Applied Psycholinguistics **24** (2003), 495–522
Printed in the United States of America
DOI: 10.1017.S0142716403000250

Acquisition of second-language speech: Effects of visual cues, context, and talker variability

DEBRA M. HARDISON
Michigan State University

ADDRESS FOR CORRESPONDENCE
Debra M. Hardison, Department of Linguistics & Languages, A-714 Wells Hall, Michigan State University, East Lansing, MI 48824-1027. E-mail: hardiso2@msu.edu

ABSTRACT
The influence of a talker's face (e.g., articulatory gestures) and voice, vocalic context, and word position were investigated in the training of Japanese and Korean English as a second language learners to identify American English /ɹ/ and /l/. In the pretest–posttest design, an identification paradigm assessed the effects of 3 weeks of training using multiple natural exemplars on videotape. Word position, adjacent vowel, and training type (auditory–visual [AV] vs. auditory only; multiple vs. single talker for Koreans) were independent variables. Findings revealed significant effects of training type (greater improvement with AV), talker, word position, and vowel. Identification accuracy generalized successfully to novel stimuli and a new talker. Transfer to significant production improvement was also noted. These findings are compatible with episodic models for the encoding of speech in memory.

One of the areas of theoretical and pedagogical significance in second-language (L2) speech research involves the factors affecting modification of the adult perceptual system to promote development of robust categories for the identification of nonnative phonetic contrasts. The acquisition of American English (AE) /ɹ/ and /l/ has been the focus of numerous studies (for Japanese, e.g., Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura, 1975; Mochizuki, 1981; Yamada & Tohkura, 1992; for Japanese and Korean, Ingram & Park, 1998), including auditory training studies (for Japanese, e.g., Lively, Logan, & Pisoni, 1993; for Korean, Borden, Gerber, & Milsark, 1983; Yu & Jamieson, 1993). Results with Japanese speakers have shown better performance with natural versus synthetic speech stimuli (Mochizuki, 1981), greater experience with the L2 (e.g., MacKain, Best, & Strange, 1981), early exposure to the L2 in childhood (e.g., Cochrane, 1980; Yamada, 1995), and multiple- versus single-talker training (e.g., Lively et al., 1993). Findings have also demonstrated the influence of perceptual learning on production and retention of improved perception and production performance (Bradlow, Akahane–Yamada, Pisoni, &

© 2003 Cambridge University Press 0142-7164/03 $12.00

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

Tohkura, 1999). Most studies have also noted the effects on the perception of /ɹ/ and /l/ of word position (initial vs. final, singleton vs. cluster), first-language (L1) phonology (e.g., Ingram & Park, 1998; Lively et al., 1993), and talker (e.g., Lively et al., 1993).

However, the aforementioned studies have dealt with a single source of input: the auditory modality. This paper reports the findings of a perceptual training program focusing on /ɹ/ and /l/ conducted with Japanese and Korean intermediate-level learners of English as a second language (ESL) in the United States to investigate (a) the influence of two phonological systems on the development of perceptual accuracy, (b) the effects of adjacent vowel and word position for these sounds, (c) the contribution of visual information from a talker's face by comparing auditory–visual (AV) and auditory-only (A-only) training for both groups, (d) the direct comparison of multiple- versus single-talker training for the Koreans, and (e) the relationship between perceptual training and production improvement in terms of the influence of both word position and vowel.

The AE /ɹ/ is generally characterized by a low third formant (F3) frequency that distinguishes it from /w/ and /l/ (Lindau, 1985). Japanese listeners perform as well as Americans in discriminating differences in the F3 transition when stimuli are presented in isolation (nonspeech); however, they differ when the transitions serve as a phonetic cue in combination with first formant (F1) and second formant (F2) patterns appropriate for the perception of /ɹ/ and /l/ in speech (Liberman, Miyawaki, Jenkins, & Fujimura, 1973). Recent experiments suggest they focus on F2 frequency values to try to distinguish these sounds (Yamada, 1995). Japanese has a voiced apical flap in the dental or alveolar region occurring in utterance-initial and intervocalic positions, with acoustic, articulatory, and perceptual similarities to the AE flap (Price, 1981). It shows context-dependent variability in place, amount, and duration of contact. The flap does not appear adjacent to other consonants and may occur in the environment of all Japanese vowels (Tsujimura, 1996).

In contrast, Korean has a voiced apicoalveolar nonvelarized or clear /l/ in syllable-final position and a voiced alveodental flap intervocalically (or between vowel and glide; Kim–Renaud, 1974). Although there is no liquid in word-initial position in native Korean and Sino-Korean words, one does occur in "loanwords" (Kim–Renaud, 1974; Shin, 1997), with a tendency to be pronounced as a flap; but there is interspeaker variability.

The present study also drew on the insights of Gillette (1980) and Shimizu and Dantsuji (1983) to explore the influence of the adjacent vowel and word position on the perception and production of /ɹ/ and /l/. The vowels /i/ and /ɯ/ generally constitute more difficult contexts for perceptual accuracy than lower unrounded vowels (Hagiwara, 1995).

In addition, there is evidence that visual cues for L2 learners are a source of influence in understanding speech. Results of a survey of Japanese students indicated they used facial cues more as a result of having lived in the United States for at least 2 years due to both American customs and the necessity for nonnative speakers to "catch as much information as possible from their interlocutor" (Hattori, 1987, p. 115) to compensate for comprehension difficulty. Although in an early experimental study, Goto (1971) suggested that in the

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

auditory presentation of sounds, "there was the disadvantage of not being able to read the lips of the speaker" (p. 321), experimental evidence on the contribution of visual cues was not reported until much later. In a study on the McGurk effect[1] (Hardison, 1999), visual /ɹ/ paired with auditory /ɹ/ in a consonant–vowel (CV) syllable with /ɑ/ significantly increased identification accuracy (87%) for Japanese ESL learner in the United States, compared to A-only presentation (65%). Accuracy for the Koreans rose from 80% to 95%. In contrast, Japanese speakers in Japan with an A-only intelligibility of 67% for AE /ɹ/ showed no improvement when the visual cue was added (Sekiyama & Tohkura, 1993); however, Sekiyama (1997) later found a positive correlation between the magnitude of the McGurk effect and the length of stay in the L2 environment for Chinese speakers living in Japan.

Although the speechreading (lipreading) literature documents that less information is communicated visually versus auditorily (e.g., Erber, 1974), listeners make use of visual information from a talker's face on a daily basis, especially when attempting to compensate for noise (e.g., Sumby & Pollack, 1954; Summerfield, 1979) or impaired hearing ability (e.g., Walden, Prosek, Montgomery, Scherr, & Jones, 1977). Generally /ɹ/ and /l/ constitute separate visual categories (e.g., Berger, 1972; Binnie, Jackson, & Montgomery, 1976). However, the visual intelligibility of a consonant is reduced in contexts with rounded vowels, especially /ʊ/ (e.g., Benguerel & Pichora–Fuller, 1982), and in clusters with bilabial stops and /f/ versus clusters with velar stops and singleton positions (Franks & Kimble, 1972), and it is subject to talker variability (Kricos & Lesner, 1982).

Several reports indicate that the information value of articulatory gestures can be improved with training. Visual recognition of /ɹ/ presented in the syllable /ɹɑ/ increased from 36.1 to 88.6% for hearing-impaired adults after 14 hr of training, although maximum improvement was reached after the first 5 hr (Walden et al., 1977). More recently, training with CV syllables facilitated unimodal and bimodal speech perception for native speakers of English (Massaro, Cohen, & Gesi, 1993).

Findings from other areas of research have suggested the existence of a mechanism or representation common to the processing of speech input from the auditory and visual modalities (Campbell, 1987; Watson, Qiu, Chamberlain, & Li, 1996). Cross-modal interaction at some level has demonstrated that information from different sensory modalities can be combined in perception, as in the McGurk effect (e.g., McGurk & MacDonald, 1976), and that input to one modality can influence processing in another (see, e.g., Robert–Ribes, Schwartz, & Escudier, 1995, for a review). Using magnetoencephalographic recordings, visual input specifically from lip movements was found to influence auditory cortical activity (Sams, Aulanko, Hämäläinen, Hari, Lounasmaa, Lu, & Simola, 1991). de Sa and Ballard (1997) suggested that multimodal integration is useful not only for improved recognition in general but also for the development of recognition abilities in the individual modalities. They demonstrated computationally that two modalities simultaneously trained, providing mutual feedback, reached a greater performance level than when treated independently. Separate modalities independently processing different sets of input dimensions together

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

were able to "teach each other" (p. 342) through integration (i.e., access to each other's output) at a higher level and feedback within each processing pathway.

Despite the potential contribution of visual cues to L2 speech processing as a second channel of input, perceptual training studies in the past have focused on the auditory modality; yet, they have served to demonstrate the modification capability of the adult perceptual system. Studies of developmental changes in cross-language speech perception have shown that infants' perceptual capacities extend to phonetic distinctions not present in their immediate language-learning environment (e.g., Jusczyk, 1992). Although these broadly tuned capacities decline relatively quickly with L1 contact in favor of attunement to the ambient language resulting in the ability to discriminate only language-specific phonetic contrasts (e.g., Aslin & Pisoni, 1980), this decline does not appear to represent an unrecoverable loss (e.g., Pisoni, Lively, & Logan, 1994; Werker, 1994).

A series of studies involving auditory training of adult Japanese speakers to identify /ɹ/ and /l/ demonstrated that training with a large and variable stimulus set using natural speech and minimal pairs contrasting /ɹ/ and /l/ in a variety of word positions produced by multiple talkers with feedback resulted in significant improvement in identification accuracy, generalization to novel stimuli and a new voice (Lively et al., 1993), transfer to production improvement, and retention of improved abilities (Bradlow et al., 1999). From the above studies, the intervocalic and initial singleton and cluster positions were the most difficult for Japanese speakers; however, the final cluster position was the most difficult for five Korean speakers who were trained using a similar stimulus set (Yu & Jamieson, 1993). In addition, in forced-choice perception tasks such as those involving minimal pairs with a relatively reliable sound–grapheme correspondence, neither word familiarity (de Jonge, 1995) nor real word status (Lively et al., 1993) had a significant influence on perception or on production in a reading task (Flege, Takagi, & Mann, 1995).

A recent experiment suggested an alternative training technique for L2 learners using computer-enhanced speech (exaggerated stimuli) and adaptive training (McCandliss, Conway, Fiez, Protopapas, & McClelland, 1998). This is reminiscent of the perceptual fading technique (Jamieson & Moroson, 1986) used to train Canadian francophones to identify /θ/ and /ð/ by exaggerating the amount of frication. In that study, after 90 min of training, identification accuracy improved with generalization to natural speech and new voices but not to other positions in the word nor to the task of identifying /ð/ versus /d/ (Moroson & Jamieson, 1989). Although adaptive training can facilitate learners' noticing of the critical features distinguishing two sounds and subsequent awareness of the contrast (Ellis, 2002), the question remains as to the generalizability and retention of improvements in accuracy as a result of brief periods of training with exaggerated stimuli.[2]

Successful training conditions have overcome previous limitations on auditory perceptual training (Pisoni et al., 1994) by demonstrating that stimulus variability, multiple talkers, an identification paradigm, and feedback during training contribute to the development of perceptual categories that are robust across phonetic environments and talkers, thus enabling transfer to novel stimuli and better retention. A high variability identification paradigm using natural

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

speech has been more effective than a low variability discrimination paradigm using synthetic speech such as the one used by Strange and Dittman (1984). Pisoni (1973) had argued that because discrimination training requires listeners to attend to small within-category differences (low-level acoustic information in sensory memory), it does not promote robust category formation. Previous studies had also found that feedback versus no-feedback perceptual training was consistently superior, especially under conditions involving a high degree of within-category variability (Homa & Cultice, 1984) that characterizes the phonetic variants of AE /ɹ/ and /l/.

Throughout development, the perceptual system becomes adapted to the environment and to tasks through the mechanism of attention weighting; that is, attention is allocated to dimensions and features important for a task, in this case, those relevant for categorization of speech sounds. Nosofsky (1986) describes these attention shifts as the "stretching" and "shrinking" of perceptual distances in the psychophysical space that is useful for particular tasks. Representations are stretched along attended dimensions so that items from different categories are made less similar, and thus more discriminable, whereas those shrunk along unattended dimensions are made more similar. This alters distances between exemplars according to their perceived similarity on an attended dimension; therefore, similarity relations are context dependent.

This stretching and shrinking process is compatible with the development of the attention weighting scheme that underlies the modification of phonetic categories during L1 development (e.g., Jusczyk, 1993, 1997). After the first few months of life, as a result of linguistic experience, attention becomes focused on features critical to the identification of L1 sounds. A similar process may occur in the training of native Japanese speakers to identify /ɹ/ and /l/ (e.g., Lively et al., 1993). During perceptual learning, attention weights may be reorganized to accommodate new phonetic contrasts, perhaps causing a reorganization of the learners' phonological spaces.

Variable performance across phonetic contexts and 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 (Goldinger, 1997; Johnson & Mullennix, 1997; Nygaard, Sommers, & Pisoni, 1995; Pisoni, 1993). Language learners may rely on context-dependent exemplars as memory representations to which input can be matched for identification (Lively et al., 1993) rather than an abstract description of the category (prototype concept) to which they belong (e.g., Brown, 1999). Multiple-trace memory theory (e.g., Hintzman, 1986) incorporates both prototypes and storage of individual episodes at different stages of processing, which can be extended to account for the variable effects on bimodal perceptual development by L2 learners (Hardison, 2000).

To date, perceptual category development has been discussed in the literature relative to one modality of input (auditory), one aspect of phonetic environment (word position), and primarily one L1 phonological system (Japanese). Therefore, the goals of the present study were to assess the contribution of visual information from a talker's face in the perception of AE /ɹ/ and /l/ by comparing AV and A-only training for Japanese and Korean ESL learners, the influence of

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

the phonetic environment (word position and adjacent vowel), the issue of multiple- versus single-talker training for the Koreans, and the transfer of perceptual training to production improvement.

I hypothesized that visual information would facilitate perceptual accuracy by providing a second set of stimulus dimensions on which similarities and distinctions could be drawn to facilitate perceptual category development, but would be of less value when the liquids were presented in clusters with labial consonants (Franks & Kimble, 1972) or contexts with rounded vowels (e.g., Benguerel & Pichora–Fuller, 1982). In the word-initial position, anticipatory rounding before /o/ and /ʊ/ was expected to mask consonantal articulation and perhaps suggest an /ɹ/ versus an /l/. Although previous L2 studies only alluded to a role for the adjacent vowel in the perception of liquids (Gillette, 1980; Shimizu & Dantsuji, 1983), speechreading reports documented the effects on consonantal identification of vowels (e.g., Benguerel & Pichora–Fuller, 1982), as well as word position (Franks & Kimble, 1972).

Previous auditory training studies with native Japanese speakers in the United States have also suggested that multiple-talker training results in better performance on tests of generalization to novel stimuli and a new talker (Lively et al., 1993). However, the question arises as to the potential effect of individual talker characteristics, including the discernability of articulatory movements (Gesi, Massaro, & Cohen, 1992; Kricos & Lesner, 1982). Comparable performance in tests of generalization for both a familiar talker from training and an unfamiliar talker may be due, at least in part, to comparable articulatory gestures and not solely to the benefits of the type of training.

Auditory training studies have also demonstrated improvement in production as a result of perceptual training (e.g., Bradlow et al., 1999). I hypothesized that significant effects of the adjacent vowel would emerge in the production and perception data and that those environments that posed problems in perception based on L1 phonology would also produce lower ratings in production (i.e., initial and intervocalic positions for the Japanese, final positions for the Koreans, and consonant clusters for both).

EXPERIMENT 1: JAPANESE SPEAKERS—PERCEPTUAL TRAINING AND PRODUCTION

Experimental design

The perception phase of the experiment consisted of the following sequence: pretest, training, posttest, and two tests of generalization (one with a familiar talker from training and one with an unfamiliar talker). There were two experimental groups: those receiving AV training and those receiving A-only. A control group was given the pretest and posttest but no training. As the objective was not speechreading training, there was no visual-only (V-only) training group. Testing and training for both groups involved a two-alternative forced-choice task with minimal pairs contrasting /ɹ/ and /l/ in a balanced design; word position and adjacent vowel were independent variables. All stimuli were presented via videotape. For the AV training group, all tests involved three condi-

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

tions in the order AV, A-only, and V-only, to assess the contribution made by each modality. Training stimuli for this group were AV. For the A-only group, testing and training stimuli were A-only (no video of talker).

For both groups, the training sequence for each stimulus involved the following components on videotape: warning tone, utterance, 4 s of black to circle the correct answer on the response sheet, feedback when the correct word appeared on the screen, and repetition of the utterance to promote the development of an association between the sound and its phonetic realization (and, for the AV group, its visual realization) in context.

Following the posttest, there were two tests of generalization involving novel stimuli, one using a familiar talker from training and the other, a new talker. To assess improvement in production as a result of perceptual training, pretest and posttest recordings were made of each participant's productions (experimental and control groups) of a randomized sequence of 100 words taken from the perception testing stimulus set. These were edited and presented to native speaker (NS) listener–judges for evaluation.

Method

Participants. Experiment 1 involved 16 NSs of Japanese, ranging in age from 18 to 25 years, with 8 assigned randomly to each experimental group (AV and A-only). These were intermediate-level learners in an intensive English program (IEP). All had been in the United States for 1–7 weeks when the training began. All had normal or corrected 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 oral skills. Each participant was paid $30 at the conclusion of the experiment when a brief interview was conducted. A control group of 8 participants from the same IEP level and with the same background received the testing (AV, A-only, V-only) but no training.

Materials

PERCEPTION TESTING AND TRAINING. Stimuli were minimal pairs contrasting /ɹ/ and /l/ in a balanced design (see Hardison, 1998a, for a complete list). For the pretest–posttest, word positions were initial singleton (e.g., *road–load*), initial cluster (e.g., *crime–climb*), final singleton (e.g., *hear–heal*), and final cluster (e.g., *tires–tiles*). Four minimal pairs were selected for each position with each of the following adjacent vowels, /ʊ/, /o/, /ɑɪ/, /e/ or /ɛ/, and /i/ or /ɪ/, as determined by the dimensions of height and rounding that influence the visual articulation of the surrounding consonants (e.g., Benguerel & Pichora–Fuller, 1982). Intervocalic position was included (e.g., *arrive–alive*) but analyzed separately due to the two vowels. The total number of trials was 510 (8 words × 4 positions × 5 vowels × 3 conditions = 480, plus 10 words in intervocalic position × 3 conditions = 30).

For the training set (e.g., *crew–clue*), there was an insufficient number of

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

words to continue all vowels; thus, they were selected following the speechreading literature: /ʊ, o/, /aɪ/, and /i, ɪ/. Within-subjects variables were vowel, word position, talker, and week. There was a total of 80 trials in training (6 words × 4 positions × 3 vowels = 72, plus 8 words in intervocalic position). Stimuli for both training groups involved multiple talkers: three female (Talkers 1–3) and two male (Talkers 4 and 5). The number and gender followed the selections made by Lively et al. (1993) for purposes of some comparability. All were NSs of General AE. Talkers were selected to provide variability in articulatory movements.

Following the posttest, there were two tests of generalization involving novel stimuli (e.g., *red–led*), one with a familiar talker from training (TG1) and one with a new talker (TG2). As with the pretest and posttest, for the AV group, each word was presented once in each of three conditions (AV, A-only, V-only). There were two vowel groups: rounded /ʊ, o/ and unrounded /e, ɛ/. The latter pair permitted testing of generalization performance on an untrained context. For each test, there was a total of 168 trials (6 words × 4 positions × 2 vowels, plus 8 words in intervocalic position = 56, × 3 conditions).

To prepare the stimuli, the audio and video signals were recorded as the talkers produced the words in a randomized order, prompted by cue cards. No specific articulation instructions were given. Recordings were made in a 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 illuminating it from the right. Talkers were seated comfortably in front of the camera. A Sony Hi-8 videocamera (EVO9100) was used with an Electrovoice (EVC090) lavaliere microphone. A full-sized image of each talker's head was obtained with a fully visible lower jaw drop. The recordings were dubbed onto Betacam SP tape, downloaded for editing into Avid Media Composer (MC 8000) Version 5.51 for MacIntosh, and digitized at a sampling rate of 44.1 kHz. Each utterance was saved as a file to permit random ordering for each modality presentation. The talker's face appeared on tape 1 s before and after each utterance. To edit each file, a template was created containing a warning tone (on a separate audio track), followed by the stimulus, and then 4 s of black for response time. For feedback, the correct word appeared on the screen in 96 point Times Black type on a 90% white background. To provide the repetition component of the training sequence, the file was dubbed. To create a V-only or A-only presentation, the respective audio or video track was deleted. Edited stimuli were transferred to broadcast quality VHS format for presentation.

Pretesting of all auditory stimuli for intelligibility with a group of eight NSs revealed no errors involving /ɹ/ and /l/. Visual intelligibility of stimuli in the V-only conditions for the pretest/posttest tape and those for the tests of generalization (familiar and unfamiliar talker) provided by another group of eight NSs revealed a similar mean percent accuracy for all three talkers: 84% for the pretest/posttest, 85% for the familiar talker from training, and 87% for the unfamiliar talker. The majority of identification errors were made when /ɹ/ and /l/ occurred in final cluster position, followed by final singleton with rounded vowels and initial cluster with /ʊ/. Given that an articulatory movement does not have a unique context- and talker-invariant sound correspondence, no word was ex-

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

cluded from the stimulus set based on these results that provided target data for comparison purposes.

PRODUCTION TESTING. Materials were taken from the perception testing stimulus set and consisted of 100 words contrasting /ɹ/ and /l/ with nine phonetic environments (3 positions [initial singleton, initial cluster and final singleton] × 3 vowels [/ʉ, o/, /ɑɪ/, /i, ɪ/] of 10 words each = 90, plus 10 contrasting the sounds in intervocalic position).

Procedure

PERCEPTION TESTING AND TRAINING. Stimuli were presented via a 27-in. Sony Trinitron color TV monitor and a Sony VHS videocassette recorder (SLV-920HF). Participants were tested and trained in small groups in a sound-attenuated room. As noted, in the testing phases for the AV group, each word was presented once in each of three conditions (AV, A-only, V-only) to provide a measure of the contribution of input from each modality to the precept. This group received AV training stimuli. For the A-only group, both testing and training stimuli were presented auditorily only. Throughout the sessions, a warning tone preceded the next trial. In the AV and V-only conditions of testing and during AV training, participants were instructed to look at the screen when they heard the tone and then to circle the word on their answer sheets that the talker said. I observed all phases of the experiment to ensure comprehension and, during training, to ensure that responses were made prior to feedback. Each training session lasted about 30 min. During training, 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.

Production testing. Prior to the perception pre- and posttest, recordings were made of each participant's productions of a randomized sequence of words to assess improvement in production as a result of perceptual training. Participants were shown a card on which a test word was printed, then the card was turned over and they were asked to produce the word. No acoustic prompt was given to avoid imitation. Productions were stored as individual files on computer hard disk using Kay Elemetrics Computerized Speech Lab Model 4300B (10-kHz sampling rate). These were then played out from the computer in a randomized order through a studio JBL speaker to groups of five NS listener–judges each, recruited from outside the university community and paid $20. Each group rated the productions of a total of eight participants over a period of several days. Productions for each were blocked in presentation. Judges were provided with the target words on response sheets and instructed to rate each production on a 7-point scale (7 representing a good example of the target) and to disregard other sounds. The presentation of individual tokens for evaluation most closely matches listeners' evaluations of nonnative speech in the natural language environment.

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

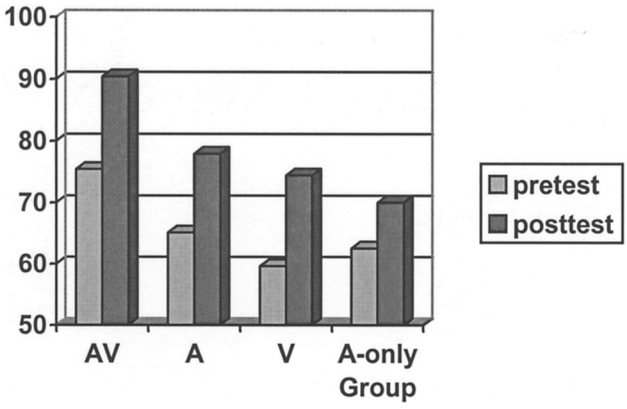


Figure 1. The Japanese groups mean percentage correct pre- and posttest by training type and modality. The scores for the AV training group for auditory–visual (AV), auditory-only (A-only) and visual-only (V-only). The A-only scores for the A-only training group.

Results and discussion

Identification scores (number correct) for each experimental group (AV, A-only) were tabulated. The results are discussed in the following order: the effectiveness of training per group (AV, A-only) (a comparison of pre- and posttest scores within each group), a comparison of AV versus A-only training (based on the shared modality, i.e., A-only pre- and posttest scores for each group), the effects of phonetic context (word position and vowel) and talker in training, tests of generalization (comparison of novel words produced by a familiar vs. unfamiliar talker), and the influence of perceptual training on production. Recall that results involving /ɹ/ and /l/ in intervocalic position were analyzed separately.

Effectiveness of training per group (*AV, A-only*). To assess the effects of training, analyses of variance (ANOVAs) were conducted on pretest and posttest scores for the AV and A-only training groups separately. In Figure 1, the first three pairs of bars represent the pretest and posttest mean scores of the three presentation conditions (AV, A-only, V-only) for the AV training group; the last pair of bars shows the scores for the A-only training group whose stimuli in testing and training were presented in only one condition (A-only).

For the AV training group, the variables were time (pretest–posttest), modality (AV, A, V), position (4), and vowel (5). All had significant main effects, F_t (1, 833) = 308.41, p = .0001; F_m (2, 833) = 137.23, p = .0001; F_p (3, 833) = 59.63, p = .0001; F_v (4, 833) = 14.82, p = .0001. The higher pretest AV (75.38%) versus A-only (65.13%) score for the AV group indicates the use of visual cues even before training. The interaction of Time × Modality, F (2, 833) = 0.77, p = .46, indicated that no single modality of presentation improved significantly more than the others. Posttest performance showed an increase of about 15% for the

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

AV (to 90.38%) and V-only conditions (from 59.63 to 74.38%), and about 12% for A-only (to 77.88%). Averaged across modalities, scores were higher for /ɹ/ and /l/ in final singleton, followed by final cluster, initial singleton, and then initial clusters, and higher for contexts with /e, ɛ/ and /ɑɪ/ compared to /u, o/.

Further comparisons revealed that, when visual information was present (AV and V-only conditions), the increase in scores after training was significantly greater in initial clusters versus singleton, F (2, 833) = 7.21, $p < .01$, and scores for initial positions showed greater improvement than those for final positions, F (1, 833) = 10.84, $p = .001$. Thus, the visual modality contributed strongly to the AV percept in the identification of /ɹ/ and /l/ in the more difficult initial positions for the Japanese and became more informative following training.

Changes in identification accuracy over time also varied with modality and vowel, F (8, 833) = 3.31, $p = .001$. Compared to the AV and A-only scores, the increase in V-only scores was smaller for /ɹ/ and /l/ with /o/ but greater with /ɑɪ/ and /e, ɛ/. In fact, V-only posttest scores with /ɑɪ/ exceeded those for A-only. Thus, the salience and information value of the visible articulatory movements for /ɹ/ and /l/ were greater in contexts with vowels produced with a more open-mouth position. The highest AV and A-only scores were for final singleton and cluster with all vowels. Results for intervocalic position revealed significant main effects for time and modality, F_t (1, 35) = 34.68, $p = .0001$; F_m (2, 35) = 30.42, $p = .0001$, and no significant interaction. The greatest improvement in accuracy for that position was in the AV condition from 68.7 to 90%.

For the A-only training group, the variables were time, position, and vowel. All had significant main effects, F_t (1, 273) = 178.43, $p = .0001$; F_p (3, 273) = 204.02, $p = .0001$; F_v (4, 273) = 44.08, $p = .0001$. The patterns of the results were similar to those of the AV group. Tukey's honestly significant difference (HSD) tests revealed that scores for /ɹ/ and /l/ were significantly lower for initial versus final positions and for contexts with /o/ versus /e, ɛ/. Overall, accuracy increased from 62.5 to 70.03%. This A-only pretest accuracy was comparable to the A-only pretest score for the AV training group (differing by only 2.63%). The interaction of Position × Vowel was also significant, F (12, 273) = 11.77, $p = .0001$, with scores lower in initial positions with /o/. The results for intervocalic position revealed a significant increase in accuracy from 63.0 to 72.6%, F (1, 7) = 81.00, $p = .0001$. The results for the control group revealed no significant improvement between pretest and posttest, F (1, 833) = 0.77, $p = .46$.

Comparison of AV versus A-only training. To compare the effects of training type, the A-only testing scores (pre- and posttest) for the AV group were compared with those from the A-only group (see Figure 1) as the auditory modality was the only one they shared. As noted above, the mean pretest scores differed by only 2.63%. Overall, AV training provided significantly greater improvement in perceptual accuracy, F (1, 14) = 7.87, $p < .05$.

Effects of phonetic context and talker in training. For the AV group, the variables were week of training (1–3), talker (1–5), position (4), and vowel (3); all had significant main effects, F_w (2, 1253) = 366.09, $p = .0001$; F_t (4, 1253) = 3.76, $p < .01$; F_p (3, 1253) = 225.75, $p = .0001$; F_v (2, 1253) = 38.27, $p = .0001$.

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

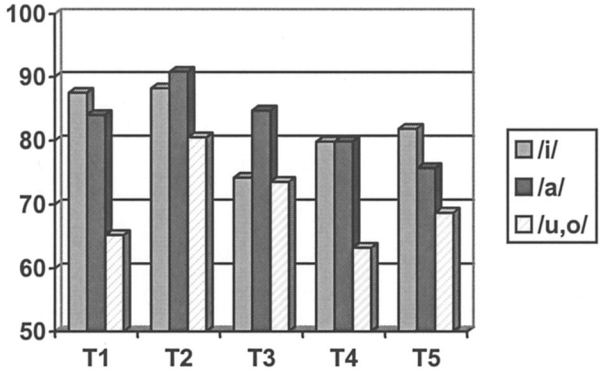


Figure 2. The Japanese AV group mean percentage correct in training for initial singleton position by talker (T) and vowel.

Tukey's HSD tests revealed that scores for each week differed significantly and climbed consistently from 72.17% at the end of Week 1 to 92.67% at the end of Week 3. Identification accuracy (mean percent correct in training) varied for tokens produced by the five talkers (Talkers 1–3 were female and 4–5 were male). Talker 1 (84.0%) was the highest followed by Talkers 2 (83.83%), 5 (81.83%), 4 (81.33%), and 3 (81.16%). The significant difference was between Talkers 1 and 3.

Overall scores for the four word positions differed significantly: final singleton (93.17%), final cluster (86.33%), initial singleton (78.50%), and initial cluster (71.83%). Scores for /ɹ/ and /l/ with rounded vowels (78.67%) were significantly lower than those with /i, ɪ/ (84.0%) and /ɑɪ/ (84.67%). Identification accuracy for initial singleton and cluster positions showed a greater increase over time than that for the final positions, especially during the first 2 weeks of training, F (6, 1253) = 16.40, p = .0001. In addition, during this time, improvement was greater for contexts with /ɑɪ/, whereas those with /ʊ, o/ increased the most between Weeks 2 and 3.

Improvement during training varied significantly according to talker, word position, and vowel, F (24, 1253) = 3.37, p = .0001, as shown in the contrast between Figures 2 (initial singleton) and 3 (final singleton). Initial singleton scores for Talker 2 were higher, perhaps because of her very rounded articulation of /ɹ/ and relatively salient apical contact for /l/. Across talkers, identification accuracy was lower for initial positions with rounded vowels and higher in final positions, especially with /ɑɪ/. Variability in identification accuracy was also evident within the tokens produced by a single talker. Scores for Talkers 4 and 5 for initial clusters with /ʊ, o/ were the lowest; however, the scores for these talkers for final clusters with /i, ɪ/ were the highest. Analysis of results for intervocalic position showed a significant main effect of week, F (2, 98) = 36.81, p = .0001, but not of talker, F (4, 98) = 2.15, p = .07.

For the A-only group, the variables were the same and all had significant main effects, F_w (2, 1253) = 425.32, p = .0001; F_t (4, 1253) = 16.26, p = .0001;

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

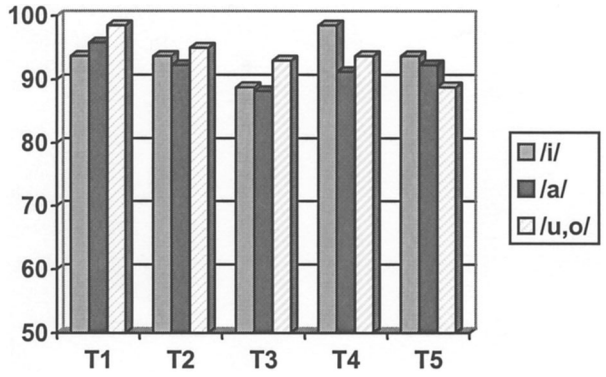


Figure 3. The Japanese AV group mean percentage correct in training for final singleton position by talker (T) and vowel.

F_p (3, 1253) = 654.44, p = .0001; F_v (2, 1253) = 23.16, p = .0001. Tukey's HSD tests revealed that scores for each week differed significantly and rose from 64% at the end of Week 1 to 80.17% at the end of Week 3. The interaction of Talker × Position × Vowel was also significant, F (24, 1253) = 3.31, p = .0001. As with the AV group, identification accuracy varied most across talkers for initial versus final positions. Perceptual accuracy was lowest for initial positions with /u̱, o̱/ and higher for final singleton position. Again for intervocalic position, there was a significant main effect of week, F (2, 98) = 60.59, p = .0001.

Tests of generalization (*familiar vs. unfamiliar talker*). Both training groups were given two tests of generalization to investigate differences in their ability to generalize to novel words spoken by a familiar talker (Talker 1) from training (TG1) and an unfamiliar talker (TG2). For the AV group, mean identification accuracy for TG1 (84.83%) was similar to the mean of 83.5% from Week 2 of training but was less than 92.83% for Week 3. AV scores were significantly higher than those for a single modality for both TG1 and TG2. Analysis of the results per modality demonstrated a significant main effect of talker, but only in the V-only condition, F (1, 105) = 9.79, $p < .01$, where accuracy rates were 69.5% for TG1 and 63.17% for TG2. This is also the case for intervocalic position for which the accuracy rates were 64% (TG1) compared to 50% (TG2). Significantly lower scores were found for initial cluster position in all conditions and for rounded vowel contexts when audio was present. When video was present, a significant Talker × Position interaction obtained. Scores for initial singleton with TG1 were higher but were almost identical to TG2 scores for initial cluster.

For the A-only group, the mean score for TG1 (75.33%) fell between the scores for this talker for Weeks 2 and 3 of training (Week 3 = 79.67%), and was significantly greater than that for TG2 (71%). Patterns were similar to the AV group with regard to position and vowel.

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

Effects of perceptual training on production

The pretest and posttest scores were tabulated for each participant in each training group. Results were analyzed by a four-factor repeated measures ANOVA with time (pretest–posttest) as the repeated measure and position (3), vowel (3), and phone (/ɹ/, /l/) as within-group factors.[3] Mean ratings (7-point scale) for the AV training group improved from 5.17 to 6.25. There was a significant main effect of time, F (1, 245) = 228.83, p = .0001; position, F (2, 245) = 19.08, p = .0001; vowel, F (2, 245) = 14.48, p = .0001; and phone, F (1, 245) = 10.71, p = .001. Post hoc tests revealed a significantly higher mean score for initial singleton versus final singleton and initial cluster and for contexts with /ɑ, ɑɪ/ compared to /i, ɪ/ and /u, o/. Scores for /ɹ/ were higher than those for /l/ (before and after training). A significant Position × Time interaction, F (4, 245) = 5.89, $p < .001$, revealed greater improvement for initial clusters. After perceptual training, there was much less variability in production ratings across word positions. Scores for intervocalic position again were analyzed separately and revealed significant improvement, F (1, 21) = 15.18, $p < .001$, but no significant difference between /ɹ/ and /l/.

Mean ratings for the A-only group were generally lower but improved significantly from 4.76 to 5.65, F (1, 245) = 10.25, $p < .01$. Other findings paralleled those of the AV group. Scores for the control group showed no significant improvement over the 3-week training period.

EXPERIMENT 2: KOREAN SPEAKERS—PERCEPTUAL TRAINING AND PRODUCTION

Experimental design

Experiment 2 explored the effect of visual input on perceptual training for a second L1 group—Korean ESL learners. As noted earlier, unlike Japanese, Korean has a voiced nonvelarized /l/ in syllable-final position, which represents an acoustic difference compared to the AE velarized postvocalic lateral. There is also a voiced apicoalveolar flap intervocalically. As with Experiment 1, factors under investigation included L1 influence, a comparison of AV and A-only training, and the effects of visual cues and phonetic environment. A comparison was made of the effect of multiple- versus single-talker training in tests of generalization using novel stimuli with a familiar and an unfamiliar talker as had been done in a previous study with Japanese speakers from the same population (Lively et al., 1993). Therefore, the AV and A-only training groups were further divided into two groups each: one received training with multiple talkers and the other with a single talker. In order for the single-talker training groups to be familiar with Talker 1 used in TG1, she had to serve as the training talker for that group.

Method

Experiment 2 involved eight NSs of Korean (from Seoul) assigned randomly to each of the four experimental groups plus a control group of eight who received the testing (AV, A-only, V-only) but no training. Background information such

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

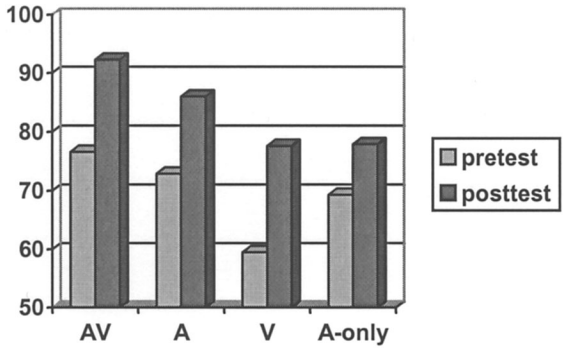


Figure 4. The Korean mean percentage correct pre- and posttest by training type and modality. The scores for the AV multiple-talker training group for auditory–visual (AV), auditory-only (A-only), and visual-only (V-only). The A-only scores for the A-only training group.

as their language program proficiency level, age range, time in the United States, and previous English experience was compatible with that of the Japanese. The same videotapes were used for both experiments. The single-talker groups received the same training stimuli as the others, but these were produced by only one talker. For these groups, five different tapes were prepared so that each training session throughout a week presented a different stimulus order.

As in Experiment 1, pretest and posttest recordings were made of each participant's productions of a randomized sequence of words taken from the perception testing stimulus set. The materials and procedure were identical, but different judges were used.

Results and discussion

Effectiveness of training per group (*AV, A-only*). Identification scores for each training group were tabulated. To assess the effects of training, ANOVAs were conducted on pretest and posttest scores for each group separately.

MULTIPLE-TALKER GROUPS. For the AV multiple-talker group, variables were time (pretest–posttest), modality (AV, A, V), position (4), and vowel (5). All had significant main effects, F_t (1, 833) = 308.36, p = .0001; F_m (2, 833) = 111.41, p = .0001; F_p (3, 833) = 33.08, p = .0001; F_v (4, 833) = 13.41, p = .0001. Tukey's HSD tests revealed significant differences in the scores for the three conditions. In Figure 4, the first three pairs of bars represent the pretest and posttest mean scores for the three conditions of stimulus presentation (AV, A-only, V-only) for the AV training group; the last pair of bars provides the scores for the A-only group. As shown, AV mean identification accuracy rose from 76.63 to 92.25%, V-only from 59.5 to 77.63%, and A-only from 72.88 to 86%.

In contrast to the Japanese, scores for the Koreans across modalities were significantly higher for /ɹ/ and /l/ in initial positions (singleton = 81.88%, clus-

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

ter = 80.13%) versus final positions (singleton = 77.5%, cluster = 70.25%). Similar to the Japanese though, scores overall were significantly higher for contexts with /e, ɛ/ (83.38%), and generally lower for the rounded vowels (74.76%). Lowest scores obtained for /ɹ/ and /l/ with rounded vowels in initial positions and /i, ɪ/ in final singleton, F (12, 833) = 5.33, p = .0001.

The increase in mean scores for final singleton and cluster positions when visual input was present (AV and V-only) was significantly greater than the increase in the A-only condition, F (6, 833) = 3.36, $p < .01$. In both the A-only and V-only conditions, scores were lower and comparable for /ɹ/ and /l/ with /ɑɪ/ and /i, ɪ/, suggesting an equal contribution from both modalities to the AV percept in these contexts. Results for intervocalic position revealed significant main effects for time, F (1, 35) = 14.63, p = .0005, and modality, F (2, 35) = 59.86, p = .0001. Identification accuracy for this position averaged across modalities increased from 75.83 to 86.2%.

For the A-only multiple-talker group, there were significant main effects for all variables: time, F (1, 273) = 109.52, p = .0001; position, F (3, 273) = 53.79, p = .0001; and vowel, F (4, 273) = 27.18, p = .0001. Patterns of results followed those for the AV group. Tukey's HSD tests revealed that scores were significantly higher in initial versus final positions, and /e, ɛ/ appeared again as the easiest vocalic environment. Mean identification accuracy increased from 69.25 to 77.88%. Identification accuracy for the intervocalic position improved significantly with training, F (1, 7) = 9.00, $p < .05$.

SINGLE-TALKER GROUPS. For the AV group, there were significant main effects for time, F (1, 833) = 797.84, p = .0001; modality, F (2, 833) = 45.06, p = .0001; position, F (3, 833) = 103.92, p = .0001; and vowel, F (4, 833) = 51.13, p = .0001. Scores increased from 71.75 to 89.75% in the AV condition, 65 to 82.75% in A-only, and 66.63 to 83.88% in V-only. In AV, the largest gain was for final cluster position, F (6, 833) = 5.73, p = .0001. The V-only scores exceeded those for A-only in pretest for initial singleton and final cluster positions and were higher in posttest for all positions except initial cluster. Across modalities, the areas of greatest improvement were noted for /ɹ/ and /l/ with /ɑɪ/ in initial singleton, /i, ɪ/ in initial cluster and final singleton, and /e, ɛ/ in final cluster.

For the A-only group, significant main effects were found for time, F (1, 273) = 151.44, p = .0001; position, F (3, 273) = 44.69, p = .0001; and vowel, F (4, 273) = 8.57, p = .0001. Overall mean identification accuracy increased from 64 to 76.75%. Scores were lower for /ɹ/ and /l/ with rounded vowels and /i, ɪ/ in final clusters and higher with /i, ɪ/ and /e, ɛ/ in initial positions, F (12, 273) = 2.41, $p < .01$. The results for the control group revealed no significant improvement between pretest and posttest, F (1, 833) = 0.93, p = .33.

Comparison of AV versus A-only training. As in Experiment 1, the A-only testing scores for the AV and A-only groups were compared. Comparison was made between the two multiple-talker groups and then between the two single-talker groups. Variables were time, group (AV, A-only), position, and vowel. Results for the multiple-talker groups (see Figure 4) revealed significantly greater im-

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

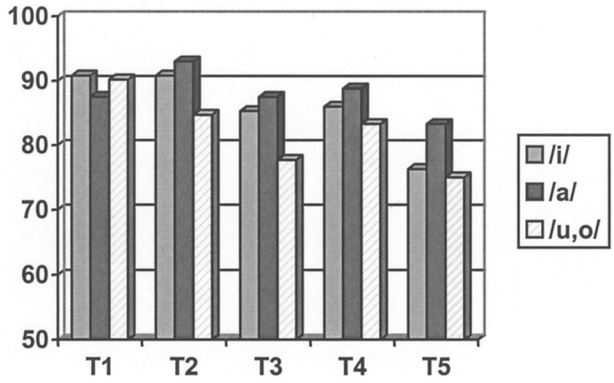


Figure 5. The Korean AV multiple-talker group mean percentage correct in training for initial singleton position by talker (T) and vowel.

provement in identification accuracy following AV training, F (1, 14) = 7.46, $p < .05$. This was also the case for the single-talker groups, F (1, 14) = 9.94, $p < .01$, and for the analysis of the intervocalic position.

Effects of phonetic context and talker in training

MULTIPLE-TALKER GROUPS. Results indicated significant main effects for week of training, F (2, 1253) = 203.51, $p = .0001$; talker F (4, 1253) = 14.23, $p = .0001$; position, F (3, 1253) = 27.05, $p = .0001$; and vowel, F (2, 1253) = 6.88, $p = .001$. Tukey's HSD tests indicated that scores differed significantly from week to week and rose from 78.0 to 95.17% identification accuracy (mean percent correct in training). Accuracy for tokens produced by Talker 1 (89.67%) was the highest, followed by Talkers 2 (88.33%), 3 (86.5%), 4 (84%), and 5 (82.5%), as averaged across 3 weeks of training. The significant difference was between Talker 1 and Talkers 3, 4, and 5.

Overall scores for /ɹ/ and /l/ in initial clusters (91.5%) were significantly higher than those for the other positions (initial singleton = 85.33%, final singleton = 84.17%, final cluster = 83.83%). Mean scores for contexts with /ɑɪ/ were higher (87.83%) than those with /ʉ, o/ (86.17%), which were significantly higher than those with /i, ɪ/ (84.67%).

The areas of greatest talker difference were the more difficult final positions, especially with the rounded vowels and /i, ɪ/, as shown by the contrast between Figures 5 (initial singleton) and 6 (final singleton), F (24, 1253) = 2.60, $p = .0001$. In general, identification accuracy varied significantly within and across talkers of both genders. Results for /ɹ/ and /l/ in intervocalic position revealed significant main effects of week, F (2, 98) = 23.82, $p = .0001$, and talker, F (4, 98) = 8.94, $p = .0001$. Accuracy increased from 85.63 (Week 1) to 97.13% (Week 3).

For the A-only group, significant main effects were found for week, F (2, 1253) = 299.10, $p = .0001$; talker, F (4, 1253) = 40.26, $p = .0001$; position,

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

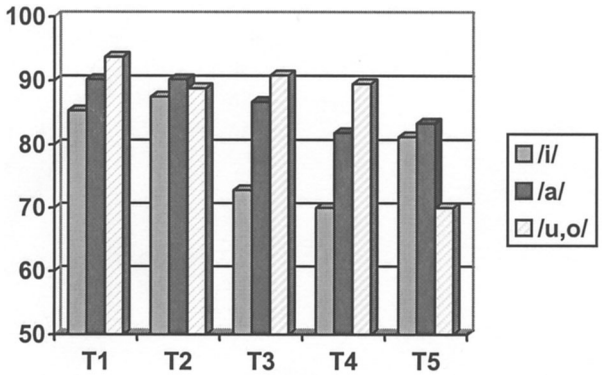


Figure 6. The Korean AV multiple-talker group mean percentage correct in training for final singleton position by talker (T) and vowel.

F (3, 1253) = 57.46, p = .0001; and vowel, F (2, 1253) = 7.38, p < .001. Tukey's HSD tests revealed significant increases in accuracy from week to week. Identification accuracy for /ɹ/ and /l/ in initial clusters was higher than in final positions, and scores were higher for contexts with /ɑɪ/ than /i, ɪ/. Significant interactions followed the patterns found for the AV group.

SINGLE-TALKER GROUPS. Results for the AV group revealed significant main effects for all variables: week, F (2, 245) = 124.87, p = .0001; position, F (3, 245) = 32.07, p = .0001; and vowel, F (2, 245) = 17.91, p = .0001. Identification accuracy increased significantly with each training week, rising from 78% after Week 1 to 92.03% at the end of Week 3. The greatest improvement in accuracy for final singleton and cluster occurred between Weeks 2 and 3, whereas larger increases for the initial positions occurred between Weeks 1 and 2, F (6, 245) = 2.94, p < .01. Again, scores were significantly lower for /ɹ/ and /l/ in final positions with /i, ɪ/, F (6, 245) = 8.16, p = .0001. Variables for the A-only single-talker group were the same and all had significant main effects, F_w (2, 245) = 555.58, p = .0001; F_p (3, 245) = 95.38, p = .0001; F_v (2, 245) = 131.73, p = .0001. Interactions followed the pattern of results for the AV group.

Tests of generalization (*familiar vs. unfamiliar talker*). All training groups were given two tests of generalization involving novel words produced by a familiar talker from training (TG1) and a new talker (TG2). Recall that Talker 1 from training was used for TG1. In this study, multiple- versus single-talker training results were directly compared within the AV groups and then within the A-only groups to determine if multiple-talker training results in better generalization to novel stimuli and a new talker as reported for Japanese speakers by Lively et al. (1993). Recall also that the number of vowels was reduced to two groups: rounded /u, o/ and unrounded /e, ɛ/ (an untrained context).

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

Table 1. *Korean mean percentage correct for Training Group* (*multiple vs. single talker*) × *Talker* (*familiar TG1 vs. unfamiliar TG2*) *per modality*

	Auditory–Visual		Auditory Only		Visual Only	
	TG1	TG2	TG1	TG2	TG1	TG2
Multiple	94.67	91.83	89.00	87.67	79.17	66.67
Single	88.17	83.83	76.17	79.33	70.00	64.00

AV TRAINING GROUPS. As with the pretest and posttest, stimuli for the tests of generalization were presented in AV, A-only, and V-only conditions. Results for each modality were analyzed separately. Variables were group (multiple- vs. single-talker training), talker (familiar TG1 vs. unfamiliar TG2), position, and vowel. As shown in Table 1, the generalization performance of both training groups was similar across modalities.

The effect of training group (multiple vs. single talker) was marginally significant in the AV condition, F (1, 14) = 4.58, p = .0505. However, in the A-only condition, identification accuracy for the single-talker group was lower for tokens produced by the unfamiliar talker in the untrained vowel context /e, ɛ/, F (1, 14) = 5.97, $p < .05$, and for the relatively difficult contexts for the Koreans: initial singleton with /u, o/, and final singleton with /e, ɛ/, F (3, 42) = 3.30, $p <$ .05. In the V-only condition, the single-talker group showed a greater difference between the scores for tokens produced by the familiar versus the unfamiliar talker, especially for initial and final singleton positions, F (3, 42) = 3.44, $p <$ .05. Also, in the V-only condition, the mean score for the intervocalic position was significantly higher, F (1, 14) = 7.83, $p < .05$, for the multiple-talker group.

A-ONLY TRAINING GROUPS. Results were similar to those found for the A-only condition for the AV groups described above. There was no effect of group, but there were two significant interactions involving talker. Scores for /ɹ/ and /l/ in initial singleton were lower for the unfamiliar versus the familiar talker, F (3, 42) = 3.76, $p < .05$. Overall scores also differed more between the two vocalic contexts for the unfamiliar talker, F (1, 14) = 5.70, $p < .05$.

Effects of perceptual training on production. The pretest and posttest scores were tabulated for participants in each training group. Within the multiple-talker training category, mean production ratings (7-point scale) increased from 5.01 to 6.39 (AV group) and from 4.84 to 5.92 (A-only group). For the single-talker groups, ratings increased from 4.88 to 6.12 (AV group) and 4.83 to 5.90 (A-only group). Results for the groups were analyzed separately by a four-factor repeated measures ANOVA with time (pretest–posttest) as the repeated measure and position (3), vowel (3), and phone (/ɹ/, /l/) as within-group factors.

For the AV multiple-talker group, there was a significant main effect of time, F (1, 245) = 167.27, p = .0001; position, F (2, 245) = 20.33, p = .0001; and

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

vowel, F (2, 245) = 6.20, $p < .01$. In contrast to the Japanese, there was no significant main effect of phone for the Koreans, F (1, 245) = 0.16, $p = .69$. However, there were two significant interactions involving this variable. Scores for productions of /ɹ/ were lower in initial singleton whereas those for /l/ were lower in final singleton, F (2, 245) = 25.93, $p = .0001$. The variation noted in pretest scores for the production of both sounds as a function of position significantly decreased with the greatest improvement noted for /l/ in final singleton, F (2, 245) = 3.39, $p < .05$. Improvement was also found for the intervocalic position, F (1, 21) = 6.95, $p < .05$, with no significant difference between the sounds. As with the Japanese, scores were significantly higher for /ɹ/ and /l/ with /ɑɪ/ compared to the other vowels.

Results for the AV single-talker group followed the same pattern with the addition of a Position × Vowel × Time interaction, F (4, 245) = 2.10, $p < .05$. Scores improved significantly for final singleton with /ɑɪ/ and /i, ɪ/. For the groups receiving A-only training, there were also significant main effects of time, position, and vowel with a significant Position × Vowel interaction ($p < .05$) and the greatest improvement was noted for final singleton with /ɑɪ/. Scores for the control group showed no significant improvement from pretest to posttest, F (1, 245) = 0.86, $p = .49$.

GENERAL DISCUSSION

The findings of this study demonstrated the role of the visual modality as a second channel of input for L2 perceptual learning, the interaction of both vowel and word position on the perception and production of /ɹ/ and /l/, the influence of talker characteristics on perceptual accuracy, and the contrasting effects of two L1 phonological systems on L2 development. Although both the Japanese and Korean groups showed improvement in identification accuracy with A-only training, results revealed a significant advantage for AV training. Visual information contributed significantly to the percept in the most challenging phonetic environments for each L1 group: initial clusters with /ʊ, o/ for the Japanese and final singleton with /i, ɪ/ for the Koreans. The difficulty of these environments shows some influence from L1 phonology (i.e., Japanese has an utterance-initial flap and Korean has a syllable-final nonvelarized lateral).

In addition, Experiment 1 (Japanese) revealed generalization of training to novel stimuli spoken by a familiar and an unfamiliar talker; however, perceptual accuracy was generally greater for the familiar talker, especially in terms of familiarity with visible articulatory gestures. Experiment 2 (Koreans) addressed the issue of multiple- versus single-talker training through direct comparison. Lively et al. (1993) reported that Japanese speakers receiving single-talker training showed poor performance in generalization to novel words spoken by a familiar talker and failed to generalize to new tokens produced by an unfamiliar talker. In the present study, both types of training showed successful generalization performance; however, in the AV condition, the effect of group (multiple vs. single talker) was only marginally significant, suggesting that information from both modalities facilitated comprehension of unfamiliar speech.

The study also replicated the transfer of perceptual training to significant

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

production improvement in the absence of explicit production instruction. For the Japanese groups, scores were higher for initial singleton and contexts with /ɑ, ɑɪ/. In both perception and production, the greatest improvement was noted for /ɹ/ and /l/ in initial clusters. For the Korean groups, the greatest improvement was found for /l/ in final singleton, similar to the perception results when visual cues were present. The variability in production accuracy as a function of context decreased with training.

Findings of experiments demonstrating the context- and talker-dependent nature of speech processing support the view that sources of variability or complexities in the speech signal are not merely noise discarded from the signal during processing, but are a part of subsequent neural representations (e.g., Goldinger, 1997). This perspective challenges traditional reductionist (or abstractionist) phonological theories. Quotes from two contemporary researchers illustrate these contrasting viewpoints on the issue of noise.

Brown (1999) wrote:

> . . . perceiving speech in terms of phonemic categories undoubtedly aids processing and facilitates comprehension of the linguistic signal. Native speakers are continually faced with variable realizations of segments, due to coarticulation, sloppy articulation or interspeaker variability. By filtering out this irrelevant "noise" in the acoustic signal, the memory load put on the auditory system is greatly reduced and processing can proceed more quickly. Those variations in the acoustic signal that do not contribute to differences in meaning are simply not perceived by the listener. (p. 19)

In contrast, Goldinger (1997) commented:

> By an abstractionist view, perceptual idiosyncrasies of words are normalized, allowing listeners to identify tokens as types. By an episodic view, words are recognized against a background of countless, detailed episodes. The speech signal is not a noisy vehicle of linguistic content; the signal itself is an integral aspect of later representation. (pp. 59–60)

The results of the present study are consistent with an episodic view that memory encoding of speech involves storage of the attended perceptual details of individual episodes or traces preserving aspects of contextual variability. Whereas the phonemic representations of some phonological theories involve an abstraction process at the time of encoding in memory, an episodic view such as multiple-trace theory (see Hintzman, 1986) does not stipulate that abstract knowledge be stored; rather, it can be derived from a composite of episodic traces at the time information is retrieved from memory. A category is regarded as an aggregate of individual exemplars activated together at the time information is retrieved.

Learning in terms of an episodic model consists of copying the features of an experience into a trace. The features that comprise the perceptual representations depend upon the attention given to the auditory and visual attributes of the stimulus relevant to the task. In perceptual training studies such as the current one, multiple exemplars plus immediate feedback enhance the learning process by allowing observations of within-category similarities and between-category distinctions across contexts and talkers. Learners are able to attend to those stimulus dimensions that have provided useful input in their categorization deci-

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

sions, increasing the salience and information value of relevant characteristics from both modalities in the case of AV training and adding traces to memory. As a consequence of learning, new L2 traces should be less ambiguous in content, specifically, less confusable or at greater psychological distance from existing L1 traces. In the case of AE /ɹ/ and /l/, L2 learners must focus attention on F3 transitions to distinguish them auditorily and on the articulatory gestures that distinguish them visually. For the Japanese, this involves a shift in attention from F2 to F3 (Yamada, 1995).

Therefore, the greater the number and variety of L2 exemplars used in training as in a high-variability stimulus set, the greater will be the opportunity to create a weighting scheme that is able to focus attention on the critical stimulus properties (auditory and/or visual) for the encoding of less ambiguous L2 traces in memory. In turn, probes to memory for perceptual identification will provide better matches to stored L2 traces.

The number of relevant dimensions over which learners must distribute attention to optimize categorization performance perhaps contributes to an explanation of why spectral distinctions (e.g., /ɹ/–/l/) are more difficult to acquire than temporal ones such as voice-onset time (Strange & Jenkins, 1978). In addition, the variability present in successful stimulus training sets no doubt is an advantage for training /ɹ/ and /l/ as it mirrors the variability that exists for these sounds across contexts and talkers in the natural language environment.

The objective of L2 perceptual learning, then, is to create a situation in which the response from an aggregate of L2 traces acting in concert overshadows that from L1 traces. Retention and generalizability to new exemplars in perceptual learning found in recent studies may be accounted for by the advantage of redundancy captured by the category prototype concept as a representation of the shared features of multiple traces. This also suggests an explanation for the limitations of relatively brief periods of training using fewer exemplars or those not from natural speech.

For the L2 learner, several studies have now shown that perceptual training transfers to production improvement in the absence of explicit production training. It is perhaps the case that L2 learners attempt to coordinate information about perception and production in category development, as Jusczyk (1993) suggested children do in L1 acquisition. I did note that learners in the AV training sessions imitated the lip movements of the talkers they were observing, although the training involved perception only, and their task was to circle the word from a minimal pair that they thought the talker on the screen was saying.

As is evident in the literature (Borden et al., 1983; Goto, 1971; Mochizuki, 1981; Sheldon & Strange, 1982), perception and production do not often follow parallel development in L2 learning. Production skills may precede perception skills. Unlike perception, production may be mediated by knowledge of how the sounds are produced and proprioceptive feedback. Production ability may also benefit from learners' specific attention to it because it helps them to integrate into the host community (Sheldon, 1985).

To relate the processes of speech perception and production, models have either required recourse to a special phonetic module, as in the motor theory (e.g., Liberman & Mattingly, 1985, 1989), or predicted parallel development of

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

perception and production abilities, as in the perceptual assimilation model (e.g., Best, 1995) and the speech learning model (e.g., Flege, 1995). A correspondence between perception and production is supported in the neurophysiological literature. For example, evidence from electrical stimulation mapping techniques and recordings of changes in neuronal activity indicate that some of the same neuronal populations are involved in speech production and perception (Ojemann, Cawthon, & Lettich, 1990). Pulvermüller and Shumann (1994) point out that neurons causing articulation of a given sound by controlling the contraction of particular muscles become active together with the sensory neurons in the auditory cortex that have responded to the features of that sound through the auditory system, which provides feedback based on the utterances. Following Hebbian principles (i.e., neurons become associated more strongly when they are frequently active at the same time), they conclude that neurons involved in production and perception strengthen their connections. This relationship is consistent with the improvement of production following perceptual training and with the similarities in contextual influence and patterns of development found in the perception and production data for each L1 group in this study.

CONCLUSION

In conclusion, recent studies in perceptual training have demonstrated that natural speech, stimulus variability, an identification paradigm, and feedback during training contribute to the development of perceptual categories that are robust across phonetic environments and talkers, thus enabling generalization to novel stimuli and talkers and transfer to production improvement. These results have established the criteria for successful learning with the objectives of maximum generalization and retention.

In addition to replicating the above findings, the present study revealed that training two modalities (auditory and visual) simultaneously was superior in improving perceptual accuracy to training only one, consistent with the computational studies of de Sa and Ballard (1997). The contribution of visible speech was most evident in the perception of /ɹ/ and /l/ in the more difficult contexts based on L1 phonology. It is also evident that both adjacent vowel and word position have significant effects on perception and production of /ɹ/ and /l/.

The benefit of multiple- versus single-talker training is less clear. The present study compared the two training group types directly for the Korean speakers and found only a marginally significant effect on performance in tests of generalization when bimodal information was present. Although multiple talkers likely offer benefits in perceptual training because they are a component of a variable stimulus set, it is also likely that performance in generalizing to familiar and unfamiliar talkers is related to the intelligibility of an individual's articulatory gestures.

Viewing the acquisition of L2 speech as a bimodal process has obvious pedagogical implications. Massaro and colleagues (1998) suggest that their animated talking head "Baldi" can facilitate language learning by providing visual input such as displays of articulatory movements inside the mouth. However, at this juncture, the question arises as to the value of segmental-level perceptual train-

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

ing using words spoken in isolation because such stimuli are devoid of the sandhi phenomena typical of the connected speech generally encountered. A subsequent study applying the gating paradigm to auditory–visual speech revealed transfer of such training with minimal pairs, including /ɹ/ and /l/, to earlier word identification in connected speech by L2 learners (Hardison, 1998b).

Although the findings of studies over the past decade have established many of the requirements for successful L2 perceptual training, further research is needed in such areas as the link between perception and production.

ACKNOWLEDGMENTS

This article is based on a PhD dissertation submitted to the Department of Linguistics at Indiana University. The work was supported by a Grant in Aid for Research to the author from Indiana University, the Fred W. Householder Fund, and NIH NIDCD Research Grant DC00111 to Indiana University (Speech Research Laboratory). I am deeply grateful for comments on this project from David Pisoni, Bob Port, Stuart Davis, and Harry Gradman; to Dave Derkacy and Dave Rust for invaluable technical production; and to Harry So and Jonathan Margolin for assistance in statistical analysis. Earlier versions of various aspects of this research were presented at the Second Language Research Forum (University of Minnesota, 1999), New Sounds 2000 (Amsterdam), the American Association for Applied Linguistics (Vancouver, 2000), and the International Conference on Spoken Language Processing (Denver, 2002).

NOTES

1. The McGurk effect refers to the influence of a talker's lip movements on an observer's perception of speech sounds (e.g., McGurk & MacDonald, 1976). Typically, experiments examining this effect involve the dubbing of the audio production of one sound onto the video of another sound to create a discordant condition (e.g., visual /pɑ/ and auditory /kɑ/). Experiments with L2 learners have also included matching visual and auditory cues (Hardison, 1999; Sekiyama & Tohkura, 1993).
2. A similar approach has been used to train language-learning impaired children by initially emphasizing speech components the children had been unable to process (e.g., Merzenich, Jenkins, Johnston, Schreiner, Miller, & Tallal, 1996). However, as of April 2000, retention had not been tested (Merzenich, 2000).
3. The differences in errors for /ɹ/ and /l/ were not examined in the statistical analysis of the perception data because of the number of existing variables.

REFERENCES

Aslin, R. N., & Pisoni, D. B. (1980). Some developmental processes in speech perception. In G. H. Yeni–Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: Vol. 2. Perception* (pp. 67–96). New York: Academic Press.

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.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.),

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

Speech perception and linguistic experience: Issues in cross-language research (pp. 171–204). Timonium, MD: York Press.

Binnie, C. A., Jackson, P. L., & Montgomery, A. A. (1976). Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation. *Journal of Speech and Hearing Disorders, 41,* 530–539.

Borden, G., Gerber, A., & Milsark, G. (1983). Production and perception of the /r/–/l/ contrast in Korean adults learning English. *Language Learning, 33,* 499–526.

Bradlow, A. R., Akahane–Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61,* 977–985.

Brown, C. A. (1999). The interrelation between speech perception and phonological acquisition from infant to adult. In J. Archibald (Ed.), *Second language acquisition and linguistic theory*. Malden, MA: Blackwell.

Campbell, R. (1987). Lip-reading and immediate memory processes or on thinking impure thoughts. In B. Dodd & R. Cambell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 243–255). London: Erlbaum.

Cochrane, R. M. (1980). The acquisition of /r/ and /l/ by Japanese children and adults learning English as a second language. *Journal of Multilingual and Multicultural Development, 1,* 331–360.

de Jonge, C. (1995). *Interlanguage phonology: Perception and production*. Unpublished doctoral dissertation, Indiana University, Bloomington.

de Sa, V., & Ballard, D. H. (1997). Perceptual learning from cross-modal feedback. In R. L. Goldstone, P. G. Schyns, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 36, pp. 309–351). San Diego, CA: Academic Press.

Ellis, N. C. (2002). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition, 24,* 297–339.

Erber, N. P. (1974). Discussion: Lipreading skills. In R. E. Stark (Ed.), *Sensory capabilities of hearing impaired children* (pp. 69–73). Baltimore, MD: University Park Press.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). Timonium, MD: York Press.

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English /r/ and /l/ accurately. *Language and Speech, 38,* 25–55.

Franks, J. R., & Kimble, J. (1972). The confusion of English consonant clusters in lipreading. *Journal of Speech and Hearing Research, 15,* 474–482.

Gesi, A. T., Massaro, D. W., & Cohen, M. M. (1992). Discovery and expository methods in teaching visual consonant and word identification. *Journal of Speech and Hearing Research, 35,* 1180–1188.

Gillette, S. (1980). Contextual variation in the perception of L and R by Japanese and Korean speakers. *Minnesota Papers in Linguistics and the Philosophy of Language, 6,* 59–72.

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.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "l" and "r." *Neuropsychologia, 9,* 317–323.

Hagiwara, R. E. (1995). Acoustic realization of American /r/ as produced by women and men (Doctoral dissertation, University of California, Los Angeles). *UCLA Working Papers in Phonetics, 90*.

Hardison, D. M. (1998a). *Acquisition of second-language speech: Effects of visual cues, context and talker variability*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Hardison, D. M. (1998b). Spoken word identification by native and nonnative speakers of English: Effects of training, modality, context and phonetic environment. In *Proceedings of the 5th International Conference on Spoken Language Processing* (Vol. 5, pp. 1875–1878). Sydney, Australia: Causal Productions PTY Ltd. [CD-ROM]

Hardison, D. M. (1999). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language Learning, 49*(Suppl. 1), 213–283.

Hardison, D. M. (2000). The neurocognitive foundation of second-language speech: A proposed

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

scenario of bimodal development. In B. Swierzbin, F. Morris, M. E. Anderson, C. A. Klee, & E. Tarone (Eds.), *The interaction of social and cognitive factors in SLA* (pp. 312–325). Somerville, MA: Cascadilla Press.

Hattori, T. (1987). A study of nonverbal intercultural communication between Japanese and Americans—Focusing on the use of the eyes. *Japan Association of Language Teachers, 8,* 109–118.

Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology, 10,* 83–94.

Ingram, J. C. L., & Park, S.-G. (1998). Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners. *Journal of the Acoustical Society of America, 103,* 1161–1174.

Jamieson, D. E., & Moroson, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/–/θ/ contrast by francophones. *Perception & Psychophysics, 40,* 205–215.

Johnson, K., & Mullennix, J. W. (1997). *Talker variability in speech processing*. San Diego, CA: Academic Press.

Jusczyk, P. W. (1992). Developing phonological categories from the speech signal. In C. A. Ferguson, L. Menn, & C. Stoel–Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 17–64). Timonium, MD: York 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.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kim–Renaud, Y.-K. (1974). *Korean consonantal phonology*. Unpublished doctoral dissertation, University of Hawaii.

Kricos, P. B., & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *Volta Review, 84,* 219–225.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science, 243,* 489–494.

Liberman, A. M., Miyawaki, K., Jenkins, J. J., & Fujimura, O. (1973). Cross-language study of the perception of the F3 cue for [r] vs. [l] in speech- and nonspeech-like patterns. *Journal of the Acoustical Society of Japan, 29,* 315.

Lindau, M. (1985). The story of /r/. In V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 157–168). Orlando, FL: Academic Press.

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.

MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics, 2,* 369–390.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., Cohen, M. M., & Gesi, A. T. (1993). Long-term training, transfer, and retention in learning to lipread. *Perception & Psychophysics, 53,* 549–562.

McCandliss, B. D., Conway, M., Fiez, J. A., Protopapas, A., & McClelland, J. L. (1998). Eliciting adult plasticity: Both adaptive and non-adaptive training improves Japanese adults' identification of English /r/ and /l/. *Society for Neuroscience Abstracts, 24,* 1898.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Merzenich, M. (2000, April). *Brain plasticity contributing to origins of language and reading impairments, and their training based remediation*. Lecture given at the Michigan State University Development and Learning Workshop, East Lansing.

Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., & Tallal, P. (1996). Temporal processing deficits of language-learning impaired children ameliorated by training. *Science, 271,* 77–81.

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

Miyawaki, K., Strange, W., Verbrugge, R. R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of /r/ and /l/ by native speakers of Japanese and English. *Perception & Psychophysics, 18,* 331–340.

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics, 9,* 283–303.

Moroson, D. E., & Jamieson, D. G. (1989). Evaluation of a technique for training new speech contrasts: Generalization across voices, but not word-position or task. *Journal of Speech and Hearing Research, 32,* 501–511.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics, 57,* 989–1001.

Ojemann, G. A., Cawthon, D. F., & Lettich, E. (1990). Localization and physiological correlates of language and verbal memory in human lateral temporoparietal cortex. In A. B. Scheibel & A. F. Wechsler (Eds.), *Neurobiology of higher cognitive function* (pp. 185–202). New York: Guilford Press.

Pisoni, D. B. (1973). Auditory and phonetic codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13,* 253–260.

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.

Pisoni, D. B., Lively, S. E., & Logan, J. S. (1994). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In J. Goodman & H. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 121–166). Cambridge, MA: MIT Press.

Price, P. J. (1981). *A cross-linguistic study of flaps in Japanese and in American English*. Unpublished doctoral dissertation, University of Pennsylvania.

Pulvermüller, F., & Schumann, J. H. (1994). Neurobiological mechanisms of language acquisition. *Language Learning, 44,* 681–734.

Robert–Ribes, J., Schwartz, J-L., & Escudier, P. (1995). A comparison of models for fusion of the auditory and visual sensors in speech perception. *Rapport de recherche de l'ICP, 4,* 121–139.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T., & Simola, J. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 127,* 141–145.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics, 59,* 73–80.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21,* 427–444.

Sheldon, A. (1985). The relationship between production and perception of the /r/–/l/ contrast in Korean adults learning English: A reply to Borden, Gerber, and Milsark. *Language Learning, 35,* 107–113.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics, 3,* 243–261.

Shimizu, K., & Dantsuji, M. (1983). A study on the perception of /r/ and /l/ in natural and synthetic speech sounds. *Studio Phonologica, 17,* 1–14.

Shin, S.-H. (1997). Correspondence in Kyungsang Korean truncation. In S. Davis (Ed.), *Optimal viewpoints* (pp. 115–138). Bloomington: Indiana University Linguistics Club Publications.

Strange, W., & Dittman, S. (1984). Effects of discrimination training on the perception of /r–l/ by Japanese adults learning English. *Perception & Psychophysics, 36,* 131–145.

Strange, W., & Jenkins, J. J. (1978). Role of linguistic experience in perception of speech. In R. D. Walk & H. L. Pick (Eds.), *Perception and experience* (pp. 125–169). New York: Plenum Press.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

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

Tsujimura, N. (1996). *An introduction to Japanese linguistics*. Oxford: Blackwell.

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press

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. *Journal of Speech and Hearing Research, 20,* 130–145.

Watson, C. S., Qiu, W. W., Chamberlain, M. M., & Li, X. (1996). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *Journal of the Acoustical Society of America, 100,* 1153–1162.

Werker, J. F. (1994). Cross-language speech perception: Developmental change does not involve loss. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from sounds to spoken words* (pp. 93–120). Cambridge, MA: MIT Press.

Yamada, R. A. (1995). Age and acquisition of second language speech sounds: Perception of American English /r/ and /l/ by native speakers of Japanese. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 305–320). Timonium, MD: York Press.

Yamada, R. A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & Psychophysics, 52,* 376–392.

Yu, K., & Jamieson, D. G. (1993). Training of the English /r/ and /l/ speech contrasts in Korean listeners. *Canadian Acoustics, 21,* 107–108.

https://doi.org/10.1017/S0142716403000250 Published online by Cambridge University Press