

RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 23 (1999)

Indiana University

**A Voice is a Face is a Voice: Cross-Modal Source Identification of
Indexical Information in Speech¹**

Lorin Lachs

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

¹ This work was supported by a grant from the NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University Bloomington. Special thanks go to David Pisoni and Luis Hernández for excellent suggestions and brainstorming sessions. Thanks also to Tyler Emley, Patrick Kelley and Jaime Brumfield for all their help in collecting and processing this data.

A Voice is a Face is a Voice: Cross-Modal Source Identification of Indexical Information in Speech

Abstract. Recent evidence from experiments using sinewave speech shows that the linguistic content of a message, as well as the indexical characteristics of the talker can be perceived from the isolated kinematic form of speech utterances. Similarly, isolated visual kinematic information in the form of point-light displays has been shown to behave in much the same way that full visual displays of a talker articulating do (e.g., by enhancing intelligibility in noise). If the isolated kinematic visual form of speech is informative in speech perception, and the isolated kinematic acoustic form of speech can carry indexical information, then visual information should also be able to carry information regarding the indexical properties of the talker. If this is true, then perceivers should be able to use the information about an utterance obtained through one sensory modality (e.g., vision) and use it to identify the same utterance in the other sensory modality (e.g., audition). The present study examined the ability of participants to perceive and use either auditory or visual information about articulation across sensory modalities in identifying source characteristics of a talker's voice.

Optical information about articulation has been shown to have substantial effects on speech perception (Massaro & Cohen, 1995). In the absence of auditory stimulation, visual information is sufficient for accurate speech perception (Bernstein, Demorest, & Tucker, in press). In conjunction with auditory information, visual information can enhance speech intelligibility in noise by +15 dB (Sumbly & Pollack, 1954). Alternatively, incongruent information in the auditory and visual aspects of multimodal stimuli can interact to form illusory percepts (the “McGurk” effect McGurk & MacDonald, 1976).

Because visual information about articulation can have such profound effects on speech perception, some theorists have proposed that the perceptually useful information in speech signals must be transmittable via acoustic as well as optic media. Indeed, some researchers have gone so far as to propose that the information is amodal; that is, the information for speech is not constrained to any particular sensory modality. In fact, the McGurk effect has been replicated by using auditory and tactile information about speech (Fowler & Dekle, 1991) demonstrating that some degree of useful information about speech can be obtained through sensory modalities other than audition.

In another experiment designed to demonstrate that speech information can be carried in multiple sensory modalities, Green and Kuhl (1989) showed that the perceived VOT boundary for a synthetic /bi-/ /pi/ continuum was shifted toward the VOT boundary for a /di-/ /ti/ continuum when the stimuli were paired with the visual specification of a talker uttering the syllable /gi/. That is, the VOT boundary shifted in an appropriate manner for the illusory percept invoked by the McGurk illusion. In an earlier study, Green and Miller (1985) showed that the speaking rate information in optical displays influenced the identification of voiced or voiceless segments on an acoustic continuum that remained constant. These studies demonstrate that the dynamic aspects of visual information play a role in the perception of speech, and that the same auditory information can be perceived differently depending on the kinds of visual information available during perception.

The exact form of such information remains the subject of some debate, but a growing body of research points to the possibility that the acoustic or optical forms of speech signals carry kinematic or dynamic information about the articulation of the vocal tract, and that such information drives the

perception of linguistically relevant utterances (Fowler, 1986; Fowler & Rosenblum, 1991; Liberman & Mattingly, 1985; Rosenblum & Saldaña, 1996; Summerfield, 1987). The use of dynamic information has been demonstrated across multiple contexts. For example, Green and Gerdeman (1995) showed that cross-modal discrepancies in the *vowel* portion of McGurk stimuli influenced the degree to which the *consonant* portion was susceptible to the McGurk effect. The findings suggest that the perceptual system must be sensitive to non-segmental, coarticulatory information when it attempts to make sense of multimodal inputs.

Another method used to study the problem of audiovisual integration in speech perception is the point-light technique (Johansson, 1973). By placing small reflective patches at key positions on a talker's face and darkening everything else in the display, one can isolate the kinematic aspects of visual displays of talkers articulating speech (Rosenblum, Johnson, & Saldaña, 1996; Rosenblum & Saldaña, 1996). Such "kinematic primitives" have been shown to behave much like unmodified, full visual displays of speech (Rosenblum & Saldaña, 1996). For example, the McGurk illusion can be induced by dubbing visual point-light displays onto phonetically discrepant auditory syllables (Rosenblum & Saldaña, 1996). In addition, an extension of Sumby and Pollack's (1954) findings has demonstrated that providing point-light information about articulation in conjunction with auditory speech embedded in noise can result in increased intelligibility (Rosenblum et al., 1996).

All of the studies reviewed above, and indeed, most of the previous investigations of the effects of audiovisual information on speech perception have focussed on what are commonly referred to as the *linguistic* aspects of the signal: phoneme or syllable identification and spoken word recognition. However, a growing body of literature has shown that speech signals also carry information about the *indexical* properties of the talker, and that this information is perceived, stored in memory and used during speech perception and spoken word recognition (see Goldinger, 1998; Pisoni, 1997, for a review).

Numerous recent studies have shown that the indexical properties of a talker's voice are stored in long-term memory (Bradlow, Nygaard, & Pisoni, 1999; Goldinger, Pisoni, & Logan, 1991; Martin, Mullennix, Pisoni, & Summers, 1989). For example, using a continuous recognition task, Palmeri, Goldinger, and Pisoni (1993) showed that repeating a word in the same voice that produced it during study facilitated later recognition of that word. Furthermore, the size of this effect did not change depending on the number of talkers uttering test items. This suggested that the encoding of voice attributes in memory is automatic and not controlled by strategic processes.

The link between memorial encoding of fine-grained details of spoken words and perceptual processes has also been established (Nygaard, Sommers, & Pisoni, 1994; Nygaard, Sommers, & Pisoni, 1995). In one experiment, Nygaard and Pisoni (1998) trained participants to identify a set of novel talkers from their voices alone. Once the participants had learned the voices using a set of training stimuli, Nygaard and Pisoni found that the knowledge of talker characteristics obtained also generalized to new stimuli. Furthermore, the perceptual learning of the trained voices transferred to a novel task: words spoken by familiar voices were recognized more accurately in noise than words spoken by unfamiliar voices.

But what kind of information about a talker is contained in speech, and how does that information contribute to speech perception? In an examination of the acoustic correlates of talker intelligibility, Bradlow, Torretta and Pisoni (1996) showed that while global characteristics such as fundamental frequency and speaking rate had little effect on intelligibility, acoustic-phonetic properties of voice, such as vowel space reduction and "articulatory precision", were strong indicators of overall intelligibility. These findings suggest that indexical properties of a talker may be completely intermixed with the

phonetic realization of the utterance, with no real dissociation between the two sources of information in the speech signal.

More direct evidence for this hypothesis comes from recent studies using sinewave replicas of speech. Sinewave speech is made by generating sinusoidal tones that trace the center frequencies of the three lowest formants produced during a natural utterance. The resulting tone complex sounds completely unnatural, but can be perceived by listeners as speech (Remez, Rubin, Pisoni, & Carrell, 1981). Indeed, not only is the linguistic content of the utterance perceptible, but specific aspects of a talker's unique identity are also preserved in sinewave replicas. Remez, Fellowes, and Rubin (1997) showed that participants could identify specific familiar talkers from sinewave replicas of their utterances. This finding is remarkable because sinewave speech patterns preserve none of the traditional stimulus aspects that cue vocal identity, such as fundamental frequency, or the average long-term spectrum. Furthermore, Fellowes, Remez, and Rubin (1997) also showed that while the gender of talkers could be perceived from sinewave replicas, the correct perception of gender was not a precondition for the identification of a specific talker. In the words of Fellowes et al., (1997), "Personal information is available in an aspect of the signal that does not arise through anatomical variation alone" (p. 848).

These findings raise some important questions about the domain of speech perception. Sinewave speech strips an utterance of all information except the time-varying properties of the resonances generated by articulatory motion. In this way, sinewave speech is much like a point-light display; it isolates the kinematic information in an acoustic display that relates to the underlying articulation. If the common metric for auditory and visual information about speech is kinematic, and kinematic auditory information has been shown to provide information for the identity of a talker, then visual kinematics should also provide the same kind of information. However, evidence for the visual perception and use of nonlinguistic information from articulatory activity has been reported only recently. In one study, Gagné, Masterson, Munhall, Bilida, and Querengesser (1994) showed that the visual intelligibility of tokens spoken in the "clear speech" speaking style was higher than the visual intelligibility of tokens spoken in a conversational style. In another study, Rosenblum, Yakel, Baseer, and Panchal (1999) showed that visual point light displays of a talker articulating could be matched accurately to unaltered visual displays of the same talker articulating.

If the important information for talker identity and source recognition in speech is contained in kinematic and dynamic aspects of articulating vocal tracts, then perceivers should be able to match the identity of a speaking face across sensory modalities, because the criterial information is not necessarily carried *solely* by the acoustic specification of speech. In theory, kinematic and dynamic information is modality-neutral, and can be conveyed by both optical and acoustic media. The question, then, is whether perceivers can actually use this information to make judgments of source identity across sensory modalities.

In order to answer this question, an experiment was designed that used a 2-alternative forced choice paradigm. Perceivers were required to match a talker presented in one modality with the same talker presented in the other modality. The two sensory modalities used for this experiment were visual and auditory. Several factors might play a role in the perceiver's ability to perform such a task. First, hit rates would be expected to be very high if the two alternatives were of different genders. Accordingly, all the talkers identified by a particular participant were of the same gender. Different groups of participants identified male or female talkers, in order to test for differences in performance based on the gender of the talkers. A second factor that might also play a role in this matching task was the order (or direction) in which the judgment was made. For example, it might be the case that seeing a face and then judging which of two voices matched it is easier than the converse situation: hearing a voice and judging which of two faces matched it. Both conditions were tested as repeated measures to find any difference in

performance dependent on the order in which the modalities were presented. In addition, it is possible that fine-grained details of the stimulus 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 in audio-only identification tests, stimulus items were split into low and high groups according to their visual intelligibility based on visual-only identification tests (Lachs & Hernández, 1998). The visual-only (VO) intelligibility of the stimulus items was manipulated as a repeated-measures variable. Finally, confidence ratings were collected in order to determine whether participants were aware of the particular trials on which they performed well. If confidence ratings were higher on correct trials, and lower on incorrect trials, then participants must have a good estimate of their ability to perform this unusual task.

Method

Experimental Design

A two-alternative forced choice procedure was used in a 2 x 2 x 2 repeated measures factorial design. The two levels for the between-subjects Gender factor were “male” and “female”. The two levels for the within-subjects Direction factor were “A-V” (where participants identified the correct visual stimulus after viewing the test auditory stimulus) and “V-A” (where participants identified the correct auditory stimulus after viewing the test visual stimulus). The levels of this factor were blocked and counterbalanced across participants for the order in which they were presented. 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 HAVMD (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). The levels of this factor were randomly distributed in each block for the Direction factor.

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.

Stimulus Materials

Two Apple Macintosh computers, each equipped with a 17” Sony Trinitron Monitor (0.26 dot pitch) and a TARGA 2000 video processing board were used to present the visual stimuli to subjects. The video processing boards were each capable of handling clips digitized at 30fps with a size of 640 x 480 and 24-bit resolution. Auditory stimuli were presented over Beyer Dynamic DT100 headphones.

The stimuli were a subset of tokens selected from the Hoosier Audiovisual Multitalker Database ((HAVMD, Lachs & Hernández, 1998; Sheffert, Lachs, & Hernández, 1996). The video portion of HAVMD tokens was digitized at 30 fps with 24-bit resolution with 640 x 480 pixel size. The audio portion of HAVMD tokens was digitized at 22 kHz with 16-bit resolution. Movie clips from eight talkers were used in this study (F1, F2, F3, F4, M1, M2, M3, and M4).

Procedures

Participants were told that they would be seeing two blocks of stimulus trials. In one block (the “V-A” level of the Direction factor), they would see a video clip of a talker uttering an isolated, English word, but they would not be able to hear it. Shortly after seeing this video display, they would be presented with two audio clips. One of the clips would be the same talker they had seen in the video, while the other clip would be a different talker. Participants were instructed to choose which audio clip matched the talker they had seen. The same instructions were provided for the trials in which the participant heard the audio clip first, and had to make their decision based on two video displays (the “A-V” level of the Direction factor).

On each trial, the test stimulus was either the video or audio portion of one movie token based on an isolated word spoken by one talker. The correct target choice was the other (cross-modal) portion of the same movie. The distractor choice was the same word spoken by one of the other three talkers, presented in the same modality as the target alternative. The order in which the target and distractor choices were presented was randomly determined on each trial. All responses were made with the mouse and recorded in a log file for further analysis.

After a response was made using dialog boxes and mouse inputs, participants were asked to record a confidence judgment for their response. The ratings were made on a scale of 1 to 7, with 1 marked as “not confident at all” and 7 marked as “very confident”. At the completion of the session, participants were asked to briefly describe any strategies they used in order to accomplish the task.

Results

Determining Chance Performance

The data from each participant was analyzed first to determine if his/her performance in either the “A-V” or “V-A” conditions differed significantly from chance. A binomial distribution was used to calculate the probability that the number of successful trials in each block was due to chance performance ($p(\text{correct}) = 0.5$). Chance performance was rejected if $p < 0.05$.

Because the Direction of Judgment factor was manipulated within-subjects, the data from 40 participants was available for the V-A and A-V conditions. The accuracy of 11 participants in the “V-A” condition did not significantly differ from chance. However, of those 11 participants, the performance of 8 participants was only marginally attributable to chance ($p < 0.1$). Similarly, the accuracy of 15 participants in the “A-V” conditions were not significantly different from chance, although 7 participants were within marginal range ($p < 0.1$). These data indicate that a little less than 75% of the participants were able to accomplish this task, with slightly fewer being able to perform as well in the “A-V” conditions.

Of the 40 participants, 33 were able to respond above chance in at least one of the “V-A” and “A-V” conditions. This indicates that most participants were able to do the matching task in one form or the other. The 7 participants who did not perform above chance in either condition were eliminated from the final data analysis.

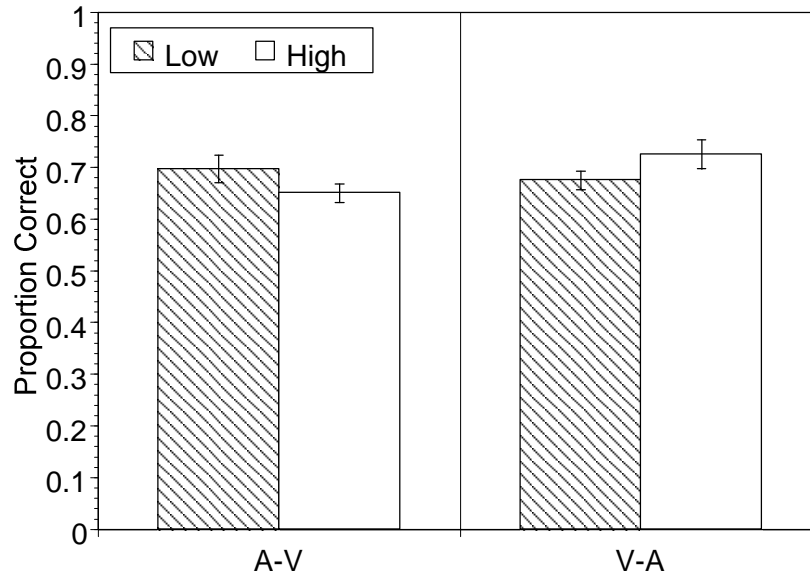


Figure 1. The interaction between visual intelligibility and the direction of the judgment. The striped bar represents performance when intelligibility is low, and the open bar represents performance when intelligibility is high.

Accuracy Analyses

The data obtained from 33 participants were then submitted to a 2 (Direction) x 2 (Visual Intelligibility) x 2 (Gender) x 2 (Order) repeated-measures ANOVA using percent correct as the dependent variable. Differences were evaluated with an α of 0.05. No main effects for any of these variables were observed, but the ANOVA did reveal a significant interaction between Direction and Intelligibility, $F(1, 29) = 5.743$, $MSE = 0.075$, $p = 0.023$, $\eta^2 = 0.165$. This interaction is illustrated in Figure 1. The set of bars on the left represent performance in the A-V trials, while the set of bars on the right represent performance on the V-A trials. Within each set of bars, the shaded bar represents performance when stimuli were of low visual intelligibility and the open bar represents performance when stimuli were of high visual intelligibility. Examination of this figure shows that participants performed better on V-A trials when the words were of high visual intelligibility. In contrast, there was a small difference in the degree to which low visual intelligibility words influenced participants' ability in the A-V direction. Post-hoc pairwise comparisons revealed that the source of this interaction could be localized in a significant difference between performance on low and high intelligibility words in the V-A direction.

The analysis also revealed a significant interaction between Direction and Order, $F(1, 29) = 9.438$, $MSE = 0.152$, $p = 0.005$, $\eta^2 = 0.246$. Differences in performance in either direction depended on whether they received that direction in the first or second block of the experiment. Figure 2 illustrates this interaction. The left panel shows the average performance of participants in the A-V block, and the right panel shows average performance in the V-A block. Within each panel, the shaded bar represents performance when the participant received the A-V block first, and the open bar represents performance when the participant received the V-A block first. As shown in the figure, participants performed better in the A-V block when they had received the V-A block first. In contrast, participants performed better in

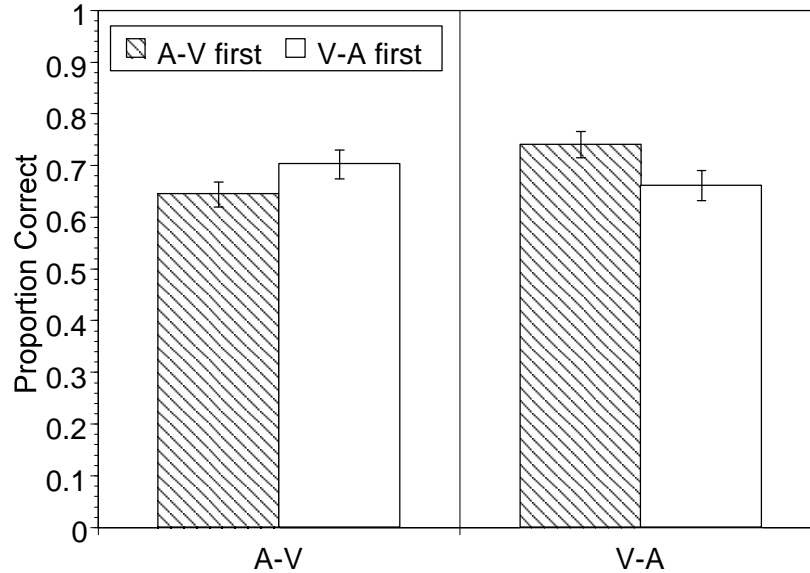


Figure 2. The interaction between the direction of judgment and the counterbalancing order. The striped bar shows performance when the A-V condition was presented first, and the open bar shows performance when the V-A condition was presented first.

the V-A block when they had received the A-V block first. That is, participants tended to be more accurate in the second Direction block they received. Post-hoc pairwise comparisons showed that significant improvement in the second block was only seen when participants received the V-A block first.

Because the counterbalancing order affected performance, the Direction factor was replaced with another within-subjects factor (“Block”), which consisted of two levels: “first” and “second”. This analysis was carried out in order to examine the effects of block order regardless of the specific conditions implemented in the block. The reapporioned data were submitted to a 2 (Block) x 2 (Intelligibility) x 2 (Gender) x 2 (Order) repeated-measures ANOVA. As reported earlier, there was a main effect of Block, $F(1, 29) = 9.438$, $MSE = 0.152$, $p = 0.005$, $\eta^2 = 0.246$. Performance in the second block ($M = 0.721$, $S.E. = 0.019$) was always better than performance in the first block ($M = 0.652$, $S.E. = 0.019$), regardless of the direction in which the judgments were made.

In addition, the analysis revealed a significant interaction between Intelligibility and Block, $F(1, 29) = 7.636$, $MSE = 0.1$, $p = 0.01$, $\eta^2 = 0.208$. The degree to which the visual intelligibility of the stimuli affected performance depended on whether the trial was in the first or second block. Figure 3 shows the average performance of participants on low and high visual intelligibility stimuli in the first and second blocks. There was clearly an improvement in performance across the blocks in the ability of participants to perform this task when given low visual intelligibility stimuli. However, no improvement is observed for the high visual intelligibility stimuli. Post-hoc pairwise comparisons confirmed that this interaction was due to a difference in performance across blocks for low intelligibility words, but not for high ones. Interestingly, this improvement occurred *regardless of the direction in which the matching judgments were made*.

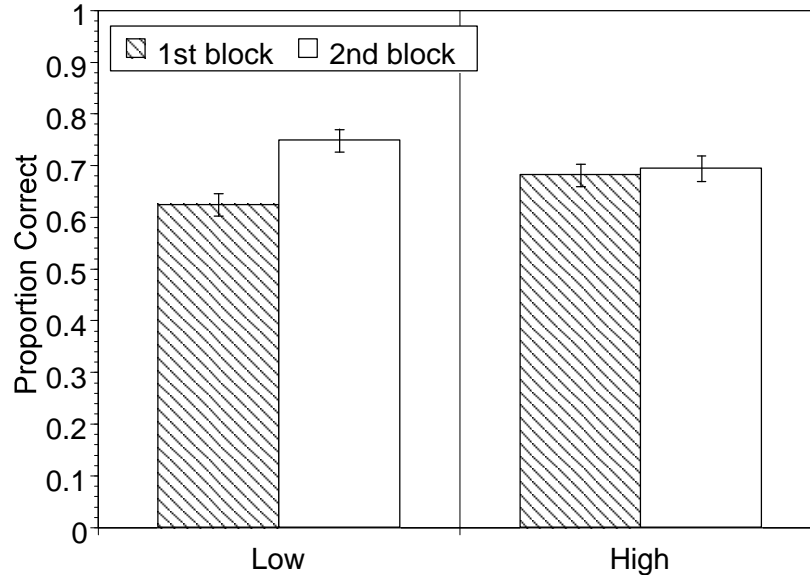


Figure 3. The interaction between Block order and visual intelligibility. Striped bars represent performance on the first block of trials, while the open bar represents performance on the second block of trials.

Finally, this analysis revealed a significant three-way interaction between Block, Intelligibility and Order, $F(1, 29) = 5.743$, $MSE = 0.075$, $p = 0.023$, $\eta^2 = 0.165$. This is shown in Figure 4. The improvement in performance from the first to second block for low intelligibility words was affected by which direction was presented first. Simple effects analyses were conducted to examine this interaction at each level of Counterbalance Order. These analyses showed a significant interaction between Block and Intelligibility when the V-A block was presented first (right panel of Figure 4; $F(1, 12) = 6.957$, $MSE = 0.152$, $p = 0.022$, $\eta^2 = 0.367$). Post-hoc pairwise comparisons between the conditions at this level of Order revealed that the interaction was due to the difference in performance between blocks for stimuli with low visual intelligibility. The difference between blocks for high intelligibility words was not significant. No interaction between Block and Intelligibility when the A-V block was presented first (left panel of Figure 4; $F < 1$, n.s.).

Summary of accuracy scores. In summary, the accuracy analyses revealed several interesting facts about this unusual task. First, the majority of participants were able to make judgments about talker identity across sensory modalities. The ability to perform this task did not seem to be affected by the direction in which judgments were made. In addition, participants got better at making cross-modal judgments in the second block with which they were presented. This improvement was mainly for stimulus items with low visual intelligibility. Furthermore, participants only showed this improvement in the second block for low intelligibility items when the V-A block was presented first and the A-V block was presented second.

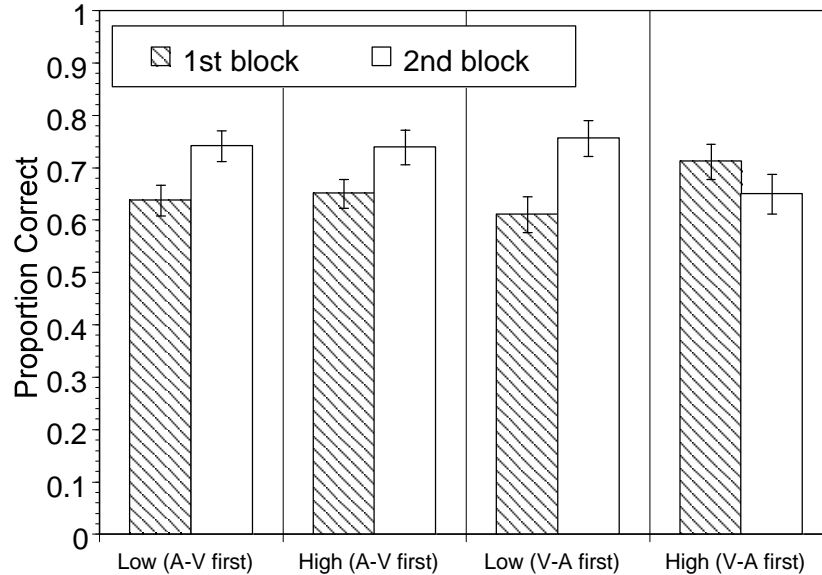


Figure 4. The interaction between Block order and visual Intelligibility. Striped bars represent performance on the first block of trials, while the open bar represents performance on the second block of trials.

Confidence Ratings

Because each participant used a different range of the 7-point confidence scale, confidence ratings were transformed relative to the scale across which each participant made his/her judgments. Before normalization, there was more variability across participants in the average confidence rating given ($M = 4.47$, $SD = 0.74$), than there was in the average variability of confidence ratings ($M = 1.2$, $SD = 0.37$). Thus, the normalized confidence ratings adjusted for differences in the “midpoint” of the scale used by different participants, without drastically changing the “range” over which confidence judgments were made.

In analyzing the confidence ratings, both correct and incorrect responses were examined. The “Score” factor had two levels: “correct” and “incorrect” and was used to determine whether there were differences in confidence based on the accuracy of responses. The normalized confidence ratings were then submitted to a 2 (Direction) \times 2 (Visual Intelligibility) \times 2 (Score) \times 2 (Gender) \times 2 (Order) repeated-measures ANOVA. A significant main effect of Score was observed, $F(1, 29) = 56.068$, $MSE = 2.612$, $p < 0.01$, $\eta^2 = 0.659$. As shown by the effect size statistic (η^2), this main effect accounted for approximately 66% of the variance in confidence ratings. Confidence ratings on correct trials ($M = 0.133$, $SE = 0.018$) were significantly higher than confidence ratings on incorrect trials ($M = -0.069$, $SE = 0.01$). Apparently, participants were aware of their ability to perform this task. However, the relatively low average values for both correct and incorrect trials indicate that participants remained equivocal on most trials. The large effect size with such low values also indicates that the extremes of each participant's confidence scale did not vary much around their average confidence rating.

The 5-way ANOVA on normalized confidence ratings also revealed a significant interaction between Direction and Visual Intelligibility, $F(1, 29) = 5.985$, $MSE = 0.038$, $p = 0.021$, $\eta^2 = 0.171$. This interaction is illustrated in Figure 5. Post-hoc pairwise comparisons revealed that the confidence ratings in the V-A/Hi intelligibility condition were significantly different ($p < 0.05$) from those in either A-V condition. As shown in the figure, confidence ratings in the V-A/Hi intelligibility condition were on than

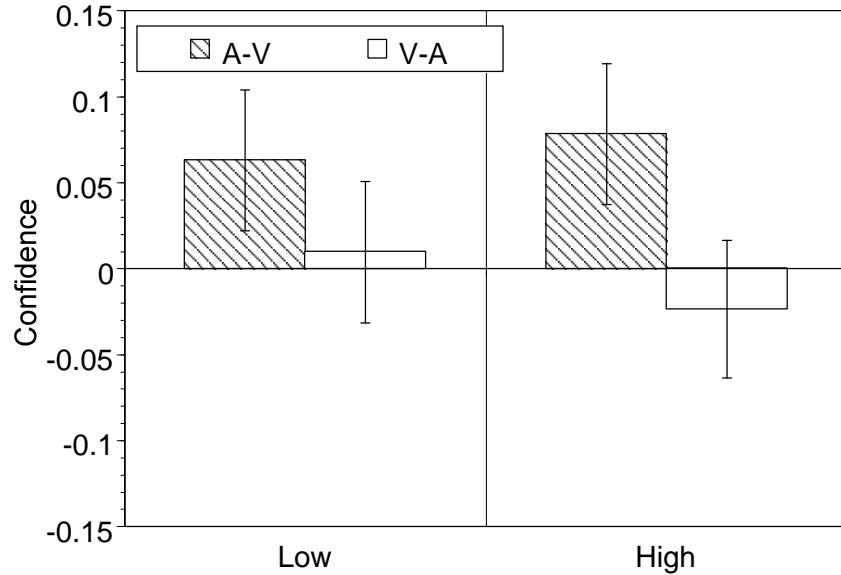


Figure 5. The interaction between Direction and Visual Intelligibility. Striped bars represent confidence on the A-V trials, while the open bar represents confidence on the V-A trials.

the average confidence ratings given in any of the other conditions, regardless of whether or not the trial was correct. Paradoxically, participants were less confident of their performance on V-A trials when they were making their decision based on highly useful visual information for spoken word recognition. This unusual result is discussed more fully in the discussion section below.

Interestingly, there was also a significant 3-way interaction between Direction, Intelligibility and Score, $F(1, 29) = 5.316$, $MSE = 0.264$, $p = 0.028$, $\eta^2 = 0.155$. Figure 6 shows this interaction. The left panel in the figure shows scores in the A-V conditions; the right panel shows scores in the V-A conditions. Within each panel, the left set of bars shows performance when the visual intelligibility of the stimulus was low; the right set of bars shows performance when the intelligibility was high. Shaded bars show performance on trials that were correct and open bars show performance on trials that were incorrect. It is very clear from the figure that confidence ratings for incorrect trials were generally much lower for V-A trials than they were for A-V trials. This pattern indicates that participants were more sensitive to their performance in the V-A block than in the A-V block. For trials they got incorrect, they were less confident. This pattern was more pronounced for low visual intelligibility stimuli than it was for high visual intelligibility stimuli.

In order to understand the nature of this interaction more fully, the difference in confidence ratings between correct trials and incorrect trials for each participant were computed. These difference scores were then submitted to a 4-way repeated measures ANOVA using Direction, Intelligibility, Gender and Order as factors. Figure 7 illustrates the significant interaction between Direction and Intelligibility ($F(1, 29) = 5.316$, $MSE = 0.529$, $p = 0.028$, $\eta^2 = 0.155$) revealed by this analysis. As shown here, the difference in average confidence ratings for correct and incorrect trials was greatest in the V-A condition when stimulus items were of low intelligibility. Post-hoc pairwise comparisons showed that the difference in this condition was significantly larger than the differences in either the V-A high intelligibility condition or the A-V low intelligibility condition.

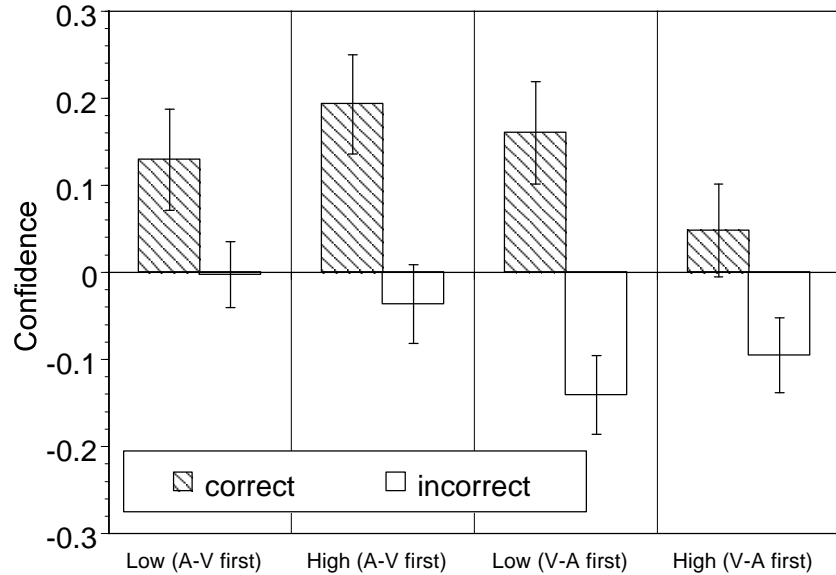


Figure 6. The interaction between Direction, Visual Intelligibility and Score. Striped bars represent confidence on correct trials, while the open bar represents confidence on incorrect trials.

Discussion

The present cross-modal source identification experiment examined the ability of participants to perceive and use auditory or visual information about articulation across sensory modalities. Participants were presented with the unimodal form of a spoken word token and required to choose which of a pair of cross-modal tokens specified the same talker. Roughly three-fourths of the participants tested were able to perform this task with better than chance performance. In addition, participants were aware of the tokens that they correctly matched, as shown by the higher confidence ratings for correct trials.

Perceptual learning was also observed in this experiment. Participants tended to perform better on the second block of trials than on the first. This learning effect was due to an increase in accuracy when the stimuli were low VO intelligibility words. Furthermore, improvement in the second block was only seen when the first block a participant experienced was in the V-A direction. Experiencing the V-A block first may have focused participants' attention on fine-grained details of optical movement in the articulator area. Stimuli with low visual intelligibility may have increased attention to these fine-grained details than those with high intelligibility. Thus, when the matching task switched to the A-V direction, participants had already learned to attend to those aspects of low visual intelligibility stimuli that would aid in the completion of the task (possibly including their acoustic correlates).

However, no such knowledge was acquired by participants who received the A-V direction first, because the low visual intelligibility of the words was only apparent *after* the test stimulus was presented. That is, the *acoustic* form of a low *visual* intelligibility word does not necessarily focus attention on fine-grained details of the stimulus, since it is not necessarily hard to perceive. Because of this, participants who experienced the A-V block first may have approached the V-A block unprepared to deal with the ambiguous or noisy information in low intelligibility words for which they were asked to make a choice.

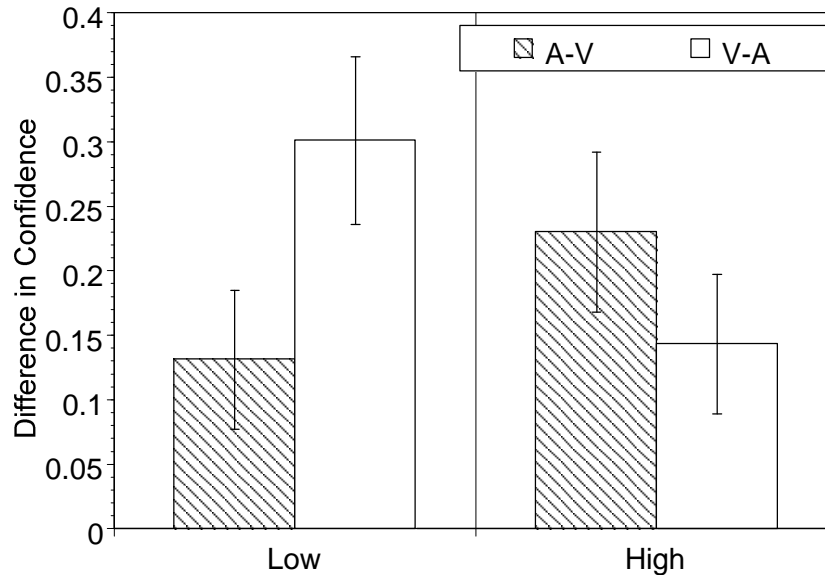


Figure 7. The interaction between Direction and visual Intelligibility for the difference in confidence ratings for correct and incorrect trials. Striped bars represent confidence on the A-V trials, while the open bars represent confidence on the V-A trials.

The patterns observed for the confidence ratings seem quite paradoxical. Although confidence was higher for trials scored correct than for those scored incorrect, overall confidence ratings in the experimental conditions did not seem to match performance levels. Despite the fact that performance was not higher in the A-V conditions than in any of the others, confidence ratings for the A-V direction of judgment were generally higher than average, with the highest ratings observed when the visual intelligibility of the tokens was high. Surprisingly, although performance was generally *best* in the V-A conditions where tokens were of high visual intelligibility, confidence ratings to these trials were on average lower than those given in any of the other conditions. Confidence ratings in the V-A/low condition were about average. This may be because participants expected that their performance in the V-A condition would be very good. If participants also expected that low intelligibility words would not be as useful as high intelligibility words, then their confidence might be placed at a “midway point” on the scale.

Interestingly, the *difference* in confidence ratings between correct vs. incorrect trials was greatest for the V-A low condition. This pattern indicates that, for this condition more than any other, there was almost a perceived dichotomy in the ability to perform the task. Remember that accuracy scores in the V-A/High intelligibility condition were significantly higher than scores in the V-A/Low intelligibility condition. Since low and high intelligibility words were interspersed throughout the blocks, it is tempting to speculate that the relatively “easy” nature of trials in the V-A/High condition set up an expectation for continued high performance on all V-A trials, exaggerating differences in confidence ratings for correct and incorrect trials.

Of course, the current experiment was not designed to test for these possibilities, so the explanations offered above concerning learning and bias in this task remain speculative at best. What is incontrovertible about the present investigation's findings, however, is that participants *can* perform this unusual matching task. Furthermore, they can do so with surprising accuracy. This provides further

support for the notion that the information about a talker's voice is inherently amodal. It can be carried by either visual or acoustic energy and it can be perceived by either the visual or auditory systems.

One surprising result was the absence of asymmetries in the direction over which the matching judgments were made. Because there *are* differences in the amount of information that acoustic and optic displays can carry about the motion of the vocal articulators, this result deserves further study. According to the Source-Filter model of speech production (Fant, 1960), the resonances in the frequency spectrum characteristic of the speech signal are directly related to the configuration of the vocal tract. The motion of the articulators causes changes in the formants' central frequencies over time. Accordingly, the acoustic form of speech can carry information about the positions and movements of the vocal articulators from the lips to the larynx. But, the same is not true for the optic form of speech. Visual displays can carry information about the configuration of the lips, tongue tip, and jaw, but it is very unlikely that they can carry information about the configuration of the velum, or show that there is a closure in the glottal area (Dodd & Campbell, 1987; Summerfield, 1987).

It is possible, however, that the effects of these asymmetries interact with the amount of phonetic information in the signal itself. After all, if the visual signal is already known to be sufficiently informative for spoken word recognition, the limitations in its ability to carry information about the non-visible articulators may be irrelevant. The data above showed that changing the visual intelligibility of the test stimulus affected performance when making judgments in the V-A direction. It is possible that adjusting the intelligibility of the audio portion of these tokens will affect performance in the A-V direction, as well.

Additional support for the proposal that the useful information in the current task is kinematic or dynamic in nature is provided by responses from a post-hoc analysis of the interview conducted at the end of each participant's session. All participants were asked to briefly describe any strategies they were using to accomplish the task. Most participants listed at least two strategies, frequently citing a preference for one or the other. A tally was made of the different types of strategies mentioned in the exit interviews and is shown in Table 1.

The "Enunciation/Emphasis" category refers to strategies that made reference to the use of information relating to the emphasis with which words were spoken or the enunciation of segments within a word. Typical responses included "I looked at their mouths to see if the way their mouths were moving was like the way I heard it". One interesting aspect of these responses was that they almost naturally referred to the optical display as "sounding like" something. Clearly, most of the participants involved in the study used this information in forming their responses. Such information refers to the dynamics of articulatory movement: emphasis is related to the forcefulness of articulator movement, enunciation is related to the forcefulness and precision with which articulatory movements are made. Without explicit instructions, participants apparently focused their attention on what is hypothesized to be the relevant information in multimodal displays. "Duration", the second most popular strategy used, is also a dynamic cue relating to the kinematics of articulatory motion. Responses in this category suggest that duration differences could often distinguish one candidate from another.

The remaining strategies mentioned fall in another class entirely from the two most frequent ones. While the most frequent ones rely on the use of stimulus-driven properties, the rest seem to be more "top-down" or "heuristically-driven". "Learning" strategies specifically described how the task got easier once associations had been learned between specific voices and faces. "Expectation" strategies noted that some of the talkers seemed to "look like" they possessed certain vocal characteristics such as a high fundamental frequency. Participants who used the "self-repetition" strategy stated that they repeated the words back to themselves (silently or out-loud) as closely to the stimulus as possible.

Strategy	Number of responses
1. Enunciation/Emphasis	29
2. Duration	14
3. Learning	10
4. Expectation based on facial characteristics	7
5. Self-repetition	6
6. Process of elimination	4
7. Emotion/expression	2

Table 1. Frequency of occurrence for the various strategies mentioned in the exit interview. (See text for an explanation of the various strategy names.)

A few participants explained that they were able to figure out the correct pairings of voices and faces by the “process of elimination”, i.e., by determining which pairs co-occurred most often. While this strategy is feasible, it seems like an extraordinarily complicated task, since it requires the explicit recognition of each voice and face in the experiment, and the maintenance of a frequency table of cooccurrence. Assuming that using this strategy for one block was sufficient to carry performance during the other block, a participant would have to keep track of the frequency of occurrence of 16 different response pairs (there were 4 talkers used for all participants). Unfortunately, the design of this experiment does not permit the elimination of this response strategy as a possibility, but the low number of participants who reported using it and the relatively heavy computational load it would impose cast doubts upon its usefulness. Finally, two participants reported the use of emotional expression as a cue for cross-modal identity. This is interesting because it points to the use of another, unanticipated non-linguistic aspect of the speech signal.

The ability to perceive the identity of the source of acoustic events has been demonstrated in other domains besides speech. 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%). Furthermore, perceivers were able to recognize the sound of their own clapping with almost 50% accuracy.

More evidence of the ability of perceivers to identify source characteristics from acoustic stimuli comes from Li, Logan, and Pastore (1991), who asked participants to identify the gender of a person whom they heard walking. Remarkably, judgments of gender were well above chance. Furthermore, anthropomorphic differences (such as weight and height) between walkers were found to be highly correlated with judgments of gender, indicating that the acoustics generated by different body-types contained information that allowed the accurate perception of these attributes.

Both the Repp (1987) and Li et al. (1991) studies indicate that detailed 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 (Fellowes et al., 1997; Remez et al., 1997). 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. The current study has demonstrated that this information can be used in judgments of source variation *across sensory modalities*.

As pointed out by Vatikiotis-Bateson and his colleagues: "...the motor planning and execution associated with producing speech necessarily generates visual information as a by-product." (Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzepoulos, 1997, p. 221). As a consequence, it is entirely possible that any information of relevance in the acoustic signal is also carried, in some form, by the visual signal. Because kinematic and dynamic sources of information about speech articulators are inherently amodal, they are good candidates for the form in which this information can be transmitted. Studies using point-light displays (Rosenblum & Saldaña, 1996) and sinewave replicas of speech (Remez, Rubin, Berns, Pardo, & Lang, 1994) demonstrate that speech information can be perceived from highly impoverished stimulus patterns that isolate kinematic and dynamic information. The present investigation has extended previous findings that indexical information about the source of spoken events is carried in the time-varying information about the motion of the articulators. Such information is modality-neutral, and as such can be perceived and used to make accurate judgments of identity across sensory modalities.

References

- Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (in press). Speech perception without hearing. *Perception & Psychophysics*.
- Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory. *Perception & Psychophysics*, *61*(2), 206 - 219.
- 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.
- Dodd, B. E., & Campbell, R. (1987). *Hearing by eye: the psychology of lip-reading*. London; Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton and Co.
- Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics*, *59*(6), 839 - 849.
- 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., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, *17*(3), 816 - 828.
- Fowler, C. A., & Rosenblum, L. D. (1991). The perception of phonetic gestures. In M. S.-K. Ignatius G. Mattingly (Ed.), *Modularity and the motor theory of speech perception*. (pp. 33-59): Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.
- 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.
- Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 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*(1), 152-162.
- 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*(6), 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*(1), 34 - 42.
- 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*(3), 269 - 276.
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, *14*, 201 - 211.
- 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.
- Li, X., Logan, R. J., & Pastore, R. E. (1991). Perception of acoustic source characteristics: Walking sounds. *Journal of the Acoustical Society of America*, *90*(6), 3036 - 3049.
- Lieberman, A., & Mattingly, I. (1985). The motor theory revised. *Cognition*, *21*, 1 - 36.
- Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *15*(4), 676-684.
- Massaro, D. W., & Cohen, M. M. (1995). Perceiving talking faces. *Current Directions in Psychological Science*, *4*(4), 104-109.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746 - 748.
- Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355 - 376.
- Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42-46.
- Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics*, *57*, 989 - 1001.
- Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309 - 328.
- 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.

- 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(5), 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(1), 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(4), 1100 - 1109.
- 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 & Performance*, 22(2), 318 - 331.
- Rosenblum, L. D., Yakel, D. A., Baseer, N., & Panchal, A. (1999). Visual speech information for face recognition. Manuscript submitted for publication.
- 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.
- 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. (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.