SPOKEN WORD IDENTIFICATION BY NATIVE AND NONNATIVE SPEAKERS OF ENGLISH: EFFECTS OF TRAINING, MODALITY, CONTEXT AND PHONETIC ENVIRONMENT

Debra M. Hardison

University of California, Davis

ABSTRACT

Several experiments explored the contribution of visual information (lip movements) to spoken word identification by Japanese and Korean learners of English as a second language (ESL) and native speakers (NSs) of English, and its interaction with sentence context, phonetic environment and, for ESL learners, perceptual training (involving /ɹ,l,p,f,θ,s/) using minimal pairs.

Parallel tests (pre- and posttraining), applying the gating technique to videotaped stimuli, were given to the learners (presented audiovisually (AV) or audio-only (A-only) depending upon training group type). Stimuli were familiar, bisyllabic words beginning with the following visual categories: bilabial (/p/), labiodental (/f/), /ɹ/, /l/, and nonlabials (/s,t,k/) combined with high, low and rounded vowels. Test (pretest-posttest), initial consonant-vowel (CV) sequence, modality of presentation (AV vs. A), and condition (context or excised word) were independent variables. Groups of NSs were presented with the same stimuli.

NNS results revealed word identification was significantly earlier after perceptual training, in AV vs. A-only presentation, in context vs. excised word condition, and varied significantly with initial CV sequence. NS results also revealed significant effects of modality (earlier identification in AV vs. A-only), context, and initial CV sequence. Findings indicate the transfer of perceptual training from segment identification to the process of word identification in connected speech, and are consistent with a multiple-trace model of spoken language processing incorporating visual input.

1. INTRODUCTION

In a recent study, Japanese and Korean ESL learners were trained using minimal pairs to identify American English (AE) /ɹ/ and /l/ [1,2]. The contribution of visual information from talkers' lip movements, phonetic environment (adjacent vowel and word position), talker familiarity, and the transfer of perceptual training to production were investigated. Previous training studies had focused only on auditory input [3,4]. Results indicated that for both groups, AV training (vs. A-only) provided significantly greater improvement in identification accuracy of /ɹ/ and /l/. Visual input contributed the most to the bimodal percept for the more difficult phonetic environments based on similar L1 sounds (i.e., initial singleton and cluster positions for speakers of Japanese in which a flap occurs utterance-initially; final positions for speakers of Korean in which a nonvelarized /l/ occurs in syllable-final position). Significant effects of talker (in training), word position and vocalic context were found. Tests of generalization revealed successful transfer to novel stimuli and a new talker. In addition, production of /ɹ/ and /l/ improved significantly as a result of perceptual training. Learners' perceptions of these sounds were indicative of the development of highly context- and talker-dependent neural representations consistent with the encoding of all attended perceptual details in multiple traces in memory. Perceptual training involving contrasts /p/-/f/ and /θ/-/s/ revealed a similar benefit of AV vs. A-only presentation, and production improvement. This training study demonstrated that the adult perceptual system can be modified to promote development of perceptual categories robust across stimulus and talker variability, the superiority of AV training over A-only, the second-language (L2) learners' use of visual cues prior to training [5], and the enhancement of their information value with training.

Nevertheless, a training procedure which utilizes words spoken in isolation raises the question of whether learners can transfer the perceptual abilities developed in minimal pair training to the task of identifying words in connected speech - a task more often encountered in the language environment. The remainder of this paper focuses on experiments addressing this issue.

Few studies have explored spoken word identification by L2 learners [6], and none have considered bimodal input and its potential interaction with variables such as sentence context, the phonetic and visible articulatory structure of words, and the effects of training. To explore the influence of these factors, the gating paradigm was used. The standard auditory gating format involves successive presentations of increasing amounts of a target word [7]. With each presentation, subjects write down the word they think is being presented. Although not considered an on-line task, gating is well-suited for use with special populations [8], and AV tasks. As the onset of articulatory movements for a sound generally precedes its acoustic component in AV research, gating allows an assessment of the contribution of visual information to word identification at each gate. Its application requires some modifications to the standard unimodal format: gate durations correspond to numbers of video frames and the gating onset is determined by the acoustic signal to ensure that only information pertaining to the target word is presented. Previous auditory studies revealed that successive presentations of stimulus information did not significantly influence performance [9,10]. As the effects of word length, word frequency and amount of preceding sentence context have been well documented with this paradigm in the literature [7], these variables were not investigated further and were held constant. Within-subjects variables in the present study were test (pretest-posttest only for NNSs using parallel tests) and

initial CV sequence of the target words. Between-subjects variables were modality of presentation (AV vs. A-only) and condition (sentence context vs. excised word). It was hypothesized that words would be identified earlier with AV input but that this would vary, according to the initial CV sequence and, for NNSs, as a result of training. In general, visual information was expected to contribute most when auditory intelligibility was lowest, that is, in the A-only presentation of excised words. Context was expected to facilitate the identification process; in fact the ability of NNSs to utilize context in spoken language processing has not been extensively investigated.

2. EXPERIMENT ONE - JAPANESE

A total of 32 Japanese ESL learners, intermediate level, were given parallel gating tests before and after 3 weeks of perceptual training. (Training stimuli involved minimal pairs contrasting /ɹ/-/l/, /p/-/f/, /θ/-/s/ produced by multiple talkers.) The modality of presentation for the gating tests matched that of their training. For 16 subjects, stimuli were presented AV and for 16, they were A-only. Each of these groups was further divided into two equal groups: one presented with the preceding sentence context plus the gated target word, and one with the target word only excised from context.

2.1 Materials and Procedure

Bisyllabic targets (judged highly familiar by a peer group) began with the following visual categories: bilabial (/p/), labiodental (/f/), /ɹ/, /l/, and nonlabials (/s,t,k/). Each consonant appeared before three types of vowels: high, low, rounded. This created 15 categories of 3 words each for a total of 45 words in each of the two tests (pretest and posttest). Stimuli were recorded by a female NS of AE in a TV studio using a SONY Hi-8 videocamera and an Electrovoice lavaliere microphone. Recordings were digitized at 44.1 kHz and edited using AVID Media Composer version 5.51 for MacIntosh. The duration of each gate was two frames. Stimuli were presented using a 27" inch SONY Trinitron color TV monitor and SONY VHS video cassette recorder (SLV-920HF). Two warning tones signaled the presentation of a new stimulus, and one indicated the next gate of the same stimulus. Subjects were tested in small groups, and were instructed to write down the word they thought the talker was saying at each gate. They were given 4 seconds to respond. Experimental sessions were observed to ensure the subjects understood the instructions, and those in AV groups looked at the screen.

2.2 Results

In the analysis of responses, the identification point was determined by the gate at which the correct word was written with no subsequent changes [10]. This gate was converted to a percentage of the word's total duration in gates. For analysis purposes, tabulations for words not correctly identified were based on the total duration of the word plus 1 gate (producing percentages over 100). As shown in Table 1, words were identified significantly earlier in AV vs. A-only presentation, [$F(1,30.15)=25.97$, $p<.0001$] and in the posttest vs. pretest as a result of perceptual training [F(l,51.36)=259.1,p<.000l]. There was a significant interaction between modality and initial CV sequence [$F(14,392)=3.251,p<.0001$]. Test x CV was also significant [$F(14,34.59),p=.0002$]. Planned comparisons indicated that the identification of words beginning with /ɹ/ and /l/ showed greater improvement from pretest to posttest than other CV combinations [F(I,34.59)=21.17,p<.000l]. The interaction between modality, test and CV sequence was also significant. Planned comparisons revealed that as a result of training, there was a greater accentuation of the advantage of AV over A-only presentation in the identification of words beginning with /ɹ/ and /I/ compared to other CV sequences [$F(1,l4)=14.58,p=.002$]. The CV sequences showing the earliest identification were: bilabial + high vowel (e.g., *picnic*), /ɹ/ + low vowel (e.g., *rocket),* and nonlabials + high vowel (e.g., *ticket).* There was also a significant main effect of context [F(l,32.99)=28.01,p<.0001].

	Excised Words		Sentence Context	
	AV	A-only	AV	A-only
Pretest	100.5	107.6	88.9	99.9
Posttest	76.9	85.3	64.7	76.2

Table 1. Mean percentage of word needed for identification before and after training for Japanese AV and A-only groups in excised word and sentence context conditions.

3. EXPERIMENT TWO - KOREAN

A total of 32 Korean ESL learners, intermediate level, were given the same parallel gating tests. As with Experiment One, subjects who received AV perceptual training were given AV gating tests (pretest and posttest), and those who received A-only training were given A-only gating tests. These subjects were also divided further into two equal groups: one presented with the preceding sentence context plus the gated target word, and one with the excised gated word only.

3.1 Materials and Procedure

Stimuli and procedure were the same as those used in Experiment One.

3.2 Results

Analysis of the results was the same as in Experiment One. As shown in Table 2, words were identified significantly earlier in AV vs. A-only presentation [$F(1,28.93)=33.13,p<.0001$], and in the posttest vs. pretest as a result of perceptual training [F(1,48.86)=259.8,p<.0001]. There was a significant interaction between modality and initial CV sequence [$F(14,392)=4.824$, $p<.0001$]. Planned comparisons revealed that for words beginning with /ɹ/ and /1/, identification was earlier in the AV presentation compared to A-only [$F(1,14)=5.88,p<.05$]. Test x CV was also significant [$F(14,34.39)=2.09,p=.01$]. Planned comparisons indicated that identification of words beginning with /ɹ/ and /l/ showed greater improvement between pretest and posttest than those beginning with other CV sequences [$F(1,34.39)=16.08,p=.001$]. The CV patterns of earliest identification were similar to those of the Japanese. There was

also a significant main effect of context [$F(1,34.6)=15.59$, $p<.001$].

	Excised Words		Sentence Context	
	AV	A-only	AV	A-only
Pretest	96.3	104.4	84.4	100
Posttest	74.9	84.8	64.4	78

Table 2. Mean percentage of word needed for identification before and after training for Korean AV and A-only groups in excised word and sentence context conditions.

4. EXPERIMENT THREE - NSs

Eight NSs of English participated in each of the same four experimental group types as the NNSs.

4.1 Materials and Procedure

Two additional groups of 8 subjects each were tested with the same stimuli as follows: for one group, only the video signal of the excised word was presented; for the second, these words were preceded by the sentence context which was presented audiovisually so as to indicate the onset of the target word. This method assessed the visual discernibility of the target words and the contribution of context to their identification. (The number of NNSs was insufficient for this video-only testing). The within-subjects variable was the CV sequence. Between-subjects variables were modality and context.

4.2 Results

Results were tabulated as in the previous experiments. As shown in Table 3, words were identified significantly earlier in AV vs. A-only presentation [$F(1,30)=129.57, p<.0001$]. There was a significant Modality x CV interaction [$F(14,406)=5.00$, $p<.0001$]. The effect of the modality of presentation varied according to the initial CV sequence. Comparisons revealed that identification of words beginning with /ɹ/ and /l/ compared to the other CV sequences was more influenced by the modality of presentation [$F(1,406)=179.5, p<.001$]; that is, there was a significantly greater difference between the percentage of the word necessary for identification in the AV vs. A-only presentation.

Excised Words		Sentence Context	
AV	A-only	AV	A-only
82.3	93.2	62.9	81.6

Table 3. Mean percentage of word needed for identification for NSs comparing AV and A-only presentations in excised word and sentence context conditions.

A significant Modality x Context x CV interaction [$F(14,406)=2.65, p=.001$] further qualifies this comparison. Only with sentence context does the identification of words beginning with /ɹ/ and /l/ show a greater difference according to modality of presentation than other words. Additional comparisons indicated that identification of words beginning with /l/ was more influenced by the presence of visual input than those beginning with /ɹ/ [$F(1,406)=6.64, p=.01$].

In the video-only excised word condition, 33% of the words were identified. In a separate measure of visual distance (a comparison in terms of the homophenous or visually indistinct categories of the stimulus and response), following adjustment for variable word length, 74% of the visual categories were matched accurately.

5. DISCUSSION

The initial training experiment in this series demonstrated that for both Japanese and Korean ESL learners, AV perceptual training provided significantly greater improvement in the identification accuracy of /ɹ/, /l/, /f/ (vs. /p/), and /θ/ (vs. /s/) when compared to A-only. The well-documented differential effect of the position of /ɹ and /l/ in the word for the Japanese was also found (i.e., perceptual accuracy was lower in initial and intervocalic positions). In contrast to the abundance of research findings related to Japanese speakers learning English, relatively little previous work had been done with Korean learners. The study revealed their scores were generally lower for /ɹ/ and /l/ in final positions where the first language (L1) has an acoustically different (nonvelarized) /l/. Furthermore, results for both groups varied significantly in the perception of the liquids as a function of adjacent vowel and talker in training.

The significant effects produced by stimulus and talker variability in perceptual learning, and results of studies with NSs demonstrating an interdependence in the processing of words and voices [11] suggest a view of the mental lexicon that differs sharply from the traditional linguistic view of the storage of abstract canonical forms of words in memory. These findings are consistent with the view that the memory encoding of speech involves storage of individual episodes, preserving both contextual variability and the indexical properties of speech [12,13]. Both multiple-trace theory [13] and the stages of development in speech processing in the WRAPSA model [12] suggest a scenario for second-language (L2) speech development. In WRAPSA, speech input is assumed to undergo preliminary analysis by a set of auditory analyzers. The output of this analysis is weighted to give prominence to phonetic features that serve a contrastive function in the language environment. Further processing takes place on the weighted output. The sound pattern is extracted from this output and refines the description of the processed signal to probe memory where, following exemplar models, it is matched against traces of previously analyzed episodes which constitute the representation of the sound structure of words in the lexicon.

For L2 learners, a new weighting scheme must be developed for the L2 in order to shift attention from the settings that are optimal for the LI to those feature values distinctive in the L2. In the case of /ɹ/ and /l/, attention must be focused primarily on the F3 transitions that distinguish them. In this model, learning consists of copying the features of an experience onto a trace. The features that comprise the perceptual representations that probe memory depend upon the attention given to the auditory and visual attributes of the stimulus relevant to the particular task. Training with multiple exemplars plus feedback enhances the process. The objective of L2 perceptual learning, then, is to create a situation in which the response provided by an aggregate of L2 traces overshadows that of L1 traces. As a

record of an experience or episode, a memory trace is a composite of all properties to which one attends including modality-specific sensory features. The model leaves open the question of the locus of integration of information from multiple sources and different processing pathways. One possibility is that a single probe, as a perceptual representation, stems from the early processing, in the form of pattern extraction, of the weighted output of the initial analysis of information from one modality. If information from one modality is related to others top-down through feedback pathways from higher-level multisensory integration areas [14], then the trace may be the final representation of integrated information as a network of potential synaptic relations [15]. The more informative source would contribute the most information to stimulus identification.

From the perspective of multiple-trace theory, long-term memory is treated as a collection of episodic memory traces suggesting an episodic lexicon as the basis of word recognition processes, and perhaps also production [16]. This view is supported by the findings of the present series of experiments which have shown the success of perceptual training with adult L2 learners, and its transfer to both production and spoken word identification in connected speech.

6. REFERENCES

1. Hardison, D. M. "Bimodal Input in Second-Language Speech: Focus on /r/ and /l/". In J. Leather and A. James (Eds.), *New Sounds 97: Proceedings of the Third International Symposium on the Acquisition of Second-Language Speech,* 125-134, University of Klagenfurt, Austria, 1997.

2. Hardison, D. M. *Acquisition of Second-Language Speech: Effects of Visual Cues, Context and Talker Variability,* Diss. Indiana University, 1998.

3. Lively, S. E., Logan, J. S., and Pisoni, D. B. "Training Japanese Listeners to Identify English /r/ and /l/. II: The Role of Phonetic Environment and Talker Variability in Learning New Perceptual Categories", *J Acoustic. Soc. Amer., Vol. 94, 1993, pp. 1242-1255.*

4. Yu, K., and Jamieson, D. G. "Training of the English /r/ and /l/ Speech Contrasts in Korean Listeners", *Can. Acoustics, Vol. 21, 1993, pp. 107-108.*

5. Hardison, D. M. "Bimodal Speech Perception by Native and Nonnative Speakers of English: Factors Influencing the McGurk Effect," *Lang. Learn., Vol. 46, 1996, pp. 3-73.*

6. Koster, C. J., *Word Recognition in Foreign and Native Language: Effects of Context and Assimilation,* Foris Publications, Dordrecht, 1997.

7. Grosjean, F. "Spoken Word Recognition Processes and the Gating Paradigm", *Percept. Psychophys., Vol. 28, 1980, pp. 267-283.*

8. Walley, A. C., Michela, V. L., and Wood, D. R. "The Gating Paradigm: Effects of Presentation Format on Spoken Word Recognition by Children and Adults", *Percept. Psychophys., Vol. 57, 1995, pp. 343-351.*

9. Cotton, S., and Grosjean, F. "The Gating Paradigm: A Comparison of Successive and Individual Presentation Formats", *Percept. Psychophys., Vol. 35, 1984, pp. 41-48.*

10. Salasoo, A., and Pisoni, D. B. "Interaction of Knowledge Sources in Spoken Word Identification", *J. Mem. Lang., Vol 24,1985, pp. 210-231.*

11. Mullennix, J. W., and Pisoni, D. B. "Stimulus Variability and Processing Dependencies in Speech Perception", *Percept. Psychophys., Vol. 47, 1990, pp. 379-390.*

12. Jusczyk, P. W., *The Discovery of Spoken Language,* MIT Press, Cambridge, Mass., 1997.

13. Hintzman, D. L. "Schema Abstraction in a Multiple-Trace Memory Model", *Psych. Rev., Vol. 93, 1986, pp.411-428.*

14. de Sa, V., and Ballard, D. H. "Perceptual Learning from Cross-Modal Feedback". In R. L. Goldstone, P. G. Schyns, and D. L. Medin (Eds.), *The Psychology of Learning and Motivation* 36:309-351, Academic Press, San Diego, 1997.

15. Brown, J. W. "Overview". In A. B. Scheibel, and A. F. Wechsler (Eds.), *Neurobiology of Higher Cognitive Function,* 357-365, The Guilford Press, New York 1990.

16. Goldinger, S. D. "Words and Voices: Perception and Production in an Episodic Lexicon". In K. Johnson, and J. W. Mullennix (Eds.), *Talker Variability in Speech Processing,* 33-66, Academic Press, San Diego, 1997.