic properties of a talker's voice, such as the amount of vowel-space reduction and the degree of articulatory precision, turned out to be strong predictors of overall speech intelligibility. Their findings on both speech perception and production suggest that the indexical properties of an utterance may be completely intermixed with the actual phonetic realization of an utterance and that there may be no real physical dissociation between the two sources of information in the speech signal itself.

More direct evidence for the parallel encoding of linguistic and indexical information in the speech signal comes from several studies that have used sine-wave replicas of speech. Sine-wave speech is created by generating independent sinusoidal signals that trace the center frequencies of the three lowest formants in naturally produced utterances (Remez, Rubin, Pisoni, & Carrell, 1981). The resulting pattern sounds perceptually unnatural, but the signal can be perceived by listeners as speech and the original linguistic message can be recovered (Remez, Rubin, Berns, Pardo, & Lang, 1994). Indeed, not only is the linguistic content of the utterance perceptible, but several recent studies have shown that specific aspects of a talker's unique individual identity and speaking style are also preserved in sine-wave replicas of speech. In the first report of this phenomenon, Remez, Fellowes, and Rubin (1997) found that listeners could explicitly identify specific familiar talkers from sine-wave replicas of their utterances. Their results on familiar talker recognition are remarkable because sine-wave speech patterns preserve none of the traditional “speech cues” that were generally thought to support the perception of vocal identity, such as fundamental frequency or the average long-term spectrum (Bricker & Pruzansky, 1976; Hollien & Klepper, 1984). In creating sine-wave speech patterns, an utterance is essentially stripped of all of the redundant acoustic information in an utterance except the time-varying properties of the vocal resonances generated by articulatory motion.

The underlying articulatory basis for spoken language has also been implicated as a source of information in the integration of the auditory and visual forms of speech (Fowler, 1996; Grant, 2001; Rosenblum, 1994). If information about the source of an utterance (i.e., the talker) is available in the acoustic signal because the signal is structured by the underlying vocal articulation and if the relationship between acoustic and optical speech is also available because the patterns in each modality are structured by the same articulation, then it should also be the case that visual indexical information is encoded and available in optic displays of speech as well.

Evidence for the visual transmission of talker-specific articulatory information has been reported in a recent study using point-light displays of speech. Point-light displays eliminate from a visual stimulus all information other than kinematic movement information (see Johansson, 1973). Rosenblum et al. (2002) asked participants to match point-light displays of talkers articulating with one of two fully illuminated speaking faces. Despite the lack of traditional cues to facial identity (e.g., configuration of facial features, color, shading, and shape; see Bruce, 1988), Rosenblum et al. found that participants were able to correctly match the corresponding fully illuminated face with the point-light display, indicating that information for face recognition is specified in movement alone. Rosenblum et al. suggested that face matching between isolated kinematic displays and their fully illuminated counterparts may be due to a common underlying source of information that supports both face recognition and visible speech perception and that indicates that visible articulatory information may incidentally specify talker-specific information that is subsequently used for face recognition. As Rosenblum et al. pointed out, this is exactly analogous to the recent findings reported by Remez et al. that talker-specific information is contained in sine-wave speech replicas. Taken together, the implication of these recent findings is that indexical source information about speech is also modality neutral, just as previous studies of audiovisual speech integration have shown linguistic information to be (Fowler & Galantucci, in press; Rosenblum, in press).

Recently, we investigated the relationship between the acoustic and optical specifications of talker-specific information using a “cross-modal matching task” (Lachs, 2002; Lachs & Pisoni, in press). Figure 1 illustrates the cross-modal matching task in a schematic form. Using an XAB matching paradigm, participants were first presented with a visual-alone, dynamic display of a talker speaking an isolated English word. Subsequently, participants were presented with two auditory-alone response alternatives. One of the alternatives was the acoustic display of the same event that generated the visual test pattern, and one was the acoustic display of a different talker speaking the same word. In another condition, participants were first presented with an acoustic display and then shown two visual-alone alternatives. The results of the study by Lachs and Pisoni showed that participants performed above chance in matching the specification of a spoken event presented to one sensory modality with the specification of the same event presented to another sensory modality. Lachs (2002) concluded, therefore, that information specifying the relationship between acoustic and optic displays of speakers (i.e., cross-modal source information) must be contained in the acoustic and optic displays of speech.

Lachs (2002) also carried out three additional experiments using the same XAB paradigm to investigate the ways in which acoustic and optic displays carry cross-modal source information in speech. In one experiment, visual displays were converted into static pictures of faces and used as optical stimuli in a cross-modal matching experiment with acoustic stimuli. Without the dynamic visual information about articulation, participants were unable to perform the cross-modal matching task above chance.

In the second experiment, acoustic stimuli were converted into noise-band “chimaeric” stimuli (Smith, Delgutte, & Oxenham, 2002). The noise-band transformation eliminates the fundamental frequency, a traditional cue to vocal identity, from the acoustic specification of speech and replaces it with white noise. The resulting signal provides a description of the evolution of the vocal-tract transfer function over time, excited by a white-noise source instead of the normal glottal source. Even without information about f_0 , participants were able to perform the cross-modal matching task above chance, indicating that the availability of f_0 information is not necessary for cross-modal source matching.

In the third experiment, the visual and auditory stimuli were played backward in time. By playing the patterns backward, Lachs (2002) was able to assess the hypothesis that duration cues, which are available acoustically and optically, were used by participants to match the displays across sensory modalities. Performance did not differ from chance with backward displays, indicating that cross-modal matching of source information is not based on the use of overall duration cues, nor is it invariant with respect to temporal structure.

In another study, the same acoustic stimulus materials used in the earlier cross-modal matching experiment were also presented to participants in a word recognition experiment (see Lachs & Pisoni, in press). As expected, word identification was successful only under the same acoustic transformations that preserved cross-modal matching ability. The observed relationship between cross-modal matching and word recognition performance raises several important questions about the auditory form of cross-modal source information. Cross-modal source information and linguistically significant information needed for word identification appear to be encoded together and are carried in parallel in the dynamic pattern of formants as they vary over time.

The cross-modal matching task provides a unique methodology for studying the way an acoustic signal can specify the underlying vocal-tract articulatory events that produced it. By systematically transforming the acoustic signal and then assessing whether performance on the cross-modal matching task is affected by the transformation, it should be possible to describe the ways in which formant structure carries cross-modal source information—and by extension, articulatory information—about speech events. If a given acoustic transformation of the signal disrupts the ability to perform the cross-modal matching task, then it can be assumed that such a transformation eliminates cross-modal source information from the acoustic pattern and that this particular source of information is critical to speech perception and spoken word recognition performance.

The cross-modal matching task as a method for uncovering the form of cross-modal source information was already used with one specific transformation: temporal reversal (Lachs & Pisoni, in press). The present investigation was designed to extend and elaborate this research strategy using several other transformations of the acoustic signal. In addition to studying cross-modal matching, all the acoustic transformations were tested using an additional word identification task to assess whether the transformation preserved the information necessary for word recognition. We predicted that acoustic transformations that preserve cross-modal source information would also preserve the acoustic–phonetic information needed for word recognition, because both types of information make use of the same underlying articulatory behavior of the talker’s vocal tract. Thus, for each of the transformations tested below, two tasks were performed: a cross-modal matching task and a word identification task.

Because the frequency relationships between formants vary over time, manipulations of these attributes of the acoustic signal can be carried out in both the spectral and temporal domains. Each of these domains was manipulated in the current experiment. The possible implications for the form of cross-modal information are outlined below under the description of each transformation.

Spectral Domain Transformations

Several transformations in the frequency domain were used to produce perturbations in the acoustic signal that would selectively preserve or disrupt various aspects of formant patterning. In the *frequency-shifting conditions*, the frequency spectrum of the original speech utterance was shifted up along the frequency axis. For example, Figure 2 shows the spectrogram for one talker (M2) speaking the word *wail*. Figure 3 shows four different

frequency-shifting transformations of the same token. Note that this transformation preserves the absolute spacing between the formants but changes the absolute frequencies themselves. If cross-modal information about speech is specified in terms of the patterning of the formants, irrespective of their absolute frequencies, then performance on the cross-modal matching task should remain unaffected by this acoustic transformation. If the hypothesized relationship between cross-modal source information and word identification information is correct, then word identification performance should not be affected either.

A related frequency domain transformation is *linear frequency scaling*. Under this acoustic transformation, the components of the frequency spectrum were transformed such that any arbitrary frequency distance in the original signal was linearly related to the frequency distance in the transformed signal. Figure 4 shows spectrograms for the two linear frequency-scaling transformations used in our experiments. If cross-modal source information is carried in the relative spacing of the formants as they unfold over time, and not in their absolute relationships, then the ability to perform the cross-modal matching task under linear frequency-scaling conditions should remain unaffected by this transformation because relative frequency-spacing information is preserved.

The third frequency domain transformation that we tested was *nonlinear frequency scaling*. Under this transformation, the components of the frequency spectrum are transformed nonlinearly, such that the frequency distances in the original spectrum were nonlinearly related to frequency distances in the transformed spectrum (see Figure 5). This transformation preserves neither the absolute nor the relative frequency spacing of the formants. We predicted that the ability to perform above chance on the cross-modal matching and word identification tasks should be eliminated under these conditions.

Finally, some of these acoustic transformations may bias performance by introducing into the signal high frequencies that are not normally produced by naturally articulating human vocal tracts. There is another transformation of the auditory signal that does not preserve the relative spacing of the formants and that also does not introduce new frequency components: a transformation called “frequency rotation” (Blessner, 1972). Under this transformation, frequency components are rotated around a center frequency in the signal (see Figure 6). Thus, the range of frequencies is preserved, while the original relationship among formants is destroyed. We expected that this transformation would eliminate participants’ ability to perform on both the cross-modal matching task and the word identification task, because it removes the critical information about speech articulation that is carried by the temporal and spectral patterning of the formants as they vary over time.

Temporal Domain Transformations

Because the formant frequencies of speech vary over time, several temporal transformations of the signal may also selectively preserve or attenuate cross-modal source information. Thus, linear scaling can be carried out in the temporal domain as well, speeding up or slowing down the original speaking rate. The spectral relationships between formants over time would remain invariant, however, across this transformation (see Figure 7). If cross-modal information is specified in the spectral relationships among formants, then this transformation should not disrupt cross-modal matching performance at all. However, if the temporal scale of formant patterning is important for conveying cross-modal information, performance should decrease on both the cross-modal matching and word identification tasks.

Acoustic patterns can also be nonlinearly scaled in the temporal domain. This transformation would produce a signal that sounds like accelerating or decelerating speech. Under this acoustic transformation, the relationship among the formants changes over time, but the

temporal information about articulation is degraded (see Figure 8). Consequently, we would expect cross-modal matching and word identification performance to be reduced or eliminated.

Method

Stimulus Materials

Four Apple Macintosh G4 computers, each equipped with a 17-in. Sony Trinitron Monitor (0.26 dot pitch) were used to present the visual stimuli to participants. The stimuli were a set of 96 tokens selected from a previously recorded audiovisual-stimulus set (Lachs & Hernández, 1998; Sheffert, Lachs, & Hernández, 1996). Each stimulus was a digitized, audiovisual movie of one talker speaking an isolated English word. The video portion of each stimulus was digitized at 30 fps with 24-bit color resolution and 640×480 pixel size. The audio portion of each stimulus was sampled and digitized at 22 kHz with 16-bit mono-resolution. Movie clips from eight talkers were used in this study. Auditory stimuli were presented over Beyer Dynamic DT 100 headphones at 74 dB SPL.

Because the stimulus items used in this study were all highly intelligible under audio-alone identification tests, stimulus items were split into *low* and *high* groups based on their visual intelligibility using data from visual-alone identification tests (Lachs & Hernández, 1998). Stimuli in the low group were words whose average visual-only (VO) intelligibility was in the bottom 1% of the distribution of VO intelligibilities for the stimulus set (Lachs & Hernández, 1998). Stimuli in the high group were taken from the top 5% of the same distribution. The percentages are different because of the extreme leftward skew of the VO intelligibility distribution (i.e., relatively few words had better than average accuracy scores). An equal number of low and high visual intelligibility words were randomly distributed in the first and second halves of each experiment.

Auditory Stimulus Transformations

The procedures used for signal processing under various transformations are described below. For two of the transformations (linear and nonlinear scaling), it was necessary to manipulate the source and filter characteristics of the sound files independently. This was accomplished using the STRAIGHT digital-signal-processing package developed by Hideki Kawahara (Kawahara, 1997; Kawahara, Masuda-Kastuse, & Cheveigné, 1999). STRAIGHT is a MATLAB-based analysis and synthesis package designed to extract the fine-structure vocal source from a sound file, leaving, in essence, a description of the vocal-tract transfer function as it evolves over time for a particular utterance. The method by which STRAIGHT accomplishes f0 extraction is very similar to the method used to create noise-band chimaeric stimuli (Smith et al., 2002). The method uses the Hilbert transform to extricate the fine-structure and modulation envelopes of the outputs of a bank of bandpass filters equally spaced on a logarithmic frequency scale. STRAIGHT also uses an additional computational component that assesses the instantaneous frequency of f0 and its harmonics in each channel, thereby allowing for fine-structure manipulation independent of envelope manipulation, and vice versa. Additionally, this method allows for the resynthesis of the manipulated fine-structure and envelope parameters (for details, see Kawahara, Katayose, de Cheveigné, & Patterson, 1999). Manipulations can be made independently to either the source or transfer function and can be reintegrated and resynthesized to produce highly natural-sounding speech signals. All other signal processing on the stimuli was carried out using custom-designed routines in MATLAB (Version 5; MathWorks, Natick, MA). Each participant was only presented with one type of stimulus from one stimulus class.

Spectral Shift

Spectral shifts were carried out by taking the Fast Fourier Transform (FFT) of each auditory stimulus and adding a constant to each frequency component. Stimuli were generated with shift magnitudes of 100 Hz, 250 Hz, 500 Hz, and 1000 Hz. Transformed frequencies that exceeded the Nyquist frequency were eliminated from the FFT. An inverse FFT transform was then used to resynthesize the acoustic stimulus. One way of representing these spectral transformations is with a *frequency map*, which shows the relationship between original frequencies and their corresponding output frequencies after transformation.

Figure 9A shows the frequency maps used for this class. Each dashed line represents the relationship between input frequencies from the original signal and output frequencies after transforming the signal. The various dashed lines correspond to shifts of different frequencies. The solid line shows a shift of 0 Hz (null transformation map). As shown in the figure, the shifted frequencies were always higher than the corresponding original frequencies.

Linear Scaling

Spectral domain—In the linear-scaling class of transformations, the ratio of an original frequency to its transformed frequency was held constant across all frequencies. For purposes of this transformation class, the transfer function of each sound file was manipulated independently of the source used to excite it using STRAIGHT. The transfer function was transformed such that the original frequency range was scaled to either 75% or 150% of its original range (see Figure 9B). Transformed frequencies that exceeded the Nyquist frequency were eliminated from the transformed signal. After spectral linear scaling was complete, the original source was reintegrated with the transformed filter to yield the transformed stimulus.

Temporal domain—Temporal scaling was also carried out using STRAIGHT, in order to preserve the fundamental frequency of the original stimuli. The filter function was extracted from the signal and adjusted such that its duration was either 75% or 150% of its original duration. The transformed-filter function was then reintegrated with the original source to make the new signal. This transformation was linear, such that the ratio of the duration of any segment in the original stimulus to its transformed counterpart was constant for all segments.

Nonlinear Stretching

Spectral domain—The STRAIGHT digital-signal-processing package was also used to perform nonlinear spectral scaling. As with linear scaling, the transfer function was extracted from its source and manipulated independently. Figure 9C shows the frequency maps for the two kinds of nonlinear stretching transformations used in these experiments. As shown in the figure, one map weighted frequencies more heavily toward high frequencies (dotted line), and one map weighted frequencies more heavily toward low frequencies (dot-dash line).

Temporal domain—Nonlinear temporal scaling was accomplished by breaking the signal into 10 approximately equal, continuous chunks. Each chunk was then linearly scaled at a different temporal rate. The chunks were then concatenated to yield the nonlinearly scaled stimulus. Although this transformation did not provide a continuous nonlinear function relating temporal patterning in the original to patterning in the transformed signal, it did approximate one. The temporal rate for each chunk was calculated such that the overall duration of the transformed signal was either 75% or 150% of the original duration.

Inversion Class

For spectral inversion, the frequency range of an acoustic signal was rotated about some center frequency (Blesser, 1969, 1972). Spectral inversion required several transformations of the original signal. First, the sound file was low-pass filtered at twice the center frequency, to eliminate any high-frequency components that would not be rotated in the transformation. Second, an FFT was obtained. The complex components of the FFT were then treated separately. The real component was rotated such that the transformed component was rotated about the chosen center frequency. The imaginary component, however, was not changed, so that phase information remained intact with the rotated frequencies. The resulting stimulus pattern preserved the original temporal information but rotated spectral information (see Figure 9D). To make sure that performance with the spectrally rotated stimuli was not due to disproportionate concentrations of acoustic energy in unfamiliar regions of frequency space, we used two center frequencies to examine the effect of the range over which the inversion takes place. Spectral inversion was performed around 2500 Hz and 1500 Hz.

Participants

Participants were 420 undergraduate students enrolled in introductory psychology who received partial credit for participation. All of the participants were native speakers of English. None of the participants reported any hearing or speech disorders at the time of testing. In addition, all participants reported having normal or corrected-to-normal vision. None of the participants in this experiment had any previous experience with the audiovisual speech stimuli used in this experiment. For each auditory transformation, 20 participants received the cross-modal matching task, and a separate group of 10 participants received the word identification task.

Procedure

Cross-modal matching—Figure 1 shows a schematic description of the cross-modal matching task. Participants in the *V–A condition* were first presented with a visual-alone movie clip of a talker uttering an isolated, familiar English word. Shortly after seeing this video display (500 ms), participants were presented with two transformed auditory-alone movie clips. One of the clips was the same talker they had seen in the video, while the other clip was a different talker. Participants were instructed to choose which audio clip matched the talker they had seen (“First” or “Second”). Similar instructions were provided for participants in the *A–V condition*. These listeners heard the audio clip first, and then had to make their decision based on two video displays.

On each trial of the experiment, the test stimulus was either the video or audio portion of one movie token. Each movie token presented an isolated word spoken by a single talker. The order in which the target and distractor choices were presented was randomly determined on each trial. For each participant, talkers were randomly paired with each other, such that each talker was compared with one and only one other talker for all trials in the experiment. For example, “Bob” was always compared with “John,” regardless of whether “Bob” or “John” was the target alternate on the trial. In addition, all the talkers viewed by any particular participant were of the same gender. The gender of the talkers was counterbalanced across participants, such that an equal number of participants made judgments using male or female speakers. Responses were made by pressing one of two buttons on a button box and were recorded in a log file for later analysis.

A short familiarization period (eight trials) preceded each participant’s session. During this period, the participant was presented with a cross-modal matching trial and asked to pick the correct response alternative. During training, the response was followed with feedback. The

feedback consisted of playing back the entire audiovisual movie clip of the target item. The feedback clip was also transformed under the relevant transformation. After training, participants judged matches with an entirely new set of talkers, so that feedback could not play a role in their final performance during testing.

Word identification—In the word identification task, each participant heard the 96 words transformed by one acoustic transformation. Only stimuli spoken by Talker F1 were presented in this task. F1 was chosen because previous data (Sheffert et al., 1996) showed that she was the most intelligible talker in auditory-alone contexts out of the eight talkers used in the cross-modal matching task. On each trial, participants were presented with an auditory token of F1 speaking an isolated English word. After presentation of the stimulus, the participant was asked to enter the word they heard using the keyboard. Participants were asked to make sure each word they typed contained no typos or spelling errors before pressing the ENTER button.

Before scoring, each participant's responses were hand-screened for typing and spelling errors by two reviewers who worked independently. A typing error was defined as a substituted letter within one key on a standard keyboard of the target key or an inserted letter within one key of an adjacent letter in the response. Spelling errors were only accepted if the letter string did not form a word in its own right. Using this very conservative method of assessment, the reviewers had a 100% agreement rate on classifying responses as typing and spelling errors. Responses were scored as correct if and only if they were homophonous with the target word in a standard American English dialect (e.g., “bare” for “bear,” but not “pin” for “pen”).

Results

Frequency Shifting

Cross-modal matching task—The left panel of Figure 10 displays the distribution of cross-modal matching scores for participants who heard frequency-shifted acoustic tokens. Each boxplot represents the distribution for one shift magnitude. For this and all boxplots presented in this article, the shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the standard error bars represent the range from the highest to the lowest score within the sample, excluding outliers. The figure shows that participants were, in general, successful at making cross-modal matching judgments using frequency-shifted stimuli. Performance in each condition was significantly different from chance (.5). Thus, frequency shifting did not prevent participants from matching faces and voices across modalities. Table 1 summarizes the descriptive statistics for performance in each condition. As shown here, performance in all conditions was significantly different from chance (.5). The data were also submitted to a $2 \times 2 \times 4$ repeated-measures analysis of variance (ANOVA; visual intelligibility, order, and shift magnitude). No main effects or interactions were revealed by this analysis.

Word identification task—The right panel of Figure 10 shows the distribution of performance on the word identification task for the four frequency shift conditions used above. Performance was good overall, but accuracy in identifying the words declined as the magnitude of the frequency shift increased. A one-way ANOVA on the accuracy scores revealed a strong effect of shift magnitude, $F(1, 36) = 101.35, p < .05, \eta^2 = .894$. Post hoc comparisons between the conditions (adjusted for multiple comparisons using Bonferroni's method) showed that performance with shift magnitudes of 100 Hz and 250 Hz did not differ from one another. All other comparisons, however, did differ from one another.

Summary—Under all four frequency shift values tested, cross-modal matching performance was above chance, and word identification performance varied over a wide range. The results support the proposal that cross-modal source information and phonetic information about the linguistic content of an utterance are carried in the parallel patterning of formant frequencies as they evolve over time and is not conditional on their absolute frequency values.

Linear Frequency Scaling

Cross-modal matching task—The left panel of Figure 11 shows the range of cross-modal matching scores for participants who heard linearly scaled acoustic tokens. Each boxplot represents the data obtained under one transformation. Inspection of the figure reveals a reduced range of performance compared with the frequency shifted tokens. However, cross-modal matching was still possible even after linear frequency scaling. Table 2 summarizes the descriptive statistics for performance in each condition. As shown here, performance in all conditions was significantly different from chance (.5).

A $2 \times 2 \times 2$ repeated-measures ANOVA (visual intelligibility, order, and scaling magnitude) also revealed a significant effect of scaling magnitude, $F(1, 36) = 7.06$, $p < .05$, $\eta^2 = .164$. Overall, performance with the 150% scaling magnitude ($M = 0.602$, $SE = 0.016$) exceeded performance with the 75% scaling magnitude ($M = 0.543$, $SE = 0.016$). No other main effects or interactions tested by the ANOVA reached significance.

Word identification task—The right panel of Figure 11 shows word identification performance for the two linear frequency-scaling magnitudes used in the cross-modal matching task. Performance under both scaling magnitudes was good, suggesting that the words were still intelligible after being subjected to linear frequency scaling. A t test for independent samples revealed that performance between the two scaling magnitudes differed from each other, $t(18) = 9.17$, $p < .05$, $d = 4.32$. Word identification performance with the 150% scaling magnitude ($M = 0.728$, $SE = 0.020$) was worse than performance with the 75% scaling magnitude ($M = 0.926$, $SE = 0.008$).

Summary—The results from the linear frequency-scaling transformation indicate that cross-modal source information and word identification information are preserved under conditions that alter the absolute frequencies of formants but preserve the relationships among them. One puzzling result arose from the two values of the transformation used. Although cross-modal matching performance was better with the transformation that extended the range of frequencies covered by the three lowest formants (the 150% magnitude), word identification performance was better with the transformation that reduced the range of frequencies covered by the formants (the 75% magnitude). It is possible that the 150% scaling magnitude emphasized those details of the acoustic signal particularly useful for making cross-modal matching judgments, while the 75% scaling magnitude exaggerated those details of the acoustic signal useful for recognizing words. A preliminary inspection of the errors made by listeners in the word identification experiment revealed an interesting pattern. Errors on stimuli transformed at the 150% magnitude were most often errors in vowel quality. In contrast, errors on stimuli transformed at the 75% magnitude were most often errors in the initial phoneme. It is possible that the fine-grained articulatory details needed for making cross-modal matching judgments are more discriminable in patterns of consonantal articulation than in vowel articulation. If the 150% scaling magnitude affected vowel quality relatively more strongly than the 75% scaling magnitude affected consonantal information, this would explain the current pattern of results. However, definitive conclusions about the extent to which different linear spectral-scaling magnitudes influence

the perception of linguistic and indexical properties of articulation cannot be drawn with the present results, and the topic remains an important avenue for future investigation.

Nonlinear Frequency Scaling

Cross-modal matching task—The left panel of Figure 12 shows the range of cross-modal matching scores obtained from participants who heard auditory stimuli transformed with nonlinear frequency scaling. Each boxplot represents the data obtained under one transformation and one order. As shown in the figure, it is clear that many of the participants had a great deal of difficulty under these conditions. Table 3 shows a summary of the descriptive statistics for performance in each condition. Inspection of the table reveals that the participants who heard the signal after nonlinear scaling toward the Nyquist frequency did not significantly differ from chance performance. However, the performance of participants who heard acoustic stimuli scaled away from the Nyquist frequency did differ significantly from chance performance. This pattern of results indicates that the nonlinear frequency-scaling transformation eliminated cross-modal source information, but only when the scaling was in one direction (toward the Nyquist frequency).

A $2 \times 2 \times 2$ repeated-measures ANOVA (visual intelligibility, order, and scaling direction) explored this relationship further. The ANOVA revealed a significant main effect of scaling direction, $F(1, 36) = 4.47, p < .05, \eta^2 = .110$, confirming that cross-modal matching performance with stimuli scaled away from the Nyquist frequency was better than performance with stimuli scaled toward the Nyquist frequency. The ANOVA also revealed a significant interaction between scaling direction and visual intelligibility (see Figure 13), $F(1, 36) = 4.68, p < .05, \eta^2 = .115$. Simple effects analysis revealed that the interaction was supported by a difference between the two scaling directions with low visual intelligibility tokens, $t(38) = 2.93, p < .05$. Performance on low visual intelligibility tokens was better when scaling was away from the Nyquist frequency than when scaling was toward it. Thus, the overall advantage in cross-modal matching for nonlinear scaling away from the Nyquist was most evident for low visual intelligibility words. Because the visual information about articulation was relatively poor under these conditions, perceivers may have been forced to pay closer attention to the acoustic details in the transformed signal to perform the task, thereby overcoming any disadvantages in performing the matching task due to the acoustic transformation itself.

Word identification task—The right panel of Figure 12 shows word identification performance for the two nonlinear frequency-scaling transformations. The figure shows a large difference in word identification scores under these two transformations. For the nonlinear scaling toward the Nyquist frequency, performance is at the floor ($M = 0.017, SE = 0.004$). However, performance with scaling away from the Nyquist frequency was very high ($M = 0.881, SE = 0.015$). These findings are consistent with the results of the cross-modal matching task for this transformation. Nonlinear scaling away from the Nyquist preserved the ability to perform the cross-modal matching task as well as the ability to accurately identify words. In contrast, nonlinear scaling toward the Nyquist completely eliminated the ability to perform both the cross-modal matching task and the word identification task.

Summary—The results demonstrate that acoustic transformations that significantly alter the linear relationships between formant frequencies eliminate the information necessary for both cross-modal matching and word identification. The link between word identification and cross-modal matching is especially evident from the results using this acoustic transformation. One of the values of this transformation preserved cross-modal matching ability, whereas the other eliminated cross-modal matching task. However, the same value

that preserved cross-modal source information also preserved word identification information, whereas the value that did not preserve cross-modal source information eliminated word identification information. Thus, word identification information and cross-modal source information appear to be encoded together in the same properties of the acoustic signal.

Frequency Rotation

Cross-modal matching task—The left panel of Figure 14 shows the range of cross-modal matching scores obtained from participants who heard the spectrally rotated acoustic stimuli. Inspection of the figure reveals that most participants were unable to perform the matching task above chance (.5). Table 4 shows the descriptive statistics for performance in each condition. Across the conditions, performance did not differ significantly from chance.

The results were also submitted to a $2 \times 2 \times 2$ repeated-measures ANOVA (visual intelligibility, order, and center frequency). None of the main effects or interactions tested in the ANOVA reached significance ($p < .05$). Frequency rotation therefore appears to eliminate cross-modal source information from an acoustic stimulus.

Word identification task—The right panel of Figure 14 shows word identification performance with spectrally rotated speech with the two center frequencies (1500 Hz and 2500 Hz). The results show that performance in this task was quite poor. Average performance was at the floor for speech rotated around 1500 Hz ($M = 0.016$, $SE = 0.004$) and 2500 Hz ($M = 0.020$, $SE = 0.006$). Performance in the two conditions was not significantly different, $t(18) < -0.59$, *ns*.

Summary—For both center frequencies used, the spectral rotation of acoustic signals eliminated the information necessary for both cross-modal matching and word recognition. The frequency rotation transformation preserves the long-term spectral characteristics of the acoustic signal but destroys the particular patterns of formant structure in time. The poor performance observed on both tasks is again consistent with the proposal that cross-modal source information and the information necessary for word recognition are carried together in the spectral structure of the formants as they evolve over time.

Linear Temporal Scaling

Cross-modal matching task—The left panel of Figure 15 shows the range of cross-modal matching scores obtained from participants who heard acoustic stimuli that were subjected to linear temporal scaling. Inspection of the figure reveals that over half of the participants in each condition performed above the statistical criterion for chance performance, indicating that participants were able to perform the task. Table 5 shows the descriptive statistics for performance in each condition. The results of the *t* tests revealed that performance in all the conditions was significantly above chance (.5), indicating that participants were able to perform the cross-modal matching task when the acoustic information had been transformed with linear temporal scaling. The results were also submitted to a $2 \times 2 \times 2$ repeated-measures ANOVA (visual intelligibility, order, and scaling magnitude). None of the main effects or interactions tested in the ANOVA reached significance ($p < .05$). Thus, linear temporal scaling of the acoustic signal appears to preserve the information necessary for cross-modal source judgments.

Word identification task—The right panel of Figure 15 shows the word identification performance with acoustic stimuli transformed with linear temporal scaling. Once again, performance was near the ceiling for shortened stimuli ($M = 0.940$, $SE = 0.009$) as well as for the lengthened stimuli ($M = 0.942$, $SE = 0.009$). There was no significant difference

between performance with the types of stimuli, $t(18) < -0.15$, *ns*. Participants therefore had little trouble identifying the words correctly in temporally scaled stimuli.

Summary—The pattern of results indicates that linearly scaling the acoustic stimuli to be shorter or longer in duration than their original recordings did not adversely affect cross-modal matching performance or word recognition performance in any of the conditions. The results support the proposal that cross-modal source information, as well as phonetic information used in spoken word recognition, is carried in the relative temporal structure of the formants, and not on their absolute temporal relationships.

Nonlinear Temporal Scaling

Cross-modal matching task—The left panel of Figure 16 shows a summary of cross-modal matching scores obtained from participants who heard acoustic stimuli that were subjected to nonlinear temporal scaling. The figure shows that, under some conditions, some participants were able to perform the cross-modal matching task. Table 6 shows the descriptive statistics for performance in each condition. Inspection of the table reveals that overall performance on cross-modal matching with the nonlinear temporal-scaling transformation only differed from chance for the V–A order. However, a $2 \times 2 \times 2$ repeated-measures ANOVA (visual intelligibility, order, scaling direction) did not show a main effect of order, $F(1, 36) = 3.13$, *ns*. Collapsing across the order variable, mean performance for the participants was above chance for both the *accelerated*, $t(19) = 2.32$, $p < .05$, and *decelerated*, $t(19) = 2.87$, $p < .05$, conditions. Thus, although performance on the cross-modal matching task was very poor, the group as a whole did differ significantly from chance.

Word identification task—The right panel of Figure 16 shows word identification performance with acoustic stimuli transformed using nonlinear temporal scaling. Performance was good for the decelerated condition ($M = 0.880$, $SE = 0.011$) and only slightly worse for the accelerated condition ($M = 0.832$, $SE = 0.019$). There was also a significant difference between performance in the two conditions, $t(18) = 2.21$, $p < .05$, $d = 1.04$.

Summary—Overall, the nonlinear temporal-scaling transformations reduced cross-modal matching performance substantially but did not ultimately eliminate cross-modal matching ability or word identification performance.

Discussion

In the present investigation, spoken utterances of single talkers producing isolated English words were subjected to several different spectral and temporal transformations that systematically manipulated the relationships between formant frequencies as they changed over time. The 14 transformations were used in both a cross-modal matching task and a word identification task, to assess whether each transformation class preserved or eliminated cross-modal source information (i.e., the information relating voices and faces across sensory modalities) and phonetic information used for spoken word recognition. If performance on the cross-modal matching task was disrupted by any particular transformation class, then it was assumed that the information necessary for matching faces and voices across sensory modalities would be attenuated or eliminated. We predicted that as long as information about the relative spacing of the formants was preserved in either the spectral or temporal domains, cross-modal source information would also be preserved and participants would be able to successfully carry out the matching task as well as the word identification task.

Because of the strong links between cross-modal source information and word identification information reported by Lachs (2002), acoustic stimuli from each transformation class were also presented in a word identification task. We expected that accurate word identification performance would be preserved only under those transformations that supported cross-modal matching. For transformations that disrupted cross-modal matching performance, we predicted that word identification accuracy would also be severely reduced. Indeed, in an overall comparison of the average performance on each task under each of the 14 acoustic transformations, we found that there was a significant correlation between performance on the cross-modal identification task and the word identification task, across transformations, $r(14) = .74, p < .05$.

This is not a trivial result. Explicit identification of the target word is not necessary for successful cross-modal matching to be performed; at no point in the cross-modal matching task was a participant ever asked to identify the word being spoken. In fact, for all three stimuli presented over the course of a trial, the word was identical. Because both cross-modal source information and word identification information were degraded by exactly the same types of signal transformations and manipulations, the results of both procedures are consistent with the proposal that indexical and linguistic information are carried in parallel components of the acoustic signal and do not appear to be dissociable from one another.

The results from the cross-modal matching task using the four spectral domain transformations showed that cross-modal source information is retained when the ratio of spacing between formants remains invariant across the transformation. With manipulations in the spectral domain, however, cross-modal matching performance was different from chance under the shifting and linear scaling transformations, but not under the nonlinear scaling and inversion transformations. Because the former two classes of transformation preserve the relative spacing of formant frequencies with respect to the frequency axis and because the latter two transformations do not, this pattern of results provides support for the hypothesis that cross-modal source information is specified in the relative spacing of formant frequencies as they change over time.

In addition, the results from the spectral domain transformations confirmed the hypothesized link observed earlier between cross-modal source information and word identification information. Those transformations that preserved cross-modal matching also provided reliable acoustic–phonetic information needed for accurate word identification, and transformations that reduced cross-modal matching also reduced word identification performance.

Both temporal domain transformations preserved the ability to do the cross-modal matching task. As expected, linear temporal scaling did not prevent participants from performing the cross-modal matching task above chance. In addition, performance on the word identification task remained quite good. Together, these results indicate that cross-modal source information and word identification information is specified in the spectral patterning of formants and is not contingent on specific, absolute temporal relationships but is rather specified over relative temporal relationships in the patterns as they change over time.

Although we expected that nonlinear temporal scaling would eliminate cross-modal source information, it is possible that because the temporal domain scaling was not continuous (the signal was split into 10 parts, each of which was linearly scaled at a different rate), the desired warping of the temporal information was not achieved at the necessary points in the signal (e.g., at formant transitions or within the duration of a vowel segment). Good word identification performance was also observed under both transformations, confirming that cross-modal source information and word identification are intimately tied together in the

acoustic signal. Any further work investigating the temporal nature of cross-modal source information will have to await the development of a method of nonlinearly scaling temporal information continuously over the duration of the utterance.

It is possible that our word identification results were somehow peculiar to the particular female talker we chose to use as the speaker in all the word identification experiments. To ensure that the pattern of results observed in the word identification tasks was not simply a function of the particular talker we selected, we conducted an additional, follow-up word identification task using a male talker. Four participants were tested in all 14 transformation conditions. The order in which the transformation conditions were presented was randomized across the participants. The results are presented in Table 7. Performance with the transformed male talker matched the findings observed with the transformed female talker. Indeed, there was a significant correlation, $r(14) = .77$, $p < .05$, between the average identification scores for the female talker and the male talker across the various acoustic transformations.

The findings presented here demonstrate that information about speech articulation is encoded acoustically on a relative scale; absolute differences in frequency and temporal patterning were not necessary for either cross-modal matching or word identification, as long as the linear relationships were still preserved. Thus, a search for acoustic cues to linguistic structure needs to be informed by the relative and global spectral structure of the acoustic signal, and not the absolute frequency values or relationships.

An interesting property of speech perception uncovered here is that formant relationships do not appear to be constrained to a narrow region of the frequency space near 0 Hz; even when the speech spectrum was shifted up by 1000 Hz, cross-modal matching (and word identification) was possible. Simple algebra shows that, for all real numbers a , b , and c ,

$$(a+c)/(b+c) \neq (a/b).$$

If the key information for relating an acoustic speech signal to an optical one lies in the ratio of formant frequency a to formant frequency b , then simply shifting the spectrum by a constant c should have eliminated cross-modal source information.

This implies that the informative formant ratios are scaled with respect to an origin intrinsic to the spectrum itself, and not to the absolute periodicity of the component frequencies. It is possible that the fundamental frequency (f_0) might provide this origin point; the absolute ratios of formant frequencies are correlated with f_0 . However, one argument against this explanation comes from an earlier study on cross-modal matching (Lachs & Pisoni, in press): Although the noise-band transformation eliminates f_0 from the signal, accurate cross-modal matching performance was still possible with noise-band stimuli that did not have an actual f_0 present. The fundamental frequency, then, does not appear to be necessary for scaling the ratios between formants.

Another possibility is that formant ratios are evaluated with respect to the region of frequency space within which formants are normally experienced. That is, the interpretation of formants may be conditional on the experience of formant relationships within a particular range (e.g., 0–5 kHz). The finding that word identification performance declined linearly with the extent of the frequency shift supports this explanation, although it is by no means conclusive. Future work on the acoustic transmission of articulatory information will need to elaborate on the ways in which formant ratios are scaled relative to each other and

may need to take into account the frequency regions over which formants are normally scaled and experienced by listeners.

It is interesting to note that a parallel line of investigation has been conducted recently by Harnsberger et al. (2001) with users of cochlear implants. Cochlear implants are surgically implanted prosthetic devices that directly stimulate the auditory nerve with electrical impulses corresponding to incoming auditory signals. Because an array of electrodes must be threaded by hand by the surgeon into the cochlea and because the electrode array encounters resistance in the many turns of the cochlea, it is frequently observed that after the procedure is completed, only the basal 25 mm (at best) of the neural substrate inside the cochlea can be stimulated electrically. This, in effect, leads to a frequency shift of the type examined here. Neurons with high characteristic frequencies are stimulated by lower veridical frequencies. For example, a tone of 300 Hz will stimulate the most apical electrode in the array, regardless of where the array is placed in the cochlea. Thus, neurons with higher characteristic frequencies (e.g., 1000 Hz) are stimulated by lower veridical frequencies.

Despite the so-called “basalward” shift in frequency due to their implants, cochlear implant recipients are generally able to categorize vowels effectively. Harnsberger et al. (2001) used a novel task to explore this issue (see Johnson, Flemming, & Wright, 1993). First, participants were presented with an orthographically presented word. Listeners were then asked to locate an auditory token corresponding to the vowel in the word in a 2×2 matrix of synthesized F1–F2 combinations. The stimulus space was arranged such that each row represented a frequency step of 0.377 Bark for F1 and each column represented a frequency step for F2 of the same magnitude. Thus, each cell represented a different vowel that was closely related to the vowels in its neighboring cells. By clicking the mouse button repeatedly around the cells in a specific region of this graphical vowel space, listeners could find the best token that matched their perception of the vowel in the visually presented word. Harnsberger et al. found that the perceptual vowel spaces of the listeners with cochlear implants (all of whom had acquired or progressive deafness) were not distorted relative to the vowel spaces of normal hearing listeners. This finding suggests that the cochlear implant recipients were able to adapt to the sensory frequency shift imposed by their implant for the purposes of speech perception.

The results of the present study indicate that frequency shifts of up to 1000 Hz can also be tolerated by normal-hearing listeners. Taken together with the results of the other acoustic transformations, the frequency shift results presented here indicate that the development of cochlear implant processing strategies may need to focus more on the preservation of key spectral relationships among formants, without concentrating as heavily on the veridical representation of the signal by stimulation of the cochlea.

In discussing their results on the perception of talker specific information from sine-wave speech replicas, Remez et al. (1997) proposed that a common form or representation of speech must simultaneously support both lexical (linguistic) and indexical analyses, and they argued for a phonetic, rather than acoustic or phonemic, level of description. Although the present set of experiments were not designed to directly address the relevance of a phonetic level of description, the present results indicate that acoustic descriptions of speech alone may lead to the misleading conclusion that sounds are important to the process of speech perception, in and of themselves, and not as carriers of more general, modality-neutral information about the underlying articulation and events and sources that generated the sounds (Fowler & Galantucci, in press; Gaver, 1993; Gibson, 1966; Rosenblum, in press). The present findings show that those aspects of an acoustic speech signal that carry both indexical and lexical variation are also carried in visual speech signals. Moreover, these

distinctive perceptual aspects of speech appear to be modality-neutral because participants are able to match similar patterns accurately across two different sensory modalities.

Taken together, the present findings provide additional support for the hypothesis that acoustic displays of speech specify their optical correlates by virtue of their ability to specify common underlying articulatory information. Because the articulatory behavior of a talker has acoustic as well as optical correlates, the common origin of patterns in either sensory modality suggests a natural and principled basis for making cross-modal judgments of talker identity based on articulatory activity.

Is it possible that the form of speech representation is modality-neutral and based on common underlying articulatory properties? Although it is true that acoustic and optic displays can covary arbitrarily (Massaro, 1998), such covariation is only made possible by virtue of current cinematic technology. In everyday, real-world situations, a perceiver is never confronted with a stimulus event in which the visual display of a particular talker's lips specifies one set of articulatory behaviors and the acoustic display specifies an incompatible set. In fact, the relationship between auditory and visual speech in natural contexts is precisely the opposite of arbitrary; it is lawful and highly structured. The motor patterns used for shaping acoustic energy into communicative speech signals necessarily create correlated visual patterns in the structure of optic energy (Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzepoulos, 1997). Because optical and acoustic patterns are each lawfully related to the underlying articulatory behaviors that generated them, there is a natural relationship between visual and auditory displays of speech across sensory modalities: Acoustic patterns of speech are caused by the same articulatory behaviors that lawfully structure optical patterns of speech.

Another avenue of investigation into the informative properties of acoustic and optic displays for speech concerns the precise aspects of articulatory behavior that can be perceived. One possibility is that observers perceive the dynamic properties (mass, stiffness, force, etc.) of the articulatory event as it unfolds over space and time (Bingham, Schmidt, & Rosenblum, 1995; Runeson & Frykholm, 1983). Investigations in the field of visual perception have shown that the dynamics of events are indeed perceived. For example, the mass of a lifted weight can be perceived visually from the movements of the person doing the lifting (Bingham, 1995). With specific reference to audiovisual speech perception, Summerfield (1987) has suggested that auditory and visual inputs might be evaluated along a common metric, based on the kinematics and dynamics of the underlying articulatory behavior.

Runeson and Frykholm (1983) have argued that dynamics are recovered through their specification by kinematic properties, like the position, velocity, and acceleration of key points involved in the event. Experiments using visual displays that isolate kinematic information with the point-light technique are sufficient for the perception of biological motion (Johansson, 1973) and, as discussed above, can be used in speech perception studies (Rosenblum, Johnson, & Saldaña, 1996). Analogously in the acoustic domain, sine-wave speech preserves and extracts kinematic information about the configuration of speech articulators by tracing the centers of the three lowest formants with three sinusoidal tones (Remez et al., 1981). These highly unusual patterns can also be perceived as speech. It is possible that these two, domain-specific forms of kinematic information specify a common, underlying dynamic description of the articulatory behavior of the vocal tract.

The value of such a description for representing the idiosyncratic properties of the talker is, on the face of it, obvious; talker differences, without inference or computation, would be encoded in the instantaneous kinematics of a particular vocal tract behaving in a particular

way. In addition, a dynamic, gestural description for linguistic stimulus and processing has also been proposed: Browman and Goldstein (1995) outline an “articulatory phonology” in which dynamic, gestural constraints on the activity of the vocal tract describe linguistically significant phonological variation and underlying linguistic rules. There is also evidence that a dynamic description of events can support categorical judgments. In one study, Bingham, Schmidt, and Rosenblum (1995) showed that observers categorized point-light displays varying in their gross dynamic properties (hydrodynamics vs. aerodynamics, animate vs. inanimate motion, rigidity vs. elasticity, etc.). Dynamic properties such as liquidity, elasticity, and rigidity were all easily distinguished from each other, even in the context of the isolated kinematic information.

Further work in this area will need to evaluate the optical displays of speech with respect to the dynamic articulatory events that produce them (see Bingham, 1995). Although a dynamic description of vocal-tract activity is likely to be exceedingly complex, recent developments in articulatory modeling may be able to provide new insights into the kinematic trajectories traced by isolated points or formants (Löfqvist & Gracco, 1997; Lucero & Munhall, 1999; Vatikiotis-Bateson et al., 1997; Vatikiotis-Bateson, Munhall, Kasahara, Garcia, & Yehia, 1996). In fact, such work has already led to satisfactory synthesis of facial motion driven by an acoustic input (Yehia, Kuratate, & Vatikiotis-Bateson, 2002). With continued refinement, a modality-neutral description of speech framed in terms of the underlying dynamics of articulatory gestures may provide a simple and powerful conceptual framework and explanatory device for observed audiovisual speech perception and spoken-language-processing phenomena and may also lead to more powerful theoretical descriptions of speech perception in general.

Acknowledgments

This research was submitted by Lorin Lachs in partial fulfillment of the requirements for a doctoral dissertation at Indiana University and was supported by National Institutes of Health–National Institute of Child Health and Human Development Training Grant DC00012 and National Institute on Deafness and Other Communication Disorders Research Grant DC-00111 to Indiana University. We would like to thank Luis R. Hernandez, Jeff Karpicke, and Jeff Reynolds for invaluable assistance during all phases of this research. In addition, we thank Robert Port and Thomas Busey for their valued input and advice.

References

- Bingham, GP. Dynamics and the problem of visual event recognition. In: Port, RF.; Van Gelder, T., editors. *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press; 1995. p. 403–448.
- Bingham GP, Schmidt RC, Rosenblum LD. Dynamics and the orientation of kinematic forms in visual event recognition. *Journal of Experimental Psychology: Human Perception and Performance*. 1995; 21:1473–1493. [PubMed: 7490589]
- Blessner, BA. Unpublished doctoral dissertation. Massachusetts Institute of Technology; Cambridge: 1969. Perception of spectrally transformed speech.
- Blessner B. Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *Journal of Speech and Hearing Research*. 1972; 15(1):1–41.
- Bradlow AR, Torretta GM, Pisoni DB. Intelligibility of normal speech: I. Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*. 1996; 20:255–273. [PubMed: 21461127]
- Bricker, PD.; Pruzansky, S. Speaker recognition. In: Lass, NJ., editor. *Contemporary issues in experimental phonetics*. New York: Academic Press; 1976. p. 295–326.
- Browman, CP.; Goldstein, L. Dynamics and articulatory phonology. In: Port, RF.; Van Gelder, T., editors. *Mind as motion*. Cambridge, MA: MIT Press; 1995.
- Bruce, V. *Recognising faces*. Hove, United Kingdom: Erlbaum; 1988.

- Creelman CD. Case of the unknown talker. *Journal of the Acoustical Society of America*. 1957; 29:655.
- Fourcin AJ. Speech-source interference. *IEEE Transactions in Audio Electroacoustics*. 1968; ACC-16:65–67.
- Fowler CA. Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*. 1996; 99:1730–1741. [PubMed: 8819862]
- Fowler, CA.; Galantucci, B. The relation of speech perception and speech production. In: Pisoni, DB.; Remez, RE., editors. *Handbook of speech perception*. Oxford, United Kingdom: Blackwell; (in press)
- Gaver WW. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*. 1993; 5(1):1–29.
- Gibson, JJ. *The senses considered as perceptual systems*. Boston: Houghton Mifflin; 1966.
- Goldinger SD. Echoes of echoes? An episodic theory of lexical access. *Psychological Review*. 1998; 105:251–279. [PubMed: 9577239]
- Grant KW. The effect of speechreading for masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America*. 2001; 109:2272–2275. [PubMed: 11386581]
- Green, KP. The use of auditory and visual information in phonetic perception. In: Stork, DG.; Hennecke, ME., editors. *Speechreading by humans and machines*. Berlin, Germany: Springer-Verlag; 1996.
- Green KP, Gerdeman A. Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception and Performance*. 1995; 21:1409–1426. [PubMed: 7490588]
- Harnsberger JD, Svirsky MA, Kaiser AR, Pisoni DB, Wright R, Meyer TA. Perceptual “vowel spaces” of cochlear implant users: Implications for the study of auditory adaptation to spectral shift. *Journal of the Acoustical Society of America*. 2001; 109:2135–2145. [PubMed: 11386565]
- Hirahara, T.; Kato, H. The effects of F0 on vowel identification. In: Tohkura, Y.; Vatikiotis-Bateson, E.; Sagisaka, Y., editors. *Speech perception, production, and linguistic structure*. Tokyo: Ohmsha; 1992. p. 89-112.
- Hollien H, Klepper B. The speaker recognition problem. *Advances in Forensic Psychology & Psychiatry*. 1984; 1:87–111.
- Johansson G. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*. 1973; 14:201–211.
- Johnson K, Flemming E, Wright R. The hyperspace effect: Phonetic targets are hyperarticulated. *Language*. 1993; 69:505–528.
- Kawahara, H. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. Paper presented at the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing; Munich, Germany. 1997 Apr.
- Kawahara, H.; Katayose, H.; de Cheveigné, A.; Patterson, RD. Fixed points analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. Paper presented at EUROSPEECH '99; Budapest, Hungary. 1999 Sep.
- Kawahara H, Masuda-Kastuse I, Cheveigné A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*. 1999; 27:187–207.
- Lachs, L. Unpublished doctoral dissertation. Indiana University; Bloomington: 2002. Vocal tract kinematics and crossmodal information.
- Lachs, L.; Hernández, LR. Research on spoken language processing progress report. Bloomington, IN: Speech Research Laboratory, Indiana University; 1998. Update: The Hoosier Audiovisual Multitalker Database; p. 377-388.
- Lachs, L.; McMichael, K.; Pisoni, DB. Speech perception and implicit memory: Evidence for detailed episodic encoding. In: Bowers, J.; Marsolek, C., editors. *Rethinking implicit memory*. Oxford, United Kingdom: Oxford University Press; 2003. p. 215-239.
- Lachs L, Pisoni DB. Crossmodal source identification in speech perception. *Ecological Psychology*. (in press).

- Löfqvist A, Gracco VL. Lip and jaw kinematics in bilabial stop consonant production. *Journal of Speech, Language, and Hearing Research*. 1997; 40:877–893.
- Lucero JC, Munhall KG. A model of facial biomechanics for speech production. *Journal of the Acoustical Society of America*. 1999; 106:2834–2842. [PubMed: 10573899]
- Massaro, DW. *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press; 1998.
- McGurk H, MacDonald J. Hearing lips and seeing voices. *Nature*. 1976 Dec 23;264:746–748. [PubMed: 1012311]
- Mullennix JW, Pisoni DB. Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*. 1990; 47:379–390. [PubMed: 2345691]
- Mullennix JW, Pisoni DB, Martin CS. Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*. 1989; 85:365–378. [PubMed: 2921419]
- Pisoni DB. Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*. 1993; 13:109–125. [PubMed: 21461185]
- Pisoni, DB. Some thoughts on “normalization” in speech perception. In: Johnson, K.; Mullennix, JW., editors. *Talker variability in speech processing*. San Diego, CA: Academic Press; 1997. p. 9-32.
- Remez RE, Fellowes JM, Rubin PE. Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*. 1997; 23:651–666. [PubMed: 9180039]
- Remez RE, Rubin PE, Berns SM, Pardo JS, Lang JM. On the perceptual organization of speech. *Psychological Review*. 1994; 101:129–156. [PubMed: 8121955]
- Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. *Science*. 1981 May 22;212:947–950. [PubMed: 7233191]
- Rosenblum LD. How special is audiovisual speech integration? *Current Psychology of Cognition*. 1994; 13(1):110–116.
- Rosenblum, LD. The primacy of multimodal speech perception. In: Pisoni, DB.; Remez, RE., editors. *Handbook of speech perception*. Oxford, United Kingdom: Blackwell; (in press)
- Rosenblum LD, Johnson JA, Saldaña HM. Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research*. 1996; 39:1159–1170.
- Rosenblum LD, Yakes DA, Baseer N, Panchal A, Nodarse BC, Niehaus RP. Visual speech information for face recognition. *Perception & Psychophysics*. 2002; 64(2):220–229. [PubMed: 12013377]
- Runeson S, Frykholm G. Kinematic specification of dynamics as an informational basis for person- and action perception: Expectation, gender recognition, and deceptive intention. *Journal of Experimental Psychology: General*. 1983; 112:585–615.
- Sheffert, SM.; Lachs, L.; Hernández, LR. Research on spoken language processing progress report. Bloomington, IN: Speech Research Laboratory, Indiana University; 1996. The Hoosier Audiovisual Multitalker Database; p. 578-583.
- Smith ZM, Delgutte B, Oxenham AJ. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*. 2002 Mar 7;416:87–90. [PubMed: 11882898]
- Sumby WH, Pollack I. Visual contribution of speech intelligibility in noise. *Journal of the Acoustical Society of America*. 1954; 26:212–215.
- Summerfield, Q. Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd, B.; Campbell, R., editors. *Hearing by eye: The psychology of lip-reading*. Hillsdale, NJ: Erlbaum; 1987. p. 3-51.
- Vatikiotis-Bateson, E.; Munhall, KG.; Hirayama, M.; Lee, YV.; Terzopoulos, D. The dynamics of audiovisual behavior in speech. In: Stork, DG.; Hennecke, ME., editors. *Speechreading by humans and machines*. Berlin, Germany: Springer-Verlag; 1997. p. 221-232.
- Vatikiotis-Bateson, E.; Munhall, KG.; Kasahara, Y.; Garcia, F.; Yehia, H. Characterizing audiovisual information during speech. Paper presented at the International Conference on Spoken Language Processing; Philadelphia, PA. 1996 Oct.

- Verbrugge RR, Strange W, Shankweiler DP, Edman TR. What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*. 1976; 60:198–212. [PubMed: 956527]
- Yehia H, Kuratate T, Vatikiotis-Bateson E. Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*. 2002; 30:555–568.

Figure 1.

Schematic of the cross-modal matching task. The top row illustrates the task in the V–A direction (participants saw visual-alone movie clip then had to choose between two auditory-alone clips). The bottom row illustrates the task in the A–V direction (participants heard auditory-alone movie clip then had to choose between two visual-alone clips). Faces denote stimuli that are presented visual alone. Waveforms denote stimuli that are presented auditory alone. Δt refers to change in time and is always 500 ms.

Figure 2.
Wide band spectrogram of Talker M2 uttering the word *wail* with no acoustic transformation.

Figure 3.

A: M2 uttering the word *wail* after a spectral shift transformation at 100 Hz. B: The same token after a spectral shift transformation at 250 Hz. C: The same token after a spectral shift transformation at 500 Hz. D: The same token after a spectral shift transformation at 1000 Hz.

Figure 4.

A: M2 uttering the word *wail* after a linear spectral-scaling transformation at 75%. B: The same token after a linear spectral-scaling transformation at 150%.

Figure 5.

A: M2 uttering the word *wail* after a nonlinear spectral-scaling transformation toward the Nyquist frequency. B: The same token after a nonlinear spectral-scaling transformation away from the Nyquist frequency.

Figure 6.

A: M2 uttering the word *wail* after a spectral inversion transformation around a center frequency of 1500 Hz. B: The same token after a spectral inversion transformation around a center frequency of 2500 Hz.

Figure 7.

Examples of the linear temporal-scaling transformation. A: The untransformed acoustic signal of M2 uttering the word *wail*. B: The same signal after linear temporal scaling at 75% (speeded up). C: M2 uttering the word *wail* after linear temporal scaling at 150% (slowed down).

Figure 8.

Examples of the nonlinear temporal-scaling transformation. A: The untransformed acoustic signal of M2 uttering the word *wail*. B: The same signal after decelerating nonlinear temporal scaling. C: M2 uttering the word *wail* after accelerating nonlinear temporal scaling.

Figure 9.

Frequency transformation maps for the various transformations. The solid lines in Panels A–C show the frequency map for the untransformed condition. A: Frequency shifting. Dashed lines indicate the various levels of shift used. B: Linear frequency scaling. The dotted line shows the map used for linear scaling at 150%. The dot–dash line shows linear scaling at 75%. C: Nonlinear frequency scaling. The dotted line shows the map used for nonlinear stretching toward the Nyquist frequency. The dot–dash line shows the map used for nonlinear stretching toward low frequencies. D: Frequency rotation. The dashed line shows the frequency map for inversion around a center frequency of 1500 Hz. The solid line shows the frequency map for inversion about a center frequency of 2500 Hz. The dotted line shows the frequency map for the untransformed condition. All rotated stimuli were low-pass filtered at twice the center frequency.

Figure 10.

Results of the experiments with frequency-shifted stimuli. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the error bars represent the range from the highest to the lowest score within the sample, excluding outliers. Left panel: Boxplots of performance on the cross-modal matching task. The circles in the cross-modal 100-Hz group indicate 2 outlier participants. The dot-dash line represents chance performance. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of .05. Right panel: Boxplots of word identification performance.

Figure 11.

Results of experiments with linear frequency scaling. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the error bars represent the range from the highest to the lowest score within the sample, excluding outliers. Left panel: Boxplots of performance on the cross-modal matching task. The dot-dash line represents chance performance. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of .05. Right panel: Word identification performance when the acoustic stimuli were transformed with linear frequency scaling.

Figure 12.

Results from experiments with nonlinear frequency scaling. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the error bars represent the range from the highest to the lowest score within the sample, excluding outliers. Left panel: Boxplots of performance on the cross-modal matching task. The dot-dash line represents chance performance. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of .05. Right panel: Boxplots of word identification performance.

Figure 13. Marginal means for the interaction between scaling direction and visual intelligibility found on the cross-modal matching task with nonlinear frequency-scaled acoustic stimuli. The solid line shows chance performance. Error bars show standard errors.

Figure 14.

Results from experiments with frequency rotation. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the error bars represent the range from the highest to the lowest score within the sample, excluding outliers. Left panel: Boxplots of performance on the cross-modal matching task. The dot-dash line represents chance performance. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of .05. Right panel: Boxplots of word identification performance. The circle represents an outlier for the group, and the asterisk represents an extreme outlier for the group. Note that the scale of the dependent axis ends at 0.1.

Figure 15.

Results from experiments with linear temporal scaling. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the error bars represent the range from the highest to the lowest score within the sample, excluding outliers. Left panel: Boxplots of performance on the cross-modal matching task with acoustic stimuli transformed with linear temporal scaling. The circle shows an outlier, and the asterisk represents an extreme outlier. The dot-dash line represents chance performance. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of .05. Right panel: Boxplots of word identification performance.

Figure 16.

Results of experiments with nonlinear temporal scaling. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the error bars represent the range from the highest to the lowest score within the sample, excluding outliers. Left panel: Boxplots of performance on the cross-modal matching task with acoustic stimuli transformed with nonlinear temporal scaling. The dot-dash line represents chance performance. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of .05. Right panel: Boxplots of word identification performance.

Table 1

Descriptive Statistics for Participants Who Made Cross-Modal Matching Judgments Using Frequency-Shifted Acoustic Stimuli

Order	Shift magnitude	<i>M</i>	<i>SE</i>	<i>t</i> vs. 0.5
A–V	100 Hz	0.572	0.026	2.79 *
	250 Hz	0.579	0.021	3.84 *
	500 Hz	0.556	0.022	2.54 *
	1 kHz	0.581	0.031	2.64 *
V–A	100 Hz	0.595	0.023	4.15 *
	250 Hz	0.598	0.025	4.00 *
	500 Hz	0.569	0.015	4.53 *
	1 kHz	0.535	0.018	1.98 *

Note. A–V: participants heard auditory-alone movie clip then had to choose between two visual-alone clips; V–A: participants saw visual-alone movie clip then had to choose between two auditory-alone clips.

* $p < .05$.

Table 2

Descriptive Statistics for Participants Who Made Cross-Modal Matching Judgments Using Linear-Frequency-Scaled Acoustic Stimuli

Order	Scale magnitude (%)	<i>M</i>	<i>SE</i>	<i>t</i> vs. 0.5
A–V	75	0.532	0.035	2.93 *
	150	0.601	0.088	3.65 *
V–A	75	0.553	0.060	2.81 *
	150	0.602	0.084	3.84 *

Note. A–V: participants heard auditory-alone movie clip then had to choose between two visual-alone clips; V–A: participants saw visual-alone movie clip then had to choose between two auditory-alone clips.

* $p < .05$.

Table 3

Descriptive Statistics for Participants Who Made Cross-Modal Matching Judgments With Acoustic Stimuli That Had Been Nonlinearly Scaled in the Frequency Domain

Order	Scaling direction	<i>M</i>	<i>SE</i>	<i>t</i> vs. 0.5
A–V	Toward Nyquist	0.494	0.015	–0.43
	Away from Nyquist	0.525	0.014	1.84 *
V–A	Toward Nyquist	0.518	0.014	1.29
	Away from Nyquist	0.552	0.019	3.84 *

Note. A–V: participants heard auditory-alone movie clip then had to choose between two visual-alone clips; V–A: participants saw visual-alone movie clip then had to choose between two auditory-alone clips.

*
 $p < .05$.

Table 4

Descriptive Statistics for Participants Who Made Cross-Modal Matching Judgments With Spectrally Rotated Acoustic Stimuli

Order	Center frequency (Hz)	<i>M</i>	<i>SE</i>	<i>t</i> vs. 0.5
A–V	1500	0.507	0.012	0.62
	2500	0.483	0.033	–0.50
V–A	1500	0.524	0.017	1.46
	2500	0.473	0.038	–0.72

Note. All *t* tests failed to reach significance ($\alpha = .05$). A–V: participants heard auditory-alone movie clip then had to choose between two visual-alone clips; V–A: participants saw visual-alone movie clip then had to choose between two auditory-alone clips.

Table 5

Descriptive Statistics for Participants Who Made Cross-Modal Matching Judgments With Acoustic Stimuli That Had Been Transformed With Linear Temporal Scaling

Order	Scaling magnitude (%)	<i>M</i>	<i>SE</i>	<i>t</i> vs. 0.5
A–V	75	0.618	0.020	6.02 [*]
	150	0.552	0.016	3.26 [*]
V–A	75	0.578	0.028	2.82 [*]
	150	0.585	0.024	3.47 [*]

Note. The scaling magnitude corresponds to the transformed duration of the utterance. Thus, acoustic stimuli transformed under the 75% scaling magnitude were shortened to 75% of their original duration. Likewise, the 150% scaling magnitude corresponded to a lengthening of the overall duration. A–V: participants heard auditory-alone movie clip then had to choose between two visual-alone clips; V–A: participants saw visual-alone movie clip then had to choose between two auditory-alone clips.

^{*}
p < .05.

Table 6

Descriptive Statistics for Participants Who Made Cross-Modal Matching Judgments With Acoustic Stimuli That Had Been Transformed With Nonlinear Temporal Scaling

Order	Scaling direction	<i>M</i>	<i>SE</i>	<i>t</i> vs. 0.5
A–V	Decelerated	0.517	0.010	1.52
	Accelerated	0.509	0.011	0.90
V–A	Decelerated	0.531	0.020	2.46 [*]
	Accelerated	0.544	0.013	2.25 [*]

Note. A–V: participants heard auditory-alone movie clip then had to choose between two visual-alone clips; V–A: participants saw visual-alone movie clip then had to choose between two auditory-alone clips.

^{*}
p < .05.

Table 7

Descriptive Statistics for the Follow-Up Participants Who Performed the Word Identification Task in All 14 Transformation Conditions

Transformation	<i>M</i>	<i>SE</i>
Frequency shift		
100 Hz	.870	.015
250 Hz	.840	.014
500 Hz	.748	.079
1 kHz	.525	.082
Linear Frequency Scale		
75%	.865	.003
150%	.743	.092
Nonlinear Frequency Scale		
Toward Nyquist	.053	.019
Away from Nyquist	.760	.034
Frequency inversion		
1500 Hz	.002	.002
2500 Hz	.013	.013
Linear Temporal Scale		
75%	.865	.010
150%	.863	.008
Nonlinear Temporal Scale		
Decelerated	.830	.015
Accelerated	.873	.008

Note. The mean and standard error are expressed in terms of proportion correct.