Perception & Psychophysics
2001, 63 (4), 746-758

Ability in perceiving nonnative contrasts: Performance on natural and synthetic speech stimuli

PETER C. GORDON, LISA KEYES, and YIU-FAI YUNG
University of North Carolina, Chapel Hill, North Carolina

The perception of the distinction between /r/ and /l/ by native speakers of American English and of Japanese was studied using natural and synthetic speech. The American subjects were all nearly perfect at recognizing the natural speech sounds, whereas there was substantial variation among the Japanese subjects in their accuracy of recognizing /r/ and /l/ except in syllable-final position. A logit model, which additively combined the acoustic information conveyed by *F*1-transition duration and by *F*3-onset frequency, provided a good fit to the perception of synthetic /r/ and /l/ by the American subjects. There was substantial variation among the Japanese subjects in whether the *F*1 and *F*3 cues had a significant effect on their classifications of the synthetic speech. This variation was related to variation in accuracy of recognizing natural /r/ and /l/, such that greater use of both the *F*1 cue and the *F*3 cue in classifying the synthetic speech sounds was positively related to accuracy in recognizing the natural sounds. However, multiple regression showed that use of the *F*1 cue did not account for significant variance in natural speech performance beyond that accounted for by the *F*3 cue, indicating that the *F*3 cue is more important than the *F*1 cue for Japanese speakers learning English. The relation between performance on natural and synthetic speech also provides external validation of the logit model by showing that it predicts performance outside of the domain of data to which it was fit.

A phonetic contrast that seems completely unmistakable to a speaker whose native language contains that contrast may seem completely unintelligible to an otherwise competent speaker whose native language does not contain that contrast. This striking phenomenon has provided the basis for theorizing about whether speech perception abilities are innate or acquired (e.g., Best, 1994; Strange, 1995; Werker, 1994) and for practical efforts to improve the abilities of nonnative speakers (e.g., Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Pisoni, Lively, & Logan, 1994). The present research examines the perception of both natural and synthetic speech in order to precisely measure the perceptual difficulty experienced by some native Japanese speakers in the perception of English /r/ and /l/. A psychophysical model is developed that characterizes the degree to which native and nonnative speakers base their phonetic classifications of synthetic speech stimuli on *F*3-onset frequency and on *F*1-transition duration and how they integrate information from these two cues. The ability of individual Japanese speakers to use these cues is shown to strongly predict their accuracy in identifying natural speech stimuli.

English /r/ and /l/

Acoustically, English /r/ and /l/ are distinguished by several acoustic features that vary depending on context (Dalston, 1975; Espy-Wilson, 1992; Olive, Greenwood, & Coleman, 1993). Studies on the perception of these segments have focused on the spectral and temporal cues that are strongly related to the /r/–/l/ distinction in syllable-initial position. The primary spectral cue is *F*3-onset frequency; it starts low for /r/ and subsequently rises for the following vowel, whereas it starts high for /l/ and remains constant or falls slightly in transition to the following vowel (O'Connor, Gerstman, Liberman, Delattre, & Cooper, 1957). *F*2 onset is similar to *F*3 in that it is lower for /r/ and higher for /l/. The primary temporal cue is *F*1-transition duration. For /r/, the initial *F*1 steady state is relatively short, and the *F*1 transition into the following vowel is relatively long. For /l/, the initial *F*1 steady state is relatively long, and the *F*1 transition is more abrupt. In contrast to English, Japanese has only a single sound classified as a liquid. This liquid is described as an apico-alveolar tap, or flap, that does not correspond to the alveolar retroflex /r/ or lateral /l/ found in English (Vance, 1987). The Japanese flap has an *F*3-onset frequency that can vary over the range of both /r/ and /l/ in English, though usually the *F*3 value is closer to that of English /r/ (Miyawaki et al., 1975).

Many studies have shown that Japanese speakers learning English have difficulty in perceiving the /r/–/l/ contrast (Logan, Lively, & Pisoni, 1991; Miyawaki et al., 1975; Sheldon & Strange, 1982; Yamada & Tohkura, 1991). Even Japanese speakers who are able to successfully produce /r/ and /l/ may be unable to perceive the contrast (Sheldon & Strange, 1982). Age of acquisition and experience influence the ability of Japanese speakers to perceive /r/ and /l/ (Best & Strange, 1992; Yamada, 1995).

The research reported here was supported by Grant IIS-9811129 from the National Science Foundation. Yiu Fai Yung is now at SAS Institute, Cary, NC 27513. Correspondence should be addressed to to P. C. Gordon, Department of Psychology, University of North Carolina, Chapel Hill, NC 27599-3270 (e-mail: pcg@email.unc.edu).

American English speakers perceive a synthetic /r/–/l/ continuum categorically, but many Japanese speakers do not (MacKain, Best, & Strange, 1981; Miyawaki et al., 1975; Yamada & Tohkura, 1991). This pattern is systematic with Japanese speakers who have had more experience with English, perceiving an /r/–/l/ continuum more categorically than less experienced speakers, but still not as categorically as Americans (MacKain et al., 1981). Japanese speakers also tend to hear some synthetic /r/–/l/ stimuli as /w/; when that judgment is taken into account, their perception of a synthetic /r/–/l/ continuum is more categorical (Yamada & Tohkura, 1991).

The role of the *F*3-onset frequency and *F*1-transition duration cues in categorization and discrimination has been studied within the trading-relations framework (Polka & Strange, 1985; Underbakke, Polka, Gottfried, & Strange, 1988). In this framework (Repp, 1983), the effect of differences on two acoustic dimensions is compared when those differences both have the same effect on the categorization of the two to-be-discriminated stimuli (e.g., making one more /r/-like and the other more /l/-like) and when they have opposite effects. The paradigm provides a way of determining whether discrimination occurs before or after categorization. Polka and Strange (1985) found that American speakers discriminate /r/ and /l/ after integrating information from *F*3-onset frequency and *F*1-transition durations. Underbakke et al. (1988) found that Japanese speakers who are skilled at /r/–/l/ perception also perform discriminations after phonetic integration, whereas less skilled Japanese speakers differ from Americans in their discrimination of /r/ and /l/.

Japanese speakers' ability to perceive /r/ and /l/ depends on position within a word (Pisoni et al., 1994; Sheldon & Strange, 1982). For Japanese speakers, discriminating /r/ and /l/ in consonant clusters causes the greatest difficulty, whereas discrimination of /r/ and /l/ in word-final position is very good, almost as good as it is for Americans. Word-initial position and medial intervocalic position are also difficult for native Japanese speakers. Response times to identify /r/ and /l/ are also slower for clusters and medial positions, whereas response times are fastest for word-final position (Pisoni et al., 1994). This pattern of response times further supports the conclusion that consonant clusters are the hardest environment for native Japanese speakers to perceive /r/ and /l/, whereas word-final position is the easiest.

Training in the laboratory is effective in helping native Japanese speakers perceive the /r/–/l/ distinction. To be effective, the training must employ tokens of natural /r/ and /l/ from several talkers (Lively, Logan, & Pisoni, 1993; Logan et al., 1991) rather than from synthetic speech (Strange & Dittmann, 1984). This improvement in the perception of /r/ and /l/ that results from training also leads to improvement in the ability to produce /r/ and /l/ (Bradlow et al., 1997). Such improvements due to laboratory training in the ability to perceive and produce /r/ and /l/ have been found to persist over several months (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999). The contrast between results of training with synthetic speech and results of training with multiple tokens of natural speech has supported the conclusion that training does not cause Japanese subjects to learn context-independent information, such as phonemes, but, instead, causes them to learn context-specific information about the properties of /r/ and /l/ (Pisoni et al., 1994). An alternative possibility is that synthetic speech, because it does not sound completely natural, does not fully tap the processing mechanisms that are used in categorizing natural speech.

The Combination of Acoustic Cues in Phonetic Perception

In experiments on the perception of synthetic speech, listeners are presented with a series of trials on which they are asked to classify a stimulus as belonging to one of a small number of linguistic categories; across trials, the acoustic composition of the stimuli is varied. The acoustic basis of the linguistic distinction is then studied using the proportion of responses as a function of the acoustic characteristics of the stimulus. Some researchers have interpreted the proportion classification directly in making inferences about perceptual processes (e.g., Miller, 1977; Polka & Strange, 1985). However, an important paper on modeling the perception of speech by Oden and Massaro (1978) showed that different acoustic features contribute independently to the identification of phonetic segments, despite apparent interactions of those cues when the response proportions are examined directly.

When modeling how acoustic cues contribute to the perception of the binary distinction that makes up a minimal linguistic contrast, response proportions emerge from placing a criterion on an S-shaped function of an underlying phonetic scale (Nearey, 1990). The S-shaped function may be either a cumulative normal distribution, if the model is based on signal detection theory (Green & Swets, 1966), or a logistic function, if the model is based on choice theory (Luce, 1959). For practical purposes in psychological research, these two types of functions usually yield equivalent fits (Macmillan & Creelman, 1991). Because the underlying phonetic scale has interval properties, it can be far easier to interpret than the observed response proportions. For example, the demonstration by Oden and Massaro (1978) that, in many cases, information from different acoustic cues contributes independently to phonetic classification results from those acoustic cues being additive on the underlying phonetic scale (McClelland, 1991).

The issue of whether different sources of evidence about the identity of phonetic segments are combined in an independent (additive) or interactive manner has been a source of empirical controversy.[1] Extensive work by Oden and Massaro (1978; Massaro, 1987; Massaro & Oden, 1980) has found evidence that different acoustic cues are combined independently across a wide range of types of information (i.e., acoustic cues, speaking rate, and lexical identity) to many phonetic categories. In contrast, Pitt (1995b) has presented evidence that lexical evidence and acoustic evidence combine in an interactive manner. The question of who is correct depends in part on

the technique that is used for parameter estimation in fitting independent and interactive models (Massaro & Oden, 1995; Pitt, 1995a). Two additional points are relevant. The first is that there need not be a general answer: Combination may be independent in some cases and interactive in others. The second has to do with the particular goals in assessing the two types of models: Independent combination may account for most of the variance in classification patterns, whereas interactive combination accounts for relatively little. The question of whether interactive combination contributes little to phonetic classification or whether it contributes nothing at all is important if the goal of the work is to falsify a general class of models. It is less important if the goal is to use classification performance to predict other aspects of perception.

The Present Study

In this experiment, we examined the identification of natural and synthetic speech by native speakers of American English and native speakers of Japanese who became immersed in English through residence in the United States only after adolescence. The experiment addressed three goals. The first was to evaluate models for how information from *F*3-onset frequency and *F*1-transition duration is used by American English speakers in identifying synthetic speech stimuli; the second was to evaluate models for how native Japanese speakers use information about *F*3-onset frequency and *F*1-transition duration in identifying synthetic speech stimuli. Progress on these two goals provides evidence about the applicability of independent (additive) and interactive models of cue combination in the perception of the contrast between /r/ and /l/. The third goal was to examine whether variation in ability among native Japanese speakers in correctly identifying naturally produced tokens of /r/ and /l/ is related to their use of *F*3-onset frequency and *F*1-transition duration in identifying synthetic speech stimuli. Success in this goal provides a way of validating models of cue combination in synthetic speech in an independent domain. It also provides evidence about the perceptual basis of phonetic classification of nonnative contrasts and about the validity of using synthetic speech stimuli to study perception of nonnative contrasts.

METHOD

Subjects

Twelve Americans and 12 Japanese were recruited from the University of North Carolina community for participation in the experiment. The American subjects were recruited with posted notices; the Japanese subjects were recruited through contacts in the community and in an ESL class. The subjects were paid $5 for their participation in the experiment.

Responses to a language-experience questionnaire, given to the Japanese subjects, are shown in Table 1. As can be seen, there was substantial variation among the subjects in English language experience, but none of the subjects had experience using English conversationally prior to 14 years of age.

Natural Speech Stimuli

The stimuli consisted of four minimal pairs contrasting /r/ and /l/ in four different positions within the word, for a total of 16 minimal pairs. These were the same minimal pairs used in Sheldon and Strange (1982) and Strange and Dittmann (1984), with the exception of one new pair, *car–call*. It replaced a pair in Sheldon and Strange's study, *war–wall*, which is not actually a minimal pair in some dialects, including the dialects of the speakers in this experiment. The minimal pairs contrasted /r/ and /l/ in initial prevocalic, consonant cluster, medial intervocalic, and final postvocalic positions. Eight other minimal pairs not contrasting /r/ and /l/ were used as fillers. The minimal pairs are listed in Table 2.

Four speakers of American English, 2 males and 2 females, pronounced the words. All 4 speakers were graduate students at the University of North Carolina. Each speaker read each word three times, and the experimenter chose the clearest sounding token of each word from each speaker for use in the experiment. Each speaker produced 32 test words plus 16 fillers, for 48 tokens in all. The

Table 1
Characteristics of the Japanese Subjects and Some Factors That Indicate Their Level of Experience With Spoken English

Subject	Age (years)	Sex	% of Day English Spoken	Age Moved to U.S. (years old)	Total Months in U.S.	Months of Conversation Experience
1	22	F	75	22	7	24
2	21	F	25	21	8	8
3	54	F	25	53	6	120
4	26	F	90	24	18	72
5	21	F	75	20	8	8
6	23	M	100	19	48	60
7	32	F	75	29	43	43
8	24	M	5	24	8	8
9	19	F	50	14	60	60
10	19	M	50	19	1	1
11	22	M	25	21	8	20
12	22	M	75	22	8	32
Average	25.4		55.8	24.0	18.6	38.0
SD	9.6		30.3	9.8	19.9	35.0

Note—The subjects are ordered from best performing to worst performing in terms of their performance in identifying the natural /r/–/l/ stimuli in nonfinal positions.

Table 2
The Test Words and Filler Words Used in the Natural-Speech Portion of the Experiment

Test Words			
Initial	Consonant Cluster	Medial	Final
read–lead	breed–bleed	mirror–miller	dear–deal
room–loom	broom–bloom	berry–belly	core–coal
road–load	grow–glow	correct–collect	car–call
right–light	grass–glass	arrive–alive	tire–tile
Filler Words			
Initial	Vowel	Medial	Final
deep–keep	boat–boot	swimming–swinging	him–hip
hope–soap	get–got	defend–descend	mad–man

Note—The vowel for the read–lead pair was /i/.

speakers were given no special instructions on how they should pronounce the words. The utterances were recorded using a Shure SM59 microphone and Sony DAT recorder (Model 670) onto digital audio tape. The stimuli were subsequently digitized at a 10-kHz sampling rate using 16-bit samples on a Kay Elemetrics CSL system. The test words and fillers were arranged into four blocks and were recorded onto digital audio tape. Each block contained one token of each test word, for a total of 32 test words per block. Sixteen fillers were also placed in each block. Within each word position, one speaker produced both members of a minimal pair, and each of the 4 speakers produced one minimal pair in each word position in an individual block. The speakers then rotated minimal pairs throughout the blocks. Within each block, the 48 words (test words and fillers) were randomized.

Synthetic Stimuli

A /ra/–/la/ series, based on the one described by Polka and Strange (1985), was created with a Klatt synthesizer (Sensimetrics version). The starting frequency of *F*3 and *F*2 and the duration of the *F*1 transition were varied. Both *F*3-onset frequency and *F*1 transition were varied in seven equal steps, as shown in Figure 1. *F*2 was varied concurrently with *F*3. *F*3-onset frequency ranged from 1477 to 2594 Hz and was varied in seven almost equal steps of 186 or 187 Hz. *F*2 onset varied from 1067 to 1207 Hz and was varied in seven steps of 23 or 24 Hz. *F*1 steady state was also varied in seven almost equal steps. *F*1 steady state ranged from 10 to 56 msec and was increased in roughly 8-msec steps, resulting in an *F*1-transition duration that ranged from 55 to 9 msec.

*F*3 onset and *F*1 transition were varied independently of each other to produce a total of 49 stimuli. These 49 stimuli were randomized and presented once in each of 10 blocks. Thus, each subject heard each stimulus item 10 times through the course of the experiment. The first block of 49 trials was considered a warm-up and was not analyzed in the model fitting.

Procedure

The Japanese subjects first completed the language-experience questionnaire, followed by identification of the natural speech stimuli, and then identification of the synthetic speech stimuli. The procedure was the same for the American subjects, except that they did not fill out the language-experience questionnaire. The subjects were tested individually in a quiet room. The entire session took approximately 1 h.

The natural stimuli were presented at a comfortable listening level over Sony MDR 7506 headphones. For each trial, the preprinted response sheet showed the stimulus word and its mate in the minimal pair. The word containing /r/ was always printed first. The subjects were instructed to circle the word on the response sheet that corresponded to the word they heard on the tape. In the synthetic speech section of the experiment, the subjects were told that they would hear a syllable, either *ra* or *la*, and to circle "R" or "L" on the answer sheet corresponding to which sound they heard on the tape. The subjects were instructed to guess if necessary.

Modeling the Data

Logit models. In order to provide a quantitative account of perception of the synthetic speech sounds, logit models were fit to the

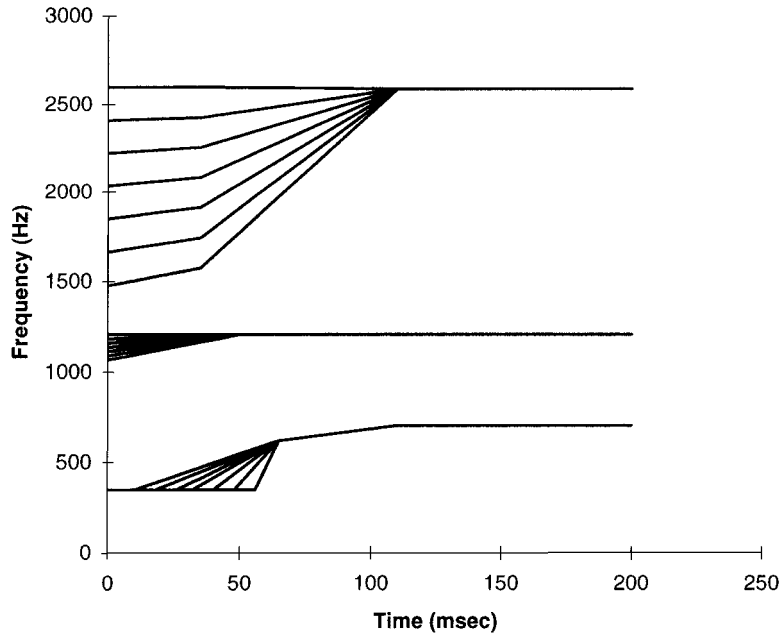

Figure 1. Schematic representation of the formant patterns of the synthetic /ra/–/la/ series. The frequencies of *F*2 and *F*3 and the temporal characteristics of *F*1 were varied in seven steps. Step 1 is the most /r/-like, and Step 7 is the most /l/-like.

data for each subject. Modeling began with the following basic form of the logit model (Wickens, 1989):

$$\text{Prob}\{/\text{r}/|S_{ij}\} = \frac{1}{1 + e^{-(a+g_i+h_j)}} \quad (i = 1,2, \ldots 7; j = 1,2, \ldots 7), \quad (1)$$

where g_i and h_j represent the effects of $F1$ and $F3$, respectively, on the probability of responding /r/. Notice that Equation 1 is the usual logistic regression model, but without the imposition that levels of $F1$ and $F3$ have a linear effect on the logit transform of probability of responding /r/. This allows for the possibility that the perceptual impact of each level difference is not exactly equivalent. If this model does not fit for a particular individual, then it implies that interaction terms may be needed in the model. However, previous results from Oden and Massaro (1978) suggest that a logit model without interaction terms may be sufficient to account for the data. If the logit model described in Equation 1 does provide a good fit of the data, then we may use the chi-square test for individual effects to examine whether individual component effects, $F1$ and $F3$, have significant contributions to the model.

To get a better understanding of Equation 1, one can rewrite it into the following equivalent form:

$$\text{logit}\{/\text{r}/|S_{ij}\} = a + g_i + h_j, \quad (2)$$

where

$$\text{logit}(\text{B}) = \log\left(\frac{\text{Prob}(\text{B})}{1-\text{Prob}(\text{B})}\right)$$

is just a monotone transformation of $\text{Prob}\{/\text{r}/|S_{ij}\}$. Accordingly, the model in Equation 1 is very much like the usual analysis of variance (ANOVA) model, except with transformed responses for $\text{Prob}\{/\text{r}/|S_{ij}\}$. However, model-fitting techniques for Equations 1 and 2 are very different from those for ANOVA. In our modeling, we will use the weighted least square (WLS) technique.[2]

If all individuals within a country could be described by Equation 2 (or Equation 1), it is tempting to fit a single model containing all individuals. A preliminary group model for this purpose is

$$\text{logit}\{/\text{r}/|S_{kij}\} = b + b_k + g_i + h_j, \quad (3)$$

where b_k is an individual parameter representing individual biases in responses (as compared with the entire group).

Alternative models. The additive logit model, described above, provides a basic accounting of the pattern of responses that is easily interpretable. However, it may not be the most precise model possible for the responses of every individual. Our modeling strategy included investigation of whether simpler models can explain the phenomena equally well. Figure 2 shows a hierarchy of possible models that could be fit to the data (Wickens, 1989). The topmost model is the most restrictive "intercept model," and the bottommost is the saturated model in which interaction effects and main effects are all included. The intercept model implies that the stimulus characteristics have no effect on classification and that a subject's responses are determined purely by response bias and chance. In the "direct" models, the levels of a stimulus feature ($F1$ or $F3$ in this case) have an impact on classification, with the difference between adjacent levels of a feature being a constant given by the slope parameter. When both features are treated in this way, then the resulting model is standard logistic regression. In the nominal models, the levels of a stimulus have an impact on classification, but, in contrast to the direct model, the difference between levels of a feature is not fixed but instead is determined by a set of parameters that are free to vary. When both features are treated in this way, the resulting model is the additive logit model. In the fully saturated model, interaction terms are added to the logit model. Essentially, this model amounts to a reparameterization of the observed data and is equivalent to a model in which there is a unique template for every stimulus.

Good fit of the data by the logit model, which is located just above the saturated model in Figure 2, indicates that no interaction for $F1$ and $F3$ need to be considered. This is shown statistically by the results of the chi-square goodness-of-f it tests, which, in essence, compare the logit model against the saturated model. The question of whether simpler models, located above the logit model, fit the data presents a problem of model search.

To test the relative fit of nested models (those connected by arrows in Figure 2), we employed the chi-square difference test defined by

$$\chi^2_{\text{diff}} = (\chi^2_0 - \chi^2_1)/(df\,0 - df\,1),$$

where χ^2_0 and $df\,0$ are the chi-square statistic and the degrees of freedom for the more restrictive model in question, and χ^2_1 and $df\,1$ are those of the less restrictive model in question. To gain support for the more restrictive model, χ^2_{diff} must be insignif icant at some α level with degrees of freedom ($df\,0 - df\,1$); otherwise, the less restrictive model is supported. Although this is a hypothesis-testing framework and the sampling distribution of the χ^2_{diff} in model searching may not follow the prescribed distribution exactly, it still is useful as an exploratory tool for finding the "best" model.

RESULTS

Natural Speech

The mean percent correct for identifying the natural speech tokens is given in Table 3 for the American and Japanese subjects as a function of position in the word for minimal pairs contrasting /r/ and /l/ and for filler pairs that did not contrast /r/ and /l/. An ANOVA was performed on the arcsine transform of the percent correct score for each subject for each type of minimal pair. There was a significant main effect of country [$F(1,22) = 25.3$, $p < .001$], and there was a significant interaction of country and type of minimal pair [$F(4,88) = 15.4$, $p < .001$]. Contrasts, adjusted with the Bonferroni method, showed that the Japanese subjects did more poorly than the American subjects for minimal contrasts of /r/ and /l/ at initial position [$t(22) = 5.72$, $p < .001$], cluster position [$t(22) = 5.99$, $p < .001$], and medial position [$t(22) = 8.80$, $p < .001$]. A marginally significant effect of the /r/–/l/ minimal contrast in final position [$t(22) = 2.65$, $p > .05$] was observed, though the difference in errors was small (100% accurate for the American subjects vs. 98.2% accurate for the Japanese subjects).[3] No significant difference was observed for filler stimuli that did not have a minimal contrast of /r/ and /l/ [$t(22) = 1.41$, $p > .15$].

The task was trivially easy for the American subjects, who made only 3 errors out of a total of 1,536 trials involving identification of pairs contrasting /r/ and /l/. The ease of the task for the American subjects indicates that the natural speech sounds were articulated clearly enough and that the quality of the recordings was high enough to support nearly perfect recognition by competent speakers of English. Relative to the American subjects, the Japanese subjects had substantial difficulty with pairs contrasting /r/ and /l/ when they were in initial, cluster, or medial position, but they had only very slightly, and nonsignificantly, greater difficulty for this distinction in final position. This pattern of results is consistent with earlier findings showing that Japanese speakers are much better able to perceive the /r/–/l/ contrast in word-final position than in other positions (Logan et al., 1991; Sheldon &

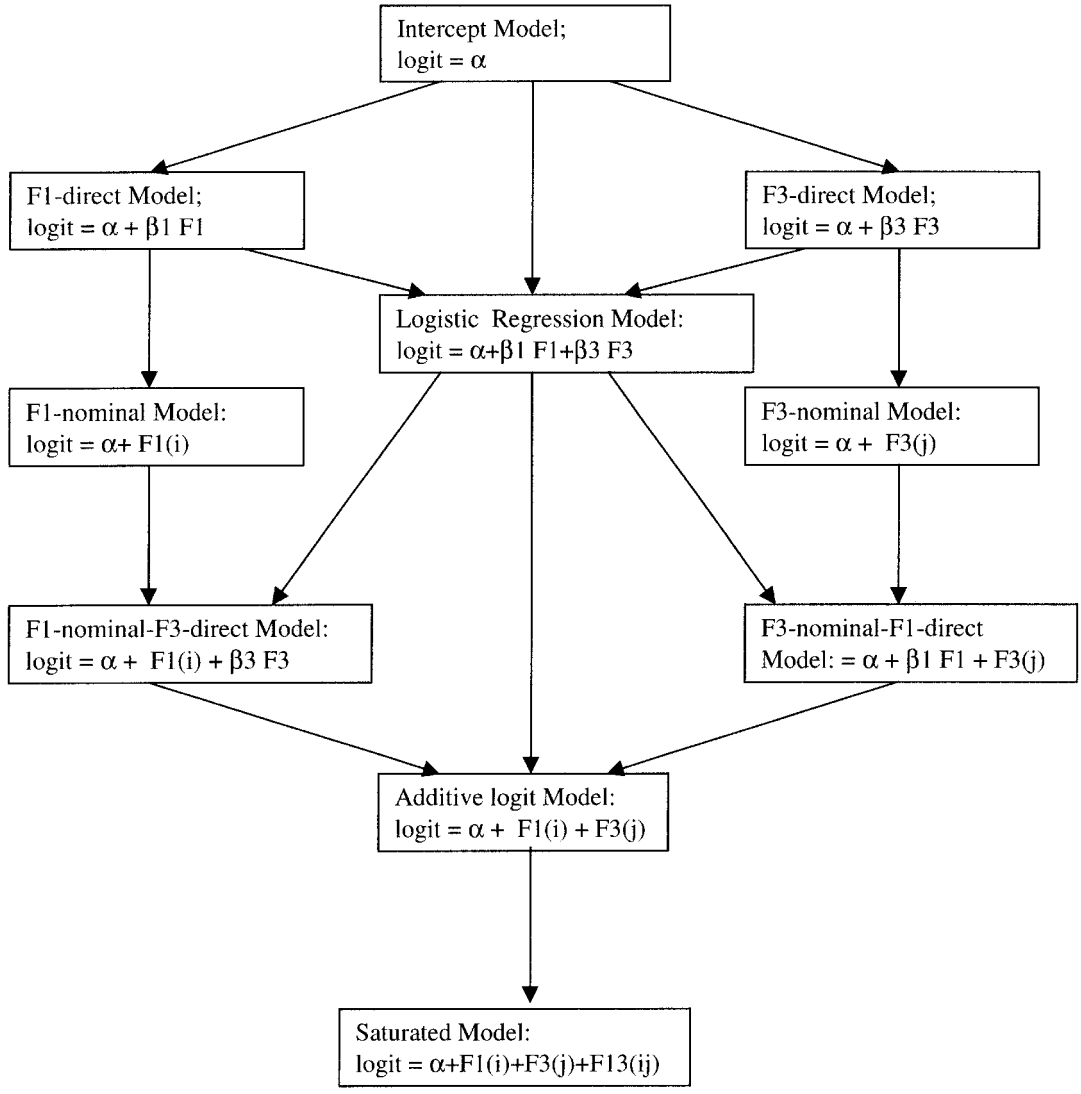

Figure 2. Hierarchy of possible models for the combination of *F*1 and *F*3 in the recognition of /r/ and /l/.

Strange, 1982; Strange & Dittmann, 1984). For present purposes, the ability of the Japanese speakers to recognize /r/ and /l/ in word-final position (as well as their ability to recognize the filler pairs) indicates that the difficulty that they exhibited in recognizing /r/ and /l/ in other positions was not due to a general difficulty with the task, a general lack of ability in recognizing spoken English, or an inability to recognize printed "R" and "L" on the response sheets.

Table 3
Average Percent Correct on Identification of Natural-Speech Tokens by American and Japanese Subjects

Position of /r/ Versus /l/	Americans	Japanese
Initial	99.7	79.7
Cluster	99.7	70.3
Medial	99.7	75.3
Final	100.0	98.2
Filler	97.8	97.1

Synthetic Speech Perception

The proportions of /r/ responses as a function of *F*3-onset frequency and *F*1-transition duration are shown in Figure 3 for the American subjects and in Figure 4 for the Japanese subjects. Examination of the data for the American subjects shows that their classifications were strongly influenced by *F*3-onset frequency (as shown by the decline in proportion /r/ responses from left to right) and were also strongly influenced by *F*1-transition duration (as shown by the separation of the lines indicating the different *F*1-transition durations). Examination of the data for the Japanese subjects shows that both of these stimulus dimensions influenced their judgments as well but that neither factor had as strong an influence as it did on the American subjects.

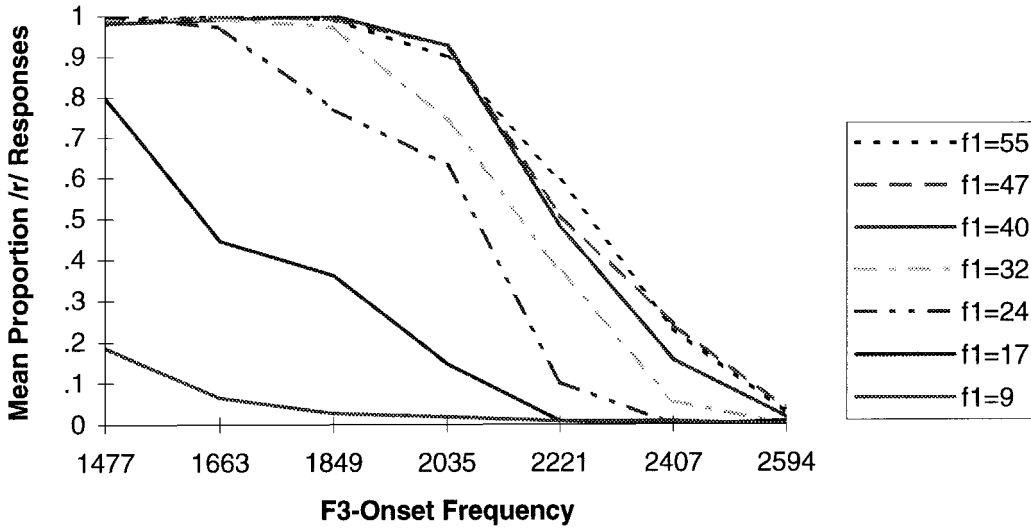

Figure 3. Effect of *F*1-transition duration and *F*3-onset frequency on the perception of the /r/–/l/ distinction by native English speakers.

Logit Models for American Subjects

Table 4A shows the results of assessing the fit of the model in Equation 1 for each American subject using the chi-square statistic, with a nonsignificant chi-square indicating a good fit. For all 12 American subjects, fitting the model described in Equation 1 produced nonsignificant chi-square values. This indicates that additive combination of the information in *F*1 and *F*3 is sufficient and that in no case is there a need to add interaction terms for a better account of the data. However, having a model that fits well simply means that no important effects have been omitted from the model. It does not exclude the possibility that unimportant effects exist in the model. To investigate this possibility, we examined the significance of the *F*1 effects and the *F*3 effects for each subject. As shown in Table 4A, both *F*1 and *F*3 effects were significant (all *p* values less than .01) for every American subject. Table 5A shows the effect sizes of *F*1 and *F*3.[4] For every subject, the effect size of *F*3 was greater than that of *F*1. This is consistent with the belief that *F*3 is a more powerful cue to the distinction between /r/ and /l/ than is *F*1 for the range of stimulus values employed in the experiment, a range that was chosen to cover the range that appears naturally in English (Polka & Strange, 1985).

A group model, as described in Equation 3, was then fit to the data of all the American subjects. Again, the chi-

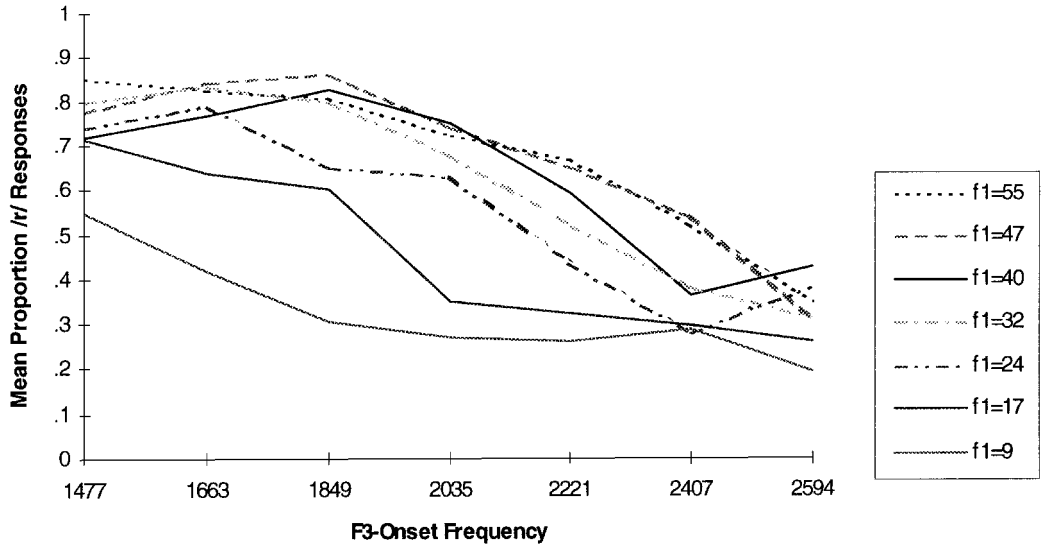

Figure 4. Effect of *F*1-transition duration and *F*3-onset frequency on the perception of the /r/–/l/ distinction by native Japanese speakers.

Table 4A
The *p* Values for Likelihood Ratio Chi-Square Model Fit Statistics and for Significance Tests of Effects for the American Subjects

Test	Subject 1	2	3	4	5	6	7	8	9	10	11	12	Group Model
Chi-square ($df = 36$)	.32	.77	.88	.68	.87	.35	.66	.84	.59	.65	.80	.86	.99 ($df = 564$)
$F1$ ($df = 6$)	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
$F3$ ($df = 6$)	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Note—Values of .00 indicate p values of less than .005.

Table 4B
The *p* Values for Likelihood Ratio Chi-Square Model Fit Statistics and for Significance Tests of Effects for the Japanese Subjects

Test	Subject 1	2	3	4	5	6	7	8	9	10	11	12	Group Model	Skilled-Group Model
Chi-square ($df = 36$)	.87	.50	1.00	.64	.83	.97	.77	.22	.98	.77	.58	.32	0.00($df = 564$)	.82 ($df = 324$)
$F1$($df = 6$)	.00	.00	.00	.19	.00	.00	.45	.00	.00	.00	.89	.68	.00	.00
$F3$ ($df = 6$)	.00	.00	.02	.00	.00	.00	.00	.00	.00	.56	.75	.33	.00	.00

Note—Values of .00 indicate p values of less than .005.

square statistic for model fit was not significant (.9981), indicating a very good account of the model for the entire set of data. Both $F1$ and $F3$ effects were significant in the group model ($p < .01$). The estimates for the $F1$ and $F3$ effects are shown in left and right panels of Figure 5. The monotone increasing trends for the estimates derive from the increase in probability of responding /r/ as $F1$-transition duration increases and as $F3$-onset frequency decreases (see Figures 3 and 4).

Logit Models for Japanese Subjects

Table 4B shows the results of assessing the model in Equation 1 for the data from each Japanese subject. As with the results for the American subjects, a nonsignificant chi-square statistic was found for every Japanese subject, indicating that the model fit well and that there was no need to consider interaction terms. In contrast with the results for the American subjects, not every Japanese subject showed significant effects for the $F1$ and $F3$ cues. These nonsignificant effects indicate that those subjects failed to use the information in the cue for purposes of classifying the synthetic speech sounds. Table 5B shows the effect sizes of $F1$ and $F3$ for each Japanese subject; the variation it reveals among individual subjects will be addressed in the next section of the paper.

A group model, as described in Equation 3, was fit for all Japanese subjects. The chi-square statistic for model fit was highly significant, indicating that the model did not provide a good account of the entire data set. This is not surprising given the substantial heterogeneity in the performance of the individual Japanese subjects. A group of Japanese subjects who were skilled at classifying the synthetic /r/ and /l/ was identified as the 7 Japanese subjects who showed significant effects of both $F1$ and $F3$. When a group model using Equation 3 was fit to this skilled group, there was a nonsignificant chi-square statistic, indicating that the model did provide a good account of the data for those skilled subjects. The parameter estimates for $F1$ and $F3$ for the skilled group of Japanese subjects are shown in Figure 5. The monotone increasing

Table 5A
Effect Size Estimates for *F*1 and *F*3 for the American Subjects

	Subject 1	2	3	4	5	6	7	8	9	10	11	12	Group Model ($N = 12$)
$F1$	1.63	1.42	1.56	1.29	1.58	1.42	1.61	1.50	1.37	1.50	1.45	1.41	1.48
$F3$	1.89	2.07	1.84	2.06	1.76	2.06	1.72	1.99	2.24	1.94	2.10	1.70	1.95

Table 5B
Effect Size Estimates for *F*1 and *F*3 for the Japanese Subjects

	Subject 1	2	3	4	5	6	7	8	9	10	11	12	Group Model ($N = 12$)	Skilled-Group Model ($N = 7$)
Natural-Speech Performance (%)	100	100	91	88	88	88	81	75	72	66	56	53		
$F1$	1.37	1.02	1.12	**0.36**	0.96	0.69	**0.28**	0.68	1.06	0.66	**0.17**	**0.22**	0.72	0.98
$F3$	1.92	1.66	0.49	1.09	0.99	0.98	0.94	0.59	0.87	**0.26**	**0.21**	**0.30**	0.86	1.07

Note—Effect size estimates associated with nonsignificant effects are in boldface.

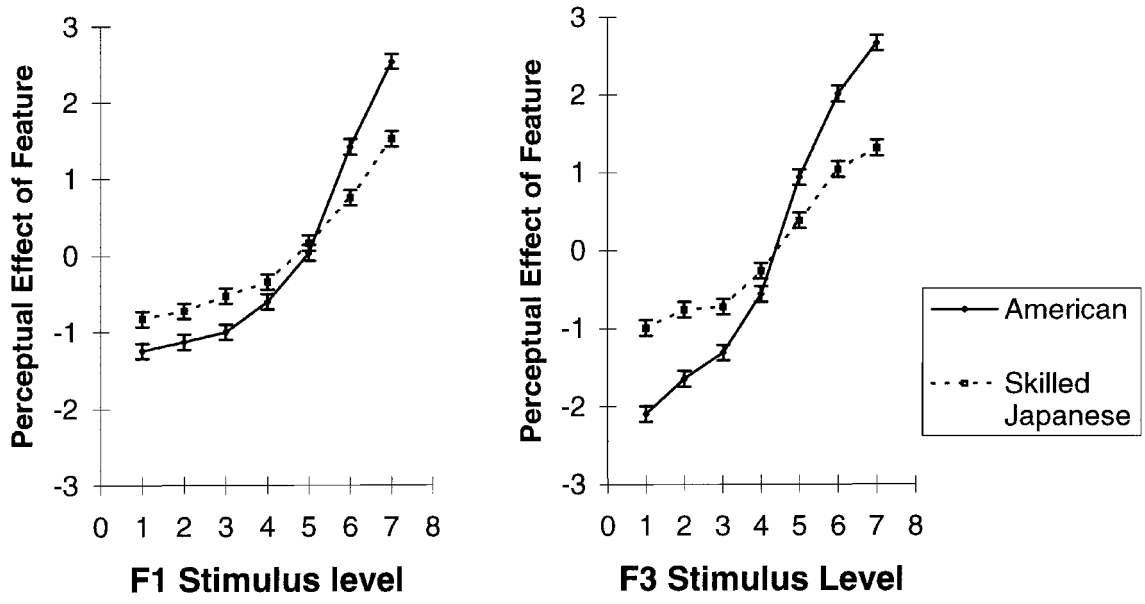

Figure 5. The perceptual effect of the features, associated with the different stimulus levels, as given by the parameter values from Equation 3 for American and skilled Japanese subjects on the *F*1 cue (left panel) and the *F*3 cue (right panel). The error bars show the standard error of the parameter estimate from the group model.

trends of the estimates are similar to those observed for the American subjects. However, the magnitudes of the increases for both *F*1 and *F*3 are substantially smaller for the skilled Japanese subjects than for the American subjects. This indicates that the American subjects made finer discriminations among levels of the *F*1 and *F*3 cues than did the skilled Japanese subjects.

Alternative Models

All the models in the Figure 2 were fit for each subject. Then, chi-square difference tests were applied for each pair of adjacent models. Using $\alpha = .05$, the most restrictive (simplest) model was found for each subject. The simplest model met the following three criteria:

1. The model itself had a nonsignificant χ^2 value.
2. The χ^2_{diff} of the model from comparison to the adjacent less restrictive model was not significant. This means that the model fit as well as the next less restrictive model, so that the more complicated model was not needed.
3. The χ^2_{diff} of the model (except for the intercept model) from comparison to the more restrictive model immediately above was significant, which means that the more restrictive model did not fit as well, so that the model found was necessary for a good account of the data.

The results of model search were are largely consistent with those of the earlier tests of the logit model for all subjects. For all American subjects, except for Subject 7, the simplest model that fit well was the logistic model, in which *F*1 and *F*3 effects are linearly related to the logit (which is a monotone transform of the probability responding /r/) responses. For Subject 7, results were fit with either a model with linear *F*1 effect and nominal *F*3 effect or a model with linear *F*3 and nominal *F*1.

For the Japanese subjects, the final models (shown in Table 6) were quite varied. The 2 subjects with the worst performance were best fit by the model with a single response tendency parameter (i.e., the intercept); they did not show discrimination in either *F*1 or *F*3. There was also a tendency for the subjects with much better performance to have discrimination in both *F*1 and *F*3. However, it also seems that many of the Japanese subjects lacked at least one of the *F*1 and *F*3 discriminations, even though some of them still attained a high performance level in natural speech.

A cautionary note about the "simplest" model obtained is in order. The logistic models found for most of the American subjects are subject to the ranges of the *F*1 and *F*3 levels studied. If these levels were extended to larger ranges, one could anticipate that the *F*1 and *F*3 effects on the logit responses would no longer be linear. There-

Table 6
Type of Best-Fitting Model for the Japanese Subjects

Subject											
1	2	3	4	5	6	7	8	9	10	11	12
logistic	logistic	*F*1-direct	*F*3-direct	logistic	*F*3-direct	*F*3-direct	logistic	logistic	*F*1-direct	intercept	intercept

fore, we emphasize here that the logistic model, even though more precise, may not apply to other situations. It is safe, however, to start with the additive logit response model, as done here, as a basic model for accounting for the response pattern. It is quite clear that the present results provide no support for the inclusion of an interaction term in fitting these data.

The finding of no need for an interaction term in modeling these data suggests that the perception of the difference between /r/ and /l/, by both native and nonnative speakers of English, can be explained by perceptual models in which information from different cues is combined in an independent (or additive) fashion (Massaro & Oden, 1995; Oden & Massaro, 1978). The modeling provides no support for perceptual models in which information is combined in an interactive fashion (e.g., Pitt, 1995a, 1995b), though, as we argued in the introduction, there is no necessity that information combination be exclusively independent or exclusively interactive. The present finding does have implications for the characterization of perception of /r/ and /l/ that has been advanced in trading-relations studies of *F*3-onset frequency and *F*1-transition duration as cues to /r/ and /l/. Polka and Strange (1985, p. 1194) argued that the dependence of discrimination results on the consistency with which *F*3 and *F*1 cues supported /r/ versus /l/ percepts indicated that the two cues are perceived in relation to each other. The finding here of support for an additive model of combining information indicates that perceiving the two cues in relation to each other is unlikely, though discrimination of tokens of phonetic segments may very well rely on the distinguishing phonetic categories that are perceived on the basis of additive cue combination.

Perception of Natural and Synthetic Speech for Japanese Subjects

Our model fits show that there is substantial variation among Japanese subjects in their use of the *F*1 and *F*3 cues for purposes of classifying the synthetic speech sounds: Some subjects use both cues, some use *F*3 but not *F*1, and some use *F*1 but not *F*3, and some use neither cue. In addition to showing effect sizes for the *F*1 and *F*3, Table 5B shows the individual subjects' accuracy in identifying natural speech /r/ and /l/ in nonfinal position. For ease of observing relations between the perception of natural and synthetic speech, the results for the subjects have been ordered from best performance to worst performance on the natural-speech measure. The results show that the three best performing subjects showed significant effects of both *F*1 and *F*3 in perception of synthetic speech. In contrast, the two worst performing subjects did not show significant effects of either *F*1 or *F*3. The subjects who lacked significant effects of only one cue were scattered through the middle range of performance on the natural-speech task, though use of *F*3 was more important than use of *F*1, as indicated by the fact that the 3 subjects who lacked *F*3 were the 3 worst performing subjects, whereas 1 of the subjects who lacked *F*1 was among the top half on performance of the natural-speech task.

A quantitative measure of the relation between use of the synthetic speech cues and accuracy in perceiving the natural-speech tokens is provided by the correlation between effect sizes for the synthetic dimensions and accuracy in natural-speech recognition. The correlation between *F*1 effect size and natural-speech accuracy is .699 ($p < .025$), and the correlation between *F*3 effect size and natural-speech accuracy is .838 ($p < .002$). Stepwise multiple regression showed that *F*3 effect size enters the model first and plays a significant role in predicting natural-speech accuracy. In contrast, *F*1 effect size does not enter into the regression model, indicating that *F*1 effect size does not account for unique variance in natural-speech accuracy beyond what is accounted for by *F*3. This pattern suggests that ability to use *F*3-onset frequency is more important than ability to use *F*1-transition duration for Japanese speakers perceiving the /r/–/l/ distinction. Our use of the logit (or the logistic) variant of the independent-cue model provides parameter estimates with scale properties that allow the use of techniques like linear regression. Crowther, Batchelder, and Hu (1995; see also McClelland, 1991, and Oden, 1979) have shown that the multiplicative form of the independent cue model (Oden & Massaro, 1978) yields nonunique parameter estimates due to a scaling indeterminancy. In the logit form of the model, this indeterminancy appears in the absolute magnitude of the bias and cue-derived parameters. For example in Equation 2, a constant could be added to the bias parameter and subtracted from all the *F*3 parameter estimates, yielding a model with identical fit to those reported above. However, the differences between the parameters for the different levels of *F*3 and *F*1 cues remain constant in such a transformation. It is these differences between levels of the parameters for *F*3 and *F*1 cues that are used in the regression analyses. Accordingly, use of the logit form of the independent-cue models allows the parameter estimates to be used in linear models.

Yamada and Tohkura (1992) observed a strong correlation between accuracy in perceiving natural /r/–/l/ contrasts and the perception of a synthetic /r/–/l/ continuum. They did so by correlating percent correct on natural /r/–/l/ with average consistency in classifying individual synthetic tokens. The present analysis provides further insight into that correlation by relating natural-speech performance to measures derived from an explicit model of how acoustic cues contribute to the classification of synthetic sounds. Doing so shows that use of *F*3-onset frequency predicts performance on natural speech. The finding of a strong relationship between the perceptual use of specific cues in synthetic speech and the accuracy of perceiving natural speech suggests that the different levels of success that have been observed in using synthetic speech (Strange & Dittmann, 1984) and natural speech (Pisoni et al., 1994) to train Japanese listeners to perceive the /r/–/l/ distinction does not occur because synthetic and natural speech tap different perceptual

mechanisms. The finding also suggests that the changes in perceptual processing due to training with natural speech could be investigated with synthetic speech.

DISCUSSION

The results of our analyses of the perception of synthetic and natural /r/ and /l/ yield progress toward the three goals of the experiment: (1) to evaluate models of the perception of synthetic /r/ and /l/ by American English speakers, (2) to evaluate such models for Japanese speakers, and (3) to determine whether variation among Japanese speakers in the ability to recognize natural /r/ and /l/ is related to their ability to encode specific acoustic cues in synthetic /r/ and /l/.

A logit model, in which the acoustic information in $F1$ and $F3$ is combined additively, provides an excellent account for the perception of synthetic /r/ and /l/ by American subjects. The model successfully fit data from each subject and also successfully fit group data from all the subjects. This supports the basic claim of Oden and Massaro (1978) concerning the independent contributions of different acoustic dimensions to the classification of speech sounds. In Oden and Massaro's analysis, information from acoustic cues is assessed independently and then combined. In the model we developed, information from the two independent acoustic cues is combined additively, because, in a logit-linear model, products are transformed into sums. The parameter estimates obtained in this way are unique to the level of an additive constant and therefore can be related by linear models to other measures, an advantage not shared by the multiplicative model of Oden and Massaro (1978; cf. Crowther et al., 1995).

No evidence was found in our analyses for interactive interpretation of the basic stimulus features. The questions of whether and when perception involves additive or interactive interpretation of cues are interesting and important. It may be that interaction is observed in cases in which processing is made difficult either through data limitations, such as those that arise in such techniques as the backward-masking paradigm used to demonstrate the word superiority effect (McClelland & Rumelhart, 1981), or through resource limitations, such as those found under speed stress (Pitt, 1995b) or reduced attention (Gordon, Eberhardt, & Rueckl, 1993). In addition, demanding tasks such as discrimination may be more likely to show interactive effects than simpler tasks such as classification (Polka & Strange, 1985). To the extent that the presence of interactive effects in the perception of different features depends on data limitations, resource limitations, and other task demands, such effects should be approached from a modeling framework that characterizes the dynamic processing mechanisms involved in perception rather than from a modeling perspective that focuses exclusively on how different types of information are combined. The present study, focusing on combinations of different types of information in a nondemanding task, shows that an additive model provides a sufficient account of performance when competent speakers make classification judgments on a distinction present in their native language.

A single logit model did not have the same uniformly high level of success in fitting results from the Japanese subjects. Seven of the 12 Japanese subjects showed significant effects of both $F1$ and $F3$, whereas the other 5 subjects failed to show effects of one of the features or of both features. A group model failed to fit the data from all the Japanese subjects, although a group model did successfully fit data from the subset of Japanese subjects who showed significant effects of both $F1$ and $F3$. Even for this skilled group of Japanese subjects, the magnitude of the effects of the cues was smaller than for the group model fit to the American subjects. The results show that an additive logit model can successfully indicate whether subjects are responsive to specific acoustic cues in attempting to perceive nonnative contrasts.

The results of the research show a strong relation between the accuracy of Japanese subjects in recognizing natural /r/ and /l/ and in using the acoustic information in the synthetic speech sounds. In particular, variation in natural-speech recognition accuracy is strongly related to ability to use $F3$-onset frequency in classifying synthetic speech. In addition to providing information about the acoustic basis of /r/–/l/ perception by Japanese speakers, the finding of a strong relation between natural and synthetic speech perception provides external validation of the model-fitting results obtained with synthetic speech. Previous work modeling the use of acoustic cues in speech perception has relied solely on the fits of models to judgments of the synthetic speech sounds. By showing that the parameters of a model fit to judgments on synthetic sounds are related to accuracy in recognizing natural speech, the present research shows that the logit model incorporating additive cue effects yields valid predictions beyond the domain of the data on which it was developed.

REFERENCES

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum, (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167-224). Cambridge, MA: MIT Press.

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, **20**, 305-330.

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.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, **101**, 2299-2310.

Crowther, C. S., Batchelder, W. H., & Hu, X. (1995). A measurement-theoretic analysis of the fuzzy logic model of perception. *Psychological Review*, **102**, 396-408.

Dalston, R. M. (1975). Acoustic characteristics of English /w, r, l/ spoken correctly by young children and adults. *Journal of the Acoustical Society of America*, **57**, 462-469.

Espy-Wilson, C. Y. (1992). Acoustic measures for linguistic features distinguishing the semi-vowels /w j r l/ in American English. *Journal of the Acoustical Society of America*, **92**, 736-757.

Gordon, P. C., Eberhardt, J. L., & Rueckl, J. G. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology*, **25**, 1-42.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

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.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, **89**, 874-886.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

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

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of perception. *Cognitive Psychology*, **21**, 398-421.

Massaro, D. W., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 3, pp. 129-165). New York: Academic Press.

Massaro, D. W., & Oden, G. C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1053-1064.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, **23**, 1-44.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effect in letter perception. Part 1. An account of basic findings. *Psychological Review*, **88**, 375-407.

Miller, J. L. (1977). Nonindependence of feature processing in initial consonants. *Journal of Speech & Hearing Research*, **20**, 519-528.

Miyawaki, K., Strange, W., Verbrugge, 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.

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, **18**, 347-373.

O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1957). Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word*, **13**, 24-43.

Oden, G. C. (1979). A fuzzy logical model of letter identification. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 336-352.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

Olive, J. P., Greenwood, A., & Coleman, J. (1993). *Acoustics of American English: A dynamic approach*. New York: Springer-Verlag.

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

Pitt, M. A. (1995a). Data fitting and detection theory: Reply to Massaro and Oden (1995). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1065-1067.

Pitt, M. A. (1995b). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1037-1052.

Polka, L., & Strange, W. (1985). Perceptual equivalence of acoustic cues that differentiate /r/ and /l/. *Journal of the Acoustical Society of America*, **78**, 1187-1197.

Repp, B. H. (1983). Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. *Speech Communication*, **2**, 341-362.

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

Strange, W. (1995). Cross-language studies of speech perception: A historical review. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 3-45). Baltimore, MD: York.

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

Underbakke, M., Polka, L., Gottfried, T. L., & Strange, W. (1988). Trading relations in the perception of /r/–/l/ by Japanese learners of English. *Journal of the Acoustical Society of America*, **84**, 90-100.

Vance, T. J. (1987). *An introduction to Japanese phonology*. Albany: State University of New York Press.

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 speech sounds to spoken words* (pp. 93-120). Cambridge, MA: MIT Press.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Erlbaum.

Yamada, R. A. (1995). Age of acquisition of second language speech sounds: Perception of American English. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 305-320). Baltimore, MD: York.

Yamada, R. A., & Tohkura, Y. (1991). Perception of American English /r/ and /l/ by native speakers of Japanese. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 155-174). Tokyo: Ohmsha.

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.

NOTES

1. The question of independent versus interactive combination has also been a source of conceptual challenge in assessing the computational properties of connectionist (interactive activation) models of information processing. Massaro (1989) demonstrated that the basic processing architecture proposed by McClelland and Rumelhart (1981) to account for interactive effects, such as the word superiority effect, did not in fact yield interactive combination of information. Rather, interaction emerged from the decision rule that was used to generate a response from the activation levels. Furthermore, use of that decision rule meant that the processing architecture was incapable of generating additive patterns of information combination, such as those demonstrated by Oden and Massaro (1978). McClelland (1991) showed that an interactive activation model with stochastic properties exhibited additive effects in a principled manner but left open the question of how an interactive activation architecture could exhibit both additive and interactive combination.

2. When fitting the logit model to the subjects, cells with observed zero probabilities were replaced by .5/9 (where 9 is the total number of responses in each cell in the present situation), a procedure automatically implemented in SAS/CATMOD, which was used for the model fitting. Such a procedure is necessary for the weighted least squares (WLS) estimation, and we believe that such a small modification should not distort our results. Alternatively, without imputing small numbers in the missing cells, one might use the maximum likelihood (ML) option for model fitting. However, we found that this procedure often led to inadmissible solutions (e.g., infinite estimates) and, when there were admissible solutions, the pattern of results was very much like that obtained using WLS estimation. Therefore, for consistency, we employed WLS estimation throughout the study.

Our use of WLS estimation contrasts with the use by Oden and Massaro (1978) of root mean square deviation (RMSD) from observed response proportions in fitting their fuzzy logical model of perception. Massaro and Oden (1995) have criticized Pitt (1995b) for using least squares estimation on normal deviates, a nonlinear transformation that is similar to the logit transform that we employ. They argue that, in such cases, least squares estimation overemphasizes small differences at extreme probabilities while underemphasizing larger differences at

midrange probabilities. WLS estimation provides a principled approach to this issue by taking into account the differing variance at different probabilities; deviations with small variance are given more weight than deviations with large variance (Wickens, 1989).

3. Accuracy for the Japanese subjects in recognizing the /r/–/l/ contrast in the speech of the 4 American speakers ranged from 77.1% to 83.4%, indicating that the difficulty in identification was reasonably general across the speakers who produced the stimuli. The pattern of effects was also reasonably general across words. For the 24 words with nonfinal contrasts, accuracy ranged from 52.1% to 93.7% and was distributed fairly evenly across that range. For the 8 words with final contrasts, 3 were recognized perfectly, 4 yielded one error, and 1 yielded three errors. The pattern of effects was again reasonably general, though more variable, when looking at individual tokens. For the 96 tokens with nonfinal contrast, accuracy ranged from a low of 41.2% to a high of 100%, which was observed for 4 tokens. For the 24 tokens with a final contrast, 7 tokens showed one error, and 17 tokens were recognized perfectly.

For the American subjects, the three errors were made on 3 different words pronounced by 3 different speakers and were committed by 3 different subjects. The high accuracy by the American subjects shows that the word stimuli were sufficient to support high accuracy in classification. The distribution of errors by the Japanese subjects across speakers, words, and tokens indicates that the errors primarily reflected the abilities of the subjects rather than characteristics of the stimuli (with the exception of the position of the contrast, word final vs. word nonfinal).

4. The effect size of the acoustic dimension is defined as the standard deviation of the parameter estimates for different levels of that dimension. It provides a good estimate of the perceptual impact of an acoustic cue when parameter values are a monotonic function of stimulus levels (see Figure 5).

(Manuscript received June 5, 1998;
revision accepted for publication August 23, 2000.)