Beyond Recognition: Visual Contributions to Verbal Working Memory

Susan Nittrouer and Joanna H. Lowenstein

Speech, Language, and Hearing Sciences
University of Florida, Gainesville, Florida

The authors have no relevant conflicts of interest to report.

Correspondence should be addressed to:
Susan Nittrouer
Speech, Language, and Hearing Sciences
University of Florida, PO Box 100174
Gainesville, FL 32610
Email: snittrouer@phhp.ufl.edu.

Abstract

Purpose: It is well recognized that adding the visual to the acoustic speech signal improves recognition when the acoustic signal is degraded, but how that visual signal affects post-recognition processes is not so well understood. The current study was designed to further elucidate the relationships among auditory and visual codes in working memory, a post-recognition process.

Design: In a main experiment, eighty young adults with normal hearing were tested using an immediate serial recall paradigm. Three types of signals were presented (unprocessed speech, vocoded speech, and environmental sounds) in three conditions (audio-only, audio-video with dynamic visual signals, and audio-picture with static visual signals). Three dependent measures were analyzed: (1) magnitude of the recency effect; (2) overall recall accuracy; and (3) response times, to assess cognitive effort. In a follow-up experiment, thirty young adults with normal hearing were tested largely using the same procedures, but with a slight change in order of stimulus presentation.

Results: The main experiment produced three major findings: (1) Unprocessed speech evoked a recency effect of consistent magnitude across conditions; vocoded speech evoked a recency effect of similar magnitude to unprocessed speech only with dynamic visual (lipread) signals; environmental sounds never showed a recency effect. (2) Dynamic and static visual signals enhanced overall recall accuracy to a similar extent, and this enhancement was greater for vocoded speech and environmental sounds than for unprocessed speech. (3) All visual signals reduced cognitive load, except for dynamic visual signals with environmental sounds. The follow-up experiment revealed that dynamic visual (lipread) signals exerted their effect on the vocoded stimuli by enhancing phonological quality.

Conclusions: Acoustic and visual signals can combine to enhance working memory operations, but the source of these effects differs for phonological and non-phonological signals. Nonetheless, visual information can support better post-recognition processes for patients with hearing loss.

INTRODUCTION

Problem Statement

By now it is quite apparent that the visual speech signal can support the detection of acoustic speech signals at poorer signal-to-noise levels than is possible when it is not present (e.g., Bernstein et al., 2004; Grant, 2001; Grant & Seitz, 2000; Plyler et al., 2015), and can evoke more accurate recognition of speech in degraded listening conditions (e.g., Bernstein & Grant, 2009; Grant et al., 1998; Taitelbaum-Swead & Fostick, 2017; Walden et al., 1974) or when listeners have hearing loss (e.g., Erber, 1972; Lachs et al., 2001). Based on this evidence, professionals involved in aural rehabilitation routinely advocate for the provision of visual information when the acoustic signal does not support optimal speech recognition (e.g., Lalonde & McCreery, 2020; Tye-Murray et al., 2005). But participants in communication exchanges must do much more than simply detect and recognize speech. They must store the speech long enough to interpret it, as well. Where post-recognition processes are concerned, there have been far fewer investigations into the role of visual signals in those processes (Bielski et al., 2020). The purpose of the current study was to examine whether visual signals can support one post-recognition process, namely working memory, and if so, to identify the associated mechanisms.

A multicomponent model of working memory

Speech signals are ephemeral. Listeners must rapidly and accurately recover critical information from complex, but fleeting spectral-temporal patterns, integrate that information, and store it long enough to compile the message. These compulsory functions mean that working memory plays a critical role in the processing of acoustic speech signals, and that role is even greater when signal quality is degraded, as it is for listeners with hearing loss (Lyxell et al., 1996; Lyxell et al., 1998; Peelle, 2018; Rönnberg, 2003).

An often-cited and widely accepted model of working memory is the multicomponent model that over the years has been described most notably by Baddeley (e.g., 2000; 2012;

Baddeley & Logie, 1999), starting with the original work of Baddeley and Hitch (1974). This model serves as the foundation of much work where speech is concerned, as is the case here, because its original focus was on the storage of verbal material. As conceptualized and illustrated in Figure 1, this model consists of three parts: a central executive, largely responsible for attentional control, and two support systems. The focus of the study reported here was on the two support components (the phonological loop and the visuospatial sketch pad), with a primary emphasis on the nature of representation in the phonological loop, but also with reference to the way that sensory input across the two components may be integrated.

In the original characterization of the Baddeley and Hitch model, the description of the phonological loop was uncertain. It continued to evolve over iterations of the model, particularly regarding the nature of the representation being stored (Baddeley & Hitch, 2019). The idea generally espoused in the early versions of the multicomponent model characterized the phonological loop as a sort of continuous tape, transmitting an uninterrupted acoustic stream. At the early stages of transmission of information through the working memory system the representation was strictly acoustic, but it was transduced into a phonological form shortly after the start of processing. The idea of a *pre-categorical acoustic store* was adopted, based on evidence from experiments involving immediate serial recall. Digits or words were typically used in these experiments, presented auditorily or visually (in print) and either with or without stimulus suffixes (Aaronson, 1968; Conrad & Hull, 1968; Crowder & Morton, 1969; Murray, 1966; Penney, 1975). These experiments required subjects to recall the order of presentation for lists of digits or words. It was observed that subjects were more accurate at recalling words in early list positions (the primacy effect), regardless of modality of presentation: auditory or visual. Differences based on modality of presentation, however, were observed for serial positions near the ends of lists. Subjects showed enhanced recall for these positions compared to recall for middle-list positions, as long as presentation was auditory rather than visual. This was known as the recency effect, and it was robustly demonstrated – except under one condition. When a verbal suffix was presented at the ends of lists with instructions that it should be ignored, the

recency effect was reduced almost to the point of being eliminated entirely. This effect – termed the suffix effect – was purportedly due to the acoustic information in the suffix entering that initial stage of processing and displacing the acoustic information currently stored there, before it could be converted to a durable phonological representation (e.g., Crowder & Morton, 1969). But later research would cast into doubt the purely acoustic basis of this interpretation.

The nature of representations in the memory buffer

It was around the time that these modality-specific outcomes were being discovered that questions regarding the fundamental nature of the representation used in speech perception arose. In particular, Liberman and colleagues were developing a model of speech perception termed the Motor Theory, in which it was proposed that articulatory movements are the elements recovered in speech perception, rather than a strictly acoustic representation (e.g., Liberman et al., 1967). Although details of that earliest model were not sustainably embraced by the psycholinguistic community, the broader notion of phonological structure as sensorimotor in nature continued to be propagated (e.g., Studdert-Kenney, 1987), along with proposals that articulatory gestures form the bases of phonological representations (e.g., Browman & Goldstein, 1986; 1989). In fact, for a time while Baddeley and Hitch were contemplating the structure of the phonological loop, they adopted the term *articulatory* loop, based on the finding that articulatory rehearsal (whether overt or subvocal) could reinforce representations in that component of their memory model (Baddeley et al., 1984). Further support for the idea that representations in this loop are articulatory in form came from experiments in which it was observed that only spoken suffixes could evoke the suffix effect, and it did not matter whether that suffix was heard or lipread (Campbell & Dodd, 1980; Crowder, 1983; Salter & Colley, 1977). For example, Spoehr and Corin (1978) presented lists of digits, and used *zero* as a suffix. This *zero* suffix was presented in a variety of forms: *(1)* a spoken version with only auditory information; *(2)* a spoken version with only lipread information; *(3)* a version with the numeral *0* written on a card; and (4) a version with the word *zero* written on a card. Results showed that

both forms of the spoken suffix (auditory and lipread) reduced the recency effect, but the written versions did not. Other studies supplemented this outcome, showing that the suffix effect could not be elicited when an acoustic tone was appended to a list of spoken digits or words, even when a spoken suffix appended to the same lists successfully elicited the effect (de Gelder & Vroomen, 1992; Nairne & Crowder, 1982; Rowe & Rowe, 1976). In fact, a spoken suffix apparently does not need to be heard or seen at all, but can instead be silently articulated by research subjects themselves to achieve the effect (Nairne & Crowder, 1982; Nairne & Walters, 1983). A general conclusion reached as a consequence of these studies was that the suffix must be articulatory in nature to have its effect, so by extension, representations in the phonological loop itself must similarly be articulatory in nature.

Thus, the requisite condition for evoking both the recency and the suffix effects appeared to be that stimuli needed to retain dynamic articulatory structure, although it did not matter whether those stimuli were presented through the auditory modality, as in heard materials, or through the visual modality, as in lipread materials. (e.g., Campbell & Dodd, 1980; Conrad & Hull, 1968; Greene & Crowder, 1984). There remained, however, one challenge to the conclusion that these effects arose explicitly because stimuli were articulated, and that was that all articulated signals involve movement; graphic materials do not. Consequently, it seemed plausible that movement itself (i.e., the dynamic nature of the stimuli) was responsible for enhanced recollection at the ends of stimulus sequences, rather than the articulatory nature of the stimuli. Campbell et al. (1983) attempted to resolve this issue in a series of four experiments involving heard, lipread, and signed stimuli (with the latter presented to native signers), along with graphic displays of digits and invented symbols. Although elegantly designed, outcomes across experiments were ambiguous: Primary linguistic codes (signed or lipread materials) were clearly found to evoke robust recency effects. With graphic displays, however, it was found that adding movement even in an unnatural manner to the display (such as with bars moving across the display) created recency effects when none were found otherwise. Moreover, for speech stimuli, simply displaying a static image of a face producing the stimulus was enough to evoke a

recency effect. In the end, no firm conclusions could be reached regarding whether it was explicitly the articulating vocal tract or dynamic structure that accounts for the recency effect, and there has been no follow up to that study, until now. The first goal of the current investigation was to examine the nature of the code used for storage of verbal material.

Memory for nonverbal acoustic stimuli

The earliest work on short-term or working memory used digits or words as stimuli, and Baddeley and colleagues built their multicomponent model of working memory around just those sorts of stimuli. Thus, when scientists began studying working memory for nonverbal acoustic stimuli, existing models were not necessarily a good fit for their results. If verbal materials are processed in the phonological loop and visual materials are processed in the visuospatial sketch pad, the question could be asked of where nonverbal acoustic stimuli are processed. The answer generally arrived on – when a multicomponent model is applied – is that these kinds of stimuli are also thought to be processed in the phonological loop, even though phonological structure cannot be recovered from such stimuli. Instead, it is believed that acoustic attributes are recovered and stored as what is termed *auditory sensory memory* (e.g., McKeown et al., 2011; Nees, 2016). Based on the extant literature, however, it is difficult to evaluate similarity in mechanisms for serial recall of verbal and nonverbal materials, because short-term recall for nonverbal stimuli has typically not been investigated by presenting lists of nonverbal items for serial recall. Instead, this phenomenon has been explored in discrimination experiments (e.g., Deutsch, 1972; McKeown et al., 2011; Nees, 2016; Starr & Pitt, 1997). Overall the mechanisms underlying working memory for nonspeech acoustic stimuli remain unclear (Jeong & Ryu, 2016).

One effort to study working memory for nonspeech acoustic signals with a recall, rather than a discrimination task was conducted by Nittrouer and Lowenstein (2014). In that study serial recall was examined for three kinds of stimuli: *(1)* real words presented in their unprocessed form; *(2)* those same words degraded by noise vocoding; and *(3)* nonverbal environmental sounds. Words or sounds were presented in lists of eight items, with the same

words or sounds used in each of the ten lists comprising each of the three conditions (i.e., closed sets). Eight channels were used in the vocoding process – a fairly high number – so the resulting stimuli were easily recognized. The use of environmental sounds as the nonverbal stimuli diverged from most research into short-term memory for nonverbal sounds, not only because a serial recall task was used, but also because most of the previous work had involved tones rather than ecologically valid sounds (Nees, 2016). Nonetheless, environmental sounds cannot be coded in the phonological loop with phonological structure. Articulatory rehearsal was constrained by instructing subjects to keep their mouths closed and not move any part of their mouths; responses were recorded by subjects tapping on pictures representing the stimuli in the order recalled. Some results from the Nittrouer and Lowenstein study are replotted in Figure 2, and reveal that subjects demonstrated the classic primacy and recency effects for verbal material, even though overall recall was diminished for the vocoded items. The recency effect was greatly reduced for the nonverbal material, supporting the proposal that only articulated stimuli evoke the recency effect. Conversely then it might be concluded that any set of stimuli presented in a serial recall experiment that evokes a strong recency effect is being coded in a short-term memory buffer with an articulatory (phonological) code. A still-unanswered question, however, is the role, if any, that movement plays in evoking that effect. The current study addressed that question.

Auditory-visual integration of nonverbal stimuli

The multicomponent model of working memory explicated by Baddeley and colleagues (e.g., 2000; 2012; Baddeley & Hitch, 2019; Baddeley & Logie, 1999) served as an appropriate basis for addressing the first goal of this study, concerning the nature of representation of verbal material in working memory. The second goal of the current study was to examine the interaction of sensory signals across modalities; other models of working memory were able to inform this goal. Baddeley's multicomponent model of working memory posits separate components, or modules, for phonological and visual signals, and no explicit account is offered

for how signals across modalities might interact when presented simultaneously: Would we expect interference or integration (also known as binding)?

The embedded-process model of Cowan (1988; 1999; Adams et al., 2018) is not modular, so is able to handle sensory input from disparate modalities without difficulty. According to this account, information comes in through a brief sensory store, activating features in long-term storage associated with that information. Phonological and visual information would not be separated in this model. However, the model makes no special predictions for how storage with a phonological code might differ from storage with other acoustic codes, as in the recency effect described above. Much of the work examining the interaction of signals across modalities according to the embedded-process model has been done with nonverbal signals (e.g., Li & Cowan, 2021). Nonetheless, support for a model with general storage, such as Cowan's embedded-process model, could be obtained if recall was enhanced in audio-visual conditions in the current study.

Cognitive Effort

The third and final goal of the study reported here was to assess how cognitive effort in service to working memory is affected by the addition of different kinds of visual signals to different kinds of acoustic signals. Although cognitive effort has been defined in a number of ways, the traditional account is that it refers to how much central processing is required to complete a task (Tyler et al., 1979). One factor that can affect cognitive effort is how clear the sensory signal is. For example, individuals who have access only to degraded acoustic signals because of hearing loss must expend greater cognitive effort just to recognize speech (Peelle, 2018; Pichora-Fuller et al., 2016; Wright & Gagné, 2020). Accordingly, if the addition of the visual signal to the acoustic signal during presentation enhances the representation of those signals – making storage and recall more efficient – response times should decrease in the audio-visual condition, compared to the audio-only condition. However, sensory input across modalities or input of signals with different features could interfere with each other during

processing, if they are competing for limited attention in a general-storage system, such as that of Cowan (Li & Cowan, 2021). This situation would make response times longer. No such interference would be expected in multicomponent systems, such as that of Baddeley, because acoustic and visual signals would be processed by separate components. Of course, by the same reasoning, no reduction in cognitive effort would be predicted by the multicomponent model, at least when nonverbal material is being presented in the visual domain, because there is no interaction between signals presented in different modalities. However, according to the multicomponent model of Baddeley, reduced cognitive effort might be observed when the visual speech signal is presented in combination with the vocoded signal, because both are inherently articulatory (phonological) in nature, so both would be processed in the phonological loop.

A valid metric of cognitive effort is response time (DeLeeuw & Mayer, 2008; Sali & Egner, 2020), and that metric has previously been applied to studies of serial recall. In particular, Nittrouer et al. (2013) tested subjects in a serial recall paradigm using phonologically dissimilar, monosyllabic nouns that could transparently be represented with pictures (e.g., *rake*, *ham*, *soap*) or adjectives that shared the traits of phonological dissimilarity and monosyllabicity, but could only indirectly be represented with pictures (e.g., a picture of a pool for *deep* and a picture of a child crying for *sad*). Accuracy of serial recall was similar across the two sets of stimuli, but response times were significantly slower for the adjectives than for the nouns, demonstrating that the process of responding required greater cognitive effort for the adjectives than for the nouns; in that case the effect was attributed to the effort needed to connect the words to the less-transparently related pictures. That experiment supported the validity of using response time as an index of cognitive effort, and the same procedures were implemented in the current study.

The Current Study

The current study was undertaken with three major goals in mind. One goal was to assess the nature of representation in storage for speech signals; in particular, what quality of

an articulated signal influences the coding of spoken materials in a short-term memory buffer so as to evoke the recency effect: Is it the articulatory nature of the stimuli or the movement that is inherent in all articulated signals? To address this goal, three types of stimuli were presented in a classic serial recall paradigm: (1) non-rhyming, phonologically dissimilar words in their natural (*unprocessed*) form; (2) the same words presented in a *vocoded* form, applying greater spectral degradation than that used by Nittrouer and Lowenstein (2014) to render them less speech-like; and (3) the nonverbal *environmental sounds* from Nittrouer and Lowenstein. All words were monosyllabic, concrete nouns that could be easily represented by pictures. Similarly, the environmental sounds were all ones that could be represented by pictures.

These three types of stimuli were presented in each of three conditions: an *audio-only* condition and two audio-visual conditions. In one of the audio-visual conditions, the movement that generated the stimulus was provided in conjunction with the stimulus itself. For the two sets of speech stimuli (unprocessed and vocoded words) this meant that the talker's face was seen producing these words. For the environmental sounds, this meant that the object that would generate the sound in the real world was shown, moving in synchrony with the sound being produced. This condition is termed the *audio-video* condition, because a brief video was used to display the movement in each case.

The other audio-visual condition paired presentation of the acoustic stimulus with the picture that was used for responding. As in Nittrouer and Lowenstein (2014), subjects registered their responses by tapping on pictures shown on a computer monitor in the order recalled. In the case of the speech stimuli, this involved pictures of the words themselves, such as a picture of a bar of soap for *soap*. For the environmental sounds, the pictures were the objects that would generate the sounds, such as a picture of a broken glass for the sound of glass breaking. This was termed the *audio-picture* condition. Different groups of subjects were tested in the audio-video and the audio-picture conditions to prohibit transfer effects from one condition to the other. Stimuli in the audio-only condition were always presented first, and stimuli in the audio-visual

condition (either video or picture) were presented second, in order to avoid subjects being able to incorporate the visual codes into their stored representations later (Rosenblum et al., 2007).

The primary question addressed with these stimuli was whether it is specifically the articulated nature of the stimuli that evokes the recency effect or rather movement more generally? The expectation was that in the audio-only condition a recency effect would be observed only for the unprocessed words, because only in this condition do stimuli provide clear articulatory structure. The four-channel vocoded stimuli provided a signal that lacked apparent articulatory structure, or phonological form, at least upon first exposure to these signals. Increased difficulty in mapping acoustic signals onto phonological form, as happens with such vocoded signals, has been found to impair serial recall of vocoded material (Bosen et al., 2020; Nittrouer & Lowenstein, 2014). Environmental sounds, the other signals used in this study, have already been shown not to demonstrate a recency effect, which is commonly attributed to their lack of articulatory attributes (de Gelder & Vroomen, 1992; Nittrouer & Lowenstein, 2014). With the addition of movement in the audio-video condition (a talking face or the object producing the environmental sound), a recency effect should be evoked for one or both of the vocoded speech and environmental sounds stimuli, depending on the true basis of that effect. If it is specifically the dynamic component of articulation that is responsible for the recency effect then both sets of stimuli should show the effect in this condition. If instead it is the case that stimuli need to possess essential articulatory structure to evoke a recency effect, then only the vocoded speech stimuli should show a novel recency effect in the audio-video condition. This outcome would be expected because the addition of visual information in the form of the lipread signal should provide the same articulatory structure that is available through the auditory channel in the unprocessed speech signal, strengthening the phonological quality of the stimuli. That is, the sum of the vocoded signal plus the lipread signal should equal what is available through unprocessed acoustic speech signals.

The second goal addressed with this work was to examine the interaction of signals across modalities. In particular, the question was asked of whether signals need to be of the

same nature in order to enhance recognition, or can static, nonspeech visual information enhance serial recall simply due to the additional information it provides. This goal was addressed by comparing outcomes for the audio-video and the audio-picture stimuli, but unlike Aim #1 the main dependent measure was *overall* accuracy, rather than just the recency effect. Specifically the question was asked if accuracy across all list positions was differentially affected by the presentation of audio-video versus audio-picture stimuli. In order to make interpretation under this aim as transparent as possible, a decision tree was developed, as shown in Figure 3. In the first step (top) it would be determined if the addition of any form of visual information enhanced overall performance. If the addition of any kind of visual signal was found to improve overall recall, the second step would be to ascertain if that visual signal needed to be dynamic in nature to evoke its effect, or if static pictures could have the same effect. If it was found that the visual signal did not need to be dynamic, it would mean that any visual information is facilitative; it does not need to be strictly phonological or dynamic in nature. However, if gains in performance were obtained only for the video condition that outcome would mean that the visual information needs to be dynamic. In this case the third and final step in the decision-making process would be to ascertain if gains in the audio-video condition were observed only for the lipread signals and only for the vocoded speech signals, not for the environmental sounds. If so that outcome would suggest that it is only visual signals that are dynamic and articulatory in nature that can be integrated across modalities. An assumption in all of this work was that the unprocessed speech would provide sufficient articulatory structure, so the addition of lipread signals would be only redundant.

Finally, the third goal of the current study was to investigate whether dynamic and static visual information affect cognitive effort for processing these signals differently. For this goal, response time for reporting serial recall was measured. If visual information facilitates encoding and storage of stimuli in the working memory buffer then response times should decrease with the addition of visual information. This outcome would be expected especially for the addition of the lipread signal in the case of vocoded stimuli, because the information provided by that signal

should be precisely the same as that provided by acoustic speech signals in that both provide articulatory structure. For the unprocessed speech stimuli the lipread signal should be largely redundant, because that acoustic signal is clear (thus minimizing the potential contribution of visual lipread information). It is harder to predict outcomes for the audio-picture condition overall or for when environmental sounds serve as stimuli. It could be that a picture interferes with the processing of an acoustic stimulus, at least if these signals are processed in a general-storage component, such as that proposed by Cowan. Unlike the videos, these pictures are not related to the acoustic signals in an essential way, so might not bind as readily. If there is interference, response times should increase. Similarly, if visual information is not readily integrated with nonspeech acoustic signals, response times might increase when either dynamic or static visual signals are presented with environmental sounds.

These three goals were examined primarily with a main experiment, but a follow-up experiment was also planned. This additional experiment was designed to address a concern that could arise regarding the audio-video condition with vocoded speech. The expectation was that the addition of the visual speech signal (i.e., lipread information) would improve working memory performance by enhancing the phonological quality of the vocoded signal; this effect was predicted to manifest primarily as a stronger recency effect. In the main experiment, the audio-visual condition was always presented last – after the audio-only condition – to ensure that any such enhancement of phonological quality would not extend to the audio-only condition. However, listeners gradually acquire a stronger phonological representation for vocoded signals through sustained exposure alone. That means that the possibility could not be ruled out that if a stronger recency effect were observed for the vocoded signal in the audio-video condition it could be due to exposure to vocoded speech in the audio-only condition. To address this concern, a follow-up study was planned in which vocoded signals would be presented in the audio-video condition first – if indeed the main experiment revealed an enhanced recency effect for vocoded signals in the audio-video condition.

In summary, three effects were examined in the current study: (1) the recency effect that is proposed to arise for articulatory signals only; (2) improvement in overall recall accuracy for acoustic signals that may arise with the addition of visual information; and (3) response time as an metric of cognitive effort. The data to be collected could extend our understanding of how well multicomponent and general-storage models explain working memory for stimuli presented across modalities, as well as inform interventional strategies for patients with hearing loss by defining the visual stimuli that can best support working memory in the face of degraded acoustic signals.

MAIN EXPERIMENT

Methods

Subjects

Eighty adults between the ages of 18 and 35 years participated. The mean age of subjects was 22 years (*SD* = 3 years). All subjects were native speakers of American English, an inclusionary criterion implemented to help ensure that vocabulary items would be equally probable across subjects and that sensitivity to phonological structure in these items would be equivalent across subjects, as well. None of the subjects reported any history of speech, language, or hearing disorder, and all had normal or corrected normal vision. None of the subjects reported having heard vocoded speech stimuli before.

Forty-one percent of the subjects were male, and were equally distributed across the two groups. Subjects were given the reading subtest of the Wide Range Achievement Test – 4 (WRAT; Wilkinson & Robertson, 2006) and needed to demonstrate better than a 12th-grade reading level to have their data included. This brief test was an additional way to help ensure that subjects had reasonable phonological sensitivity: The WRAT presents words in isolation for reading, so performance is dependent on sensitivity to word-internal phonological structure. Subjects also needed to pass hearing screenings consisting of the pure tones 0.5, 1.0, 2.0, 4.0, and 6.0 kHz presented at 20 dB hearing level to each ear separately.

Equipment

All testing took place in a sound-treated booth. Hearing was screened with a Welch Allyn TM262 audiometer using TDH-39 headphones. Stimuli were stored on a computer server and presented through a Creative Labs Soundblaster soundcard, Samson C-Que 8 headphone amplifier, and AKG-K141 headphones.

A 22-inch wide touchscreen monitor (HP E220t) presented the videos and the line drawings that represented the words and environmental sounds, and served to collect responses. For generation of the word stimuli, video samples (with sound) were recorded using a Sony HDR-XR550V video recorder and a Sony FM transmitter microphone. The receiver for this FM system was connected directly to the audio input of the video recorder.

Subjects' responses to the WRAT test instrument were recorded using the same Sony HDR-XR550V video recorder and Sony FM transmitters to ensure good sound quality on the recordings. These recordings were used for offline scoring of the WRAT.

Stimuli

Three sets of acoustic stimuli were created: unprocessed speech stimuli, vocoded versions of those stimuli, and environmental sounds. All stimuli were newly created versions of ones used previously (e.g., Nittrouer & Lowenstein, 2014), and all were between 600 ms and 700 ms in length. Appropriate test-retest reliability has been found for these materials and these procedures (e.g., Nittrouer & Miller, 1999). The speech stimuli consisted of eight non-rhyming nouns that can be transparently represented with pictures: *ball, cub, dog, ham, pack, rake, soap,* and *teen*. The vocalic nuclei of these words spanned the range of the vowel quadrilateral in terms of first and second formant frequencies, making them optimally distinguishable from one another on the basis of vowel identity. The nonspeech stimuli consisted of eight sounds that would be familiar to most people: a bird chirping, a hand drill, glass breaking, a helicopter, knocking on a door, a single piano note (one octave above middle C), a sneeze, and a sports whistle being blown.

The speech stimuli were video recorded in a sound booth, using a 44.1-kHz sampling rate with 16-bit digitization for audio recording, and a 3600-kbps sampling rate for video recording. The talker was a man with no interfering facial hair and a fundamental frequency of roughly 100 Hz. He sat on a chair with the back of his head against a cushion positioned between his head and the wall behind him to ensure that his head remained still. He produced the words one at a time, each with a falling inflection. His lips began in a closed position at the start of each word and returned to a closed position at the end of each word. In this way the sequences of eight words appeared as one continuous string when presented for testing, regardless of order. The camera was zoomed so that the talker's face filled the frame vertically. Five repetitions of each word were recorded, and subsequently edited into separate files. The tokens that best matched selected tokens of the other words in duration, fundamental frequency, and intonation were used. The total length of each file was 1300 ms, with the actual acoustic waveform occupying the middle 600-700 ms. These files served as the unprocessed audio-video stimuli. The audio track of each file was saved separately, and presented as the unprocessed audio-only stimuli, with 1300-ms onset to onset.

The vocoded stimuli were created from the audio track of each word file. A MATLAB routine was used to create 4-channel vocoded stimuli. All signals were first band-pass filtered with a high-frequency cutoff of 8000 Hz and a low-frequency cutoff of 250 Hz. Next that filtered signal was divided into four channels, with boundaries at 709 Hz, 1676 Hz and 3712 Hz, which were based on the Greenwood (1990) function for frequency-place maps along the basilar membrane. All filtering used in the generation of these stimuli was done with digital filters that had greater than 50-dB attenuation in stop bands, and had 1-Hz transition bands between pass- and stop-bands. Each channel was half-wave rectified and filtered below 20 Hz to remove fine structure. The temporal envelopes derived for separate channels were subsequently used to modulate white noise, limited to the same channels as those used to divide the speech signal. The resulting bands of amplitude-modulated noise were combined with the same relative amplitudes across channels as measured in the original speech signals. Root-mean-square

amplitude was equalized across all stimuli. These vocoded stimuli were presented by themselves in the audio-only condition, and the audio files were combined with each corresponding video file to create the vocoded audio-video stimuli. Care was taken to realign the vocoded signal with the video signal precisely. Again, onset-to-onset was 1300 ms.

For the environmental sounds, the stimuli from Nittrouer and Lowenstein (2014) were used, and were combined with videos of the actions that generate these sounds to create the audio-video stimuli. For example, the sound of a bird chirping was combined with a video of a bird chirping and the knocking sound was combined with a video of someone knocking on a door. Specific components of the sounds were aligned with corresponding components of the videos. All files were 1300 ms in length, matching the length of files for the speech stimuli, but again the acoustic waveforms occupying those files were 600-700 ms in length. For example, the whistle stimulus began with a person raising the whistle to his mouth, the whistle was seen and heard being blown for the middle portion of the video, and the last portion showed the whistle being lowered.

Audio-picture stimuli were created for each speech (unprocessed and vocoded word) and environmental sound stimulus. To do this, the same picture used to represent the word or sound in the response procedure was combined with the audio file so that the picture displayed during the entire 1300-ms stimulus.

In summary, nine sets of stimuli were created from one set of words and one set of environmental sounds. These sets consisted of the three types of stimuli (unprocessed speech, vocoded speech, and environmental sounds) crossed with the three conditions (audio only, audio-video, and audio-picture).

Procedures

All procedures were approved by the local Institutional Review Board. Subjects were tested in a single session that lasted about 60 minutes. Subjects were divided into two groups, with half of them tested in the audio-video condition and the other half tested in the audio-

picture condition; all were tested in the audio-only condition. Thus subjects in each group were tested in a 2 x 3 repeated-measures design: 2 conditions (audio-only and one of the audio-visual conditions) x 3 stimulus types (unprocessed speech, vocoded speech, and environmental sounds).

After obtaining informed consent, the hearing screening was administered. Next, baseline response times were measured as a way to obtain an index of general motoric response speeds. Differences in the time it takes to tap on the pictures showing on the computer monitor could affect analysis of response times. To assess possible differences in these response speeds, a series of eight blue squares were arranged across the top of the computer monitor with equal spacing. Participants were instructed to keep the hand that they would be responding with flat on the table until the pictures appeared. Then they needed to tap on the blue squares in order from left to right as quickly as possible. Response time was the interval between when the pictures appeared and when the last tap was issued. This task was performed five times, and the mean time across the five trials was used as the baseline response time.

All subjects were tested first in the audio-only condition. That was followed by administration of the reading subtest of the WRAT. Then subjects were tested in the second condition, either audio-video or audio-picture, depending on group. Order of audio-only and audio-visual presentation was not randomized because of evidence that subjects can use information about a talker gathered from lipreading that talker to support later processing of speech from that talker presented in the audio-only modality (Rosenblum et al., 2007).

Order of presentation of stimulus type was randomized across subjects within both the audio-only and audio-visual conditions. Environmental sounds were always presented as the second set in the trilogy, with the unprocessed and vocoded speech stimuli evenly distributed across subjects as either the first or third set presented. If a subject heard the unprocessed stimuli as the first set in the audio-only condition, then these stimuli were presented as the third set in the second condition (either audio-video or audio-picture). Stimuli were presented at a

peak intensity of a 68 dB sound pressure level, and the touchscreen monitor was positioned directly in front of the subject, 9 inches from the edge of the table.

Software written in Visual Basic controlled the presentation of stimuli and collection of responses. This software randomized the order of words or environmental sounds across the ten lists presented in each condition for each type of stimulus. The software kept track of the order of presentation, collected data on the order recalled by the subject, and provided a summary of the numbers of errors made for each list position across the ten lists. It also computed the time between appearance of the pictures on the screen and the subject's last tap on a picture.

Before testing in each condition commenced, subjects were trained to match the response pictures to the stimuli. This was done by displaying all eight response pictures in a set across the top of the monitor, and presenting each stimulus in isolation. Subjects needed to tap on the picture representing each stimulus after it was presented. During the first time through, subjects were corrected if they tapped on the wrong picture. Next this process was then repeated without feedback to ensure that subjects had all correct matches. Subjects' abilities to match stimuli to pictures were checked before and after testing with each stimulus set to ensure accurate matches throughout testing. This ensured that any incorrect responses in the serial recall task could not be due to a subject failing to know which picture matched each stimulus.

During testing in each condition, the eight words or sounds were played with an onset-to-onset interval of 1300 ms. Pilot testing had revealed that the usual rate of one per 1000 ms was too fast in the audio-video condition. For the audio-video and audio-picture conditions, the videos or pictures were shown in the middle of the monitor in a roughly 4" x 4" display during stimulus presentation. After presentation of a stimulus list, the response pictures (2" x 2" displays) immediately appeared across the top of the monitor in random order, signally subjects to respond by tapping on the pictures in the order recalled. As each image was touched, it dropped to the vertical center of the monitor into the next position going from left to right. The

order of pictures could not subsequently be changed. Ten lists were presented in each condition with each stimulus type.

Subjects were instructed to keep their hands flat on the table in front of the monitor during stimulus presentation. They were told to keep their mouths closed so there could be no articulatory movement of any kind (voiced or silent) during presentation of the words or sounds, or between hearing the words or sounds and tapping the pictures.

After testing, the software automatically compared the order in which words or sounds were recalled with the order actually presented for each stimulus condition. A response was considered wrong if the word or sound was recalled in the wrong list position. Mean response time (from appearance of pictures to the last tap) across the ten lists presented for each stimulus type in each condition was used to index cognitive effort.

RESULTS

All statistical analyses were conducted using SPSS Software (Version 25). In all analyses, a traditional alpha of .05 was used, although exact *p* is reported when $p < .10$; for $p > .10$ results are reported simply as *not significant*, or *ns*.

Preliminary analyses

Data for both accuracy of serial recall and response times were screened for homogeneity of variance and normality of distribution in all nine sets of stimuli. All data were found to meet these criteria. Accordingly, parametric statistics were used in all analyses. With 40 subjects in each group, power was 95% for finding a group difference of 0.80 with an alpha of 0.05.

The next analyses performed for this experiment were meant to ensure there were no fundamental differences between the two groups of subjects: those who were presented with the audio-video stimuli as the second condition and those who were presented with the audio-picture stimuli as the second condition. To that end, percent correct scores across the eight list

positions and the ten trials for the audio-only conditions were examined. Table 1 shows these scores for each group, and *t* tests were performed to examine potential mean differences across the two groups. None of the three comparisons was statistically significant, $p > .10$ for all, so it can be concluded there were no differences between the two groups on overall accuracy.

Next the two groups were compared on basic response time by comparing the time it took to tap on the eight blue squares. Mean times were 2.0 s (SD = 0.48 s) for the audio-video group and 1.9 s (SD = 0.34 s) for the audio-picture group. This difference was not statistically significant ($p > .10$), so it was concluded that subjects in the two groups had equivalent motor response speeds.

Recency Effect

The next measure examined in this analysis was the recency effect, with the primary goal of comparing this effect in the audio-video and audio-picture condition, especially for the vocoded speech and environmental sounds. The expectation was that the unprocessed speech stimuli would show the strongest recency effect, in all conditions. The environmental sounds were predicted to show no recency effect, or at most only a weak recency effect, regardless of condition. It was the vocoded speech stimuli that were expected to show the greatest change across conditions. The recency effect was predicted to be weak for these stimuli in the audio-only condition, because these stimuli would not be as strongly phonological in nature as the unprocessed stimuli. The addition of the pictures upon presentation were not expected to affect this quality of these stimuli. However, the addition of the visual, lipread signal was expected to enhance the phonological quality of the stimuli, resulting in a much larger recency effect for the audio-video condition.

Figure 4 displays results for serial recall across stimulus types, conditions, and groups. Responses from subjects in the audio-video group are shown in the top of the figure, and responses from subjects in the audio-picture group are shown in the bottom of the figure. Responses for both groups in the audio-only condition, shown on the left of the figure, support

predictions: Strong recency effects are seen for the unprocessed speech stimuli. Weaker recency effects are apparent for the vocoded speech stimuli, and even weaker effects are apparent for the environmental sounds. Looking at the right panels displaying results for the second condition, these patterns of recency effects remain unchanged for the audio-picture condition (bottom-right panel), but are changed for the audio-video condition (top-right panel). For this condition there is a clear gain in the magnitude of the recency effect for the vocoded stimuli, but not for the environmental sounds. These trends match predictions.

Figure 5 provides another illustration of this finding. In order to quantify the magnitude of the recency effect, it is necessary to have a baseline against which to compare recall accuracy in that last list position. Although the recency effect is defined by performance on the last list position, the last few positions can show some effects, so they cannot be used as a baseline. Therefore, the mean of the two middle positions (4 and 5) was used as the baseline, and the recency effect was operationally defined as the improvement in recall for the last list position compared to that baseline (position 8 – positions 4/5). Figure 5 clearly shows that for unprocessed speech (left-most panel) the recency effect is large and remains unchanged in magnitude across groups (audio-video or audio-picture, squares or circles, respectively) and conditions (audio-only or audio-visual, open or filled symbols, respectively).

For environmental sounds (right-most panel), there appears to be no (or little or backwards) recency effect, regardless of group or condition. To examine that impression more carefully, paired-sample *t*-tests were done comparing recall accuracy for the last list position and the baseline (position 8 versus positions 4/5). None was found to be significant for these environmental sounds; that is, there was no difference in recall accuracy for the last position, compared to the mean of the middle two positions. Thus, there was no recency effect for environmental sounds, in either condition or for either group. (In contrast, all *t*-tests comparing recall accuracy for the last list position and the baseline done on unprocessed and vocoded speech signals (left and middle panels) were significant, $p < .001$, indicating that listeners

recalled items for the last position more accurately than for the middle positions in all cases that involved speech-like stimuli.)

For the vocoded speech stimuli (middle panel), however, there is an obvious change in the magnitude of that recency effect across conditions and groups. In the audio-visual condition for the audio-video group (red, filled squares) the recency effect is greater in magnitude than that observed for the audio-only condition (red, open squares) and greater than that observed in either condition for the audio-picture group (black circles). That observation is supported by the fact that when paired-sample *t*-tests were done comparing recency effects (position 8 – positions 4/5) across conditions (audio-only versus the second condition), only listeners in the audio-video group showed a significant effect, and only for vocoded stimuli, $t(39) = 2.428$, $p = 0.020$; this difference remained significant when a Bonferroni correction was applied. That is, the difference represented by the red, filled squares is greater in magnitude than the difference represented by the red, open squares. Thus, only listeners in the audio-video group showed an enhanced recency effect for the second (audio-visual) condition, and only for the vocoded stimuli.

Overall Performance

The second purpose of this experiment was to examine whether dynamic and static visual information affect serial recall differently, and whether visual signals that are explicitly articulatory in nature would have an especially large effect on recall accuracy overall. The expectation going into this study was that signals of an articulatory nature would likely have a disproportionately large effect. As already seen, these signals were the only ones to evoke a novel recency effect for the degraded acoustic speech signal. This finding suggests that lipread signals provide precisely the same articulatory structure as clear acoustic speech signals. Thus, the addition of lipread signals should boost the phonological qualities of these signals, enhancing coding and storage of these materials in a short-term memory buffer. More generally,

outcomes for static visual signals and for environmental sounds could inform conclusions regarding requirements for binding of acoustic and visual information, as delineated in Figure 3.

Figure 6 shows overall recall accuracy, for both groups of listeners. Here the accuracy across all eight positions for ten list presentations was the dependent measure. The most apparent observation from this figure is that increases in recall accuracy were obtained for the second condition compared to the audio-only condition, regardless of whether that second condition involved dynamic visual (audio-video) or static visual (audio-picture) information. Furthermore, it appears that the magnitude of these effects were roughly equivalent for the vocoded speech and environmental sounds, with perhaps some diminishment for the unprocessed speech stimuli. To examine these effects, a three-way, repeated-measures ANOVA was performed, with group as the between-subjects factor and condition and stimulus type as the within-subjects factors. A significant main effect of condition would indicate that the addition of visual signals affected overall accuracy of serial recall. Because the initial analysis revealed no differences between these two groups in their recall accuracy for the audio-only condition, a significant group effect would indicate that dynamic or static visual information had a greater influence on recall: One or the other group would need to perform significantly better on that second condition to evoke a significant main effect of group. A significant Group x Condition interaction could indicate that the addition of different visual signals had different effects on recall, such as one showing improvement in recall and the other showing a decrement in recall or no change at all.

Outcomes of this ANOVA are shown in Table 2. Both within-subjects factors were found to be significant: condition and stimulus type. These results indicate that subjects displayed better serial recall overall when visual information was available (mean correct = 59.2%, *SD* = 12.5%) than when it was not (mean correct = 50.2%, *SD* = 10.7%), and displayed differences across stimulus types, with the best overall performance for unprocessed speech (mean correct = 60.4%, *SD* = 12.6%), followed by vocoded speech (mean correct = 53.6%, *SD* = 12.5%), and finally by environmental sounds (mean correct = 50.1%, *SD* = 12.5%). Paired-sample *t*-tests

showed that each of these comparisons was significant; $p < .001$ for the comparisons of unprocessed speech with each of the other stimulus types; $p = .005$ for the comparison of vocoded speech with environmental sounds. All comparisons were significant when Bonferroni corrections were applied. The main effect of group was not significant, indicating that there were no differences in performance for the group that was presented with the dynamic visual information and the group that was presented with the static visual information.

The only significant interaction was Condition x Type, reflecting the fact that increases in recall accuracy arising from the addition of visual information (i.e., condition 2) differed across the types of stimuli: these increases were larger for vocoded speech (mean change = 10.3%, *SD* = 13.3%) and environmental sounds (mean change = 12.0%, *SD* = 11.9%) than for unprocessed speech (mean change = 4.8%, *SD* = 13.0%). To examine this interaction more closely, difference scores were computed across conditions (condition 2 – audio-only) to quantify the magnitude of the condition effect for each stimulus type. Using those difference scores, paired-samples *t*-tests were done, comparing each stimulus type to each other type. Results revealed that the condition effect for unprocessed stimuli was significantly smaller than that of both vocoded speech and environmental sounds: $t(79) = 2.53$, $p = .013$ and $t(79) = 3.87$, $p < .001$, respectively. Both results remained significant when Bonferroni corrections were applied. Thus, the source of the Condition x Type interaction was a diminished effect for unprocessed speech, relative to the other two types of stimuli. Recall was more accurate in the audio-only condition for these unprocessed stimuli than for vocoded speech or environmental sounds. Accordingly, the visual signal contributed less in the second condition for the unprocessed stimuli.

Of course, the measure of overall accuracy included responses for the late list positions, so the question arises as to whether there is improvement in recall that is independent of the recency effect. The advantage for the most recently presented list items can extend backwards from the final position for several positions. Therefore, recall accuracy for the first four list positions only was examined as a way to verify the trends observed across entire lists. If similar

trends were observed for just the initial positions stronger support would be provided for the conclusion that both kinds of visual information contribute similarly to recall accuracy. Figure 7 displays mean recall accuracy across these initial positions, and reveals similar results to those displayed in Figure 6 for whole lists. Overall accuracy, however, is found to be generally higher when only the first half of the lists are considered, as is done here in Figure 7. A three-way, repeated-measures ANOVA identical to the one described above was performed on these values. Results of this ANOVA are shown on Table 3, and reveal the same trends as those observed for the full lists: The main effects of condition and type are again significant, indicating better recall when visual information was available (mean correct = 72.2%, *SD* = 14.3%) than when it was not (mean correct = 62,6%, *SD* = 13.4%), and the best overall performance for unprocessed speech signals (mean correct = 71.1%, *SD* = 13.4%), compared to vocoded speech (mean correct = 65.2%, *SD* = 14.8%) and environmental sounds (mean correct = 65.9%, *SD* = 16.5%). Paired-sample *t*-tests showed that the comparisons of unprocessed speech and each of the other stimulus types were significant; $p < .001$ in both cases, which remained significant with Bonferroni corrections. The comparison of vocoded speech and environmental sounds was not significant.

A significant two-way interaction for Condition x Type was obtained again. Difference scores were again computed across conditions (condition 2 – audio-only), and paired-sample *t*-tests done using those difference scores, comparing each stimulus type to each other type. In this case results showed significant differences for all three comparisons: for unprocessed versus vocoded speech, $t(79) = 2.13$, $p = .037$; for unprocessed speech versus environmental sounds, $t(79) = 4.67$, $p < .001$; and for vocoded speech versus environmental sounds, $t(79) = 2.29$, $p = 0.25$. When Bonferroni corrections were applied, only the comparison of unprocessed speech versus environmental sounds remained significant. Nonetheless, outcomes of this additional analysis for only the early list items largely replicate outcomes for the entire lists: Recall accuracy was generally better for unprocessed speech than for either vocoded speech or environmental sounds, and recall accuracy was better in the audio-visual than in the audio-only

condition. The unprocessed stimuli showed smaller benefits from the addition of visual signals, but that reflects the fact that recall was highest for these stimuli in the audio-only condition.

In summary, analyses of these results for recall accuracy across all list positions and for the first four positions indicate that both kinds of visual information enhanced serial recall to a similar extent. Regardless of whether the visual signal was dynamic or static, that visual information led to greater improvements in recall accuracy for the vocoded speech and the environmental sounds than for the unprocessed speech.

Cognitive Effort

Response times were used to assess cognitive effort. Although the preliminary analysis showed that there was no difference between groups in basic response times, it was still possible that the time it took for subjects to perform this motor response could influence analyses of response times for serial recall. Therefore, Pearson product-moment correlation coefficients were computed for basic response times and time for completing the serial recall task for each stimulus type, for each condition and group separately. None of these twelve correlation coefficients was significant, and it was concluded that basic response times would not influence outcomes of statistical analyses for times to complete the serial recall task.

Mean response times are shown in Figure 8 for each stimulus type x condition, for each group separately. The most apparent outcomes are that all visual stimuli led to decreases in response times, indicating less cognitive effort being exerted, except for videos of the environmental sounds being produced. A three-way, repeated-measures ANOVA was performed on these response times, and outcomes are shown in Table 4. Inspection of this table reveals that a significant three-way interaction was obtained, likely due to the fact that environmental sounds alone failed to show a decrease in response time, and only for the audio-video signals. In order to test this suggestion, separate two-way, repeated-measures ANOVAs were performed for each stimulus type. The expectation was that results for the unprocessed and vocoded speech stimuli would show a significant main effect of condition only, indicating

that response times were shorter for the audio-visual stimuli. A lack of a significant group effect would indicate audio-video and audio-picture stimuli had similar effects. It would be a significant Group x Condition interaction that would indicate that the pattern of response times across conditions was different for the two groups, and only for environmental sounds was a significant Group x Condition interaction predicted.

Outcomes of these ANOVAs are shown in Table 5, and match impressions from Figure 8: Results for the unprocessed and vocoded speech signals showed a significant main effect of condition only, meaning that response times were shorter when visual signals were present, regardless of whether they were dynamic (audio-video group) or static (audio-picture group). Environmental sounds showed a significant Group x Condition interaction, meaning that the effect of adding a visual signal differed for dynamic and static visual signals: response times were longer when a dynamic visual signal was added and briefer when a static visual signal was added. These findings support the general conclusion that the addition of any kind of visual signal – dynamic or static – to an acoustic speech signal facilitated working memory for those speech signals, such that listeners expended less cognitive effort. In this case, environmental sounds showed a different pattern of results. For these stimuli, only static visual signals were found to reduce the cognitive effort expended in working memory operations; that is, only listeners in the audio-picture group had reduced response times with the addition of a visual signal. Dynamic visual signals interfered with working memory operations, such that listeners in the audio-video group needed to expend more cognitive effort when a visual signal was present, as indicated by their longer response times.

FOLLOW-UP EXPERIMENT

Results of the main experiment revealed a significant increase in the magnitude of the recency effect for the vocoded speech stimuli when presented in the audio-visual condition, compared to the audio-only condition. This outcome was interpreted as demonstrating that the

visual speech signal contributes precisely the same kind of articulatory information to the listener as the acoustic speech signal, thus enhancing the phonological quality of the signal when the acoustic signal is degraded, as it is for vocoded speech. However, an alternative explanation is that subjects in the audio-video group came to perceive the vocoded signal as more phonological in nature, simply by exposure in the audio-only condition. In order to decide between these alternative explanations this follow-up experiment was conducted in which subjects were presented with vocoded speech stimuli in the audio-video condition before they heard them in the audio-only condition. If these new subjects showed a recency effect comparable to that observed for unprocessed speech signals it could be concluded that the visual speech signal indeed was enhancing the phonological quality of the signal.

But one potential confound existed with this follow-up experiment. We predicted that exposure to the vocoded speech stimuli in the audio-video condition would permanently enhance the phonological quality of these signals for these subjects, such that they would subsequently show an equivalent recency effect for vocoded speech stimuli in the audio-only condition. An alternative to that explanation, however, would be that subjects acquired a mental image of the talker producing these words and used that image when they were later presented with the vocoded speech stimuli in the audio-only condition. To address that concern, a new set of stimuli was added to this follow-up experiment: the same words produced by a female talker. If indeed the subjects were using a sustained mental image of the (male) talker from the original stimulus set that image should interfere with recall of stimuli from the female talker. If instead, however, it is the case that any stimuli that possess a phonological quality evoke a strong recency effect, that effect should be observed in similar magnitude for all stimulus sets, including those spoken by the female talker.

Methods

Subjects

Thirty subjects participated in this follow-up experiment. These subjects met all the same criteria as those in the main experiment, and had a mean age of 22 years (*SD* = 3 years).

Equipment

The same equipment was used in this experiment as in the main experiment.

Stimuli

Four sets of stimuli were used in this follow-up experiment. Three of these sets were from the main experiment: unprocessed speech in the audio-only condition; vocoded speech in the audio-only condition; and vocoded speech in the audio-video condition. One set of stimuli were added, and that consisted of the same words used for the unprocessed and vocoded speech stimuli, only spoken by a woman. These stimuli were presented as unprocessed speech in the audio-only condition. Unprocessed speech was used for this additional set of stimuli, rather than vocoded stimuli, because fundamental frequency is not well preserved in vocoded speech. Therefore, using unprocessed speech maximized the discrepancy between gender identity of the acoustic speech stimuli and what the listeners were presented with visually in the audio-video condition.

Procedures

Procedures were the same in this follow-up experiment as in the main experiment, except the order of stimulus presentation was modified. In this follow-up experiment, all subjects were presented with the vocoded speech stimuli in the audio-video condition first. Next they were presented with these vocoded speech stimuli in the audio-only condition. Lastly they were presented with the two sets of unprocessed speech stimuli. The order of presentation of male and female stimuli was randomized across listeners, such that half heard each order.

Results

Figure 9 displays results for all four sets of stimuli, and reveals that a recency effect was obtained for all sets, although it may have been slightly reduced for the vocoded speech stimuli. As in the main experiment, means of recall accuracy for the middle two positions (4 and 5) were used as baselines against which to compare recall accuracy for the final list position, in order to quantify the magnitude of the recency effect. Recall accuracy for these baselines and for the final list positions are shown in Figure 10. In this figure it appears as if the unprocessed speech stimuli spoken by the male talker and the vocoded speech stimuli in the audio-video condition evoked slightly stronger recency effects. To test that impression, paired-sample *t*-tests were done testing difference scores (position 8 – positions 4/5) for all combinations of stimulus sets. Only the comparison of the unprocessed (male) speech versus the vocoded speech in the audio-only condition was significant, $t(29) = 2.24$, $p = .033$, but it ceased to be significant when a Bonferroni correction was applied. Consequently, it can be concluded that all of these conditions elicited a similar recency effect. Therefore, the suggestion is supported that the visual speech (i.e., lipread) signal serves to enhance the phonological quality of degraded signals, rather than to create a mental image of a talker that provides a sustained effect.

DISCUSSION

The current study had three goals. First, this study was designed to examine whether visual speech (i.e., lipread) information influences serial recall for words because it provides explicitly articulatory structure of the same nature as acoustic speech signals, or because lipread signals are dynamic. This goal was accomplished by examining the magnitude of the recency effect across conditions. The second goal of the study was to assess the source of contributions, if any, made by the addition of visual information to overall working memory performance: Does visual information regardless of its nature aid the coding and storage of items in short-term memory? This question was addressed by examining overall recall accuracy across list positions. The third goal of the study was to examine whether one specific kind of

visual information (dynamic or static) best reduces cognitive load, and if that reduction in cognitive load would be found for both speech and nonspeech stimuli. This goal was accomplished by measuring response times.

Recency effect

The results of this study support the proposal that it is specifically the articulatory nature of the speech signal that results in the recency effect, and that articulatory structure can be derived from the visual as well as the acoustic speech signal. Moreover, the speech-relevant information available in the visual and auditory domains integrate seamlessly to evoke a single phonological percept. The bases of these conclusions rest largely on the finding that the recency effect was observed only for signals providing articulatory structure, and the more salient that structure was, the stronger the recency effect was found to be. Thus, the unprocessed speech signals showed large recency effects in all conditions (audio-only, audio-video, and audio-picture). The vocoded speech signals showed recency effects of reduced magnitude compared to the unprocessed speech, except in the condition that had visual speech information simultaneously available – the audio-video condition. Vocoded speech signals are highly degraded, so differ in substantive ways from the signals from which they were derived. These differences render vocoded speech more like nonspeech signals, as revealed by the diminished recency effect in the other conditions. With the addition of the lipread information, the combined signals regained sufficient phonological quality to evoke a stronger recency effect, one equivalent to that found for unprocessed speech. The follow-up experiment further demonstrated that once this enhanced phonological quality was obtained for vocoded signals it was sustained.

Overall performance

The second major finding of this study was that any additional visual information enhanced the overall performance of working memory to a similar extent. This conclusion is

based on the observation that the percent of items recalled in the correct position increased (compared to the audio-only condition) to a similar extent for both audio-visual conditions (audio-video and audio-picture) for all three types of stimuli (unprocessed speech, vocoded speech, and environmental sounds). When this general outcome is compared to the decision tree shown in Figure 3, it is found that progression is made to the second step where the question is asked of whether improved performance is obtained only for *dynamic* visual information. The answer is a resounding *no*, so the conclusion must be that any visual information can facilitate working memory.

Where the two sets of speech stimuli are concerned (unprocessed and vocoded) this conclusion was not necessarily anticipated. Instead it was considered likely that dynamic speech would exert a more robust effect than static pictures of the words being pronounced. The fact that both kinds of visual signals supported overall recognition to a similar extent contradicted that prediction. Nonetheless, the finding that a specific recency effect was evoked for the vocoded stimuli only when the visual signal was in the form of dynamic articulatory structure suggests that even though the effects may have been similar in magnitude for the dynamic and static visual information, they likely were based on different mechanisms. In the audio-video condition, it appears that dynamic articulatory information of the same nature was provided by both the visual and auditory signals. Accordingly, there was only a small improvement observed for the unprocessed speech stimuli when the video signal was added, but for the vocoded speech stimuli an enhanced recency effect was obtained. In the audio-picture condition, the static visual display appears to have added new information that was not articulatory in form. Consequently, the increase in recall accuracy was slightly greater for unprocessed speech stimuli in this condition than in the redundant audio-video condition. Effects across the two visual conditions were similar for the vocoded speech stimuli, even though they may have had different bases. Thus, in the audio-video condition the effect may best be described as one of *integration* of the audio and visual signals, very much like the effect traditionally termed the McGurk effect (McGurk & MacDonald, 1976) in which the audio-visual

phonetic percept is distinct from either the audio or visual input alone. In this case the integrated signal is most phonetically detailed, so can be robustly stored in a working memory buffer. For the audio-picture condition, the effect might best be described as one of *addition* of the two types of signals, whereby the signals remain distinct, but each contributes information for storage in a working memory buffer. When it comes to the environmental sounds, presentation in both the audio-video and audio-picture conditions might best fit the description of addition of informational sources.

Cognitive effort

With the exception of environmental sounds in the audio-video condition, response times decreased by similar amounts across all stimulus types, in both audio-visual conditions. This finding indicates that the provision of visual information reduced the cognitive effort associated with performing this task. The finding that there was a reduction in response time for environmental sounds when static visual signals were presented, but not when dynamic visual signals were presented suggests that even though the overall accuracy of serial recall improved by similar amounts across conditions for both groups, those improvements exacted different tolls on cognitive effort. In one case it was easy for listeners to combine acoustic and visual signals (when pictures were presented); in the other case it was more difficult (when dynamic visual signals were presented).

Additional findings

In addition to addressing these major questions, other findings were observed that inform our understanding of working memory in fundamental ways. The first additional finding that is of interest has to do with the notion of visual imagery. For most of the history of research into working memory, distinct response patterns – especially at the ends of lists – have been associated with modality of stimulus presentation (Conrad & Hull, 1968; Corballis, 1966; Cowan et al., 2004; Greene & Crowder, 1984; Grenfell-Essam et al., 2017; Spoehr & Corin, 1978). But

that assumption has recently been challenged. In a study by Guitard and Cowan (2020), phonologically dissimilar words that were either similar or dissimilar visually – meaning in terms of orthographic representation – were presented in an audio-only condition. These same words were also presented visually, in their orthographic form, with no audio presentation. Results showed that visual similarity affected recall accuracy, regardless of whether words were presented in the auditory or visual modality. Consequently it was concluded that subjects were accessing visual codes, even when they were being presented with only auditory stimuli. The subjects in that study were all adults, presumably with normal language and literacy. Thus, in an almost automatic manner they may have accessed both the phonological and orthographic representations of words when presented with them. That may not be the case for other subjects such as children or adults with dyslexia – or individuals with hearing loss. Future studies will need to answer this question for these populations.

Orthographic representations were not used in the current study. Nonetheless, it may be that the design of the current experiment promoted the use of visual codes by implementing a nonverbal response procedure in which subjects needed to touch pictures of the words named or of the items that would produce the sound heard. Pre-test training designed to familiarize subjects with the stimulus-response picture matches may have further encouraged the application of visual codes for use in the storage of these items in a short-term memory buffer, even in the audio-only condition. In particular, use of pictures as the response mode could have diminished the opportunity to observe a difference, if one existed, in the magnitude of visual contributions in the dynamic and static audio-visual conditions (i.e., audio-video and audio-picture). Again, future studies will have to examine this issue in more depth.

Another finding that provides useful information in regards to how visual codes are integrated with auditory stimuli concerns the greater improvement in overall performance observed for the unprocessed speech stimuli in the audio-picture condition, compared to the audio-video condition. Although not large, this finding suggests that in the audio-video condition the lipread information was redundant with the already phonologically rich auditory signal that

subjects heard. In the audio-picture condition, on the other hand, the picture provided additional information that subjects could combine with the auditory signal to generate stronger representations in short-term storage.

These findings have some implications for deciding whether a multicomponent or general-storage model of working memory best explains outcomes. Clearly a component akin to the phonological loop described by Baddeley can explain the finding that only stimuli of a phonological nature exhibited a recency effect, and the magnitude of that recency effect was related to how strongly phonological the signal structure was. However, a multicomponent model cannot explain the finding that any kind of visual signal strengthened recall of the acoustic signals, regardless of whether those signals were speech like or not. That finding is more effectively explained by a general-storage model of working memory, such as that of Cowan, because sensory inputs across modalities are processed in a single-purpose component.

Clinical Implications

The ability to recognize speech signals is typically the focus of diagnostic procedures. Outcomes of the current study suggest that language processing beyond recognition should be evaluated, in order to ensure that patients are receiving optimal sensory inputs. Although the subjects in this study were all able to recognize the words that served as stimuli in their vocoded form, these stimuli were not perceived as fully phonological in form, a situation that constrained their ability to store these words in working memory. Combining those vocoded speech stimuli with the visual speech signal, however, promoted the phonological quality of the signals, improving working memory operations. And that enhanced phonological quality was sustained, even when the subjects subsequently heard the stimuli in an audio-only format. These findings suggest that new recipients of cochlear implants should be afforded ample opportunity to hear speech through their new devices in an audio-visual format. This suggestion extends to children who receive cochlear implants.

Limitations of the Current Study and Future Directions

This study provided important observations for understanding the operations of working memory for auditory signals, especially speech signals. The predominant view has long been that speech is stored exclusively with phonological codes, which in turn derive from articulatory primitives. Results of this study challenge that exclusivity by showing that visual codes other than lipread signals can enhance the storage of speech in a short-term memory buffer. Of course, not all conditions were presented in the current experiment to construct a complete understanding of the interaction between auditory and visual codes in working memory for speech. For example, a response mode that did not utilize pictures would have provided more detailed evidence regarding whether audio-video and audio-picture presentation led to similar outcomes, because then static pictures would have been completely absent from the procedures used with the audio-video stimuli. The response method used here of tapping on pictures in the order recalled was implemented so that articulatory responses were not required, thus opportunities for rehearsal were greatly diminished, if not eliminated entirely. This same response method, however, may have facilitated the use of visual imagery in storage of items to be recalled. Furthermore, no information was gathered in the current study regarding how subjects would have performed with pictures only: Would subjects be found to invoke phonological codes for picture materials? This question critically needs to be examined because most working memory studies using visual stimuli have depended on orthographic symbols for presentation, and included subjects familiar with those orthographic symbols. Thus, the auditory and visual stimuli shared a common phonological base.

Finally, the finding that cognitive effort increased with the addition of visual information for one type of stimuli in one condition only – environmental sounds in the audio-video condition – raises questions about why that may have been. Does the integration of acoustic and visual information for nonspeech signals truly require more effort than either the integration of those information sources for a speech signal or the summation of acoustic and visual information for speech/nonspeech signals and pictures? One possible confounding factor in this case is that all

of the videos used to represent the environmental sounds had distinct backgrounds, unlike the videos used of the talker producing the speech stimuli in which a common background was used. Perhaps greater cognitive effort was required to process each new video that appeared every 1300 ms. To examine this question, a future study would need to generate videos for the audio-video condition that allowed seamless presentation across changes in stimuli.

Summary

In sum, the current experiment was undertaken to extend our collective understanding of visual contributions to a post-recognition process; namely, working memory for verbal (speech) material. This is an important area for research, because speech carries a heavy informational load, yet the signals are ephemeral. Listeners must be able to store early-arriving information in order to combine it with later-arriving information if the message intended by the talker is to be recovered. That task is especially difficult under some circumstances, such as if the listener receives only a degraded signal or if the acoustic environment is sub-optimal. The current study was undertaken to examine the mechanisms underlying visual contributions to working memory. Results of the this experiment showed that all visual information can be utilized to support coding and storage of speech signals in a short-term memory buffer, and this process is largely similar to that implemented for nonverbal signals. Nonetheless, lipread signals were found to provide a kind of support that appears different in kind from other visual signals. While dynamic visual signals can enhance coding and storage for both verbal and nonverbal signals, only lipread signals specifically strengthen the phonological quality of the signal. Overall, this information can help shape auditory rehabilitation for patients with hearing loss by emphasizing the benefits obtained with visual speech signals.

Acknowledgements

This work was supported by grant number R01 DC000633 from the National Institute on Deafness and Other Communication Disorders, the National Institutes of Health, to Susan Nittrouer, and with funding from the College of Public Health and Health Professions at the University of Florida. The authors are grateful to Shay Fitzgerald and Yi Yuan for help with data collection and to Harrison Walker for software development.

References

Aaronson, D. (1968). Temporal course of perception in an immediate recall task. *Journal of Experimental Psychology, 76*(1), 129-140. https://doi.org/10.1037/h0025290

Adams, E. J., Nguyen, A. T., & Cowan, N. (2018). Theories of working memory: Differences in definition, degree of modularity, role of attention, and purpose. *Language, Speech, and Hearing Services in Schools, 49*(3), 340-355. https://doi.org/10.1044/2018_LSHSS-17-0114

Baddeley, A., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *The Quarterly Journal of Experimental Psychology, 36*(2), 233-252. https://doi.org/10.1080/14640748408402157

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417-423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1-29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 47-89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Baddeley, A. D., & Hitch, G. J. (2019). The phonological loop as a buffer store: An update. *Cortex, 112*, 91-106. https://doi.org/10.1016/j.cortex.2018.05.015

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28-61). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.005

Bernstein, J. G., & Grant, K. W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 125*(5), 3358-3372. https://doi.org/10.1121/1.3110132

Bernstein, L. E., Auer Jr., E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*, 5-18. https://doi.org/10.1016/j.specom.2004.10.011

Bielski, L. M., Byom, L., Seitz, P. F., & Grant, K. W. (2020). Modality effects on lexical encoding and memory representations of spoken words. *Ear and Hearing, 41*(4), 825-837. https://doi.org/10.1097/AUD.0000000000000801

Bosen, A. K., Monzingo, E., & AuBuchon, A. M. (2020). Acoustic-phonetic mismatches impair serial recall of degraded words. *Auditory Perception & Cognition, 3*(1-2), 55-75. https://doi.org/10.1080/25742442.2020.1846012

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*(2), 201-251. https://doi.org/10.1017/S0952675700001019

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252. https://doi.org/10.1017/S0952675700000658

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology, 32*(1), 85-99. https://doi.org/10.1080/00335558008248235

Campbell, R., Dodd, B., & Brasher, J. (1983). The sources of visual recency: Movement and language in serial recall. *Quarterly Journal of Experimental Psychology Section A, 35*(4), 571-587. https://doi.org/10.1080/14640748308402147

Conrad, R., & Hull, A. J. (1968). Input modality and the serial position curve in short-term memory. *Psychonomic Science, 10*(4), 135-136. https://doi.org/10.3758/BF03331446

Corballis, M. C. (1966). Rehearsal and delay in immediate recall of visually and aurally presented items. *Canadian Journal of Psychology, 20*(1), 43-51. https://doi.org/10.1037/h0082923

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104*(2), 163-191. https://doi.org/10.1037/0033-2909.104.2.163

Cowan, N. (1999). An embedded-process model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory* (pp. 62-101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N., Saults, J. S., & Brown, G. D. (2004). On the auditory modality superiority effect in serial recall: separating input and output factors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(3), 639-644. https://doi.org/10.1037/0278-7393.30.3.639

Crowder, R. G. (1983). The purity of auditory memory. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 302*(1110), 251-265. https://doi.org/10.1098/rstb.1983.0053

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics, 5*, 365-373. https://doi.org/10.3758/BF03210660

de Gelder B., & Vroomen, J. (1992). Abstract versus modality-specific memory representations in processing auditory and visual speech. *Memory & Cognition, 20*(5), 533-538. https://doi.org/10.3758/bf03199585

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*(1), 223-234. https://doi.org/10.1037/0022-0663.100.1.223

Deutsch, D. (1972). Mapping of interactions in the pitch memory store. *Science, 175*(4025), 1020-1022. https://doi.org/10.1126/science.175.4025.1020

Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research, 15*(2), 413-422. https://doi.org/10.1044/jshr.1502.413

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America, 109*(5 Pt 1), 2272-2275. https://doi.org/10.1121/1.1362687

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108*(3 Pt 1), 1197-1208. https://doi.org/10.1121/1.1288668

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103*(5), 2677-2690. https://doi.org/10.1121/1.422788

Greene, R. L., & Crowder, R. G. (1984). Modality and suffix effects in the absence of auditory stimulation. *Journal of Verbal Learning and Verbal Behavior, 23*(3), 371-382. https://doi.org/10.1016/S0022-5371(84)90259-7

Greenwood, D. D. (1990). A cochlear frequency-position function for several species--29 years later. *Journal of the Acoustical Society of America, 87*(6), 2592-2605. https://doi.org/10.1121/1.399052

Grenfell-Essam, R., Ward, G., & Tan, L. (2017). Common modality effects in immediate free recall and immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(12), 1909-1933. https://doi.org/10.1037/xlm0000430

Guitard, D., & Cowan, N. (2020). Do we use visual codes when information is not presented visually? *Memory & Cognition*. https://doi.org/10.3758/s13421-020-01054-0

Jeong, E., & Ryu, H. (2016). Nonverbal auditory working memory: Can music indicate the capacity? *Brain and Cognition, 105*, 9-21. https://doi.org/10.1016/j.bandc.2016.03.003

Lachs, L., Pisoni, D. B., & Kirk, K. I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: a first report. *Ear and Hearing, 22*(3), 236-251. https://doi.org/10.1097/00003446-200106000-00007

Lalonde, K., & McCreery, R. W. (2020). Audiovisual enhancement of speech perception in noise by school-age children who are hard of hearing. *Ear and Hearing, 41*(4), 705-719. https://doi.org/10.1097/AUD.0000000000000830

Li, Y., & Cowan, N. (2021). Attention effects in working memory that are asymmetric across sensory modalities. *Memory & Cognition, 49*(5), 1050-1065. https://doi.org/10.3758/s13421-021-01142-9

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431-461. https://doi.org/10.1037/h0020279

Lyxell, B., Andersson, J., Andersson, U., Arlinger, S., Bredberg, G., & Harder, H. (1998). Phonological representation and speech understanding with cochlear implants in deafened adults. *Scandinavian Journal of Psychology, 39*(3), 175-179. https://doi.org/10.1111/1467-9450.393075

Lyxell, B., Andersson, J., Arlinger, S., Bredberg, G., Harder, H., & Rönnberg, J. (1996). Verbal information-processing capabilities and cochlear implants: implications for preoperative predictors of speech understanding. *Journal of Deaf Studies and Deaf Education 1*(3), 190-201. https://doi.org/10.1093/oxfordjournals.deafed.a014294

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748. https://doi.org/10.1038/264746a0

McKeown, D., Mills, R., & Mercer, T. (2011). Comparisons of complex sounds across extended retention intervals survives reading aloud. *Perception, 40*(10), 1193-1205. https://doi.org/10.1068/p6988

Murray, D. J. (1966). Vocalization-at-presentation and immediate recall, with varying recall methods. *Quarterly Journal of Experimental Psychology, 18*(1), 9-18. https://doi.org/10.1080/14640746608400002

Nairne, J. S., & Crowder, R. G. (1982). On the locus of the stimulus suffix effect. *Memory & Cognition, 10*(4), 350-357. https://doi.org/10.3758/bf03202427

Nairne, J. S., & Walters, V. L. (1983). Silent mouthing produces modality- and suffix-like effects. *Journal of Verbal Learning and Verbal Behavior, 22*(4), 475-483. https://doi.org/10.1016/S0022-5371(83)90300-6

Nees, M. A. (2016). Have we forgotten auditory sensory memory? Retention intervals in studies of nonverbal auditory working memory. *Frontiers in Psychology, 7*, 1892. https://doi.org/10.3389/fpsyg.2016.01892

Nittrouer, S., Caldwell-Tarr, A., & Lowenstein, J. H. (2013). Working memory in children with cochlear implants: problems are in storage, not processing. *International Journal of Pediatric Otorhinolaryngology, 77*(11), 1886-1898. https://doi.org/10.1016/j.ijporl.2013.09.001

Nittrouer, S., & Lowenstein, J. H. (2014). Separating the effects of acoustic and phonetic factors in linguistic processing with impoverished signals by adults and children. *Applied Psycholinguistics, 35*(2), 333-370. https://doi.org/10.1017/S0142716412000410

Nittrouer, S., & Miller, M. E. (1999). The development of phonemic coding strategies for serial recall. Applied Psycholinguistics, 20(4), 563-588. https://doi.org/10.1017/S0142716499004051

Peelle, J. E. (2018). Listening Effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing, 39*(2), 204-214. https://doi.org/10.1097/AUD.0000000000000494

Penney, C. G. (1975). Modality effects in short-term verbal memory. *Psychological Bulletin, 82*(1), 68-84. https://doi.org/10.1037/h0076166

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing, 37 Suppl 1*, 5S-27S. https://doi.org/10.1097/AUD.0000000000000312

Plyler, P. N., Lang, R., Monroe, A. L., & Gaudiano, P. (2015). The effects of audiovisual stimulation on the acceptance of background noise. *Journal of the American Academy of Audiology, 26*(5), 451-460. https://doi.org/10.3766/jaaa.14084

Rönnberg, J. (2003). Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: a framework and a model. *International Journal of Audiology, 42 Suppl 1*, S68-S76. https://doi.org/10.3109/14992020309074626

Rosenblum, L. D., Miller, R. M. & Sanchez, K. (2007). Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects. *Psychological Science, 18*(5), 392-396. https://doi: 10.1111/j.1467-9280.2007.01911.x

Rowe, E. J., & Rowe, W. G. (1976). Stimulus suffix effects with speech and nonspeech sounds. *Memory & Cognition, 4*, 128-131. https://doi.org/10.3758/bf03213153

Sali, A. W., & Egner, T. (2020). Declarative and procedural working memory updating processes are mutually facilitative. *Attention, Perception, & Psychophysics, 82*(4), 1858-1871. https://doi.org/10.3758/s13414-019-01887-1

Salter, D., & Colley, J. G. (1977). The stimulus suffix: A paradoxical effect. *Memory & Cognition, 5*(2), 257-262. https://doi.org/10.3758/BF03197371

Spoehr, K. T., & Corin, W. J. (1978). The stimulus suffix effect as a memory coding phenomenon. *Memory & Cognition, 6*(6), 583-589. https://doi.org/10.3758/bf03198247

Starr, G. E., & Pitt, M. A. (1997). Interference effects in short-term memory for timbre. *Journal of the Acoustical Society of America, 102*(1), 486-494. https://doi.org/10.1121/1.419722

Studdert-Kennedy, M. (1987). The phoneme as a perceptuomotor structure. In A. Allport, D. G. MacKay, W. Prinz, & E. Scheerer (Eds.), *Language perception and production: Relationships between listening, speaking, reading, and writing* (pp. 67-84). Academic Press.

Taitelbaum-Swead, R., & Fostick, L. (2017). Audio-visual speech perception in noise: Implanted children and young adults versus normal hearing peers. *International Journal of Pediatric Otorhynolaryngology, 92*, 146-150. https://doi.org/10.1016/j.ijporl.2016.11.022

Tye-Murray, N., Sommers, M., & Spehar, B. (2005). Speechreading and aging: How growing old affects face-to-face speech perception. *The ASHA Leader, 8-9*, 28-29. https://doi.org/10.1044/leader.FTR5.10092005.8

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory, 5*(6), 607-617. https://doi.org/10.1037/0278-7393.5.6.607

Walden, B. E., Prosek, R. A., & Worthington, D. W. (1974). Predicting audiovisual consonant recognition performance of hearing-impaired adults. *Journal of Speech and Hearing Research, 17*(2), 270-278. https://doi.org/10.1044/jshr.1702.270

Wilkinson, G. S., & Robertson, G. J. (2006). *The Wide Range Achievement Test (WRAT)* (4th ed.). Psychological Assessment Resources.

Wright, D., & Gagné, J. P. (2020). Acclimatization to hearing aids by older adults. *Ear and Hearing, 42*(1), 193-205. https://doi.org/10.1097/AUD.0000000000000913

Figure Legend

FIGURE 1: *Schematic of the Multicomponent Model of Working Memory.* Note. More recently an episodic memory component has been added to the classic version of the model, but it is not relevant to the current study.

FIGURE 2: *Classic U-shape Function for Serial Recall of Verbal Material.* Note. Data replotted from Nittrouer and Lowenstein (2014), from adult subjects for unprocessed speech, 8-channel vocoded speech, and environmental sounds.

FIGURE 3: *Decision Tree for Interpreting Outcomes Regarding Overall Recall Accuracy.*

FIGURE 4: *Recall Accuracy for Main Experiment.* Note. Recall accuracy shown for each list position for subjects in the audio-video group (top) and subjects in the audio-picture group (bottom). AO: Audio-only. AV: Audio-visual.

FIGURE 5: *Recency Effect for All Conditions and Stimulus Types in Main Experiment.* Note. Recall accuracy for the last list position (P8) is compared to mean recall accuracy across the middle two list positions (P4-P5) for each stimulus type (unprocessed speech, vocoded speech, and environmental sounds), each group (audio-video or audio-picture), and in each condition (audio-only [AO] or audio-visual [AV]). Results for vocoded speech for the audio-video group are highlighted in red.

FIGURE 6: *Overall Recall Accuracy across All Positions in Main Experiment.* Note: Mean recall accuracy summed across all list positions (8 positions x 10 trials) for each group (audio-video on left; audio-picture on right) for each stimulus type (unprocessed speech, vocoded speech, and environmental sounds) and in each condition (audio-only [AO] or audio-visual [AV]).

FIGURE 7: *Overall Recall Accuracy across First Four Positions in Main Experiment.* Note: Mean recall accuracy summed across the first four list positions (4 positions x 10 trials) for each group (audio-video on left; audio-picture on right) for each stimulus type (unprocessed speech, vocoded speech, and environmental sounds) and in each

condition (audio-only [AO] or audio-visual [AV]). Scale of this figure differs slightly from that of Figure 6.

FIGURE 8: *Mean Response Times in Main Experiment.* Note: Mean response times (seconds) for recalling item order for each group (audio-video on left; audio-picture on right) for each stimulus type (unprocessed speech, vocoded speech, and environmental sounds) and in each condition (audio-only [AO] or audio-visual [AV]).

FIGURE 9: *Recall Accuracy for Follow-Up Experiment.* Note: AO: Audio-only. AV: Audio-visual.

FIGURE 10: *Recency Effect for Follow-Up Experiment.* Note: Recall accuracy for the last list position (P8) is compared to mean recall accuracy across the middle two list positions (P4-P5). AO: Audio-only. AV: Audio-visual.

Table 1

Means and Standard Deviations for Percent Correct Recall for the Three Types of Stimuli in the Audio-Only Condition, for the Two Groups of Participants

Type of Stimulus	Audio-Video Group		Audio-Picture Group	
	M	*SD*	M	*SD*
Unprocessed Speech	58.1	*11.0*	58.1	*14.5*
Vocoded Speech	46.4	*14.9*	50.7	*12.5*
Environmental Sounds	42.8	*13.6*	45.4	*13.1*

Table 2

Results of a Three-Way, Repeated-Measures ANOVA for Percent Correct Recall across All List Positions

	df	*F*	*p*	η_p^2
Main Effects				
Condition	1,78	121.79	<.001	.610
Type	2,78	42.49	<.001	.353
Group	1,78	1.75	NS	---
Two-way Interactions				
Group x Condition	1,78	0.92	NS	---
Group x Type	2,78	0.44	NS	---
Condition x Type	2,156	6.98	.001	.082
Three-way Interaction				
Group x Condition x Type	2,156	1.39	NS	---

Table 3

Results of a Two-Way, Repeated-Measures ANOVA for Percent Correct Recall across the First Four List Positions

	df	*F*	*p*	η_p^2
Main Effects				
Condition	1,78	104.07	<.001	.572
Type	2,78	10.53	<.001	.119
Group	1,78	0.14	NS	---
Two-way Interactions				
Group x Condition	1,78	0.89	NS	---
Group x Type	2,78	0.45	NS	---
Condition x Type	2,156	10.08	<.001	.114
Three-way Interaction				
Group x Condition x Type	2,156	0.909	NS	---

Table 4

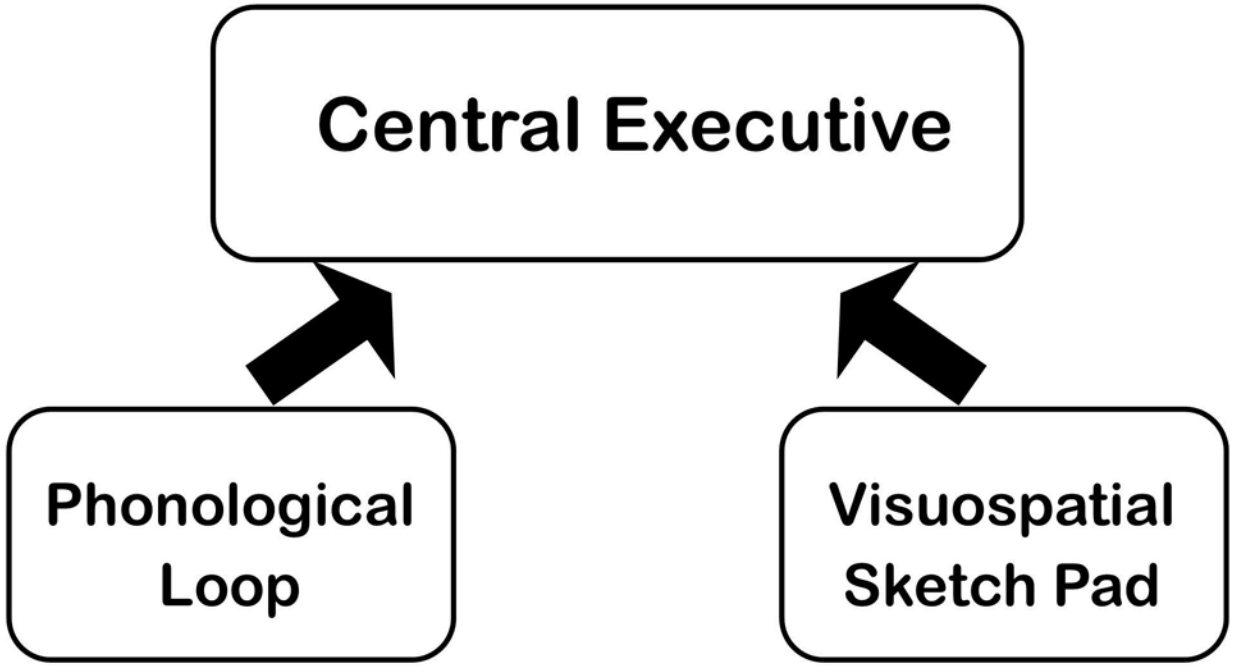
Results of a Three-Way, Repeated-Measures ANOVA for Response Times

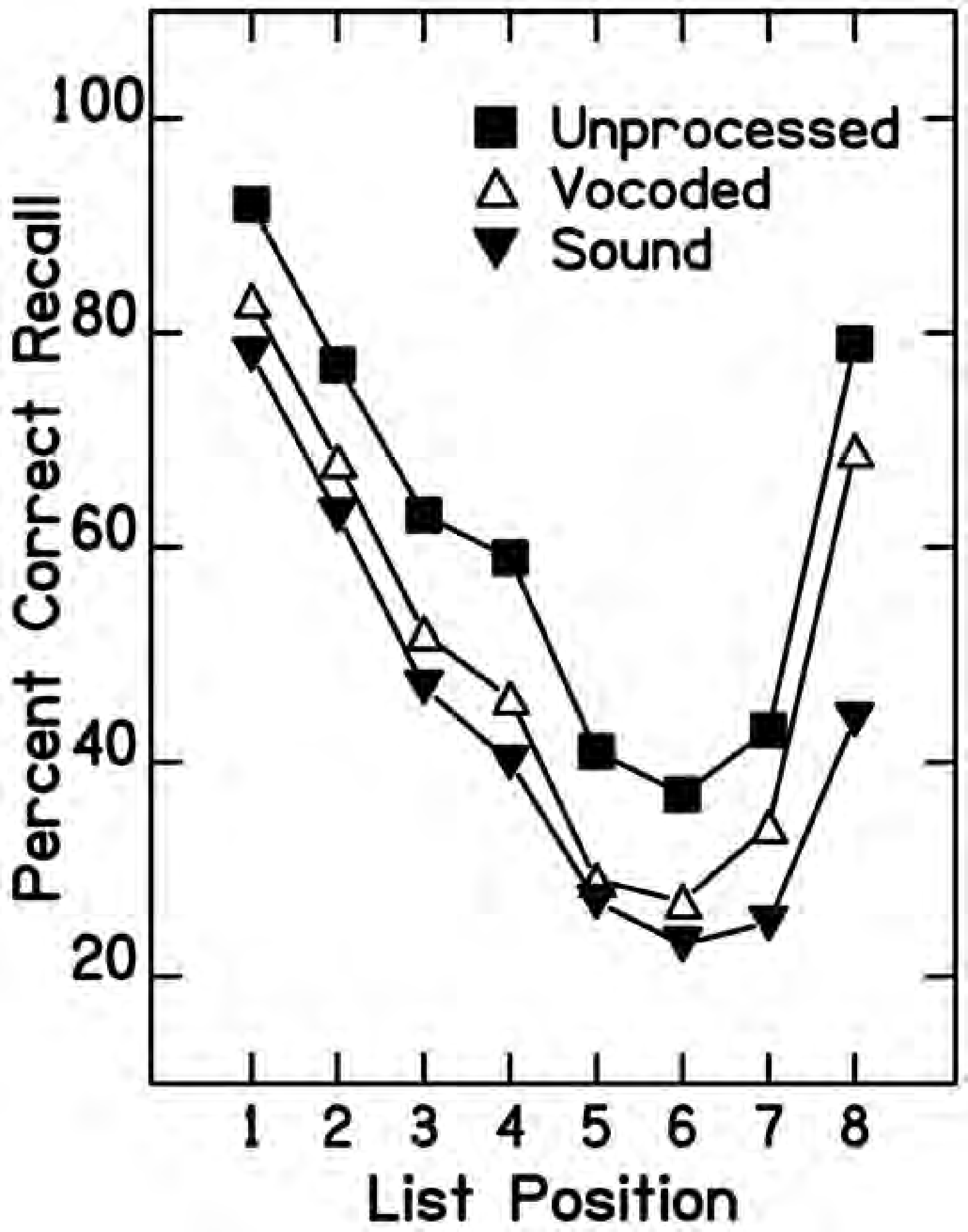
	df	*F*	*p*	η_p^2
Main Effects				
Condition	1,78	33.12	<.001	.298
Type	2,78	28.23	<.001	.266
Group	1,78	1.30	NS	---
Two-way Interactions				
Group x Condition	1,78	7.20	.009	.085
Group x Type	2,78	2.10	NS	---
Condition x Type	2,156	5.12	.007	.062
Three-way Interaction				
Group x Condition x Type	2,156	6.46	.002	.076

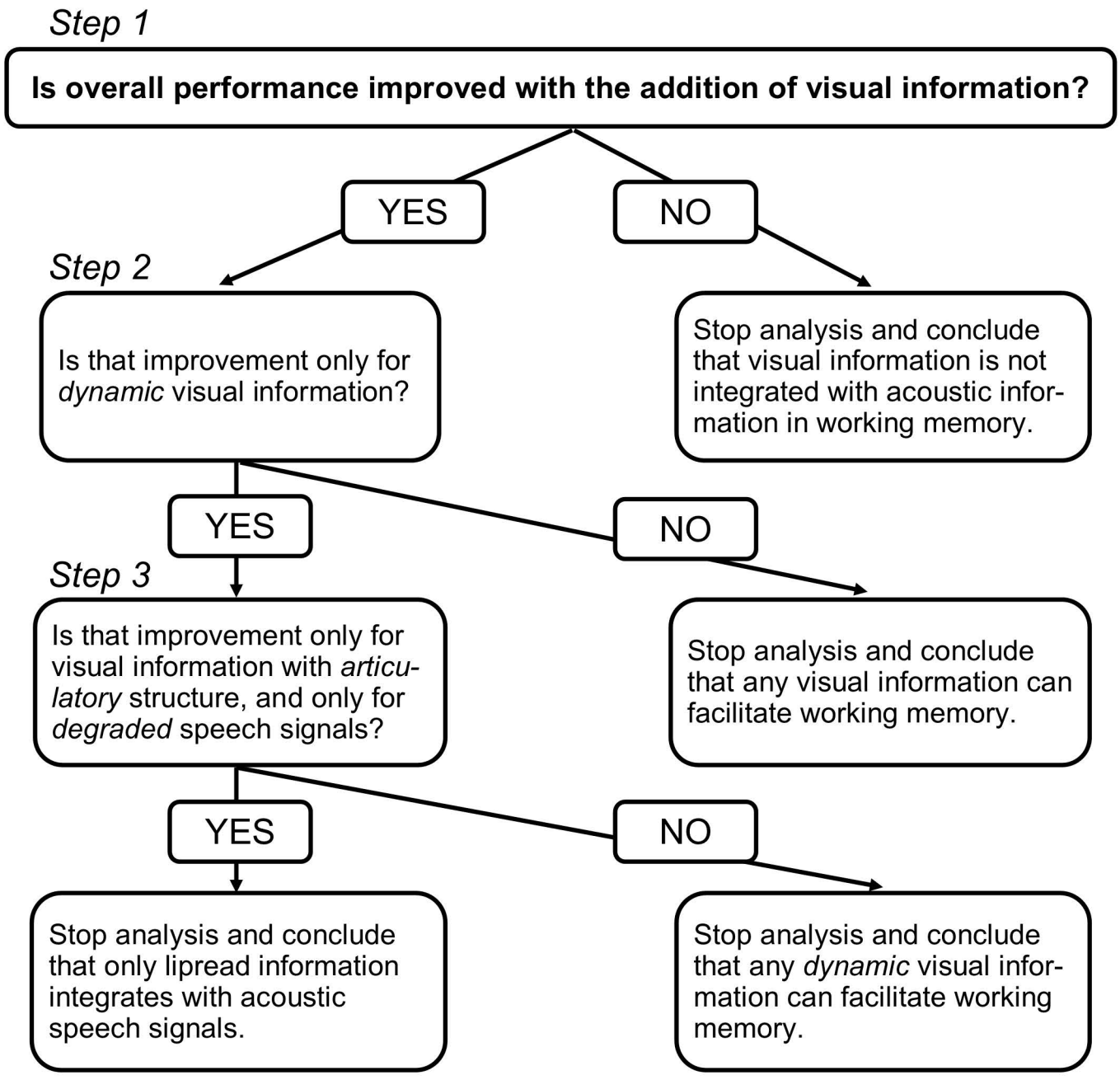
Table 5

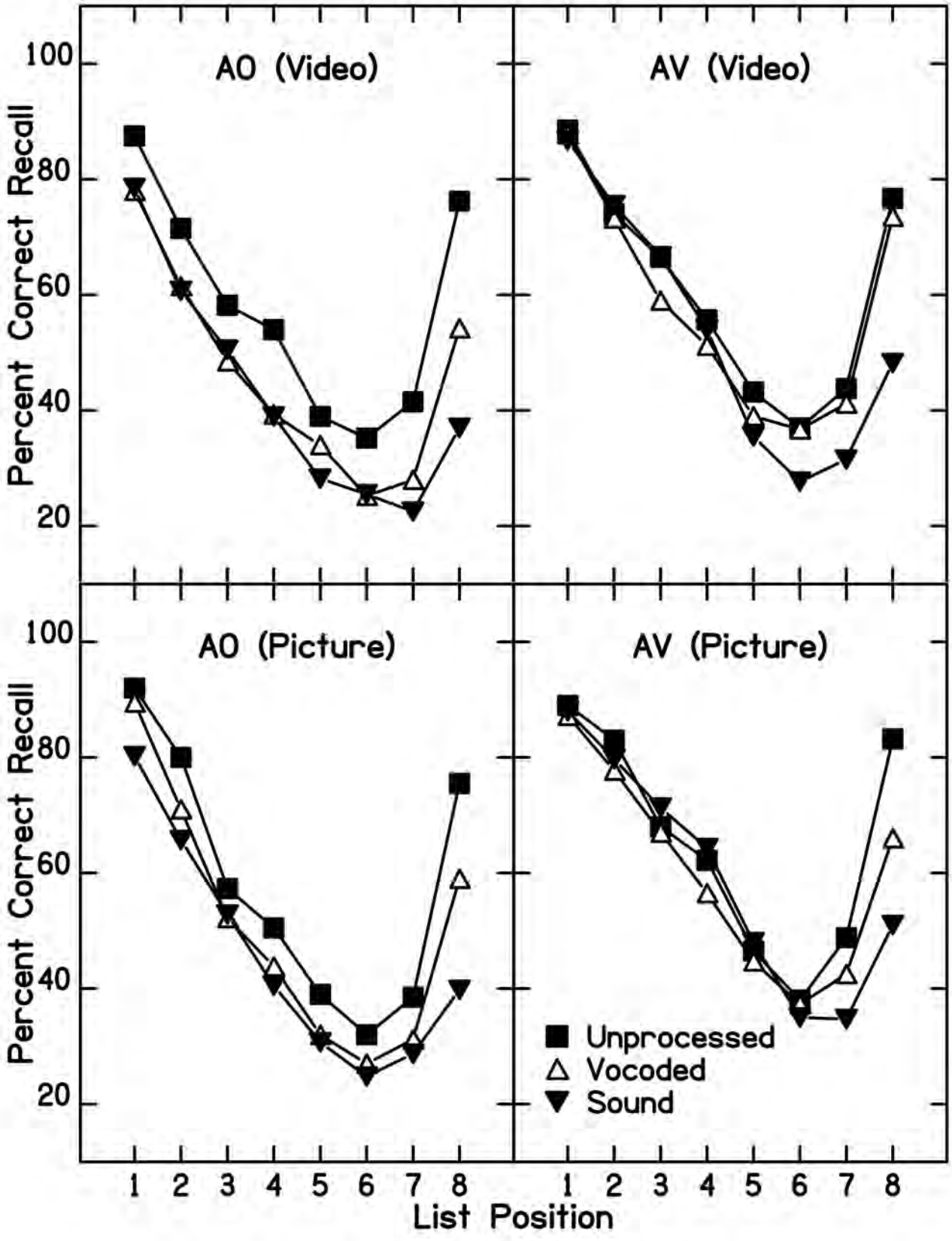
Results of a Two-Way, Repeated-Measures ANOVA for Response Time, for Each Stimulus Type Separately

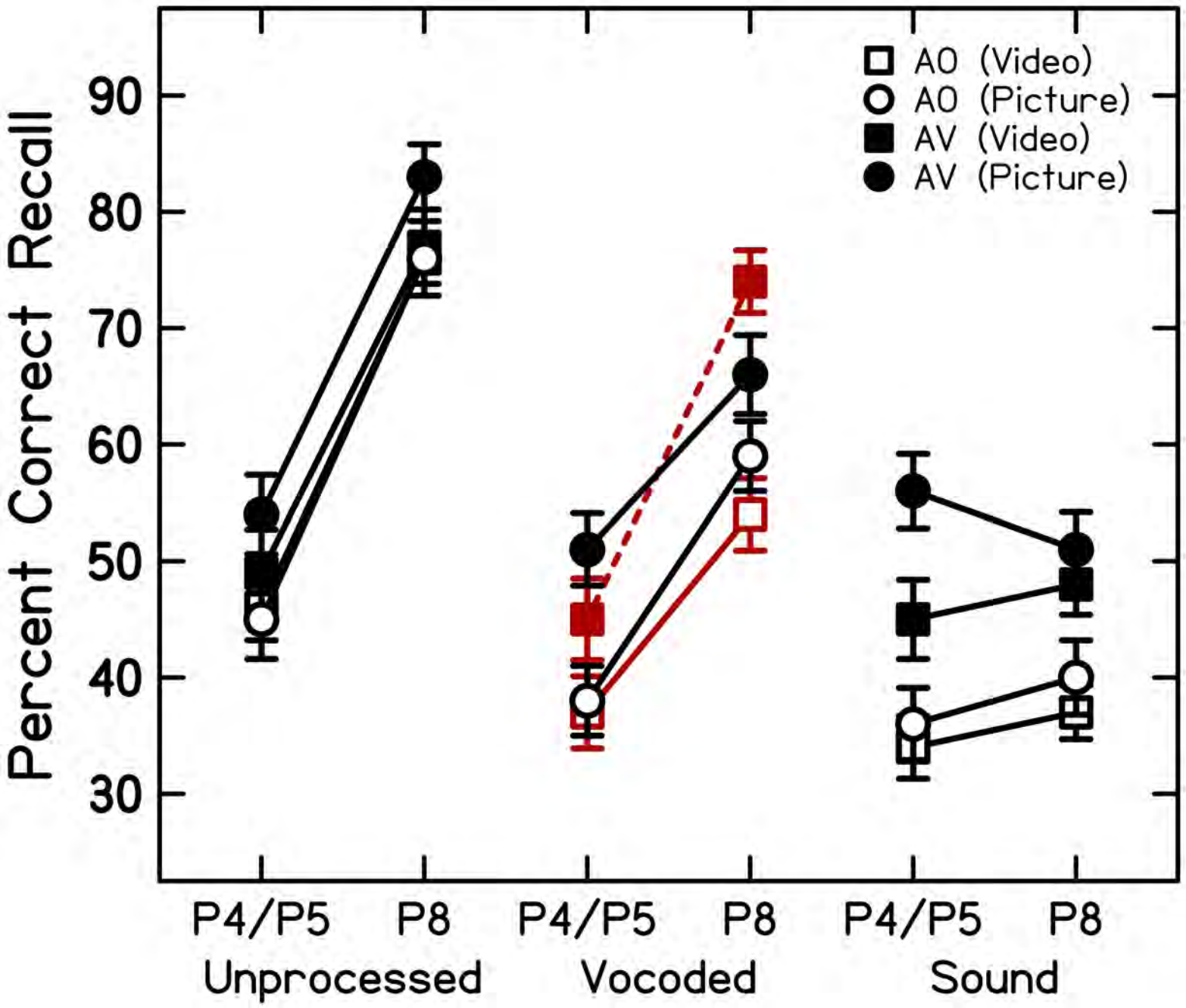
	df	*F*	*p*	η_p^2
Unprocessed Speech				
Condition	1,78	24.13	<.001	.236
Group	1,78	.311	NS	---
Group x Condition	1,78	1.36	NS	---
Vocoded Speech				
Condition	1,78	24.53	<.001	.239
Group	1,78	1.43	NS	---
Group x Condition	1,78	.121	NS	---
Environmental Sounds				
Condition	1,78	6.69	.012	.079
Group	1,78	2.12	NS	---
Group x Condition	1,78	28.35	<.001	.267

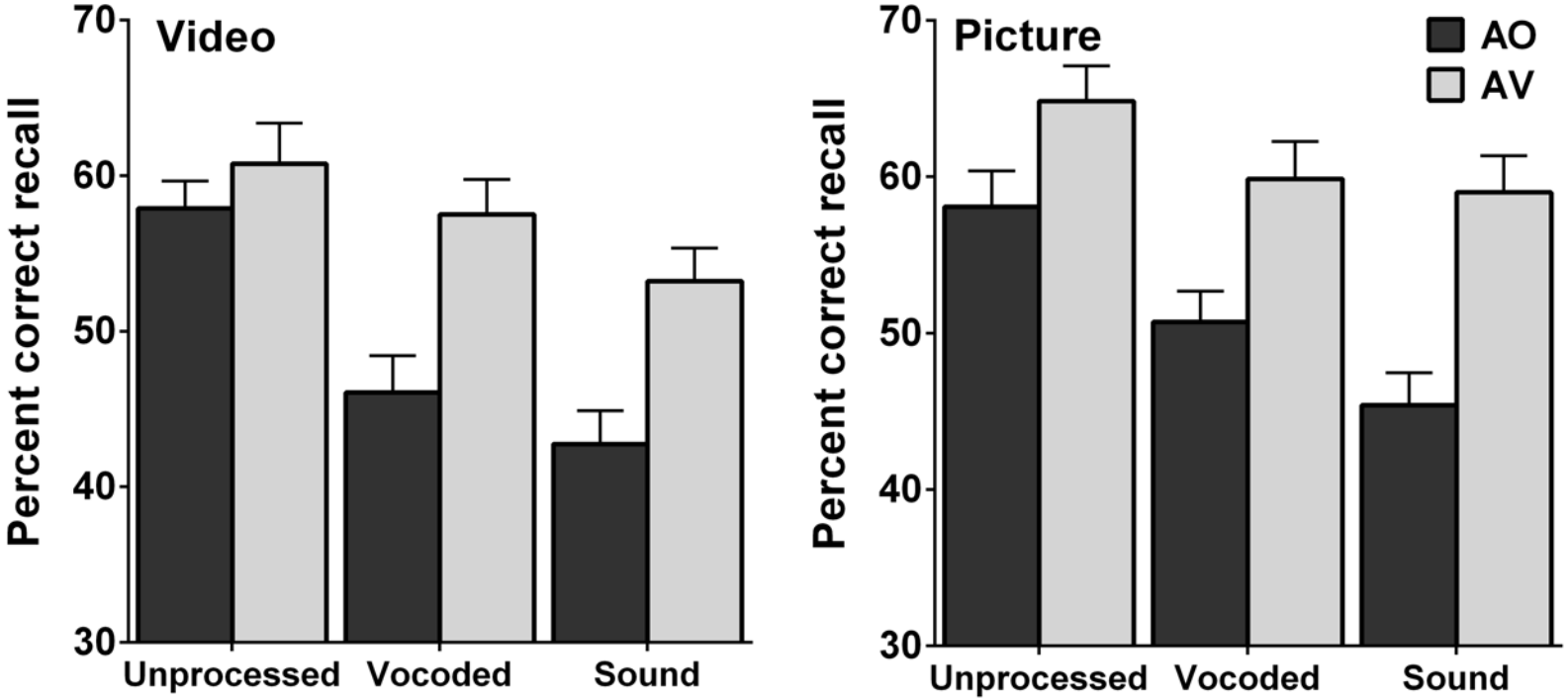

Central Executive
Phonological Loop
Visuospatial Sketch Pad

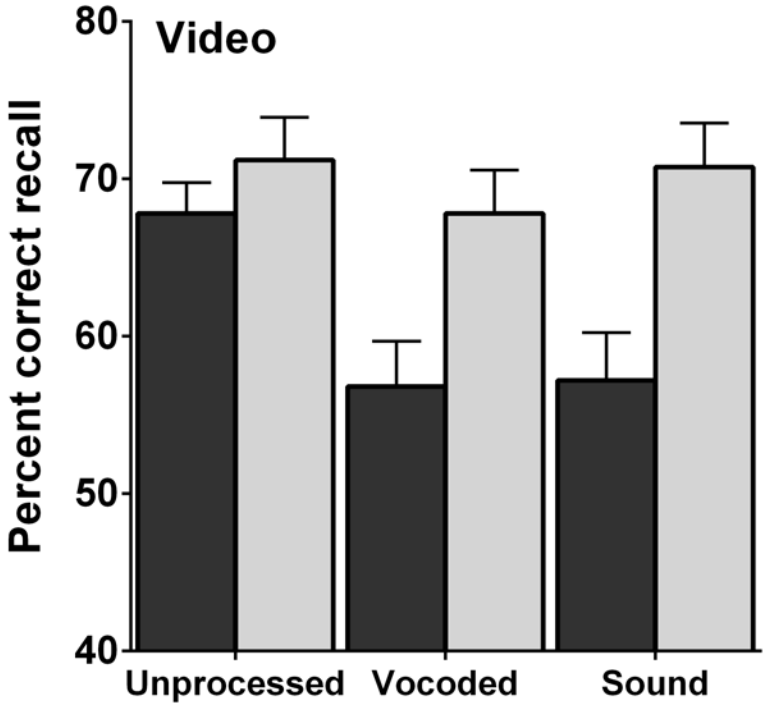

Unprocessed
Vocoded
Sound
Percent Correct Recall
100
80
60
40
20
1
2
3
4
5
6
7
8
List Position

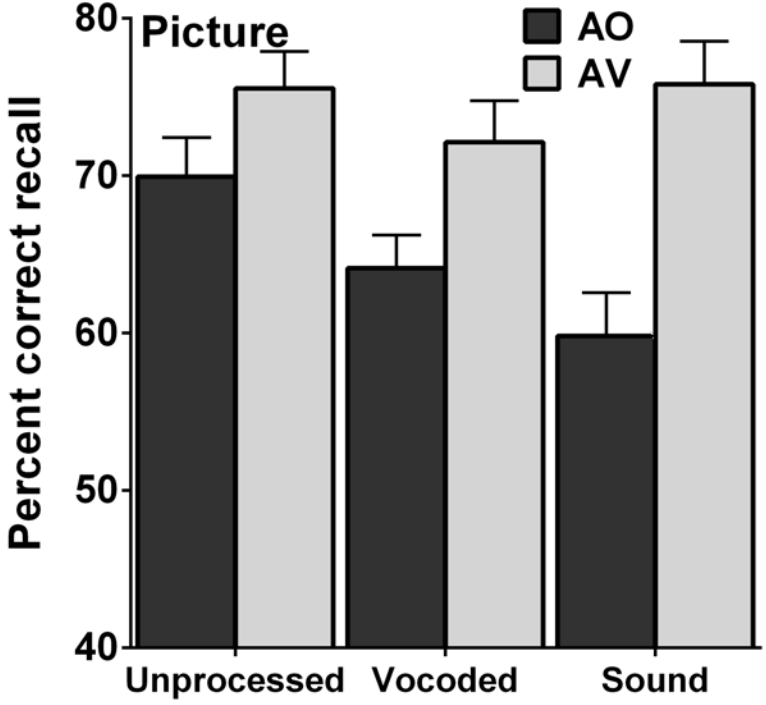

Step 1
Is overall performance improved with the addition of visual information?
YES
NO
Step 2
Is that improvement only for *dynamic* visual information?
Stop analysis and conclude that visual information is not integrated with acoustic information in working memory.
YES
NO
Step 3
Is that improvement only for visual information with *articulatory* structure, and only for *degraded* speech signals?
Stop analysis and conclude that any visual information can facilitate working memory.
YES
NO
Stop analysis and conclude that only lipread information integrates with acoustic speech signals.
Stop analysis and conclude that any *dynamic* visual information can facilitate working memory.

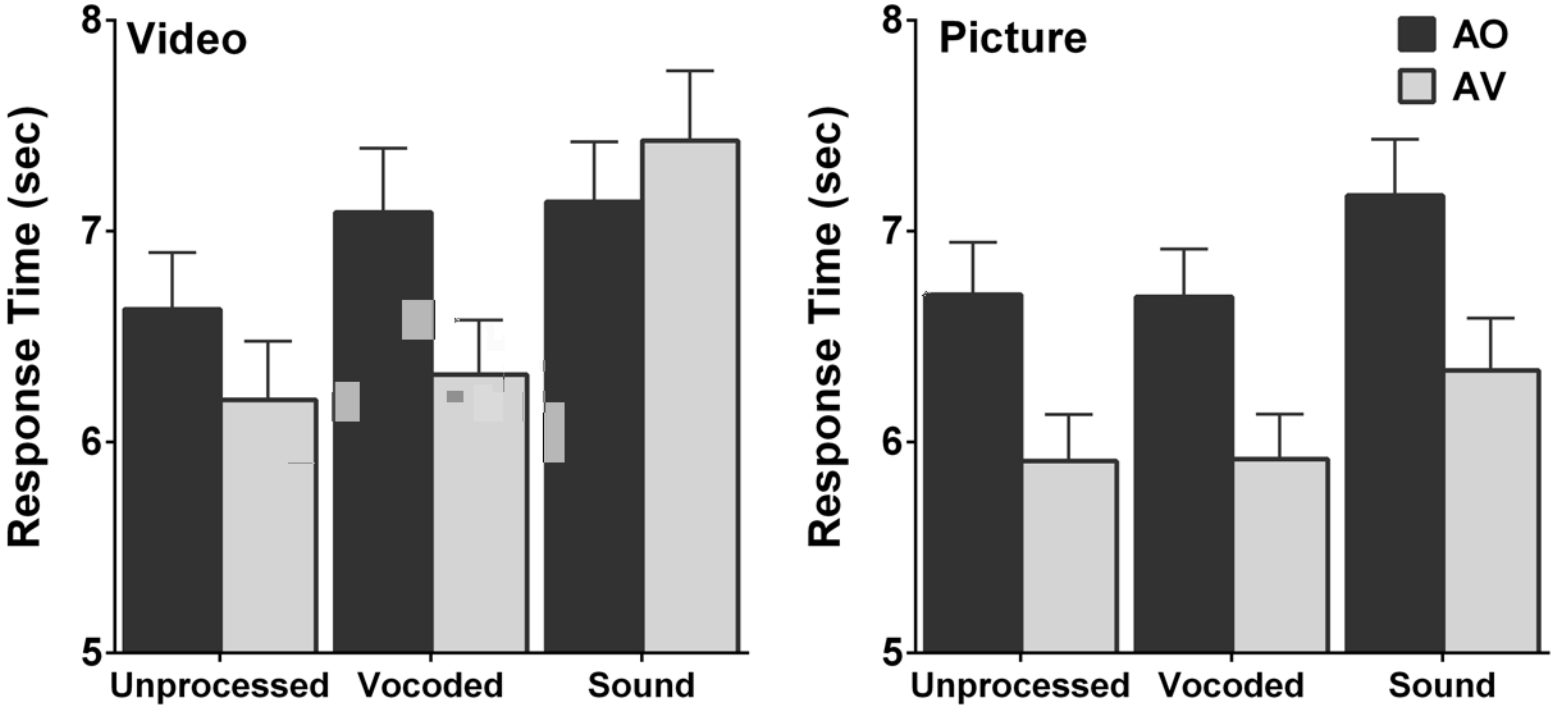

AO (Video)
AV (Video)
AO (Picture)
AV (Picture)
Percent Correct Recall
100
80
60
40
20
1 2 3 4 5 6 7 8
List Position
Unprocessed
Vocoded
Sound

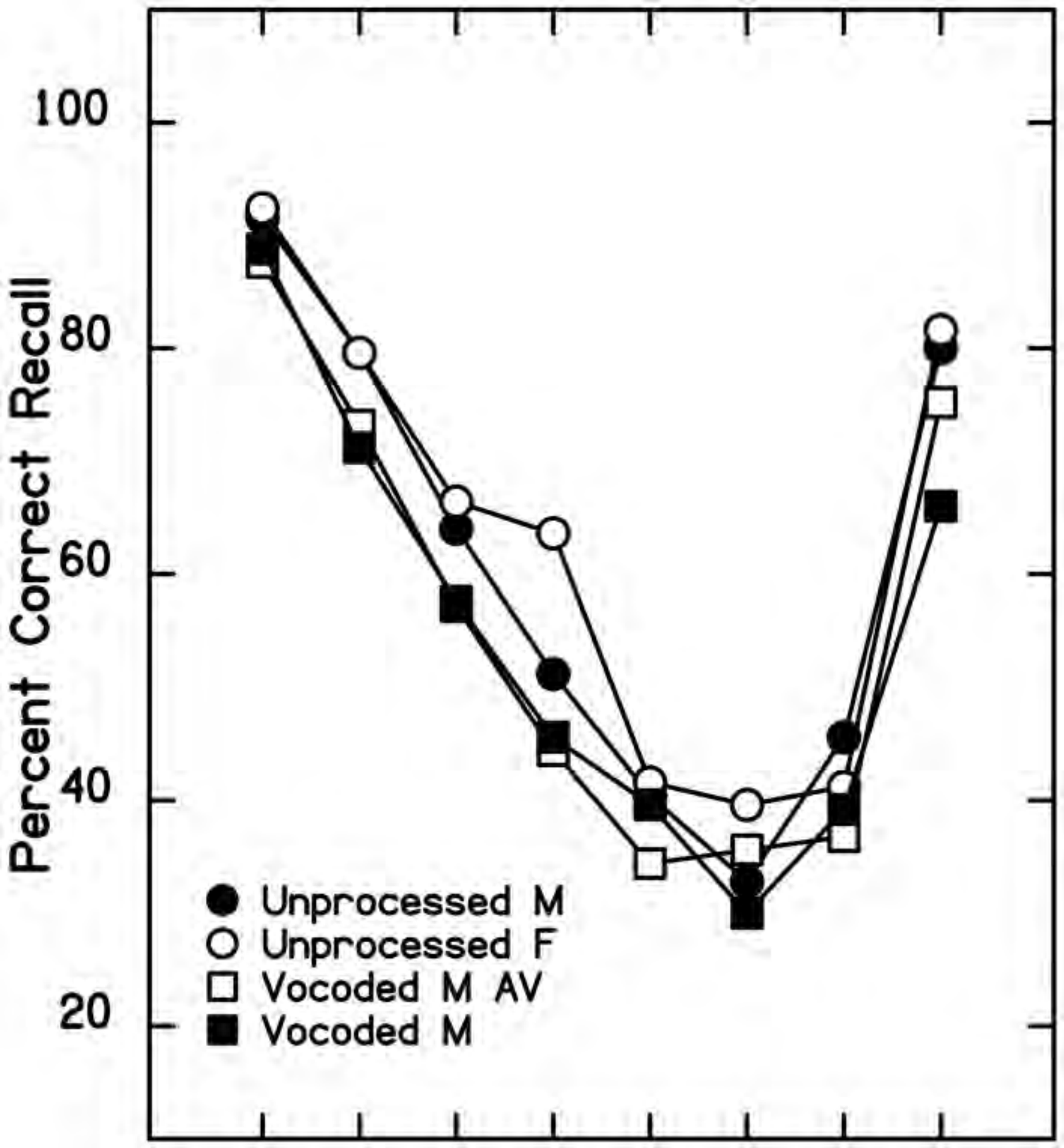

Percent Correct Recall
90
80
70
60
50
40
30
A0 (Video)
A0 (Picture)
AV (Video)
AV (Picture)
P4/P5
P8
Unprocessed
P4/P5
P8
Vocoded
P4/P5
P8
Sound

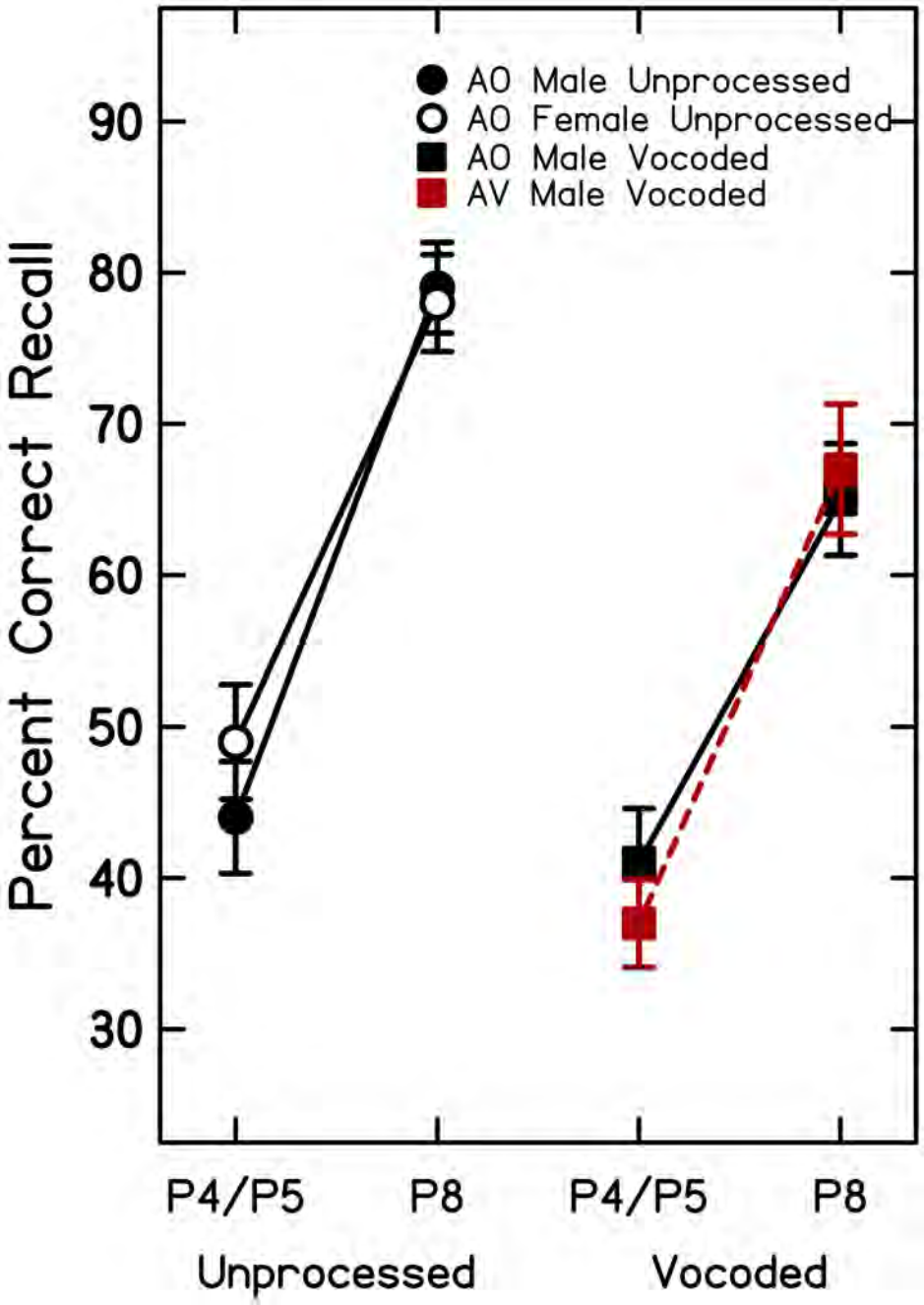

Video
Picture
AO
AV
Percent correct recall
70
60
50
40
30
Unprocessed
Vocoded
Sound


Video
Percent correct recall
80
70
60
50
40
Unprocessed
Vocoded
Sound


Picture
AO
AV
Percent correct recall
80
70
60
50
40
Unprocessed
Vocoded
Sound


Video
Response Time (sec)
8
7
6
5
Unprocessed
Vocoded
Sound
Picture
AO
AV
Response Time (sec)
8
7
6
5
Unprocessed
Vocoded
Sound


100
80
60
40
20
Percent Correct Recall
Unprocessed M
Unprocessed F
Vocoded M AV
Vocoded M


A0 Male Unprocessed
A0 Female Unprocessed
A0 Male Vocoded
AV Male Vocoded
Percent Correct Recall
90
80
70
60
50
40
30
P4/P5
P8
P4/P5
P8
Unprocessed
Vocoded