Applied Psycholinguistics **26** (2005), 579–596
Printed in the United States of America
DOI: 10.1017.S0142716405050319

Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment

DEBRA M. HARDISON
Michigan State University

ADDRESS FOR CORRESPONDENCE
Debra M. Hardison, Michigan State University, Department of Linguistics and Languages, A-714 Wells Hall, East Lansing, MI 48824. E-mail: hardiso2@msu.edu

ABSTRACT
Experiments using the gating paradigm investigated the effects of auditory–visual (AV) and auditory-only perceptual training on second-language spoken-word identification by Japanese and Korean learners of English. Stimuli were familiar bisyllabic words beginning with /p/, /f/, /ɹ/, /l/, and /s, t, k/ combined with high, low, and rounded vowels. Results support the priming role of visual cues in AV speech processing. Identification was earlier with visual cues and following training, especially for words beginning with /ɹ/ and /l/, which also showed significant effects of adjacent vowel. For the Japanese, the AV advantage in identifying /ɹ/- and /l/-initial words was accentuated following training. Findings are discussed within a multimodal episodic model of learning.

In face to face communication, an important source of information about the sounds a speaker is producing comes from visual cues such as lip movements. We are perhaps most aware of the contribution of visual information when compensating for noise or impaired hearing ability (e.g., Summerfield, 1979); however, some researchers have proposed that multimodal speech is the primary mode of speech perception (e.g., Rosenblum, 2002). The influence of lipread information on speech perception has been demonstrated through numerous studies over the past several decades (see Massaro, 1998, for a review). In many of these experiments, stimuli involved consonant–vowel (CV) syllables in which the visual and acoustic cues did not match. For example, when presented with an acoustic /ba/ dubbed onto a visual /ga/ native speakers of English often report the perceptual illusion "da," a phenomenon that came to be known as the McGurk effect (e.g., McGurk & MacDonald, 1976). The magnitude of this effect is influenced by the observer's familiarity with the talker (Walker, Bruce, & O'Malley, 1995), intelligibility of the cues, which varies across talkers (Kricos & Lesner, 1982), and phonetic environment. For example, anticipatory rounding of the mouth may extend over as many as four consonants preceding a rounded vowel (Daniloff & Moll, 1968). The effect is also subject to the degree to which observers perceive

© 2005 Cambridge University Press 0142-7164/05 $12.00

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

the acoustic and visual cues to match. This assessment of the cues' relationship is influenced by the native or nonnative status of the observer's perceptual categories (Hardison, 1999).

In Hardison (1999), both matched and mismatched stimulus conditions were created using American English (AE) /p, f, ɹ, w, t, k/ combined with /a/ in a CV syllable, and presented to native speakers and learners of English as a second language (L2) whose first languages (L1) were Japanese, Korean, Spanish, and Malay. Results indicated a significant increase in identification accuracy of /f/ and /ɹ/ for the Japanese and Korean learners when corresponding visual cues were available. In mismatched conditions, visual nonlabials /t, k/, generally not problematic for learners, contributed significantly to the percept when combined with auditory labial /p/; however, when nonlabials represented the acoustic cue paired with visual /p/, the acoustic cues were correctly identified. These examples demonstrate that visual cues can be informative for L2 learners, and the more informative cue contributes the most to the perceptual outcome.

In a recent study, auditory–visual (AV) versus auditory-only (A-only) training using minimal pairs produced significantly greater improvement in the perceptual accuracy of AE /ɹ/, /l/, /f/ (vs. /p/) and /θ/ (vs. /s/) by intermediate-level Japanese and Korean learners (Hardison, 2003). The current study's objective was to investigate the effect of this perceptual training on the L2 learners' identification of words beginning with those sounds, and to compare AV versus A-only input.

Identification accuracy of acoustic and visual cues is based on the assessment of similarity between input and stored memory representations. Following multiple-trace memory theory, these representations are traces or episodes that preserve all attended details of individual perceptual events. Such a view is compatible with the variable performance across phonetic environments and talkers found in previous AV and A-only L2 perceptual training studies (e.g., Hardison, 2000, 2003; Lively, Logan, & Pisoni, 1993; Pisoni, 1993), suggesting that language learners utilize context-dependent exemplars or memory traces as representations to which input can be matched for identification. This is consistent with an episodic lexicon as the basis of the word identification process (e.g., Goldinger, 1997; Jacoby & Brooks, 1984; Jusczyk, 1993). In this process, traces are activated simultaneously in response to a stimulus on the basis of feature similarity (e.g., Hintzman, 1986). Therefore, improved stimulus identification by L2 learners has the potential to reduce competing activation among traces and reduce the number of candidates to the word identification process.

Competition for identification among activated items is influenced by similarity neighborhood (e.g., Luce, 1986), that is, the words phonetically similar to each other and to any given stimulus word. In general, identification is earlier for high-frequency words with few neighbors (e.g., Goldinger, Luce, & Pisoni, 1989) because a sparse neighborhood reduces confusability with the stimulus. The neighborhood for an L2 lexical item is subject to inhibitory or facilitative influence from the L1 depending upon the learner's accuracy in assessing similarity.

Of particular interest in the present study was the role that visual cues might play in word identification. In the production of speech, there is a natural temporal precedence for the articulatory gesture over the associated acoustic cue. Visual cues for place of articulation can precede the associated acoustic signal by more

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

than 100 ms (Smeele, 1994). Sekiyama and Sugita (2002) reported a delay of 600 ms between the completion of the lip closure for /b/ in the syllable /ba/ and the beginning of the acoustic signal. Results of perception experiments with native speakers using VCV stimuli suggested that this temporal precedence gives visual cues a priming role in AV speech processing (Munhall & Tohkura, 1998). The question arises as to whether training that increases L2 learners' identification accuracy of AV speech cues might enhance this priming role.

In the word identification process, the initial responsibility is on lower level discrimination and selection of the appropriate acoustic–phonetic pattern (e.g., Luce, 1986; Marslen–Wilson & Welsh, 1978) and articulatory gesture information (e.g., Massaro, 1987) stored in memory to match the input, regardless of the presence of semantic context (Tyler & Wessels, 1985). This early stage of processing of word-initial acoustic–phonetic and visual information was the focus of the present study.

The stimuli in this study were familiar bisyllabic content words that began with sounds that were included in perceptual training: /ɹ/, /l/, /p/, /f/, and /s/ combined with high, low, and rounded vowels. Words beginning with /θ/ (i.e., a sound contrasted with /s/ in training) were not included due to an insufficient number with a high rating of familiarity by a comparable group of learners. The nonlabial category was expanded to /s, t, k/. Because these sounds generally are not problematic for learners, they offer some baseline data, and a greater opportunity for the learners to have a sense of accomplishment with the task. The above consonants also represent a range of visual categories: bilabial /p/, labiodental /f/, and nonlabial /s/ (e.g., Owens & Blazek, 1985; Woodward & Barber, 1960). Generally, /ɹ/ and /l/ constitute separate visual categories (e.g., Berger, 1972), although there is no consensus on the categorization of all articulations as studies differ in talkers, methodologies, stimuli, and so forth. The selection of high, low, and rounded vowels was guided by the significant effect of adjacent vowel on the AV and A-only identification accuracy of /ɹ/ and /l/ by Japanese and Korean learners (Hardison, 2003). Phonetic environment also reduces the visual discernibility of a consonant (e.g., Benguerel & Pichora–Fuller, 1982; Kricos & Lesner, 1982). Bisyllabic words were chosen because monosyllabic ones produced in continuous speech are not always identified by native speakers before their acoustic offsets especially in A-only presentation (Bard, Shillcock, & Altmann, 1988; Grosjean, 1985).

In addition, the selection of consonants was influenced by their status in Japanese and Korean phonology (see Table 1 for details), and the problems these sounds (excluding the nonlabial group) pose for speakers of these two L1s. As numerous studies in the literature attest, the most problematic of these sounds are /ɹ/ and /l/, due to their talker and contextual variability, and status in the above L1 phonological systems. For the Japanese speakers, perceptual accuracy of /ɹ/ and /l/ in word-initial position is the most difficult where the L1 has a voiced apical flap. For the Korean speakers, the most difficult position is word final, where a voiced nonvelarized lateral occurs in the L1 in contrast to the English velarized /l/ (e.g., Hardison, 2003). In the present study, this difference permitted some observations about the influence of L1 phonology on word identification, a process heavily dependent on word-initial versus word-final sounds.

To investigate word identification, the gating paradigm was adapted for use with AV as well as A-only speech stimuli. The A-only gating format as implemented

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

Table 1. *Selected elements of Japanese and Korean phonology*

	Japanese	Korean
Flaps, liquids	Voiced apical flap in dental or alveolar region occurs in utterance-initial and intervocalic positions with acoustic, articulatory, and perceptual similarities to the AE flap; does not appear adjacent to other consonants; may occur in the environment of all Japanese vowels.	No liquid in word-initial position in native Korean and Sino-Korean words; may occur in loanwords with tendency to be pronounced as a flap; voiced alveodental flap intervocalically (or between vowel and glide).
	No lateral	Voiced apico-alveolar nonvelarized or clear /l/ in syllable-final position
Obstruents (/p, t, k/, /s/)	Unaspirated stops; weak lip involvement in the production of /p/	Voiceless lax, aspirated, and reinforced obstruents (e.g., /p′/, /p^h/, /P/); voicing is contextually and prosodically conditioned.
Labiodental fricative (/f/)	Weak bilabial fricative that may be labiodental in less careful speech	None
Vowels	Five phonemically distinct vowels /i, e, a, o, ɯ/ that may be long or short. Midfront vowel is between [e] and [ɛ]; variation in production of /a/. Midback vowel is between [o] and [ɔ]. High back unrounded /ɯ/ is transliterated as "u".	Phoneme inventory includes /i, e, ɛ, ɨ, ə, a, u, o/. Vowels /u, o/ are produced with lip protrusion and rounding. Word-initial syllables may have long vowels.

Note: References for Japanese are Price (1981), Tsujimura (1996), and Vance (1987). References for Korean are Cho (1987), Kim–Renaud (1974), Shin (1997), and Silva (1989). AE, American English.

by Grosjean (1980) involves successive presentations of increasing amounts of a target word. With each presentation, participants write down the word they think is being presented. Studies have shown no significant effect of the successive presentation format on subjects' performance (e.g., Cotton & Grosjean, 1984). In the current study, to assess change in word identification as a result of perceptual training, a pretest–posttest design was used involving parallel gating tasks. These tasks were comparable in content but not identical to avoid repetition effects (Hatch & Lazaraton, 1991). Stimuli were words excised from sentences spoken at a conversational rate rather than words produced in isolation. This provides stimuli that are more like the natural language environment, which is characterized by

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

coarticulatory phenomena, vowel reduction, and a faster rate. Although gating is not generally considered an on-line task, it has implications for spoken language processing. Gating experiments have produced results that converge with on-line paradigms (e.g., Tyler & Wessels, 1985), including the demonstration of typical effects on word identification such as word frequency and length (see Grosjean, 1996, for a review).

The present study extended the investigation of the role of AV input to L2 spoken word identification with the following four research questions. Does perceptual training focused on specific sounds in minimal pairs facilitate identification of words beginning with these sounds? Does AV input result in earlier identification compared to A-only? Do words beginning with some sounds show earlier identification than others following training? Do results for Japanese and Korean speakers suggest an influence of L1 phonology?

I hypothesized that word identification by learners (a) would be earlier following the perceptual training described above, (b) would be earlier with AV versus A-only input, and (c) would vary according to the initial CV sequence of the words. Earlier word identification following training was expected based on the robust nature of this type of perceptual training using multiple natural exemplars with feedback that has shown generalization to novel stimuli and new talkers, and transfer to production improvement (e.g., Hardison, 2003; Lively et al., 1993). The benefit of AV stimuli (vs. A-only) was expected to be accentuated for sounds such as /ɹ/ and /l/ for both Japanese and Korean learners, given their attention to them during training. Variation based on the vowel in the initial CV sequence was anticipated because of the influence of the phonetic environment on the visual discernibility of consonants (e.g., Kricos & Lesner, 1982) and on the formant structure of /ɹ/ and /l/ (e.g., Hagiwara, 1995).

Results for the Japanese and Korean learners are discussed in Experiments 1 and 2, respectively. They were not directly compared statistically because of the existing number of factors and levels in the study.

EXPERIMENT 1: JAPANESE SPEAKERS

Method

Participants. A total of 32 Japanese speakers participated. Eight were presented with AV-gated stimuli and received AV perceptual training, 8 were given A-only-gated stimuli and received A-only training, and 8 were assigned to each of two control groups (8 with AV stimuli and 8 with A-only) who received testing but no training. This assured that the modality of presentation for gating matched that for training. All participants ranged in age from 18 to 25 years and were high intermediate level learners in an Intensive English Program at a university in the United States. Program level placement was based on a battery of tests. All had been in the United States for 1–7 weeks when the study began. All had corrected to normal vision and reported no hearing problems. None were receiving focused pronunciation instruction or tutoring. Previous English study had begun in middle school and had emphasized grammar and reading rather than aural/oral skills. Each participant was paid US$30 following the posttest when a brief interview was conducted.

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

Materials

Stimulus selection: Gating. Gating involved words beginning with /ɹ/, /l/, /p/, /f/, and /s, t, k/ combined with high, low, and rounded vowels. Targets were bisyllabic content words with initial-syllable stress that had been rated as being highly familiar by a group of learners of English as a second language (ESL; primarily Japanese and Korean speakers) in the same language program one proficiency level lower. Selected stimuli were those to which the raters had assigned a 6 (*I know the word and am fairly sure of the meaning*) or a 7 (*I know the word and its meaning*) on a 7-point scale. These words were also from relatively sparse neighborhoods defined as no more than 10 neighbors.[1] Each consonant appeared before three types of vowels: high, low, and rounded. This created 15 categories (5 Consonant Groups × 3 Types of Vowels) each with 3 words, for a total of 45 words in each of the two tests (pretest and posttest). (See Appendix A for stimuli and duration in gates.)

The target words were embedded in sentences.[2] The word preceding the target ended in a relatively neutral lip position (e.g., a schwa or an alveolar stop, e.g., /t/) to control its acoustic and visual influence. The same approach was taken for the preposition beginning the phrase that followed the target word in recording (e.g., *I forgot to pick up the package at the post office; The children wanted a rabbit as a pet; She decided to order a salad at the deli*.). The prepositional phrase was used to avoid sentence-final effects (e.g., falling intonation) on the production of the target word.

Stimulus recording and editing: Gating. All stimuli were recorded by a female native speaker of general AE. Recordings were made in a TV 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 was seated comfortably in front of the camera (Sony HI-8 videocamera) providing a full-sized image of the head with a fully visible lower jaw drop. She was shown a cue card for each sentence and was instructed to look directly at the camera and then produce the sentence at a conversational rate of speech. She wore an Electrovoice lavaliere microphone.

The recordings were dubbed onto Betacam SP tape and then downloaded for editing by AVID Media composer (MC 8000) version 5.51 for MacIntosh and were digitized at a sampling rate of 44.1 kHz. Each utterance was saved as a file. Determining the point at which to excise a target word produced in a sentence was guided by the acoustic signal to avoid including an acoustic cue from the previous word. This would have misdirected the identification process for the A-only groups by suggesting an incorrect word-initial sound. The audio and video signals were simultaneously gated. The video signal imposes some editing restrictions; for example, the minimum unit is one frame. After reviewing sample stimuli with various gate durations, a two-frame gate (approximately 66 ms) was selected, taking into consideration the amount of information per frame, the number of target words to be included in the study, and potential fatigue for the participants. Following editing, stimuli were transferred in a randomized order to broadcast

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

quality videotapes for both the AV and A-only groups. For the A-only groups, the screen remained black.

Summary of stimuli: Perceptual training. For the AV and A-only perceptual training (see Hardison, 2003, for details), stimuli were minimal pairs produced by five native speakers of AE (three female, two male) recorded on videotape. There was a total of 80 trials for /ɹ/ and /l/ (6 Words × 4 Positions × 3 Adjacent Vowels [/u, o/, /aɪ/, /i, ɪ/] plus eight words in intervocalic position), 57 trials for /p/-/f/ (52 contrasts in initial singleton and cluster positions, 5 in final singleton with a variety of vowels), and 25 trials for /θ/-/s/ (initial and final positions, with a variety of vowels). Each trial was preceded by a warning tone, followed by the stimulus, and then 4 s of black for response time. The talker's face appeared on the screen 1 s before and after each utterance. For the A-only group, the talker's face did not appear. For feedback, participants saw the correct word on the screen in 96 point Times Black type on a 90% white background. The stimulus was then repeated.

Procedure

Gating. To obtain some measure of pretraining word identification performance between the AV and A-only groups in both L1s, participants were presented with 12 A-only gating trials before the experimental trials. All stimuli were produced by the same talker to control for talker effect (e.g., Hardison, 2003; Lively et al., 1993). These stimuli involved three target words (bisyllabic content words with high familiarity ratings) for each of the three most problematic sounds involved in the perceptual training (i.e., /f/, /ɹ/, /l/) plus three filler words beginning with nonlabial sounds /t/ or /k/ in randomized order. They were presented only auditorily (i.e., the modality shared by both groups) to avoid repetition effects.

Participants were tested in small groups, seated comfortably in front of an elevated 27-in. TV monitor. To familiarize them with the gating task, two practice trials were presented using a talker unrelated to the experiment. Participants were instructed to write down the word they thought the talker was saying at each gate and not to change any previous responses. They were given 4 s to respond after each gate. This 4-s interval had been determined from pilot testing with a comparable population. Observation of experimental sessions ensured that participants understood and followed the instructions, and those in the AV groups looked at the screen. Each gating task took about 30 min.

After the first gate (i.e., the first two frames or 66 ms) of a target word, subsequent presentations provided the first gate plus an additional gate, and this process continued until the acoustic offset of the target word had been presented. Two warning tones signaled a new stimulus, and one tone indicated the next gate of the same stimulus.

During postexperiment interviews, participants were shown the list of stimuli used in the study and asked to rate them on a 7-point scale to confirm their familiarity.

Perceptual training. Perceptual training was conducted in the same room and with the same equipment as that described for the gating tasks. During AV training,

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

participants were instructed to look at the screen when they heard the warning tone that preceded each trial, and then to circle the word on their answer sheets that the talker had said. I observed all training sessions to ensure that participants made their responses prior to the appearance of feedback on the screen. The stimuli recorded by each of the five talkers were presented a total of three times over a period of 15 training sessions (3 weeks). Each session involved only one talker and lasted approximately 45 min.

RESULTS AND DISCUSSION

Following previous gating studies, the identification point for each target word was determined by the gate at which the correct word was written with no subsequent changes. This was converted to a percentage of the word's total duration in gates to compensate for differences in the absolute duration (in milliseconds) of the stimuli (e.g., Salasoo & Pisoni, 1985). Although all words were bisyllabic, they ranged from five to eight gates in duration (see Appendix A). For analysis purposes in gating studies, tabulations for words not correctly identified by their acoustic offset were based on the total duration of the word plus one gate (producing percentages over 100; e.g., Grosjean, 1980). These data were submitted to analyses of variance. Planned comparisons based on an a priori hypothesis (cf. hypothesis 3) and L1 phonology (see Table 1) focused on /ɹ/, /l/, and /f/. The large number of degrees of freedom in the CV factor made finding a significant effect by post hoc tests highly improbable.

The 12 preliminary A-only gating trials revealed that the mean percentage of the target words that the participants needed to hear to identify them correctly exceeded 100% for the AV and A-only experimental groups. This represented a failure to identify the words prior to their acoustic offset. These groups were also comparable in terms of pretraining perceptual identification accuracy of /p, f, ɹ, l, θ, s/ in minimal pairs; their A-only mean scores differed by less than 3% (Hardison, 2003).

For gating, within-subjects variables were Time (pretest, posttest) and Initial CV Sequence (15 levels: 5 Initial-Consonant Visual Categories × 3 Vowel Categories). The between-subjects variable was Modality (AV, A-only). Results for the experimental groups revealed significant main effects of modality, $F\,(1, 30.15) = 25.97$, $p < .0001$, time, $F\,(1, 51.36) = 259.1$, $p < .0001$, and CV sequence, $F\,(14, 30.55) = 2.38$, $p < .05$. Mean identification points were earlier (i.e., less information was required) in the AV presentation (88.67% of the word needed) compared to A-only (96.46%). Mean identification was also earlier in the posttest (81.09% of the word needed following training) compared to the pretest (104.05% prior to training). Omega-squared calculations revealed these effects accounted for 56% of the variance. Participants confirmed their familiarity with the stimuli by rating them as 6 or 7 on the 7-point scale during the postexperiment interview. The control groups showed no significant effect of time, for example, for A-only, $F\,(1, 30.65) = 1.36$, *ns*), indicating that posttest target words were not identified earlier than pretest ones; therefore, no further analyses were conducted on these data.

Figure 1 shows the mean percentage of the target words needed for correct identification. Because of the number of levels of the CV factor, the data here

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

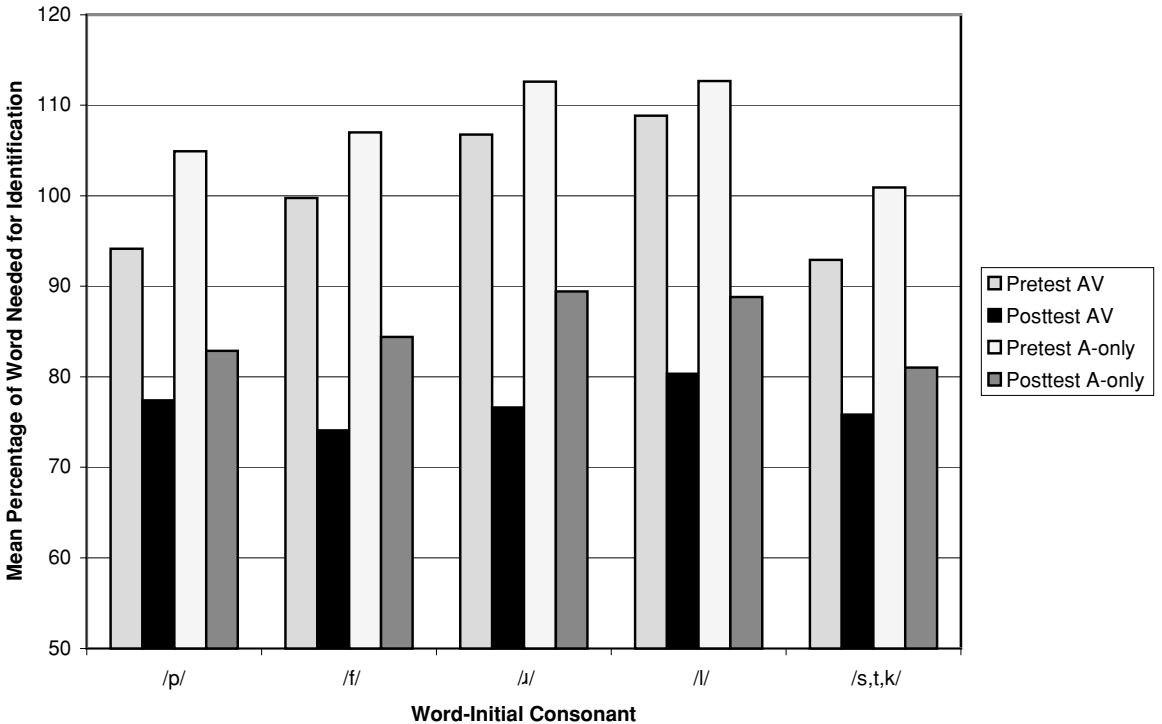


Figure 1. Japanese groups' mean percentage of words needed for identification by time (pretest, posttest), modality (AV, A-only), and CV sequence (averaged across vowels).

are combined for the high, low, and rounded vowels within each word-initial consonant category. For each consonant group in the figure, there are four bars: the pretest AV, posttest AV, pretest A-only, and posttest A-only. The lower the percentage, the less information participants needed to identify the target word accurately.

In general, the mean pretest A-only scores indicated failure to identify the words before their acoustic offsets (i.e., more than 100% of the words was needed for correct identification). This was also the case for AV words beginning with /ɹ/ and /l/ in the pretest. Those beginning with /f/ showed a mean of 99.8%. Significantly earlier identification in both AV and A-only presentations as a result of perceptual training was evident across consonant categories. The best performance was found for AV stimuli in the posttest, indicating earlier identification following training and in the presence of visual cues. Less information was required to identify words with the following initial CV sequences: /p/ + high vowel (e.g., *picnic*), and /ɹ/ + a low vowel (e.g., *rocket*, or the low vowel as the first element in the diphthong in *writer*).

There was a significant Time × CV Sequence interaction, $F\ (14, 34.59) = 2.93$, $p < .001$. Planned comparisons revealed that identification of words beginning with /ɹ/ and /l/ was earlier as a result of perceptual training compared to other word-initial sequences, $F\ (1, 34.59) = 21.17$, $p < .0001$. The posttest findings for words such as *rocket* may be attributed to the focus of learner attention during training on /ɹ/ and /l/, and to the more visually salient gesture for /ɹ/ in the environment of a low versus a rounded vowel.

There was also a significant Modality × CV × Time interaction, $F\ (14, 30) = 2.41$, $p < .05$. For some word-initial CV sequences, the amount of information needed for identification between AV and A-only presentations differed significantly as a result of training. Planned comparisons revealed that identification of /ɹ/- and /l/-initial words showed greater accentuation of the advantage of AV over A-only input as a result of perceptual training compared to other word-initial sequences, $F\ (1, 14) = 14.58$, $p < .01$. Comparisons also revealed a significant effect of adjacent vowel on the identification of AV and A-only /ɹ/- and /l/-initial words, $F\ (2, 143) = 18.61$, $p < .001$. As shown in Figure 2, identification was later for words with a rounded vowel. This is compatible with the visible anticipatory influence of rounding that can obscure consonant articulation (e.g., Benguerel & Pichora–Fuller, 1982), and lower identification accuracy scores in the perception of /ɹ/ and /l/ when followed by rounded vowels (Hardison, 2003).

EXPERIMENT 2: KOREAN SPEAKERS

Method

A total of 32 Korean speakers (from Seoul) participated. Eight were assigned to each of two experimental groups for the gating task according to the modality of their perceptual training (AV and A-only). An additional eight were assigned to each of two control groups (AV and A-only). Other details of the method are the same as for Experiment 1. These participants received the same training as the Japanese speakers.

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

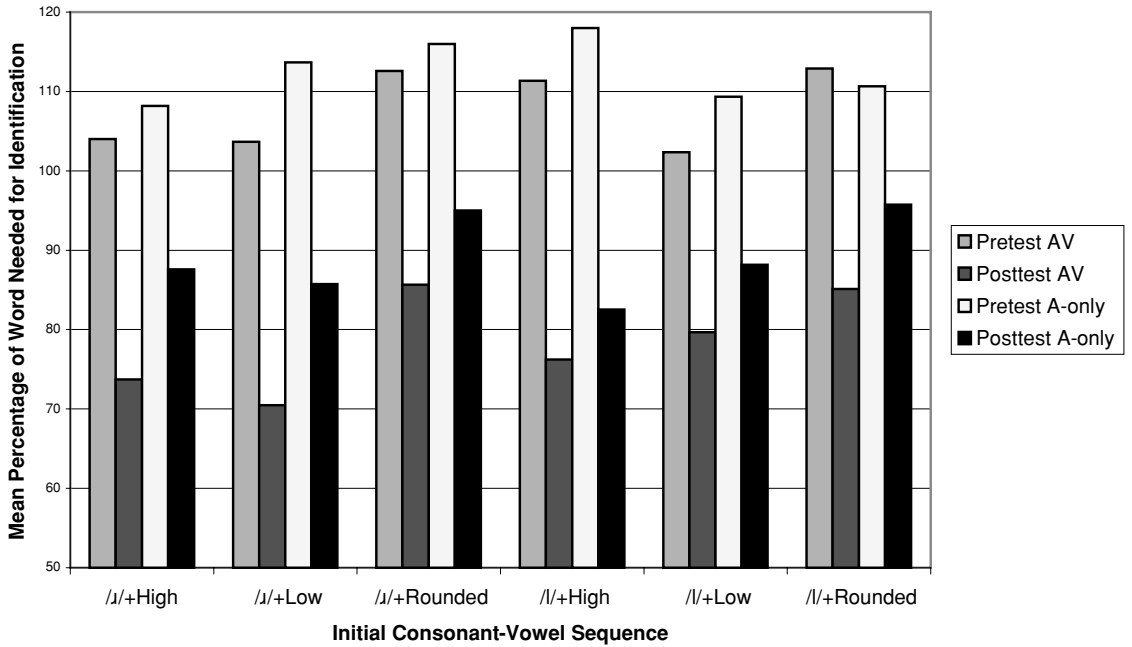


Figure 2. Japanese groups' mean identification of words beginning with /ɹ/ and /l/ as a function of time (pretest, posttest), modality (AV, A-only), and CV sequence.

RESULTS AND DISCUSSION

The gating results from Experiment 2 were analyzed with the same approach as those in Experiment 1. The 12 preliminary A-only gating trials revealed that the mean percentage of the target words the participants needed to identify them correctly exceeded 100% across groups, indicating a failure to identify the words prior to their acoustic offset. Results of the experimental data revealed significant main effects of modality, $F\ (1, 28.93) = 33.13,\ p < .0001$, and time, $F\ (1, 48.86) = 259.8,\ p < .0001$, but not of CV sequence, $F\ (14, 30.65) = 1.36$, *ns*. Mean identification points were earlier in AV presentation (85.61% of the word needed) compared to A-only (94.57%), and in the posttest (79.84% after training) versus the pretest (100.34% before training). Omega-squared calculations revealed these effects accounted for 49% of the variance. Participants confirmed their familiarity with the stimuli by rating them as 6 or 7 on the 7-point scale during the postexperiment interview. The control groups showed no significant effect of time, for example, for A-only, $F\ (1, 34.39) = 2.64$, *ns*, indicating that posttest target words were not identified earlier than pretest ones; therefore, no further analyses were conducted on these data.

Figure 3 shows the mean percentage of the target words needed for correct identification combined here across the high, low, and rounded vowels for each CV sequence. For all categories except the nonlabials (/s, t, k/), the mean A-only pretest results showed that participants failed to identify the target words before their acoustic offset. Identification of AV words prior to perceptual training required over 90% of the word except for the nonlabial-initial words, which required about 87%. Significant improvement in both AV and A-only presentations occurred across consonant categories. The lowest percentages (i.e., earliest identification points) were found for posttest AV presentations. As was the case for the Japanese, the Korean speakers also demonstrated earlier identification of words following perceptual training, and with visual cues. This was especially evident for words with the following initial CV sequences: /p/ + high vowel (e.g., *picnic*), /f/ + high vowel (e.g., *finger*), /ɹ/ + high vowel (e.g., *reason*), /ɹ/ + low vowel (e.g., *writer*), and /l/ + low vowel (e.g., *laughter*). Note that none of these words includes a rounded vowel.

There was a significant Modality × CV interaction, $F\ (14, 392) = 4.824,\ p < .0001$. The advantage of AV over A-only input varied according to CV sequence for the Korean as well as the Japanese speakers. Planned comparisons revealed a significant effect of adjacent vowel for words beginning with /ɹ/ and /l/ in both AV and A-only presentations, $F\ (2, 143) = 6.25,\ p < .01$. As shown in Figure 4, identification was later for /ɹ/ + rounded vowels and for /l/ + high or rounded vowels. In addition, for /ɹ/- and /l/-initial words, especially /l/ + low vowel (e.g., *language*), AV input was particularly effective, $F\ (1, 14) = 5.88,\ p < .05$. This finding may be attributed to a combination of learner attention to these sounds during training, the use of visual cues, and the visual salience of the articulatory gesture in the environment of the low vowel.

There was also a significant Time × CV interaction, $F\ (14, 34.39) = 2.09,\ p = .01$, indicating that some word-initial CV sequences showed greater improvement as a result of perceptual training than others. Planned comparisons revealed that

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

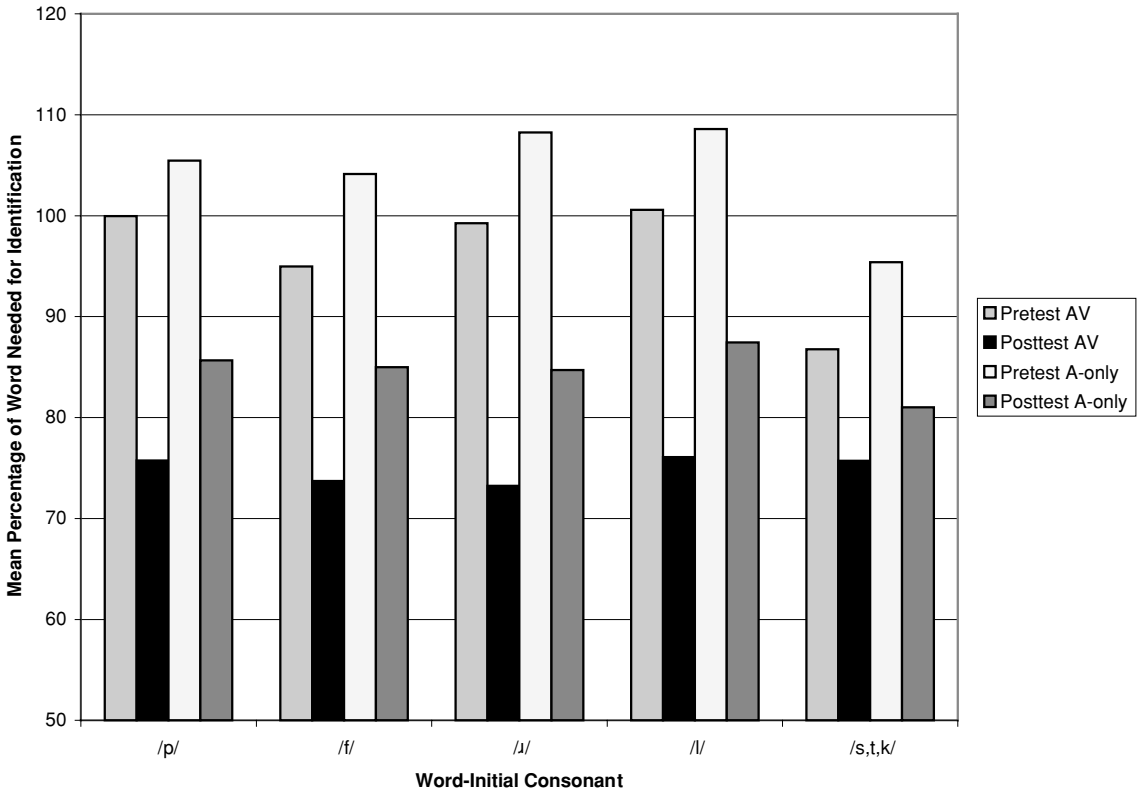


Figure 3. Korean groups' mean percentage of words needed for identification by time (pretest, posttest), modality (AV, A-only), and CV sequence (averaged across vowels).

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

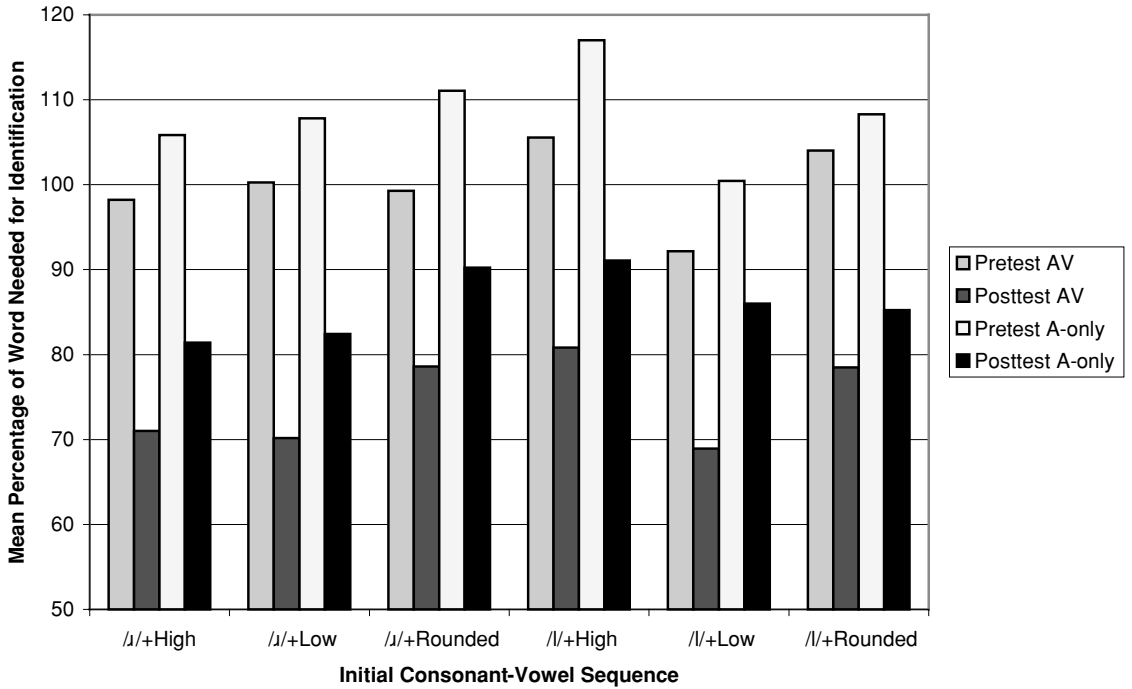


Figure 4. Korean groups' mean identification of words beginning with /ɹ/ and /l/ as a function of time (pretest, posttest), modality (AV, A-only), and CV sequence.

identification of /ɹ/- and /l/-initial words showed more improvement with training than the other word-initial sequences, $F\ (1, 34.39) = 16.08$, $p = .001$. This was particularly evident for /ɹ/-initial words followed by a low vowel. However, in contrast to the findings for the Japanese, the Modality × Time × CV interaction was not significant, $F\ (14, 30) = .851$, *ns*. Although the advantage of AV over A-only input varied across CV sequences as noted above, it was not significantly affected by the training for the Korean speakers. This may be due to the fact that the gating task focuses on the initial versus the final part of the word, and the effect of training is more noticeable on the area of greatest perceptual difficulty. This area for Korean speakers is word-final liquids, especially /l/ following high vowels (Hardison, 2003).

GENERAL DISCUSSION

Results of the present study involving Japanese and Korean ESL learners demonstrated significantly earlier identification of words beginning with /ɹ/, /l/, /p/, and /f/ following perceptual training focused on these sounds in minimal pairs. In addition, AV versus A-only input facilitated identification both before and after training with some variation according to word-initial CV sequence. This variation was sensitive to influence from L1 phonology. For the Japanese speakers, identification of words beginning with /ɹ/ and /l/ showed greater accentuation of the advantage of AV versus A-only input following training compared to other word-initial consonants. For the Korean speakers, words beginning with /ɹ/ and /l/ were identified earlier than others after training, and earlier with AV input; however, in contrast to the Japanese speakers, the AV benefit for the Koreans did not change significantly with training. This difference between the L1 groups may be attributable to the importance of the initial CV sequence in the word identification process and differences in L1 phonology. Japanese has a dental or alveolar flap in initial position; Korean has no word-initial liquid. During perceptual training, the Japanese had lower accuracy scores for word-initial liquids compared to the Koreans for whom liquids in word-final position were more difficult to identify (Hardison, 2003).

The first two sounds in a word are crucial to the word identification process, as these activate all stored representations of words having this initial sequence, suggesting potential word candidates (Marslen–Wilson & Welsh, 1978). As additional information is presented, the initial pool of candidates or cohort is reduced in size. Although semantic context facilitates processing when available and comprehensible, studies have shown that initial activation is based on acoustic–phonetic input (Tyler & Wessels, 1985). In a situation where visual cues are also present, results of the present study suggest that activation may begin even earlier. This priming role for visual speech cues is enhanced through training to facilitate L2 word identification. This is particularly helpful for learners such as the Japanese speakers in the current study whose perceptual difficulties with /ɹ/ and /l/ are greatest in word-initial position.

Following multiple-trace theory, acoustic and visual signals processed in primary memory constitute probes or retrieval cues that activate in parallel stored traces in long-term memory. The contribution of each trace is weighted according to its similarity to the features of the probe. Ultimately, the probe returns an echo to primary memory. In identifying a spoken word, a process that extends

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

over time, echo intensities are obtained over time until a decision is reached. For L2 learners, perceptual training results in stronger echoes from L2 traces that are less ambiguous in content, in better defined lexical neighborhoods, and at greater psychological distance from existing L1 traces. This process is further enhanced by the temporal precedence of identifiable visual speech cues, which can influence the rate at which initial word candidates are reduced ultimately to one choice as increasing amounts of a stimulus word are made available. Although native speakers identify some monosyllabic words after their acoustic offset when embedded in sentence context (Bard et al., 1988; Grosjean, 1985), the postexperiment interviews conducted in the current study suggest that this is unlikely to be as successful with intermediate-level L2 learners when they are confronted with a continuous stream of speech. One Korean learner expressed it best when he commented, "If I don't know a word, I just get stopped." However, word identification after acoustic offset remains a possibility for nonnative speakers, perhaps at an advanced level of proficiency, and is a question for further research. It is evident that focused training facilitates L2 word identification by increasing the information value of the acoustic–phonetic and visual cues to speech sounds. As a result, learners are better able to take advantage of the priming role of visible speech in spoken language processing.

APPENDIX A

The stimuli used in the gating tasks

	High Vowel	Low Vowel	Rounded Vowel
/p/	Picture (5), pillow (6), peanut (5), picnic (7), people (5), pizza (6)	Package (7), panic (6), pilot (6), pocket (6), passport (8), party (6)	Poster (7), poison (7), pony (5), poem (5), portion (7), postcard (8)
/f/	Finger (8), figure (6), female (7), feature (6), fiction (7), feedback (7)	Farmer (6), fashion (8), finance (8), factor (6), father (7), family (7) (produced with two syllables /fæmli/)	Focus (7), folder (6), format (7), photo (5), forest (8), future (8)
/ɹ/	Reason (6), river (7), region (6), research (7), refund (7), reader (7)	Rocket (7), rabbit (7), riot (7), racket (7), writer (5), roster (7)	Ruler (5), rumor (6), robot (8), royal (5), roommate (8), romance (8)
/l/	Leader (5), limit (5), liquid (6), leisure (7), liter (5), listener (7)	Laughter (6), language (7), lion (7), lighter (6), laptop (6), license (8)	Laundry (6), lawyer (6), lotion (7), loyal (5), loser (6)
/s, t, k/	Singer (7), silver (7), ticket (6), teacher (7), kingdom (8), kitchen (7)	Salad (8), sandwich (8), target (6), tower (6), carrot (6), contrast (7)	Sofa (8), soldier (7), toaster (8), tourist (7), coffee (8), cola (6)

The duration (ms) in number of gates is given in parentheses after each word.

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

ACKNOWLEDGMENTS

I am deeply grateful to Dave Derkacy and Dave Rust for technical production assistance, Terese Thonus for her participation, and statistician Jonathan Margolin for assistance in data analysis. I appreciate comments on an earlier version from Andy McCullough and three anonymous reviewers. Preliminary findings from this study were presented at the meeting of the American Association for Applied Linguistics in Vancouver in March 2000.

NOTES

1. The data for neighborhood density were obtained from the Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN.
2. These stimuli were used in another study to investigate the contribution of context to L2 word identification; therefore, semantically plausible contexts were used for each target word instead of the same carrier sentence.

REFERENCES

Bard, E., Shillcock, R., & Altmann, G. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics, 44*, 395–408.

Benguerel, A.-P., & Pichora–Fuller, M. K. (1982). Coarticulation effects in lipreading. *Journal of Speech and Hearing Research, 25*, 600–607.

Berger, K. W. (1972). *Speechreading: Principles and methods*. Baltimore, MD: National Education Press.

Cho, Y.-M. Y. (1987). Phrasal phonology of Korean. In S. Kuno, I.-H. Lee, J. Whitman, S.-Y. Bak, & Y.-S. Kang (Eds.), *Harvard studies in Korean linguistics II* (pp. 328–340). Cambridge MA: Harvard University, Department of Linguistics.

Cotton, S., & Grosjean, F. (1984). The gating paradigm: A comparison of successive and individual presentation formats. *Perception & Psychophysics, 35*, 41–48.

Daniloff, R. G., & Moll, K. (1968). Coarticulation of lip rounding. *Journal of Speech and Hearing Research, 11*, 707–721.

Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 33–66). San Diego, CA: Academic Press.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28*, 501–518.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28*, 267–283.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics, 38*, 299–310.

Grosjean, F. (1996). Gating. *Language and Cognitive Processes, 11*, 597–604.

Hagiwara, R. E. (1995). Acoustic realizations of American /r/ as produced by women and men (Doctoral dissertation, University of California, Los Angeles, 1995). *UCLA Working Papers in Phonetics No. 90*.

Hardison, D. M. (1999). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language Learning, 49*, 213–283.

Hardison, D. M. (2000). The neurocognitive foundation of second-language speech: A proposed scenario of bimodal development. In B. Swierzbin, F. Morris, M. E. Anderson, C. A. Klee, & E. Tarone (Eds.), *Social and cognitive factors in second language acquisition* (pp. 312–325). Somerville, MA: Cascadilla Press.

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context and talker variability. *Applied Psycholinguistics, 24*, 495–522.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411–428.
Jacoby, L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 18, pp. 1–47). New York: Academic Press.
Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics, 21*, 3–28.
Kim–Renaud, Y.-K. (1974). Korean consonantal phonology (Doctoral dissertation, The University of Hawaii, 1974). *Dissertation Abstracts International, 36*, 869.
Kricos, P. B., & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *Volta Review, 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. *Journal of the Acoustical Society of America, 94*, 1242–1255.
Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon. Research on speech perception* (Tech. Rep. No. 6). Bloomington, IN: Indiana University.
Marslen–Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology, 10*, 29–63.
Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.
Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.
Munhall, K. G., & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America, 104*, 530–539.
Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research, 28*, 381–393.
Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication, 13*, 109–125.
Price, P. J. (1981). A cross-linguistic study of flaps in Japanese and in American English (Doctoral dissertation, University of Pennsylvania, 1981). *Dissertation Abstracts International, 42*, 1128A.
Rosenblum, L. D. (2002). The perceptual basis for audiovisual speech integration. In J. H. L. Hansen & B. Pellom (Eds.), *International Conference on Spoken Language Processing 2002* (pp. 1461–1464). Sydney, Australia: Causal Productions Pty, Ltd.
Salasoo, A., & Pisoni, D. B. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language, 24*, 210–231.
Sekiyama, K., & Sugita, Y. (2002). Auditory–visual speech perception examined by brain imaging and reaction time. In J. H. L. Hansen & B. Pellom (Eds.), *ICSLP 2002* (pp. 1693–1696). Sydney, Australia: Causal Productions PTY, Ltd.
Shin, S.-H. (1997). Correspondence in Kyungsang Korean truncation. In S. Davis (Ed.), *Optimal Viewpoints* (pp. 115–138). Bloomington, IN: Indiana University Linguistics Club Publications.
Silva, D. J. (1989). Korean obstruent voicing: Investigations into prosodic structuring. In S. Kuno, I.-H. Lee, J. Whitman, S.-Y. Bak, & Y.-S. Kang (Eds.), *Harvard studies in Korean linguistics III* (pp. 177–188). Cambridge, MA: Harvard University, Department of Linguistics.
Smeele, P. M. T. (1994). *Perceiving speech: Integrating auditory and visual speech.* Unpublished doctoral dissertation, Delft University of Technology, Delft.
Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica, 26*, 314–331.
Tsujimura, N. (1996). *An introduction to Japanese linguistics*. Oxford: Blackwell.
Tyler, L., & Wessels, J. (1985). Is gating an on-line task? Evidence from naming latency data. *Perception & Psychophysics, 38*, 217–222.
Vance, T. J. (1987). *An introduction to Japanese phonology*. Albany, NY: State University of New York Press.
Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics, 57*, 1124–1133.
Woodward, M. F., & Barber, C. G. (1960). Phoneme perception in lip reading. *Journal of Speech and Hearing Research, 3*, 212–222.

https://doi.org/10.1017/S0142716405050319 Published online by Cambridge University Press