Published in final edited form as:

Ecol Psychol. 2004 ; 16(3): 159–187. doi:10.1207/s15326969eco1603_1.

Crossmodal Source Identification in Speech Perception

Lorin Lachs and

Department of Psychology California State University, Fresno

David B. Pisoni

Department of Psychology Indiana University, and Department of Otolaryngology, Head, and Neck Surgery Indiana University School of Medicine

Abstract

Four experiments examined the nature of multisensory speech information. In Experiment 1, participants were asked to match heard voices with dynamic visual-alone video clips of speakers' articulating faces. This cross-modal matching task was used to examine whether vocal source matching can be accomplished across sensory modalities. The results showed that observers could match speaking faces and voices, indicating that information about the speaker was available for cross-modal comparisons. In a series of follow-up experiments, several stimulus manipulations were used to determine some of the critical acoustic and optic patterns necessary for specifying cross-modal source information. The results showed that cross-modal source information was not available in static visual displays of faces and was not contingent on a prominent acoustic cue to vocal identity (f0). Furthermore, cross-modal matching was not possible when the acoustic signal was temporally reversed.

Research on audiovisual speech perception has demonstrated that the domain of speech perception is not limited to the auditory sensory modality. The visual correlates of speech can be perceived accurately by adults (Berger, 1972; Bernstein, Demorest, & Tucker, 2000; Campbell & Dodd, 1980; Jeffers, 1971; Walden, Prosek, Montgomery, Scherr, & Jones, 1977) and children (Erber, 1972, 1974). Furthermore, auditory and visual stimuli can combine to elicit illusory perceptions. In a classic demonstration of this effect, McGurk and MacDonald (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976) combined the auditory form of a person saying /baba/ with the visual form of the same person saying /gaga/. When asked to identify the multimodal stimulus display, 98% of subjects responded /dada/, indicating that the different sources of information from the two sensory modalities were integrated at some point during the process of speech perception. This effect has been replicated many times and under many circumstances (see Massaro, 1998).

More practically, visual information about speech has also been shown to enhance auditory speech perception in noise (Erber, 1969; Middleweerd & Plomp, 1987). In their pioneering study, Sumby and Pollack (1954) found that the addition of visual information about articulation to an auditory signal can improve speech intelligibility performance in noise; these gains were equal to a +15-dB gain in signal-to-noise (S/N) ratio under auditory-alone conditions (MacLeod & Summerfield, 1987; Summerfield, 1987). They found that absolute gains in speech perception accuracy were most dramatic at S/N ratios where auditory-alone performance was low. However, when the gains were expressed relative to possible improvement over auditory-alone performance, the contribution of visual information to

speech perception accuracy remained constant over the entire range of S/N ratios tested (from -30 dB to 0 dB). In addition, Reisberg, McLean, and Goldberg (1987) showed that concurrently presented visual information facilitated the repetition of foreign-accented speech and semantically complex sentences. These findings demonstrate that visual information about speech is useful and informative and is not simply compensatory in situations where auditory information is insufficient to support perception.

The discovery of these “audiovisual speech phenomena” has raised several general theoretical questions about the domain of speech perception (Bernstein et al., 2000). Clearly, any comprehensive theory of speech perception must be able to explain the utility and importance of visual speech information (Summerfield, 1987). In an effort to construct such a theory, investigators have compiled a large and growing body of work concerning the nature of the phonetic information contained in the visual signal (Brancazio, Miller, & Paré, 1999; Green & Kuhl, 1991; Green & Miller, 1985; Jordan & Bevan, 1997; Jordan, McCotter, & Thomas, 2000; Kanzaki & Campbell, 1999). Frequently, these studies use susceptibility to the McGurk effect or degree of auditory enhancement (as in Sumbly & Pollack, 1954) as their dependent variable. Based on these results, investigators have drawn several general conclusions about the nature of visual speech information and how it combines with auditory speech information during the process of speech perception.

Several models of audiovisual speech perception, however, specifically incorporate assumptions about the independence of information arriving from disparate modalities (Braida, 1991; Massaro, 1998). The job of the perceptual system, on these views, is therefore to assemble the independent signals into a coherent, multimodal perceptual object. For example, one account of audiovisual integration, the Fuzzy Logical Model of Perception (FLMP; Massaro, 1998; Massaro & Cohen, 1995), relies on a priori assumptions that the perceptual system has tacit knowledge of the relations that exist across sensory modalities, by virtue of audiovisual representations of speech sounds stored in memory. Speech information is assessed by determining the presence or absence of visual or auditory features independently. The continuously valued features obtained by sensory stimulation are then compared to stored feature templates that contain both auditory and visual features, and the best matching template is selected for perception. According to Massaro, auditory and visual information are only linked via their combination in stored multimodal representations. Under the FLMP's formalization, the objective of the speech perception system is simply to recover the informative aspects of the auditory and visual signals independently, so that they can be integrated at some later point in the process.

However, the auditory and visual properties of speech are not independent. Despite the incontrovertible body of evidence generated using the McGurk paradigm, it should be emphasized here that the McGurk effect is based on an illusion, created artificially in the laboratory to demonstrate the role of audiovisual integration during the process of speech perception. Under normal everyday listening conditions, a perceiver is never confronted with a situation in which a single talker produces speech that generates conflicting patterns of acoustic and optic energy.

Lawful relations exist between acoustic and optic displays that can potentially be useful for the process of speech perception. For example, Yehia, Rubin, and Vatikiotis-Bateson (1998) presented a detailed analysis of facial motion and its relation to vocal-tract motion and the acoustical structure of spoken events. Their analysis showed that a great deal of the variability in facial motion can be accounted for by the concomitant motion of the vocal tract. Conversely, variation in vocal-tract motion can be predicted by facial motion. Furthermore, both sources of information can account for variation in acoustic parameters associated with speech utterances. Remarkably, Yehia, Kuratate, and Vatikiotis-Bateson

(2002) used a subset of these acoustic parameters to drive an animated talking head. In another example, regions of peak amplitude in auditory speech filtered in the F2 region are strongly correlated with the area function of lip opening, available visually (Grant & Seitz, 2000). This relation appears to be useful for detecting auditory speech in noise; when presented along with visual displays of articulation, auditory-speech-detection thresholds improve by as much as +2.2 dB (Grant, 2001; Grant & Seitz, 2000).

Even infants as young as 4 months of age are sensitive to the relations between simultaneously presented auditory and visual information for many natural events. In one investigation of cross-modal sensitivity in infants, Spelke (1976) presented 4-month-olds with two visual films of different events, while simultaneously playing the sound track of one of the films through a central speaker. The events displayed were either a woman playing “peekaboo,” or a woman rhythmically beating a tambourine and wood-block with a stick. Spelke found that infants fixated their gaze on the visual display that matched the auditory display more often than on the visual display that did not correspond to the auditory display. Furthermore, infant sensitivity to cross-modal structure is not limited to general events, but extends to the speech domain. Kuhl and Meltzoff (1984) showed that, given two visual displays of a talker articulating a vowel, infants looked longer at the display that matched the phonetic content of a simultaneously presented auditory vowel than at the display that contained a mismatch.

These observations provide support for a different approach to audiovisual integration: an approach that has had profound consequences for the way we conceptualize audiovisual speech information and the process by which the information from two sensory modalities are “integrated.” By acknowledging the fact that auditory and visual speech are generated by a common source—the talker engaged in the act of speaking—the locus of audiovisual integration moves from inside the head to outside of it, into the real world. This approach is compatible with a direct realist view of perceptual systems (Gibson, 1966), in which the object of perception is not the pattern of sensory stimulation impinging on the eyes or ears, but is rather the event in the real world that initially shaped the pattern of stimulation. Although much work from this approach has concentrated on the visual perception of events, Gaver (1993) extended the idea conceptually to the domain of auditory perception. The structure of acoustic energy produced during a sound-making event is lawfully tied to the event that produced it. Gaver claimed that the human auditory system may be structured in a way to exploit these relations. This approach has also been applied to the study of speech and speech perception in the direct realist theory of speech perception (Fowler, 1986, 1996; Fowler & Rosenblum, 1991).

According to this view of speech perception, acoustic and optical displays of speech are integrated because they are simply two different sources of information about the same distal event. The visual display transmits information about the event in one way, and the acoustic display transmits information about the event in another way. Nevertheless, the object of perception remains the same—the underlying articulatory event that produced the speech being perceived. Information, therefore, is said to be *amodal*; that is, it is not specific to transmission through any one particular sensory medium (Fowler, 1986).

Confirmation of the amodal nature of phonetic information in speech has been obtained in several experiments over the last few years. Fowler and Dekle (1991), for example, had naive subjects listen to spoken syllables while using their hands to obtain information about the articulation of another syllable, in much the same way that deaf-blind users of the Tadoma method (Schultz, Norton, Conway-Fithian, & Reed, 1984) do. Because incongruent information about speech was perceived simultaneously through disparate sensory modalities, this procedure can be viewed as another kind of McGurk stimulus, albeit

involving different sensory modalities (auditory and tactile vs. auditory and visual). Fowler and Dekle found that even the *haptic* specification of a spoken utterance (as obtained via Tadoma) was enough to influence the perception of the auditory signal. Furthermore, this McGurk effect was found with naive subjects who had no training in Tadoma at all; Fowler and Dekle interpreted their findings as evidence that the ability to “integrate” information about speech is not based on matching features to learned representations, but is rather based on the detection of amodal information about speech articulation that is carried in different energy patterns.

In a similar study, Bernstein, Demorest, Coulter, and O'Connell (1991) examined the lipreading performance of an observer who was wearing a vibrotactile vocoder, a device that transmits information about the auditory frequency spectrum over time via a band of mechanical stimulators that rest against the skin on the forearm. Bernstein et al. (1991) found that several naive normal-hearing and hearing-impaired participants were able to use the vocoded tactile information together with optical information to significantly improve lipreading accuracy above baseline visual-alone scores. The results reported by Fowler et al. and Bernstein et al. (1991) clearly demonstrate that the information needed for audiovisual speech perception is not tied to the auditory, or even the visual, modality alone. Rather, the sensory patterns of auditory or visual stimulation convey information about the same underlying articulatory events occurring in the world—the articulatory motion of the vocal tract.

EXPERIMENT 1: CROSS-MODAL MATCHING OF AUDITORY AND VISUAL PATTERNS

Because the object of perception—the dynamics of articulation—is assumed to be the same, regardless of the particular sensory domain being used, the modality through which a perceiver makes a judgment should be to some extent irrelevant (barring, of course, the idiosyncratic ways in which particular sensory domains carry information about the relevant dynamics of the event to be perceived). Perhaps counterintuitively, this theoretical standpoint predicts that a perceiver should be able to match a display of a particular event in one sensory modality with a display of the same event in another sensory modality, even though the specific sensory patterns of optic or acoustic energy are never experienced by the perceiver twice. As shown by Spelke's (1976) findings, infants are sensitive to the correspondences across sensory modalities. Can adult perceivers use these same correspondences to explicitly match events across modalities? We tested this question using a cross-modal matching task.

The cross-modal matching task (Lachs, 1999) is a two-alternative forced-choice procedure designed to test the theoretical prediction that modality-neutral information is available in both optical and acoustic displays of speech, and that such information can be used to match speech events presented to different sensory modalities (see Figure 1). In the task, a participant is presented with the auditory or visual form of a particular talker speaking an isolated English word. The target stimulus (or “test pattern”) is presented to only one sensory modality. After the test pattern is presented, the participant is presented with two response alternatives in another sensory modality. One of the two alternatives is the same event that generated the test pattern, presented in a different modality. Thus, if the test pattern is visual-alone, then the response alternatives are auditory-alone (the “V-A order”). The task can also be carried out in the reverse order, where the test pattern is acoustic, and the response alternatives are optical (the “A-V order”). The participant's task is to choose the alternative that was generated by the same event as the test pattern (i.e., to choose the target alternative).

Figure 1 shows a schematic representation of the cross-modal matching task. In the top row of the figure, which illustrates the “V -A order”, the test pattern contains the *visual* form of Bob saying the word “cat”. The target alternative is the *auditory* form of Bob saying the word “cat”. The distracter alternative is the auditory form of a *different talker* (John) saying the *same* word (“cat”). The participant's task in this procedure is to choose the response alternative that matches the test pattern presented. The bottom row shows an example of the task carried out in the “A-V order”.

Experiment 1 was designed to assess participants' ability to match speaking faces and speaking voices across sensory modalities. If perceivers are indeed sensitive to the modality-neutral form of phonetic information, then participants should be able to perform the cross-modal matching task at above chance levels of performance.

Method

Experimental design—Two factors may play a role in the perceiver's ability to perform the matching task successfully above chance. The first is the order in which the patterns are presented. It might be the case that seeing a face and then judging which of two voices (V-A) matched it is easier than the converse situation: hearing a voice and judging which of two faces matched it (A-V). For example, sensory memory for the fine-grained details of an utterance may be more robust for auditory speech than for visual speech, enabling easier comparisons of two acoustic alternatives. On the other hand, there might be an advantage when the target stimulus is auditory because comparisons to the target stimulus are made throughout the duration of the trial, which can last up to 3 sec. If the target is auditory, and auditory details are preserved for comparison longer than visual details, then there might be an advantage for A-V trials over V-A trials. To assess any differences based on these factors, both conditions were tested in this experiment.

In addition, it is possible that fine-grained details of the stimulus pattern will be lost if the stimulus is unintelligible in one or the other modality. To test this possibility, stimulus items were balanced for their intelligibility. Because the stimulus items used in this study were all highly intelligible under V-A identification tests, stimulus items were split into low and high groups based on their visual intelligibility using data from A-V identification tests (Lachs & Hernández, 1998).

A two-alternative forced-choice matching procedure was used in a 2×2 factorial design. The two levels for the between-subject “order” factor were “A-V” (where participants identified the correct visual stimulus after viewing the test auditory stimulus) and “V-A” (where participants identified the correct auditory stimulus after viewing the test visual stimulus). The two levels of the within-subjects “visual intelligibility” factor were “low” and “high.” 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.

Participants—Participants were 20 undergraduate students enrolled in an introductory psychology course at Indiana University 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.

Stimulus materials—Four Apple Macintosh G4 computers, each equipped with a 17-inch Sony Trinitron Monitor (0.26-mm dot pitch) were used to present the visual stimuli to subjects. The stimuli consisted of a set of 96 tokens selected from a previously recorded audiovisual database used for multimodal experiments (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 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 DT100 headphones at 74 dB SPL.

Procedures—Participants in the V-A condition were first presented with a A-V movie clip of a talker uttering an isolated, English word. Shortly after seeing this video display (500 msec), they were presented with two acoustic signals. One of the signals was the same talker they had seen in the video, whereas the other signal was a different talker. Participants were instructed to choose which acoustic signal matched the talker they had seen (“First” or “Second”). Similar instructions were provided for participants in the A-V condition, who heard an acoustic signal first, and then had to make their decision based on two video displays.

On each trial, the test stimulus was either the video or audio portion of one movie token. Each movie token displayed an isolated word spoken by a single talker. The order in which the target and distracter 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 contrasted with “John,” regardless of whether “Bob” or “John” was the target alternative 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 recorded by pressing one of two buttons on a response box and were entered into a log file for further analysis.

A short training period (eight trials) preceded each participant's session. During the training period, the participant was presented with a cross-modal matching trial and asked to pick the correct alternative. During training only, the response was followed with feedback. The feedback provided was a presentation of the combined audiovisual movie display of the target. 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.

Results

Figure 2 shows box plots of the performance scores obtained in Experiment 1. Each shaded box represents the interquartile range for the observed data, the bold line within each box represents the median score for the group, and the whiskers show the range from the lowest to the highest score in the group, excluding outliers. The box on the left represents the participants in the A-V group and the box on the right represents the V-A group. As shown, the majority of the participants were able to perform the matching task above chance, regardless of the presentation order. A one sample *t* test showed that average performance on this task was significantly greater than chance performance (0.50) for the A-V group ($M = 0.607$, $SE = 0.035$, $t(9) = 3.06$, $p < 0.01$) and for the V-A group ($M = 0.651$, $SE = 0.017$, $t(9) = 8.75$, $p < 0.001$).

The results were also submitted to a two-way (Visual intelligibility $\times$ Order) analysis of variance (ANOVA) to examine any effects of the manipulated factors. The ANOVA showed

only a marginal effect of visual intelligibility, $F(1, 18) = 3.094, p < 0.10$. Performance on high visual intelligibility words ($M = 0.643, SE = 0.021$) was better than performance on low visual intelligibility words ($M = 0.616, SE = 0.021$). There was no effect of presentation order, $F(1, 18) = 1.25, ns$, and no interaction between presentation order and visual intelligibility, $F(1, 18) < 1, ns$.

Discussion

The results from the cross-modal matching experiment indicate that participants are able to make use of information about a spoken event in one modality and use it to match the specification of the same event in another modality. On average, participants were quite successful at making cross-modal matching judgments above chance. One surprising result was the absence of asymmetries in performance as a function of the order in which the matching judgments were made. Because there are differences in the extent to which acoustic and optic displays carry information about the motion of the vocal articulators, this result deserves further study. The acoustic form of speech can carry information about the positions and movements of the vocal articulators from the lips to the larynx (Fant, 1960). However, the same is not true for the optic form of speech. Visual displays of speech can carry reliable information about the configuration of the lips, tongue tip, and jaw, but it is unlikely that they can carry information about the configuration of the velum, or show that there is a closure near the glottis (Dodd & Campbell, 1987; Summerfield, 1987). Despite these differences, however, recent findings on speechreading have shown that the visual signal is not as impoverished with respect to speech perception as previously thought (Bernstein, Auer, & Moore, in press; Bernstein et al., 2000; Lachs & Pisoni, 2004).

We also found a marginal effect of visual intelligibility indicating that it may have been easier to make cross-modal judgments when the word itself was more visually intelligible. Although the effect was marginal, this relation is of some interest because the cross-modal matching task does not require participants to recognize the words or make explicit judgments of word identity: Every stimulus presented during a given trial contains the same word. It is not clear why an increased ability to identify a word would lead to an increased ability to discriminate between cross-modal alternatives. One possibility is that the ability to identify the word allows for a more fine-grained discrimination (visually) of the idiosyncratic speaking style of a particular talker, leading to an enhanced ability to discriminate the response alternatives. Another possibility is that the cross-modal information specifying talker identity is inextricable from the phonetic information specifying word identity. This possibility is supported by recent evidence suggesting a close association between linguistic and indexical properties of auditory speech (Bradlow, Torretta, & Pisoni, 1996; Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989).

EXPERIMENT 2: STATIC FACES

The results of Experiment 1 indicate that sufficient information is present in the visual and auditory speech signal to allow participants to make reliable comparisons of talker identity across sensory modalities. The cross-modal information that supports these comparisons must be specified in both the auditory and visual signals or else such comparisons could not be made above chance. From a direct realist perspective, cross-modal information arises because the optic and acoustic specifications of phonetic events are lawfully tied to one another by virtue of their common origin in a single articulatory event.

However, a far simpler explanation of cross-modal matching abilities exists. It may be that participants make cross-modal matching judgments based on the expectation that specific faces should be paired with specific voices. This explanation of the results is certainly plausible. It is possible that static facial features related to identity (e.g., relative age, relative

size, hair style) set up cognitive expectations about the tone, accent, or speaking style of the talker. Such a strategy is clearly distinct from making judgments based on the cross-modal specification of identical phonetic events.

To test whether participants relied on cognitive strategies for relating *static* features of the visual display with expectations about voice characteristics, Experiment 2 tested participants' ability to make cross-modal matching judgments with static pictures of faces as visual displays. By definition, a static display of a face eliminates all optical information about the movement of speech articulators over time. If participants are able to match static pictures with voices, then cross-modal information for matching faces and voices must be contained in static features of faces. However, if the ability to match faces and voices is eliminated by the use of static pictures, then it can be concluded that cross-modal information about the time-varying acoustic structure of speech is not presented in static displays of faces.

Method

Participants—Participants were 20 undergraduate students enrolled in an introductory psychology course 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.

Stimulus materials—The presentation equipment and display parameters in this experiment were identical to those reported previously in the Methods section for Experiment 1. The stimulus materials were also taken from the same set as those used in the experiments previously conducted. However, the video portion of the stimulus was no longer a dynamic stimulus pattern that changed over time. Instead, the visual displays were a single static frame taken from the original movie clip. Because the static frame was taken from the original movie, the appearance of a given talker differed slightly on each presentation. This provided each participant with multiple, static views of the same face over the entire experiment. Each static frame was displayed on the computer monitor for the duration of the original video track.

Procedures—The procedures used in this experiment were identical to those used in Experiment 1, with the exception that visual displays were static pictures, rather than dynamic video clips.

Results

Figure 3 shows box plots of performance in Experiment 2 for the A-V and V-A groups separately. As shown in the figure, participants performed very poorly when asked to match static pictures of faces with voices. Overall, the group's performance did not differ significantly from chance (0.50), $t(19) = 1.06$, *ns*. This was true for the A-V group, $M = 0.504$, $SE = 0.038$, $t(9) < 1$, *ns*, as well as the V-A group, $M = 0.546$, $SE = 0.029$, $t(9) = 1.29$, *ns*, when analyzed separately.

A two-way ANOVA revealed no main effects or interactions due to the visual intelligibility and order factors.

Discussion

The results of Experiment 2 show that participants could not make cross-modal matching judgments between faces and voices when the faces were static pictures of the original

talkers. This finding rules out the hypothesis that cross-modal matching judgments are made based on cognitive strategies or expectations set up by static features of appearance in the visual display. When the dynamic structure of visual speech was eliminated from the visual display, participants were no longer able to perform the matching task above chance.

EXPERIMENT 3: NOISE-BAND STIMULI

Another source of information that could be used to set up cognitive strategies about the correspondence of particular voices and faces are traditional cues to vocal identity, such as fundamental frequency (f_0). The fundamental frequency of an utterance is the frequency of vibration of the vocal folds, the harmonics of which are selectively amplified or attenuated by the shape of the vocal tract as it is deformed by the movements of the vocal articulators over time (Ladefoged, 1996). F_0 varies substantially among talkers, especially across gender. It is possible that the pitch of a talker's voice, his or her inflection, or even his or her prosody could be used as a reliable cue for distinguishing among talkers. For example, a participant in the cross-modal matching task may make the decision that all "deep" voices should be matched to older or bigger males, or that rising prosodic contours should be matched with visual displays in which the eyebrows move up. Such strategies would have less to do with detection of cross-modal information and intersensory correspondences about articulation and more to do with expectations set up by social norms, prior experience, and so forth.

To test whether f_0 plays a role in the judgment of cross-modal correspondences, Experiment 3 used noise-band "chimaeric" speech (Smith, Delgutte, & Oxenham, 2002) as an acoustic stimulus. To make noise-band speech, the acoustic waveform is filtered with a number of evenly spaced band-pass filters. The output of each filter is then subjected to the Hilbert transform, which separates an acoustic waveform into two parts: the rapidly varying fine structure (i.e., the source) and the slower-changing envelope that modulates the fine structure. The envelope from each filter is then used to modulate white noise. Finally, each filtered channel of white noise is reassembled into one "chimaeric" stimulus. The resulting pattern incorporates the fine structure of white noise and the envelope modulation of speech. The result is a stimulus pattern that can be understood as speech; with 16 or more channels, sentence recognition for these types of stimuli was close to 100% (Smith et al., 2002). However, the noise-band stimulus does not contain any of the superficial acoustic characteristics of the original vocal source (i.e., vocal fold vibration or f_0). Rather, it can be thought of as preserving only the vocal-tract transfer function, as it evolves over time, by exciting it with a completely novel voicing source (e.g., white noise).

Noise-band filtering of speech makes it possible to test the hypothesis that the observed cross-modal matching judgments are based on cognitive strategies and expectations set up by traditional cues to vocal identity, such as f_0 . If participants are still able to match auditory and visual displays of speech when the auditory stimuli are noise-band stimuli, then we can conclude that f_0 is not a reliable cue for making cross-modal matching judgments.

Method

Participants—Participants were 40 undergraduate students enrolled in an introductory psychology course 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.

Stimulus materials—The presentation equipment and display parameters in this experiment were identical to those reported previously in the Methods section for Experiment 1. The stimulus materials were also taken from the same set as those used in the experiments previously conducted. However, the auditory portion of each stimulus (that is, the audio track of the movie) was manipulated using the digital signal processing methods to create noise-band stimuli (Smith et al., 2002).

Figure 4 illustrates the noise-band stimulus creation process. First, the audio track of each movie was band-pass filtered between 80 Hz and 8820 Hz with a number of channels that were equally spaced on a cochlear (nearly logarithmic) scale. The overlap of adjacent filters was set to 25% the width of the lowest frequency channel. Two sets of noise-band stimuli were made: one with 16 channels and one with 32 channels. After band-pass filtering, the amplitude envelope of each resulting channel was extracted from the fine structure using the Hilbert transform (see Smith et al., 2002, for details). Finally, the channels were summed together, yielding the final, noise-band stimulus. Figure 5 shows an example of an untransformed auditory token and a 16-channel noise-band stimulus created from it.

After undergoing noise-band transformation, the audio tracks were dubbed back on to the video tracks of their original movies. The resulting movie resembled the original in all ways, except that the sound track was a noise-band “chimaeric” stimulus.

Procedures—The task procedures used in this experiment were identical to those used in Experiment 1, with the exception that the auditory stimuli were transformed into noise-band stimuli. An equal number of participants made judgments with the 16-channel noise-band stimuli and with the 32-channel noise-band stimuli. The number of channels used in stimulus creation was a between-subject factor.

Results

Figure 6 shows box plots of scores for participants who made cross-modal matching judgments using noise-band auditory tokens. The figure displays participants who made judgments in the V-A presentation order separately from those who made judgments in the A-V presentation order. It also displays separately the data from participants who made judgments with 16-channel noise-band stimuli, and those who made judgments with 32-channel noise-band stimuli.

As shown in Figure 6, most participants were able to perform above chance. This was true for both groups, regardless of the number of channels used to make the noise-band stimuli, 16-channel: $t(19) = 3.98, p < 0.001$; 32-channel: $t(19) = 6.42, p < 0.001$. Within each stimulus group, participants making judgments in either direction were also significantly different from chance (see Table 1). Overall, participants were generally able to match faces and voices when the voices were transformed into noise-band stimuli.

The results were also submitted to a three-way (Number of channels $\times$ Visual intelligibility $\times$ Order) ANOVA to examine any differences in performance based on the manipulated factors. The ANOVA revealed a significant two-way interaction between number of channels and order, $F(1, 36) = 6.90, p < 0.05$. Figure 7 displays the means and standard errors for the relevant cells in this interaction. Post hoc analyses revealed that the interaction was supported by the high score obtained by the 32-channel group in the V-A order. Performance in this group was significantly greater than performance with noise-band stimuli by any other group. None of the other main effects or interactions reached significance.

Discussion

Despite the direct removal of f_0 , a traditional cue to vocal identity, participants in Experiment 3 were able to perform in the cross-modal matching task above chance levels of accuracy, indicating that participants were able to match optical and acoustic displays of speech across modalities. Although performance was only slightly above chance, it should be noted that performance in Experiment 1 under the best and most naturalistic conditions was only 0.629, very close to chance performance. Stimulus degradations such as the noise-band transformation can reasonably be expected to decrease performance due to unnatural listening conditions. However, close examination of the data presented in Figure 6 shows that the interquartile range for performance with chimaeric stimuli was entirely above chance, as in Figure 2 for the untransformed stimuli. In contrast, Figure 3 shows that the interquartile range for performance in the static faces experiment spanned across chance performance. This pattern, along with the statistical analysis, is consistent with the proposal that decreased performance with chimaeric stimuli was due to stimulus degradation, but not to an underlying inability to perform the cross-modal matching task.

The results from this experiment also showed that cross-modal matching performance was not affected by the spectral resolution of the channels used to create the noise-band tokens (both the 16- and 32-channel groups performed above chance), nor was it affected by the order in which the modalities were presented (participants in the A-V and V-A conditions for both channel resolution groups also performed above chance).

There was evidence that the 32-channel V-A task provided the easiest conditions for making cross-modal matching judgments. However, it is unclear why this might be the case. It is possible that the detailed spectral information contained in the 32-channel stimuli provided for better comparisons between the auditory response alternatives in the V-A task than 16-channel stimuli did. At the same time, this increased resolution may not have aided judgments in the A-V direction because they only served to enhance acoustic differences, not to help specify the cross-modal information more clearly. Regardless of these differences, the results demonstrate clearly that f_0 is not a necessary source of auditory information for making reliable cross-modal matching judgments between voices and faces.

Participants in this experiment were still able to match optical and acoustic patterns of speech, despite the removal of direct f_0 information. As mentioned previously, the noise-band stimulus creation process removes f_0 but preserves information in the acoustic signal that specifies the vocal-tract transfer function as it evolves over time. Detailed information about the formant resonances is preserved in a noise-band stimulus, and this source of information is apparently sufficient to support cross-modal matching judgments. As long as the acoustic signal can specify the dynamic articulations of the vocal tract and how they change over time, cross-modal information is apparently preserved and correct cross-modal matches can be made.

EXPERIMENT 4: TEMPORAL REVERSAL

The results of the three experiments reported previously demonstrate that cross-modal information about speech is available in acoustic and optical displays of a talker and is not contained in static visual features about the talker's identity, or in traditional auditory cues to vocal identity, such as f_0 . Rather, it appears that cross-modal information refers to the dynamic movements of the talker's vocal tract. These time-varying movements appear to structure acoustic and optic energy in such a way as to preserve information about their common origin.

What are the dynamic properties of the spoken event that are used for making cross-modal matching judgments? One possibility is stimulus duration. Even when different talkers say the same word, the duration of the utterance is different from token to token, due to idiosyncratic properties of the talker's speaking style, such as accent or hyperarticulation. Thus, differences in duration may serve to distinguish one talker from another during matching.

In addition, the duration of the auditory utterance must necessarily be identical to the duration of the visual utterance. The duration of an articulatory event is constant, regardless of the sensory modality that is measured. The duration, then, can be seen as a kind of amodal information about the relation between the auditory sensory stimulation and the visual sensory stimulation. As such, it is possible that participants could use this source of information to effectively match patterns across sensory modalities.

To test whether duration cues were used to match patterns across modalities, Experiment 4 manipulated the temporal patterns of the video and audio tracks of the stimulus movies. To accomplish this, all audio and video signals were simply played backward in time. Figure 8 illustrates this transformation with a sample auditory stimulus. The top panel shows the acoustic waveform of an untransformed auditory stimulus. The bottom panel shows the same waveform after it has been played backward in time. This temporal reversal transformation destroys the information necessary for accurate word recognition (Kimura & Folb, 1968), but it preserves the duration of the stimulus. If observers are able to make cross-modal judgments based on signal duration, then performance on the cross-modal matching task should not suffer as a result of the transformation.

Method

Participants—Participants were 20 undergraduate students enrolled in an introductory psychology course 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.

Stimulus materials—The presentation equipment and display parameters in this experiment were identical to those reported previously in the Methods section for Experiment 1. The stimulus materials were also taken from the same set as those used in the experiments previously conducted. However, on all trials, video and audio clips were temporally reversed.

Procedures—The procedures used in this experiment were identical to those used in Experiment 1.

Results

Figure 9 shows the box plot of performance in Experiment 4 for the A-V and V-A groups separately. The results show that participants found it extremely difficult to make cross-modal matching judgments using backward video and audio clips. Regardless of the presentation order, average performance in the group did not differ statistically from chance (0.5), $t(19) = 1.47$, *ns*. This was true separately for the A-V group, $M = 0.525$, $SE = 0.017$, $t(9) = 1.50$, *ns*, and for the V-A group, $M = 0.529$, $SE = 0.034$, $t(9) = 0.86$, *ns*, when analyzed separately.

A two-way ANOVA revealed no main effects or interactions of the visual intelligibility and order factors.

Discussion

The results from Experiment 4 show that participants were unable to match audio and video displays of speech when those displays were played backward in time. Thus, it is unlikely that duration alone was used as a cue for making cross-modal matching judgments of speech. The duration of a spoken event is the same whether measured in an auditory or visual display, and there is typically a great deal of intertalker variation in the duration of a spoken word. However, this source of information about the relation between auditory and visual displays is not useful as a cross-modal cue for matching faces and voices.

GENERAL DISCUSSION

This set of cross-modal matching experiments assessed the ability of participants to perceive and use auditory and visual information about vocal articulation to match talkers across sensory modalities. In all of these experiments, participants were presented with the unimodal form of a spoken word and were required to choose which of a pair of cross-modal tokens specified the same talker. Roughly three fourths of the participants tested were able to perform this task with better than chance performance when the visual and auditory displays were the original, unaltered, dynamic displays of speech.

The results from the four experiments demonstrate that sufficient information exists in visual and auditory displays of speech to specify their relation to each other. Furthermore, the cross-modal information that supports these matching judgments is not based in static visual features about face identity (e.g., relative age, hairstyle), but must be dynamic in nature (see also Rosenblum & Saldaña, 1996, for a discussion of the role of dynamic information in visual speech perception); in Experiment 2, static faces could not be matched with cross-modal voices. The results showed that cross-modal information is not well specified in the fine structure (f0) of an utterance because participants could match band-pass, chimaeric auditory stimuli with dynamic faces (Experiment 3). Finally, although duration cues could serve as a potential source of cross-modal information, these properties cannot be used for making cross-modal matching judgments; temporally backward speech in the visual and auditory domains could not be matched across modalities in Experiment 4. Interestingly, Kamachi, Hill, Lander, and Vatikiotis-Bateson (2003) also recently reported a similar set of experiments with similar results. In their series of experiments, it was found that cross-modal matching was possible when the visual and auditory components of a spoken sentence were played forward in time, but not when the stimuli were played backward in time. Our results differ from theirs, however, in the length of the stimuli used. Kamachi et al. used sentence-length stimuli, but we used isolated words. Our results with isolated words indicate that cross-modal source information is detectable from even short samples of speech. Furthermore, cross-modal matching with isolated words provides support for the proposition that cross-modal source information arises from very basic perceptual processes and is not necessarily dependent on the more complex linguistic processing required during sentence comprehension.

The ability to perceive the identity of the source of acoustic events has been demonstrated in several other domains in addition to speech perception. For example, in an especially clever experiment, Repp (1987) presented the sound of hand clapping for identification by participants. Some of the claps were generated by the participants themselves and the others were generated by people with whom the participants were acquainted. Perceivers performed above chance on this task, although their absolute identification accuracy was rather low (11%). In another related study, Li, Logan, and Pastore (1991) asked participants

to identify the gender of a person whom they heard walking down a hallway. Remarkably, judgments of gender were well above chance. Furthermore, anthropomorphic differences (such as weight and height) between walkers were highly correlated with judgments of gender, indicating that the acoustics generated by different body types contains reliable information that allows for the accurate perception of these attributes.

Both the Repp (1987) and Li et al. (1991) studies indicated that detailed acoustic information about sound-producing events can be perceived and used to identify the idiosyncratic minutiae associated with the person producing them. This is also true in the domain of speech perception. The subtle variations exhibited by different talkers during the process of speech production can be used to identify the specific talker uttering a speech event and appear to be integrally encoded with linguistic information (Mullennix & Pisoni, 1990; Mullennix et al., 1989). Indeed, recent findings reported by Remez and his colleagues (Fellowes, Remez, & Rubin, 1997; Remez, Fellowes, & Rubin, 1997) have provided converging evidence for the integral nature of linguistic and indexical information in speech signals using sine-wave speech replicas. Sine-wave speech (Remez, Rubin, Pisoni, & Carrell, 1981) is an acoustic transformation of speech that replicates the center frequencies of the three lowest formants with sinusoidal tones that vary in frequency over time. These sinusoidal replicas of speech therefore eliminate all traditional acoustic cues to phonetic and indexical information (Remez, Rubin, Berns, Pardo, & Lang, 1994). However, in a series of experiments, Fellowes et al. showed that sine-wave speech can support the identification of the gender and even the identity of a talker. The results of these experiments demonstrate that fine-grained differences in the speaking styles of different talkers can also be used in judgments of source variation across sensory modalities.

It is clear from these findings that there is sufficient information about the spoken event encoded in the optical or acoustic signals that allows subjects to make reliable cross-modal comparisons, even for isolated spoken words. For accurate cross-modal judgments to be made, auditory and visual information about the movement of speech articulators had to be preserved in some form. As noted previously, this is precisely the same information that has been shown to be relevant to the perception of the linguistic properties of speech (Remez et al., 1981). In fact, evidence of the ability to match speech across modalities on the basis of isolated articulatory information comes from another experiment reported by Kamachi et al. (2003). Their results showed that fully illuminated faces could be matched to sine-wave speech with an average accuracy of 61%. In another recent study, Rosenblum, Smith, Nichols, Hale and Lee (2004) also used sentence-length stimuli and showed that point-light faces (Johansson, 1973; Rosenblum & Saldaña, 1996) could be matched to untransformed speech at above chance levels of performance. Finally, Lachs and Pisoni (in press), using isolated words, reported that cross-modal matching of sine-wave speech with point-light faces is possible (although extremely difficult), indicating that isolated information about the kinematics of articulatory activity in the visual and auditory modalities is sufficient to carry cross-modal source information.

Linguistic Versus Indexical Properties of Speech

The observed relations between lexical and indexical speech information, and their common basis in vocal-tract articulation, suggests a link between cross-modal matching and word recognition. That is, if both word-identification information and cross-modal source information are supported by patterns of sensory stimulation that relate to the articulatory events that produced them, then performance on word-identification tasks should be relatively high under the same acoustic transformations that support cross-modal matching. In a supplementary study of the acoustic transformations used in this investigation, 30 additional undergraduates were recruited to serve as participants in a word-recognition experiment that used the two noise-band transformations used in Experiment 3 plus the

temporal reversal transformation used in Experiment 4. On each trial, listeners were presented with an auditory token of one talker speaking an isolated English word. After presentation of the stimulus, the participant was asked to enter the word he or she 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. After the response was entered, the next trial was presented. No feedback was given to the participants at any point in the procedure. All eight talkers were used in this study, and all participants heard all eight talkers an equal number of times. The number of presentations of a particular talker speaking a particular word was counterbalanced across participants.

Overall, the two types of acoustic transformations yielded different results. On average, only 0.93% ($SE = 0.29\%$) of the 96 backward stimuli were identified correctly. Out of the 10 participants in this condition, 3 participants identified 2 words correctly, 3 identified 1 word correctly, and the rest did not identify any words correctly. In contrast, participants who heard the noise-band stimuli had little trouble with the task, extending the earlier word-recognition findings of Smith et al. (2002) who showed that sentences transformed with the noise-band transformation were also highly intelligible. Participants who heard the 16-channel stimuli identified an average of 89.2% ($SE = 0.75\%$) of the words correctly. Participants who heard the 32-channel stimuli identified 90.8% ($SE = 0.98\%$) of the words correctly. Pairwise comparisons using the Bonferroni adjustment showed that performance in both noise-band conditions differed significantly from performance in the backward condition (both $p < 0.001$), but performance did not differ significantly based on the number of channels used in the noise-band conditions.

Taken together, the results of this supplemental study suggest that the auditory form of cross-modal source information is carried in parallel with linguistically relevant information needed for word recognition, as predicted. This finding is consistent with earlier research showing that indexical information is inextricably linked to lexical information in V-A speech (Mullennix & Pisoni, 1990). A transformation of the auditory stimulus that destroys the time-varying properties of the acoustic spectrum necessary for word identification is also likely to disrupt the ability to perceive the idiosyncratic attributes of the talker.

The results also provide support for the proposal that talker-specific, indexical information is carried in the pattern of formants as they evolve over time (Remez et al., 1997) and not necessarily in traditional speech cues to vocal identity (e.g., f_0). Noise-band stimuli substitute white noise for the fine-structure vocal-fold vibrations whose harmonics are normally amplified or attenuated by the movements of the vocal tract. As such, f_0 is stripped away from the acoustic form of the word. However, cross-modal matching of talkers is still possible using these transformed stimuli. Thus, information about the talker is still present in the pattern of harmonic modulation preserved by the transformation. Thus, both cross-modal source information and linguistically relevant phonetic information appear to be carried in parallel and encoded in the pattern of formants as they vary over time. The cross-modal matching findings, taken together with the word-recognition findings, raise intriguing possibilities about the auditory form of cross-modal source information that will be investigated further in future work.

In summary, this set of results indicates that detailed information about the vocal source of an utterance is available in both optical and acoustic displays of speech, and this information is only available in dynamic displays of speech that preserve the spectral and temporal relations among vocal-tract resonances. In addition, the phonetic information necessary for spoken word recognition that is contained in acoustic displays of speech is preserved under the same acoustic transformations that preserve cross-modal source information. The results are consistent with the theory of direct perception, which predicts that cross-modal matching

judgments of the kind reported here should be possible by virtue of the common origin of auditory and visual patterns in the articulatory movements of a particular vocal tract as they unfold over time. Both the auditory and visual forms of a phonetic event specify the same underlying dynamics of articulation, and this common origin is necessarily reflected in the patterning of acoustic and optic displays of speech. Cross-modal information about speech, on this view, arises as a direct result of the lawful structuring of optic and acoustic energies by a unitary spoken event (Gaver, 1993). As pointed out by Vatikiotis-Bateson and his colleagues, “The motor planning and execution associated with producing speech necessarily generates visual information as a by-product” (Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzepoulos, 1997, p. 221). Consequently, it is entirely possible that any information of relevance in the acoustic signal about the talker or the linguistic message is also carried, in some form, by the optical signal of speech. This investigation has extended previous findings by demonstrating that indexical information about the source of spoken events is carried in the time-varying information about the motion of the articulators. Such information appears to be modality neutral and as such can be perceived and used to make accurate judgments of identity across sensory modalities.

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 NIH–NIHCD Training Grant DC00012 and NIDCD Research Grant DC–00111 to Indiana University. We thank Luis R. Hernández, Jeff Karpicke, and Jeff Reynolds for invaluable assistance during all phases of this research. In addition, we thank Geoffrey Bingham, Robert Port, Thomas Busey, Eric Vatikiotis-Bateson, and one anonymous reviewer for their valued input and advice.

REFERENCES

- Berger, KW. Speechreading: Principles and methods. National Educational Press; Baltimore: 1972.
- Bernstein, LE.; Auer, ET., Jr.; Moore, JK. Modality-specific perception of auditory and visual speech. In: Calvert, GA.; Spence, C.; Stein, BE., editors. Handbook of multisensory processes. MA: MIT Press; Cambridge: (in press)
- Bernstein LE, Demorest ME, Coulter DC, O'Connell MP. Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. *Journal of the Acoustical Society of America*. 1991; 90:2971–2984. [PubMed: 1838561]
- Bernstein LE, Demorest ME, Tucker PE. Speech perception without hearing. *Perception & Psychophysics*. 2000; 62:233–252. [PubMed: 10723205]
- 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]
- Braida LD. Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology*. 1991; 43A:647–677. [PubMed: 1775661]
- Brancazio, L.; Miller, JL.; Paré, MA. Perceptual effects of place of articulation on voicing for audiovisually-discrepant stimuli; Paper presented at the 138th meeting of the Acoustical Society of America; Columbus, OH. November. 1999
- Campbell R, Dodd B. Hearing by eye. *Quarterly Journal of Experimental Psychology*. 1980; 32:85–99. [PubMed: 7367580]
- Dodd, BE.; Campbell, R. Hearing by eye: The psychology of lip-reading. Lawrence Erlbaum Associates, Inc.; Hillsdale, NJ: 1987.
- Erber NP. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*. 1969; 12:423–424. [PubMed: 5808871]
- Erber NP. Auditory, visual and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech, Language, and Hearing Research*. 1972; 15:413–422.
- Erber NP. Visual perception of speech by deaf children: Recent developments and continuing needs. *Journal of Speech & Hearing Disorders*. 1974; 39:178–185. [PubMed: 4596705]

- Fant, G. Acoustic theory of speech production. Mouton; The Hague, Netherlands: 1960.
- Fellowes JM, Remez RE, Rubin PE. Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics*. 1997; 59:839–849. [PubMed: 9270359]
- Fowler CA. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*. 1986; 14:3–28.
- Fowler CA. Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*. 1996; 99:1730–1741. [PubMed: 8819862]
- Fowler CA, Dekle DJ. Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception & Performance*. 1991; 17:816–828. [PubMed: 1834793]
- Fowler, CA.; Rosenblum, LD. The perception of phonetic gestures. In: Mattingly, IG.; Studdert-Kennedy, M., editors. *Modularity and the motor theory of speech perception*. Lawrence Erlbaum Associates, Inc.; Hillsdale, NJ: 1991. p. 33-59.
- Gaver WW. What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology*. 1993; 5:1–29.
- Gibson, JJ. *The senses considered as perceptual systems*. Houghton Mifflin; Boston: 1966.
- 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]
- Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*. 2000; 108:1197–1208. [PubMed: 11008820]
- Green KP, Kuhl PK. Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*. 1991; 17:278–288. [PubMed: 1826317]
- Green KP, Miller JL. On the role of visual rate information in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*. 1985; 38:269–276.
- Jeffers, J.; Barley, M. *Speechreading (Lipreading)*. Thomas; Springfield, IL: 1971.
- Johansson G. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*. 1973; 14:201–211.
- Jordan TR, Bevan K. Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance*. 1997; 23:388–403. [PubMed: 9104001]
- Jordan TR, McCotter MV, Thomas SM. Visual and audiovisual speech perception with color and gray-scale facial images. *Perception & Psychophysics*. 2000; 62:1394–1404. [PubMed: 11143451]
- Kamachi M, Hill H, Lander K, Vatikiotis-Bateson E. “Putting the face to the voice”: Matching identity across modality. *Current Biology*. 2003; 13:1709–1714. [PubMed: 14521837]
- Kanzaki, R.; Campbell, R. Effects of facial brightness reversal on visual and audiovisual speech perception; Paper presented at the Audio Visual Speech Processing Conference; University of California, Santa Cruz. August. 1999
- Kimura D, Folb S. Neural processing of backwards-speech sounds. *Science*. 1968; 161:395–396. [PubMed: 5667143]
- Kuhl PK, Meltzoff AN. The intermodal representation of speech in infants. *Infant Behavior and Development*. 1984; 7:361–381.
- Lachs, L. *Research on spoken language processing*. Vol. 23. Indiana University Speech Research Laboratory; Bloomington: 1999. A voice is a face is a voice.
- Lachs, L.; Hernández, LR. *Research on spoken language processing progress report*. Vol. 22. Speech Research Laboratory, Indiana University; Bloomington: 1998. Update: The Hoosier Audiovisual Multitalker Database; p. 377-388.
- Lachs L, Pisoni DB. Specification of crossmodal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America*. (in press).
- Lachs L, Pisoni DB. Spoken word recognition without audition. 2004 Manuscript submitted for publication.
- Ladefoged, P. *Elements of acoustic phonetics*. 2nd ed.. University of Chicago Press; Chicago: 1996.

- Li X, Logan RJ, Pastore RE. Perception of acoustic source characteristics: Walking sounds. *Journal of the Acoustical Society of America*. 1991; 90:3036–3049. [PubMed: 1787243]
- MacDonald J, McGurk H. Visual influences on speech perception processes. *Perception & Psychophysics*. 1978; 24:253–257. [PubMed: 704285]
- MacLeod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*. 1987; 21:131–141. [PubMed: 3594015]
- Massaro, DW. *Perceiving talking faces: From speech perception to a behavioral principle*. MIT Press; Cambridge, MA: 1998.
- Massaro DW, Cohen MM. Perceiving talking faces. *Current Directions in Psychological Science*. 1995; 4:104–109.
- McGurk H, MacDonald J. Hearing lips and seeing voices. *Nature*. 1976; 264:746–748. [PubMed: 1012311]
- Middleweerd MJ, Plomp R. The effect of speechreading on the speech-reception threshold in noise. *Journal of the Acoustical Society of America*. 1987; 82:2145–2147. [PubMed: 3429736]
- 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]
- Reisberg, D.; McLean, J.; Goldfield, A. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In: Dodd, B.; Campbell, R., editors. *Hearing by eye: The psychology of lip reading*. Lawrence Erlbaum Associates, Inc.; Hillsdale, NJ: 1987. p. 97–114.
- 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; 212:947–950. [PubMed: 7233191]
- Repp BH. The sound of two hands clapping: An exploratory study. *Journal of the Acoustical Society of America*. 1987; 81:1100–1109. [PubMed: 3571727]
- Rosenblum LD, Saldaña HM. An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*. 1996; 22:318–331. [PubMed: 8934846]
- Rosenblum LD, Smith NM, Nichols SM, Hale S, Lee J. Hearing a face: Cross-modal speaker matching using isolated visible speech. 2004 Manuscript submitted for publication.
- Schultz M, Norton S, Conway-Fithian S, Reed C. A survey of the use of the Tadoma method in the United States and Canada. *Volta Review*. 1984; 86:282–292.
- Sheffert, SM.; Lachs, L.; Hernández, LR. *Research on spoken language processing*. Vol. 21. Indiana University Speech Research Laboratory; Bloomington: 1996. The Hoosier Audiovisual Multitalker Database; p. 578–583.
- Smith ZM, Delgutte B, Oxenham AJ. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*. 2002; 416:87–90. [PubMed: 11882898]
- Spelke E. Infants' intermodal perception of events. *Cognitive Psychology*. 1976; 8:553–560.
- 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*. Lawrence Erlbaum Associates, Inc.; Hillsdale, NJ: 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*. Springer-Verlag; Berlin: 1997. p. 221–232.

- Walden BE, Prosek RH, Montgomery AA, Scherr CK, Jones CJ. Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*. 1977; 20:130–145. [PubMed: 846196]
- Yehia H, Kuratate T, Vatikiotis-Bateson E. Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*. 2002; 30:555–568.
- Yehia H, Rubin PE, Vatikiotis-Bateson E. Quantitative association of vocal-tract and facial behavior. *Speech Communication*. 1998; 26:23–43.

FIGURE 1.

Schematic of the cross-modal matching task. The top row illustrates the task in the "V-to-A direction". The bottom row illustrates the task in the "A-to-V direction". Faces denote stimuli that are presented visual-alone. Waveforms denote stimuli that are presented auditory-alone. Δt is always 500 ms.

FIGURE 2.

Box plot of results from Experiment 1 for the A-V and V-A presentation conditions. The shaded box represents the interquartile range, the bold line indicates the median score, and the whiskers represent the range from the highest to the lowest score, excluding outliers. The circle in the A-V group indicates an outlier. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of 0.05.

FIGURE 3.

Box plot of results from Experiment 2 for the A-V and V-A presentation conditions. Experiment 2 used static pictures of faces as visual stimuli for cross-modal matching. The shaded boxes represent the interquartile range, the bold line indicates the median score for the group, and the whiskers represent the range from the highest to the lowest score, excluding outliers. The circle in the V-A group indicates an outlier. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of 0.05.

FIGURE 4.

Illustration of the noise-band stimulus creation process used in Experiment 3. (Figure adapted from "Chimaeric Sounds Reveal Dichotomies in Auditory Perceptions," by Z. M. Smith, B. Delgutte, and A. Oxenham, 2002, *Nature* (<http://www.nature.com>), 416, pp. 87–90. Copyright 2002 by Nature Publishing Group. Adapted with permission.

FIGURE 5. Narrow-band spectrograms of an untransformed auditory token (top) and the derived 16-channel noise-band equivalent (bottom). The token depicts talker M2 speaking the word “WAIL.” The noise-band stimulus preserves the structure of the formant resonances but eliminates fine-grained details of f_0 and its harmonics.

FIGURE 6.

Box plot of results from Experiment 3 for the A-V and V-A presentation conditions. Experiment 3 used noise-band stimuli for auditory tokens. The data shown are for participants who heard 16-channel stimuli (diagonal shading) and 32-channel stimuli (horizontal shading). The shaded boxes represent the interquartile range, the bold line indicates the median score for the group, and the whiskers represent the range from the highest to the lowest score, excluding outliers. The circle in the V-A group indicate two outliers. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of 0.05.

FIGURE 7.

Means and standard errors for the significant interaction between Number of Channels and Presentation Order observed in Experiment 3. The squares show means in the A-V Presentation Order, and the circles show means in the V-A Presentation Order.

FIGURE 8.

Waveform of the auditory display of a stimulus token, “BACK” spoken by talker F1. The top panel shows the waveform of the untransformed stimulus. The bottom panel shows the waveform of the stimulus when it was played backward in time.

FIGURE 9.

Box plot of results from Experiment 4 for the A-V and V-A presentation conditions. Experiment 4 used auditory and visual tokens played backward in time. The shaded boxes represent the interquartile range, the bold line indicates the median score for the group, and the whiskers represent the range from the highest to the lowest score, excluding outliers. The circle and asterisk in the V-A group indicate two outliers. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of 0.05.

TABLE 1

Statistical Results From Experiment 3 With Noise-Band Stimuli

No. of Channels	Order	Group M	SE	Student's <i>t</i> vs. 0.50
16 channel	A-V	.578	.025	$t(9) = 3.04, p < .01$
	V-A	.550	.020	$t(9) = 2.52, p < .05$
32 channel	A-V	.552	.013	$t(9) = 4.09, p < .01$
	V-A	.628	.019	$t(9) = 6.86, p < .001$