

**RESEARCH ON
SPOKEN LANGUAGE PROCESSING**

Technical Report No. 10

June 20, 2002

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405*

Supported by:

Department of Health and Human Services
U.S. Public Health Service
National Institutes of Health
Research Grant DC-00111
and
Training Grant DC-00012

**VOCAL TRACT KINEMATICS
AND
CROSSMODAL SPEECH INFORMATION**

Lorin Lachs

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the Department of Psychology and Cognitive Science Program
Indiana University

June 20, 2002

Copyright 2002

Lorin Lachs

ALL RIGHTS RESERVED

Acknowledgments

This work was supported by NIH-NIDCD grants R01 DC00064 and T32 DC00012 to Indiana University. This dissertation would never have made it onto paper without the support of the various people it has been my privilege to work with in my time at the Speech Research Laboratory at Indiana University: Rose Burkholder, Allyson Carter, Steve Chin, Connie Clarke, Miranda Cleary, Cynthia Clopper, Caitlin Dillon, Stefan Frisch, Winston Goh, Jimmy Harnsberger, Jeff Karpicke, Jeff Reynolds, Mike Vitevitch, and Richard Wright. My graduate experience has been enriched by their willingness to trade ideas, discuss issues, and listen to my rants. I also thank Luis Hernandez, whose contributions to this dissertation go well beyond the merely technical. For teaching me the fundamentals of technical lab administration; for being a constant source of creative solutions to all manner of technological problems; and not least of all, for frequent coffee and lunch breaks, he deserves all the thanks I can give. For their contributions to this dissertation and my overall education, I also thank Bob Port and Tom Busey who served as advisors on this dissertation. I also thank Bob for training me in phonetics and the foundations of dynamic systems. I'd also like to thank John Kruschke for valued statistical advice throughout graduate school. My gratitude also goes to Geoff Bingham, whose good-natured support and advice have been invaluable over the course of my graduate career. Most importantly, it was his truly remarkable course in Perception and Action that set the foundation for what was to become a fundamental change in my theoretical perspective on psychology. I can only hope to one day inspire that kind of experience in one of my own students. Of course, none of this would have been possible without the dedicated, indefatigable, interminable and unflagging support of my advisor, David Pisoni. I truly appreciate his devotion and commitment to my education: the countless hours of editing our collaborative projects; the weekly office meetings; the office space and the considerable amount of equipment needed to develop this program of audiovisual speech research; the innumerable brainstorming sessions; the frequent dog 'n' pony shows with the luminaries of the field. I thank him especially for spurring me on to reach the potential he always saw and encouraged in me. Several other people in the department are due thanks for their general, all-around helpfulness and support. Darla Sallee has been especially helpful with navigating all the paperwork necessary in a university setting. Ralph Zuzolo and Alan Mauro in the Media Production division of Instructional Support Services, whose technical assistance in filming the point-light displays was invaluable. Andrew Cohen, Matt Crawford, Bill Rodawalt, and Mary Wyman have been especially fun to hang out with during those times when a break from academics was necessary. Sean McCrea deserves additional thanks for carpooling from Indianapolis with me at the ungodly hour of 7 am. I would also like to thank my family for all their support: Ken, Renda, and Tiphonie Rosenblatt whose pride in my achievements is evident. Seth Lachs has kept me in touch with the "cultural center of the universe" through daily chat sessions over the internet. Andy, Mona and Geoff Lachs who always liked hearing the war stories from school. Rolly, Judy, Doug and Carla Dreher whose good humor and good nature brightened many a weekend. Finally, I'd like to thank my wife, Kris Dreher Lachs for her love, support, jokes, antics, and unending patience. Her willingness to leave the mountains of Denver to provide me with emotional and moral support while I finished this dissertation is admired and tremendously appreciated. This dissertation is for her.

Abstract

Visual attributes of speech (lipreading) and auditory attributes of speech influence each other in the process of speech perception. Traditional accounts of audiovisual speech perception propose that cognitive mechanisms integrate the independent visual and auditory information during perception. However, the multimodal correlates of speech are not independent of each other, but are lawfully related by virtue of their common origin in the production of speech. Alternative accounts of perceptual integration build on this fact, proposing that acoustic and optical properties of speech are integrated because the relevant phonetic information is not constrained to transmission via optic or acoustic energy, but is instead modality-neutral. Under this “direct realist theory,” perceivers should be able to match visual and auditory speech patterns presented to different sensory modalities. The present series of investigations examined this hypothesis in detail using a two-alternative forced choice crossmodal matching task. The results of Experiment I indicated that observers could match speaking faces and voices, indicating that information about the speaker was available for crossmodal comparisons. This crossmodal source information was not available in static visual displays of faces and was not contingent upon a prominent acoustic cue to vocal identity (f_0). Furthermore, crossmodal matching was not possible when the acoustic signal was temporally reversed. Experiment II tested 6 different acoustic transformations to see whether they preserved or eliminated crossmodal source information. In addition, word recognition performance was tested under the same acoustic transformations. The results showed that crossmodal source information was preserved under the same conditions that preserved information needed for word recognition. Finally, Experiment III used novel point-light displays of speech to assess perceivers’ ability to match faces and voices under conditions when only isolated kinematic information about vocal tract articulation was available. The results of all three experiments were consistent with the hypothesis that the form of speech information is not contingent upon transmission via acoustic or optic energy, but is based upon the dynamics of vocal tract activity as they change over time. These properties are common to both sources of information because they reflect the same underlying articulatory events employed in speech production.

**VOCAL TRACT KINEMATICS
AND
CROSSMODAL SPEECH INFORMATION**

TABLE OF CONTENTS

Chapter I : Introduction and Overview	1
Chapter II : Crossmodal Source Identification in Speech Perception.....	4
Experiment 1: Crossmodal Matching of Auditory and Visual Patterns	6
Experiment 2: Static faces	11
Experiment 3: Noise-band Stimuli	13
Experiment 4: Temporal Reversal	18
General Discussion	21
Chapter III : Crossmodal Source Information and Spoken Word Identification	24
Experiment 5: Transformations of the Acoustic Signal	30
General Discussion	46
Chapter IV : Specification of Crossmodal Source Information in Isolated Kinematic Displays of Speech.....	51
Experiment 6: Crossmodal Matching of Kinematic Primitives.....	55
Experiment 7: Word Identification with Kinematic Primitives.....	62
General Discussion	68
Chapter V : Summary and Conclusions	71
Modality-neutral Representations vs. Evidence from Neuropsychology ...	72
Future Directions	74
Conclusion	75
References	76
Appendix A	84

Chapter I: Introduction and Overview

The ability to comprehend spoken language is one of the most extraordinary capabilities of the human species. The basis for this hallmark psychological skill lies in two behavioral components: the mind's capacity for controlling and manipulating the human vocal tract in order to produce psychologically meaningful utterances, and its complementary ability to perceive those utterances within a culturally-based communicative framework. At the heart of psychological investigations into these two capabilities lies a central question: How do the sensory patterns produced in the act of speaking carry linguistically relevant information about the communicative intent of the speaker?

For many years, the focus of inquiry into this experimental question was the acoustic speech signal. Speech scientists investigated acoustic cues to linguistic identity in the pattern. For example, in a pioneering study on speech perception, Liberman (1957) presented findings, based on spectrographic analyses and a synthesis experiment, that certain aspects of the speech signal could be found which corresponded to consonants. In addition, using speech synthesis findings, he reported that certain aspects of the patterning of these formants were shown to correlate directly with the perception of particular consonants. Thus, Liberman (1957) asserted "that there are, for each consonant, characteristic frequency positions, or loci, at which formant transitions begin, or to which they may be assumed to point." These cues could thus be easily detected by the speech perception system, and used to convert acoustic signals into a symbolic language.

However, although these results were encouraging, numerous findings over the last 45 years have shown that the goal of linearly relating the acoustic waveform to the perceived speech message is an unrealistic one. In two complementary studies, Liberman and colleagues demonstrated that the perception of speech is more closely tied to articulatory variables than to acoustic cues. In one study, Liberman, Delattre, and Cooper (1952) demonstrated that the same acoustic cue (a noise burst at 1440 Hz) can alternatively be perceived as the phoneme /p/ if it is followed by the vowel /i/ or as the phoneme /k/ if it is followed by the vowel /a/. Thus, depending on the context, the same acoustic pattern can specify different phonetic content. Conversely, Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) demonstrated that, depending on context, *different* acoustic cues can specify the *same* phonetic content. The second formant transition in the syllable /di/ is acoustically different from the second formant transition in the syllable /du/; however, in both cases, the initial phoneme is perceived as /d/. Although the acoustics differ, the articulation of the initial phoneme is assumed to be constant – a voiced alveolar closure. Thus, the perception of speech seems to rely on perceptual access to vocal tract articulatory activity, rather than to simple properties of the acoustic signal itself (Fowler, 1996; Liberman & Mattingly, 1985).

The debate between acoustic cue theories of speech perception and articulatory or gesturally-based theories of speech perception, however, is far from resolved (see, for examples, Diehl & Kluender, 1989; Fowler, 1996; Ohala, 1996; Remez, 1989). At issue in this debate is the fundamental question of whether the object of perception is the proximal acoustic signal or the distal articulatory events that produced the speech signal.

A relatively new area of investigation in speech perception has generated much theoretical interest recently because of a very basic finding: perceptual access to spoken language can be obtained *visually*. The value of visual speech information for the hearing impaired has been known for many years (Berger, 1972; Dodd & Campbell, 1987), but recent findings have demonstrated that even normal-hearing listeners are able to lipread (or "speechread") at quite high levels of accuracy (Auer, in press; Bernstein, Demorest, & Tucker, 2000; Lachs & Pisoni, submitted). Furthermore, the visual

properties of speech have been shown to influence auditory information about speech. In their pioneering study on the multimodal perception of speech, Sumbly and Pollack (1954) found that spoken word recognition was dramatically improved by simply allowing their participants to see as well as hear the words as they were spoken. Sumbly and Pollack manipulated the speech-to-noise ratio of the acoustic waveform and asked participants to identify spoken words under auditory-alone and audiovisual presentation conditions. They found that combined audiovisual stimulation was equivalent to a +15 dB gain in signal-to-noise ratio relative to the auditory-alone presentation condition. In an interesting extension of this study, Erber (1969) found that, even at signal-to-noise ratios where auditory-alone spoken word recognition was below threshold, audiovisual performance was over 50%. Furthermore, the effects of visual speech information are not limited to degraded listening conditions. In a computational analysis of their results, Sumbly and Pollack showed that the gain due to combined auditory and visual information was constant across the various signal-to-noise ratios tested. In addition, Reisberg, McLean, and Goldberg (1987) showed that concurrently presented visual information facilitated the repetition of foreign-accented speech and semantically complex sentences. The results of these studies confirm the basic finding that visual information influences the auditory processing of speech.

Another important finding demonstrates that optical and acoustic information about speech are perceptually “integrated” when auditory and visual information are presented under conflicting conditions (Massaro & Cohen, 1995). In their seminal study of this phenomenon, McGurk and MacDonald (1976) asked participants to identify the linguistic content of an incongruent audiovisual stimulus. The stimulus consisted of a visual display of a talker uttering a disyllable (/gaga/) and a simultaneously presented auditory display of a talker uttering a different disyllable (/baba/). In 98% of the cases, McGurk and MacDonald found that participants reported hearing the syllable /dada/, a “fused” response that was not contained in either the acoustic or optical signal alone. The McGurk effect has been replicated many times and under many circumstances (see Massaro & Cohen, 1995 for a comprehensive review). The McGurk effect demonstrates that the perceptual system must integrate information from the disparate sensory modalities during speech perception.

The published audiovisual speech perception literature provides additional evidence that the cues to linguistic identity do not necessarily have to be acoustic in nature to support speech perception. If the objects of speech perception are the acoustic patterns themselves, then acoustic-cue theories of speech perception must explain how visual information can be combined and integrated with auditory information to enhance or alter the perception of speech. In order to account for these findings, revised cue-based theories of speech perception have been developed that do not constrain the modality from which cues can come. One formalization of this type of theory is Massaro’s Fuzzy Logical Model of Perception (“FLMP,” Massaro, 1987; Massaro & Cohen, 1995). In FLMP, auditory and visual speech signals are analyzed with respect to multimodal feature templates stored in memory. These feature templates represent the prototypical auditory and visual features associated with various syllables. The perceptual analysis system is then assumed to weigh the degree of support from each sensory modality for each feature template. Finally, the system outputs the syllable that is most heavily supported by the incoming information. Thus, particular acoustic and optic patterns are matched directly to abstract, linguistic units (syllables), and no reference is made to the distal articulatory event that produced the patterns.

In contrast, gestural accounts of speech perception appear to be more naturally suited to accommodate these multisensory findings. Because the object of perception is assumed to be the vocal tract gesture itself, information about these gestures, whether it is available optically or acoustically, can support speech perception. For example, according to the Motor Theory (Liberman & Mattingly, 1985), auditory and visual information is analyzed by a specialized speech module that can relate both sources of information to the underlying (and intended) articulation of a vocal tract used to produce

them. Although gesturally-based, the Motor Theory does not offer a detailed account of the way in which auditory and visual information are combined and integrated together to support perception.

In contrast, the relationship between auditory and visual information about speech is made explicit in the Direct Realist theory of speech perception proposed by Fowler (1986). This ecological perspective on speech builds on the observation that the public actions and movements of the vocal tract structure acoustic and optic media in lawful ways (Gibson, 1979). Thus, articulatory activity is directly specified by the structure of patterns in both the acoustic and optic media. In some sense, perceptual information about speech is *modality-neutral*, because it can be carried by the energy detected by the visual *and* auditory perceptual systems. Indeed, even the haptic system is sensitive to speech information (see Fowler & Dekle, 1991). Under this direct realist theory, optical and acoustic properties of speech are “integrated” because they ultimately refer to the same common event: the underlying gestures of the vocal tract that produce speech.

The present investigation was designed to examine the hypothesis that the linguistically significant information in speech is modality-neutral and articulatory in nature. To explore this question, a crossmodal matching task was used to examine the acoustic and optical correlates of speech. In this procedure (Lachs, 1999), a subject is presented with the auditory or visual form of a particular talker speaking an isolated English word. The “test pattern” is presented to only one sensory modality (e.g., the visual-only form of a person speaking a word). After the test pattern is presented, the subject is presented with two response alternatives in the *complementary* sensory modality (i.e., auditory-only). One of the two alternatives is *the same event that generated the test pattern*, but presented in a different modality. Thus, if the test pattern is visual-alone, then the two response alternatives are auditory-alone. The matching task can also be carried out in the reverse order, where the test pattern is auditory, and the two response alternatives are visual. The subject’s task is to choose the alternative that was generated by the same event as the test pattern (i.e., to choose the target alternate). In essence, the crossmodal matching task asks participants to match information about the vocal source of an utterance across sensory modalities.

The experiments reported below in this dissertation used the crossmodal matching task to examine the nature of multisensory speech information and to assess the specific proposal that this information is fundamentally modality-neutral in nature. In Chapter II, the crossmodal matching task is used to examine whether vocal source matching can be accomplished across sensory modalities. In addition, several stimulus manipulations were used to determine some of the critical acoustic and optic patterns necessary for specifying crossmodal source information. In Chapter III, the acoustic signal was systematically manipulated to examine those properties of acoustic patterns that can convey crossmodal source information. In addition, the link between crossmodal source information and the information necessary to support word recognition was investigated. Finally, in Chapter IV, kinematic information about the movement of the vocal tract was isolated in optic and acoustic displays of speech by constructing point-light visual displays of speech and sinewave speech replicas of the acoustic signal. These multimodal kinematic primitives were then used in several experiments to examine whether crossmodal source information and the information necessary for word recognition is specified in the isolated motions of the vocal tract, regardless of the particular sensory modality through which this information is obtained. The results from this series of experiments are discussed in terms of implications for the nature of speech information and current theories of speech perception and spoken word recognition.

Chapter II: Crossmodal Source Identification in Speech Perception

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 et al., 2000; Campbell & Dodd, 1980; Jeffers, 1971; Walden, Prosek, Montgomery, Scherr, & Jones, 1977) and children (Erber, 1972; Erber, 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/ and simultaneously presented it 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, Sumbly 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 ratio under auditory-alone conditions (MacLeod & Summerfield, 1987; Summerfield, 1987). They found that *absolute* gains in speech perception accuracy were most dramatic at signal-to-noise 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 (Braidia, 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 (Massaro, 1998; Massaro & Cohen, 1995), relies on *a priori* assumptions that the perceptual system has tacit knowledge of the

relationships 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 together 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 relationships exist between acoustic and optic displays that can potentially be useful for the process of speech perception. For 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 relationship 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).

In fact, the visual and auditory correlates of speech are always lawfully tied to one another. Even infants as young as four months of age are sensitive to the relationships between simultaneously presented auditory and visual information for many natural events. In one investigation of crossmodal sensitivity in infants, Spelke (1976) presented four-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 crossmodal 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 a 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 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) has 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 claims that the human auditory system may be structured in a way to exploit these relationships. 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 naïve 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 naïve 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. found that several naïve normal-hearing and hearing-impaired subjects 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. 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: Crossmodal 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 counter-intuitively, 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 crossmodal matching task.

The crossmodal matching task (Lachs, 1999) is a 2-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 subject 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 in only one sensory modality. After the test pattern is presented, the subject 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 auditory, and the response alternatives are visual (the “A-V order”). The subject’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).

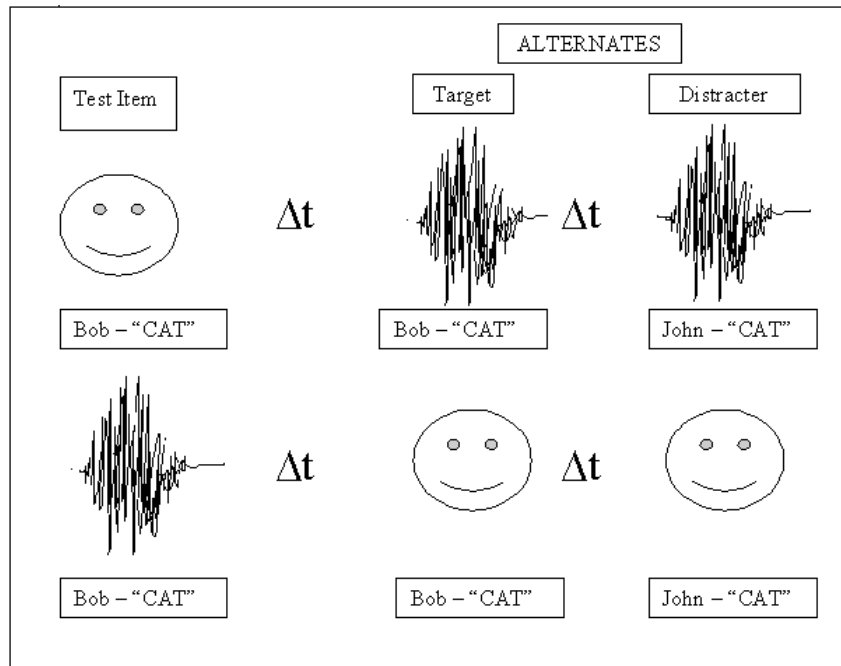


Figure 1. Schematic of the cross-modal matching task. The top row illustrates the task in the V-A order. The bottom row illustrates the task in the A-V order. Faces denote stimuli that are presented visual-alone. Waveforms denote stimuli that are presented auditory-alone. Δt is always 500 ms.

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 subject’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 crossmodal 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 matched it (V-A) 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, since comparisons to the target stimulus are made throughout the duration of the trial, which can last up to 3 seconds. In order 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. In order to test this possibility, stimulus items were balanced for their intelligibility. 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).

A 2-alternative forced choice matching procedure was used in a 2 x 2 factorial design. The two levels for the between-subjects "Order" factor were "A-V" (where participants identified the correct visual stimulus after hearing the auditory test stimulus) and "V-A" (where participants identified the correct auditory stimulus after viewing the visual test stimulus). The two levels of the within-subjects "Visual Intelligibility" factor were "low" and "high." Stimuli in the "low" group were words whose average 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" Sony Trinitron Monitor (0.26 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 x 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 visual-alone 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, while 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 (8 trials) preceded each participant's session. During the training period, the participant was presented with a crossmodal matching trial and asked to pick the correct alternative. During training *only*, the response was followed by feedback. The feedback provided was a presentation of the combined audio-visual 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 boxplots 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 2-way (Visual Intelligibility and Order) 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 ($F(1,18) = 1.25$, n.s.) and no interaction between presentation order and visual intelligibility ($F(1,18) < 1$, n.s.).

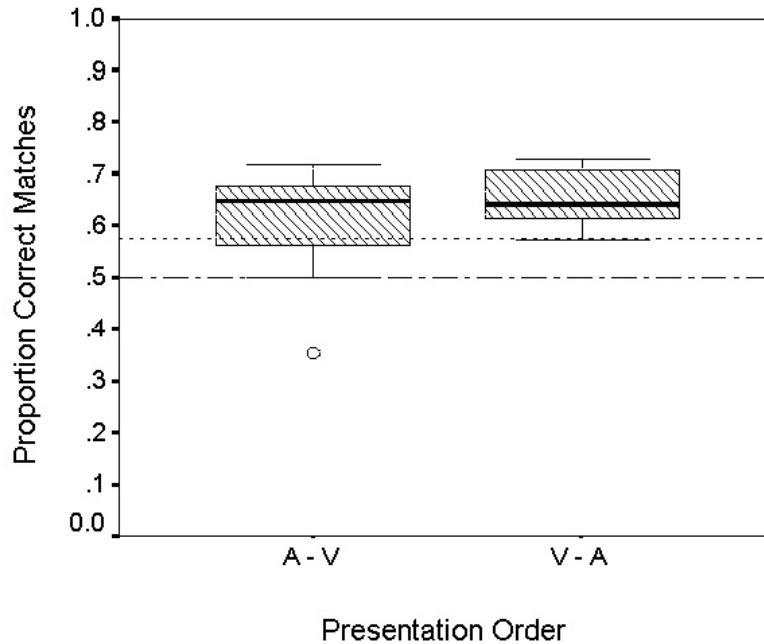


Figure 2. Boxplot of results from Experiment 1 for the A-V and V-A presentation conditions. The shaded box represents the interquartile range, the bold line indicates the median score, and the whiskers represent the range from the highest to the lowest score, excluding outliers. The circle in the A-V group indicates an outlier. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of 0.05.

Discussion

The results from the crossmodal 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 crossmodal 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, submitted).

We also found a marginal effect of visual intelligibility indicating that it may have been easier to make crossmodal judgments when the word itself was more visually intelligible. Although the effect was marginal, this relationship is of some interest because the crossmodal 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 crossmodal 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 crossmodal 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 “crossmodal 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, crossmodal 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 crossmodal matching abilities exists. It may be that participants make crossmodal 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 crossmodal specification of identical phonetic events.

In order 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 crossmodal 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 crossmodal 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 crossmodal information is not present 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 above in the Methods section for Experiment 1. The stimulus materials were also taken from the same set as those used in the Experiment above. 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 boxplots 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), $t(19) = 1.06$, n.s. This was true for the A-V group ($M = 0.504$, $SE = 0.038$, $t(9) < 1$, n.s.) as well as the V-A group ($M = 0.546$, $SE = 0.029$, $t(9) = 1.29$, n.s.) when analyzed separately. A 2-way ANOVA revealed no main effects or interactions due to the Visual Intelligibility and Order factors.

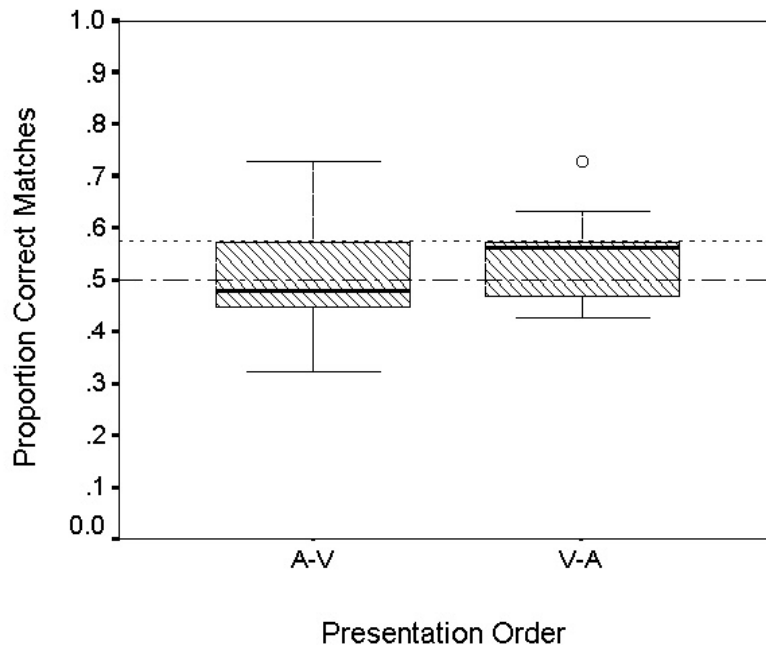


Figure 3. Boxplot of results from Experiment 2 for the A-V and V-A presentation conditions. Experiment 2 used static pictures of faces as visual stimuli for crossmodal matching. The shaded boxes represent the interquartile range, the bold line indicates the median score for the group, and the whiskers 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.

Discussion

The results of Experiment 2 show that participants could not make crossmodal matching judgments between faces and voices when the faces were static pictures of the original talkers. This finding rules out the hypothesis that crossmodal 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 between talkers, especially across gender. It is possible that the pitch of a talker's voice, his/her inflection, or even his/her prosody could be used as a reliable cue for distinguishing between talkers. For example, a participant in the crossmodal 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 crossmodal information and intersensory correspondences about articulation and more to do with expectations set up by social norms, prior experience, etc.

In order to test whether f_0 plays a role in the judgment of crossmodal 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 bandpass 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 crossmodal 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 the only reliable cue for making crossmodal 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 above in the Methods section for Experiment 1. The stimulus materials were also taken from the same set as those used in the Experiments above. However, the auditory portion of each stimulus (that is, the audio track of the movie) was manipulated using digital signal processing methods to create noise-band stimuli (Smith et al., 2002).

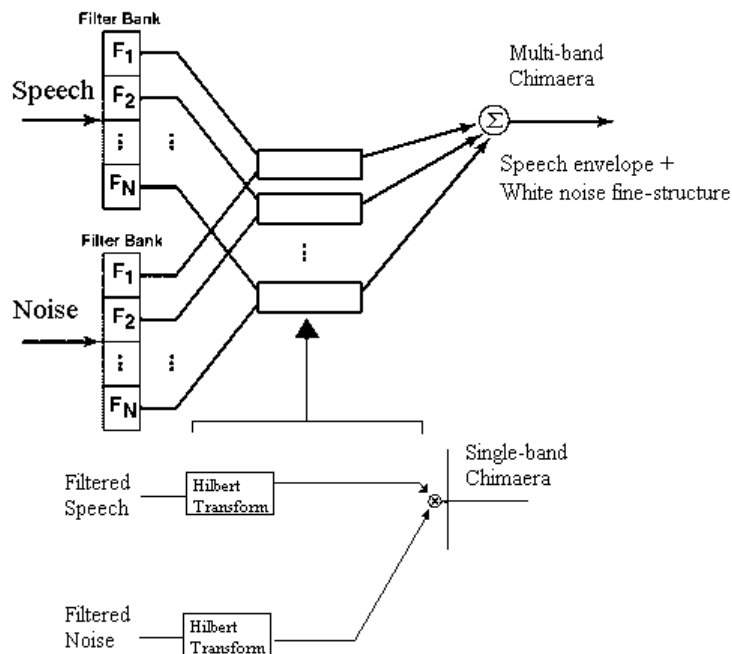


Figure 4. Illustration of the noise-band stimulus creation process used in Experiment 3 (Figure adapted from Smith et al., 2002).

Figure 4 illustrates the noise-band stimulus creation process. First, the audio track of each movie was first 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% of 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-subjects factor.

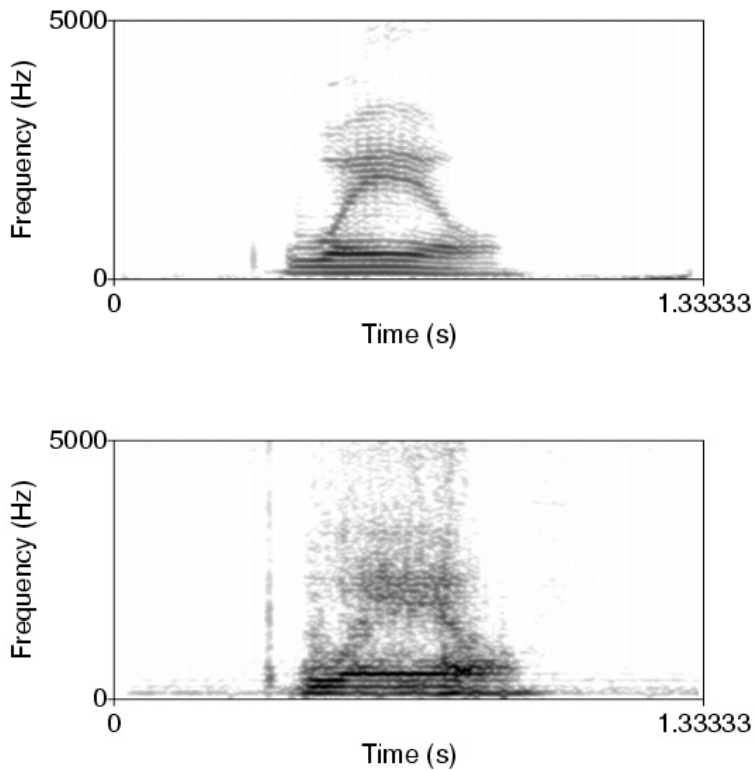


Figure 5. Narrow-band spectrograms of an untransformed auditory token (top) and the derived 16-channel noise-band equivalent (bottom). The token depicts talker M2 speaking the word “WAIL.” The noise-band stimulus preserves the structure of the formant resonances while eliminating fine-grained details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

Results

Figure 6 shows boxplots of scores for participants who made crossmodal 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 presentation order 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.

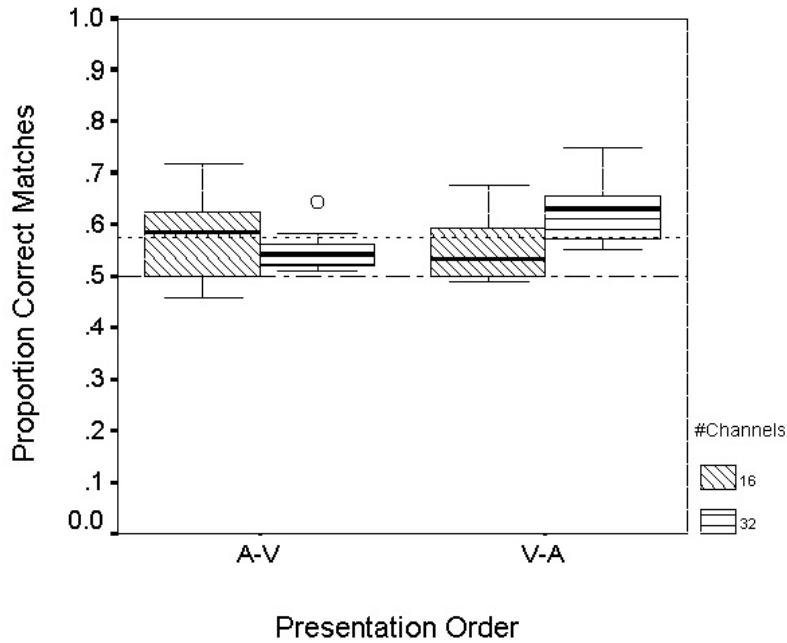


Figure 6. Boxplot of results from Experiment 3 for the A-V and V-A presentation conditions. Experiment 3 used noise-band stimuli for auditory tokens. The data shown are for participants who heard 16-channel stimuli (diagonal shading) and 32-channel stimuli (horizontal shading). The shaded boxes represent 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 A-V 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.

#Channels	Order	Group Mean	Std. Error	Student's t vs. 0.50
16-channel	A-V	0.578	0.025	$t(9) = 3.04, p < 0.01$
	V-A	0.550	0.020	$t(9) = 2.52, p < 0.05$
32-channel	A-V	0.552	0.013	$t(9) = 4.09, p < 0.01$
	V-A	0.628	0.019	$t(9) = 6.86, p < 0.001$

The results were also submitted to a 3-way (Number of Channels, Visual Intelligibility and 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.

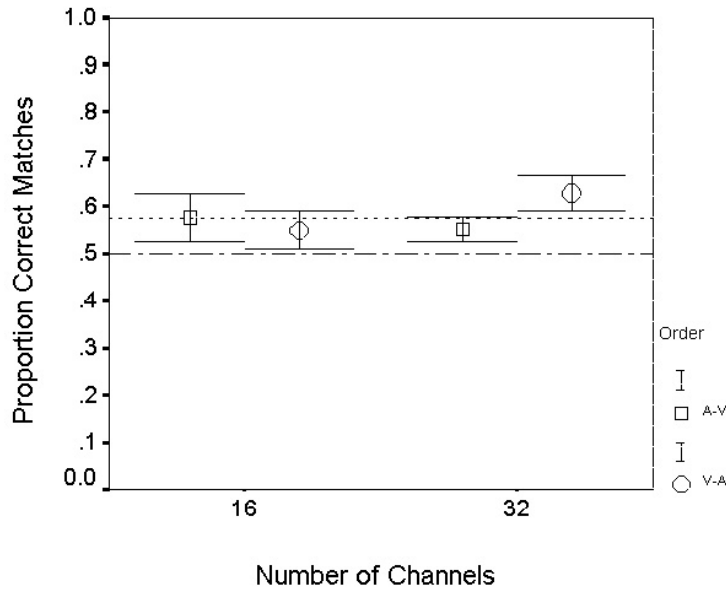


Figure 7. Means and standard errors for the significant interaction between Number of Channels and Presentation Order observed in Experiment 3. The squares show means in the A-V Presentation Order, and the circles show means in the V-A Presentation Order.

Discussion

Despite the removal of f_0 , a traditional cue to vocal identity, participants in Experiment 3 were able to perform in the crossmodal 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 crossmodal matching task.

The results from this experiment also showed that crossmodal 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 crossmodal 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, but did not help to specify the crossmodal information any more clearly. Regardless of these differences, the results demonstrate clearly that f_0 is not a necessary source of auditory information for making reliable crossmodal 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 f_0 information. As mentioned above, 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 crossmodal matching judgments. As long as the acoustic signal can specify the dynamic articulations of the vocal tract and how they change over time, crossmodal information is apparently preserved and correct crossmodal matches can be made.

Experiment 4: Temporal Reversal

The results of the three experiments reported above demonstrate that crossmodal 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 crossmodal 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 crossmodal 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 in which it is measured. The duration, then, can be seen as a kind of amodal information about the relationship 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.

In order 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 backwards 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 backwards in time. This temporal reversal transformation destroys the information necessary for accurate word recognition (Kimura & Folb, 1968), but preserves the duration of the stimulus. If observers are able to make crossmodal judgments based on signal duration, then performance on the crossmodal matching task should not suffer as a result of the transformation.

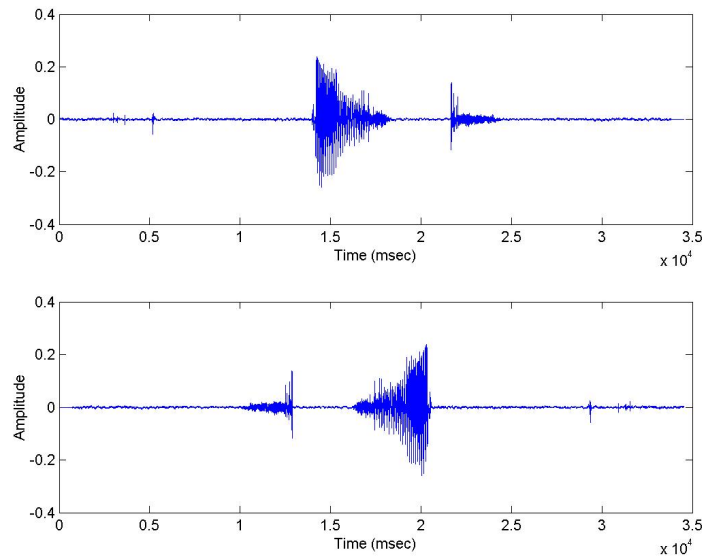


Figure 8. Waveform of the auditory display of a stimulus token, “BACK” spoken by talker F1. The top panel shows the waveform of the untransformed stimulus. The bottom panel shows the waveform of the stimulus when it was played backwards in time.

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 above in the Methods section for Experiment 1. The stimulus materials were also taken from the same set as those used in the Experiments above. 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 boxplot 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 crossmodal matching judgments using backwards video and audio clips. Regardless of the Presentation Order, average performance did not differ statistically from chance (0.5), $t(19) = 1.47$, n.s. This was true separately for the A-V group ($M = 0.525$, $SE = 0.017$, $t(9) = 1.50$, n.s.) and for the V-A group ($M = 0.529$, $SE = 0.034$, $t(9) < 1$, n.s.) when analyzed separately. A 2-way ANOVA revealed no main effects or interactions of the Visual Intelligibility and Order factors.

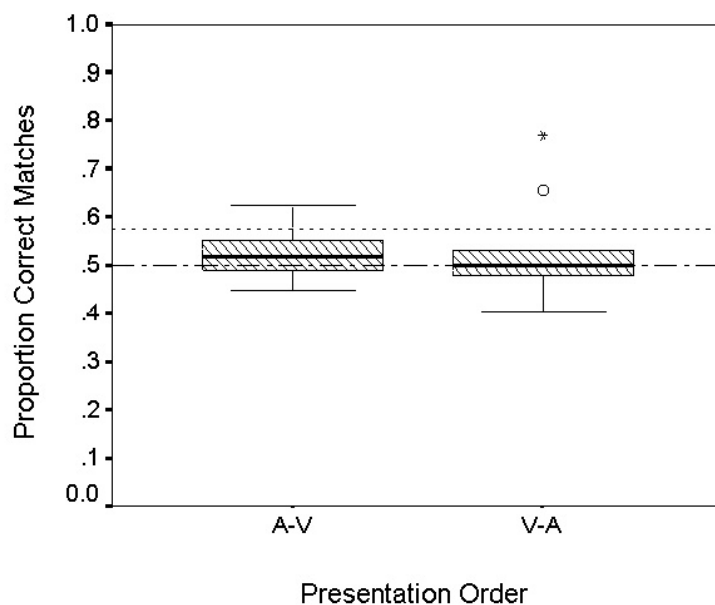


Figure 9. Boxplot of results from Experiment 4 for the A-V and V-A presentation conditions. Experiment 4 used auditory and visual tokens played backwards in time. The shaded boxes represent the interquartile range, the bold line indicates the median score for the group, and the whiskers represent the 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.

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 was used as a cue for making crossmodal 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 inter-talker variation in the duration of a spoken word. However, this source of information about the relationship between auditory and visual displays is not useful as a crossmodal cue for matching faces and voices.

General Discussion

The present set of crossmodal 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 crossmodal 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 relationship to one another. Furthermore, the crossmodal information that supports these matching judgments is not based in static visual features about face identity (e.g., relative age, hair style), 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 crossmodal voices. The results showed that crossmodal information is not well-specified in the fine-structure (f_0) of an utterance, because participants could match bandpass, “chimaeric” auditory stimuli with dynamic faces (Experiment 3). Finally, although duration cues could serve as a potential source of crossmodal information, these properties cannot be used for making crossmodal matching judgments; temporally backwards speech in the visual and auditory domains could not be matched across modalities in Experiment 4.

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 contain reliable information that allow for the accurate perception of these attributes.

Both the Repp (1987) and Li et al. (1991) studies indicate 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 sinewave speech replicas. Sinewave 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. (1997) showed that sinewave speech can support the identification of the gender and even the identity of a talker. The results of the present 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 the present findings that there is sufficient information about the spoken event encoded in the optical or acoustic signals that allows subjects to make reliable crossmodal comparisons, even for isolated spoken words. For accurate crossmodal judgments to be made, auditory and visual information about the movement of speech articulators had to be preserved in some form. As noted above, 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).

Linguistic vs. Indexical Properties of Speech. The observed relationships between lexical and indexical speech information, and their common basis in vocal tract articulation, suggests a link between crossmodal matching and word recognition. That is, if both word identification information and crossmodal source information are supported by patterns of sensory stimulation that relate to the articulatory events which produced them, then performance on word identification tasks should be relatively high under the same acoustic transformations that support crossmodal matching. In a supplementary study of the acoustic transformations used in the present 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 they heard using the keyboard. Participants were asked to make sure each word they typed contained no typographical 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 backwards 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, Delgutte, and Oxenham (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 Backwards condition (both $p < 0.001$), but that 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 crossmodal 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 auditory-alone 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, crossmodal 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 crossmodal 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 crossmodal matching findings, taken together with the word recognition findings, raise intriguing possibilities about the auditory form of crossmodal source information which will be investigated further in future work (Lachs, 2002).

In summary, the present 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 relationships 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 crossmodal source information. The results are consistent with the theory of direct perception, which predicts that crossmodal 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. Crossmodal 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, & Terzopoulos, 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. The present 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.

Chapter III: Crossmodal Source Information and Spoken Word Identification

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). The existence of the McGurk effect demonstrates that integration of information in the auditory and visual sensory modalities occurs during the process of speech perception. Exactly how this integration occurs has been a topic of some interest over the last few years.

Much of the work on the integration issue has concerned the effects of conflicting auditory and visual cues on phonetic identification (e.g., Green, 1996; Green & Gerdeman, 1995; Massaro, 1998). This research effort has produced a large body of knowledge concerning the ways in which acoustic and optic information influence each other during the process of extracting phonetic information from the signal. For example, it has been shown that the vowel context in which a segment is presented affects the degree to which visual information influences the McGurk illusion (Green, 1996; Green, Kuhl, & Meltzoff, 1988), that inverted faces reduce the effects of McGurk integration (Green, 1994; Jordan & Bevan, 1997; Massaro & Cohen, 1996), that temporal asynchrony between the auditory and visual displays has no effect on McGurk integration (Munhall, Gribble, Sacco, & Ward, 1995; Smeele, Sittig, & Heuven, 1994), and even that separating the spatial location of the auditory and visual aspects of the stimulus makes little difference on the extent of the McGurk effect (Bertelson, Vroomen, Wiegeraad, & De Gelder, 1994; Fisher & Pylyshyn, 1994; Jones & Munhall, 1997).

Other studies have shown that the auditory and visual information in a multimodal stimulus are evaluated together from a very early point in the process, such that the information in each channel is evaluated relative to the information present in the other channel. For example, Green and Kuhl (1989) showed that the perceived phoneme boundary in voice onset time along an /ibi/ to /ipi/ continuum was dependent on whether or not there was concurrent visual information at the time the auditory signal was presented. Presentation of the visual portion of the word caused the VOT boundary to shift as though it were along an /idi/ to /iti/ continuum, precisely the continuum specified by the McGurk effect.

Green and Kuhl's (1989) findings showed that low-level auditory cues are evaluated in the context of the *combined* audiovisual stimulus, indicating that integration of the information in the various modalities happens before linguistic classification. Another study by Green and Kuhl showed interference in the processing of one sensory modality based on variation in the other (Green & Kuhl, 1991). Using a Garner-type speeded classification task (Garner, 1974), they showed that variation in visual place information affected the classification of simultaneously presented *auditory* voicing information. Importantly, the classification times for nonspeech pitch information was not affected by variation in concurrently presented visual speech information, indicating that the crossmodal interference observed for auditory and visual speech occurred because both signals were speech, and not because they were presented synchronously. Green and Kuhl interpreted their evidence as further support for the proposal that speech information in both sensory channels is processed in tandem.

Much of the research on audiovisual integration has been gathered using experiments that assess phonetic identification. With the exception of Dekle, Fowler and Funnel's (1992) study of audiovisual "McGurk" integration in real words, nonsense syllables have been used to examine the

McGurk effect. Sumbly and Pollack (1954) showed that audiovisual speech effects are not confined to segmental (or illusory) contexts. Their results demonstrated that the intelligibility of spoken words can be enhanced by as much as +15 dB in noisy environments, a gain that surpasses even the best hearing aid devices currently available. Moreover, a recent set of studies concerning the effects of visual speech on detection thresholds for auditory speech in noise showed that speech signals can be detected at much lower signal-to-noise ratios when simultaneously presented with visual speech displays (see Grant, 2001; Grant & Seitz, 2000).

Grant (2001) suggests that auditory and visual speech information integrate at the level of signal detection well before any subsequent linguistic processing occurs. Interestingly, he found a strong correlation between the visual area circumscribed by lip opening and the rms amplitude envelope of acoustic speech (especially F2) over the duration of an utterance. Grant suggested that synchronous, crossmodal correspondences between acoustic and optical phonetic displays might underlie the perceptual system's tendency to integrate information across sensory modalities. Moreover, these crossmodal relationships appear to combine and to be founded upon the common origin of optical and acoustic patterns in the same articulatory event; changes in the frequency of F2 are associated with changes in the area of the front oral cavity (Ladefoged, 1996; Stevens, 1998), an articulatory event which is visible in the area of lip-opening.

Multimodal relationships such as these are consistent with the direct realist theory of speech perception (Fowler, 1986), in which the articulatory event forms the underlying basis for speech information. Whenever a human vocal tract moves and changes shape in order to produce speech, it structures both the acoustic and optic media in lawful ways that are dictated by the physical properties of light and sound propagation. These dynamic patterns provide direct information about the underlying events that structured them, and thereby provide information about the event to be perceived (Gibson, 1966). According to this conceptualization, information is said to be modality-neutral; as long as an energy medium can be structured by an event, and as long as such structure is detectable by an animal, the specific modality through which information is obtained does not matter.

Convincing evidence for the modality-neutral form of audiovisual speech information comes from a study investigating the integration of *haptic* and auditory phonetic information carried out by Fowler and Dekle (1991). In their experiment, normal-hearing participants were briefly instructed in Tadoma, a method of obtaining phonetic information by active touch, developed primarily for use with blind and hearing impaired individuals (Schultz et al., 1984). The participants in Fowler and Dekle's study were then presented with McGurk-type stimuli: they simultaneously *heard* the syllable [ba] and *felt* the syllable [ga]. Despite this unusual way of obtaining speech information, participants showed an effect on the perception of the heard syllable based on the *tactile* information presented concurrently, indicating that they had integrated the acoustic and haptic phonetic information during perception. Fowler and Dekle interpreted their findings as support for the proposal that the perceptual system is not constrained by the particular sensory modality through which it obtains information; even haptic information about articulatory movements can be "integrated" with acoustic information in the process of perceiving speech.

Linguistic vs. Indexical Properties of Speech

All of the studies reviewed above, and indeed, most of the previous investigations of the effects of audiovisual information on speech perception have focused on what are commonly referred to as the *linguistic* properties of the signal: that is, the attributes of the signal that support the recognition and identification of phonemes, syllables, and words in the message. However, a growing body of literature has shown that speech signals also simultaneously carry important information about

the *indexical* properties of the talker: properties such as gender, dialectal variation and idiosyncratic speaking style are present in the acoustic display of speech (Hirahara & Kato, 1992; Pisoni, 1993). In contrast to traditional conceptualizations of the speech signal (Abercrombie, 1967; Ladefoged, 1996), the time-varying acoustic signal is not discontinuously divided neatly into linguistic and indexical channels of information; the speech signal carries information about the linguistic utterance as well as information about both the source of the utterance and the listener's communicative circumstances. In other words, linguistic and indexical information are mixed together and 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 in several important ways (see Goldinger, 1998; Lachs, McMichael, & Pisoni, in press, for reviews). Early studies showed that simply changing the voice of the talker from one trial to the next affected the identification of vowels (Verbrugge, Strange, Shankweiler, & Edman, 1976), consonants (Fourcin, 1968), and words (Creelman, 1957; Mullennix et al., 1989). In addition, changes in the talker's voice also affected speed of processing (Cole, Coltheart, & Allard, 1974). In one study, Mullennix and Pisoni (1990) assessed the codependencies of processing linguistic and indexical information in the same set of stimuli. Using a Garner-type speeded classification task (Garner, 1974), they constructed sets of stimuli that varied along two dimensions. One dimension, the "word" dimension, varied the cues to phonetic categorization of the initial segment of a word (e.g., "b" vs. "p"). The other dimension, the "voice" dimension, varied the cues to the identity of the talker uttering the word (e.g., "male" vs. "female"). Mullennix and Pisoni asked their subjects to make several judgments about the stimuli using one dimension at a time, while manipulating the variation along the irrelevant dimension. They found systematic differences in reaction time that depended on the variation in the "irrelevant" dimensions. The pattern of speeded classification data they observed under these conditions was consistent with the proposal of mutually dependent processing of the two stimulus dimensions. 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.

These studies have shown that stimulus variability has an effect on speech perception and spoken word recognition. More importantly, the information about a talker's voice in the acoustic signal is processed in a dependent or contingent fashion along with the information specifying the linguistic content of the message. What kind of information about a talker's voice is encoded in the speech signal, and how does that information contribute to speech perception? In a measurement study of the acoustic correlates of talker intelligibility, Bradlow, Torretta and Pisoni (1999) found that while 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 suggest that the indexical properties of a talker may be completely intermixed with the actual phonetic realization of an utterance and 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 other studies using sinewave replicas of speech. When speech is produced, the vocal cords generate an acoustic energy source which has a power spectrum that extends over a range of frequencies. These frequencies are selectively damped or amplified into bands of concentrated energy, or "formants" by the supralaryngeal vocal tract. The frequency damping is based on the motion of the vocal articulators, which generate, at a rough approximation, two cavities in the vocal tract with selective resonant frequencies (Fant, 1960). Thus, as a spoken event unfolds and

evolves over time, the spectral pattern of the formants (i.e., the vocal tract transfer function) carries information about the motion of the articulators as they change over time.

Sinewave speech is created by generating independent sinusoidal signals that trace the center frequencies of the three lowest formants in naturally produced utterances (Remez et al., 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 et al., 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 sinewave 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 sinewave replicas of their utterances. Their results on familiar talker recognition are remarkable because sinewave 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 sinewave speech patterns, an utterance is essentially stripped of all of the redundant acoustic information in the utterance except the time-varying properties of the vocal resonances generated by articulatory motion. While these “skeletonized” versions of speech have been shown to be sufficient for accurate identification of the linguistic content of a message, the recent findings reported by Remez and his colleagues demonstrate that sinewave speech patterns are also sufficient for the accurate identification of indexical information about familiar voices as well. Time-varying sinewave speech patterns preserve individual, talker-specific cues needed for voice recognition, as well as the linguistic content of the message.

Thus, even in its most basic forms, linguistic and indexical sources of information appear to be inextricably bound to one another, and encoded in the time-varying speech signal. Because sinewave speech patterns preserve little of the original signal other than the acoustic variation corresponding to the kinematics of articulatory motion that are reflected in the vocal tract transfer function, the findings suggest that the links between linguistic and indexical information derive from the common underlying articulatory events and movements of the speech articulators that produce speech.

Crossmodal Source Information in Speech

The underlying articulatory origins of spoken language have been implicated as potential sources of information in the integration of multimodal sensory inputs (Fowler, 1996; Grant, 2001; Rosenblum, 1994) and as the basis of the inextricable link between linguistic and indexical information in acoustic displays of speech (Pisoni, 1997; Remez et al., 1997). 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. Point-light displays eliminate from a visual stimulus all information other than kinematic movement information (see Johansson, 1973). Rosenblum, Yakel, Baseer, Panchal, Nodarse, and Niehaus (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 correct fully illuminated face with the point-light display, indicating that information for face recognition is specified in isolated movements alone. Rosenblum et al. suggest that face matching between isolated kinematic displays and their fully-illuminated counterparts may be due to a common source of information that supports both face recognition and visible speech perception and that visible articulatory information may “incidentally” specify talker-specific information that is subsequently used for face recognition. As Rosenblum, et al. point out, this is exactly analogous to the findings that talker-specific information is contained in sinewave 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.

A recent study investigated the relationship between the acoustic and optical specifications of talker-specific information using a “crossmodal matching task” (Lachs, 2002). Figure 10 illustrates the crossmodal 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 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 (2002) 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 (that is, “crossmodal source information”) must be contained in the acoustic and optic displays of speech.

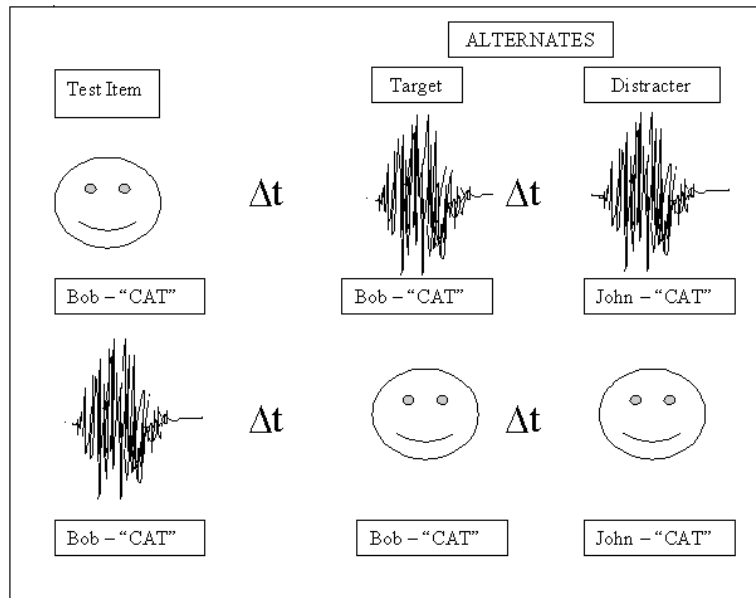


Figure 10. Schematic of the crossmodal matching task. The top row illustrates the task in the “V-A direction”. The bottom row illustrates the task in the “A-V direction”. Faces denote stimuli that are presented visual-alone. Waveforms denote stimuli that are presented auditory-alone. Δt is always 500 ms.

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

In the second experiment, acoustic stimuli were converted into noise-band “chimaeric” stimuli (Smith et al., 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 crossmodal matching task above chance, indicating that f_0 is not implicated in crossmodal source information.

In the third experiment, all of the visual and auditory stimuli were played backwards in time. By playing the patterns backwards, Lachs (2002) tested 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 backwards displays, indicating that crossmodal source information is not based on the use of overall duration cues.

In another study, the same acoustic stimulus materials used in the earlier crossmodal matching study were also presented to participants in a word recognition experiment. As expected, word identification was successful only under the same acoustic transformations that preserved crossmodal matching ability. This relationship between crossmodal matching and word recognition raises intriguing possibilities about the auditory form of crossmodal source information. Crossmodal source information and linguistically significant information needed for word identification appear to be encoded together and are carried in parallel in the pattern of formants as they vary over time.

The importance of perceptual access to vocal tract articulatory activity in theoretical accounts of speech perception has long been acknowledged (Fowler, 1996). In two complementary studies, Liberman and colleagues demonstrated that the perception of speech is more closely tied to articulatory variables than to acoustic cues. In one study, Liberman, Delattre, and Cooper (1952) demonstrated that the same acoustic cue (a noise burst at 1440 Hz) can alternatively be perceived as the phoneme /p/ if it is followed by the vowel /i/ or as the phoneme /k/ if it is followed by the vowel /a/. Thus, depending on the context, the same acoustic pattern can specify different phonetic content. Conversely, Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) demonstrated that, depending on context, *different* acoustic cues can specify the *same* phonetic content. The second formant transition in the syllable /di/ is acoustically different from the second formant transition in the syllable /du/; however, in both cases, the initial phoneme is perceived as /d/. Although the acoustics differ, the articulation of the initial phoneme is constant – a voiced alveolar closure.

Thus, according to Liberman and his colleagues speech perception is more closely tied to articulatory activity than it is to acoustic patterns. The job of a theory of speech perception based on articulation is to explain the way that acoustic patterns can signify articulatory gestures. For the motor theorists (Liberman & Mattingly, 1985), acoustic patterns are evaluated with respect to an innate neural speech perception module that recruits the perceiver’s motor control system and translates the incoming sound patterns into the intended gestures that must have produced them.

In contrast, the direct realist theory of speech perception (Fowler, 1986) posits that lawful physical constraints on the production of speech sounds will necessarily lead to invariant properties of

the acoustic signal that are directly related to the underlying gestures of the vocal tract that produced them. These invariant physical properties support direct perceptual access to the events that produced them.

Experiment 5: Transformations of the Acoustic Signal

The crossmodal 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 crossmodal matching task is affected by the transformation, it should be possible to describe the ways in which formant structure carries crossmodal source information – and by extension, articulatory information – about speech events. If a given transformation of the signal disrupts the ability to perform the crossmodal matching task, then we can assume that such a transformation eliminates crossmodal source information from the acoustic pattern and that this particular source of information is critical to speech perception and spoken word recognition performance.

The crossmodal matching task as a method for uncovering the form of crossmodal source information was already used with one specific transformation: temporal reversal. From the “backwards” studies presented above, in which the visual and/or auditory displays were simply played backward in time, we concluded that the information necessary to support crossmodal judgments was not temporally invariant. That is, the crossmodal information is only transmitted when auditory displays are played forward in time, preserving the original temporal structure of the recorded event. Furthermore, the results of the backwards experiments showed that overall duration cues did not play a role in specifying crossmodal source information. The present investigation was designed to extend and elaborate this research strategy using several other transformations of the acoustic signal.

In addition to studying crossmodal matching, all the acoustic transformations were tested using an additional word identification task to see if the transformation preserved the information necessary for word recognition. We predicted that transformations that preserve crossmodal 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 vocal tract. Thus, for each of the transformations tested below, two tasks were performed: a crossmodal 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 crossmodal information are outlined below under the description of each transformation. Appendix A (p. 84) also shows spectrograms of the same token across all the different transformations.

Spectral Domain Transformations

Several transformations in the frequency domain were used to produce perturbations in the 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. Note that this transformation preserves the absolute spacing between the formants, but changes the absolute frequencies themselves. If crossmodal information about speech is specified in terms of the patterning of the formants, irrespective of their absolute frequencies, then performance on the crossmodal matching task should remain unaffected by this transformation. Because of the hypothesized relationship between crossmodal source information and word identification information, word identification performance should also not be affected.

A related frequency domain transformation is “linear frequency scaling”. Under this transformation, the components of the frequency spectrum were transformed such that any arbitrary frequency distance in the original signal is linearly related to the frequency distance in the transformed signal. If crossmodal 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 crossmodal matching task under linear frequency scaling conditions should remain unaffected by this transformation, since relative frequency spacing information is preserved.

The third frequency domain transformation that we tested was “nonlinear frequency scaling”, in which the components of the frequency spectrum are transformed non-linearly, such that the frequency distances in the original spectrum were nonlinearly related to frequency distances in the transformed spectrum. This transformation preserves neither the absolute *nor* the relative frequency spacing of the formants. Thus, the ability to perform above chance on the crossmodal matching and word identification tasks should be eliminated.

Finally, some of these transformations may bias performance by introducing into the signal high frequencies that are not normally produced by human vocal tracts. There is another transformation of the auditory signal that does not preserve the relative spacing of the formants, but also does not introduce new frequency components: “frequency rotation” (Blessner, 1972). In this transformation, frequency components are rotated around a center frequency in the signal. Thus, the range of frequencies is preserved, while the original relationship among formants is disrupted. We expected that this transformation would eliminate subjects’ ability to perform on both the crossmodal matching task and the word identification task, because it removes the information about articulation that is carried by the temporal and spectral patterning of the formants.

Temporal Domain Transformations

Because the formant frequencies of speech vary over time, several temporal transformations of the signal may also selectively preserve or attenuate crossmodal source information. Thus, linear scaling can be carried out in the temporal domain, as well, speeding up or slowing down the original speech. The spectral relationships between formants over time would remain invariant, however, across this transformation. If crossmodal information is specified in the spectral relationships among formants, then this transformation should not disrupt crossmodal matching performance at all. However, if the temporal scale of formant patterning is important for conveying crossmodal information, performance should decrease on both the crossmodal 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 transformation, the relationship among the formants changes over time, but the temporal information about articulation is degraded. Consequently, crossmodal matching and word identification performance should also be reduced or eliminated.

Method

Stimulus Materials

Four Apple Macintosh G4 computers, each equipped with a 17” Sony Trinitron Monitor (0.26 dot pitch) were used to present the visual stimuli to subjects. The stimuli were a set of 96 tokens selected from a previously recorded audiovisual stimulus set (Lachs & Hernández, 1998; Sheffert et al., 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 x 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 DT100 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 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-Katuse, & Cheveigne, 1999). STRAIGHT is a MATLAB-based analysis/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 f_0 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 f_0 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 (v. 5). 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 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 11a 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.

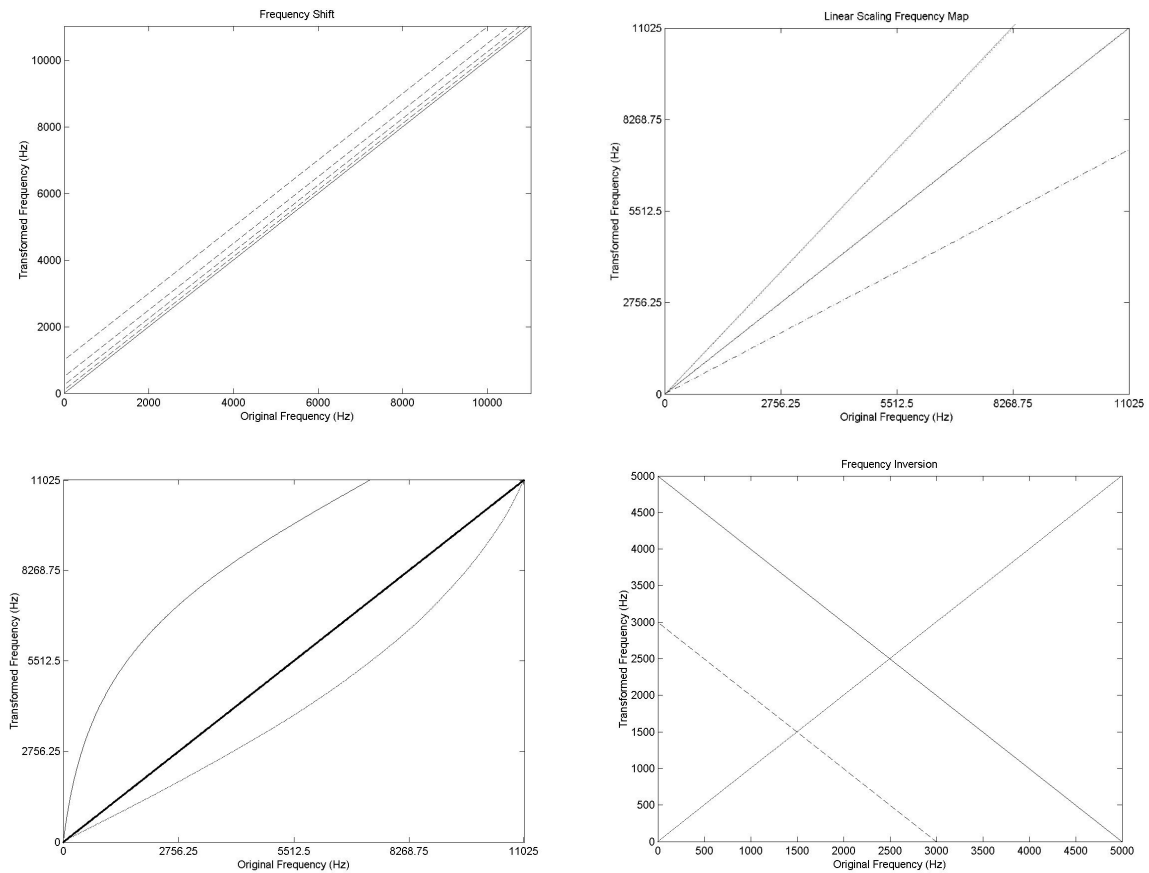


Figure 11. Frequency transformation maps for the various transformations. The solid line in each panel shows the frequency map for the untransformed condition. (a) Frequency Shifting. Dashed lines indicate the various levels of shift used. (b) Linear Frequency Scaling. The dashed line shows the map used for linear scaling at 150%. The dot-dash line shows linear scaling at 75%. (c) Nonlinear frequency scaling. The thin solid line shows the map used for nonlinear stretching towards the Nyquist frequency. The dashed line shows the map used for nonlinear stretching towards low frequencies. (d) Frequency rotation. The dashed line shows the frequency map for inversion around a center frequency of 1500 Hz. The dot-dash line shows the frequency map for inversion about a center frequency of 2500 Hz. All rotated stimuli were low-pass filtered at twice the center frequency.

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 11b). 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 11c 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 towards high frequencies (dotted line), and one map weighted frequencies more heavily towards 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. While 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 (Blessner, 1972; Blessner, 1969). 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, a Fast Fourier Transform 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 11d). In order 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, two center frequencies were used 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 crossmodal matching task, and a separate group of 10 participants received the word identification task.

Procedures

Crossmodal Matching. Figure 10 shows a schematic description of the crossmodal 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 msec), they 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 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 compared with "John," regardless of whether "Bob" or "John" was the target 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 recorded in a log file for later analysis.

A short familiarization period (8 trials) preceded each participant's session. During this period, the participant was presented with a crossmodal 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 crossmodal 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 typographical 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 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

Crossmodal Matching Task. Figure 12a displays the distribution of crossmodal matching scores for participants who heard frequency shifted acoustic tokens. Each boxplot represents the distribution for one shift magnitude. The figure shows that participants were, in general, successful at making crossmodal matching judgments using frequency shifted stimuli. As shown in Figure 12a, performance in each condition was significantly different from chance (0.5). Thus, frequency shifting did not significantly affect the ability to match faces and voices across modalities.

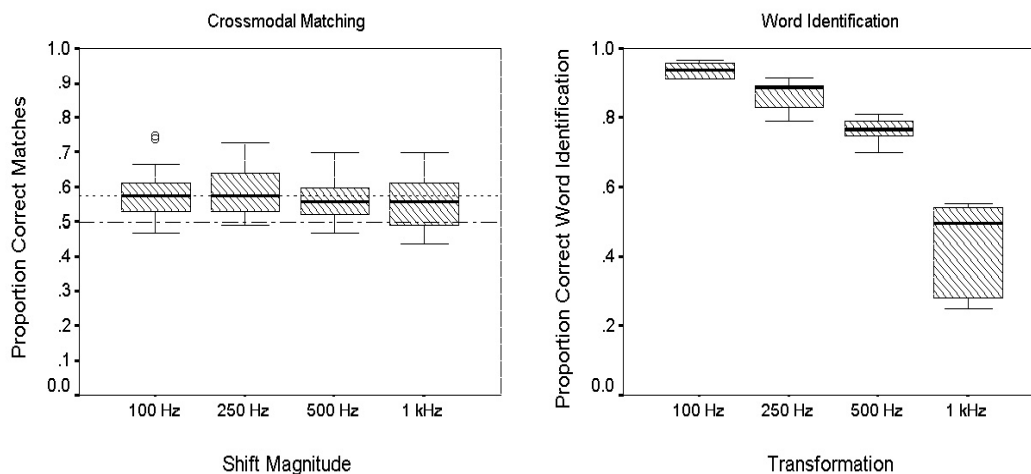


Figure 12. Results of the experiments with frequency shifted stimuli. (a) Boxplots of performance on the crossmodal matching task. Each boxplot represents the sample for a particular shift magnitude and a particular order. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the whiskers represent the range from the highest to the lowest score within the sample, excluding outliers. The circle in both 100 Hz groups indicates an outlier. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of 0.05. (b) Boxplots of word identification performance.

The data were also submitted to a $2 \times 2 \times 4$ repeated-measures ANOVA (Visual Intelligibility, Order, and Shift Magnitude). The ANOVA revealed a marginal interaction between visual intelligibility and order, $F(1, 72) = 2.79, p < 0.10$. Paradoxically, in the V-A order, there was a tendency for participants to perform better on low visual intelligibility words ($M = 0.591, SE = 0.014$) than on high visual intelligibility words ($M = 0.558, SE = 0.012$). However, in the A-V order, performance did not differ for high visual intelligibility words ($M = 0.571, SE = 0.012$) and low visual intelligibility words ($M = 0.571, SE = 0.014$). Table 2 summarizes the descriptive statistics for participants who made crossmodal matching judgments using frequency shifted acoustic stimuli.

Word Identification Task. Figure 12b 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, $F(1, 36) = 101.35, p < 0.001$. 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. Overall, frequency shifting of the acoustic signal did not affect either crossmodal matching judgments or word identification performance. Under all four frequency shift values tested, crossmodal matching performance was above chance, and word identification performance varied over a wide range. The results support the proposal that crossmodal 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.

Table 2. Descriptive statistics for participants who made crossmodal matching judgments using frequency shifted acoustic stimuli (also see Figure 12).

Order	Shift Magnitude	Mean	SE	t 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*

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Linear Frequency Scaling

Crossmodal Matching Task. Figure 13a shows the range of crossmodal 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 to the linearly shifted tokens. However, crossmodal matching was still possible even after linear frequency scaling. Table 3 summarizes the descriptive statistics for performance in each condition. As shown here, performance in all conditions was significantly different from chance (0.5).

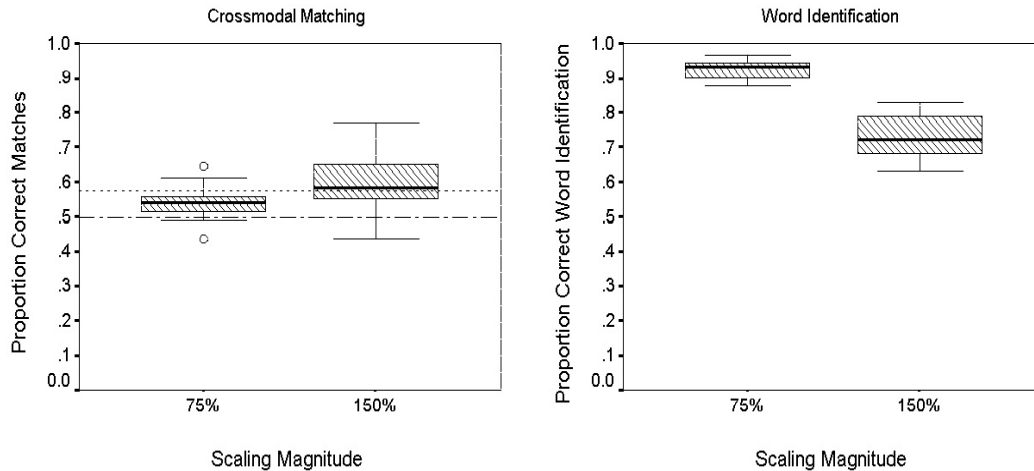


Figure 13. Results of experiments with linear frequency scaling. (a) Boxplots of performance on the crossmodal matching task. Each boxplot represents the sample for a particular scaling constant. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the whiskers represent the range from the highest to the lowest score within the sample, excluding outliers. The circles represent outliers. The dotted line represents the statistical

threshold for chance performance using a binomial test with an α of 0.05. (b) Word identification performance when the acoustic stimuli were transformed with linear frequency scaling.

A 2 x 2 x 2 repeated-measures ANOVA (Visual Intelligibility, Order, and Scaling Magnitude) also revealed a significant effect of Scaling Magnitude, $F(1, 36) = 7.06, p < 0.01$. 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.

Table 3. Descriptive statistics for participants who made crossmodal matching judgments using linear frequency scaled acoustic stimuli (also see Figure 13).

Order	Scale Magnitude	Mean	SE	t 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**

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Word Identification Task. Figure 13b shows word identification performance for the two linear frequency scaling magnitudes used in the crossmodal 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, $t(18) = 9.17, p < 0.001$. 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 crossmodal 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 crossmodal matching performance was better with the transformation that *extended* the range of frequencies covered by the three lowest formants (150% magnitude), word identification performance was better with the transformation that *reduced* the range of frequencies covered by the formants (75% magnitude). It is possible that the 150% scaling magnitude emphasized those details of the acoustic signal particularly useful for making crossmodal 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 crossmodal 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

Crossmodal Matching Task. Figure 14a shows the range of crossmodal matching scores obtained from participants who heard auditory stimuli transformed with nonlinear frequency scaling. Each boxplot represents the data obtained under one transformation. As shown in the figure, it is clear that many of the participants had a great deal of difficulty under these conditions. Table 4 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 *towards* 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 crossmodal source information, but only when the scaling was in one direction (towards the Nyquist frequency).

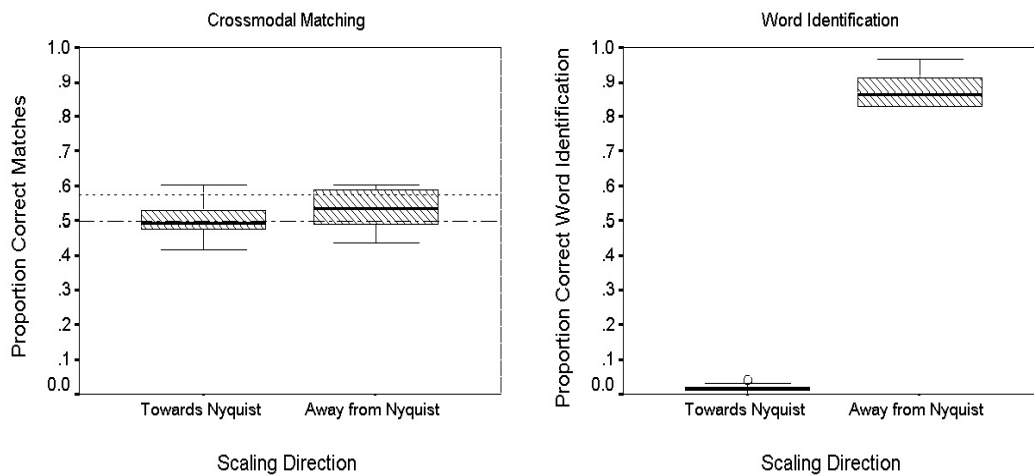


Figure 14. Results from experiments with nonlinear frequency scaling. (a) Boxplots of performance on the crossmodal matching task. Each boxplot represents the sample for a particular scaling constant. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the whiskers represent the range from the highest to the lowest score within the sample, excluding outliers. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of 0.05. (b) Boxplots of word identification performance.

Table 4. Descriptive statistics for participants who made crossmodal matching judgments with acoustic stimuli that had been nonlinearly scaled in the frequency domain.

Order	Scaling Direction	Mean	SE	t vs. 0.5
A-V	Towards Nyquist	0.494	0.015	-0.43
	Away From Nyquist	0.525	0.014	1.84*
V-A	Towards Nyquist	0.518	0.014	1.29
	Away From Nyquist	0.552	0.019	3.84*

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

A 2 x 2 x 2 repeated-measures ANOVA (Visual Intelligibility, Order, and Scaling Direction) explored this relationship further. The ANOVA revealed a significant main effect of Scaling Direction, $F(1, 36) = 4.47$, $p < 0.05$, confirming that crossmodal matching performance with stimuli scaled away

from the Nyquist frequency was better than performance with stimuli scaled towards the Nyquist frequency. The ANOVA also revealed a significant interaction between Scaling Direction and Visual Intelligibility (Figure 15), $F(1, 36) = 4.68, p < 0.05$. Simple effects analyses revealed that the interaction was supported by a difference between the two scaling directions for low visual intelligibility tokens, $t(38) = 2.93, p < 0.05$. Performance on low visual intelligibility tokens was better when scaling was away from the Nyquist frequency than when scaling was towards it. Thus, the overall advantage in crossmodal 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.

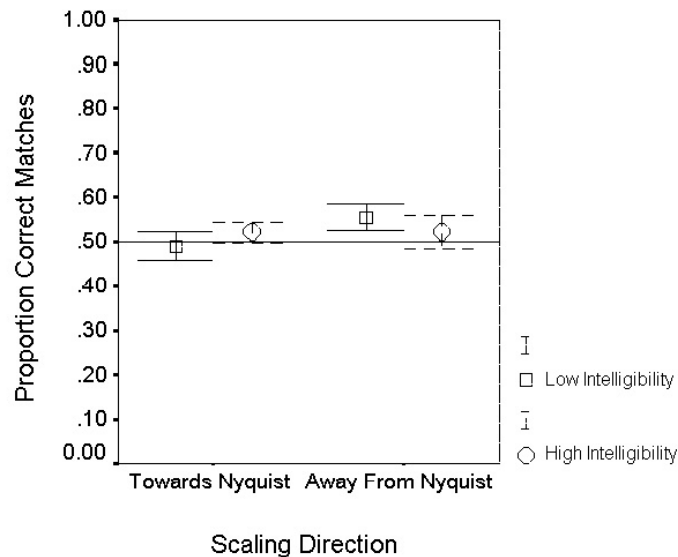


Figure 15. Marginal means for the interaction between Scaling Direction and Visual Intelligibility found on the crossmodal matching task with nonlinear frequency scaled acoustic stimuli. Error bars show standard errors.

Word Identification Task. Figure 14b 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 towards 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 crossmodal matching task for this transformation. Nonlinear scaling away from the Nyquist preserved the ability to perform the crossmodal matching task *as well as* the ability to accurately identify words. In contrast, nonlinear scaling towards the Nyquist completely eliminated the ability to perform both the crossmodal 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 crossmodal matching and word identification. The link between word identification and crossmodal matching is especially evident from the results using this acoustic transformation. One of the values of this transformation preserved crossmodal matching ability, while the other eliminated crossmodal matching task. However, the same value that preserved crossmodal source information also preserved

word identification information, while the value that did not preserve crossmodal source information eliminated word identification information. Thus, word identification information and crossmodal source information appear to be encoded together in the same properties of the acoustic signal.

Frequency Rotation

Crossmodal Matching Task. Figure 16a shows the range of crossmodal 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 (0.5). Table 5 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 x 2 x 2 repeated-measures ANOVA (Visual Intelligibility, Order, and Center Frequency). None of the main effects or interactions tested in the ANOVA reached significance ($p < 0.05$). Frequency rotation therefore appears to eliminate crossmodal source information from an acoustic stimulus.

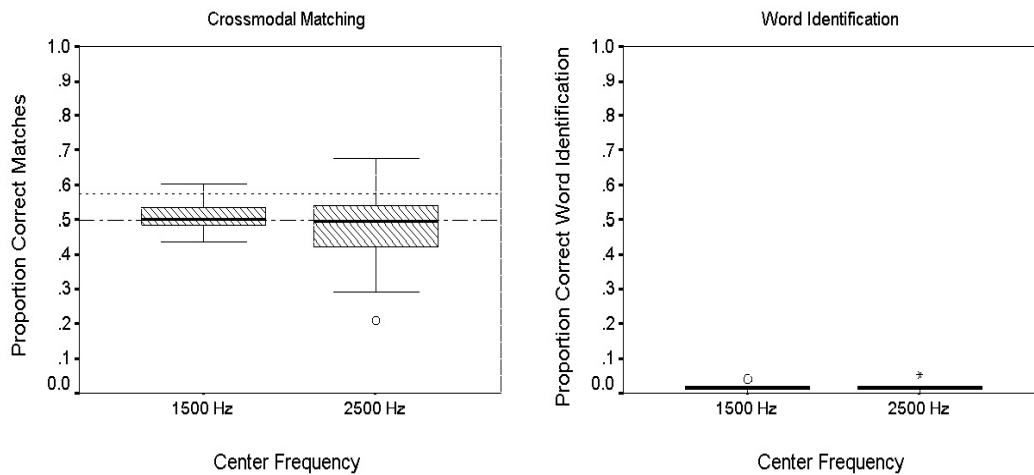


Figure 16. Results from experiments with frequency rotation. (a) Boxplots of performance on the crossmodal matching task. Each boxplot represents the sample for a particular center frequency. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the whiskers represent the range from the highest to the lowest score within the sample, excluding outliers. The circle shows an outlier. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of 0.05. (b) Boxplots of word identification performance. The circle represents an outlier and the asterisk represents an extreme outlier. Note that the scale of the dependent axis ends at 0.1.

Table 5. Descriptive statistics for participants who made crossmodal matching judgments with spectrally rotated acoustic stimuli. All t-tests failed to reach significance ($\alpha = 0.05$).

Order	Center Frequency	Mean	SE	t vs. 0.5
A-V	1500 Hz	0.507	0.012	0.62
	2500 Hz	0.483	0.033	-0.50
V-A	1500 Hz	0.524	0.017	1.46
	2500 Hz	0.473	0.038	-0.72

Word Identification Task. Figure 16b 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 around 2500 Hz ($M = 0.020$, $SE = 0.006$). Performance in the two conditions was not significantly different, $t(18) < 1$, n.s.

Summary. For both center frequencies used, the spectral rotation of acoustic signals eliminated the information necessary for both crossmodal 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 crossmodal 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

Crossmodal Matching Task. Figure 17a shows the range of crossmodal 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 6 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 (0.5), indicating that participants were able to perform the crossmodal matching task when the acoustic information had been transformed with linear temporal scaling. The results were also submitted to a 2 x 2 x 2 repeated-measures ANOVA (Visual Intelligibility, Order, and Scaling Magnitude). None of the main effects or interactions tested in the ANOVA reached significance ($p < 0.05$). Thus, linear temporal scaling of the acoustic signal appears to preserve the information necessary for crossmodal source judgments.

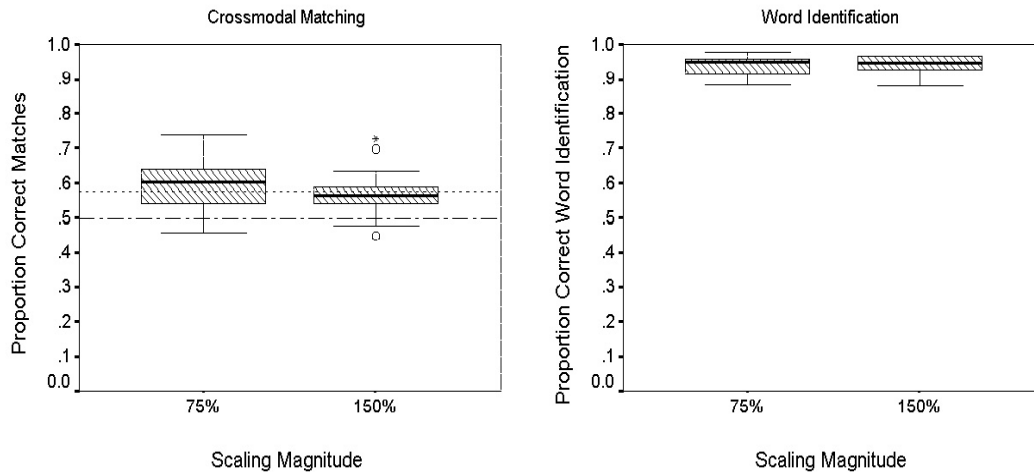


Figure 17. Results from experiments with linear temporal scaling. (a) Boxplots of performance on the crossmodal matching task with acoustic stimuli transformed with linear temporal scaling. Each boxplot represents the sample for a particular center frequency. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the whiskers represent the range from the highest to the lowest score within the sample, excluding outliers. The circle shows an outlier and the asterisk represents an extreme outlier. The dotted line

represents the statistical threshold for chance performance using a binomial test with an α of 0.05.
 (b) Boxplots of word identification performance.

Word Identification Task. Figure 17b 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, $t(18) < 1$, n.s. Participants therefore had little trouble identifying the words correctly in temporally scaled stimuli.

Table 6. Descriptive statistics for participants who made crossmodal matching judgments with acoustic stimuli that had been transformed with linear temporal scaling. 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.

Order	Scaling Magnitude	Mean	SE	t 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**

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

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 crossmodal matching performance or word recognition performance in any of the conditions. The results support the proposal that crossmodal source information, as well as phonetic information used in spoken word recognition, is carried in the relative temporal structure of the formants, and not their absolute temporal relationships.

Nonlinear Temporal Scaling

Crossmodal Matching Task. Figure 18a shows a summary of crossmodal 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 crossmodal matching task. Table 7 shows the descriptive statistics for performance in each condition. Inspection of the table reveals that overall performance on crossmodal matching with the nonlinear temporal scaling transformation only differed from chance for the V-A order. However, a 2 x 2 x 2 repeated-measures ANOVA (Visual Intelligibility, Order, Scaling Direction) showed only a marginal effect of Order, $F(1, 36) = 3.13$, $p = 0.09$. Collapsing across the order variable, mean performance for the participants was above chance for both the “accelerated” ($t(19) = 2.32$, $p < 0.05$) and “decelerated” ($t(19) = 2.87$, $p < 0.01$) conditions. Thus, although performance on the crossmodal matching task was very poor, the group as a whole did differ significantly from chance.

Table 7. Descriptive statistics for participants who made crossmodal matching judgments with acoustic stimuli that had been transformed with nonlinear temporal scaling.

Order	Scaling Direction	Mean	SE	t 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*

* $p < 0.05$

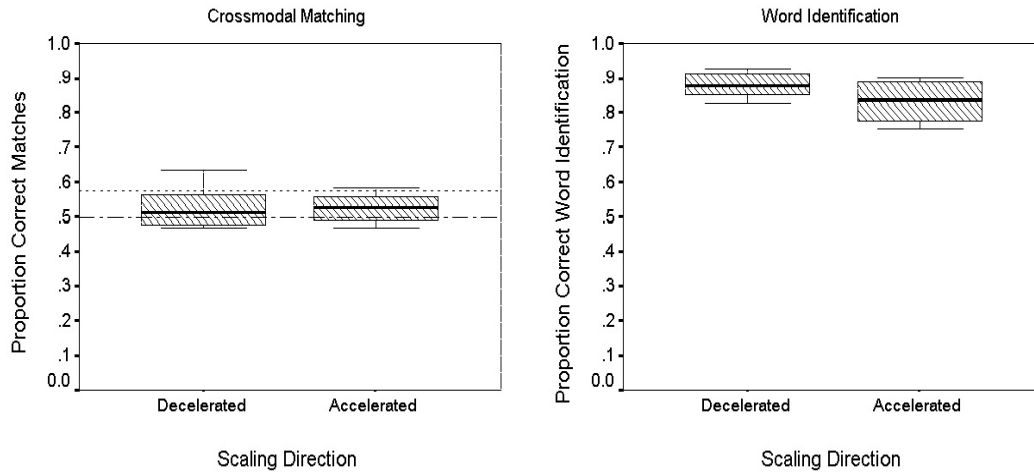


Figure 18. Results of experiments with nonlinear temporal scaling. (a) Boxplots of performance on the crossmodal matching task with acoustic stimuli transformed with nonlinear temporal scaling. Each boxplot represents the sample for a particular center frequency. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the whiskers represent the range from the highest to the lowest score within the sample, excluding outliers. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of 0.05. (b) Boxplots of word identification performance.

Word Identification Task. Figure 18b 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, $t(18) = 2.21$, $p < 0.05$.

Summary. Overall, the nonlinear temporal scaling transformations reduced crossmodal matching performance severely, but overall did not affect crossmodal matching ability or word identification performance.

Discussion

In the present investigation, spoken utterances of single talkers speaking 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. Four

transformations were used to manipulate relationships in the frequency domain, while preserving relationships in the temporal domain. Under frequency shifting, absolute differences in formant frequency were preserved, but were moved to a different region of the frequency space. Under the linear frequency scaling transformation, only the *relative ratios* of formant frequencies were preserved, while the absolute differences between frequencies were linearly scaled up or down. Under nonlinear frequency scaling, both the absolute and relative spacing of formant frequencies were eliminated while the orientation of the formants was maintained with respect to the frequency axis, such that the first formant always had a lower frequency than the second formant, and so on. Finally, under the frequency rotation transformation, the orientation of the formant frequencies with respect to the frequency axis was rotated.

Two other transformations were also used to manipulate relationships in the temporal domain, while preserving relationships in the frequency domain. Under linear temporal scaling, the overall duration of the original utterance was extended or reduced. This preserved the relative relationships among temporal components. Under nonlinear temporal scaling the overall duration was also extended or reduced, but in such a way that the rate of extension or reduction increased over the course of the original utterance.

The 14 transformations were used in both a crossmodal matching task and a word identification task, to assess whether each transformation class preserved or eliminated crossmodal 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 crossmodal 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 was 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, crossmodal source information would also be preserved and participants would be able to successfully carry out the matching task and the word identification task.

Because of the strong link between crossmodal 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 support crossmodal matching. For transformations that disrupted crossmodal matching performance, we predicted that word identification accuracy would also be severely reduced.

The results from the crossmodal matching task using the four spectral domain transformations showed that crossmodal source information is retained when the ratio of spacing between formants remains invariant across the transformation. With manipulations in the spectral domain, however, crossmodal 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 the latter two transformations do not, this pattern of results provides support for the hypothesis that crossmodal 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 tight link observed earlier between crossmodal source information and word identification information. Those transformations that preserved crossmodal matching also provided information needed for accurate word identification, and transformations that reduced crossmodal matching also reduced word identification performance.

Both temporal domain transformations preserved the ability to do the crossmodal matching task. As expected, linear temporal scaling did not degrade performance on the crossmodal matching task or the word identification task, indicating that crossmodal source information and word identification information in the spectral patterning of formants 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 crossmodal source information, it is possible that, because the temporal domain scaling was not continuous (the signal was split into ten 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 (for example, at formant transitions, or within the duration of a vowel segment). Good word identification performance was also observed under both transformations, confirming that crossmodal source information and word identification are intimately tied together in the acoustic signal. Any further work investigating the temporal nature of crossmodal source information will have to await the development of a method of nonlinearly scaling temporal information continuously over the duration of the utterance.

Taken together, the results from the six acoustic transformations provide new insights into how crossmodal source information is specified in acoustic patterns of speech, and the ways in which source information is used in spoken word recognition.

General Discussion

The results of the present set of perturbation experiments provide support for the proposal that acoustic and optical displays of speech carry specific, modality-neutral information about speech that reflects the underlying dynamics of vocal tract articulation. We found that crossmodal matching was only possible under signal transformations that preserved the relative spectral and temporal patterns of formant frequencies as they evolved over time, indicating that articulatory information is specified in *relative*, and not *absolute*, spectral and temporal relationships.

Crossmodal matching performance remained unchanged under precisely the same acoustic transformations that supported word identification performance. This is not a trivial result. Explicit identification of the target word is not necessary for crossmodal matching; at no point in the crossmodal 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 is *identical*. Because both crossmodal 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.

Modality-specific constraints on the transmission of crossmodal source information. The ability of an energy array to carry information about an event is necessarily limited by the extent to which the perceiver is sensitive to structure in the energy and the extent to which the energy can be structured by the unique active components of the event itself (Fowler, 1996; Stoffregen & Bardy, 2001). This is perhaps nowhere more apparent than in the perception of speech and processing of spoken language. Optic displays may be able to carry information relating to the positions of the lips, jaw and tongue tip, but they are inherently limited in their ability to specify the positions and movements of more internal articulators, such as the glottis, velum, and tongue body. In contrast, acoustic displays are necessarily shaped by articulators from the vocal cords to the lips. However, the acoustic display is also limited in its ability to specify differences between similar articulatory events;

[f] and [T] are perceptually nearly identical when considered acoustically. Indeed, one of the most remarkable features of audiovisual speech is the surprising complementarity exhibited across auditory and visual sensory modalities: contrasts that are particularly hard to discriminate in the auditory domain are relatively easy to discriminate visually, and vice versa (Massaro, 1998; Summerfield, 1987).

These modality-specific constraints are, however, more valuable than they at first appear, because they delimit the bounds of unimodal sensory stimulation that can be perceptually relevant. The fact that [f] and [T] are nearly identical in auditory-alone contexts is an extremely important piece of evidence about the ability of acoustic patterns to carry perceptually relevant information about gestural distinctions between the labiodental and interdental place of articulation. In the same manner, the transformations used in the present experiments are also relevant to the ability of acoustic patterns to carry perceptually relevant information about the gestural activity used in speech production. The findings presented here demonstrate that information about articulation is encoded acoustically on a relative scale; absolute differences in frequency and temporal patterning were not necessary for either crossmodal matching or word identification, so 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 to absolute frequency values or relationships.

Another interesting property of speech perception 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, crossmodal 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 crossmodal 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 earlier work on crossmodal matching (Lachs, 2002): although the noise-band transformation eliminates f_0 from the signal, accurate crossmodal matching performance was possible with noise-band stimuli which did not have an actual f_0 included. 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 formant relationships may be conditional upon the experience that formant relationships within a particular range (say, 0 – 5 kHz) are related. 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.

Interestingly, a parallel line of investigation (Harnsberger, Svirsky, Kaiser, Wright & Meyer, 2001) has been conducted with users of cochlear implants - 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 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 complete, 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 “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. First, participants were presented with an orthographically presented word. Listeners were then asked to locate an auditory token corresponding to the vowel of the orthographic word in a two by two matrix of synthesized F1-F2 combinations. The matrix was arranged such that each row represented a frequency step of 0.377 Barks 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 orthographic vowel. Harnsberger et al. found that the perceptual vowel spaces of the listeners with cochlear implants were not shifted 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 up to 1000 Hz can also be tolerated by normal hearing listeners. Taken together with the results of the other 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.

Modality-neutral forms of crossmodal source information. In discussing their results on the perception of talker specific information from sinewave 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 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 (Gaver, 1993; Gibson, 1966). The present findings show that those aspects of an acoustic speech signal that relate to both indexical and lexical variation are also carried in visual speech signals. Moreover, these distinctive perceptual aspects 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 articulatory information. Because the articulatory behavior of a talker has acoustic as well as optical correlates, the common origin of patterns in either modality suggests a natural and principled basis for making crossmodal judgments of talker identity based on articulatory activity.

Is it possible that the form of speech representation is modality-neutral and based on common 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 specify 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 et al., 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 are caused by the same articulatory behaviors which lawfully structure optical patterns.

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 also 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 displays that isolate kinematic information with the point-light technique are sufficient for the perception of biological motion (Johansson, 1973) and, as explained above, can be used in speech perception studies (Rosenblum, Johnson, & Saldaña, 1996). Analogously in the acoustic domain, sinewave 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. These highly unusual patterns can also be perceived as speech. It is possible that these two, domain-specific forms of kinematic information may 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 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 apparently 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). 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 explanatory device for observed audiovisual speech perception phenomena, and may also lead to more powerful theoretical descriptions of speech perception in general.

Chapter IV: Specification of Crossmodal Source Information in Isolated Kinematic Displays of Speech

Optical information about articulation has been shown to have substantial effects on speech perception (Massaro & Cohen, 1995; McGurk & MacDonald, 1976; Sumbly & Pollack, 1954). In the absence of auditory stimulation, visual information is sufficient to support accurate spoken word recognition (Bernstein et al., in press; Bernstein et al., 2000; Lachs & Pisoni, submitted). Combined with acoustic information, visual stimulation can also enhance speech intelligibility in noise by +15 dB (Middleweerd & Plomp, 1987; Sumbly & Pollack, 1954; Summerfield, 1987). Conflicting information in the acoustic and optic displays of an audiovisual speech stimulus can also interact to form illusory percepts in speech perception (the "McGurk" effect MacDonald & McGurk, 1978; McGurk & MacDonald, 1976).

These core phenomena in audiovisual speech perception are robust and reliable (Summerfield, 1987), and provide strong evidence that some components of the speech perception process must also be involved in integrating different sources of information across sensory modalities. A great deal of research, especially within the past two decades, has focused on the nature of the perceptual integration process – how sensory stimulation from disparate and seemingly incommensurate modalities (cf., Stoffregen & Bardy, 2001) can influence the perception of a multimodal stimulus.

The precise nature of the integration process has been a matter of considerable debate and several theories of perceptual integration have emerged from these research efforts. Schwartz, Robert-Ribes, and Escudier (1998) have proposed a taxonomy of the various theories of multisensory integration that focuses on three hierarchically embedded, empirical questions. The first question deals with whether the optical and acoustical information remains in its raw sensory form during perceptual processing, or is first translated into some common representational format for later analysis. The consensus on this issue seems to be that some intermediate format of shared representation is necessary to explain several basic findings in the audiovisual speech perception literature. For example, infants can detect discrepancies in acoustic and optic displays of speech, indicating that some comparison of information in the two modalities is possible even at early developmental stages (Kuhl & Meltzoff, 1984; Lewkowicz, 2001).

The second question in the Schwartz et al. taxonomy concerns whether intermodal integration occurs early (i.e., before linguistic processing) or late (i.e., after linguistic processing). In one experiment designed to investigate this question, Green and Kuhl (1989) showed that the perceived VOT boundary for a synthetic /bi-/pi/ continuum, when paired with a visual display of a talker uttering the syllable /gi/, was shifted toward the VOT boundary for a /di-/ti/ continuum. The VOT boundary for the continuum shifted in a manner that was consistent with the *integrated* percept invoked by the McGurk illusion, and was inappropriate for the actual syllables presented in the acoustic signal. Their findings suggested that linguistic analysis occurred *after* perceptual processing had been completed.

In another study, Green and Miller (1985) synthesized an acoustic continuum along the voicing dimension from /bi/ to /pi/. The stimuli were then presented in conjunction with visual speech information that varied in speech rate to observers who were asked to categorize the initial phoneme of the syllable. The results showed that the perceived voicing boundary was conditional on the speaking rate presented in the visual signal, even though the speaking rate of the acoustic signal never changed. Thus, the linguistic process of phoneme identification was evaluated in the context of *both* auditory and visual information about speech. Again, these results suggest that linguistic processing occurs

after the combination and integration of optic and acoustic information, supporting an early integration model.

Finally, in a recent series of studies, Grant (Grant, 2001; Grant & Seitz, 2000) has shown that concurrent visual speech stimulation decreased the detection thresholds of auditory speech signals that were masked with white noise, suggesting that acoustic and optic information interact at early levels of signal detection, which presumably takes place before any linguistic analysis the occurred (however, see Grant, 2001 for appropriate caveats).

Despite the wealth of available evidence (see also Radeau & Colin, 2001), the question of early vs. late integration in audiovisual speech perception has not been resolved and the matter still remains a topic of considerable debate in the literature. However, even late-integration models like FLMP (Massaro & Oden, 1995), have had to be altered considerably to accomodate the need for the pre-linguistic evaluation of some interactive, crossmodal information (see Massaro, 1998).

The third and final question proposed by Schwartz et al. (1998) deals with the precise form of any common representational format. According to their analysis, there are two alternatives to this question. In one representational format, information from one modality is translated or recoded into a format that is compatible with another, more familiar or “dominant” modality. Unfortunately, support for the dominant modality recoding hypothesis is scarce or non-existent, and Schwartz et al. present it simply as a logical possibility resulting from their taxonomic scheme rather than as a viable alternative supported by empirical evidence. They assert that models based on this type of a representation necessarily predict that visual influences on speech perception will only be observed when optical information conflicts with the auditory stimulus, or when the auditory stimulus is degraded. This prediction was disconfirmed, however, many years ealier by Sumby and Pollack (1954), who found that auditory-alone performance improved as the signal-to-noise ratio increased. After scores in the audiovisual condition were normalized relative to this increasing baseline, Sumby and Pollack discovered that the amount by which performance improved due to audiovisual stimulation remained constant over the entire range of signal-to-noise ratios tested. This important finding which is often overlooked by researchers and theorists outside the domain of audiovisual speech perception indicates that the effect of the additional visual information is not conditional on the degree of ambiguity in the auditory signal.

The current available evidence therefore indicates that translation of information into a dominant modality format is not a viable option. Instead, Schwartz et al.’s alternative “common format” model assumes that acoustic and optic information about speech is analyzed with reference to some kind of amodal or modality-neutral (Fowler, 1986; Fowler & Rosenblum, 1991) representational space. In most theoretical approaches of this type, the perceptual space relates to the vocal tract articulation that underlies the production of auditory or visual speech signals (Fowler, 1996; Liberman & Mattingly, 1985; Summerfield, 1987).

Convincing evidence for a modality-neutral form of speech information comes from a study of the McGurk effect using auditory and *tactile* information about speech. Fowler and Dekle (1991) had naïve participants listen to spoken syllables synthesized along a /ba-/pa/ continuum while using their hands to obtain information about the articulation, in much the same way that deaf-blind users of the Tadoma method (Schultz et al., 1984) do. The tactile information on every trial was either a /ba/ or a /pa/, articulated by a talker who was unable to hear the auditory syllable. Observers were then asked to categorize the heard stimulus as either a /ba/ or a /ga/ in a forced choice task. Fowler and Dekle found that the auditory perceptual boundary for /ba/ and /ga/ was shifted by the phonetic content of the tactile pattern. Interestingly, this effect was found with subjects who had no training in Tadoma at all,

indicating that stored associations between tactile and auditory speech gestures could not be involved in the integration results.

In a second experiment, observers were simultaneously presented with the acoustic syllables and *orthographic* displays of “BA” or “PA.” Because all of the observers were literate college undergraduates, Fowler and Dekle (1991) assumed that they had stored robust associations between the orthographic symbols and their phonetic counterparts in memory. However, the visual orthographic displays did not affect the position of the category boundary along the continuum at all. Fowler and Dekle interpreted these 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. Their results also demonstrate that some degree of useful information about speech can be obtained through sensory modalities other than audition and vision.

Further support for the modality-neutral form of speech information comes from a recent series of studies using a novel crossmodal matching task (Lachs, 2002). In the crossmodal matching task, participants view a visual-alone, dynamic display of a talker speaking an isolated English word and are asked to match the pattern with one of two auditory-alone displays. One of the alternative auditory displays is the acoustic specification of the *same articulatory events* that produced the visual-alone pattern. The other alternative is the acoustic specification of a different talker saying the same word that was spoken in the visual-alone pattern. When matching a visual-alone token to one of two auditory-alone tokens, the “order” is said to be “V-A.” In contrast, when participants hear an auditory-alone display first, and then are asked to match to one of two visual-alone alternatives, this is referred to as the “A-V” order.

Lachs (2002) found that participants were able to correctly match the same phonetic events across sensory modalities, regardless of the order in which the modalities were presented. That is, observers could correctly match a speaking face with one of two voices, or correctly match a voice with one of two speaking faces. Lachs argued that the results provided evidence for the existence of “crossmodal source information,” that is, modality-neutral information about the source or speaker of an utterance that is specified by sensory patterns of both acoustic and optic energy. In a series of further experiments, Lachs (2002) elaborated on this finding using a series of manipulations to the acoustic and optical displays that were presented for crossmodal matching. One experiment found that participants were not able to match across modalities when optical and acoustic patterns were played backwards in time, indicating that crossmodal source information is sensitive to the normal temporal order of spoken events. Another experiment showed that static visual displays of faces could not be matched to acoustic displays. This result suggested that crossmodal source information is specified in the dynamic structure of visual displays, and not in their static features. Finally, an additional experiment showed that noise-band, “chimaeric” stimuli (Smith et al., 2002) preserved crossmodal source information, despite the elimination of a traditional acoustic cue to vocal identity: fundamental frequency (f_0). This result was interpreted as evidence that crossmodal source information is specified in the pattern of formant resonances as they evolve over time, irrespective of the f_0 used to excite them. The results of this series of experiments suggest that crossmodal matching can be carried out because of a common articulatory basis for acoustic and optic displays. Lachs proposed that, because both visual and auditory displays of speech are lawfully structured by the same underlying articulatory events, matching is accomplished by comparison to a common, underlying source of information about vocal tract activity.

The proposal that audiovisual integration reflects a common underlying articulatory basis for audiovisual speech integration is also supported by a correlational analysis of the stimuli in Grant’s (2001) signal detection experiments. His analysis showed that the rms amplitude of the acoustic signal

(especially in the bandpass filtered F2 region) was strongly related to the area of the opening circumscribed by the lips. Grant proposed that this correlation between an acoustic variable and an optical variable may have been responsible for his finding that auditory speech detection thresholds were reduced with concurrent visual stimulation. It is interesting to note that the relationship reported by Grant between the auditory and visual variables is also related to a common, underlying kinematic variable based on jaw and mouth opening.

An articulatory foundation for modality-neutral information is also consistent with the observation that the visual correlates of spoken language arise as a direct result of the movements of the vocal tract that are necessary for producing linguistically significant sounds (Munhall & Vatikiotis-Bateson, 1998; Vatikiotis-Bateson et al., 1997). An articulating vocal tract structures acoustic and optic energy in space and time. As a consequence, the auditory and visual sensory information present in those displays is lawfully structured by a common unitary articulatory event. Thus, the disparate patterns of acoustic and optic energy are *necessarily* and *lawfully* related to each other by virtue of their common origin in articulation.

These observations recast the problem of audiovisual integration in speech perception in a form that is consistent with the direct realist approach to speech perception (Fowler, 1986; Gibson, 1966; Gibson, 1979). "Ecological" theories of perception are based on the findings that patterns in light and sound are lawfully related to the physical events that cause them. For example, the ambient light in a room is reflected off multiple surfaces, like the floor, walls, tables, chairs, etc. before it stimulates the retina. The pattern of light observable from the point-of-view of an active observer, then, is highly *structured* with respect to the relationship between the observer and his or her environment. Furthermore, this structure is lawfully determined by the physical properties of light. In the same way, events in the environment that have acoustic consequences structure patterns in pressure waves by virtue of the physical properties of sound (Gaver, 1993). Direct realists therefore assert that direct and unmediated perceptual access to the causes of ecologically important events is possible by virtue of the lawful relation between the patterns of energy observable by a perceiver and the events that caused them.

According to this theoretical perspective, the question for audiovisual speech research is not to determine how auditory and visual signals are translated into a common representational format, but rather to discover the ways in which optical and acoustical information structure or specify the informative properties of vocal tract articulation. A growing body of research suggests that both the acoustic and optical properties of speech carry *kinematic* or *dynamic* information about the articulation of the vocal tract (Rosenblum, 1994), and that these sources of information drive the perception of linguistically significant utterances (Fowler, 1986; Fowler & Rosenblum, 1991; Liberman & Mattingly, 1985; Rosenblum & Saldaña, 1996; Summerfield, 1987). Kinematic variables refer to position and its time derivatives: velocity, acceleration, etc.; dynamic variables refer to forces and masses (Bingham, 1995). These variables are necessarily grounded in spoken events: they do not refer to specific properties of visual or auditory stimulation, but rather to the events that structure light or sound. As such, they are prime theoretical candidates for the perceptible properties of events.

The point-light technique developed by Johansson (1973) has been used extensively to investigate the perception of kinematic variables (e.g., Bingham, 1987; Cutting & Kozlowski, 1977; Kaiser & Proffitt, 1984; Kozlowski & Cutting, 1977, 1978; Runeson & Frykholm, 1981). By placing small reflective patches at key positions on a talker's face and darkening everything else in the display, researchers have been able to isolate the kinematic properties of visual displays of talkers articulating speech (Rosenblum et al., 1996; Rosenblum & Saldaña, 1996). These "kinematic primitives" have been shown to behave much like unmodified, fully-illuminated visual displays of speech (Rosenblum & Saldaña, 1996). Indeed, Rosenblum and Saldaña found that the McGurk illusion can be induced by

dubbing visual point-light displays onto phonetically discrepant auditory syllables. In addition, Roseblum, Johnson, and Saldaña (1996) also demonstrated that providing point-light information about articulation in conjunction with auditory speech embedded in noise can result in increased speech intelligibility scores, just as fully-illuminated visual displays can improve speech intelligibility (Middleweerd & Plomp, 1987; Sumbly & Pollack, 1954).

As with optic displays, kinematic information can also be selectively isolated and studied in acoustic displays of speech. “Sinewave speech replicas” (Remez et al., 1981) are time-varying acoustic signals that contain three sinusoidal tones generated at the frequencies traced by the centers of the three lowest formants in a speech signal. Because the formants are resonances of the vocal tract, that is, bands of energy at frequencies specified by the evolving configuration of the vocal tract transfer function (Stevens, 1998), these “sinewave speech replicas” can be said to isolate kinematic information in the auditory domain. Pioneering experiments by Remez et al. using sinewave speech replicas have demonstrated that subjects can perceive relevant linguistic information from these minimal kinematic displays in many of the same ways that they can from untransformed acoustic displays (Remez et al., 1994; Remez et al., 1981).

In summary, kinematic information about articulation is carried by both acoustic and optic energy, and these sources of information appear to be sufficient to support speech perception in each modality alone. The results reported in the literature support the novel proposal that the object of the speech perception system is to recover the dynamics of articulation via the transmission over various sensory media of kinematic properties of articulator movement as they unfold over time (Rosenblum, 1994; Runeson & Frykholm, 1983; Summerfield, 1979).

Experiment 6: Crossmodal Matching of Kinematic Primitives

The conclusion that the information relevant to speech perception is modality-neutral and fundamentally articulatory in nature is consistent with the crossmodal matching results reported by Lachs (2002). Because the object of perception – the articulatory movement of the vocal tract - is the same, regardless of the particular sensory domain being used, the particular modality through which a perceiver receives information about articulation should be, to some extent, irrelevant. According to this view, reliable crossmodal matching judgments are possible because the object of perception under visual-alone or auditory-alone conditions is ultimately *identical* – that is, kinematics are a property of the event that produced the stimulus, and information about those kinematics is contained in any energy pattern structured by the event, whether it be acoustic, optical, or even haptic. Of course, there are inherent limitations in the kind of articulatory information that can be carried by either display (Summerfield, 1987): visual displays can no more distinguish whether the vocal cords are vibrating (as in /b/ vs. /p/) than auditory displays can distinguish between friction produced with a labiodental or interdental place of articulation (as in /f/ vs. /T/). However, factoring in the differential ability of various energy media to carry information about the vocal tract, the crossmodal matching task is theoretically no different from a task in which participants are presented with a “unimodal matching task,” in which an acoustic display of speech is matched to one of two acoustic displays of speech, one of which is identical to the target¹. In both the crossmodal matching task and the traditional unimodal matching task, the perceiver is matching kinematic information about the vocal tract with kinematic information about the vocal tract.

¹ Note the similarity of this proposed experiment to the classic matching experiments used by Posner and colleagues to examine the representation of letters in visual memory (Posner, Boies, Eichelman, & Taylor, 1969; Posner & Keele, 1967).

The present experiment was designed to test the hypothesis that crossmodal source information is specified in the isolated kinematic behavior of the vocal tract. To accomplish this goal, we used the crossmodal matching task and asked participants to crossmodally match point-light displays with sinewave speech replicas. If the object of perception is modality-neutral and based on the articulation of the vocal tract, we expected that observers would be able to match faces and voices with only the minimal, isolated kinematic information about articulation available in point-light displays and sinewave replicas of speech.

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

Point-light displays of speech. To capture and record the isolated movement of the articulators, it was necessary to videotape an additional set of 4 talkers from those used in the previous experiments. Because the motion of the selected points must be recorded directly, the motion of the articulators cannot be extracted from a fully-illuminated display of the talker's face (Runeson, 1994). Four females between the ages of 23 – 30 volunteered as talkers for the recording session. These four females were not used as talkers in any of the stimulus sets used in the previous experiments. Before recording, each talker glued glow-in-the-dark dots, each approximately 3 mm in diameter, to her face in the pattern outlined in Figure 19. For dots on the outside of the mouth, the adhesive used was a spirit gum commonly used for adhesion of latex masks (Living Nightmare® Spirit Gum). Dots on the lips, teeth, and tongue were affixed with an over-the-counter dental adhesive (Fixodent® Denture Adhesive Cream).

Stimuli were recorded using a digital video camera and microphone (Sony AKGC414) in a sound-attenuated room onto Sony DVcam PDV-124ME digital media tapes. Each talker was seated approximately 56 inches from the camera lens, and the zoom control on the camera was adjusted such that the visual angle subtended by the distance between the talker's ears was equal across talkers. The camera lens and face of the talker were placed at a height of approximately 44 inches. Two black lights (15.7 cm long, 15 watts each) were secured on permanent fixtures placed 13.5 inches on either side of the camera lens, on an axis perpendicular to the line between the talker's face and the camera lens. The two black lights were the only source of illumination used during videotaping, and did not significantly illuminate the skin of the talker being recorded. However, the glow in the dark dots reflected the black light in the normal visible spectrum. The video track of the recorded movies thus recorded only the movement of the glow-in-the-dark points in isolation of the face to which they were affixed.

During videotaping, talkers read a list of 96 English words off a teleprompter configured with a green font and black background so that it provided no additional ambient illumination in the recording session. There was a 3 second delay between the presentation of successive words on the teleprompter. The digitized stimuli were segmented later such that each word was preceded and followed by 10 silent frames ($10 * 30 \text{ fps} = 333.3 \text{ msec}$) with no speech sounds.

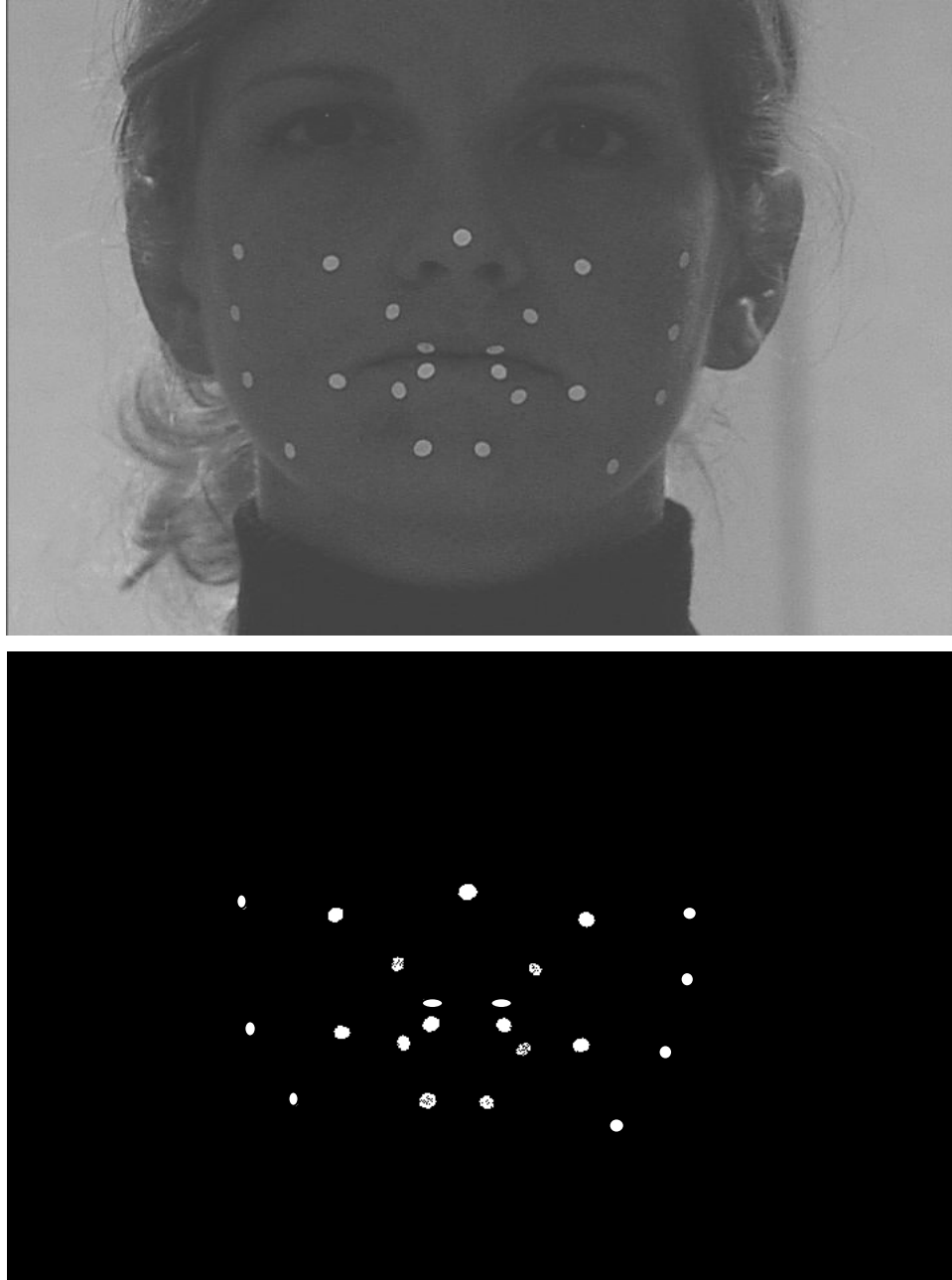


Figure 19. Dot configuration for point-light displays. All dots were of a uniform diameter. Five dots are not visible due to occlusion by the lips: upper teeth (2 dots), lower teeth (2 dots), and tongue tip (1 dot).

Sinewave speech replicas. The audio track of each utterance was extracted from the videotape and stored in a digital audio file for transformation into sinewave replicas of speech (Remez et al., 1981). Sinewave speech transformations were created by tracking the frequencies and amplitudes of the first four formants as they varied over time.

These acoustic measurements were obtained in a two-step process. First, each sound file was resampled to 10 kHz. The resampled sound file was then broken into windows of 10 msec each. Each

window was the subject of an 8th order LPC analysis, and the four coefficients with the highest magnitudes were then converted to frequencies and magnitudes and stored in a data file. Each sound file thus had an associated data file with 8 parameters (4 frequencies and 4 magnitudes) per 10 msec window, corresponding to the change in the formants in the original sound file over time.

Sometimes the LPC analysis would output erroneous or spurious noises in the original sound file. To deal with this problem, the automated output file was scrutinized visually by hand to determine the accuracy of the automated process. Overall, the automated process proved to be an excellent starting point, although some slight adjustments in frequency and amplitude were necessary in several windows for most of the stimuli. The automated process had some difficulty in tracking formants during unvoiced portions of the speech waveform corresponding to periods of aspiration, consonant transitions, and frication. Adjustments in frequency and amplitude were made by hand using MATLAB.

The output of the two-stage process was a new data file specifying eight parameters (four frequencies and four associated amplitudes) for each 10 msec window in the original sound file. These data were then submitted to a synthesis routine (Ellis, 2001) that produced four sinusoidal tones that varied over time according to the parameters output by the measurement process.

Figure 20 shows examples of the spectrograms of an original sound file from the stimulus set and its corresponding sinewave speech replica. As shown in the figure, the sinewave replica eliminated all extraneous information from the sound file other than the variation of the formants over time.

Visual Intelligibility. The words spoken by the four talkers used to generate the point-light displays were the same words used in previous investigations of crossmodal source matching (Lachs, 2002). These words were divided equally into high visual intelligibility words and low visual intelligibility words. However, it should be noted that no measure of the visual-only intelligibility of the point-light displays was taken. The classification of a word as high or low visual intelligibility was based on lipreading performance obtained earlier with an entirely different set of *fully-illuminated* visual displays of speech (see Lachs & Pisoni, submitted).

Familiarization Phase Stimuli. There was also a short familiarization phase so that participants could become accustomed to the unnaturalness of the sinewave speech replicas. The stimulus materials used during the familiarization phase were 20 isolated, monosyllabic English words that were not used in the final test list. These 20 words were spoken by a different talker who was not used to create the stimuli used in the point-light display database. The 20 familiarization words were converted into sinewave speech replicas using the same methods outlined earlier.

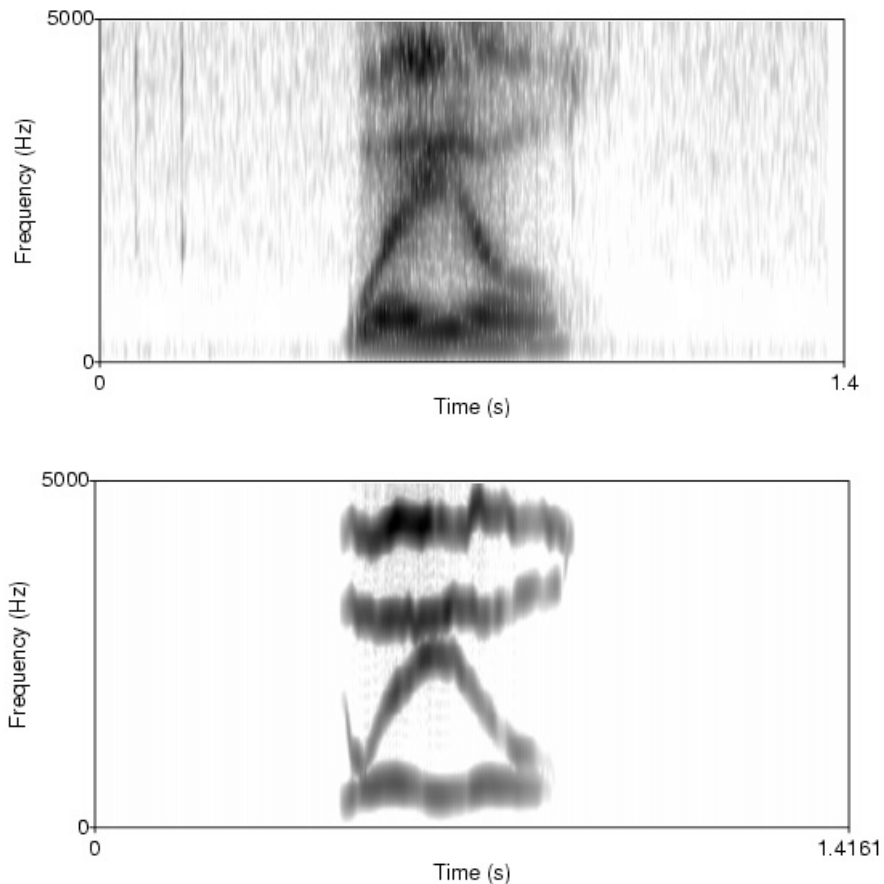


Figure 20. The top panel shows a spectrogram of talker F1 speaking the word “WAIL.” The bottom panel shows a spectrogram of a synthetic sinewave replica of this token.

Procedure

During the familiarization phase, participants heard a sinewave speech replica of a spoken word and then the original, untransformed utterance from which that replica was constructed. This sinewave-original pair was presented 3 times in succession. After the third presentation of the stimulus pair, participants were asked to rate on a three-point scale how closely the sinewave replica matched the natural utterance. No feedback was provided. The familiarization task gave participants some exposure to the unusual nature of sinewave speech replicas without explicitly instructing them on how to identify words from them.

Figure 21 shows a schematic description of the crossmodal matching task. Participants in the “V-A” condition were first presented with a visual-alone point-light display video clip of a talker uttering an isolated English word. Shortly after seeing this video display (500 msec), they were presented with two auditory-alone, *sinewave replicas* of speech. 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 the audio clip that matched the talker they had seen (“First” or “Second”). Similar instructions were provided for participants in the “A-V” condition. These participants heard the audio clip first, and had to make their decision based on two video displays.

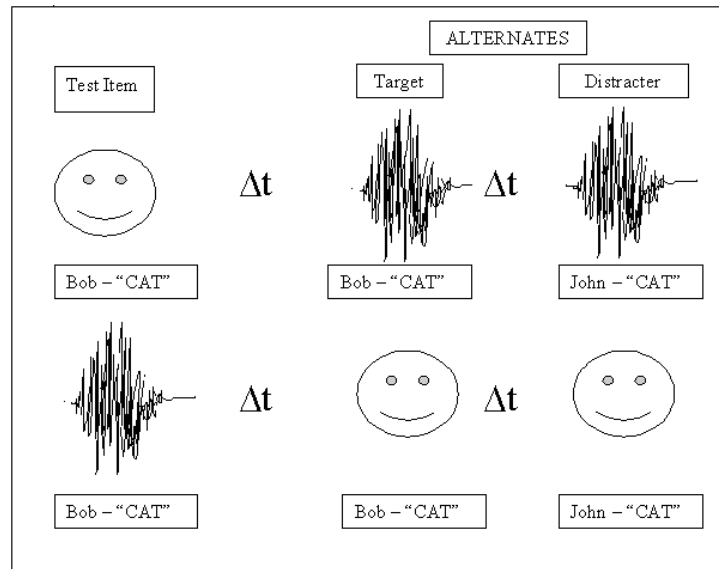


Figure 21. Schematic of the crossmodal matching task. The top row illustrates the task in the “V-A order”. The bottom row illustrates the task in the “A-V order”. Faces denote stimuli that are presented visual-alone. Waveforms denote stimuli that are presented auditory-alone. Δt is always 500 ms.

On each trial of the experiment, 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 contrasted against one and only one other talker for all trials in the experiment. For example, “Mary” was always contrasted with “Jane,” regardless of whether “Mary” or “Jane” was the target on the trial. In addition, all the talkers viewed by a particular participant were of the same gender. Responses were entered by pressing one of two buttons on a response box and transferred to a log file for further analysis.

A short training period (8 trials) preceded each participant’s test session. During the training period, the participant was presented with a crossmodal matching trial and asked to pick the correct alternate. During training only, the response was followed immediately by feedback. The feedback consisted of playing back the entire audiovisual movie clip of the test word. Half of the participants matched point-light displays with the original, untransformed sound files; the other half of the participants matched point-light displays with sinewave replicas of speech.

Results

Figure 22 shows performance on the crossmodal matching task when point-light displays (PLD) were matched to untransformed auditory-only (AO) speech tokens (PLD – AO) or sinewave speech (SWS) tokens (PLD – SWS). The figure shows that crossmodal matching performance with point-light displays was quite poor, although the majority of participants performed better than chance (0.5). Table 8 provides a summary of the descriptive statistics for each of the conditions pictured in Figure 22. As shown in the table, performance in both matching conditions and both orders differed significantly from chance, indicating that crossmodal matching judgments can be carried out using

only isolated kinematic displays in both the visual and auditory sensory domains. The data were submitted to a 2 x 2 x 2 repeated measures ANOVA (Visual Intelligibility, Order, and Matching Condition). The ANOVA did not reveal any significant main effects or interactions, indicating that crossmodal matching performance was not significantly affected by the visual intelligibility of the words.

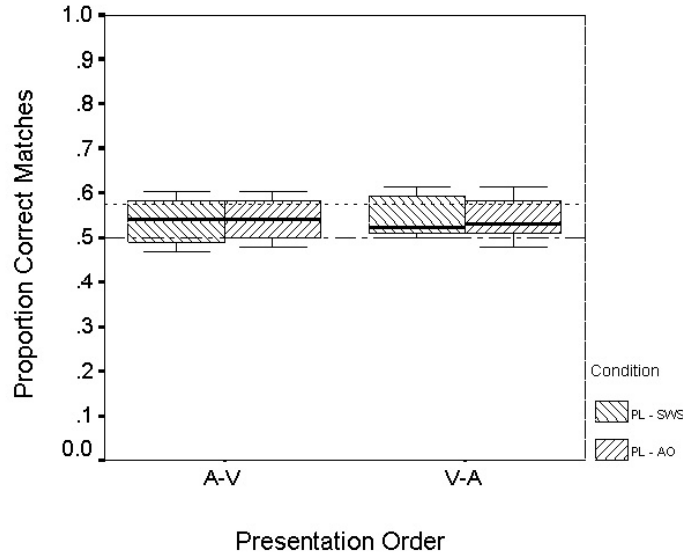


Figure 22. Boxplots of performance in Experiment 6. In the PLD-AO condition, point-light displays were matched with untransformed auditory speech tokens. In the PLD-SWS condition, point-light displays were matched with sinewave speech. Each boxplot represents the sample for a particular matching task and a particular order. The shaded boxes represent the interquartile range for the sample, the bold line indicates the median score of the sample, and the whiskers represent the range from the highest to the lowest score within the sample, excluding outliers. The dotted line represents the statistical threshold for chance performance using a binomial test with an α of 0.05.

Table 8. Descriptive statistics for crossmodal matching performance when visual stimuli were point-light displays (PLD) and auditory stimuli were either untransformed, natural utterances (AO) or sinewave speech replicas (SWS).

Matching Condition	Order	Mean	SE	t vs. 0.5
PLD-AO	A-V	0.538	0.017	2.17*
	V-A	0.547	0.015	3.19**
PLD-SWS	A-V	0.541	0.016	2.58*
	V-A	0.541	0.015	2.78*

* $p < 0.05$; ** $p < 0.01$

Discussion

Point-light displays and sinewave speech replicas of isolated spoken words can be matched correctly across sensory modalities in a crossmodal matching task. This finding suggests that isolated

kinematic displays of speech (visual and auditory) contain sufficient information for crossmodal specification. The results are consistent with the proposal that crossmodal matching performance is based on the ability to detect vocal tract movement information from unimodal displays. Furthermore, the results indicate that crossmodal comparisons are made using a common, modality-neutral metric based on the kinematics of articulatory activity. Because these extremely simplified stimulus displays were designed to eliminate all traditional cues to visual identity (e.g., configuration of facial features, shading, shape, etc.) and to auditory vocal identity (e.g., f_0 , average long-term spectrum, etc.), but preserve only the isolated movement of the vocal articulators, the results suggest that the common crossmodal source information is sensory information that is related to articulatory kinematics of vocal tract activity.

Experiment 7: Word Identification with Kinematic Primitives

In a recent investigation of the acoustic form of crossmodal source information, Lachs (2002) found that several acoustic transformations preserved the ability to match across sensory modalities while others eliminated it. Importantly, the same transformations that preserved crossmodal matching ability also preserved *word identification* performance. Lachs (2002) suggested that a common representational format might underlie both crossmodal source information, an indexical or extralinguistic property of spoken language, *and* phonetic information used to recognize spoken words. Indeed, several recent investigations have shown that the indexical properties of speech are inextricable from their linguistic or phonetic properties (see Goldinger, 1998; Lachs et al., in press; Pisoni, 1997).

Two studies using unimodal kinematic primitives have indicated that indexical information in speech signals may be carried in the very same “kinematic details” that support phonetic identification (Runeson, 1994). In the visual domain, Rosenblum et al. (2002) presented participants with point-light displays of a talker speaking a sentence and asked them to match the point-light display with one of two fully-illuminated faces speaking the same sentence. One of the fully-illuminated faces belonged to the same talker who had generated the point-light display. The results showed that visual point-light displays of a talker articulating speech could be matched accurately to fully-illuminated visual displays of the same utterances. This finding suggests that some talker-specific details are contained in the visually isolated kinematics of vocal tract activity. Analogously, in the auditory domain, Remez et al. (1997) presented listeners with sinewave replicas of English sentences and asked them to match the token with one of two untransformed auditory tokens. As above, one of the untransformed auditory tokens was spoken by the same talker who had spoken in the sinewave replica. The results showed that sinewave speech replicas carry enough idiosyncratic phonetic variation to support correct matching between untransformed, natural utterances and their sinewave speech replicas.

Taken together with the findings of Experiment 6, the available evidence supports the hypothesis that crossmodal source information is carried in parallel with phonetic information in the fine-grained kinematic details of a talker’s articulatory motions over time. These kinematic details are inherently modality-neutral because they refer to the common articulatory event, not the surface patterns of acoustic or optic energy impinging on the eyes or ears. As discussed above, a modality-neutral form for speech information has also been proposed to account for classic audiovisual speech phenomena (Fowler & Rosenblum, 1991; Rosenblum, 1994), like the McGurk effect and audiovisual enhancement to speech intelligibility. Is the isolated kinematic information available in sinewave speech and point-light displays also sufficient to support accurate word recognition?

Several parallel lines of investigation have detailed the integration of these minimal, unimodal stimulus displays with untransformed auditory stimuli or fully-illuminated visual stimuli. For example,

in one recent study, Remez, Fellowes, Pisoni, Goh and Rubin (1999) demonstrated that the intelligibility of sinewave replicas of sentences was significantly enhanced when presented in conjunction with a full visual display of the talker (cf. Breeuwer & Plomp, 1985). Similarly, in another study Rosenblum et al. (1996) showed that point-light displays of a talker speaking sentences could enhance recognition of untransformed auditory displays embedded in noise. Other than these two studies, no one has shown that sinewave speech replicas can be integrated with point-light displays to demonstrate the classic audiovisual enhancement effect – that spoken word recognition under multimodal, audiovisual conditions is better than spoken word recognition under auditory-alone conditions.

To extend these two sets of findings, Experiment 7 was designed to show that multimodal kinematic primitives are also integrated in a classic audiovisual integration task. Participants were asked to identify spoken words under two presentation conditions. In the auditory-alone condition, they were presented with sinewave replicas of isolated English words. In the audiovisual condition, they were presented with *multimodal* kinematic primitives: point-light displays of speech paired with sinewave replicas of the same utterance. Based on the earlier findings reported by Rosenblum et al. (1996) and Remez et al. (1999), we expected that word identification performance under combined, multimodal stimulation would show enhancement and would be better than performance under unimodal, auditory-alone stimulation. This finding would provide evidence for the proposal that auditory and visual displays of speech are “integrated” because they both carry information about the underlying kinematics of articulation.

Method

Experimental Design

Experiment 7 used a word recognition task to measure speech intelligibility for isolated words. Two within-subjects factors were manipulated. The first factor, “Presentation Mode,” consisted of two levels: sinewave speech alone (SWS) and point-light display plus sinewave speech (PLD + SWS). The levels of this factor were presented in blocks, which were counterbalanced for order of presentation across participants. The second factor, “Visual Intelligibility,” also consisted of two levels: low visual intelligibility and high visual intelligibility. The levels of this factor were presented in a random order within each Presentation Mode block. The number of high and low visual intelligibility words within a block was equal. In addition, an equal number of participants were presented with each word in each presentation mode.

Participants

Participants were 32 undergraduate students enrolled in an introductory psychology course who received partial credit for participation. All of the participants were native speakers of English and none of them 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 stimulus materials used during the familiarization phase were 20 isolated, monosyllabic English words that were not used in the final test list. These 20 words were spoken by a different talker who was not used to create the stimuli used in the point-light display database. The 20 familiarization words were converted into sinewave speech replicas using the same methods outlined in Experiment 6. The stimulus materials used during testing were identical to those used in Experiment

6. However, only the sinewave replicas of the spoken words were used as auditory tokens in this experiment.

Procedures

Each session began with a short familiarization phase identical to the one outlined earlier in Experiment 6 so that participants could become accustomed to the unusual nature of the sinewave speech replicas. It should be emphasized here that the familiarization task only asked for judgments of “goodness” and provided no feedback at all.

During the testing phase, each participant heard all 96 words spoken by one of the four point-light talkers. Every participant heard each talker speak on an equal number of trials, and no participant ever heard a word twice over the course of the experiment. The talker who spoke a given word was counterbalanced across participants. The 96 words recorded by the point-light talkers were also divided into two lists, each list contained 24 lexically easy words and 24 lexically hard words. For each participant, one test list was presented under audio-only (AO: SWS) conditions and one list was presented under audiovisual (AV: PLD + SWS) conditions. The assignment of a presentation context to each list and the order in which presentation contexts were viewed was counterbalanced across participants. After each stimulus item was presented, the participant simply typed in the word he/she heard on a standard keyboard and clicked the mouse button to advance to the next trial.

Results

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 conservative method of assessment, the two reviewers had a 100% agreement rate on classifying responses as typing and spelling errors. Responses on this task were scored 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”).

The data were submitted to a 2 x 2 repeated-measures ANOVA (Presentation Mode and Visual Intelligibility). The ANOVA revealed a main effect of Presentation Mode, $F(1, 30) = 57.37, p < 0.001$. Overall, performance in the audiovisual condition ($M = 0.298, SE = 0.025$) was better than performance in the audio-alone condition ($M = 0.174, SE = 0.022$), demonstrating that significant benefit in speech intelligibility was gained under multimodal presentation. The ANOVA also revealed a significant main effect of Visual Intelligibility, $F(1, 30) = 297.60, p < 0.001$. High visual intelligibility words ($M = 0.380, SE = 0.029$) were identified better than low visual intelligibility words ($M = 0.093, SE = 0.017$), *regardless* of presentation mode. This finding indicates that there may be a common, modality-neutral source of variability in speech that affects intelligibility scores (see also Lachs, Pisoni, & Kirk, 2001).

We also observed an interaction between Presentation Mode and Visual Intelligibility, $F(1, 30) = 23.05, p < 0.001$. Figure 23 illustrates this interaction. Performance on high visual intelligibility words was uniformly higher than performance on low visual intelligibility words, and this difference was larger under audiovisual (PLD + SWS) conditions than under auditory-alone (SWS) conditions.

In order to explore the visual intelligibility effect further, difference scores were calculated for each presentation mode between high and low visual intelligibility words. The average difference between performance on high visual intelligibility words and low visual intelligibility words for SWS

trials was 0.21 ($SE = 0.022$). In contrast, the difference for PLD + SWS trials was 0.36 ($SE = 0.023$). These scores were significantly different from each another, $t(30) = 5.13, p < 0.001$.

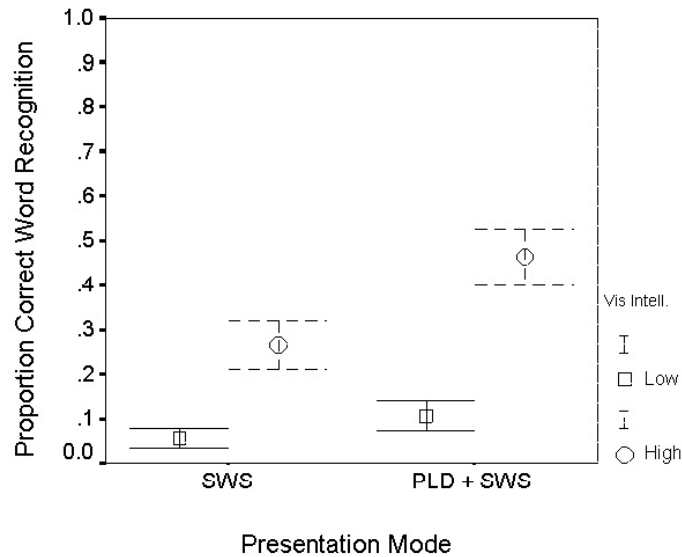


Figure 23. Average proportion correct in the auditory-alone (SWS) and audiovisual (PLD + SWS) conditions of Experiment 7. Error bars show standard errors.

Difference scores between the PLD + SWS and SWS conditions were also calculated separately for each level of visual intelligibility to assess the size of the gain due to audiovisual input. Difference scores were higher for high visual intelligibility words ($M = 0.20, SE = 0.028$) than for low visual intelligibility words ($M = 0.05, SE = 0.013$). Although close to the floor, the difference scores for low visual intelligibility words were different from zero, $t(30) = 3.78, p < 0.001$, as were the difference scores for the high visual intelligibility words, $t(30) = 7.01, p < 0.001$.

In summary, the results obtained in the word recognition task revealed that point-light displays of an articulating face can improve the intelligibility of sinewave speech. Furthermore, the results indicate that this effect was larger for words that were more visually intelligible than words that were less visually intelligible.

Audiovisual Gain (R scores). In order to examine individual differences in the extent to which combined audiovisual stimuli enhanced word intelligibility, the scores in the SWS and PLD + SWS conditions were combined into a single metric to obtain the measure R, the relative gain in speech perception due to the addition of visual information (Sumbly & Pollack, 1954). R was computed using the following formula:

$$R_a = \frac{AV - A}{100 - A}$$

where AV and A represent the speech intelligibility scores obtained in the audiovisual and auditory-alone conditions, respectively. From this formula, one can see that R measures the gain in accuracy in the AV condition relative to the accuracy in the A condition, normalized relative to the amount by which speech intelligibility could have possibly improved above the auditory-alone scores. The R

score can be used as an effective measure for comparing audiovisual gain across participants, because it effectively normalizes for the baseline, auditory-alone performance.

Figure 24 shows the R scores for all 32 participants in Experiment 7 ($M = 0.154$, $SE = 0.020$). A few participants ($N = 4$) showed no gain at all over auditory-alone performance. However, most of the participants showed some advantage in identifying words in the audiovisual condition relative to the auditory-alone condition. One participant showed a gain of 41% due to multimodal presentation.

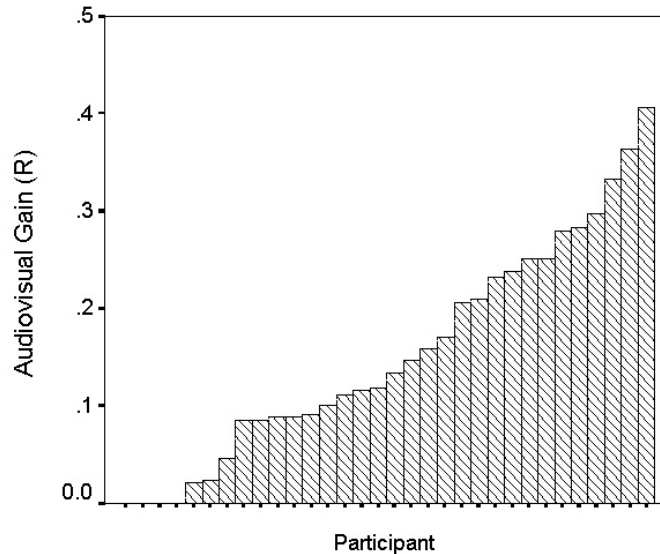


Figure 24. R scores for each participant in Experiment 7, sorted by performance. R scores represent the gain over auditory-alone performance due to additional visual information. Auditory stimuli were sinewave speech replicas and visual stimuli were point-light displays. The first four participants in the graph did not show any audiovisual gain.

R scores were also computed separately for the low visual intelligibility words and high visual intelligibility words. A paired t-test comparing high and low visual intelligibility R scores showed a significant difference in gain, $t(31) = 5.51$, $p < 0.001$. The R scores for high visual intelligibility words ($M = 0.267$, $SE = 0.028$) were much higher than the R scores for low visual intelligibility words ($M = 0.059$, $SE = 0.015$). The finding that R scores were higher for high visual intelligibility words than for low visual intelligibility words demonstrates clearly that the visual intelligibility manipulation facilitated the *integration* of sinewave speech and point-light displays, above and beyond the extent to which the manipulation increased the intelligibility of sinewave speech alone. However, both sets of R scores were significantly different from zero (High: $t(31) = 7.10$, $p < 0.001$; Low: $t(30) = 3.92$, $p < 0.001$), indicating that the point-light displays facilitated the intelligibility of all words, regardless of their baseline visual intelligibility.

Discussion

The present set of results using point-light displays and sinewave replicas of speech demonstrate that sufficient information for the combined perception of visual and auditory speech exists in highly impoverished displays that isolate the kinematics of articulatory activity that underlie

the production of speech. Even with limited, minimal information in the two sensory modalities, participants were able to perceive point-light displays and sinewave speech as integrated patterns, and they were able to exploit the complementary sources of sensory information to aid them in recognizing isolated spoken words. On average, additional point-light information increased intelligibility by 12%, almost doubling the intelligibility of the sinewave speech tokens observed in isolation. Importantly, the size of the observed gain was related to the visual intelligibility of the words. R scores for low visual intelligibility words were smaller than R scores for high visual intelligibility words. This finding is consistent with previous research that has shown that the extent to which visual information facilitates auditory-alone word and sentence identification is dependent on the visual-alone discriminability of the specific consonants and vowels in the pattern (Grant & Seitz, 1998; Grant, Walden, & Seitz, 1998).

A note on the issue of super-additivity in perceptual integration. Based on several findings, Grant and Seitz (1998) have argued that audiovisual integration is not necessarily super-additive as other researchers have suggested (cf. Massaro & Cohen, 2000), but is rather a simple linear combination of optically and acoustically discriminable phonetic information, perceived in the context of higher-order contextual information such as lexical, syntactic, and semantic constraints. Unfortunately, the present effects of visual intelligibility cannot be used to address this issue, because the visual intelligibility manipulation was based on lipreading scores obtained for an entirely different set of talkers under fully illuminated conditions. Thus, the precise visual phonetic information available in these particular point-light displays was not actually measured and cannot be used to draw conclusions about super-additivity. Future work regarding this issue will require more detailed analyses of the discriminability of point-light CV syllables and the extent to which these confusion data can predict word identification performance (as in Grant & Seitz, 1998).

Individual differences in perceptual integration. In a recent study, Grant et al. (1998) compared observed audiovisual gain scores to the gains predicted by several models of sensory integration. These models predicted optimal AV performance based on observed recognition scores for consonant identification in auditory-alone and visual-alone conditions. Grant et al. (1998) found that the models either over- or under-predicted the observed audiovisual data, indicating that not all individuals integrate auditory and visual information optimally. The present results also revealed individual differences in the extent to which sources of auditory and visual information could be combined and integrated together to support spoken word recognition. Audiovisual gain scores in the present experiment ranged from 0% to 41%.

Without question, the informative value of optic and acoustic arrays for specifying linguistic information must be interpreted in terms of a perceiver's sensitivity to relevant phonetic contrasts in the language. It is well-known that speechreading performance in hearing-impaired and normal-hearing listeners is subject to wide variation (Bernstein et al., 2000; Demorest & Bernstein, 1992). Although studies on the role of training in lipreading performance have reported conflicting conclusions (compare Bernstein, Auer, & Tucker, 2001; Walden et al., 1977), the results nevertheless indicate that experience can enhance the discriminability of linguistic contrasts (Dodd, Plant, & Gregory, 1989; Gesi, Massaro, & Cohen, 1992; Massaro, Cohen, & Gesi, 1993). It is not unreasonable to assume that the speechreading of point-light displays will be subject to even more such variation, and that such variation might have extensive effects on the ability to integrate auditory and visual displays of speech. Investigations into the role of individual differences in multisensory perception may provide important insights into the perceptual integration of audiovisual speech information and its development (Lewkowicz, 2001).

Despite the individual variation, however, the results of the present study are consistent with the hypothesis that auditory and visual speech information are combined and integrated with reference

to the underlying kinematic activity of the vocal tract. With dynamic movement information available from either sensory modality, perceivers evaluated the two kinds of information together, and exhibited the standard audiovisual gain finding that has been replicated many times in the literature (Erber, 1975; Grant & Seitz, 1998; Middleweerd & Plomp, 1987; Sumbly & Pollack, 1954). The present results extend and complement the recent findings of Rosenblum et al. (1996) and Remez et al. (1999) who found that isolated kinematic displays could facilitate word identification in conjunction with untransformed displays in the other sensory modality. The present results also provide additional support for the proposal that the evaluation of optical and acoustic information conceptualized along a common, kinematic and articulatory metric underlies classic audiovisual integration phenomena such as audiovisual enhancement (Fowler & Rosenblum, 1991; Summerfield, 1987).

General Discussion

In the present investigation, point-light displays of four talkers speaking isolated English words were recorded, and the accompanying acoustic displays were converted into sinewave speech replicas. In the first experiment, participants were asked to match point-light displays and sinewave speech replicas of the same talker using a crossmodal matching task. As with fully-illuminated visual displays and untransformed auditory signals, participants were able to correctly match the identity of the talker across different sensory modalities. These findings provide support for the proposal that modality-neutral source information is specified in the kinematic primitives of articulatory activity that underlie vocal gestures in speech production. In the second experiment, we showed that these minimal, “skeletonized” versions of speech could also be integrated across sensory modalities in an open set word recognition task and display classic audiovisual enhancement effects in speech intelligibility. Taken together, the results from both experiments using multimodal kinematic primitives show that crossmodal source information is specified in the isolated kinematics of a talker’s vocal tract activity, which is intimately tied to the phonetic realization of an utterance. A common form of phonetic representation, focusing on the articulatory origin of spoken language, may therefore be needed to account for the robust findings in the audiovisual speech perception literature (Schwartz et al., 1998; Summerfield, 1987), as well as investigations into the indexical properties of spoken language (Lachs, in preparation; Remez et al., 1997; Rosenblum et al., 2002).

It is important to note here that the modality-neutral form of crossmodal source information is restricted to sensory patterns that are lawfully structured by articulatory events, and not to static visual features that are commonly implicated in models of face recognition (e.g., configuration of facial features, color, shading, and shape, see Bruce, 1988). This observation raises an interesting dichotomy: the difference between visual kinematic information about *facial* identity vs. visual kinematic information about *vocal* identity. The point-light displays used in the present experiment contained sufficient detail about the idiosyncratic speaking style (or “voice”) of the talker to specify the auditory form of the utterance. Thus, the kinematic motions that were useful to participants in the present experiment could be called “visual voice” or “visual indexical” information. In contrast, visual motion information can be useful for identifying the person portrayed in the visual display, *irrespective* of that person’s unique vocal qualities.

Bruce and Valentine (1988), for example, showed that participants could identify familiar faces when shown point-light displays of familiar persons making different facial expressions and moving their heads. This kind of dynamic information reflects the identity of the person but is not related to the idiosyncratic speaking style of the talker, especially because the faces being identified were not speaking. Moreover, visual identity information can also be obtained from point-light displays of talkers *speaking*, as discussed above (Rosenblum et al., 2002). In another related study,

Berry (1991) showed that both children and adults could identify the gender of a speaking point-light display.

Evidence for the dichotomy between visual face and visual indexical information has also been found in neuropsychological investigations of prosopagnosia and aphasia. For example, Campbell, Landis and Regard (1986) reported on two patients: Mrs. "D." had the ability to speechread and was susceptible to the McGurk effect, despite having a profound prosopagnosia that rendered her incapable of discriminating the identity of various faces. However, another patient, Mrs. "T." showed no ability to lipread, yet could recognize faces quite well.

It should be emphasized here that the present results are only relevant to the ability of optic displays to convey voice information; modality-neutral information for voice appears to be conveyed by kinematic movements of the articulators. It remains to be seen whether such information is also useful for discriminating faces in the absence of speech. For example, it is not clear from these results whether training with point-light displays of a person engaged in the act of speaking would facilitate the subsequent recognition of that person's face engaged in *other* types of facial motion (e.g., laughing, smiling, frowning, nodding, etc.). What is clear, however, is that the idiosyncratic speaking style of a talker is conveyed in a modality-neutral form and is comparable across the visual and auditory sensory modalities as long as the underlying kinematics is preserved.

Kinematic Specification of Articulatory Dynamics. In his detailed summary of audiovisual speech research, Summerfield (1987) proposed two alternative modality-neutral metrics for the evaluation of audiovisual speech information: a kinematic metric (based on position and its time derivatives) and a dynamic metric (based on masses and forces), either of which could specify vocal tract activity. Considering modern theories of action and motor control, which propose that action is effected and coordinated by changes in the dynamic properties of muscle systems (Bernstein, 1967; Fink, Kelso, Jirsa, & de Guzman, 2000; Kelso, 1995), a dynamic metric might be a more theoretically appealing alternative at this point in time.

Under one conceptualization of the visual perception of these *causal*, dynamic properties of events, the distinction between kinematic and dynamic properties may be somewhat misleading. Indeed, a great deal of work in the visual perception of motion has shown that the kinematics of events contain information about dynamic properties (Bingham, 1995), and these are also useful in the perception of those dynamic properties. For example, slight deviations in the kinematic patterns of two colliding balls alter the perception of their relative mass (Runeson & Vedeler, 1993). Generalizing to animate events, Runeson and Frykholm (1981) showed that participants could perceive the mass of a weight being lifted by an arm with only point-light information about the arm itself. Furthermore, Bingham (1987) conducted a series of studies demonstrating that perceivers' judgments of weight were scaled based on kinematic information intrinsic to the display, and not to extrinsic "standards" provided by previous experience. Thus, optically-available kinematics provide direct, unmediated access to the dynamic, causal properties of events and thereby provide an informational "basis" for perception. These observations have been formulated in the kinematic specification of dynamics (KSD) principle (Runeson, 1994). Under this conceptualization, perception of animate motion is contingent on the pickup of information about the dynamic properties of the events that produced that motion, *via* its specification in optically available kinematic parameters. As Runeson puts it more succinctly, perceivers "can see the weight of an object handled by a person" (Runeson, 1994, p. 386 - 387).

With this conceptualization in mind, it may be possible to reformulate the perceptually relevant information available in optical and acoustic displays of speech such that a common, underlying dynamic description of speech information is obtained. Detailed analysis of the positions

and velocities of points on a face during simple vocal events, correlated with an analysis of the kinematic parameters specified in sinewave speech replicas, may be able to reveal this kind of dynamic structure. A model for such analysis was provided by Bingham (1995) who demonstrated that substantial dynamic parameters are available in the seemingly disjointed kinematic trajectories of point-lights placed on a ball rolling back and forth in a bowl.

A quantitative description of the underlying dynamic parameters relevant to speech perception may also provide a firm theoretical basis for further inquiry into the nature of speech perception in general. As demonstrated in the present investigation, kinematic parameters provide enough detailed information to support spoken word recognition and talker identification. A dynamic description of speech information may then dovetail with attempts to recast computational models of the mental lexicon in terms that incorporate both linguistic and indexical information in memory for spoken words (Goldinger, 1998). Indeed, a modality-neutral basis for lexical knowledge has been implicated in recent investigations of speechreading that show an underlying similarity in the factors relevant to visual-alone and auditory-alone spoken word recognition. For example, Auer (in press) found that phonological neighborhood density (the number of words in the lexicon that are phonologically similar to a target word) influences speechreading for isolated words. Lachs and Pisoni (submitted) reported similar results, showing that word frequency and phonological neighborhood frequency (i.e., the average frequency of phonological neighbors) also affect speechreading performance.

Such a reconceptualization of the mental lexicon might also help to integrate direct realist theories of speech perception (Fowler, 1986) with research on the traditionally higher-order, “cognitive” mechanisms involved in spoken communication (see Goldinger, Pisoni, & Logan, 1991; Lively, Pisoni, & Goldinger, 1994; Luce & Pisoni, 1998). At first glance, these abstract symbol processing approaches seem inherently incompatible with ecologically-oriented approaches to psychology. However, some recourse to lexical knowledge must be incorporated into any theory of speech perception, direct-realist or otherwise. By grounding perceptual information about speech in a modality-neutral, ecologically plausible framework, it may be possible to reformulate traditional cognitive architectures in a format more amenable to ecological principles (see also Brancazio, 1998).

The results presented here demonstrate for the first time that isolated kinematic information in the optical and acoustic displays are able to specify crossmodal source information that can be successfully integrated during the process of speech perception. Future work on audiovisual speech perception will need to formulate a more rigorous theoretical description of the nature of audiovisual speech integration effects by recasting speech information in terms of its dynamic, articulatory, and ultimately modality-neutral form.

Chapter V: Summary and Conclusions

Visual attributes of speech (lipreading) and auditory attributes of speech influence each other in the process of speech perception. Traditional cue-based accounts of audiovisual speech perception propose that cognitive mechanisms integrate the independent visual and auditory information during perception (Braidá, 1991; Massaro & Cohen, 1995). However, the multimodal correlates of speech are not independent of each other, but are lawfully related by virtue of their common origin in the production of speech (Vatikiotis-Bateson et al., 1997). Alternative accounts of perceptual integration build on this fact, proposing that acoustic and optical speech are integrated because the relevant phonetic information is not constrained to transmission via optic or acoustic energy, but is instead modality-neutral. This direct realist theory of speech perception (Fowler, 1986; Gibson, 1966; Gibson, 1979) predicts that perceivers should be able to match visual and auditory speech patterns presented to different sensory modalities because they both carry information about the same underlying object of perception – the linguistically significant gestures of an articulating vocal tract. The present investigation was designed to further clarify and elaborate the direct realist proposal that the important information necessary for speech perception is modality-neutral and is inherently based upon the articulatory behavior of a moving vocal tract. To assess this proposal, a crossmodal matching task was used to examine the acoustic and optical correlates of speech. Participants were asked to match the visual and auditory specifications of a talker speaking an isolated word. Performance above chance was taken to imply the presence of common crossmodal source information in the unimodal signals.

The results of the preliminary experiments in Chapter II indicated that observers could successfully match speaking faces and voices, indicating that information about the speaker was available for crossmodal comparisons. Crossmodal source information was not available when static visual displays of faces were used as visual stimuli and was not contingent upon a prominent acoustic cue to vocal identity (f_0). Furthermore, crossmodal matching was not possible when the acoustic signal was temporally reversed.

The series of perturbation experiments presented in Chapter III assessed six different acoustic transformations to see whether they preserved or eliminated crossmodal source information. In addition, word recognition performance was tested under the same acoustic transformations. The results showed that crossmodal source information was preserved under the same conditions that preserved information needed for word recognition. Furthermore, we found that linear transformations of the frequency spectrum (frequency shifting and linear frequency scaling) preserved both crossmodal matching ability and word identification performance. In contrast, nonlinear transformations of the spectrum (nonlinear frequency scaling and frequency rotation) eliminated or severely disrupted performance on both tasks. The results indicated that the acoustic form of crossmodal speech information (and by extension, articulatory information) is specified in the relative spectral structure of formant patterning as it evolves over time.

Finally, in Chapter IV, dynamic point-light visual displays of speech were used in conjunction with sinewave replicas of speech to assess listeners' ability to match faces and voices under conditions when only isolated kinematic information about vocal tract articulation was available in either sensory modality. The results of all three point-light experiments were consistent with the hypothesis that the form of speech information is not contingent upon transmission via acoustic or optic energy, but is based upon the underlying dynamics of vocal tract activity as it changes over time.

Modality-neutral Representations vs. Evidence from Neuropsychology

At first glance, the proposal of a modality-neutral form for speech would seem incompatible with the simple biological and anatomical fact that sensory stimulation at the retina is anatomically removed from sensory stimulation at the cochlea, and that the activation patterns that propagate from these initial “portals for the entry of sensory information” (Mesulam, 1998, p. 1015) may not interface until their anatomical confluence at more central cortical structures (Stein & Meredith, 1993) like the superior temporal sulcus (Calvert, 2001).

In a recent publication, Bernstein, Auer and Moore (in press) reviewed a large body of data concerning the neural structures implicated in both optical and acoustic phonetic perception and concluded that, rather than supporting a modality-neutral or common format (CF) view of speech information, the available evidence from neurobiological studies supports a modality-specific (M-S) theory of speech perception. Their argument revolves around three main points. First, Bernstein et al. cite behavioral evidence that supports the proposal that modality-specific neural pathways for speech perception exist in the brain. In one study, Bernstein, Demorest, and Tucker (2000) showed that, despite long-standing assumptions to the contrary, visual-only speech perception can be extremely accurate, with accuracy for words in sentences sometimes reaching 57%. In addition, lexical effects on the visual perception of speech are better predicted using visual-alone phonetic confusion patterns than auditory-alone phonetic confusion patterns (Auer, in press). Bernstein et al. claim that this evidence rules out a CF theory because the results show, *prima facie*, that neural pathways exist to process optical information relevant to speech perception. The results also demonstrate that the influence of visual speech information can extend up the cognitive architecture into suprasegmental levels of linguistic representation, a fact that Bernstein et al. claim is in direct contradiction to the predictions of a CF theory.

As a second source of evidence supporting an M-S theory of speech perception, Bernstein et al. (in press) review neurophysiological evidence supporting the proposal that specific cortical areas exist that are specialized for modality-specific processing of speech. More specifically, they claim that the propagation of neural activity along the optic nerve and into more central visual processing areas is quite literally removed and separate from the propagation of neural activity entering via the auditory nerve. Thus, there is no anatomical opportunity before linguistic processing for neural activity based on visual stimulation to interface with activity based on auditory stimulation.

For their third source of evidence, Bernstein et al. (in press) finally review literature that shows that the association of unimodal processing via co-terminal neural projections can be accomplished “without also being transformed into a common format.”

That modality-specific neural pathways exist to process optical speech information should seem immediately obvious; it would be absurd to assert that neural mechanisms do *not* exist for the processing of visual-alone speech information, especially given the evidence cited by Bernstein et al. (in press) about the lipreading capabilities of normal-hearing as well as hearing-impaired observers. However, as discussed earlier in Chapter IV, CF theories of speech perception are not necessarily incompatible with “higher-order” or “cognitive” effects of visual speech, as Bernstein et al. claim. Language is an inherently social entity; it does not rely solely on the direct perception of articulatory gestures. It also relies upon a common agreement among a particular group of speakers that certain gestures will be linguistically significant (i.e., will convey meaning), while others will not (Gibson, 1966). It is not unreasonable to assume that a person’s experience with the language will have appreciable effects on the form of perceptual information. These observations are all consistent with an ecological theory of speech perception. From the ecological perspective, energy patterns can

directly specify the events that structured them, but they are only *informative* insofar as those energy patterns have biological relevance to the animal – in this case, a human speaker of the language.

Bernstein et al.'s (in press) point is well taken, however. Prior demonstrations of a common representational format for speech have remained ecologically-oriented front ends to more traditionally symbolic linguistic architectures. That is, prior applications have explained how direct information about vocal tract activity can specify sub-segmental linguistic features or even phonemes (e.g. Fowler, 1994; Fowler, Brown, & Mann, 2000), while very rarely delving into the realm of suprasegmental linguistic phenomena (however, see Brancazio, 1998; Dekle et al., 1992; Lachs & Pisoni, submitted). Indeed, as argued above, it is imperative to extend modality-neutral formalizations of speech information into traditionally “higher order” or “cognitive” domains if the ecological approach to speech perception is to remain a viable theory of speech perception. With the eventual formalization of modality-neutral speech information as dynamic articulatory events, it should be possible to recast lexical processes in more ecologically-oriented terms, in much the same way that articulatory phonology (Browman & Goldstein, 1995) has recast the traditional symbolic processing architecture of phonological rules in terms of the dynamic parameters that govern the production of speech.

Another objection to Bernstein et al.'s (in press) claims concerns their assertion that CF theories require the *translation* of acoustic and optical information into a common representational format. This may be true of motor theory, which proposes a specialized speech processing module that specifically converts incoming sensory information into a metric relevant to the perceiver's own knowledge of the vocal tract (Lieberman & Mattingly, 1985). However, translation is decidedly *not* required by the direct realist theory of speech perception. As noted earlier, according to the direct realist approach, acoustic and optic information specify the dynamics of the vocal tract *directly*; no mediation, analysis, or translation is required because the laws of physics dictate the relationship between articulatory events and the acoustic, optic, or haptic patterns structured by those events.

Finally, Bernstein et al.'s (in press) approach implicitly assumes that proponents of a CF theory of speech information would search for neural structures or mechanisms that represent speech information in a common format by looking for patterns of neural activation, localized to a specific brain area. This is also not necessarily a valid assumption, although some CF theorists have taken this approach (Calvert et al., 1997). Alternative approaches to examining the neural basis of common format information are possible. Ironically, Bernstein et al. present just such an approach as an example of how optical and acoustic information can be related to one another “without re-representation of signal (sic) into a new amodal representation” (p. 20). They present the results of a multilinear regression analysis in which the 3-dimensional positions over time of points glued to a talker's face are compared with an rms energy and line spectral pair representation of the acoustic signal produced concurrently with the visual data. The regression analysis then optimizes the correspondence between the two sets of data and represents a solution vector that relates the two sources of unimodal data to each other. They report that the correlation between optical and acoustic patterns of four talkers speaking 23 CV (consonant-vowel) syllables, after weighting by the optimal solution vector, was between .74 and .82, a very good correspondence.

The “visual” information used in their analysis was the isolated kinematic properties of the vocal tract as it moved in time. In a sense, the regression analysis asks the question: how are these unimodal and complex signals related? If there is a lawful relationship between the acoustic and optical signals by virtue of their common instantiation in a vocal event, the solution discovered by the regression analysis cannot be anything other than a representation of the same dynamic activity of the vocal tract: it *is* the common link between the kinematics of the vocal tract and the particular acoustic representation chosen as an input to the analysis. It is possible that a sinewave speech representation of

the acoustic signal might yield even stronger correlations. Ironically, Bernstein et al. (in press) present a viable method for determining the common form of speech information while simultaneously asserting that no such thing exists! Translation of unimodal signals into a common format is not a necessary component of CF theories. What is important is to determine the ways in which a unimodal signal can specify the underlying event properties that are informative.

From the ecological perspective, information is an “emergent property” of the continuous, dynamic interaction of the perceiver and the environment in which the perceiver is situated (Gibson, 1966; Gibson, 1979). The evidence reported by Bernstein et al. (in press) may rule out literal interpretations of CF theories that search for direct neural instantiations of amodal properties, but the findings are still compatible with modality-neutral theories of speech information like direct realist theory.

Future Directions

Although the evidence presented in this dissertation is consistent with the hypothesis that the form of speech information is modality-neutral and inherently articulatory in nature, there is one problematic aspect to the data: performance in all conditions, regardless of whether they differed from chance or not, was extremely poor overall. Even when the mean performance scores for the sample differed as a whole from chance, the performance of individual participants ranged between 50% and 65% accuracy, at best.

Future work should focus on ways of increasing performance in the crossmodal matching task. One possibility would be to change some of the task dynamics from a classic XAB task. For instance, on every trial, participants only heard or saw each token once. It may be possible to increase performance by allowing participants to view each target-alternative pair multiple times (but not presented concurrently). For example, a participant might view the X-A combination, and then the X-B combination, and would be able to repeat these two combinations multiple times until a confident decision could be made. This might reduce any memory effects implicit in the task and might allow participants to better distinguish the small details of articulation that distinguish one talker from another.

Another way to improve performance would be to change the stimulus itself from isolated words to sentences. It is likely that crossmodal source information is much more salient in sentences than in isolated words. Studies in the learning of talker voices have shown that talker-specific learning is more robust if participants are trained with sentences rather than words (Nygaard & Pisoni, 1998). It is possible that salient talker-specific information extends over longer time scales than comparatively salient phonetic information. Extending the duration of exposure to articulatory activity so that it encompasses a variety of actions, gestures, and articulations of the vocal tract may increase the distinction between talkers (cf., Cutting & Kozlowski, 1977; Kozlowski & Cutting, 1977, 1978).

Finally, it might be possible to increase performance if participants had some prior familiarity with the idiosyncratic articulatory habits of the talkers in the study. Future work could pre-train participants to discriminate the auditory-alone displays of the talkers to a certain criterion and then test their ability to match the voices to faces. Again, this might serve to enhance perceptual sensitivity to the small intertalker variations necessary for discriminating across sensory modalities.

A final extension of the current findings could resolve the distinction between crossmodal source information and modality-neutral phonetic information in a single task. In the crossmodal matching task used in these studies, the participant attempted to match the *speaker* of an event, given

two alternative speakers, across sensory modalities, while the linguistic content of the utterance (the word) was held constant. However, it should be theoretically possible for participants to match the *specific articulatory events* that produced a particular speech token across modalities as well. That is, both the word *and* the talker would be held constant across the target pattern and the two response alternatives, but participants would be asked to match the crossmodal specification of the particular instantiation of the spoken utterance itself. Future work should examine this question by filming the same talker uttering the same word under two different speaking conditions: clear and conversational speech. Some evidence has been reported showing that there is sufficient intratalker variability between these two speaking modes (Gagné, Masterson, Munhall, Bilida, & Querengesser, 1994) to support crossmodal matching, even when the talker, and the linguistic content, of the message is identical. Finding that participants are sensitive to the modality-neutral form of the instance-based properties of particular utterances would provide further support for the proposal that fine-grained details of dynamic vocal tract articulatory activity underlie the perception of speech.

Conclusion

The present findings contribute to the now rapidly growing literature on audiovisual speech perception, a field of study that promises to yield new insights into the fundamental process of speech perception and spoken word recognition. In addition, the results point in new directions for scientific inquiry in more applied areas of speech research. For instance, although speech recognition software is becoming cheaper and easier to use while growing increasingly more powerful and robust, levels of performance equivalent to those exhibited by humans remain elusive, especially under noisy, degraded, or novel conditions. Perhaps the present findings, with their implications for the form of speech information, will help to guide future design in artificial speech recognition technology by providing a “job description” for the perceptual system (whether it be human or digital). In particular, Experiment 5 outlines some interesting acoustic transformations under which human speech recognition remains unperturbed; digital acoustic pattern matching strategies for speech recognition may be able to generalize over these kinds of transformations and yield more robust performance under novel or degraded conditions.

This research may also be applicable to the relatively new field of research on cochlear implantation. The biotechnology needed for implantation and direct electrical stimulation of the cochlear nerve is continually improving. However, the range of variation in clinical outcomes is very large. The results presented here could be used to guide further research into the optimal pre-processing strategies useful for communicating speech information to the neural/electrical interface, yielding better clinical outcomes.

In conclusion, the present findings support the proposal that there is a modality-neutral form for speech information and that this form is based on the underlying articulatory gestures used in spoken language. It is hoped that with continued investigation of the ways in which acoustic and optic signals can be informative about vocal tract articulation, as well as continued investigation into the ways in which the dynamic activity of the vocal tract structures acoustic and optic media, a better description of the important information for speech perception will be obtained.

References

- Abercrombie, D. (1967). *Elements of general phonetics*. Chicago, IL: Aldine Publishing Company.
- Auer, E.T., Jr. (in press). The influence of the lexicon on speechread word recognition: Contrasting segmental and lexical distinctiveness. *Psychonomic Bulletin & Review*.
- Berger, K.W. (1972). *Speechreading: Principles and methods*. Baltimore, MD: National Educational Press, Inc.
- Bernstein, L.E., Auer, E.T., Jr., & Moore, J.K. (in press). Modality-specific perception of auditory and visual speech. In G.A. Calvert & C. Spence & B.E. Stein (Eds.), *Handbook of Multisensory Processes*. Cambridge: MIT Press.
- Bernstein, L.E., Auer, E.T., Jr., & Tucker, P.E. (2001). Enhanced speechreading in deaf adults: Can short-term training/practice close the gap for hearing adults? *Journal of Speech, Language, and Hearing Research, 44*, 5-18.
- Bernstein, L.E., Demorest, M.E., Coulter, D.C., & O'Connell, M.P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. *Journal of the Acoustical Society of America, 90*, 2971-2984.
- Bernstein, L.E., Demorest, M.E., & Tucker, P.E. (2000). Speech perception without hearing. *Perception & Psychophysics, 62*, 233-252.
- Bernstein, N.A. (1967). *The coordination and regulation of movements*. London: Pergamon Press.
- Berry, D.S. (1991). Child and adult sensitivity to gender information in patterns of facial motion. *Ecological Psychology, 3*, 349-366.
- Bertelson, P., Vroomen, J., Wiegand, G., & De Gelder, B. (1994). Exploring the relationship between McGurk interference and ventriloquism. *Proceedings of 1994 International Conference on Spoken Language Processing, 13*, 559-562.
- Bingham, G.P. (1987). Kinematic form and scaling: Further investigations on the visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception & Performance, 13*, 155-177.
- Bingham, G. P. (1995). Dynamics and the problem of visual event recognition. In R. F. Port & T. Van Gelder (Eds.), *Mind as Motion: Explorations in the dynamics of cognition* (pp. 403-448). Cambridge, MA: The MIT Press.
- Bingham, G.P., Schmidt, R.C., & Rosenblum, L.D. (1995). Dynamics and the orientation of kinematic forms in visual event recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1473-1493.
- Blessner, B. (1972). Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *Journal of Speech and Hearing Research, 15*, 1-41.
- Blessner, B.A. (1969). *Perception of spectrally transformed speech*. Unpublished Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- Bradlow, A.R., Torretta, G.M., & Pisoni, D.B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20*, 255-273.
- Braida, L.D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology, 43A*, 647-677.
- Brancazio, L. (1998). *Contributions of the lexicon to audiovisual speech perception*. Unpublished Ph. D. Dissertation, University of Connecticut.
- Brancazio, L., Miller, J.L., & Paré, M.A. (1999, November). *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, Ohio.
- Breeuwer, M., & Plomp, R. (1985). Speechreading supplemented with formant-frequency information for voiced speech. *Journal of the Acoustical Society of America, 77*, 314-317.

- Bricker, P.D., & Pruzansky, S. (1976). Speaker recognition. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 295-326). New York: Academic Press.
- Browman, C.P., & Goldstein, L. (1995). Dynamics and articulatory phonology. In R.F. Port & T. Van Gelder (Eds.), *Mind as Motion*. Cambridge, MA: MIT Press.
- Bruce, V. (1988). *Recognising faces*. Hove, U.K.: Erlbaum.
- Bruce, V., & Valentine, T. (1988). When a nod's as good as a wink: The role of dynamic information in facial recognition. In M.M. Gruneberg, P.E. Morris & R.N. Sykes (Eds.), *Practical aspects of memory: Current research and issues: Vol. 1. Memory in everyday life* (pp. 169-174). New York: Wiley.
- Calvert, G.A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, *11*, 1110-1123.
- Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., Woodruff, P.W.R., Iversen, S.D., & David, A.S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*, 593-596.
- Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology*, *32*, 85-99.
- Campbell, R., Landis, T., & Regard, M. (1986). Face recognition and lipreading: A neurological dissociation. *Brain*, *109*, 509-521.
- Cole, R.A., Coltheart, M., & Allard, F. (1974). Memory of a speaker's voice: Reaction time to same- or different-voiced letters. *Quarterly Journal of Experimental Psychology*, *26*, 1-7.
- Creelman, C.D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America*, *29*, 655.
- Cutting, J.E., & Kozlowski, L.T. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, *9*, 353-356.
- Dekle, D.J., Fowler, C.A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception & Psychophysics*, *51*, 355-362.
- Demorest, M.E., & Bernstein, L.E. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech and Hearing Research*, *35*, 876-891.
- Diehl, R.L., & Kluender, K.R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*, 121-144.
- Dodd, B., Plant, G., & Gregory, M. (1989). Teaching lip-reading: The efficacy of lessons on video. *British Journal of Audiology*, *3*, 229-238.
- Dodd, B.E., & Campbell, R. (1987). *Hearing by eye: The psychology of lip-reading*. London; Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Ellis, D. (2001). *Sinewave Speech Analysis/Synthesis in MATLAB*. Retrieved January 5, 2001, from the World Wide Web: <http://www.ee.columbia.edu/~dpwe/resources/matlab/sws>
- Erber, N.P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*, 423-424.
- Erber, N.P. (1972). Auditory, visual and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech, Language and Hearing Research*, *15*, 413-422.
- Erber, N.P. (1974). Visual perception of speech by deaf children: Recent developments and continuing needs. *Journal of Speech & Hearing Disorders*, *39*, 178-185.
- Erber, N.P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, *40*, 481-492.
- Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton and Co.
- Fellowses, J.M., Remez, R.E., & Rubin, P.E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics*, *59*, 839-849.
- Fink, P.W., Kelso, J.A.S., Jirsa, V.K., & de Guzman, G. (2000). Recruitment of degrees of freedom stabilizes coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 671-692.

- Fisher, B.D., & Pylyshyn, Z.W. (1994). The cognitive architecture of bimodal event perception: A commentary and addendum to Radeau (1994). *Current Psychology of Cognition*, 13, 92-96.
- Fourcin, A.J. (1968). Speech-source interference. *IEEE Transactions in Audio Electroacoustics*, ACC-16, 65-67.
- Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3-28.
- Fowler, C.A. (1994). Invariant, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception & Psychophysics*, 55, 597-610.
- Fowler, C.A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99, 1730-1741.
- Fowler, C.A., Brown, J.M., & Mann, V.A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 877-888.
- Fowler, C.A., & Dekle, D.J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, 17, 816-828.
- Fowler, C.A., & Rosenblum, L.D. (1991). The perception of phonetic gestures. In M. Studdert-Kennedy & I.G. Mattingly (Eds.), *Modularity and the motor theory of speech perception*. (pp. 33-59), Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Gagné, J.-P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal of the Academy of Rehabilitative Audiology*, 27, 135-158.
- Garner, W.R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.
- Gaver, W.W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, 5, 1-29.
- Gesi, A.T., Massaro, D., & Cohen, M.M. (1992). Discovery and expository methods in teaching visual consonant and word identification. *Journal of Speech and Hearing Research*, 35, 1180-1188.
- Gibson, J.J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.
- Gibson, J.J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton-Mifflin.
- Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105, 251-279.
- Goldinger, S.D., Pisoni, D.B., & Logan, J.S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 17, 152-162.
- Grant, K.W. (2001). The effect of speechreading for masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America*, 109, 2272-2275.
- Grant, K.W., & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America*, 104, 2438-2450.
- Grant, K.W., & Seitz, P.F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, 108, 1197-1208.
- Grant, K.W., Walden, B.E., & Seitz, P.F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America*, 103, 2677-2690.
- Green, K.P. (1994). The influence of an inverted face on the McGurk effect. Paper presented at the 127th Meeting of the Acoustical Society of America. MIT.
- Green, K.P. (1996). The use of auditory and visual information in phonetic perception. In D. G. Stork & M.E. Hennecke (Eds.), *Speechreading by humans and machines*. Berlin: Springer-Verlag.
- Green, K.P., & Gerdeman, A. (1995). 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*, 21, 1409 -1426.

- Green, K.P., & Kuhl, P.K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, *45*, 34-42.
- Green, K.P., & Kuhl, P.K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 278-288.
- Green, K.P., Kuhl, P.K., & Meltzoff, A.N. (1988). *Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment*. Paper presented at the 148th meeting of the Acoustical Society of America, Honolulu, Hawaii.
- Green, K.P., & Miller, J.L. (1985). On the role of visual rate information in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 269-276.
- Harnsberger, J.D., Svirsky, M.A., Kaiser, A.R., Pisoni, D.B., Wright, R., & Meyer, T.A. (2001). Perceptual "vowel spaces" of cochlear implant users: Implications for the study of auditory adaptation to spectral shift. *Journal of the Acoustical Society of America*, *109*, 2135-2145.
- Hirahara, T., & Kato, H. (1992). The effects of F0 on vowel identification. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 89-112). Tokyo: Ohmsha Publishing.
- Hollien, H., & Klepper, B. (1984). The speaker recognition problem. *Advances in Forensic Psychology & Psychiatry*, *1*, 87-111.
- Jeffers, J.B.M. (1971). *Lipreading (Speechreading)*. Springfield, IL: Charles C. Thomas.
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*, 201-211.
- Jones, J.A., & Munhall, K.G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics*, *25*, 13-19.
- Jordan, T.R., & Bevan, K. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 388-403.
- Jordan, T.R., McCotter, M.V., & Thomas, S.M. (2000). Visual and audiovisual speech perception with color and gray-scale facial images. *Perception & Psychophysics*, *62*, 1394-1404.
- Kaiser, M. K., & Proffitt, D. R. (1984). The development of sensitivity to causally relevant dynamic information. *Child Development*, *55*, 1614-1624.
- Kanzaki, R., & Campbell, R. (1999, August). *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.
- Kawahara, H. (1997, April). *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 (ICASSP '97), Munich.
- Kawahara, H., Katayose, H., de Cheveigné, A., & Patterson, R.D. (1999). *Fixed points analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity*. Paper presented at the EUROSPEECH '99, Budapest, Hungary.
- Kawahara, H., Masuda-Kastuse, I. & Cheveigne, A. (1999). 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*, *27*, 187-207.
- Kelso, J.A. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: The MIT Press.
- Kimura, D., & Folb, S. (1968). Neural processing of backwards-speech sounds. *Science*, *161*, 395-396.
- Kozlowski, L.T., & Cutting, J.E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics*, *21*, 575-580.
- Kozlowski, L.T., & Cutting, J.E. (1978). Recognizing the gender of walkers from point-lights mounted on ankles: Some second thoughts. *Perception & Psychophysics*, *23*, 459.

- Kuhl, P.K., & Meltzoff, A.N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, 7, 361-381.
- Lachs, L. (1999). A voice is a face is a voice, *Research on Spoken Language Processing No. 23*. (pp. 241-258) Bloomington, IN: Speech Research Laboratory, Indiana University Bloomington.
- Lachs, L. (2002). *Vocal tract kinematics and crossmodal information*. Unpublished Dissertation, Indiana University, Bloomington, IN.
- Lachs, L. (in preparation). The acoustic form of crossmodal source information and its relationship to word identification.
- Lachs, L., & Hernández, L.R. (1998). Update: The Hoosier Audiovisual Multitalker Database, *Research on Spoken Language Processing Progress Report 22* (pp. 377-388). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Lachs, L., McMichael, K., & Pisoni, D.B. (in press). Speech perception and implicit memory: Evidence for detailed episodic encoding. In J. Bowers & C. Marsolek (Eds.), *Rethinking Implicit Memory*. Oxford University Press.
- Lachs, L., & Pisoni, D.B. (submitted). Spoken word recognition without audition. *Perception & Psychophysics*.
- Lachs, L., Pisoni, D.B., & Kirk, K.I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: A first report. *Ear & Hearing*, 22, 236-251.
- Ladefoged, P. (1996). *Elements of acoustic phonetics* (2nd ed.). Chicago: University of Chicago Press.
- Lewkowicz, D.J. (2001). Infants' perception of the audible, visible and bimodal attributes of multimodal syllables. *Child Development*, 71, 1241-1257.
- Li, X., Logan, R.J., & Pastore, R.E. (1991). Perception of acoustic source characteristics: Walking sounds. *Journal of the Acoustical Society of America*, 90, 3036-3049.
- Liberman, A., Delattre, P., & Cooper, F. (1952). The role of selected stimulus variables in the perception of the unvoiced-stop consonants. *American Journal of Psychology*, 65, 497-516.
- Liberman, A., & Mattingly, I. (1985). The motor theory revised. *Cognition*, 21, 1-36.
- Liberman, A.M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America*, 29, 117-123.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431-461.
- Lively, S.E., Pisoni, D.B., & Goldinger, S.D. (1994). Spoken word recognition: Research and theory. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics*, pp. 265-301. New York: Academic Press.
- Löfqvist, A., & Gracco, V.L. (1997). Lip and jaw kinematics in bilabial stop consonant production. *Journal of Speech, Language and Hearing Research*, 40, 877-893.
- Luce, P.A., & Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, 19, 1-36.
- Lucero, J.C., & Munhall, K.G. (1999). A model of facial biomechanics for speech production. *Journal of the Acoustical Society of America*, 106, 2834-2842.
- MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, 24, 253-257.
- MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21, 131-141.
- Massaro, D., Cohen, M.M., & Gesi, A.T. (1993). Long-term training, transfer, and retention in learning to lipread. *Perception & Psychophysics*, 53, 549-562.
- Massaro, D.W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Earlbaum Associates.
- Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
- Massaro, D.W., & Cohen, M.M. (1995). Perceiving talking faces. *Current Directions in Psychological Science*, 4, 104-109.

- Massaro, D.W., & Cohen, M.M. (1996). Perceiving speech from inverted faces. *Perception & Psychophysics*, *58*, 1047-1065.
- Massaro, D.W., & Cohen, M.M. (2000). Tests of auditory-visual integration efficiency within the framework of the FLMP: A reply to and extension of Grant and Seitz (1998). *Journal of the Acoustical Society of America*, *108*, 784-789.
- Massaro, D.W., & Oden, G. C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 1053-1064.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.
- Mesulam, M.-M. (1998). From sensation to perception. *Brain*, *121*, 1013-1052.
- Middleweerd, M.J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold in noise. *Journal of the Acoustical Society of America*, *82*, 2145-2147.
- Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*, 379-390.
- Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*, 365-378.
- Munhall, K., Gribble, P., Sacco, L., & Ward, M. (1995). Temporal constraints on the perception of the McGurk effect. *Perception & Psychophysics*, *58*, 113-131.
- Munhall, K.G., & Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 123-139). East Sussex, UK: Psychology Press, Ltd.
- Nygaard, L.C., & Pisoni, D.B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*, 355-376.
- Ohala, J.J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, *99*, 1718-1725.
- Pisoni, D.B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, *13*, 109-125.
- Pisoni, D.B. (1997). Some thoughts on "Normalization" in speech perception. In K. Johnson & J.W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9-32). San Diego: Academic Press.
- Posner, M.I., Boies, S.I., Eichelman, W.H., & Taylor, R.I. (1969). Retention of visual and name codes of single letters. *Journal of Experimental Psychology Monograph*, *79*(1, Pt. 2).
- Posner, M.I., & Keele, S.W. (1967). Decay of visual information from a single letter. *Science*, *158*, 137-139.
- Radeau, M., & Colin, C. (2001). Object identity is not a condition but a result of intersensory integration: The case of audiovisual interactions. *Current Psychology of Cognition*, *20*, 349-357.
- Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip reading* (pp. 97-114). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Remez, R.E. (1989). When the objects of perception are spoken. *Ecological Psychology*, *1*, 161-180.
- Remez, R.E., Fellowes, J.M., Pisoni, D.B., Goh, W.D., & Rubin, P.E. (1999). Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances. *Speech Communication*, *26*, 65-73.
- Remez, R.E., Fellowes, J.M., & Rubin, P.E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 651-666.
- Remez, R.E., Rubin, P.E., Berns, S. M., Pardo, J.S., & Lang, J.M. (1994). On the perceptual organization of speech. *Psychological Review*, *101*, 129-156.

- Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947-950.
- Repp, B.H. (1987). The sound of two hands clapping: An exploratory study. *Journal of the Acoustical Society of America*, *81*, 1100-1109.
- Rosenblum, L.D. (1994). How special is audiovisual speech integration? *Current Psychology of Cognition*, *13*, 110-116.
- Rosenblum, L.D., Johnson, J.A., & Saldaña, H.M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research*, *39*, 1159-1170.
- Rosenblum, L.D., & Saldaña, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 318-331.
- Rosenblum, L.D., Yakel, D.A., Baseer, N., Panchal, A., Nodarse, B.C., & Niehaus, R.P. (2002). Visual speech information for face recognition. *Perception & Psychophysics*, *64*, 220-229.
- Runeson, S. (1994). Perception of biological motion: The KSD-principle and the implications of the distal versus optimal approach. In G. Jansson, S.S. Bergström & W. Epstein (Eds.), *Perceiving events and objects* (pp. 383-405). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Runeson, S., & Frykholm, G. (1981). Visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 733-740.
- Runeson, S., & Frykholm, G. (1983). Kinematic specification of dynamics as an informational basis for person-and-action perception: Expectation, gender recognition, and deceptive intention. *Journal of Experimental Psychology: General*, *112*, 585-615.
- Runeson, S., & Vedeler, D. (1993). The indispensability of precollision kinematics in the visual perception of relative mass. *Perception & Psychophysics*, *53*, 617-632.
- Schultz, M., Norton, S., Conway-Fithian, S., & Reed, C. (1984). A survey of the use of the Tadoma method in the United States and Canada. *Volta Review*, *86*, 282-292.
- Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by eye II* (pp. 85-108). East Sussex, UK: Psychology Press, Ltd.
- Sheffert, S.M., Lachs, L., & Hernández, L.R. (1996). The Hoosier audiovisual multitalker database, *Research on Spoken Language Processing No. 21* (pp. 578-583). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Smeele, P.M.T., Sittig, A.C., & Heuven, V.J.v. (1994). *Temporal organization of bimodal speech information*. Paper presented at the International Conference on Spoken Language Processing, Honolulu, HI.
- Smith, Z.M., Delgutte, B., & Oxenham, A.J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*, 87-90.
- Spelke, E. (1976). Infants' intermodal perception of events. *Cognitive Psychology*, *8*, 553-560.
- Stein, B.E., & Meredith, M.A. (1993). *The merging of the senses*. Cambridge, MA: The MIT Press.
- Stevens, K.N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.
- Stoffregen, T.A., & Bardy, B.G. (2001). On specification and the senses. *Behavioral and Brain Sciences*, *24*.
- Sumby, W.H., & Pollack, I. (1954). Visual contribution of speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212-215.
- Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, *36*, 314-331.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3-51). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Lee, Y.V., & Terzepoulos, D. (1997). The dynamics of audiovisual behavior in speech. In D.G. Stork & M.E. Hennecke (Eds.), *Speechreading by humans and machines* (pp. 221-232). Berlin: Springer-Verlag.

- Vatikiotis-Bateson, E., Munhall, K.G., Kasahara, Y., Garcia, F., & Yehia, H. (1996). *Characterizing audiovisual information during speech*. Paper presented at the International Conference on Spoken Language Processing, Philadelphia, PA.
- Verbrugge, R.R., Strange, W., Shankweiler, D.P., & Edman, T.R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*, *60*, 198-212.
- Walden, B.E., Prosek, R.H., Montgomery, A.A., Scherr, C.K., & Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, *20*, 130-145.

Appendix A

Spectrograms of the various acoustic transformations

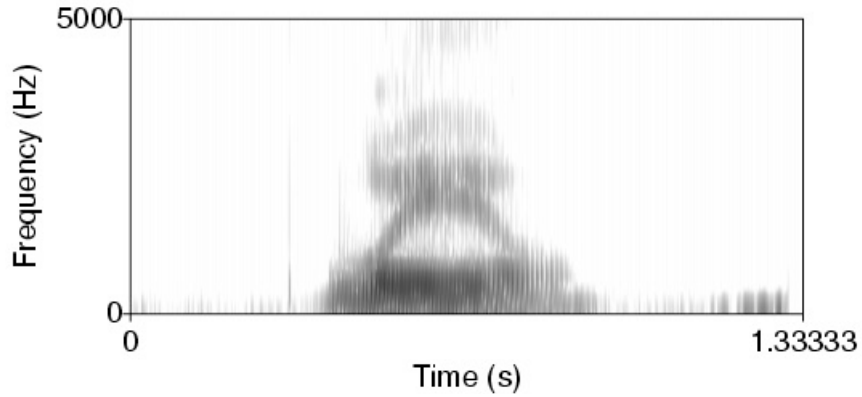


Figure A1. Wide band spectrogram of talker M2 uttering the word “wail” with no acoustic transformation.

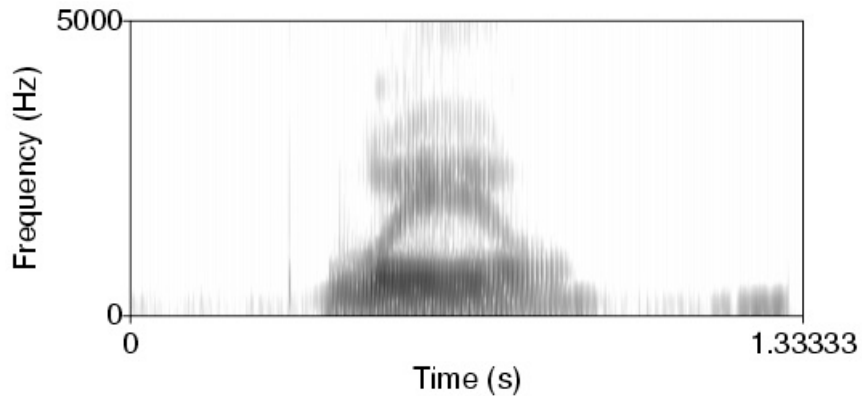


Figure A2. M2 uttering the word “wail” after a spectral shift transformation at 100 Hz.

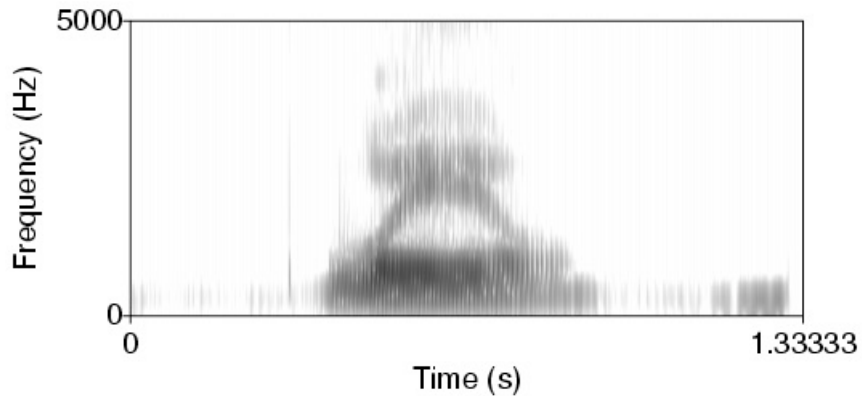


Figure A3. M2 uttering the word “wail” after a spectral shift transformation at 250 Hz.

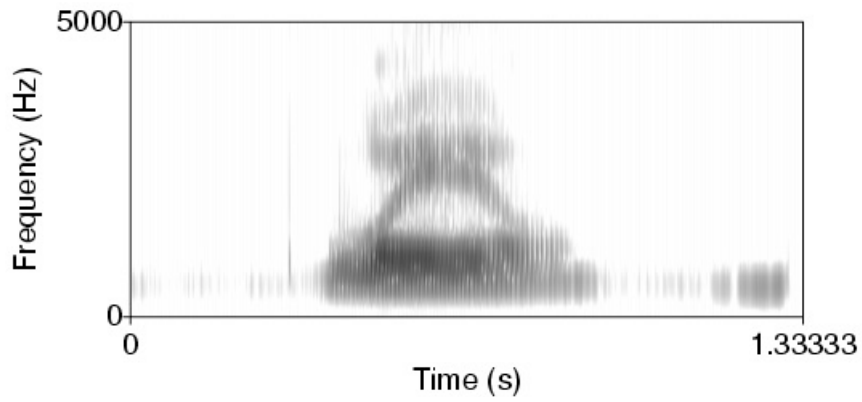


Figure A4. M2 uttering the word “wail” after a spectral shift transformation at 500 Hz.

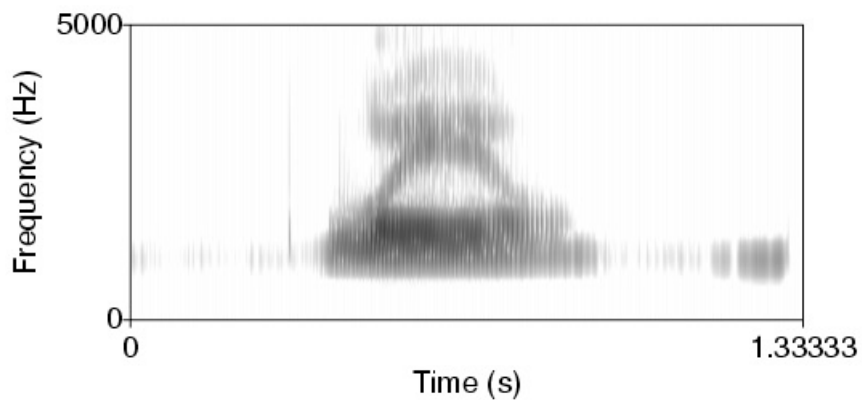


Figure A5. M2 uttering the word “wail” after a spectral shift transformation at 1000 Hz.

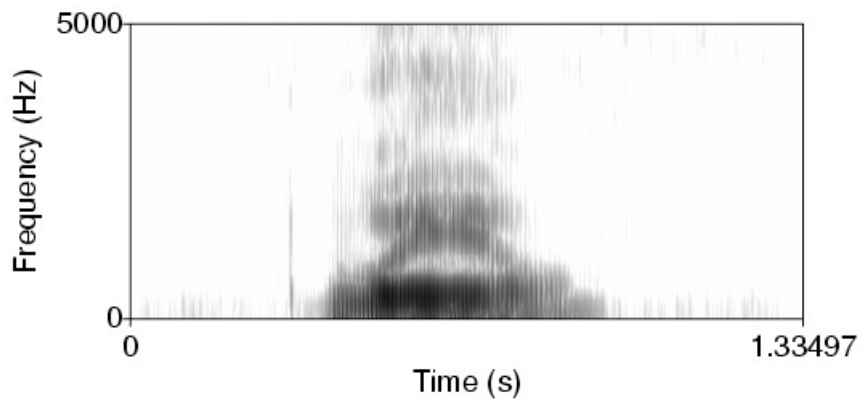


Figure A6. M2 uttering the word “wail” after a linear spectral scaling transformation at 75%.

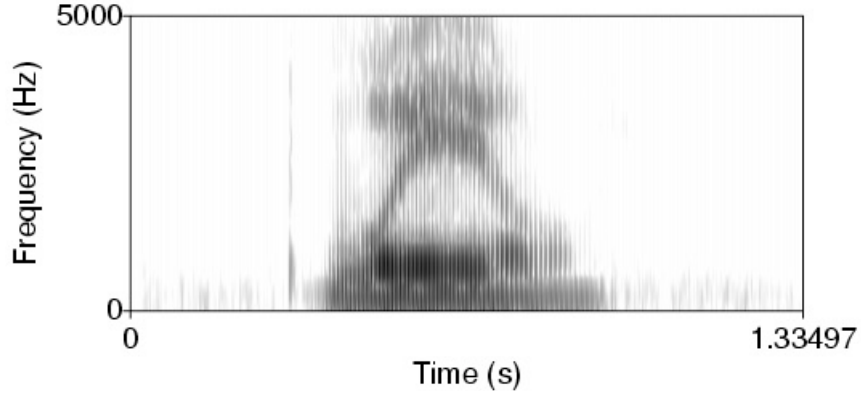


Figure A7. M2 uttering the word “wail” after a linear spectral scaling transformation at 150%.

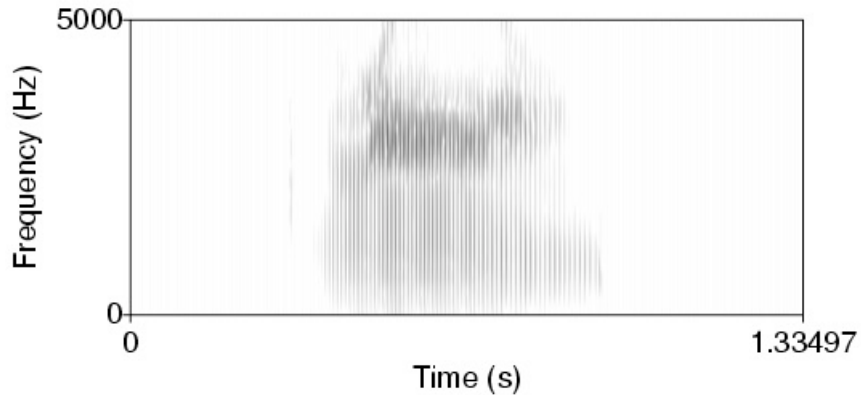


Figure A8. M2 uttering the word “wail” after a nonlinear spectral scaling transformation towards the Nyquist frequency.

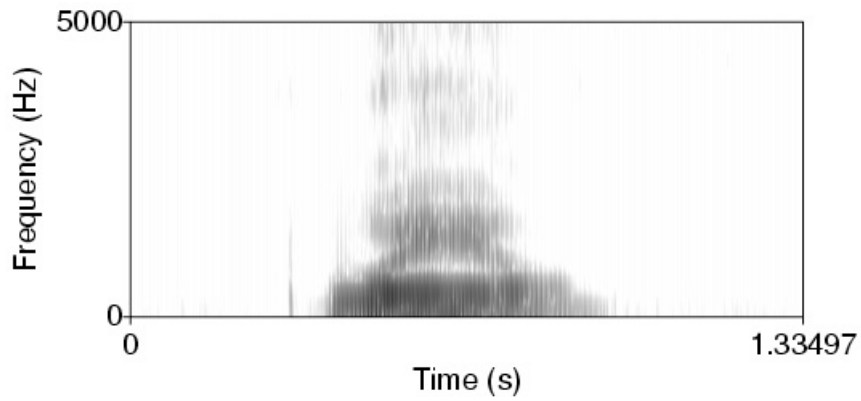


Figure A9. M2 uttering the word “wail” after a nonlinear spectral scaling transformation away from the Nyquist frequency.

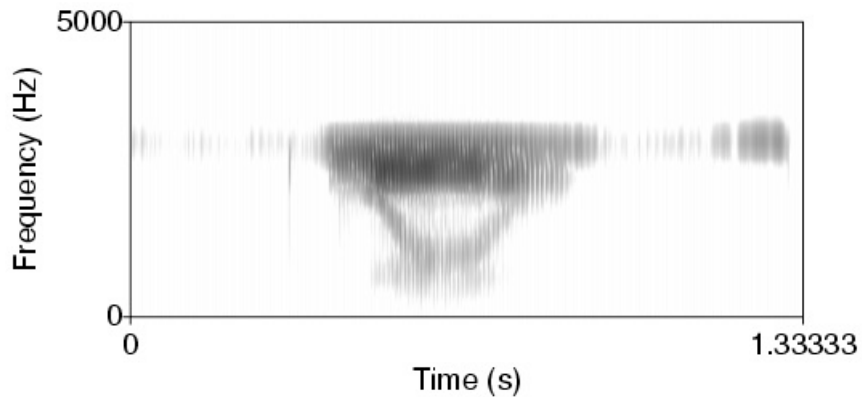


Figure A10. M2 uttering the word “wail” after a spectral inversion transformation around a center frequency of 1500 Hz.

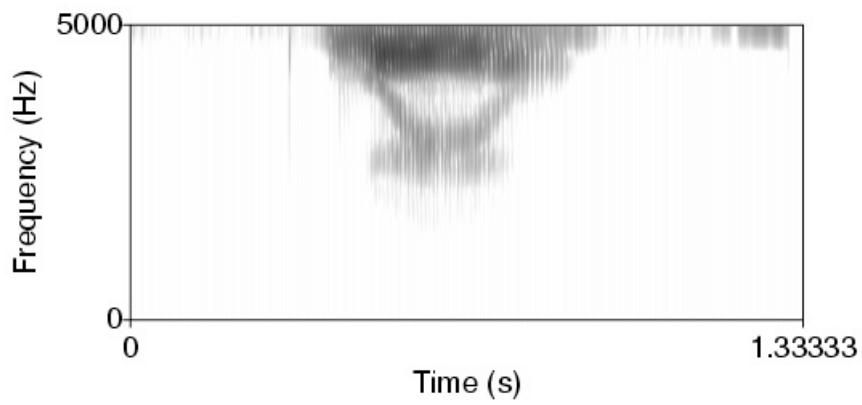


Figure A11. M2 uttering the word “wail” after a spectral inversion transformation around a center frequency of 2500 Hz.

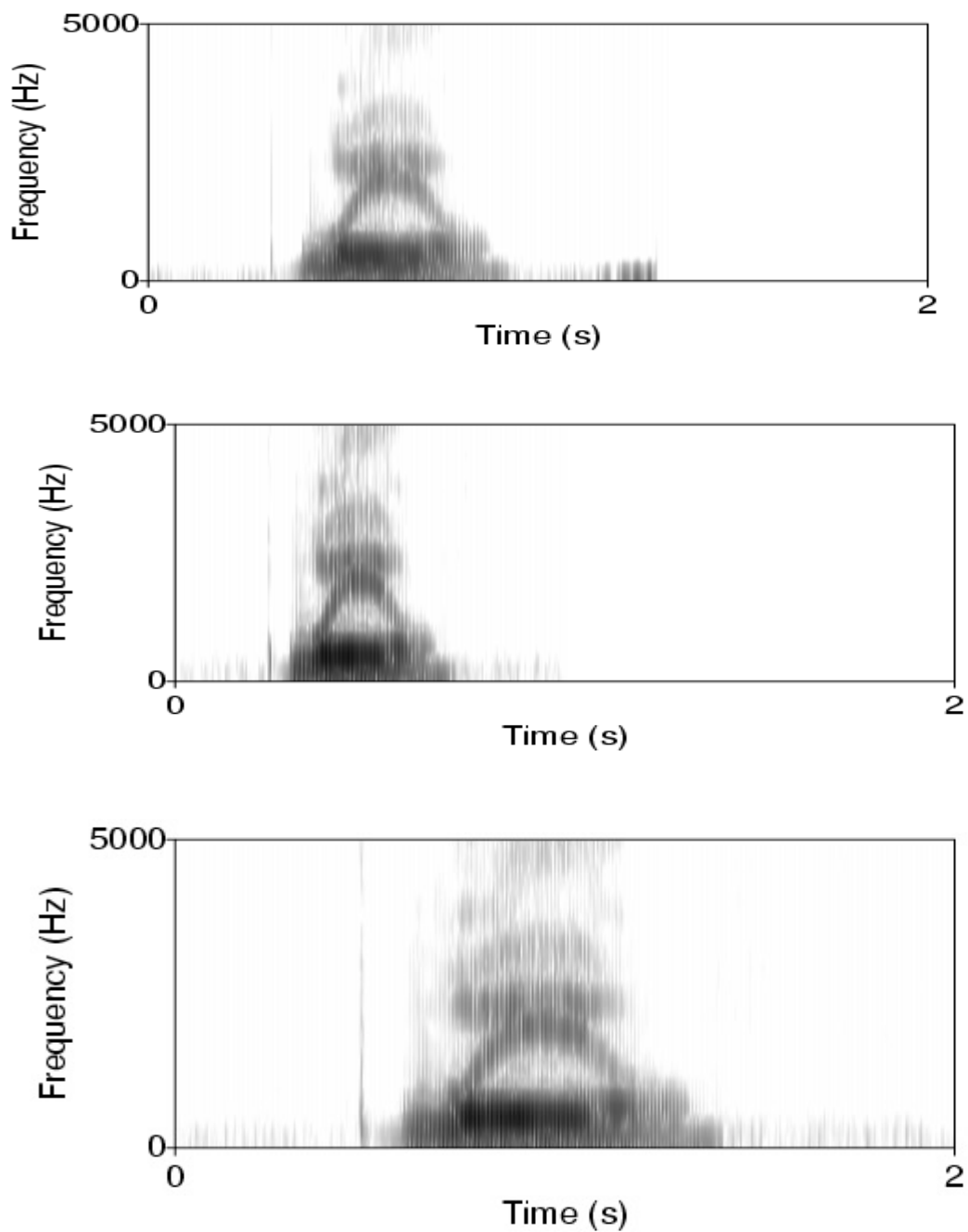


Figure A12. Examples of the linear temporal scaling transformation. The top spectrogram shows the untransformed acoustic signal of M2 uttering the word “wail.” The middle spectrogram shows the same signal after linear temporal scaling at 75% (speeded up). The bottom spectrogram shows M2 uttering the word “wail” after linear temporal scaling at 150% (slowed down).

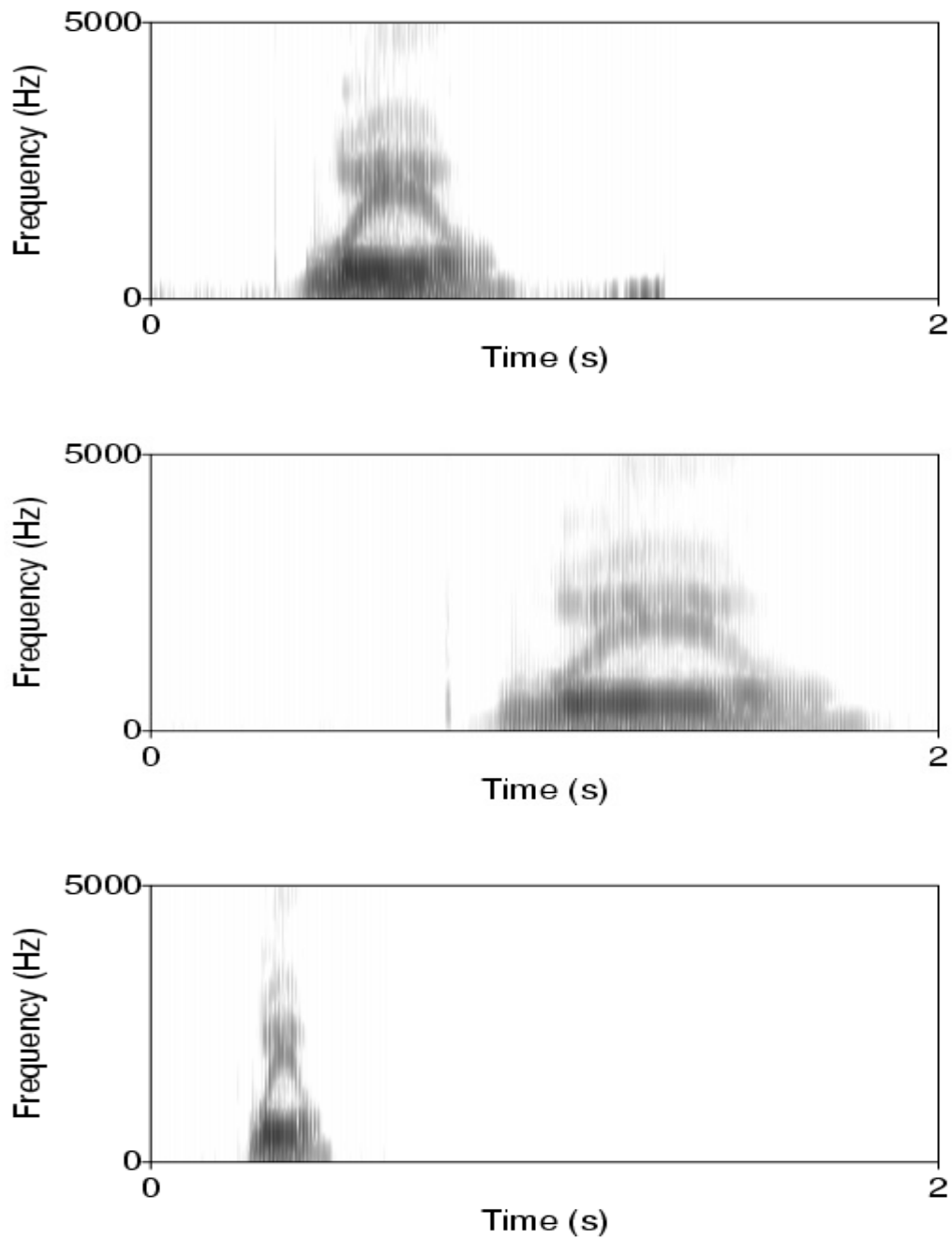


Figure A13. Examples of the nonlinear temporal scaling transformation. The top spectrogram shows the untransformed acoustic signal of M2 uttering the word “wail.” The middle spectrogram shows the same signal after decelerating nonlinear temporal scaling. The bottom spectrogram shows M2 uttering the word “wail” after accelerating nonlinear temporal scaling.