ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Speech Communication 46 (2005) 73–93

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Variability in bimodal spoken language processing by native and nonnative speakers of English: A closer look at effects of speech style

Debra M. Hardison *

Department of Linguistics and Languages, Michigan State University, A-714 Wells Hall, East Lansing, MI 48824, USA

Received 20 November 2003; received in revised form 12 August 2004; accepted 10 February 2005

Abstract

Experiments using the gating paradigm investigated the influence of speech style (unscripted vs. scripted), visual cues from a talker's face (auditory-visual vs. auditory-only presentations), word length (one vs. two syllables) and initial consonant (IC) visual category in spoken word identification by native- (NSs) and nonnative speakers (NNSs) of English. Two talkers were videotaped in separate conversations with the author on various topics (unscripted speech) with the camera focused on the talker's face. They later recorded selected sentences as scripted speech. In Experiment 1, target words were excised from sentences, gated, and presented to NSs and NNSs. Results for both groups showed significantly earlier identification with visual cues and for bisyllabic vs. monosyllabic words, and significant interaction of speech style and IC category across talkers. For Talker A, identification of unscripted monosyllabic words in some IC categories was earlier than the scripted versions. For this talker, words beginning with /ɹ/ were identified later than others; for Talker B, they were the earliest to be identified. For NNSs, the AV advantage was accentuated for words beginning with /ɹ, w, θ/ in unscripted speech by Talker A, and for /ɹ/- and /l/-initial words in both speech styles by Talker B. Experiment 2 presented the preceding sentence context with the gated word to NSs. Results revealed earlier identification in AV presentation and for unscripted vs. scripted words by Talker A. With context, word length was not significant. Findings highlight the priming role of visual cues, and the talker- and context-dependent nature of bimodal spoken language processing, but do not support a strict conversational-clear speech dichotomy.

Keywords: Auditory-visual; Spoken language processing; Speech style; Gating; Language learning

* Tel.: +1 517 353 0800; fax: +1 517 432 1149.
E-mail address: hardiso2@msu.edu

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.02.002

1. Introduction

Studies that have dealt with the effects of speech style on language processing have assumed a loss of intelligibility in conversational speech regardless of talker, phonetic environment, or modality of presentation. Researchers have varied in how they have elicited and labeled different speech styles. For example, McAllister (1991) investigated the role of syllable stress on word recognition in "spontaneous" vs. "read" speech while others used "conversational" vs. "clear" speech (Gagné et al., 1994; Helfer, 1997). In the latter case, conversational was labeled less intelligible by virtue of its contrast with a style that actually represented hyperarticulated or exaggerated speech. Given the numerous findings in the literature on talker variability in auditory language processing (e.g., Johnson and Mullennix, 1997), the present study explored further the intelligibility of conversational speech by comparing the influence of speech style (unscripted/conversational vs. scripted) on the information value of facial cues from two different talkers in a spoken word identification[1] task by native and nonnative listeners of English using the gating paradigm. Talkers were not given instructions on how to produce speech in either style. Stimuli were excised from recorded sentences and presented to both subject groups in auditory-visual (AV) and auditory-only (A-only) presentations. A visual-only (V-only) condition was also presented to a group of native listeners. This research was guided by the following questions: (1) Does the contribution of visual cues to word identification vary according to speech style, the initial consonant visual category of the target word,[2] and/or word length (i.e., one or two syllables)? (2) Do nonnative listeners show earlier word identification with visual cues across speech styles and initial consonant categories? (3) What is the visual discernibility of words excised from connected speech and presented in a visual-only condition (i.e., how well can such words be lip-read)? (4) Are findings compatible with reduced intelligibility for unscripted vs. scripted speech? A second experiment examined the interaction of preceding sentence context with modality, word length, and initial consonant category with native listeners.

The auditory gating paradigm involves the successive presentation of increasing increments or gates of a stimulus (e.g., Grosjean, 1980, 1996; Salasoo and Pisoni, 1985). It has demonstrated effects on word identification of factors such as word frequency (e.g., Grosjean, 1980; Tyler, 1984), neighborhood density (Metsala, 1997), syntactic and semantic context (e.g., Grosjean, 1980; McAllister, 1988), word length (e.g., Craig and Kim, 1990; Grosjean, 1980), and word stress (McAllister, 1991). In general, identification is earlier for high-frequency words with few neighbors (i.e., words with a similar phonetic structure) (Garlock et al., 2001; Goldinger et al., 1989; Metsala, 1997). Monosyllabic words may be identified after their acoustic offset (Grosjean, 1985) facilitated by the presence of subsequent context (Bard et al., 1988). With more constraining preceding contexts, length of the target word plays less of a role in the identification process (Grosjean, 1980). Responses are the same when the gates are presented individually or in successive format (Cotton and Grosjean, 1984; Salasoo and Pisoni, 1985).

[1] The term *identification* is used in this paper in the phrase spoken word identification rather than *recognition*. In the auditory gating paradigm developed by Grosjean (1980), subjects were asked to indicate what word they thought the talker was going to say and how confident they were of their decision. The result of the first task was referred to as the isolation point; that is, the point at which the word is isolated from other word candidates (usually expressed in terms of the amount of the stimulus needed for identification with no change in response thereafter). The amount of the stimulus needed to reach a particular confidence level was referred to as the recognition point. As Grosjean (1996) notes, this does not necessarily reflect a word's actual recognition point from a processing perspective. In the present study, subjects were given only the task of identifying the word. Following Salasoo and Pisoni (1985), the term *identification point* is used here and operationally defined as the amount of the stimulus required for consistent correct identification of a word.

[2] The term *viseme* is sometimes used to indicate a visual or homophenous category within which articulatory movements are not visually distinguishable (e.g., Fisher, 1968), e.g., the bilabials /p, b, m/. This term arose from analogy with the term phoneme. As one does not observe (or hear) an abstract concept, but a talker- and context-dependent articulatory gesture, the term *visual category* is used in this paper.

In an auditory gating study designed to investigate the role of syllable stress in word identification in spontaneous and read speech, McAllister (1991) found that (a) words with a stressed initial syllable were more likely to be identified than those with an unstressed one, (b) more read words than those produced in spontaneous speech were identified at or before their acoustic offset, and (c) in the presence of sentence context, the effect of word stress was confined to spontaneous speech.

Researchers have varied in how different styles or manners of speech have been elicited. In (McAllister's, 1991) study, audio recordings of conversations were used for spontaneous speech with subsequent recordings of scripted versions for read speech. In (Helfer's, 1997) study, both "clear" and "conversational" speech samples were distinguished based on instructions given to the talker prior to recording. For clear speech, the talker was instructed to speak as if communicating in a noisy environment or to someone with a hearing loss. She was to enunciate consonants more carefully and with greater effort and to avoid slurring words. This was in contrast to instructions to "speak as in normal conversation." The stimuli were nonsense sentences, read from a cue card, and presented to normal-hearing subjects in two conditions (AV and A-only). Results indicated stimuli were easier to understand in AV presentation for both clear and conversational speech.

Gagné et al. (1994) used mono- and bisyllabic words spoken in isolation that were presented to normal-hearing adult subjects in either degraded auditory, visual-only, or degraded auditory-visual conditions. The stimulus set was recorded twice producing "conversational" and "clear" speech with only the lower half of the talker's face visible to subjects. "Conversational" resulted from no instructions to the talkers on how to produce the utterances; "clear" described the second recording when the talker was asked to say the same words as if talking to someone who had difficulty understanding what they were saying. Intelligibility results revealed a significant interaction between talkers and the above manners of speech as talkers varied in whether they produced a significant effect of clear speech in any modality.

Clear speech has been characterized by a longer duration of utterances, more pauses, and longer pauses between words (e.g., Picheny et al., 1986, 1989). In contrast, "fast or casual speech" has been described as being reduced (e.g., having fewer syllables through deletion of unstressed vowels) or underspecified compared to "careful speech" (Dalby, 1986).

More recently, results of a study pointed to several characteristics that define a talker producing highly intelligible speech from an auditory standpoint (Bradlow et al., 1996): female, use of a relatively expanded vowel space that covers a broad range in F1, precise articulation especially of the point vowels, low degree of phonetic reduction, high precision of intersegmental timing, and at the sentence level, use of a relatively wide range in fundamental frequency. Criteria for greatest visual discernibility of a talker's articulatory gestures include absence of facial hair that could obscure lip movements, the number of changes in facial movement during the production of a word, total elapsed time from the onset of the word to each of the changes in movement, intensity of facial movement and total duration of the word (Berger, 1972).

Attention to facial cues in speech perception serves several purposes. It orients the listener to the location and identity of the talker, establishes a rapport between the interlocutors, which is important in many cultures, and provides both redundant speech cues and information that complements that which is received auditorily. Listeners are often more aware of the contribution of visual information when compensating for noise (e.g., Sumby and Pollack, 1954; Summerfield, 1979) or impaired hearing ability (e.g., Walden et al., 1977). Visual cues can also be useful for second language learners. McGurk effect experiments revealed that the presence of a talker's face significantly improved the perceptual accuracy of nonnative sounds such as /ɹ/ and /f/ for Japanese and Korean learners of English as a second language (Hardison, 1999). Further studies demonstrated the superiority of AV vs. A-only perceptual training of /ɹ/, /l/, /f/, and /θ/ for Japanese and Korean speakers, generalization to novel stimuli and talkers, and transfer to improved production of these sounds (Hardison, 2003).

In a McGurk effect experiment using separately gated auditory and visual VCV stimuli, Munhall and Tohkura (1998) found that visual information appeared to unfold linearly with articulatory movements; however, the information contributed by the acoustic signal varied across time with some points more salient in their information value than others such as the release burst of a stop. They concluded that because of the natural temporal precedence of articulatory gestures over the associated acoustic signal, visual cues could serve a priming role in AV speech processing.

A recent study demonstrated that this priming role results in earlier word identification using the gating paradigm in AV vs. A-only presentation for Japanese and Korean learners of English as a second language (Hardison, in press). Stimuli were familiar, bisyllabic words beginning with /p/, /f/, /ɹ/, /l/, and /s, t, k/ (based on visual categories) combined with high, low and rounded vowels. They were excised from experimentally controlled sentences produced by a female talker who served as one of the talkers in the current study. Subjects were assigned to AV or A-only presentations compatible with their perceptual training group type. Identification of words was significantly earlier in AV vs. A-only presentation, and was enhanced following 3 weeks of perceptual training using minimal pairs that contrasted /ɹ/-/l/, /f/-/p/, and /θ/-/s/. AV input was particularly effective for accurate identification of words beginning with /ɹ/ and /l/. A control group that did not receive any training did not show significant improvement in identification performance.

In a V-only presentation of the stimuli in the above study, at least six of eight NSs accurately identified 19% in the excised word condition and 41% in the sentence condition. The majority of these identified words began with /p/, /f/, /ɹ/ and /l/. Their identification accuracy may be attributed to the relative salience of the movement from one visual category to the next as the talker produced the word. Comparison of the visual categories of the stimuli with those of the subjects' responses revealed a closer match when the V-only gated word was preceded by the AV sentence context.

In the present study, two female talkers, both native speakers of General American English, were videotaped under two conditions: an unscripted condition (conversation with the author) from which stimuli were selected, and a scripted condition involving a second recording of the selected stimuli. The terms *unscripted* and *scripted* are used here rather than "spontaneous and read" (e.g., McAllister, 1991) or "conversational and clear" (e.g., Helfer, 1997). During the videotaping of the scripted version, the talkers did not actually read a script but were required to look at each sentence, hold it in memory, then look at the camera, and produce it. The talkers were not given any articulation instructions at any time and were not informed of the intended use of the recordings; therefore, the style of speech emerged from the unscripted (conversational) or scripted nature of the speech event as would be the case in a real-life situation. In both the Helfer (1997) and Gagné et al. (1994) studies, "clear" speech might be better characterized as hyperarticulated.

The current study was designed to investigate the potential interactions between style of speech and modality of presentation, word length (one or two syllables), initial consonant visual category, and presence of sentence context for speech produced by two different talkers. Based on earlier findings, it was hypothesized that word identification would be facilitated by the presence of visual cues for both native and nonnative listeners, but that variation would be evident with speech style and the initial consonant visual category of the target word. Performance by nonnatives was expected to follow patterns of other studies by showing later identification of words beginning with /ɹ, l, f, θ/ especially in A-only presentation as these subjects had not received any perceptual training on these sounds. For the NSs in Experiment 2 (sentence condition), AV presentation was also expected to result in earlier identification. It was hypothesized that the presence of context would reduce the effect of word length for both speech styles. In previous studies, context impacted the influence of other factors on the identification process: the effect of word stress was confined to spontaneous speech (McAllister, 1991) and the effect of word length was reduced (Grosjean, 1980).

2. Experimental design

Two female native speakers of American English participated in separate videotaped casual conversations with the author in a small television studio producing large samples of unscripted speech on a variety of topics. Each talker (henceforth Talkers A and B) was well known to the author. This rapport contributed to a comfortable atmosphere and permitted the author a knowledge of topics on which conversation could be prompted to produce extended speech samples. The goal was to obtain as many potential target words to be gated as possible. Topics of daily life included family, professional activities, hobbies, pleasant and unpleasant experiences, travel, etc. Both talkers had had some experience being videotaped and were comfortable in front of the camera.

From these recordings, target words and their sentences were selected, transcribed, randomized and printed in a large font on a script for each talker. Ten days after the first taping, they returned to the studio and were videotaped again while producing each of the sentences from the script. The recordings providing unscripted and scripted speech samples were digitally edited to create gated target words in two conditions within each speech style: excised word (target word excised from the sentence) and sentence (included preceding sentence fragment with only the target word gated).

The data for the excised word condition (Experiment 1) and sentence condition (Experiment 2) were analyzed separately because the amount of preceding sentence context could not be controlled and has a significant effect on word identification (Grosjean, 1980); thus, it was treated in analysis as a covariate. For Experiment 1, subjects included native and nonnative speakers of American English; Experiment 2 involved only native speakers.[3] The NSs were randomly assigned to a total of 24 experimental groups: three modalities of presentation (AV, A-only, V-only) × two styles of speech (unscripted, scripted) × two conditions (excised word, sentence) × two talkers. There were 14 subjects in each group. Eight NNSs were randomly assigned to each of eight experimental groups: two modalities of presentation (AV, A-only) × two styles of speech × two talkers. All NNSs were presented with the excised word condition. Between-subjects variables were modality (AV, A-only) and speech style (unscripted, scripted). Within-subjects variables were word length (monosyllabic, bisyllabic) and initial consonant (IC) category of the target word. There were eight consonant categories for Talker A and seven for Talker B. Due to the anticipated interactions involving these variables, the differing numbers of IC categories, and different subjects,[4] the data for each talker were analyzed separately.

Additional word identification data were gathered from NSs in V-only (speechreading) presentations involving the excised word and sentence conditions. In the sentence condition, the audio for only the gated word was deleted. In the absence of a unique correspondence between sound and articulatory gesture, these data were analyzed separately from the AV and A-only data.

3. Experiment 1: excised word condition

3.1. Method

3.1.1. Subjects

Subjects involved both native and nonnative speakers of American English. All were between the ages of 22 and 30. None reported any vision or hearing deficits, speechreading training, or familiarity with either talker in the study. The nonnatives were speakers of Japanese whose prior English study had begun in middle school and had focused on grammar and reading rather than aural/oral skills. They had a high-intermediate level of spoken English proficiency as judged by

[3] There was an insufficient number of nonnative speakers for both experiments. Those who participated were assigned to the excised word condition. Previous research had documented the contribution of context to the word identification process for this population with stimuli produced by Talker A in the present study (Hardison, 2004).

[4] Because of the amount of time needed for the gating task and concerns for subject fatigue, different subjects were used for each talker.

the author[5] following a brief interview. At the time of this study, they were enrolled in academic programs in the US. Native speakers of Japanese were chosen for this study to establish some comparability with previous studies, and because of the challenges faced by this population with regard to perceptual accuracy of word-initial liquids (see, e.g., Hardison, 2003; Lively et al., 1993).

3.1.2. Materials recording and selection

A conversation was videotaped between the author and each talker seated comfortably in a small television studio with a total of five LOWEL DP 750 W halogen lights: one directed at the light gray background, two illuminating the set from the left, and two from the right. The talker wore an Electrovoice (EVC090) lavaliere microphone. The camera (SONY EVO9100 Hi-8 videocamera) positioned at a distance focused on the talker providing a full-sized image of the head. As only the interlocutors were present in the studio to maintain an atmosphere compatible with a casual conversation, the camera remained in a fixed position. The objective was to produce a substantial amount of recorded material from which target words could be selected to meet the criteria described below.

From the initial videotaped conversation with each talker, sentences were selected, transcribed and presented to a group of eight NSs as a cloze task that provided the preceding sentence fragment followed by a blank line for the target word. Respondents were asked to complete the blank with a logical word. This provided some measure of the predictability of the word in context for Experiment 2 (Marslen-Wilson and Welsh, 1978; McAllister, 1991). Those words not reported were then rated on a 7-point familiarity scale by another group of eight NSs, and six NNSs with English proficiency below that of the subject population. A preliminary stimulus set was selected consisting of highly familiar (rated as 6 or 7) monosyllabic and bisyllabic content words (with initial syllable stress)[6] beginning with consonant sounds belonging to each of the following visual categories: bilabial /p, m/, labiodental /f/, nonlabial /s, t, n, k/, palato-alveolar /ʃ, t ʃ, dʒ/, /ɹ/, /l/, /w/, and for Talker A, there were also sufficient words beginning with interdental /θ/.

Additional criteria for selection included sentence focal stress, nonfinal position (to avoid sentence-final effects on intonation), and the visual influence of adjacent sounds. Targets were selected that were preceded by a relatively neutral lip position (e.g., words such as *that* ending in an alveolar consonant, or *the* ending in a schwa), with initial sounds that were followed by as little rounding as possible. Lexical neighborhood density was also taken into account although not treated as a variable. Neighborhood density is determined by the number of words that differ from a target word by a one-phoneme addition, deletion or substitution (e.g., Luce, 1986). Because of the amount of material recorded, it was possible to use only those words from relatively sparse neighborhoods (defined as 15 or fewer neighbors in this study).[7] Sentences were not used if the target word or preceding context contained a speech error, long pause, repetition, smile or laughter that distorted lip movements related to speech sounds, or involved movement of the head so that full view of the face and lower jaw drop were obscured. A final stimulus set of four monosyllabic and four bisyllabic words for each visual category was constructed (see Appendix A).

Ten days after the first taping, talkers returned to the studio and were videotaped again while producing each of the sentences from the script. Each

[5] The author also has experience in language teaching including oral proficiency testing.

[6] The number of syllables was based on actual production of a word rather than the citation form. For example, the words *family* and *restaurant* were produced with two syllables (no medial vowel) in both speech styles in this study.

[7] The validity of the density calculation for adult second-language (L2) learners is questionable much as it is for children (Metsala, 1997). In the case of an L2 learner actively engaged in the process of acquiring a language in the host environment with or without formal instruction, the neighborhood for each lexical item fluctuates potentially daily throughout interlanguage development and is subject to influence (inhibitory and facilitative) from the first language.

sentence was presented on a separate sheet of paper and the sheets were randomized so that topics were mixed to avoid intonation patterns characteristic of producing successive sentences on a related topic (McAllister, 1991). Both talkers indicated that they did not recall the specific content of the conversations.

3.1.3. Materials editing

The recordings (at 30 frames/s) were digitized at 44.1 kHz and edited using AVID Media Composer (MC8000) version 5.51 for MacIntosh to create gated stimuli in two conditions within each speech style: excised word and sentence condition (i.e., the preceding sentence fragment with only the target word gated). Each utterance was saved as a file. The duration of each gate was two frames (see Appendix B).[8] Following editing, stimuli were transferred in a randomized order to broadcast quality videotapes for AV, A-only and V-only presentations. For the A-only groups, the screen remained black.

To produce the excised word condition, the target word was edited out of the sentence. The first presentation involved no visual or acoustic information for the target in order to correspond to the first presentation in the sentence condition (see Section 4.1.2) that provided only the preceding context. The second pass presented the first gate (first two frames) of the target word, the third included the first gate plus another (i.e., two additional frames), and so on. With this approach, the amount of the target word at each presentation was the same for both the sentence and excised word conditions even though the data from these conditions could not be directly compared in statistical analysis in this study. Two warning tones signaled the presentation of a new stimulus, and one tone indicated the next gate of the same stimulus.

The determination of the gating onset was guided by the acoustic signal. Although this allows for some anticipatory lip movements belonging to the target word's production to be evident on the first pass in an AV sentence presentation, it avoids creating an auditory garden path (i.e., suggesting an incorrect word-initial sound).

3.1.4. Procedure

Subjects were tested in small groups and seated comfortably in front of an elevated 27-in. TV monitor. To familiarize them with gating, practice utterances were presented using a talker unrelated to the experiment; then an adaptation utterance was provided by the experiment talker. In the experiment, subjects were instructed to write down the word they thought the talker was saying at each gate and not to change previous responses. They were given four seconds to respond. Observation of experimental sessions ensured the subjects understood the instructions and those in AV and V-only groups looked at the screen.

3.2. Results and discussion

Results for the AV and A-only modalities of presentation are discussed first. Within this section, the native speaker results for Talker A and Talker B are followed by a discussion of the nonnative speaker results for each of these talkers. Findings are then presented for the V-only modality (NSs only) also divided into subsections for Talkers A and B.

3.2.1. AV and A-only modalities

For analysis of the AV and A-only results, the identification point of a target word was taken as the point at which the correct answer was written and not subsequently changed (e.g., Salasoo and Pisoni, 1985). That point, representing x gates of the word, was expressed in terms of the percentage of the target word needed for identification to allow for differences in the absolute duration of words. For example, a subject needing only four gates to identify a word with a total of six gates required 66.7% of that word. If a word was identified before any portion of it had been presented, the

[8] In video editing, the minimum unit is one frame. After reviewing sample stimuli with various gate durations, a duration of two frames was selected taking into consideration the amount of information per frame, potential subject fatigue, the number of words in the stimulus set, and a degree of comparability with previous studies (Hardison, in press, 2004).

figure used was zero. If all of the word was needed, the result was 100%. If the word was not identified by its acoustic offset, the result was expressed as one gate more than the total number in the word and the percentage exceeded 100. Data were analyzed with a four-factor ANOVA: modality (AV, A-only), speech style (unscripted, scripted), word length (one or two syllables), and IC category (8 levels). Tukey's HSD post hoc tests were conducted.

3.2.1.1. Native speaker results. Each of the figures in this paper indicates the mean percentage of the target word needed for accurate identification (vertical axis). For comparison with NNS findings and the results in Experiment 2 (sentence condition), both the data for the monosyllabic and those for the bisyllabic words are given; for ease of exposition, they are shown in separate figures. Results are grouped according to the initial consonant category (horizontal axis) of the targets. Each category has four sets of bars, two for AV presentation and two for A-only. These presentation modalities are further divided into unscripted and scripted speech styles.

3.2.1.1.1. Talker A. Identification of words produced by Talker A revealed significant main effects of modality [$F(1,52) = 363.62, p < .0001$], speech style [$F(1,52) = 12.01, p = .001$], word length (one or two syllables) [$F(1,48) = 4.57, p < .05$], and initial consonant category [$F(7,48) = 3.23, p < .01$]. Identification was earlier in AV vs. A-only presentation. In general, bisyllabic words required less information than monosyllabic ones for identification but this varied across IC categories. As anticipated, there was considerable variation in terms of the effects of speech style and initial consonant category, which was reflected in a significant four-way interaction involving the above factors [$F(7,3420) = 4.62, p < .0001$].

As shown in Fig. 1, identification of excised monosyllabic words in AV presentation was earlier across initial consonant categories compared to A-only. Significant differences were found for those words beginning with visually salient palato-alveolar consonants (e.g., *chef*) and /θ/ (e.g., *theme*). Differences between scripted and unscripted speech were more evident with A-only (vs. AV) presentation of words beginning with bilabial and palato-alveolar consonants. Words beginning with /ɹ/

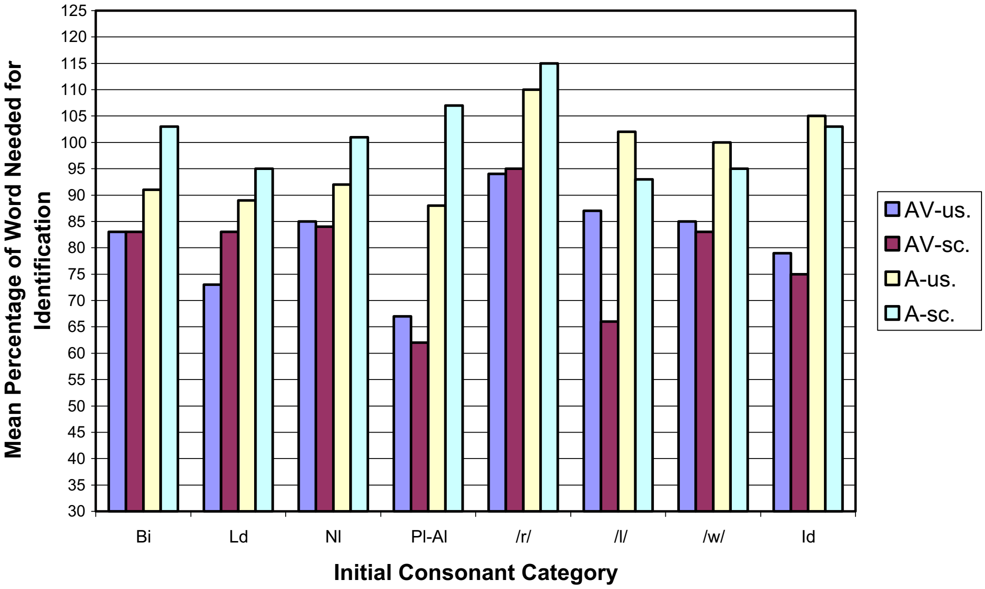


Fig. 1. Talker A: Identification of monosyllabic excised words by native speakers. Note: Bi = bilabial, Ld = labiodental, Nl = nonlabial, Pl-Al = palato-alveolar, Id = interdental, us = unscripted, sc = scripted.

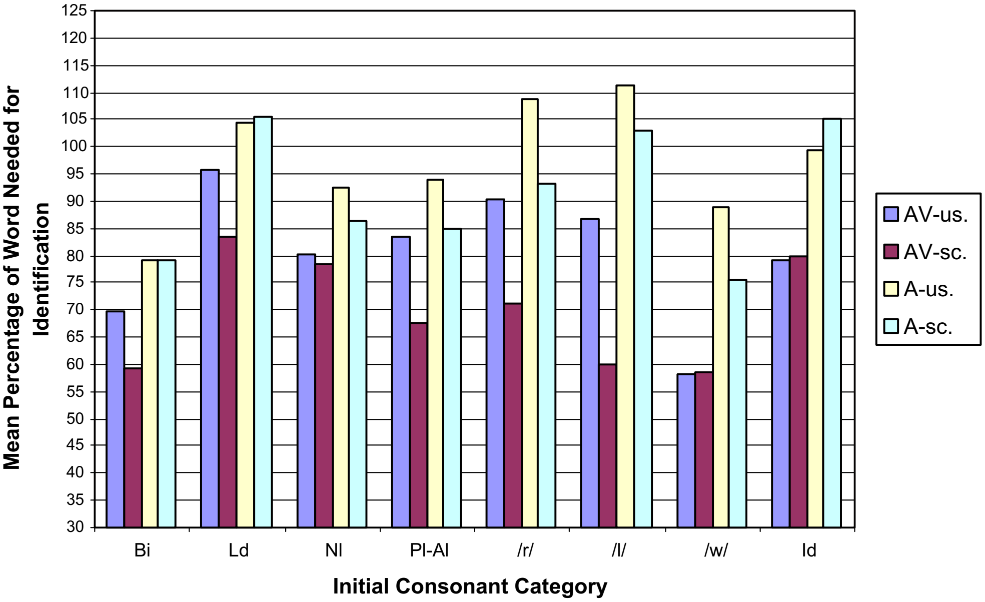


Fig. 2. Talker A: Identification of bisyllabic excised words by native speakers Note: Bi = bilabial, Ld = labiodental, Nl = nonlabial, Pl-Al = palato-alveolar, Id = interdental, us = unscripted, sc = scripted.

were identified significantly earlier in scripted vs. unscripted speech in both AV and A-only presentations. This pattern also obtained for those beginning with /w/ and /θ/; however, this trend was reversed for the remainder of the IC categories in A-only presentations where unscripted speech showed earlier identification, and to a significant degree for palato-alveolars. Bilabial, nonlabial and /ɹ/-initial words showed comparable findings for both AV speech styles. For this talker, /ɹ/-initial words required the most information for identification among the IC categories in AV presentation; without visual cues, words generally were not identified before their acoustic offset.

As shown in Fig. 2, AV vs. A-only presentation resulted in earlier identification across consonant categories; for /w/- and /θ/-initial words, identification was significantly earlier for AV presentation in both speech styles, and for AV words in scripted speech beginning with consonants from the bilabial, labiodental and palato-alveolar categories, /ɹ/, and /l/, and A-only words beginning with /ɹ/ and /w/. In contrast to Fig. 1, there was little evidence of earlier identification for words in unscripted vs. scripted speech.

The apparent influence of style, per se, on the discernibility of some lip movements may be confounded with rate of speech. Some authors regard casual/conversational/spontaneous speech as synonymous with fast speech (e.g., Dalby, 1986).[9] In the present study, word duration was measured in terms of number of gates for each word in the unscripted and scripted speech styles (see Appendix A). Duration for monosyllabic words was significantly shorter in scripted speech (mean = 5.47 gates) than unscripted (mean = 6.19 gates) [$t = -3.97, \mathrm{df} = 31, p < .05$]. For bisyllabic words, the duration was also significantly shorter for scripted speech (mean = 6.22 gates) than unscripted (mean = 6.94) [$t = -3.97, \mathrm{df} = 31, p < .05$].

3.2.1.1.2. Talker B. Identification of words produced by Talker B also revealed main effects of modality [$F(1,52) = 98.22, p < .0001$], speech style [$F(1,52) = 68.40, p < .0001$], and word length [$F(1,42) = 9.27, p < .01$], but not initial consonant category [$F(6,42) = 1.87, p = .11$]. However, there was a significant Modality × Style × IC category

[9] McAllister (1991) did not measure rate differences for spontaneous and read speech.

interaction [$F(6, 2986) = 3.31, p < .01$]. Identification of excised monosyllabic words again revealed a significant advantage for AV presentation compared to A-only for Talker B. As shown in Fig. 3, in both AV and A-only presentations, words beginning with bilabials, labiodentals, /ɹ/, /l/ and /w/ were identified significantly earlier in scripted vs. unscripted speech. Only small differences were noted across all four bars in the nonlabial category and for the palato-alveolars where unscripted speech produced earlier identification in both AV and A-only presentations. In contrast to Talker A, less information was needed to identify /ɹ/-initial words regardless of modality or style of speech.

A significant Modality × Style × IC category × Word Length [$F(6, 2986) = 2.59, p < .05$] interaction also obtained for this talker. Bisyllabic words generally required less information for identification than monosyllabic ones. As shown in Fig. 4, identification was earlier for scripted vs. unscripted speech in AV and A-only presentations with the exception of words beginning with nonlabials (in AV) and /w/.

There was also a contrast between the two talkers in terms of the relationship between rate of speech and speech style. For Talker B, the duration for monosyllabic words in scripted (mean = 5.79 gates) and unscripted speech (mean = 5.36) was significantly different [$t = 4.5, \mathrm{df} = 27, p < .05$]. The difference in duration between bisyllabic words in scripted (mean = 6.64 gates) and unscripted speech (6.25) was not significant [$t = 2.09, \mathrm{df} = 27, \mathrm{n.s.}$].

3.2.1.2. Nonnative speaker results

3.2.1.2.1. Talker A. Similar analyses were conducted on the data from the NNSs. There were significant main effects of modality [$F(1, 28) = 72.41$, $p < .0001$], speech style [$F(1, 28) = 12.23, p < .001$], initial consonant category [$F(7, 28) = 8.56$, $p < .001$], and word length (one or two syllables) [$F(1, 28) = 8.92, p < .01$]. There was a significant Modality × Style × IC category interaction [$F(7, 1949) = 7.02, p < .001$]. The discussion of results focuses on those for the bisyllabic words where there were fewer instances of mean percentages over 100 that are indicative of failure to identify the target before its acoustic (and, in AV presentation, visual) offset. As with the NSs, the results for the NNSs as shown in Fig. 5 revealed earlier identification of words across IC categories for AV vs. A-only presentation, and in general, for

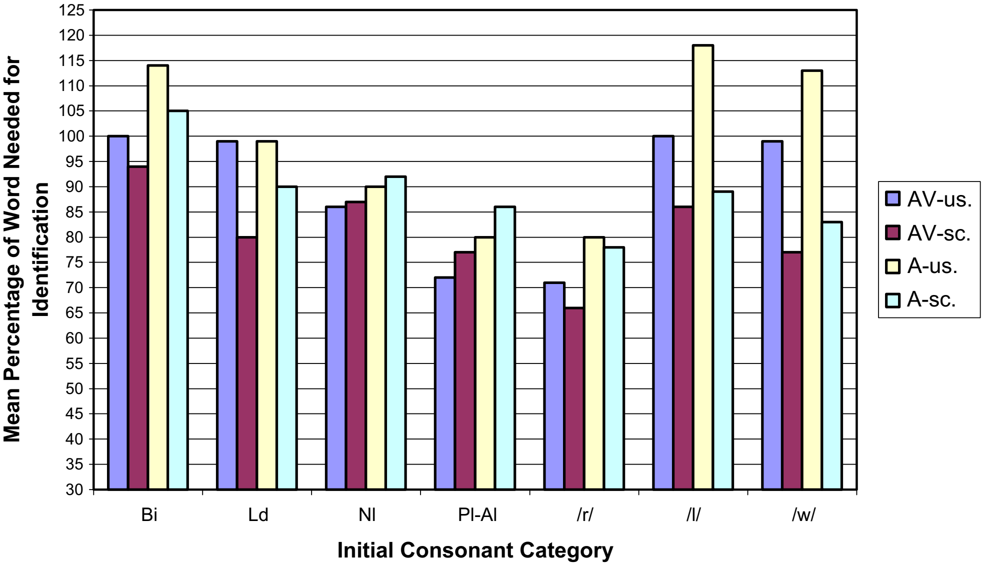


Fig. 3. Talker B: Identification of monosyllabic excised words by native speakers. Note: Bi = bilabial, Ld = labiodental, Nl = nonlabial, Pl-Al = palato-alveolar, us = unscripted, sc = scripted.

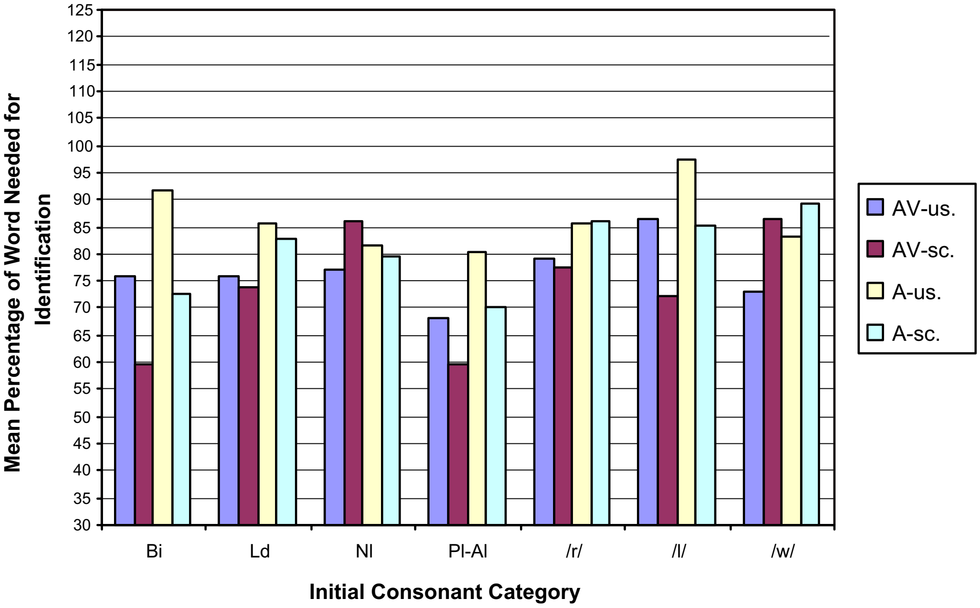


Fig. 4. Talker B: Identification of bisyllabic excised words by native speakers. Note: Bi = bilabial, Ld = labiodental, Nl = nonlabial, Pl-Al = palato-alveolar, us = unscripted, sc = scripted.

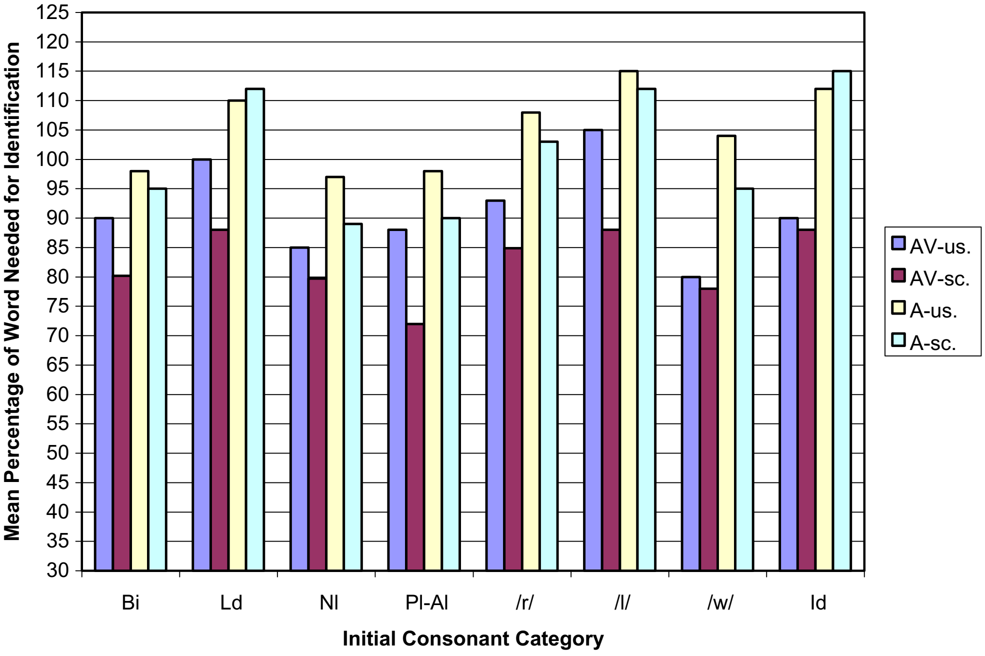


Fig. 5. Talker A: Identification of bisyllabic excised words by nonnative speakers. Note: Bi = bilabial, Ld = labiodental, Nl = nonlabial, Pl-Al = palato-alveolar, Id = interdental, us = unscripted, sc = scripted.

scripted vs. unscripted speech. The advantage of visual cues was significantly greater for words beginning with /ɹ/, /l/, /w/ and /θ/.

Bilabial sounds and nonlabials such as /t/ and /k/ are generally not problematic for this population. In contrast, mean percentages for words beginning with /f/, /ɹ/, /l/, and /θ/ indicated they were not identified before their acoustic offset in A-only presentation regardless of speech style. This is compatible with the oft-cited difficulties Japanese speakers have in perceiving these sounds and with the pretest data (prior to perceptual training) from a previous study (Hardison, in press).

3.2.1.2.2. Talker B. As with Talker A, the results for the NNSs focus on those from the bisyllabic words where again most stimuli were identified before their offsets. There were significant effects of modality [$F(1,28) = 52.30, p < .0001$], speech style [$F(1,28) = 11.76$, $p < .001$], word length [$F(1,28) = 9.04, p < .01$], and initial consonant category [$F(6,28) = 4.53$, $p < .01$].

In general, visual cues and scripted speech resulted in significantly earlier word identification. As shown in Fig. 6, the visual salience of palato-alveolar sounds continued to be an advantage, and AV presentation showed a significant advantage for /ɹ/- and /l/-initial words. There was also a significant Modality × Style × IC category interaction [$F(6,1702) = 8.96, p < .001$]. The degree of influence of speech style varied across IC categories showing a reverse pattern in AV presentation of targets beginning with nonlabials and /w/ (unscripted speech earlier).

3.2.2. V-only modality

Two different approaches were taken in the analysis of the V-only data. Recall that these data could be obtained only from NSs. In both approaches, mono- and bisyllabic words were combined. First, the percentage of words identified correctly by the end of the final gate of a stimulus was calculated for each talker. Because of the lack of a unique correspondence between articulatory gesture and sound, an estimate was also made of how well observers identified the visual categories to which the phones of the stimulus belong, and this figure was then divided by the total number of phones in the word to adjust for word length (Demorest et al., 1996).

The determination of category identification was made by a group of three judges, native speakers of English, each with a background in phonetics. Any disagreements were discussed until a

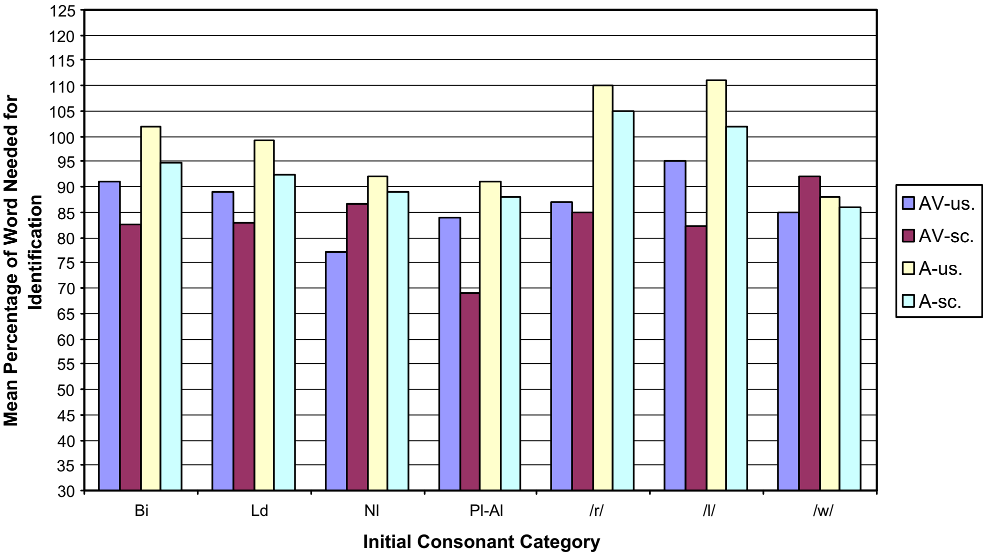


Fig. 6. Talker B: Identification of bisyllabic excised words by nonnative speakers. Note: Bi = bilabial, Ld = labiodental, Nl = nonlabial, Pl-Al = palato-alveolar, us = unscripted, sc = scripted.

consensus was reached. To compare the visual categories of a stimulus with those of an observer's response, we assigned a zero for those words correctly identified or those for which the stimulus and response phones belonged to the same visual category (e.g., stimulus *singer* and response *seeker*), and a 1 for each visual category mismatch or omission in a response. This result was converted to a mean percentage of visual categories accurately identified for each talker. However, this gross-level analysis should be regarded as an estimate only and with some caution. It assumes a sequence of somewhat normalized visual categories and phones, and there is no consensus in the literature on the sounds that comprise homophenous categories given context and talker variability (Kricos and Lesner, 1982). We used the following categories for word-initial sounds compatible with the talkers in this experiment and with other research using part of this database (Hardison, in press, 2004): /p,m/, /f/, /θ/, /ɹ/, /l/, /w/, /ʃ, tʃ, dʒ/, and nonlabials (/s, t, n, k/). The results provided a basis on which to evaluate differences in the visual discernibility of sounds in scripted vs. unscripted speech.

3.2.2.1. Talker A. Of the words produced by Talker A, 13% were identified correctly by at least six of eight subjects in both the scripted and unscripted speech styles. These words were characterized by initial consonant categories that have relatively salient lip positions such as bilabials (e.g., *mother, people*), labiodentals (e.g., *favor*) and palato-alveolars (e.g., *jacket*). Assessment of visual category identification reveals a mean accuracy of 28% (s.d. 2.39) for words produced in scripted speech and 31% (s.d. 1.62) for unscripted.

3.2.2.2. Talker B. Of the words produced by Talker B, 8% were identified correctly by at least six of eight subjects in unscripted speech and 18% in scripted. These words also began with visually salient consonants (e.g., *people, friend*). Evaluation of identification accuracy according to visual category raised the percentage for unscripted speech to 18% (s.d. 2.01), which was similar to the 19% (s.d. 2.87) accuracy result for scripted speech.

4. Experiment 2: sentence condition

4.1. Method

4.1.1. Subjects

In Experiment 2, 14 different NSs were randomly assigned to each of 12 experimental groups: the AV, A-only and V-only groups for each talker were further divided into groups presented with unscripted and scripted speech. As noted earlier (see Section 2.), an insufficient number of NNSs were available to participate in the sentence condition in the present study.

4.1.2. Materials and procedure

The recorded materials were the same as those in Experiment 1. In Experiment 2, the preceding sentence fragment was presented to subjects followed by the gated target word in AV, A-only and V-only presentations. The first presentation (or pass) of the sentence condition included the beginning of the sentence up to the point before the target word to allow for the possibility that a subject might predict the correct word without hearing and/or seeing any part of the target. The second presentation was like the first plus the first gate (first two frames) of the target word. Subsequent presentations provided the same information from the previous pass plus an additional gate until all of the target word had been presented.[10]

To measure the intelligibility of the sentence fragment preceding each target word, eight tapes were constructed (two modalities [A-only, AV] ×

[10] This study did not investigate the influence of subsequent sentence context on the identification of monosyllabic words because of time constraints. In addition, results for NSs revealed very few instances of these words not identified before their offsets, and these failures occurred only in A-only presentation of /ɹ/- and /w/-initial words produced by Talker A (see Fig. 7). Findings for the NNSs did reveal instances where monosyllabic words were not identified before their offset. It remains a question as to whether they would have continued to utilize subsequent context to identify the target. The results of exit interviews with NNSs conducted after a previous study (Hardison, in press) suggest that this strategy is unlikely at least for learners at intermediate levels of aural proficiency. Often when they do not know a word, they simply stop attending to the input.

two speech styles [unscripted, scripted] × two talkers [A, B]) using only these fragments. Each tape was presented to six different subjects who were asked to write down as much of each fragment as possible. Following McAllister (1991), each tape was played twice. After hearing the sentence fragments, the subjects in the A-only presentations were also asked to provide a global rating of style for the speech they had heard; that is, they were asked to rate the speech on a 7-point scale (1 and 2 represented "conversational"; 6 and 7 represented "reading from cue cards").

4.2. Results and discussion

Data for the sentence condition for each talker were submitted to ANCOVA with the amount of sentence context (number of words in the sentence preceding the target word) as the covariate. This was done because the amount of preceding context has a significant effect on word identification (Grosjean, 1980; McAllister, 1991), and it was not possible to confine selection of targets from the unscripted speech samples (the conversations) to those with the same number of preceding words or degree of contextual support. In addition, although the cloze task reduced the likelihood that the targets could be predicted based on context alone, contexts vary in their contribution of semantic, syntactic and pragmatic information.

4.2.1. AV and A-only modalities

4.2.1.1. Talker A. Identification of words produced by Talker A revealed significant main effects of modality [$F(1,99.96) = 135.64, p < .0001$], style [$F(1,99.96) = 135.64, p < .0001$], and initial consonant category [$F(7,48.04) = 4.00, p < .01$], but not word length [$F(1,48.16) = 0.45, p = .50$]. Because there was no significant effect of word length in the presence of context, compatible with the findings of Grosjean (1980), the results for monosyllabic words are given for comparison with the findings of the excised word condition. Recall that monosyllabic words for Talker A were the ones that showed more variability in identification performance in response to speech style differences.

As shown in Fig. 7, words were identified earlier in AV vs. A-only presentation in both speech styles and across initial consonant categories. Words beginning with /ɹ/ and /w/ were the only stimulus

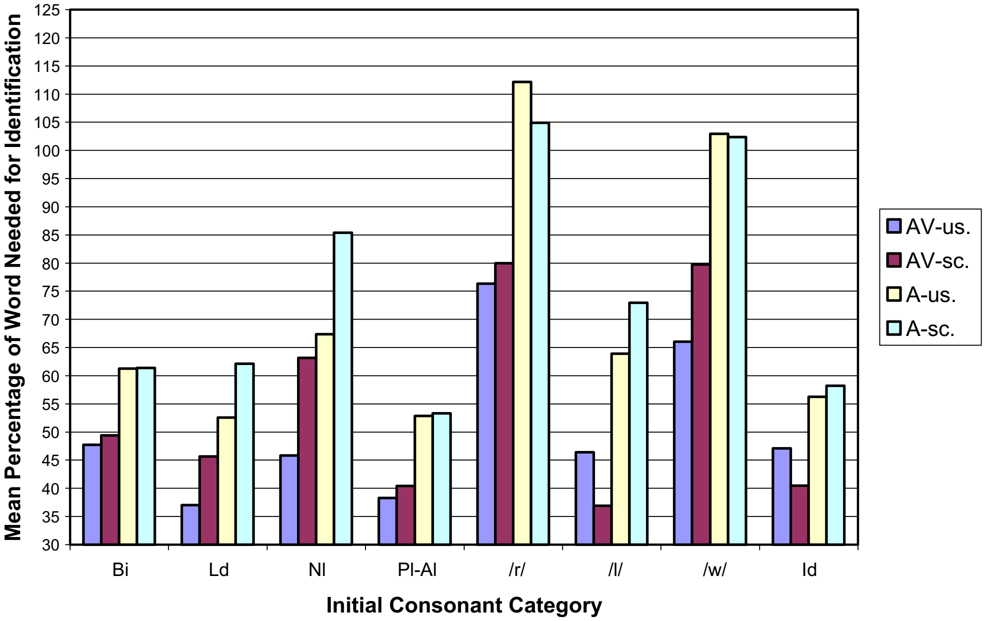


Fig. 7. Talker A: Identification of monosyllabic words in sentence context by native speakers. Note: Bi = bilabial, Ld = labiodental, Nl = nonlabial, Pl-Al = palato-alveolar, Id = interdental, us = unscripted, sc = scripted.

categories not identified in A-only presentation. Overall, identification was earlier for unscripted speech; however, as in Experiment 1, there was significant variability reflected in a Modality × Style × IC category interaction [$F(7, 3419) = 14.79, p < .0001$]. Inspection of the graph shows earliest identification for AV presentation of words beginning with palato-alveolar affricates and fricatives in both speech styles, labiodentals in unscripted speech, and /l/ and /θ/ in scripted speech. As anticipated from previous research, visual comparison of Figs. 1 and 7 revealed that word identification was earlier overall with sentence context than for excised words.

4.2.1.2. Talker B. Identification of words spoken by Talker B revealed significant main effects of modality [$F(1, 89.06) = 12.59, p < .001$] and style [$F(1, 57.62) = 20.63, p < .0001$], but not IC category [$F(6, 42.02) = 1.10, p = .38$] nor word length [$F(1, 42.18) = 1.78, p = .19$]. Words were identified earlier when visual cues were present, and in scripted vs. unscripted speech for Talker B. As in Experiment 1, there was a significant Modality × Style × IC category interaction [$F(6, 2984) = 6.04, p < .0001$]. Visual inspection of Fig. 8 shows the earliest identification points for words in AV presentation beginning with labiodentals and /ɹ/, the latter category in marked contrast to the findings for Talker A. The largest differences in terms of speech style were found for /l/ and /w/ in A-only presentation.

4.2.1.3. Sentence fragment intelligibility and perception of speech style. The intelligibility of the sentence fragment preceding the gated word was calculated as the mean percentage of words accurately identified in the correct order. In the scripted version, intelligibility for both talkers in both AV and A-only presentations was 100%. In the unscripted version, for Talker A, intelligibility was 100% for AV presentation and 96% for A-only; for Talker B, results were 98% for AV and 95% for A-only. Subjects also correctly distinguished the two speech styles on the 7-point scale by assigning a mean rating of 1.6 for Talker A and 1.2 for Talker B to the unscripted (conversational) speech fragments, and 6.3 for Talker A and 6.7 for Talker B to the scripted (reading from cue cards) version.

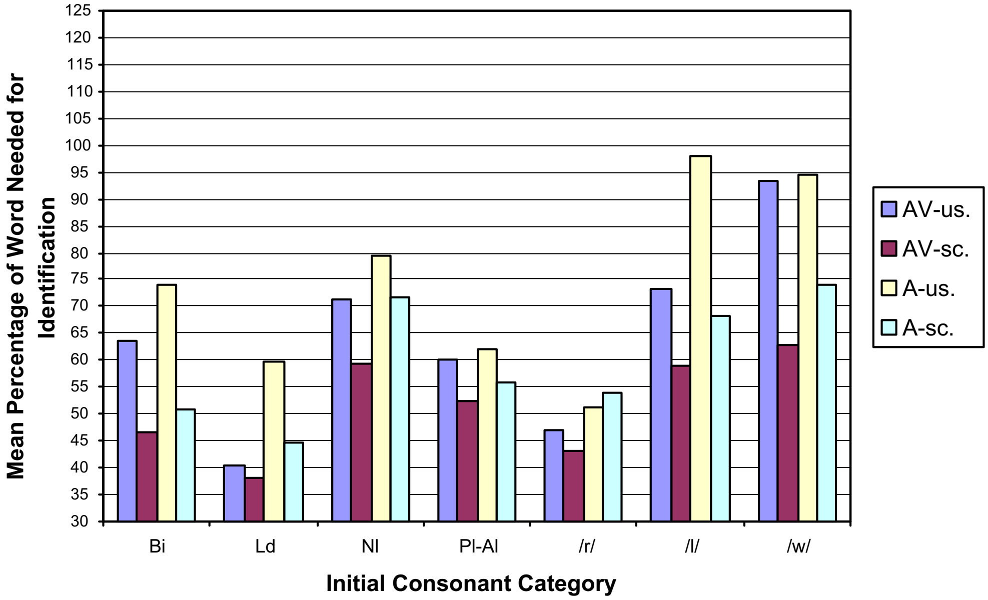


Fig. 8. Talker B: Identification of monosyllabic words in sentence context by native speakers. Note: Bi = bilabial, Ld = labiodental, Nl = nonlabial, Pl-Al = palato-alveolar, us = unscripted, sc = scripted.

4.2.2. V-only modality

In the V-only sentence condition, both auditory and visual information was presented for the sentence fragment; the acoustic signal was deleted only for the gated target word. Results indicated that visual categories were identified more accurately when sentence context preceded the target word. For Talker A, the mean accuracy rates for word identification were similar again for unscripted (49%, s.d. 1.55) and scripted speech (46%, s.d. 1.97) as found in Experiment 1. The presence of context contributed to significantly greater identification accuracy in terms of visual categories [$F(3,252) = 5.08, p < .05$]. A similar pattern was found for Talker B (unscripted: 36% accuracy, s.d. 2.95; scripted: 38% accuracy, s.d. 2.20) although overall identification accuracy of visual categories was lower than that for Talker A.

5. General discussion

This study focused on the effects of different speech styles produced by two talkers on auditory-visual and auditory-only word identification by native and nonnative speakers of English using the gating paradigm. Speech style (unscripted vs. scripted), modality of presentation (auditory-visual, auditory-only), word length (mono- vs. bisyllabic words), and initial consonant category were treated as variables in excised word and sentence conditions. Previous research documented the effect of talker variability on auditory processing (e.g., Johnson and Mullennix, 1997), and the contribution of visual cues to spoken word identification by native and nonnative speakers of English using experimentally controlled stimuli (Hardison, 2003, in press). Some experiments that have focused on the investigation of effects of style have concluded that conversational speech is less intelligible than "clear" speech; however, in those studies, "clear" represented hyperarticulated or exaggerated speech (e.g., Gagné et al., 1994; Helfer, 1997).

In the present study, talkers were not given any articulation instructions and there were no instances of hyperarticulated sounds. Style differences emerged as a consequence of whether speech was unscripted as in conversations where a talker is more focused on the meaning of an utterance, or scripted where attention may be directed toward articulation. Contrary to assumptions in the literature that regard casual or conversational speech as synonymous with rapid speech (e.g., Dalby, 1986), scripted speech for Talker A was significantly more rapid (i.e., fewer gates in duration) than the unscripted version for mono- and bisyllabic words; however, for Talker B, only the duration of monosyllabic words differed significantly between the two styles, and in this case, scripted speech was longer.

Although speech style was classified into two categories—scripted and unscripted—and the scalar ratings of subjects' perceptions of speech style corroborated these categories, their effect on word identification performance was not similarly binary but quite variable across talkers and phonetic environments for both native and nonnative listeners. In Experiment 1 (excised words), the AV advantage was particularly evident in identification by native speakers of words beginning with visually salient articulations such as palato-alveolars, /w/, /θ/, and /l/ when produced by Talker A whose gesture for word-initial /l/ approximates an interdental position (see Appendix B). Monosyllabic words showed more variability than bisyllabic ones in identification performance in response to style differences and initial consonant category. Overall visual cues resulted in significantly earlier identification for words spoken by Talker B as did scripted vs. unscripted speech. Interactions between style and modality revealed that for words beginning with /f/, /l/, and /w/, the contribution of a scripted style of speech outweighed that of visual cues in terms of identification. A marked contrast was noted in the identification of /ɹ/-initial words which were the earliest to be identified by native speakers in both styles and modalities for speech produced by Talker B and the latest for that by Talker A. For nonnative speakers, mean identification of monosyllabic words was generally not accomplished before their acoustic offset. Findings for bisyllabic words showed that AV vs. A-only presentation resulted in earlier identification of words produced by both talkers in both speech styles. Identification of words beginning with the most problematic sounds for this population benefited most from visual cues even in the absence of any training. These included /ɹ/, /w/ and /θ/ in unscripted speech by Talker A, and /ɹ/ and /l/ in both speech styles by Talker B.

In Experiment 2 (sentence condition), only native speakers were available to participate. Consistent with findings in other studies, identification of words produced by both talkers was earlier with context than for excised words, and in AV vs. A-only presentation; however, variability again was evident in significant interactions between modality, speech style, and initial consonant category. With context, word length was not significant. Word identification overall was earlier for unscripted speech produced by Talker A, and scripted speech produced by Talker B.

Despite the considerable variability found in the results of the current study, which is consistent with the recent auditory processing literature on talker variability, some clear patterns emerged: (1) visual cues contributed to earlier word identification for native and nonnative speakers of English across speech styles, talkers, and initial consonant categories, in excised word and sentence conditions; (2) visual category identification accuracy by native speakers was comparable for unscripted and scripted speech style within each talker, but higher for speech produced by Talker A compared to Talker B; (3) nonnative speakers required more of a word for identification than native speakers; (4) in general, excised bisyllabic words required less information to be identified than monosyllabic ones; however, context in Experiment 2 eliminated a significant effect of word length; (5) for nonnative speakers, the advantage of visual cues in word identification was accentuated for problematic sounds such as /ɹ/ and /l/, and /θ/ for Talker A consistent with other studies (Hardison, 2003); however, talker variability and style differences were apparent.

In conclusion, the findings of this study demonstrate complex interactions of various information sources in the processing of spoken language, and underscore both the priming role of visual cues in auditory-visual speech processing, and the context- and talker-dependent nature of spoken language processing. For words produced by both talkers, style of speech was a factor in the identification process for both native and nonnative speakers but not always in the same direction; that is, within the data for each talker, scripted speech was not always the more facilitative style as might be predicted from previous studies in the literature. Style is not necessarily indicative of rate of speech, which may also contribute to the discernibility of articulatory gestures. Variable performance across phonetic and visual contexts as well as talkers is consistent with an episodic view that memory encoding of speech involves storage of the attended perceptual details of individual episodes that preserve both contextual variability and the indexical properties of speech (e.g., Johnson and Mullennix, 1997; Nygaard et al., 1995; Pisoni, 1993). The results also emphasize the benefit of providing adult second-language learners with input involving variability in terms of talkers (voices and faces), phonetic environments, and speech styles for robust perceptual category development.

Acknowledgments

I gratefully acknowledge the technical expertise of Dave Derkacy and Dave Rust in recording and editing, the participation of Terese Thonus and Mary McCarty, and the assistance of the MSU Department of Statistics and Probability in the statistical analysis of the data. Preliminary findings were presented at AVSP 2001: International Conference on Auditory-Visual Speech Processing, Aalborg, Denmark.

Appendix A

Following are the gated words used in this study produced by Talkers A and B including the duration (in number of gates) in unscripted and scripted speech, and the number of lexical neighbors as an indication of neighborhood density (from Nusbaum et al., 1984).[11] Stimuli are

[11] The data for neighborhood density were obtained from the Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405, USA.

grouped according to initial consonant category: bilabial /p,m/, labiodental /f/, nonlabial /s, t, n, k/, palato-alveolar /ʃ, dʒ, t ʃ/, /ɹ/, /l/, /w/, and /θ / (/θ/ for Talker A only).

Talker A: one-syllable words

Gated word	Duration unscripted	Duration scripted	Neighbors
Pink	6	4	10
Plant	8	6	2
Place	8	5	4
Paint	7	5	9
Film	7	6	6
Fact	5	5	3
Fish	6	6	9
Five	5	5	10
Tasks	9	7	7
Song	7	5	10
Test	6	4	14
Skin	6	6	11
Jeans	6	6	3
Chef	7	7	6
Judge	8	8	5
Jazz	6	6	6
Wrist	5	6	11
Rent	4	4	15
Round	9	9	10
Wrong	7	5	10
Luck	4	4	1
Land	7	6	9
Love	7	5	10
Lamp	5	5	8
White	4	4	7
Watch	6	6	5
Wish	5	5	11
Web	5	5	8
Theme	7	5	8
Thing	5	5	10
Thick	5	4	12
Thumb	6	6	13

Talker A: two-syllable words

Gated word	Duration unscripted	Duration scripted	Neighbors
Mother	8	7	2
Music	7	5	0

Appendix A (*continued*)

Gated word	Duration unscripted	Duration scripted	Neighbors
People	4	4	4
Program	6	6	0
Family[a]	6	6	0
Favor	4	4	3
Feature	8	5	2
Feedback	9	6	0
College	8	7	3
Contact	7	7	0
Custom	8	8	0
Topic	5	6	2
Jacket	9	7	–
Shopping	9	8	–
Shelter	7	7	1
Shower	7	7	6
Reason	9	7	5
Razor	6	6	2
Writing	5	5	5
Raisin	6	5	5
lecture	4	4	0
Lady	5	5	5
Lazy	7	6	6
Leather	7	7	10
Water	6	6	2
Weakness	9	8	–
Wallet	7	6	0
Wagon	7	7	0
Thriller	9	6	–
Thunder	8	7	2
Thousand	9	8	0
Theory	6	6	5

Note: – indicates that the word did not appear in the database.

[a] Pronounced with two syllables.

Talker B: one-syllable words

Gated word	Duration unscripted	Duration scripted	Neighbors
Path	3	4	14
Pants	4	4	8
Park	2	2	12
Page	5	6	14

Appendix A (*continued*)

Gated word	Duration unscripted	Duration scripted	Neighbors
Field	7	7	6
Farm	6	6	4
Fault	5	6	6
Friend	5	6	3
Teach	6	6	11
Tense	5	5	9
Size	7	7	10
Smile	7	7	4
Chance	5	5	7
Choice	5	6	3
Change	7	7	2
Charge	7	7	6
Rush	7	8	12
Risk	6	6	5
Rent	5	5	15
Rinse	6	6	4
Lunch	5	6	7
Left	4	4	8
Love	4	5	10
Large	6	7	6
Wish	3	4	11
Wind (n.)	5	6	8
Wise	6	7	13
Wave	7	7	–

Note: – indicates that the word did not appear in the database.

Talker B: two-syllable words

Gated word	Duration unscripted	Duration scripted	Neighbors
People	4	5	4
Partner	7	8	1
Product	6	7	0
Pocket	5	6	7
Factor	8	8	0
Future	7	6	0
Famous	7	7	0
Feature	6	6	2
Salad	5	6	3
Tennis	7	7	3
Season	6	8	2

Appendix A (*continued*)

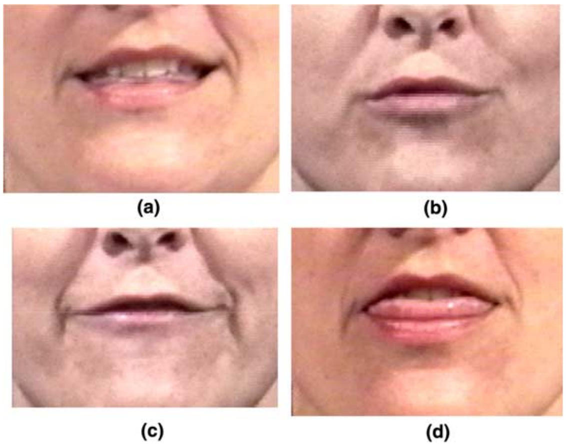
Gated word	Duration unscripted	Duration scripted	Neighbors
Nature	7	7	3
Challenge	8	9	0
Charming	7	8	–
Chapter	6	6	1
Channel	7	7	5
Reason	4	6	5
Restaurant[a]	6	8	0
Region	6	6	1
Refund	6	7	0
Lucky	5	4	1
Learning	5	6	1
Level	5	6	6
Laundry	6	7	0
Weekend	7	7	0
Witness	7	6	0
Waitress	7	6	0
Wander	8	6	5

Note: – indicates that the word did not appear in the database.

[a] Pronounced with two syllables.

Appendix B

Each of the frames below shows the articulatory gesture produced by Talker A at the beginning of the following words: (a) /l/ 'leather', (b) /ɹ/ 'reason', (c) /f/ 'feature', and (d) /θ/ 'thunder'.



References

Bard, E.G., Shillcock, R.C., Altmann, G.T.M., 1988. The recognition of words after their acoustic offsets in spontaneous speech: effects of subsequent context. Percept. Psychophys. 44, 395–408.

Berger, K.W., 1972. Speechreading: Principles and Methods. National Education Press, Baltimore.

Bradlow, A.R., Torretta, G.M., Pisoni, D.B., 1996. Intelligibility of normal speech I: global and fine-grained acoustic–phonetic talker characteristics. Speech Commun. 20, 255–272.

Cotton, S., Grosjean, F., 1984. The gating paradigm: a comparison of successive and individual presentation formats. Percept. Psychophys. 35, 41–48.

Craig, C., Kim, B., 1990. Effects of time gating and word length on isolated word-recognition performance. J. Speech Hear. Res. 33, 808–815.

Dalby, J., 1986. Phonetic Structure of Fast Speech in American English. Indiana University Linguistic Club Publications, Bloomington, IN.

Demorest, M.E., Bernstein, L.E., DeHaven, G.P., 1996. Generalizability of speechreading performance on nonsense syllables, words, and sentences: subjects with normal hearing. J. Speech Hear. Res. 39, 697–713.

Fisher, C.G., 1968. Confusions among visually perceived consonants. J. Speech Hear. Res. 11, 796–804.

Gagné, J.-P., Masterson, V., Munhall, K.G., Bilida, N., Querengesser, C., 1994. Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. J. Acad. Rehab. Audiol. 27, 135–158.

Garlock, V.M., Walley, A.C., Metsala, J.L., 2001. Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. J. Mem. Lang. 45, 468–492.

Goldinger, S.D., Luce, P.A., Pisoni, D.B., 1989. Priming lexical neighbors of spoken words: effects of competition and inhibition. J. Mem. Lang. 28, 501–518.

Grosjean, F., 1980. Spoken word recognition processes and the gating paradigm. Percept. Psychophys. 28, 267–283.

Grosjean, F., 1985. The recognition of words after their acoustic offset: evidence and implications. Percept. Psychophys. 38, 299–310.

Grosjean, F., 1996. Gating. Lang. Cogn. Proc. 11, 597–604.

Hardison, D.M., 1999. Bimodal speech perception by native and nonnative speakers of English: factors influencing the McGurk effect. Lang. Learn. 49, 213–283.

Hardison, D.M., 2003. Acquisition of second-language speech: effects of visual cues, context, and talker variability. Appl. Psycholing. 24, 495–522.

Hardison, D.M., in press. L2 spoken word identification: effects of perceptual training, visual cues, and phonetic environment. Appl. Psycholing.

Hardison, D.M., 2004. Role of context in spoken word identification by nonnative speakers. Unpublished raw data.

Helfer, K.S., 1997. Auditory and auditory-visual perception of clear and conversational speech. J. Speech Lang. Hear. Res. 40, 432–443.

Johnson, K., Mullennix, J.W. (Eds.), 1997. Talker Variability in Speech Processing. Academic Press, San Diego.

Kricos, P.B., Lesner, S.A., 1982. Differences in visual intelligibility across talkers. Volta Rev. 84, 219–225.

Lively, S.E., Logan, J.S., Pisoni, D.B., 1993. Training Japanese listeners to identify English /r/ and /l/. II: the role of phonetic environment and talker variability in learning new perceptual categories. J. Acoust. Soc. Amer. 94, 1242–1255.

Luce, P.A., 1986. Neighborhoods of words in the mental lexicon. Research on Speech Perception, Technical Report No. 6, Indiana University, Bloomington, IN.

Marslen-Wilson, W., Welsh, A., 1978. Processing interactions and lexical access during word recognition in continuous speech. Cog. Psych. 10, 29–63.

McAllister, J., 1988. The use of context in auditory word recognition. Percept. Psychophys. 44, 94–97.

McAllister, J., 1991. The processing of lexically stressed syllables in read and spontaneous speech. Lang. Speech 34, 1–26.

Metsala, J.L., 1997. An examination of word frequency and neighborhood density in the development of spoken-word recognition. Mem. Cogn. 25, 47–56.

Munhall, K.G., Tohkura, Y., 1998. Audiovisual gating and the time course of speech perception. J. Acoust. Soc. Amer. 104, 530–539.

Nusbaum, H.C., Pisoni, D.B., Davis, C.K., 1984. Sizing up the Hoosier mental lexicon: measuring the familiarity of 20,000 words, Research in Speech Perception, Progress Report 10, Indiana University, Bloomington, IN.

Nygaard, L.C., Sommers, M.S., Pisoni, D.B., 1995. Effects of stimulus variability on perception and representation of spoken words in memory. Percept. Psychophys. 57, 989–1001.

Picheny, M.A., Durlach, N.I., Braida, K.D., 1986. Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. J. Speech Hear. Res. 29, 434–446.

Picheny, M.A., Durlach, N.I., Braida, K.D., 1989. Speaking clearly for the hard of hearing: an attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. J. Speech Hear. Res. 29, 600–603.

Pisoni, D.B., 1993. Long-term memory in speech perception: some new findings on talker variability, speaking rate and perceptual learning. Speech Commun. 13, 109–125.

Salasoo, A., Pisoni, D.B., 1985. Interaction of knowledge sources in spoken word identification. J. Mem. Lang. 24, 210–231.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Amer. 26, 212–215.

Summerfield, Q., 1979. Use of visual information for phonetic perception. Phonetica 36, 314–331.

Tyler, L., 1984. The structure of the initial cohort: evidence from gating. Percept. Psychophys. 36, 417–427.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., Jones, C.J., 1977. Effects of training on the visual recognition of consonants. J. Speech Hear. Res. 20, 130–145.