

**RESEARCH ON SPOKEN LANGUAGE PROCESSING**  
Progress Report No. 21 (1996-1997)  
*Indiana University*

**Static vs. Dynamic Faces as Retrieval Cues in  
Recognition of Spoken Words<sup>1</sup>**

**Lorin Lachs**

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

---

<sup>1</sup> This research was supported by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University.

## Static vs. Dynamic Faces as Retrieval Cues in Recognition of Spoken Words

**Abstract.** Three experiments examined the integration of auditory and visual information in memory for spoken words. Across experiments, recognition of isolated words was tested in the context of studied or non-studied faces and voices. The degree to which faces were informative about the studied speech event was manipulated between experiments. In Experiment 1, faces were static pictures experimentally paired to the voice speaking the word. In Experiment 2, a control for Experiment 1, faces were presented upside down. In Experiment 3, faces were dynamic video clips of talkers articulating. Subjects were either instructed that faces were to be recognized explicitly or not. The results show that static faces can only be used as effective retrieval cues to the recognition of words when experimental conditions encourage the association between visual and auditory information. By contrast, dynamic, articulating faces are automatically encoded along with voice and word information and improve recognition performance substantially. In addition, the encoding of faces for explicit recognition interferes with the utility of dynamic faces in recognizing speech. The results are taken to imply that visual information is encoded in cross-modally integrated memory representations for speech, but only when it is informative about the speech event.

### Introduction

Although much of the research on speech perception and spoken word recognition conducted in the past has regarded speech as a purely auditory phenomenon, a growing body of evidence suggests that the visual sensory modality plays an important role in the perception of speech and the understanding of spoken language. Perhaps the most well-known phenomenon relevant to this area of inquiry is the so-called “McGurk effect”, discovered by McGurk and MacDonald (1976). They found that the perception of speech can be altered by simultaneously presenting conflicting information in the auditory and visual modalities. Specifically, they reported that the visual presentation of a talker’s face repeatedly articulating the syllable [ga], along with the dubbed audio presentation of the same talker speaking the syllable [ba], induced in subjects the perception of the syllable [da] (McGurk & MacDonald, 1976). This effect is very robust; 98% of adults who viewed stimuli of this nature reported a “fused” percept (i.e., the perception of the syllable [da] given an auditory [ba] and a visual [ga]). Furthermore, McGurk and MacDonald (1976) reported that the effect did not habituate over time; despite long experience with and full knowledge of the nature of the stimuli, the authors themselves continued to experience the effect.

Since its discovery, a great deal of research has been conducted on the nature of the McGurk effect. In study after study, the effect has been replicated, using various experimental manipulations (MacDonald & McGurk, 1978; Massaro & Cohen, 1983; Massaro & Cohen, 1990; Rosenblum & Saldaña, 1992; Summerfield & McGrath, 1984). For instance, Summerfield and McGrath (1984) found that the effect is not necessarily confined to the perception of consonants. By presenting simultaneous conflicting information in the auditory and visual modalities, Summerfield and McGrath (1984) reported that subjects perceived vowels which were combinations of the vowels specified by either modality alone. Additionally, Dekle, Fowler & Funnell (1992) found that audiovisual integration can occur in the perception of real words. In another study, Massaro (1987) found that explicitly instructing subjects to ignore one of the sensory dimensions does not eliminate the effect it has on the perception of the other dimension. Indeed, the research into this phenomenon has been extensive. Among the copious amount of research into the McGurk effect is research which shows, for example, that the asynchronous presentation of the information in the two modalities still elicits an effect (Munhall, Gribble, Sacco, & Ward, 1996) and that, for certain syllables, the effect can be evoked with inverted articulating faces (Campbell, 1994; Massaro & Cohen, 1996).

Taken together, the research findings on the McGurk effect have produced compelling evidence that the influence of the visual sensory modality on speech perception is substantial and worthy of study. Indeed, Summerfield (1987) lists the McGurk effect as one of five major phenomena that must be accounted for by any theory of speech perception. Still, doubts as to the significance of the McGurk effect may be raised, due to its extremely artificial nature. In the real world, one is hardly, if ever, confronted with a situation in which one must perceive speech in the presence of conflicting, non-degraded information from both the auditory and visual modalities. Fortunately, more naturalistic evidence of vision's role in the process of speech perception comes from a groundbreaking study which pre-dates McGurk and McDonald (1976) by over twenty years and provides the foundation upon which all audiovisual theories of speech perception may stand.

In order to assess the contribution of visual information to the process of speech perception, Sumbly and Pollack (1954) had two groups of subjects seated around a talker. Each listener wore a pair of headphones and sat rather close to the talker. Half of the listeners, however, could see the talker, while the other half, turned away, could not. The words spoken during all trials were mixed with white noise at varying signal-to-noise ratios. After each stimulus, subjects made a forced-choice response from a list of words whose length varied between subjects.

Sumbly and Pollack (1954) observed several important relations. First, they found that the intelligibility of words spoken in noise is negatively affected by the size of the possible message set. This replicated the well-known result that as the number of possible words increases, the susceptibility of those words to interference by noise increases (Miller, Heise, & Lichten, 1951).

Second, Sumbly and Pollack found that the average intelligibility of *audiovisually* presented words from different vocabulary sizes, when plotted as a function of speech-to-noise ratio, were substantially higher than the unimodal scores. For example, under the most degraded condition tested (-30 dB S/N), the average intelligibility of the eight word list went from about 15% correct in the unimodal condition to around 90% correct in the multimodal condition. Put in other terms, the size of the gain in speech intelligibility as a result of multimodal presentation was roughly equal to the gain in intelligibility afforded by an increase in signal-to-noise ratio of +15 dB (Erber, 1969; MacLeod & Summerfield, 1987; Middleweerd & Plomp, 1987; Rosenblum & Saldaña, 1996; Sumbly & Pollack, 1954).

Interestingly, this remarkable gain in accuracy as a result of multisensory presentation was found to interact with the size of the possible message set. Greater advantages were found for small vocabulary sizes under very degraded conditions. Furthermore, this interaction was not simply a result of the fact that, probabilistically, smaller set sizes lead to higher accuracies, since, when the results were plotted in terms of the percent information gained (a measure which normalizes for stimulus set size), the effect was still observed. Sumbly and Pollack found, however, that for higher signal-to-noise ratios the visual contribution to intelligibility was mainly detected with longer vocabulary lists, whereas little if any advantage was gained for smaller vocabulary lists. Although this may seem puzzling at first, the results (as Sumbly and Pollack (1954) point out) appear to be due mainly to the fact that at the higher signal-to-noise ratios (i.e., under less degraded conditions) unimodal performance for small vocabulary sizes is already near ceiling levels, thereby leaving little room for improvement as a result of additional visual information.

In light of this "ceiling effect," Sumbly and Pollack (1954) proposed that the measure of visual information's contribution to speech intelligibility should be scaled in terms of its *possible* contribution. Remarkably, when they re-analyzed the data, they found that this measure, which they called "R" (i.e., the ratio of the actual contribution of visual information to its possible contribution), remained constant across a wide range of signal-to-noise ratios. In other words, the relative contribution of visual information was found to be independent of the signal-to-noise ratio under test. This finding can be interpreted to mean that the utility of the information provided by vision is not simply "additional" to that provided by audition, but is instead somehow intrinsic to the process of speech perception itself:

degrading the auditory signal with noise only serves to tease apart the underlying contributions of both input modalities to the perceptual process.

In another intelligibility experiment, Erber (1969) replicated the initial results of Sumbly and Pollack (1954) and extended them in several ways. Erber (1969) reported that while the threshold for performance above chance in auditory-only conditions lies somewhere around -18 dB S/N, performance in audiovisual conditions begins to increase above baseline lip-reading at a much lower signal-to-noise ratio, implying that visual information can work in tandem with auditory information in the perception of speech, even when the information in the auditory signal alone would be unusable to the listener.

Taken together, these early studies have demonstrated that the information which is specified by the visual modality can have a major influence on speech perception and spoken word recognition. Subsequent studies have attempted to determine the exact nature of this influence. Many of these studies have concentrated on the fact that the aspects of acoustically transmitted speech which are most confusable in noise are precisely those aspects of the message that are reliably transmitted visually (Massaro, 1987; Summerfield, 1987). For example, while the phonemes /k/ and /p/ are confusable at signal-to-noise ratios of +12 dB and below when presented acoustically (i.e., at a very low level of noise), cluster analysis of confusions made by subjects when identifying visual only speech reveals that these particular visual phonemes ("visemes") are among the last consonants to be confused when presented visually (Summerfield, 1987). The exact process by which these stimulus properties are exploited and integrated is a topic of some debate; Summerfield (1987) takes these properties as support for the assertion that information from both modalities is integrated before categorization, while Massaro (1987) claims that these confusions point to evidence that sub-phonetic categorical judgments are made on modality specific inputs which are then integrated *after* perceptual analysis.

Since both of these possibilities seem equally likely from the viewpoint of perceptual studies, it has become necessary to examine the nature of multimodal speech representations in memory, since insight into the way in which information from disparate modalities is encoded may provide important new clues about the underlying process by which this information is obtained from sensory input.

In a recent study of the effects of multimodal input on memory, Pisoni, Saldaña, and Sheffert (1995) showed an influence of audiovisual encoding on two aspects of memory: immediate memory span and serial recall. In the immediate memory span experiment, Pisoni et al. (1995) found that the number of items which could be recalled correctly was significantly shorter when stimulus items were presented audio-visually, compared with audio-only presentations. This result was taken to imply that visual information usurps processing resources available to the limited capacity working memory system. In the serial memory experiment, Pisoni et al. found that items in the primacy portion of the list were recalled more accurately when they were originally presented audio-visually than when they were originally presented audio-only. Since performance on items in the primacy portion of serial recall lists is usually taken to reflect the degree to which those items have been rehearsed, encoded and transferred to long term memory, Pisoni et al. concluded that the additional visual information must also be encoded into long term memory along with the auditory information and can be used as an effective cue at the time of retrieval.

Taken together, these recent findings suggest that visual information about a speech event is processed and encoded in some way relating to the phonetic information provided by the same event. These findings do not, however, provide an answer to whether the information from the two modalities is stored in a multimodal, integrated form or in separate, but linked, unimodal representations appropriate to each input modality.

One study designed to provide insight into this question was conducted recently by Sheffert and Fowler (1995). Using a continuous recognition task, Sheffert and Fowler (1995) examined whether words could be more accurately recognized at test when presented along with the studied video information

about the talker. Sheffert and Fowler found that, while presentation of the word at test using the same voice always facilitated recognition of the word, little or no advantage was gained as a result of repeated video contexts. Several other studies were conducted to assess the degree to which visual information was encoded but all of these revealed the same basic pattern of results: repeated voices facilitated the recognition of words, while repeated faces did not. Sheffert and Fowler therefore concluded that voices have a “privileged” status in the mnemonic encoding of words, and that faces may play a more contextual role.

In another effort to examine the question of multimodal integration in memory, Kato, Kanzaki, Tohkura, and Akamatsu (1995) examined the recognition of spoken sentences presented in one modality given changes in the stimulus presented in the other modality. Their study was motivated by the earlier findings of Legge, Grosman, and Pieper (1984), who showed that presenting a static picture of a face during the study interval for a particular voice increased the probability that that voice would be recognized later. However, the Legge et al. (1984) study did not examine whether voices could aid in the recognition of faces, a question of critical importance to the debate over integrated multimodal encoding. If it were found that faces can aid in the recognition of voices (as demonstrated by Legge et al.) but not vice versa, then this would be strong evidence against the notion that information from the two modalities is stored in an integrated representation.

Kato et al. (1995) therefore designed a recognition memory experiment to examine this issue. During the study phase, subjects were presented with a static picture of a face and a concurrent recording of a voice speaking a sentence. Throughout the study phase, the specific face+voice pairings remained constant, although the pairing was assigned randomly between subjects (i.e., Face A was not necessarily presented along with Voice A for all subjects); each face+voice pairing was seen six times (two sentences were presented three times). During the test phase, subjects were asked whether the face (or voice) that was presented was a face that they had already seen (or, alternatively, whether the voice was a voice which they had already heard). The item in the test modality (i.e., the face or the voice, depending on whether the task was to recognize the face or the voice, respectively) was presented in one of four “other-mode contexts.” So, for example, if the test mode was the face, then an *old* (i.e., studied) face could appear with either the voice with which it had been presented at study, a different voice that had been paired with another face during study, a different voice which the subjects had not yet heard, or no concurrent voice information. Similarly, a *new* face could appear with either a studied voice, a new voice which they had never heard, or no concurrent voice information. This experimental design was employed to assess whether a face or voice could be used as a facilitatory cue for retrieval of the other mode.

Several measures of performance were obtained. First, the hit rate (the rate of correctly identifying the test mode as “old” when given a studied, “old” item) was examined. Kato et al. (1995) found that while there was no difference in accuracy between correctly recognizing voices and faces, there was a significant effect of the other-mode context on recognition accuracy. That is, subjects were equally able to recognize faces and voices, but this ability was affected by the simultaneously presented other-mode context, due to decreased performance when the other-mode was *not* studied with the test item, relative to when no other-mode context was available (Kato et al., 1995).

Consider the following example as an illustration of the above finding: say that during the study phase for a particular subject, Face A was always presented with Voice A and Face B was always presented with Voice B. If this subject’s performance was consistent with the results of Kato et al. (1995), then his recognition accuracy for Face A would be worse in the context of Voice B than when there was no voice at all. However, the recognition of Face A in the context of Voice A or Voice B would not be different. In other words, recognition of a face was best when no voice context was given; if a voice context *was* given, then whether or not that voice context was the one with which the face had been originally studied did not significantly impact recognition performance. This pattern of results suggests that static faces and voices are not stored integrally, but instead that voices can interfere with the recognition of faces.

In a second analysis, Kato et al. (1995) examined the correct rejection rate (the rate of correctly identifying the test mode as “new” when given a non-studied item). Here they found that the rejection of a new *face* was not affected by the other-mode context, while the rejection of a *voice* was facilitated in the presence of new faces (Kato et al., 1995).

Kato et al. (1995) concluded that the pattern of their results suggests that face and voice information are encoded independently in memory, since recognition of a test item was not better in the context of its correct other mode item, relative to performance in the context of incorrect or novel other mode items (i.e., there was no facilitation).

There are, however, several weaknesses in the methodology and design of this experiment that call into question their conclusion. First of all, the study used an explicit recognition memory procedure to assess the encoding of face and voice in memory. That is, subjects were asked explicitly on each presentation to determine whether or not the face or the voice was a component attribute of the stimulus item that was presented during the study phase. This task has no reference to the other mode item; in other words, there is no reason for the subject to encode both modalities in an integrated representation, since their only reference to each other is simply an arbitrary co-occurrence or association during the study phase.

Similarly, because the pairing of faces and voices was randomly assigned as a between subjects variable, there was no reason, other than the arbitrary experimental ones, that a subject would encode the face and voice together in an integrated representation in memory.

Finally, and perhaps most importantly, the use of static pictures of faces eliminates any naturally occurring dynamic optical information that links faces and voices together. That is, by eliminating dynamic, articulatory information from the optical display, all aspects of the relationship between the face and the voice speaking were eliminated from the stimulus, once again leaving only experimental conditions to signify the importance of the intermodal pairing.

The present series of experiments was motivated by the Kato et al. (1995) experiment and was designed to deal with the criticisms mentioned above by using dynamic visual displays and examining both implicit and explicit memory processes.

## Experiment 1

In Experiment 1, the task was changed from an explicit recognition memory test of faces or voices to an implicit one; on each trial during the test phase, subjects were required to recognize whether the *word* they heard was an item that was presented earlier during the study phase. Faces and voices were manipulated but their effects were never expressed explicitly. It should also be noted that while sentences acted as the carrier for voice information in the Kato et al. (1995) study, the present study utilized isolated words as both the carrier of voice information *and* the test items used for recognition.

The test items could be presented in one of four different conditions. In the (**F+V+**) condition, the face and voice which were presented with the word during the study phase were also presented during the test phase. In the (**F+v-**) condition, the face was the same as during study, but the voice was not. In the (**f-V+**) condition, the voice was the same as during study, but the face was not. Finally, in the (**f-v-**) condition, the face and voice were both different from those presented with the word during study. The prediction was that these experimental manipulations would allow a more sensitive measure of the encoding of face and voice in memory; in addition, these manipulations allowed the effects of face and voice on recognition memory to be examined independently.

The main purpose of Experiment 1 was to serve as a control by replicating the findings of Kato et al. (1995) using a new implicit memory task and different stimuli. A replication of the major results was

expected because, in this experiment, there was still no reason other than experimental necessity for a subject to associate a particular face with a particular voice: the faces were still static and semi-randomly assigned to one another at the time of study<sup>2</sup>.

Because the results of Kato et al. (1995) are not directly interpretable within the implicit memory paradigm, we should be clear about what was expected. Using subjects' explicit judgments on face and voice recognition, Kato et al. (1995) found some evidence that intermodal relationships could exist in memory (e.g., better rejection of new voices in the presence of new faces, and inhibitory effects on recognition due to incongruent other-mode stimuli) and interpreted the absence of facilitatory effects to mean that these intermodal relationships existed as links between independent, unimodal memory representations. Our task, on the other hand, used subjects' recognition of *words* to examine their knowledge of face+voice pairings. Still, the change in task should not affect the basic finding reported in Kato et al. (1995) that intermodal relationships can indeed exist in memory. As such, we expected that Experiment 1, if it were a valid extension of the findings in Kato et al. (1995) should indeed show an effect of modality context on recognition of words.

Before work could begin on this experiment, it was important to make sure that subjects were actually encoding and using the visual information during study and test. It is possible that subjects might just as well have closed their eyes and performed the task on the basis of word and voice information alone. As a result, our measures might be rendered useless. In order to control this situation, two additional manipulations were introduced into the experimental design. First, at the end of the test phase, subjects were presented with a test of explicit face recognition. Performance on this additional task was used as a criterion for inclusion into the final data analysis. Second, we also manipulated the encoding instructions. For half of the subjects, advance notice of the explicit face recognition test at the end of the session was given directly in the instructions for the experiment (the "explicit instructional condition"). For the other half of the subjects, no advance warning was given, and the explicit face recognition test came as a surprise (the "on the fly instructional condition"). We assumed that if there were no differences across the two instructional conditions between the performance of subjects meeting some criterial level of accuracy on the explicit face recognition test, then it could safely be said that the subjects had followed the instructions as intended and had watched the visual display during study and test.

## Method

### Subjects

Subjects were 80 Indiana University undergraduates who participated in partial fulfillment of course requirements for Introductory Psychology. All subjects were native speakers of English, had normal hearing, and reported no history of speech or hearing disorders at the time of testing.

### Stimulus Materials

*STUDY AND TEST FOR WORD RECOGNITION:* The stimuli consisted of 72 digitized movies of 8 talkers speaking isolated words. All of the items were taken from the Hoosier Multimodal Database (Sheffert et al., 1997). The average intelligibility of each word was 100% for all talkers used in the study, as indicated by the intelligibility data which accompanies the tokens in the HMD. All together, then, the total number of stimuli used in the experiment was  $72 \times 8 = 576$ . However, for each subject, only 72 tokens were ultimately viewed.

In order to provide the static picture of a talker's face for presentation, one movie from the set of 72 for each talker was selected as the "representative movie". The representative movie for each talker was a movie in which the second frame of the digitized sequence contained a picture of the talker with

<sup>2</sup> The assignment of faces to voices was semi-random due to the fact that, in our stimulus database, half of the talkers are male and half of them are female. Since all of the talkers in Kato et al. (1995) were male, throwing in the mixing of male voices with female faces and vice versa could be a potential confound in the replication of the original study. As such, only male voices were assigned to male faces, and only female voices were assigned to female faces. Within these limitations, however, the assignment of a particular face to a particular voice was randomized between subjects.

his/her mouth closed and which was not blurry. Thus, the second frame of each talker's representative movie was used as the static "face" stimulus for a particular talker for all subjects in Experiment 1.

Voice/word information was taken from the audio tracks of the 576 movies described above. For each subject, then, a semi-random pairing of faces and voices (preserving sex) was assigned prior to the commencement of the session. Then, for each subject, a randomly selected subset of half of the 72 words were taken to be "study" words. For each of these 36 words, one face-voice pairing was randomly assigned as the study presentation context.

For the test phase, the 36 "old" words were randomly assigned to one of four conditions. If a word was assigned to the (F+V+) condition, then it would be presented during test along with the same face-voice pairing with which it had been presented during study. If a word was assigned to the (F+v-) condition, then the face presented during study would be the same as the one with which the word was presented during study, but the voice would be a different, randomly assigned one. If a word was assigned to the (f-V+) condition, then the face would be a different randomly assigned one, but the voice would be the same as during study. Finally, if the word was assigned to the (f-v-) condition, then it would be presented with a totally new, randomly assigned face-voice pairing. However, this face-voice pairing was a pairing which had been seen consistently during study, just not with the specific word in question.

The 36 old words were then randomly mixed with the 36 new words to form the list of 72 test items. The 36 new words were randomly assigned to one of two possible conditions: one condition in which the randomly assigned face+voice pairing was one which had been a valid face+voice pairing during study, and one condition in which the face-voice pairing was not a valid one during study. Thus, the assignment of study words, face-voice pairings, the face-voice pairs which spoke each word and the conditions under which all words were tested were all randomly chosen for each subject.

*TEST OF EXPLICIT FACE RECOGNITION:* Stimuli for the explicit face recognition portion of this experiment were taken from another digitized set of video clips used as stimuli in a previous study (see Pisoni et al., 1995). Four of the talkers in the HMD were also talkers in this new database; "old" items in the explicit face recognition test were still frames taken from video clips of the four talkers who served as talkers in both databases. "New" items were still frames of four other video clips from the Pisoni et al. (1995) database. Thus, all the stimuli for the explicit face recognition test were drawn from the same set of digitized images. This set of eight explicit face stimuli were used for all subjects in all of the experiments reported here.

### **Apparatus**

All stimuli were stored on an Pinnacle Micro 4.6 GB optical disk. A control program running on a Macintosh PowerPC 8100/100 assigned all random variables and presented stimuli according to this assignment. Visual stimuli were presented on a 17" Apple Multiple Scan 17 Display color monitor, controlled by a Radius video board, while auditory stimuli were presented over BeyerDynamic DT100 headphones calibrated to 74 dB SPL.

### **Procedure**

Every other subject was assigned to one of the two instructional manipulations. For subjects in the "explicit" instructional condition, instructions were given which read as follows:

This experiment will consist of three parts.

- The first part is called the study phase. On each trial during this part, you will be shown a picture of a face on the computer monitor. At the same time, you will hear a word spoken over your earphones. Do your best to pay attention to both the word that is spoken and the face which is being shown, because you will be tested on them later.

- The second part is called the word recognition phase. On each trial, you will be shown a picture of a face and you will hear a word being spoken. After you hear the word, you should determine whether you have heard this word already in the study phase. You will then be required to indicate your response on the button box in front of you. Press the OLD button if you think you have already heard this word before. Press the NEW button if you think you have not. You should try to do this part as quickly as possible without sacrificing accuracy. Remember, you are judging whether the word you heard is “OLD” or “NEW.”
- The last part will be a test on the faces you’ve seen. On each trial, you will see a picture of a face. Some of these will be faces you already saw in the previous phases, and some will be new. On each trial, your task is to determine whether you have already seen the face and indicate your response on the button box in front of you. You should press the OLD button if you think you have already seen this face. Similarly, you should press the NEW button if you think you have not seen the face before in the experiment. You should try to do this part as quickly as possible without sacrificing accuracy.

Subjects in the “on the fly” instructional condition, read the following instructions: This experiment will consist of three parts. Instructions for each part will be posted on the computer screen before the beginning of each phase.

All subjects were given on-screen instructions before the commencement of each sub-section of the procedure.

During the test phase of the experiment, “old” or “new” responses for either words or faces, as the case may be, were collected by means of a button-box attached to a Strawberry Tree card. Responses, along with the parameters for a given trial, were recorded in data files for later analysis.

## Results

In order for a subject to be included in the final data analysis, a score of 87.5% or better on the explicit face recognition test was necessary. The data from 36 subjects were eliminated in this way. In all, 23 subjects for each instructional condition were included in the final analysis.

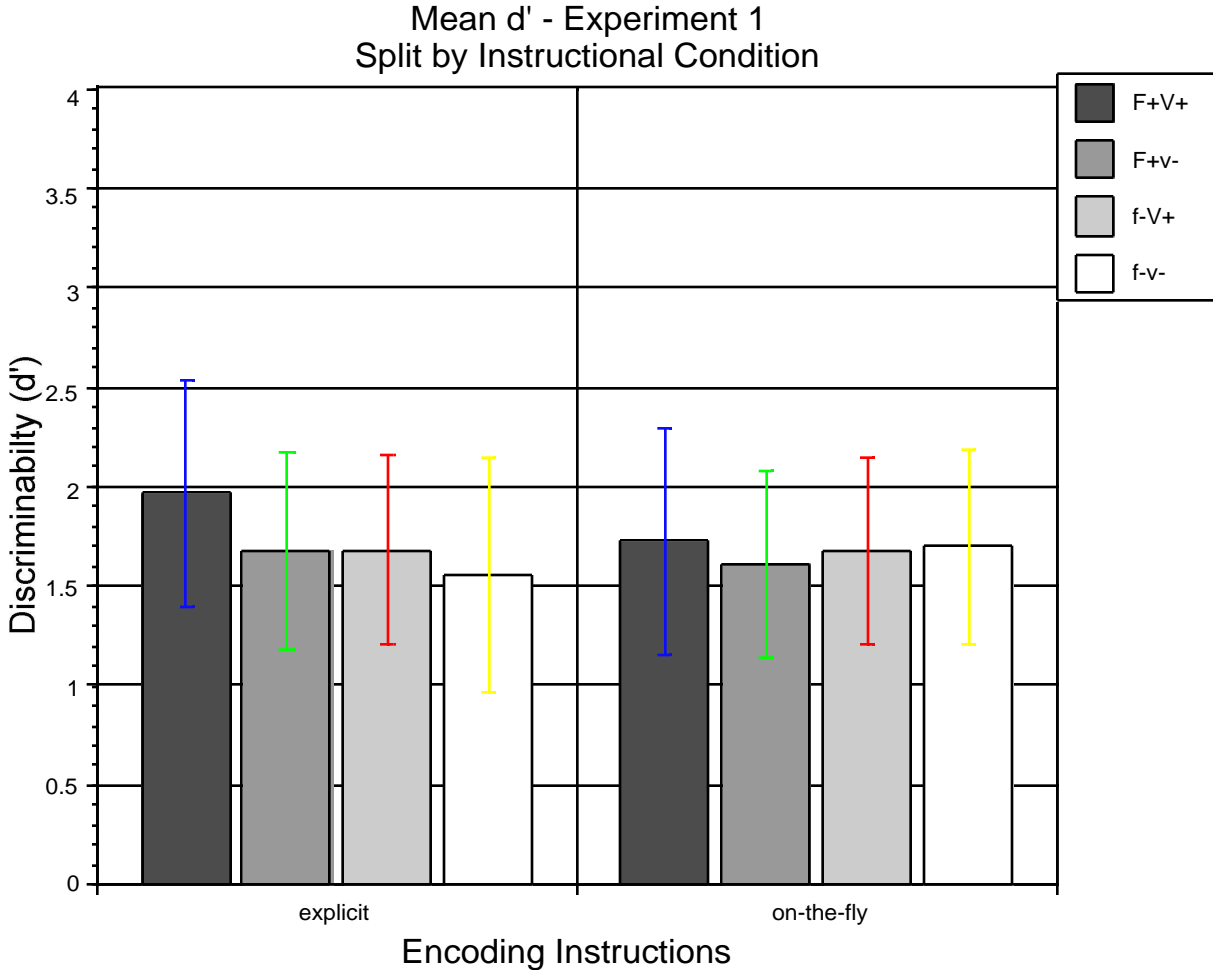
Figure 1 shows the average  $d'$  scores for recognition of *words* in each testing condition for Experiment 1, as a function of Instructional Condition. This measure was chosen to assess the discriminability of old and new items in memory, consistent with previous approaches to the study of recognition memory (Banks, 1970; Egan, 1958; Lockhart & Murdock, 1970; Parks, 1966; Pollack, 1959; Shepard, 1967). The  $d'$  measure is advantageous with respect to measures of performance such as percent correct because it takes into account all the data collected for a given session. While accuracy measures only present data for responses to “old” items,  $d'$  is calculated using responses to both “old” and “new” items. Table 1 shows the corresponding Hits and False Alarm rates used in the calculation of  $d'$  scores for the various experimental conditions.

-----  
 Insert Figure 1 about here.  
 -----

It should be noted that responses to new word items cannot be split into the same number of conditions as responses to old word items. While old word items could occur in one of four contexts (F+V+, F+v-, f-V+, and f-v-), new words could only occur in one of two contexts (studied face+voice pairing, or non-studied face+voice pairing). The calculation of  $d'$  scores was carried out such that the False Alarm rate used for the F+V+ and f-v- conditions was the rate at which, in the context of a

*previously studied* face+voice pairing, subjects responded “old” to new words. Similarly, the False Alarm rate used for the calculation of  $d'$  scores in the F+v- and f-V+ conditions was the rate at which subjects responded “old” to new words presented in the context of face+voice pairings which had *not* been previously studied.

In other words, two different False Alarm rates were used in the calculation of  $d'$  scores. One False Alarm rate was used for the calculation of  $d'$  scores for stimuli which were presented with *experimentally valid* face+voice pairings. The other False Alarm rate was used for the calculation of  $d'$  scores for stimuli which were presented in face+voice pairings which *had not been studied*.



**Figure 1:** Average d' scores for all four conditions in Experiment 1 as a function of instructional condition.

**Table 1****Average Hits and False Alarms for all Conditions in Experiment 1,  
Split by Instructional Condition.**

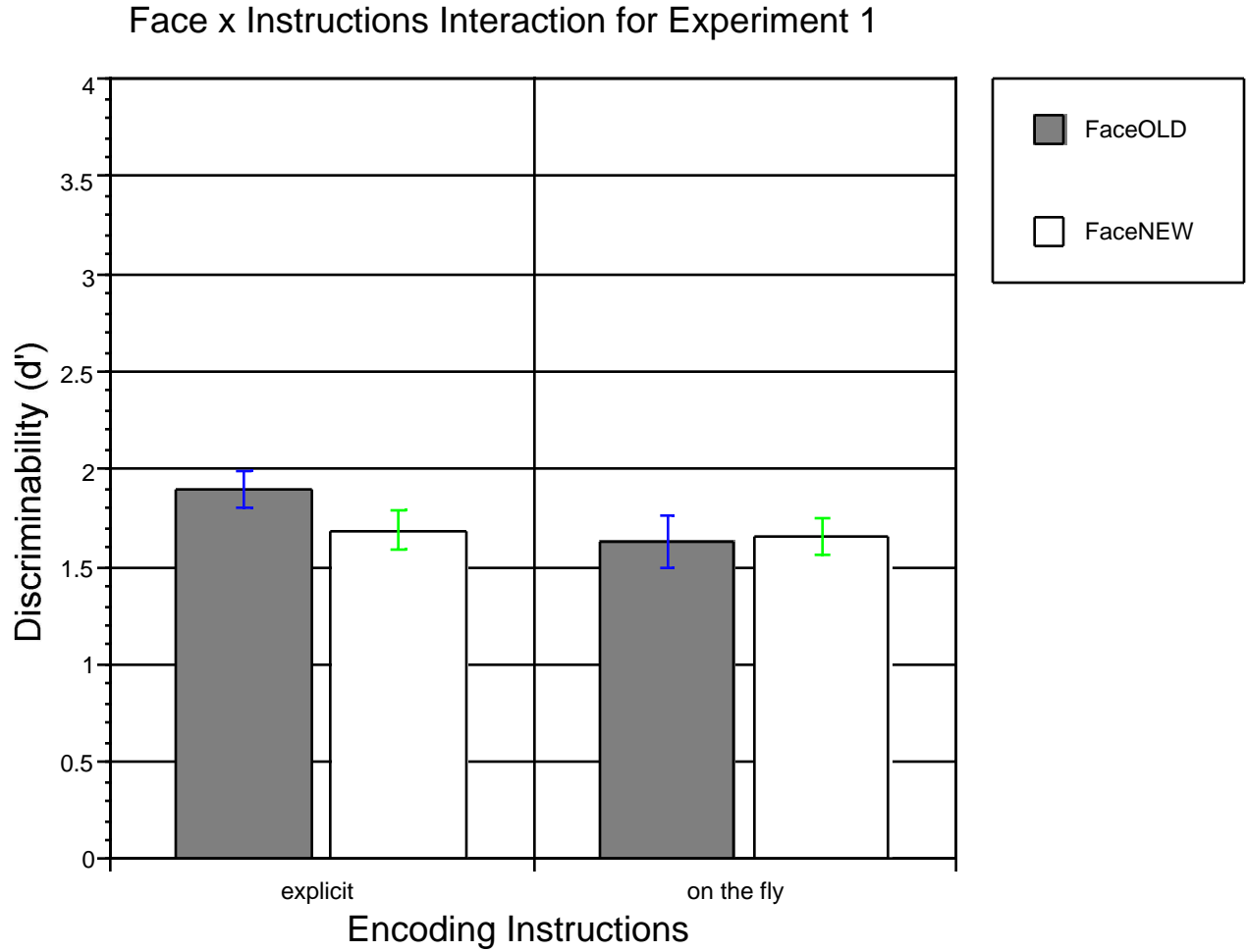
explicit face+voice	Encoding			Instructions			
	average FAs	condition	average Hits	on the fly face+voice	average FAs	condition	average Hits
non-studied	3.30	F+V+	8.00	non-studied	4.30	F+V+	6.87
		f-v-	6.35			f-v-	7.35
studied	3.70	F+v-	7.22	studied	3.61	F+v-	6.57
		f-V+	6.35			f-V+	6.61

The left panel of Figure 1 shows the average  $d'$  scores from subjects in the “explicit” instructional condition for the four experimental conditions, (F+V+), (F+v-), (f-V+) and (f-v-). The right panel of Figure 1 shows the analogous scores taken from subjects in the “on the fly” instructional condition. As can be seen from the graph of “on the fly” scores, there were virtually no differences between discriminability scores in each of the experimental conditions. However, the left panel reveals a notable difference between discriminability in the (F+V+) condition and the rest of the conditions. Similarly, there seems to be at least a numerical advantage to discriminability in the conditions (F+v-) and (f-V+) relative to the (f-v-) condition.

A repeated measures ANOVA was conducted on the  $d'$  scores for Experiment 1 with Face Context (new or old) and Voice Context (new or old) as the repeated measures and Instructional Condition (“explicit” or “on the fly”) as a between subjects variable. This analysis revealed a significant effect of Voice Context,  $F(1,44) = 3.686$ ,  $p = 0.061$ . Across instructional conditions, subjects were better able to distinguish old from new words when the old words were presented in the context of the voice with which that word was originally studied.

In addition, the interaction between Face Context and Instructional Condition approached significance,  $F(1,44) = 2.849$ ,  $p = 0.099$ . Figure 2 displays the Face Context x Instructional Condition interaction for  $d'$  scores in Experiment 1. The left panel of Figure 2 shows the discriminability scores for subjects in the “explicit” instructional condition; the left side shows scores for subjects in the “on the fly” instructional condition. Each bar represents the average discriminability of items, collapsed across voice condition. Thus, the “FaceOLD” bar on both sides of the figure represents the average discriminability for items in the (F+V+) and (F+v-) conditions. Similarly, the “FaceNEW” bar for each panel of the figure represents the average discriminability for items in the (f-V+) and (f-v-) conditions. A probe of this interaction through a 2 (Face Context) x 2 (Voice Context) repeated measures ANOVA split by Instructional Condition revealed that this effect was due to a significant main effect of Face Context in the “explicit” instructional condition  $F(1,22) = 4.831$ ,  $p = 0.039$ , but not in the “on the fly” instructional condition,  $F(1,22) = 0.048$ , n.s. Thus, for subjects in the “explicit” instructional condition only, the distinction between old and new items was significantly greater when those items were presented in the context of the face with which they were originally presented.

-----  
 Insert Figure 2 about here.  
 -----



**Figure 2:** Average  $d'$  scores for test conditions in which the face was old or new as a function of instructional condition. The FaceOLD bar represents the average  $d'$  scores for test items in the (F+V+) and (F+v-) conditions. The FaceNEW bar represents the average  $d'$  scores for test items in the (f-V+) and (f-v-) conditions.

## Discussion

The results from Experiment 1 replicated and extended the major findings of Kato et al. (1995) in accordance with our original predictions. We observed an effect of modality context on the recognition of words. Old words presented in the context of an old voice were more discriminable from new words than old words presented in a different voice context. This effect replicates the voice repetition effect found in previous studies (Goldinger, 1995; Palmeri, Goldinger, & Pisoni, 1991) and establishes the validity of the current experimental procedures.

Furthermore, a significant effect of Face Context on the recognition of words was found in the “explicit” instructional condition. In retrospect, it seems likely that the “explicit” instructional condition may represent a task which is more closely tied to the procedures used by Kato et al. (1995) than originally expected, because it may result in an encoding strategy for face information which allows for the “explicit” recall of face information at some later point. There is evidence to show that the encoding of information for “explicit” and implicit retrieval may be dissociated from one another (Schacter & Church, 1992) In light of these findings, it may be that the instructional manipulation was more fruitful than originally anticipated. To the extent that the instructions given in the “explicit” instructional condition result in a strategy of “explicit” encoding of face information *and* to the extent that the instructions given in the “on the fly” condition do not, it can be said that the instructional manipulation represents a direct test of one of our hypotheses. Namely, it tests the assertion that an “explicit” task will not foster the need for multimodal encoding since the demands of such a task do not require it. In other words, asking a subject to explicitly recall information from one modality may only invoke the information in another modality presented at study due to simple co-occurrence. By contrast, an implicit task requires encoding of an underlying event, and as such, may make use of all information which was relevant to the instantiation of the event being recognized.

In light of this interpretation of the instructional manipulation, an explanation for the Face Context x Instructional Condition interaction may be seen. We claim that conclusions about the integration of audiovisual information in memory may not be drawn when the visual information used to test this integration is static, because such displays eliminate any natural relationships which may exist between the visual and auditory information and thus obviate any perceptual processes which may serve to integrate multimodal inputs. However, we do not deny that such displays may serve as effective retrieval cues during recognition (as in Kato et al., 1995; Legge et al., 1984). Indeed, our hypothesis implies that it is the *relationship* between information in the various input modalities which effects their link in memory. Thus, if some experimentally induced manipulation lends credence to the relationship between information perceived through disparate sensory modalities, then that relationship should be encoded, and thereafter used during recognition processes. In other words, explicitly alerting subjects to the fact that a static face is related to an associated speech event - either through instructions for an explicit task, as in Kato et al. (1995) and Legge et al. (1984), or through instructions similar to our own - will result in a bond in memory between the information coming from either modality. The fact that static faces were used as an effective retrieval cue for recognition of words in the “explicit” instructional condition, but *not* in the “on the fly” condition, indicates that, under normal circumstances, there is no integral encoding of static visual information with simultaneously presented speech. With manipulation of specific task demands, though, the memory system may forge an arbitrary link where no natural one occurs.

In summary, Experiment 1 can still be considered an extension of Kato et al. (1995) because it shows that using single words spoken in isolation as the carrier of voice information (as opposed to sentences) and using an implicit memory paradigm (as opposed to an explicit one) does not change the basic finding that audiovisual relationships do exist in memory and can be constructed as a result of arbitrary pairings of spoken words with static faces (Kato et al., 1995; Legge, Grossmann, & Pieper, 1984); however, some sort of experimental manipulation must be present to force the bond if the visual information is static.

Experiment 1 therefore confirms that, through experimental manipulation, it is possible to induce in subjects a strategy for encoding events in recognition memory which in some way links co-occurring information from disparate modalities. The results, however, seem to imply that this cross-modal link in memory is due to purely arbitrary reasons: faces can only be used as effective retrieval cues for the recognition of speech events when subjects are explicitly instructed as to their relevance to the task at hand. It may be, then, that *any* arbitrary co-occurring visual information could be encoded in the same way, given the right conditions. Thus, these results do not necessarily add to our understanding of speech perception and spoken word recognition in audiovisual environments. Experiment 2 was designed to test whether the findings from Experiment 1 with static faces generalize to results which might be found with *any* visual stimulus. If so, then conclusions may not be drawn from static stimuli concerning the integration of face and voice information in representations of speech in memory.

## Experiment 2

The task for subjects in Experiment 2 was the same as in Experiment 1; once again, an explicit face recognition test was used as a criterion for entry into subsequent data analyses, and the between-subjects instructional condition was again used. However, in this experiment, the faces were rotated 180° and presented upside down. We reasoned that, because the pairing of voices and faces was arbitrary in the Kato et al. (1995) experiment, then *any* arbitrary pairing of any visual stimulus with voice information should produce essentially the same recognition memory effects.

Several earlier studies have shown that recognition of faces is impaired by upside down presentation (Valentine, 1988; Yin, 1969). By presenting the faces upside down, we controlled for visual complexity of the visual stimulus, while simultaneously reducing the cue value of the facial information. One caveat to this approach, however, is that upside down faces are unfamiliar to subjects.

The aim of this experiment was to show that, while experimentally derived conditions may well explore a link between unimodal representations, they may not be able to shed light on the matter of natural multimodal integration. Showing that static faces can be used as a retrieval cue for recognition of voices (Kato et al., 1995; Legge et al., 1984) may be no better than showing that *any* arbitrary concurrent other mode stimulus can be used as a retrieval cue, and, therefore, cannot provide new insights into the matter of audiovisual speech representations in memory.

## Method

### Subjects

Subjects were 40 Indiana University undergraduates who participated in this experiment as partial fulfillment of course requirements for Introductory Psychology. All subjects were native speakers of English, had normal hearing, and reported no history of speech or hearing disorders at the time of testing.

### Stimulus Materials

*STUDY AND TEST FOR WORD RECOGNITION:* The stimuli in Experiment 2 were identical to those used in Experiment 1 in all but one respect: in Experiment 2, the static face part of each face+voice pairing was rotated and presented upside down. This was accomplished by taking the “representative movie” (as described above) and passing it through a 180 degree rotation matrix in Adobe Premier v.4.0. The output rotated movie was then used as the “representative movie” for a particular talker across subjects.

*TEST OF EXPLICIT FACE RECOGNITION:* Stimuli for the explicit face recognition portion of this experiment were the same as in Experiment 1. This includes their orientation.

## Apparatus

Experiment 2 was carried out using the exact same apparatus as Experiment 1.

## Procedure

Procedures for Experiment 2 were analogous to those in Experiment 1, except that whenever it was necessary to mention the nature of the face information, subjects were told that the faces would be presented upside down on their video screens.

Once again, the assignment of study words, face-voice pairings, the face-voice pairs which spoke each word and the conditions under which all words were tested were all randomly assigned for each subject.

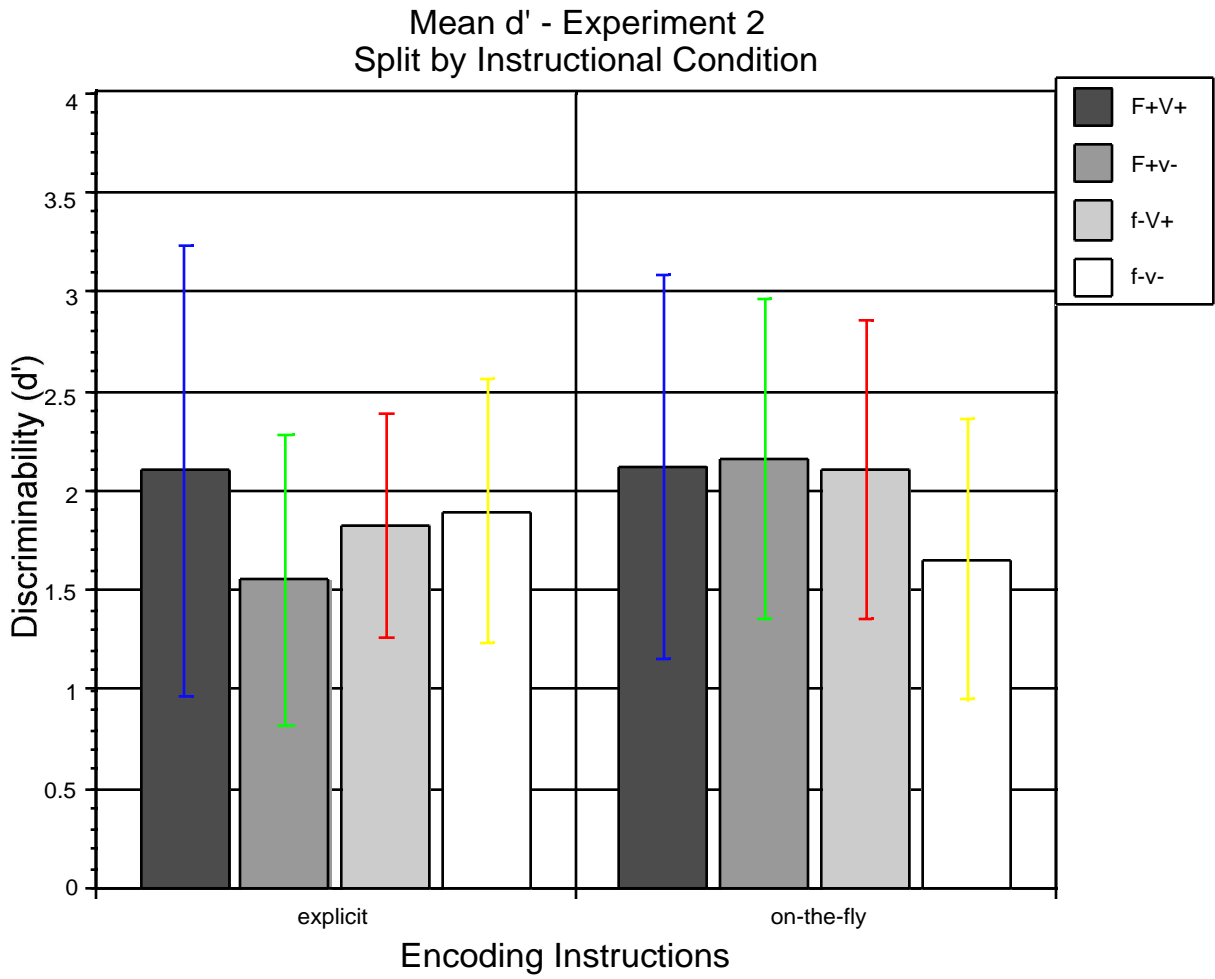
During the test phase of the experiment, “old” or “new” responses for either words or faces, as the case may be, were collected by a button-box interfaced with a Strawberry Tree card. Responses, along with the parameters for a given trial, were recorded in data files for later analysis.

## Results

Due to extremely poor performance on the explicit face recognition task, very few subjects were able to meet the criterion for entry into the analysis of 87.5% accuracy. The criterion was therefore relaxed to 75% accuracy so that statistical analyses could be carried out. Even with the weaker criterion, only 9 subjects from the “on the fly” instructional condition and 10 subjects from the “explicit” instructional condition could be included in the final analysis.

-----  
Insert Figure 3 about here.  
-----

Figure 3 shows the average  $d'$  scores in each testing condition for Experiment 2, separated as a function of Instructional Condition. Figure 3 follows the same format as Figure 1, with the left panel representing average discriminability scores in each of the experimental conditions (F+V+), (F+v-), (f-V+) and (f-v-) from subjects in the “explicit” instructional condition. The right panel of Figure 3 shows the scores from subjects in the “on the fly” instructional condition. Table 2 shows the corresponding Hits and False Alarm rates used in the calculation of  $d'$  scores for the various experimental conditions. As before, the calculation of  $d'$  scores was carried out such that the False Alarm rate used for the (F+V+) and (f-v-) conditions was the rate at which subjects responded “old” to new words presented in the context of a previously studied face+voice pairing. Similarly, the False Alarm rate used for the calculation of  $d'$  scores in the (F+v-) and (f-V+) conditions was the rate at which subjects responded “old” to new words presented in the context of face+voice pairings which had not previously been studied together.



**Figure 3:** Average  $d'$  scores for all four conditions in Experiment 2, as a function of instructional condition.

**Table 2****Average Hits and False Alarms for all Conditions in Experiment 2,  
Split by Instructional Condition.**

Encoding				Instructions			
explicit				on the fly			
face+voice	average FAs	condition	average Hits	face+voice	average FAs	condition	average Hits
non-studied	3.70	F+V+	7.60	non-studied	4.78	F+V+	5.89
		f-v-	7.60			f-v-	5.67
studied	4.20	F+v-	6.50	studied	5.22	F+v-	7.00
		f-V+	7.10			f-V+	6.67

A repeated measures ANOVA was conducted on the  $d'$  scores for Experiment 2 with Face Context (new or old) and Voice Context (new or old) as the repeated measures and Instructional Condition (“explicit” or “on the fly”) as a between subjects variable. The results of this analysis revealed no significant main effects or interactions.

**Discussion**

It seems at first glance as though Experiment 2 did not confirm the hypothesis it was designed to test. This experiment was designed to show that *any* visual stimulus which was arbitrarily paired with voice information could act as an effective cue for retrieval during recognition of a speech event, and therefore could not speak to the issue of audiovisually integrated representations in memory. In order to test this hypothesis, an upside down face was chosen as a suitable visual stimulus, since such a stimulus would, when compared to performance with static faces, control for the complexity of the visual image while simultaneously eliminating any cues to the “face-ness” of the stimulus (Valentine, 1988; Yin, 1969).

However, in retrospect, given the outcome of this study, it seems as though this particular visual stimulus may not have been an appropriate one; given the previous findings showing the difficulty subjects have with explicitly recognizing upside down faces, it is not surprising that performance on the explicit face recognition task (with upright faces) was so poor. This poor performance is potentially confounding in several respects.

First, despite the relaxation of the entry criterion, the data from very few subjects were actually included in the final analysis. As such, the low power associated with each statistical test may have acted to obscure any underlying trends.

Second, and perhaps more importantly, it may be that the explicit face recognition task did not test what it was designed to test in the context of Experiment 2. Originally, the explicit face recognition test was added to the design of the experiment so that we could have an objective measure of how well subjects followed the instructions to pay attention to both visual and auditory information during study. As such, we assumed that the measure should remain constant across experiments, so that parallels could be drawn between the data collected in each. However, it now can be seen that, in the context of Experiment 2, the explicit face recognition test also measured another variable - that is, it assessed how well subjects could transfer their knowledge of the upside down faces in the study and word test phases to rightside up faces. As stated above, such a transfer of knowledge has already been shown to be poor and was the very reason we selected upside down faces as visual stimuli in the first place.

There are two ways in which the design of Experiment 2 could be altered to eliminate this confound and simultaneously test the original hypothesis. First, the explicit face recognition phase could be altered such that the stimuli included in it are also upside down. This would eliminate the need for a transfer of knowledge about upside down faces to recognition of upright faces. In this way, the explicit recognition task would be more clearly comparable to that in Experiment 1.

However, it may be that the inappropriateness of the stimuli in Experiment 2 goes further than that. Subjects are, in general, extremely practiced at recognizing upright faces and have very little experience recognizing upside down ones. It therefore may be necessary to use a different type of visual stimulus for Experiment 2, such as animal faces, geometric shapes, houses, cars, or patches of color. Because these visual stimuli are usually well-learned in the experience of most subjects, their recognition may be more directly related to the recognition of static faces.

Still, the results from Experiments 1 and 2 make one thing perfectly clear: under normal circumstances, static visual displays, regardless of their content, are *not* integrated in memory with information about a simultaneously presented speech event. For the subjects in the “on the fly” instructional condition of both experiments, visual information presented concurrently was not used as an effective retrieval cue for recognition of a speech event. However, in Experiment 1, explicit instructions as to the importance of the visual stimulus did forge a connection in memory between visual and auditory information. These explicit instructions may well have forced a connection between visual and auditory information in Experiment 2 also, but, given subjects’ difficulty in encoding upside down faces, the encoding of upside down faces may have been insufficient to foster the use of such information as a retrieval cue. In any event, Experiments 1 and 2 show that arbitrary visual information may not serve as an effective retrieval cue for speech events in the absence of sufficient experimentally induced bias.

Can visual information act as an effective retrieval cue when it is *not* arbitrarily paired with the speech event? A dynamic visual display of a talker speaking contains information which is not arbitrarily tied to the underlying speech event. Indeed, the auditory specification of a speech event is *lawfully tied* to its articulation. Our hypothesis, therefore, was that subjects would be able to use a dynamic optic display as an effective retrieval cue during recognition of a speech event, since this type of display provides information concerning the speech event during encoding. In other words, we postulated that any transfer to memory that takes place during speech perception will *automatically* encode any information which is relevant to the event being encoded; since dynamic optical displays of articulation are informative about a speech event, then it follows that visual information will also be encoded and later used as an effective retrieval cue during recognition. As such, Experiment 3 was designed to test the integration of multimodal information in memory for speech using dynamic visual displays of talkers uttering the stimuli.

### Experiment 3

The task for subjects in Experiment 3 was the same as in the previous two experiments. However, this time, the visual stimuli were dynamic video clips of talkers articulating the words. As in the previous experiments, the explicit face recognition test and the instructional condition were used. We reasoned that because intermodal relationships in the sensory information are preserved in dynamic visual displays, the results from this experiment would provide evidence for the integration of audiovisual speech representations in memory. We expected that we would once again find evidence for the use of intermodal relationships in memory in the form of main effects of both face and voice. In addition, we hypothesized that overall levels of performance in Experiment 3 would be greater than in Experiment 1, due to increased ability of subjects to exploit the naturally occurring, lawful, intermodal relationships between faces and voices specifying speech events.

It is worth noting here that our post hoc interpretation of the instructional manipulation in Experiment 1 has ramifications for the outcome of Experiment 3, especially when considered in the light of our experimental hypothesis. Our post-hoc interpretation of the instructional manipulation assumes that

the “explicit” instructional manipulation causes a change in the strategies utilized by subjects in their encoding of face information, resulting in an experimentally induced arbitrary association between face and voice information in memory. Thus, in Experiment 1, static faces could only be used as effective retrieval cues during recognition of speech events when subjects were *explicitly* told that faces were important in the experiment and that they would be tested at the end of the experiment. Under normal circumstances, a static face would *not* be linked in the encoding of a speech event because its non-dynamic nature serves to eliminate any information which could specify the phonetic context. These hypotheses were supported by the findings of Experiment 1.

The prediction made by these hypotheses, however, is that if the visual display contained information specifying its relation to the acoustic signal (i.e., if the visual display provides dynamic information about the articulation of the speech being heard) then a link between the auditory and visual information should be forged *in the absence of explicit instructions to do so*. In other words, the integrated encoding of dynamic visual information in memory for speech events should be mandatory, since a dynamic optic display of a talker’s articulation will specify the nature of the linguistic message. As such, both face and voice context should show an effect on the recognition of words in the “on the fly” instructional condition, when only the information that is naturally encoded in a mandatory fashion will implicitly affect recognition performance.

The effects of dynamic visual information on recognition for subjects in the “explicit” instructional condition are not directly predictable from the hypotheses, but two possibilities exist. The first alternative is that explicit instructions to associate face and voice information will only serve to strengthen or reinforce the bond between multimodal information in memory, and as such, subjects in the “explicit” instructional condition should show the same pattern of results as in Experiment 1, albeit with higher recognition scores. The second alternative is that the *experimentally induced* bond between cross-modal information in memory may serve to counteract the effects of the *naturally occurring and lawful* bond which relates the information in both modalities to a common underlying perceptual event. Thus, the second alternative predicts that the effects of face context in the recognition of words will be eliminated by explicit instructions during the dynamic experiment.

## Method

### Subjects

Subjects were 40 Indiana University undergraduates who participated in partial fulfillment of course requirements for Introductory Psychology. All subjects were native speakers of English, had normal hearing, and reported no history of speech or hearing disorders at the time of testing.

### Stimulus Materials

*STUDY AND TEST FOR WORD RECOGNITION:* The stimuli used in Experiment 3 were taken from the same digital database as those for Experiments 1 and 2. However, face context information was now provided by the entire video track that corresponded to the word being presented. In addition, the digitized movie clips were truncated such that the number of frames to the onset of speech was equal for all movies showing the different talkers speaking a given word. Thus, face+voice pairings were naturally occurring ones and were not randomly assigned.

Another computer program was written to trace out the course of each subject’s experimental session. This “plotting” program randomly assigned test items to serve as old or new words, randomly determined the order in which old and new words were presented during the test phase, and randomly determined the audiovisual context in which a word could occur during test. The program then determined which stimuli were composed of face+voice pairings where the face and voice came from different people. For these stimuli, the appropriate video track was dubbed on to the corresponding audio track by four steps. First, the header file for the movie which contained the necessary audio track was obtained. Second, a new digital movie structure was created using the audio information from the

appropriate audio track. Third, the video track pointer for the new movie was changed so that it pointed to the appropriate video track. Fourth, the data were flattened and used to write a new “dubbed” movie, which was stored in a location which was accessible later. Because the lead-in time for all movies which could possibly be dubbed together was equal, synchronization of the video and audio tracks was accomplished by default, with a maximum possible error of 16.5 ms, half the duration of one movie frame. It should be noted that running *correct* face+voice pairing movies through this same process would have resulted in output movies identical to those input, and as such, only movies that needed to be dubbed were actually processed in this manner, in order to save time and space.

The output of the “plotting” program was used as input to a modified version of the control program used in Experiments 1 and 2. The modified version simply played out the movies (study, dubbed and non-dubbed) in the order specified by the plotting program. The control program also displayed on-line instructions and collected responses, as before.

*TEST OF EXPLICIT FACE RECOGNITION:* Stimuli for the explicit face recognition portion of this experiment were identical to those used in Experiments 1 and 2 (i.e., they were static faces in their upright orientation).

**Apparatus**

A Macintosh PowerPC 8100/100 was used to run both the plotting program and the control program. Due to the increased demand for fast data transfer between the drive where the movies were stored and the video processing board, a ProDirect 9.6 GB hard drive with an Ultra-SCSI connection was used to store the stimuli. As in Experiments 1 and 2, the video track of each stimulus movie was presented on a 17” Apple Multiple Scan 17 Display color monitor, controlled by a Radius video board, while the audio track of each stimulus movie was presented over Beyer dynamic DT100 headphones calibrated to 74 dB SPL.

**Procedure**

Procedures for Experiment 3 were identical to those in Experiment 1, except that subjects were always told that the stimuli they would be viewing were movies.

As in the previous experiments, “old” or “new” responses for either words or faces, as the case may be, were collected by a button-box interfaced with a Strawberry Tree card. Responses, along with the parameters for a given trial, were recorded in text files for later analysis.

**Results**

In order for a subject to be included in the final data analysis, a score of 87.5% or better on the explicit face recognition memory test was necessary; the final analysis included 15 subjects from the “on the fly” instructional condition, while 17 subjects were included from the “explicit” instructional condition.

-----  
 Insert Figure 4 about here.  
 -----

Figure 4 shows the average *d'* scores in each testing condition for Experiment 3, separated as a function of Instructional Condition. The left side of the figure shows scores obtained in each of the four within-subjects conditions from subjects participating in the “explicit” instructional condition. The right side shows the same scores obtained from subjects in the “on the fly” instructional condition. Table 3 shows the corresponding Hits and False Alarm rates used in the calculation of *d'* scores for the various experimental conditions. As before, the calculation of *d'* scores was carried out such that the False Alarm rate used for the (F+V+) and (f-v-) conditions was the rate at which subjects responded “old” to new

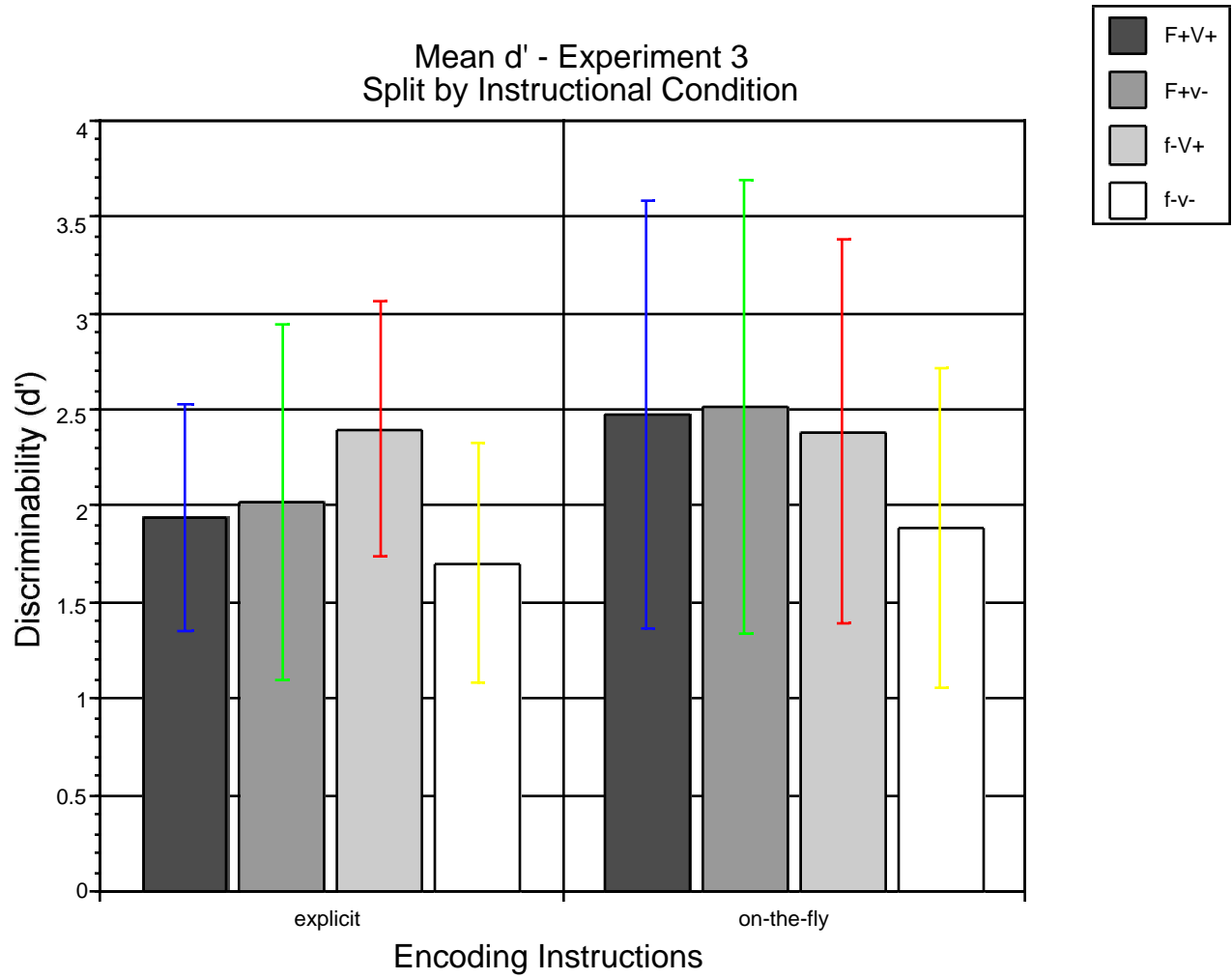
words presented in the context of a previously studied face+voice pairing (which, in the case of Experiment 3, was a naturally occurring face+voice pair). Similarly, the False Alarm rate used for the calculation of  $d'$  scores in the (F+v-) and (f-V+) conditions was the rate at which subjects responded “old” to new words presented in the context of face+voice pairings which had not been previously studied, and which do not occur naturally.

**Table 3**  
**Average Hits and False Alarms for all Conditions in Experiment 3,**  
**Split by Instructional Condition.**

Encoding				Instructions			
explicit				on the fly			
face+voice	average FAs	condition	average Hits	face+voice	average FAs	condition	average Hits
non-studied	3.06	F+V+	7.35	non-studied	2.47	F+V+	7.47
		f-v-	6.47			f-v-	6.40
studied	3.53	F+v-	6.94	studied	2.53	F+v-	7.47
		f-V+	7.88			f-V+	7.47

A repeated measures ANOVA was conducted on the  $d'$  scores for Experiment 3 with Face Context (new or old) and Voice Context (new or old) as the repeated measures and Instructional Condition (“explicit” or “on the fly”) as a between subjects variable. This analysis revealed a significant effect of Voice Context  $F(1,30) = 11.125$ ,  $p = 0.002$ . Across both instructional conditions, subjects were better able to discriminate old from new words when the old words were presented in the context of the voice with which that word was studied. In addition, the interaction between Face Context and Instructional Condition was significant,  $F(1, 30) = 4.977$ ,  $p = 0.033$ .

Figure 5 illustrates the Face Context x Instructional Condition interaction for  $d'$  scores in Experiment 3. The left panel of Figure 5 shows the discriminability scores for subjects in the “explicit”



**Figure 4:** Average d' scores for all four conditions in Experiment 3, as a function of instructional condition.

instructional condition; the left side shows scores for subjects in the “on the fly” instructional condition. As in Figure 2, each bar represents the average discriminability of items, collapsed across voice condition. A probe of this interaction using a 2 (Face Context) x 2 (Voice Context) repeated measures ANOVA split by Instructional Condition revealed that the interaction was due to a highly significant main effect of Face Context in the “on the fly” instructional condition only,  $F(1,14) = 6.827$ ,  $p = 0.020$ , but not in the “explicit” instructional condition,  $F(1,16) = 0.280$ , N.S. Thus, for subjects in the “on the fly” instructional condition, the distinction between old and new word items was significantly greater when those items were presented in the context of the face with which they were originally studied.

-----  
 Insert Figure 5 about here.  
 -----

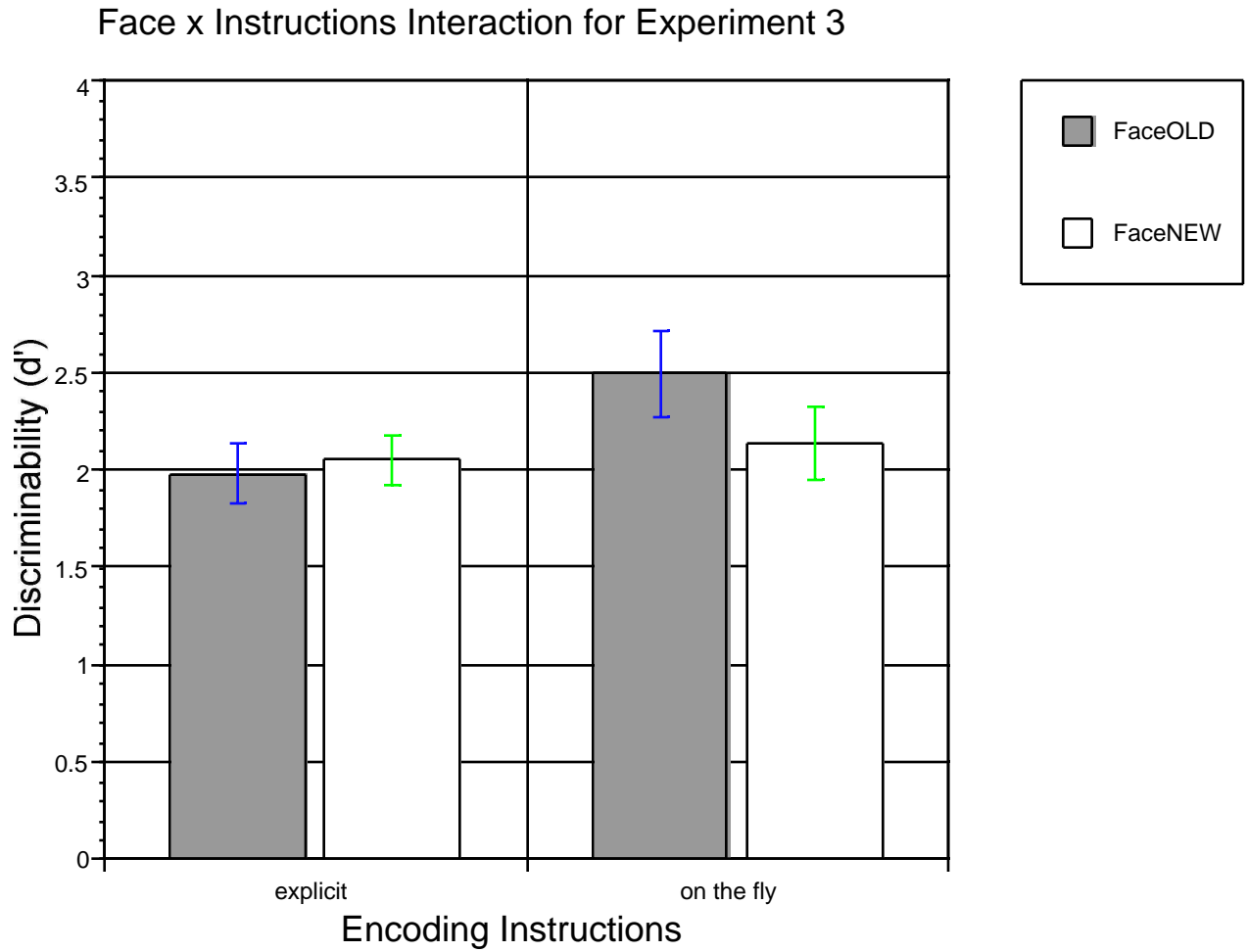
### Across Experiment Analysis

We also predicted that overall levels of performance would be higher in Experiment 3 when compared to Experiment 1 because subjects are able to make use of the naturally occurring and lawful bonds between information presented in the two separate modalities. In order to test this hypothesis, the data from both Experiments 1 and 3 were submitted to a repeated measures ANOVA with Face Context (new or old) and Voice Context (new or old) as the repeated measures and Instructional Condition (“explicit” or “on the fly”) and Experiment (1 or 3) as between subjects factors. Figure 6 illustrates all the data which was analyzed in this way. This figure is a restructuring of the data presented in Figures 1 and 4, so that the scores for each condition may better be compared across experiments. Figure 6a shows the scores from all subjects in the “explicit” instructional condition; Figure 6b shows the scores from all subjects in the “on the fly” instructional condition. The light bars represent scores obtained from subjects in Experiment 1 with static faces, while the dark bars represent scores obtained from subjects in Experiment 3 with dynamic faces.

-----  
 Insert Figure 6 about here.  
 -----

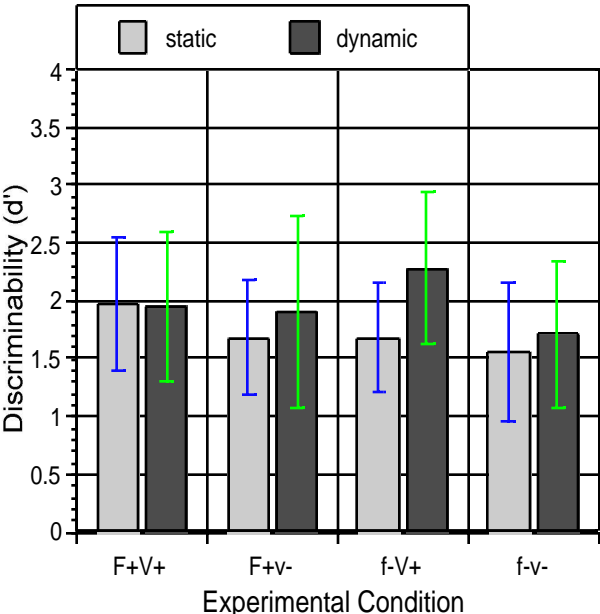
As can be seen from Figures 6a and 6b, scores were generally higher in Experiment 3 than they were in Experiment 1. Indeed, a significant main effect of Experiment was found,  $F(1,74) = 11.987$ ,  $p = .001$ , indicating that subjects in Experiment 3 with dynamic visual stimuli had better recognition scores than those in Experiment 1 with static visual stimuli. In addition, a highly significant main effect of Voice Context was found,  $F(1,74) = 14.442$ ,  $p < 0.0009$ , whereas there was no interaction between Voice Context and Experiment,  $F(1,74) = 1.827$ , n.s. This demonstrates that the voice effect was consistent across experiments. In other words, both Experiments 1 and 3 replicated previous findings showing that word recognition is affected by voice context (Goldinger, 1995; Palmeri et al., 1991); old words were more easily discriminated from new words when they were presented during test with the same voice as in test.

Finally, we observed a significant interaction between Face Context, Instructional Condition, and Experiment,  $F(1, 74) = 8.351$ ,  $p = .005$ . Figure 7 shows the interaction of Face Context and Experiment as a function of Instructional Condition. The graph in Figure 7 is a recombination of the data in Figures 2 and 5, reorganized to facilitate comparison across experiments. Thus, the left panel shows scores, collapsed across voice condition, for subjects who participated in the “on the fly” instructional condition for Experiments 1 and 3, while the right panel shows scores obtained from the “explicit” condition. Since this interaction is not easily interpretable, the data were split into two groups based on Instructional Condition,

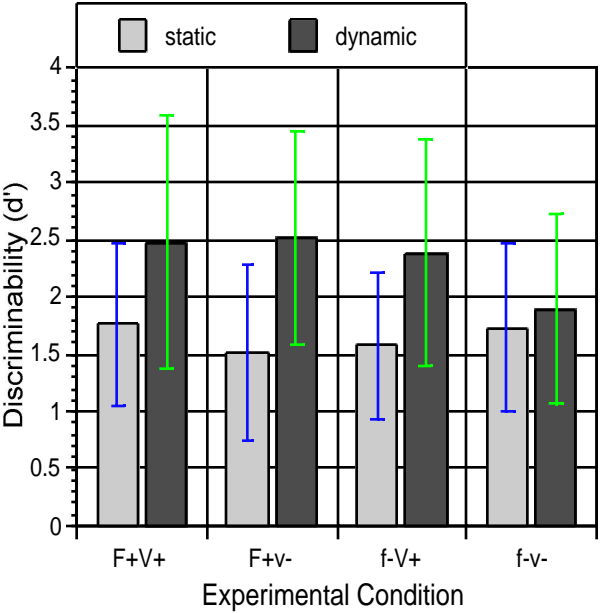


**Figure 5:** Average  $d'$  scores for test conditions in which the face was old or new as a function of instructional condition. The FaceOLD bar represents the average  $d'$  scores for test items in the (F+V+) and (F+v-) conditions. The FaceNEW bar represents the average  $d'$  scores for test items in the (f-V+) and (f-v-) conditions.

Static vs. Dynamic d' for subjects in the "explicit" Instructional Condition for Experiments 1(static) and 3(dynamic)



Static vs. Dynamic d' for subjects in the "on the fly" Instructional Condition for Experiments 1(static) and 3(dynamic)



**Figure 6a:** comparison of average d' scores for subjects in the “explicit” instructional condition of Experiments 1 and 3.

**Figure 6b:** comparison of average d' scores for subjects in the “on the fly” instructional condition of Experiments 1 and 3.

and each group's data were submitted to a repeated measures ANOVA, using Face Context and Voice Context as repeated measures and Experiment as a between subjects factor.

-----  
 Insert Figure 7 about here.  
 -----

The 2 (Face Context) x 2 (Voice Context) x 2 (Experiment) ANOVA for subjects in the “on the fly” instructional conditions of both experiments revealed a significant main effect of Experiment,  $F(1,36) = 10.575, p = 0.002$ : subjects' performance with “on the fly” instructions was always better in Experiment 3 (dynamic visual displays) than in Experiment 1. The ANOVA also revealed a significant Face Context by Experiment Interaction,  $F(1,36) = 5.469, p = 0.025$ . A post-hoc t-test confirmed that the source of this interaction was a difference in the effects of Face Context on the recognition scores for items presented in Experiment 3,  $t(14) = 2.613, p = 0.02$ . For subjects who were not explicitly instructed to attend to the faces in the experiment, Face Context only affected recognition scores when visual displays were dynamic. This finding can be interpreted as a form of implicit encoding of face information.

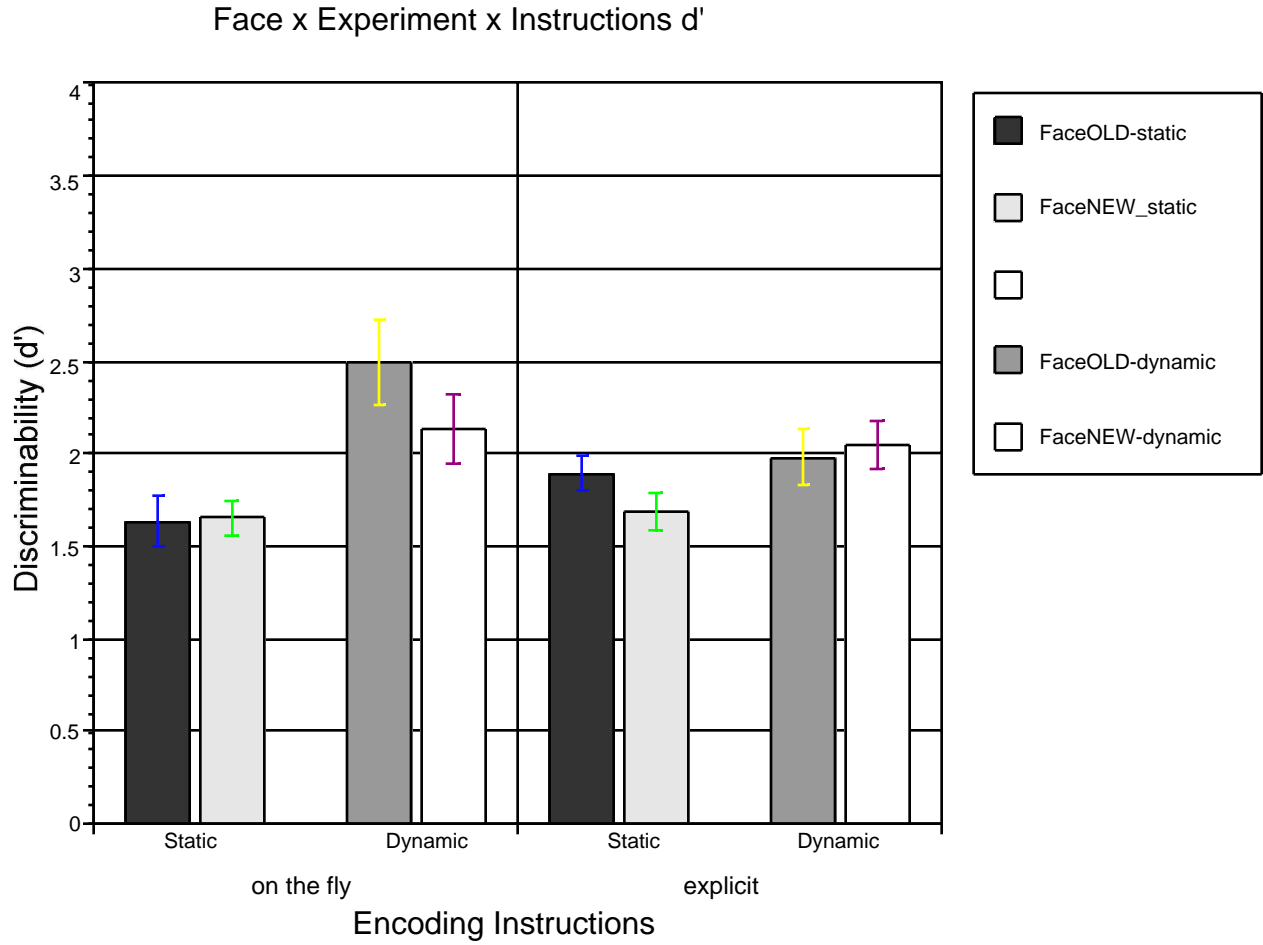
Analysis of the data from subjects in the “explicit” instructional condition also revealed a marginally significant interaction between Face Context and Experiment,  $F(1,38) = 3.036, p = 0.09$ . A post-hoc t-test confirmed that the source of this interaction was a difference in the effects of Face Context on the recognition scores for items presented in Experiment 1,  $t(22) = 2.198, p = 0.039$ : For subjects who were explicitly instructed that faces were important in the experiment, Face Context could only serve as an effective cue to recognition of words when visual displays were static.

In order to understand more fully the cross-experiment results,  $\eta^2$  values were computed for those main effects of Face Context that were found to be significant. Since  $\eta^2$  is taken to be the proportion of variance accounted for by a particular effect, it is useful as a measure of the relative magnitudes of different effects. For subjects in Experiment 1 (static visual displays) who were given the “explicit” instructions, the main effect of Face Context accounted for 18% of the total variance. For subjects in Experiment 3 (dynamic visual displays) who were given “on the fly” instructions, the main effect of Face Context accounted for 32.8% of the total variance, almost double that found in Experiment 1. Thus, for those subjects who were able to use face cues (given the experiment and instructional condition in which they participated), dynamic information was more useful than static information.

### Discussion

Experiment 3 replicated the finding that same-voice repetition of test items can facilitate later recall and recognition (Goldinger, 1995; Palmeri et al., 1995); across instructional groups and face contexts, old items were recognized better when the word was presented in the context of the voice with which it was originally presented than when it was presented in the context of a new voice. It is safe to assume then, that our experimental procedure was a valid one, because we were able to consistently find a repetition effect.

As expected, the findings from Experiment 3 are consistent with the hypothesis that dynamic visual information is implicitly encoded in memory and can be used as an effective retrieval cue in the recognition of words. For subjects in the “on the fly” instructional condition, recognition memory for words was improved when those words were presented in the context of the dynamic, articulating face with which it was originally presented. This finding is consistent with the hypothesis that the integrative encoding of visual information in memory for speech is automatic and mandatory when the face information is dynamic



**Figure 7:** comparison across Experiments 1 and 3 of average  $d'$  scores for test conditions in which the face was old or new as a function of instructional condition. Dark bars represent the average  $d'$  scores for test items in the (F+V+) and (F+v-) conditions. Lighter bars represent the average  $d'$  scores for test items in the (f-V+) and (f-v-) conditions.

(e.g., when it provides information concerning the linguistic content of the speech event), since subjects were never told that faces would be important at later stages of the experiment.

The comparison between Experiments 1 and 3 confirms another of our hypotheses: overall, performance was always better when visual information was dynamic, compared with the scores from static visual displays. Dynamic visual information about the articulation of a word is simply more useful than static information in distinguishing old from new items, regardless of the encoding strategies subjects use for faces.

Finally, it is interesting to examine the effects of our instructional manipulation on the ability of subjects to use face information as a cue to recognition of speech events across experiments. In Experiment 1, static faces could *only* be used as retrieval cues for recognition when subjects were instructed to attend to faces in the experiment. When such explicit encoding instructions were not used, subjects were unable to utilize visual information in the process of distinguishing old from new items. We assume that the explicit instructions, then, cause subjects to consciously encode the visual information in a way which will be conducive to later, explicit recognition. As a result, an arbitrary link between the explicitly encoded face and the simultaneous speech event is formed in memory. In Experiment 3, dynamic visual information was only used as a retrieval cue for recognition when subjects were *not* alerted to the necessity for explicit recognition of faces later. *Explicitly* encoded dynamic faces could not be used as cues in the recognition of speech events, indicating that this different encoding strategy actually interferes with the natural, mandatory encoding of dynamic visual displays of the talker's articulation along with the speech signal.

The overall pattern of results from Experiment 3 indicates that dynamic visual information is encoded and stored in an obligatory fashion (i.e., without additional task demand) in memory representations for speech events. In addition, the results imply that tasks that require explicit encoding of dynamic face information may actually interfere with this automatic encoding, eliminating the potential utility of visual information in the recognition of words later. Although the potential use of *static* visual displays as retrieval cues can actually be *increased* by encouraging explicit encoding strategies, this increase in utility is not nearly as beneficial as that gained naturally by dynamic presentation, as shown by the relative proportions of variance accounted for by Face Context in Experiments 1 and 3.

Taken together, the findings from this experiment support the conclusion that dynamic information about talkers' articulation of a speech event is encoded integrally in memory with information about the acoustics of that speech event. In addition, the results indicate that explicit encoding strategies can interfere with this process. Given that a static face can only be used as an effective retrieval cue when the information is encoded explicitly, we conclude that studies that utilize static visual displays may not be taken as evidence for audiovisual integration in memory representations of speech events.

## General Discussion

The present series of experiments was designed to examine the nature of multimodal speech representations in memory. In carrying out these experiments, we tested several hypotheses concerning the nature of the stimuli and the encoding tasks used to examine this important issue. Experiment 1 was designed as a control for the earlier Kato et al. (1995) experiments. Kato et al. (1995) used an explicit recognition memory procedure to determine whether static faces and arbitrarily paired voices were stored integrally in memory. Full sentences were used as the carriers of voice information. In contrast, the present set of experiments tested subjects' ability to use implicit information about face-voice pairings by explicitly recognizing isolated words spoken in a mixed factorial design of studied and non-studied faces and voices.

It is important to emphasize here that the results obtained in all three experiments replicated the basic findings of earlier studies by Goldinger (1995) and Palmeri et al. (1991); words repeated in the same voice used during study were more easily recognized than words presented in a different voice. Additionally, Experiment 1 showed that manipulating the instructions given to subjects can alter their encoding strategies and thus can affect the way in which static faces can be used as retrieval cues in the recognition of spoken words. We found that explicitly instructing subjects that faces would be important later in the experiment encouraged them to encode stimuli in a way that fostered the use of static visual displays as effective retrieval cues in the later recognition of words. When no explicit encoding instructions were given, subjects were not able to use pictures of faces as additional retrieval cues for the recognition of words.

We conclude that under normal circumstances static pictures of faces are not encoded in memory along with speech events. Because the static faces were *arbitrarily* paired with the voice speaking a word during study and test trials, the information in the optical displays were not correlated with the underlying speech events to be encoded. As a result, a relationship between the arbitrarily associated visual and acoustical information was only encoded when subjects were alerted to the importance of the artificial, experimental relationship inherent to the stimuli. In other words, our explicit encoding instructions acted to artificially increase the potential utility of static faces during the experiment and, as such, subjects encoded them visual information in faces with respect to their relationship to simultaneously occurring speech events.

In contrast, Experiment 3 used *dynamic* movies of talkers articulating the word to be remembered. Since the acoustic specification of a speech event is lawfully tied to its articulation, the visual displays in these dynamic movies were informative with respect to the underlying speech event, because they contained optical information about articulation. As a result, subjects automatically encoded visual information in representations of speech events and were able to use this information later as effective retrieval cues in the recognition of spoken words.

However, when subjects were *explicitly instructed* that faces would be important later in the experiment, faces were not used effectively. The results suggest that the aspects of a face which can be most easily used for explicit recall are not necessarily the same aspects of a face which are potentially informative about an underlying speech event. Although the kinematics of a talker's speech articulators may provide a rich source of visual information concerning the speech being heard and seen, this dynamic information may not be useful in the explicit recall of that talker's face, and vice versa. As a result, biasing subjects to treat articulating faces as items for explicit recall may actually serve to reduce the potential of those articulating faces as information carriers by drawing attention away from articulatory aspects of the visual display and towards those aspects useful in the explicit recognition of faces.

A comparison of the results in Experiments 1 and 3 also revealed that, while static faces could be used as effective retrieval cues to the recognition of words, the gain in discriminability as a result of repeated face context was greater in Experiment 3, where faces were dynamic and provided complementary information about the utterance to be recognized. This finding indicates that while the potential relevance of a static face to the recognition of simultaneously presented speech events can be manipulated experimentally, the human perceptual and memory systems are more aptly designed to encode and exploit naturally occurring and lawful relationships between simultaneously occurring intermodal events (Fowler, 1986; Gaver, 1993).

Taken together, the present results contribute to the growing body of literature indicating that the processes of speech perception and spoken word recognition should not be viewed as exclusively auditory phenomena (see Bernstein, Demorest, and Tucker, in press, for a review of this literature). As noted before, the role of visual information in the process of speech perception was first described by the landmark study of Sumby and Pollack (1954), who showed that the intelligibility of speech in noise is greatly increased when subjects are allowed to see the talker articulate. Sumby and Pollack also showed

that the relative contribution of visual presentation to overall intelligibility was independent of the speech-to-noise ratio under test, implying that the contribution of visual information is not simply *additive*, but *complimentary* to the auditory information presented. These initial conclusions were further supported by Erber (1969) who found that audiovisual information increased speech intelligibility even when speech is unintelligible when presented with auditory information alone. Both of these early studies suggest that acoustic and optical information can work in complimentary ways to support the perception of speech. The increase in speech intelligibility due to audiovisual presentation is equivalent to the gain in intelligibility afforded by an increase in speech to noise ratio of +15 dB (Erber, 1969; MacLeod and Summerfield, 1987; Middelweerd and Plomp, 1987; Rosenblum and Saldaña, 1996). This enormous gain in intelligibility clearly justifies further inquiry into the basis of multimodal speech perception and may provide important new theoretical insights into speech perception and spoken word recognition.

Further evidence of an integral role for vision in the perception of speech comes from the well-known McGurk effect (McGurk and MacDonald, 1976). With this illusion, McGurk and MacDonald showed that conflicting information in the auditory and visual modalities can significantly modify the perception of a speech sound and can even lead to a percept which is not specified by either modality alone (McGurk and MacDonald, 1976). For example, the presentation of an auditory [ba] with a visual [ga] led to the perception of [da] in 96% of the subjects tested. Because this multi-modal effect is extremely robust, subsequent studies have incorporated it into their designs in order to examine the integration of acoustic and optic information during the process of perception (Dekle et al., 1992; MacDonald and McGurk, 1978; MacLeod and Summerfield, 1987; Massaro, 1987; Massaro and Cohen, 1983; Massaro and Cohen, 1990; Munhall et al., 1996; Rosenblum and Saldaña, 1996; Rosenblum and Saldaña, 1992; Summerfield, 1984).

Taken together, these studies show that visual information, when available, is intrinsic to the process of speech perception. Indeed, a recent study by Bernstein et al. (in press) shows that visual information may even be *sufficient* for the process of speech perception, as demonstrated by the finding that some severely hearing impaired subjects can out-perform normal hearing subjects in identifying words presented with visual information only.

At the present time, several explanations have been proposed to explain the processes by which visual information affects the perception of speech. All of these explanations rest on assumptions concerning the nature of the information provided by the visual modality. For example, some explanations of the McGurk effect rest on the assumption that visual information and acoustical information allow for differing degrees of reliability in the perception of the speech sounds in question. Indeed, Summerfield (1987) presents evidence that the very speech segments which are most confusable when presented with solely auditory presentation are those which are least confusable when presented with solely visual information, and vice versa.

Massaro's (1987) formalization of this relationship, in terms of his Fuzzy Logical Model of Perception (FLMP), is based on the assumption that visual and acoustic information provide evidence with varying levels of reliability for the presence or absence of sub-phonemic features of the speech sound in question. Combining visually supported featural information with featural attributes which are supported by information from the auditory signal allows the perceptual system to derive a percept (e.g., Massaro, 1987; Massaro and Cohen, 1983). However, evidence has also been found which contradicts the proposal that information from disparate modalities is analyzed independently (Green and Kuhl, 1991). These results support a theory based on interactive processes in the perception of multimodal speech sounds.

Still, a theory of the analysis of sensory input cannot reveal much without addressing the fundamental problem of representation. As (Summerfield, 1987) points out:

“Accounts of audio-visual speech perception must suggest how a knowledge of audio-visual structure is represented, and what information exists in the acoustical and optical streams to indicate that they should be interpreted together.” (p.31)

A growing body of evidence suggests that the neural representation of speech must include information about the dynamic changes articulators make during the production of speech (Green and Gerdeman, 1995; Rosenblum and Saldaña, 1996). In a recent study, Rosenblum and Saldaña (1996) found that a point-light face display influenced the perception of speech only when it provided kinematic information about the talker’s articulators, i.e., when it was moving. Static displays of point-light stimuli did not affect the perception of speech at all. Rosenblum and Saldaña (1996) argue that their findings contradict traditional theories of speech perception based on the perception and encoding of discrete featural cues, whether those cues be analyzed independently (e.g., Massaro, 1987) or interactively (e.g., McClelland and Elman, 1986). Indeed, it is hard to imagine how theories of perception based on units which have no extent temporally, like linguistic features or phonemes, can account for these findings.

In order to solve the problem of encoding temporal information in speech, it is worthwhile to consider the proposals of several investigators who have called for a drastic overhaul of the basic assumptions of most current theories of speech perception. One approach is the direct realist theory of speech perception advocated and elucidated by Fowler (1986). From this theoretical perspective, speech is perceived with respect to the event which produced it. An event is said to “structure” an informational medium (such as sound or light) in a lawful manner. Consequently, sound is viewed as a medium through which information about distal speech events may travel. The structure of this informational medium thus specifies the perception of a speech event. From the direct realist standpoint, then, it does not matter through which sensory modality information about a speech event comes, only that the information in that channel is related in some way to the event being perceived (Fowler, 1986).

In a paper presenting a new approach to auditory perception, Gaver (1993) argues for a conceptualization of the human auditory system designed to perceive the events and sources which cause sounds, not the sounds themselves. The structure of acoustic energy produced during a sound-making event is lawfully tied to the event which produced it and Gaver (1993) claims that the human auditory system may be structured in a way to exploit these relationships. While Gaver was more concerned with building a taxonomy of the events which can be perceived through sound than with explaining the process of speech perception, it is not hard to see how his ideas may be generalized easily to some of the long-standing problems in speech perception and spoken word recognition, since the acoustic energy which transmits speech is shaped and structured by the process of producing speech through the speech motor control system and the human vocal tract.

The present set of results suggests that the integrated encoding of multimodal sensory inputs is automatic and mandatory when the optic display available during a speech event is potentially informative about the underlying speech event. In Experiment 1 which used static faces, we found that explicitly instructing subjects about the potential usefulness of faces enabled them to make use of optical cues during later recognition of speech. However, for subjects who were not instructed in this way, static faces were shown to be very poor retrieval cues. Under normal circumstances, however, static faces are not encoded integrally with the speech signal, because there is no information in the acoustic signal that specifies a connection between the information in the two sensory modalities. In contrast, in Experiment 3, subjects who were not explicitly instructed about the potential usefulness of faces were able to utilize dynamic faces as cues in their recognition of speech events; explicitly instructed subjects could not. In this condition, perceivers were able to exploit and use the natural correlation between auditory and optical sources of information about a common underlying speech event.

Direct realism also takes a stand on the nature of information containing integrated sensory inputs; the information processed in the perception and encoding of speech must be *modality-neutral*, so that any information relating to an event can be used in perception, regardless of the modality through

which it is obtained (Fowler & Rosenblum, 1991; Rosenblum and Saldaña, 1996; Summerfield, 1987). Whether this modality-neutral form is represented in articulatory terms or not, our results provide additional support for the hypothesis that the critical information is based on the time-varying properties of the speech signal (Rosenblum and Saldaña, 1996).

The findings obtained in these three experiments suggest that dynamic visual displays are automatically encoded in memory with auditory information in speech because they provide an additional source of information concerning the kinematic, time-varying properties of the underlying speech event. Dynamic visual information is lawfully tied to the form of acoustic information, and, as such, provides valuable cues about the source of the speech signal. This lawful bond between the disparate sensory modalities, to paraphrase Summerfield, (1987), serves as the information which indicates that the two input streams should be analyzed together because they are informative about the same underlying event. Explicit instructions to encode dynamic articulating faces may serve to divert the perceiver's attention away from those aspects of the visual display which are informative about the utterance and towards those aspects which can foster explicit recognition of faces in the memory task at hand.

The present set of findings, therefore, have several important implications for both memory and spoken word recognition. First, the evidence for multimodal encoding of speech presents serious conceptual problems for several well-known conceptualizations of human memory which operate on abstract, modality-specific inputs, such as Baddeley's phonological loop and visuospatial sketchpad (e.g., Baddeley, 1980). Given that all information entering long-term memory must first pass through working memory, it seems at best superfluous to assert that multimodal speech input, after being categorized using integrated representations, should once again be split into its component phonological and visuospatial parts for processing in working memory, only to be re-integrated later on for storage in long-term memory. An important new direction for future research on speech perception and spoken word recognition is to determine what the phonological loop "knows" about the speaker's face and how this kind of information is represented in working memory.

In addition, the evidence obtained here concerning different types of intermodal relationships (i.e., experimentally validated simultaneity conditions vs. natural, lawful and automatic interdependencies) and their relative value during recognition argues against any notion that information from disparate modalities is simply linked via some central executive function. In other words, it is hard to imagine within the context of Baddeley's tri-partite model of working memory, why a central executive, with the power to link representations of information from disparate sensory modalities, might do so differentially based on the relationship between them. Rather, it seems more parsimonious to posit a working memory system which operates on modality-neutral representations whose degree of integration is determined by task demands and exploitation of naturally occurring relationships. At the very least, the current conceptualization of the phonological loop must change from a system which operates on abstracted, discrete, symbolic linguistic units to a system that encodes and manipulates input which is informative about underlying speech events.

A second implication of the present findings concerns the structure of the mental lexicon. Landauer and Streeter (1973) have shown that there are important differences in the structure of similarity neighborhoods for high and low frequency words when similarity is measured by a metric that is based on the number of phonemes shared by two words. Luce (1986) and Luce and Pisoni (1998) found that these differences have ramifications for the perception and recognition of spoken words. Their findings suggest that the mental lexicon is functionally structured so as to reflect the phonemic similarity between words.

In a recent computational study that builds on the earlier work of Landauer and Streeter (1973), Auer and Bernstein (1997) analyzed the structure of the lexicon with all words recoded into strings of *visemes*. Their analysis revealed that a visually specified lexicon has a structure that "compliments" that of an acoustically specified one. That is, it seems as though spoken words which occupy dense

*phonemically* specified similarity neighborhoods reside in relatively sparse viseme-coded neighborhoods, and vice versa (Auer and Bernstein, 1997).

If, as our study suggests, the neural representation of spoken words in lexical memory contains multimodal information, then the findings of Auer and Bernstein (1997) should have ramifications for the perception of words, just as the findings of Landauer and Streeter (1973) and Luce (1986) did for studies of spoken word recognition. In the same way that phonemic similarity has been shown to have effects on the perception of words, so too should similarity along visual dimensions. Thus, a serious reconceptualization of the underlying similarity space of spoken words in the lexicon is warranted, which takes into account both acoustic and optical features of spoken words.

The results of the present study demonstrate that speech perception is not necessarily confined to the realm of audition alone, but rather can operate in multiple sensory domains (Bernstein, 1998). One consequence of this view of speech is that it is no longer sufficient to base our theories of speech perception and spoken word recognition on auditory data alone, or to motivate those theories on acoustically-biased foundations. It is time now for a re-evaluation of the information we consider relevant to speech perception and of the processes we deem necessary for successful spoken word recognition.

## References

- Auer, E. T., & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness and lexical uniqueness. *Journal of the Acoustical Society of America*, **102**, 3704 - 3709.
- Baddeley, A. D. (1997). *Working memory: Theory and Practice*. Boston: Allyn and Bacon.
- Banks, W. P. (1970). Signal Detection Theory and Human Memory. *Psychological Bulletin*, **74**, 81 - 99.
- Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (submitted). Speech perception without hearing.
- Campbell, R. (1994). Audiovisual speech: Where, what, when, how? *Current Psychology of Cognition*, **13**, 76 - 80.
- Dekle, D. J., Fowler, C. A., & Funnel, M. G. (1992). Audiovisual integration in perception of real words. *Perception and Psychophysics*, **51**, 355 - 362.
- Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Technical Note Contract No. AF 19(604)-1962). Bloomington, IN: Indiana University Hearing and Communication Laboratory.
- Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, **12**, 423 - 425.
- 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., & Rosenblum, L. D. (1991). Perception of the phonetic gesture. In Mattingly, I. G. & Studdert-Kennedy, M. (Eds.), *Modularity and the motor theory*. (33-59). Hillsdale, NJ: Erlbaum.
- Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, **5**, 1 - 29.

- Goldinger, S. D. (1996). Words and Voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 1166-1183.
- 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. (1991). Integral Processing of Visual Place and Auditory Voicing Information During Phonetic Perception. *Journal of Experimental Psychology: Human Perception and Performance*, **17**, 278 - 288.
- Kato, T., Kanzaki, R., Tohkura, Y. i., & Akamatsu, S. (1995). *Effects of other-mode context on face and voice memory* (Technical report of IEICE, PRU95-88, HIP95-15 (1995-07)). Kyoto, Japan: ATR Human Information Processing Research Laboratories.
- Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Behavior*, **12**, 119 - 131.
- Legge, G. E., Grossmann, C., & Pieper, C. M. (1984). Learning Unfamiliar Faces. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **10**, 298 - 303.
- Lockhart, R. S., & Murdock, B. B., Jr. (1970). Memory and the Theory of Signal Detection. *Psychological Bulletin*, **74**, 100-109.
- Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon* (Research on Speech Perception Technical Report No. 6). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear & Hearing*, **19**, 1 - 36.
- MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception and 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. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Massaro, D. W., & Cohen, M. M. (1983). Evaluation and Integration of Visual and Auditory Information in Speech Perception. *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 753 - 771.
- Massaro, D. W., & Cohen, M. M. (1990). Perception of synthesized audible and visible speech. *Psychological Science*, **1**, 55 - 63.
- Massaro, D. W., & Cohen, M. M. (1996). Perceiving speech from inverted faces. *Perception & Psychophysics*, **58**, 1047 - 1065.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1 - 86.

- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.
- Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *Journal of the Acoustical Society of America*, **82**, 2145 - 2147.
- Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, **41**, 329 - 335.
- Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, **58**, 351 - 362.
- Palmeri, T. J., Goldinger, S. J., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **19**, 309 - 328.
- Parks, T. E. (1966). Signal-detectability Theory of Recognition-Memory Performance. *Psychological Review*, **73**, 44-58.
- Pisoni, D. B., Saldana, H. M., & Sheffert, S. (1995). Multimodal encoding of speech in memory: a first report. In *Research on Spoken Language Processing Progress Report #20* (pp. 297-305). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Pollack, I. (1959). Identification of Elementary Auditory Displays and the Method of Recognition Memory. *The Journal of the Acoustical Society of America*, **31**, 1126-1128.
- Rosenblum, L. D., & Saldana, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception & Psychophysics*, **52**, 461 - 473.
- Rosenblum, L. D., & Saldana, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, **22**, 318 - 331.
- Schacter, D. L., & Church, B. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 915 - 930.
- Sheffert, S., Lachs, L., & Hernandez, L. (1998). Hoosier Audiovisual Multitalker Database. In *Research on Spoken Language Processing Progress Report #21*. Bloomington, IN: Speech Research Laboratory, Indiana University.
- Shepard, R. N. (1967). Recognition Memory for Words, Sentences, and Pictures. *Journal of Verbal Learning and Verbal Behavior*, **6**, 156-163.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215.
- Summerfield, A. Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, **36A**, 51-74.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Aslin, R. N., Alberts, J. and Peterson, M. J. (Eds.), *The Development of Perception: Psychobiological Perspectives*. (219 - 255). New York: Academic Press.

Valentine, T. (1988). Upside-down faces: A review of the effects of inversion upon face recognition. *British Journal of Psychology*, **79**, 471 - 491.

Yin, R. K. (1970). Face recognition: A dissociable ability? *Neuropsychologia*, **23**, 395 - 402.