Influences of phonetic identification and category goodness on American listeners’ perception of \( /r/ \) and \( /l/ \)

Paul Iverson and Patricia K. Kuhl

Department of Speech and Hearing Sciences and Virginia Merrill Bloedel Hearing Research Center, University of Washington, Box 357920, Seattle, Washington 98195-7920

(Received 7 February 1995; revised 28 August 1995; accepted 11 September 1995)

Recent experiments have demonstrated that category goodness influences the perception of vowels [Iverson and Kuhl, J. Acoust. Soc. Am. 97, 553–562 (1995)]; listeners show a perceptual magnet effect characterized by shrunken perceptual distances near excellent exemplars of vowel categories and stretched distances near poor exemplars. The present study extends this investigation by examining the relative influence of phonetic identification and category goodness on the perception of American English \( /r/ \) and \( /l/ \). Eighteen \( /ra/ \) and \( /la/ \) tokens were synthesized by varying \( F_2 \) and \( F_3 \) frequencies. Adult listeners identified and rated the goodness of individual stimuli, and rated the similarity of stimulus pairs. Multidimensional scaling analyses revealed that the perceptual space was shrunk near the best exemplars of each category and stretched near the category boundary. In addition, individual differences in \( /r/ \) identification corresponded to the degree of shrinking near the best exemplars of the \( /r/ \) category. The results demonstrate that category goodness and phonetic identification both contribute to the perception of \( /r/ \) and \( /l/ \).

PACS numbers: 43.71.An, 43.71.Es

INTRODUCTION

Research on speech perception has demonstrated that listeners are much more sensitive to acoustic differences among tokens from different phonetic categories than they are to differences among tokens from the same phonetic category, even when the physical differences separating stimuli have been equated (Liberman et al., 1957; Studdert-Kennedy et al., 1970; Repp, 1984). Increased sensitivity to acoustic differences near phonetic boundaries may initially be inherent in the auditory processing of speech; infants show increased sensitivity to between-category differences in the absence of extensive experience with language (Eimas et al., 1971; Eimas, 1974, 1975; Streeter, 1976; Swoboda et al., 1978), and nonhuman animals show increased sensitivity to consonant boundaries (Kuhl and Miller, 1975; Kuhl and Pad- den, 1982, 1983; Dooling et al., 1995). Regardless of the cause of these auditory sensitivities, there is no doubt that they change substantially with exposure to a specific language; adults become especially sensitive to their own native language phonetic contrasts (Strange and Jenkins, 1978; Best et al., 1988; Werker and Polka, 1993). Moreover, there is ample evidence suggesting that linguistic experience alters these perceptual sensitivities early in life (Kuhl et al., 1992; Kuhl, 1994; Werker and Tews, 1984; Werker and Polka, 1993). The main question for speech perception theories is what is the nature of the change brought about by exposure to a specific language?

Although the tradition of categorical perception has focused attention on linguistic experience and perceptual sensitivity at phonetic boundaries (e.g., Miyawaki et al., 1975; Best et al., 1988; Werker and Logan, 1985), more recent work has examined the internal structure of phonetic categories (Miller and Volaitis, 1989; Volaitis and Miller, 1992; Wayland et al., 1994; Kuhl, 1991; Iverson and Kuhl, 1995; cf. Sussman and Lauckner-Morano, 1995). Within phonetic categories, listeners consistently judge that certain exemplars of their native phonetic categories are particularly good, and this category goodness strongly influences sensitivity to acoustic differences. For example, Iverson and Kuhl (1995) synthesized 13 variants of the vowel \( /l/ \) (as in the word “he”), and had subjects identify and rate the goodness of each token on a scale from 1 (“bad”) to 7 (“good”). Subjects reliably judged that specific tokens with high \( F_2 \) and low \( F_1 \) frequencies were the best exemplars of the \( /l/ \) category in this stimulus set. In additional experiments, listeners’ perception of these tokens was modeled using signal detection theory (Green and Swets, 1966; Macmillan and Creelman, 1991) and multidimensional scaling (Shepard, 1962a, b). The results demonstrated that the perceptual space was shrunk (reduced sensitivity to acoustic differences) near the best exemplars of the \( /l/ \) category, and stretched (increased sensitivity to acoustic differences) near poor exemplars. This distortion of the perceptual space has been characterized as a perceptual magnet effect (Kuhl, 1991, 1992, 1993a, b; Kuhl and Iverson, in press); the best exemplars of a category pull neighboring tokens closer in the perceptual space.

The perceptual magnet effect seems critically dependent on exposure to language early in life. Kuhl et al. (1992) tested 6-month-old infants in America and Sweden on synthesized variants of the English \( /l/ \) and the Swedish \( /y/ \). Both groups of infants demonstrated a perceptual magnet effect for their native language vowel; American infants had reduced discrimination sensitivity near the best instances of English \( /l/ \) and Swedish infants had reduced discrimination sensitivity near the best instances of Swedish \( /y/ \). In addition,
Kuhl (1991) tested American adults, 6-month-old infants, and Rhesus monkeys on synthesized variants of the English /l/. Adults and infants showed reduced sensitivity for the best exemplars of /l/, but monkeys did not. Thus the perceptual magnet effect seems a result of acquiring one’s native language.

The present study evaluates whether the perceptual magnet effect influences the perception of the American English /t/ and /l/ consonant categories. Previous experiments on the perceptual magnet effect have primarily examined the /l/ vowel category, so tests of additional categories are necessary to evaluate the generality of this phenomenon. The /t/ and /l/ categories are particularly interesting because they seem substantially influenced by linguistic experience; adult native speakers of some languages (e.g., Japanese and Korean) have great difficulty learning these English categories (Goto, 1971; Logan et al., 1991; Miyawaki et al., 1975; Strange and Dittmann, 1984; Gillette, 1980). An examination of /t/ and /l/ consonants seems promising because the perceptual magnet effect has proven useful for explaining the influence of linguistic exposure.

An additional goal of this study is to better evaluate the relative contribution of basic auditory processing, phonetic identification, and category goodness to the perception of these consonants. Since the earliest categorical perception experiments, phonetic identification has been known to predict the discrimination of speech sounds (Liberman et al., 1957; Pisoni, 1975); sounds are easy to discriminate when they receive different phonetic labels and are hard to discriminate when they receive the same phonetic labels. It is thus important to examine what additional variance can be explained by the perceptual magnet effect. Previous studies of the perceptual magnet effect (Kuhl, 1991; Kuhl et al., 1992; Iverson and Kuhl, 1995; Sussman and Laukner-Morano, 1995) have tested stimuli that mostly belonged to a single phonetic category (/i/), so it has been difficult to examine the influence of category boundaries. The present study uses a stimulus set with excellent exemplars from two categories so that the relative influence of goodness and phonetic identification can be evaluated.

The stimulus set of this study is composed of 18 /ra/ and /la/ stimuli that vary in F2 and F3 frequency during the initial consonant closure. F2 varies on three levels and F3 varies on six levels to form a two-dimensional “grid” of tokens. At the beginning of the experimental session, subjects identify each token as /l/, /l/, or /l/, and rate the goodness of each token on a scale from 1 (bad) to 7 (good). Subjects then complete a longer experiment in which they rate the similarity of every pair of tokens on a scale from 1 (dissimilar) to 7 (similar).

The similarity ratings are designed for analysis by multidimensional scaling (MDS; Shepard, 1962a, b) to map the perceptual space underlying these tokens, and to assess the relative contribution of acoustic distance, phonetic identification, and category goodness to perceptual similarity. MDS assigns tokens to a geometric space where distances in the space correspond to perceived similarity; similar tokens are placed close together in MDS solutions and dissimilar tokens are placed far apart. Modeling similarity in this manner uncovers relationships among tokens that would be difficult to observe from raw similarity ratings.

For consonant perception, MDS has mostly been used to model identification errors by impaired listeners (Danahauer and Lawarre, 1979; Doyle et al., 1981; Walden et al., 1980; Gordon-Salant, 1985a) and by unimpaired listeners when speech is mixed with noise (Pols, 1983; Gordon-Salant, 1985b). These studies have used MDS to identify the dimensions that distinguish tokens from different categories (e.g., formant transitions, voicing, and frication), and to compare the dimensions used by different subject populations. In contrast, the present experiment maps the perceptual space within consonant categories, as has been accomplished for vowels (Iverson and Kuhl, 1995; Kewley-Port and Atal, 1989). MDS will be used to map listeners’ sensitivities to F2 and F3 frequencies within each category, and these solutions will be compared to judgments of phonetic identification and category goodness. The aim of this experiment is to examine whether the perceptual space underlying /t/ and /l/ is shrunk near the best exemplars and stretched near the category boundary, as predicted by the perceptual magnet effect and categorical perception.

I. METHOD

A. Subjects

Twenty-eight adult members of the University of Washington community participated in this experiment. One participant was dropped from the analysis because he did not follow the instructions for the goodness and identification portion of the experiment. All subjects were native English speakers, all reported having no known hearing impairments, and all received course credit for participating in this 1-h experiment.

B. Apparatus

The stimuli were presented by a Data Translation DT2821 digital audio board controlled by an NEC 386 microcomputer. The sounds were played to subjects using the right-ear speaker of a pair of Telephonic TDH-39P headphones while subjects sat in a sound-treated booth. The sounds were reproduced with 10,000 12-bit samples per second, and were low-pass filtered with a 4.6-kHz cutoff frequency. Responses were entered and recorded using the computer that controlled the presentation of stimuli.

C. Stimuli

Eighteen /ra/ and /la/ tokens were synthesized using the SenSyn (1992) implementation of the Klatt and Klatt (1990) speech synthesizer. The synthesis parameters modeled the speech of an individual adult female talker. This individual was recorded saying /ra/ and /la/ syllables in a clear (hyper-articulated) fashion. Two tokens (one /ra/ and one /la/) were selected from this recording session on the basis that they were closely matched in all acoustic characteristics other than F2 and F3 frequency during the initial consonant. Synthesis parameters were then chosen to produce tokens that modeled the acoustics of these two recorded syllables.
The synthesis parameters are described in detail in Appendix A, and the tokens are schematically displayed in Fig. 1. The $F_2$ and $F_3$ frequencies during the initial consonant and transition were varied, but the vowels were identical for all tokens. During the consonant, $F_2$ frequency was varied on three levels (744, 1003, and 1301 Hz) and $F_3$ frequency was varied on six levels (1325, 1670, 2067, 2523, 3047, and 3649 Hz) to create a two-dimensional grid of 18 tokens. These frequencies were equally spaced on the mel scale (Stevens et al., 1937). Fant (1973) has argued that the mel scale is appropriate for speech stimuli because difference limens for the first three formants are similar when measured in mels, and the mel scale corresponds to excitation patterns on the basilar membrane.

All other synthesis parameters were identical for all tokens. During the consonant, $F_1$ frequency was 351 Hz and $F_4$ frequency was 4512 Hz. The bandwidths during the consonant were 200, 100, 150, and 100 Hz, respectively for $F_1-F_4$. The narrow bandwidths of the higher formants were necessary to match the formant amplitudes of the natural tokens that guided synthesis. During the vowel, the formant frequencies were 796, 1221, 2973, and 4512, respectively for $F_1-F_4$. The bandwidth of $F_4$ was 400 Hz during the vowel, but the bandwidths of the other formants were the same as during the consonant.

Each token was 800 ms long. The formant frequencies did not change during the first 155 ms of the consonant or during the last 545 ms of the vowel. All formant transitions started 155 ms into the tokens; the $F_1$ and $F_2$ transition durations were 35 ms, and the $F_3$ transition duration was 100 ms. $F_0$ rose from 204 to 216 Hz over the first 245 ms, and then fell to 137 Hz during the rest of the vowel.

The tokens were equalized in rms amplitude and were played to subjects at a comfortable level. Careful listening by the experimenters verified that these amplitude-equalized tokens were equally loud.

FIG. 1. Formant frequencies of stimuli and example spectrograms. The 18 /r/ and /l/ tokens varied in $F_2$ and $F_3$ frequency during the initial consonant closure. The dimensions were varied independently in 200-mel steps to form a two-dimensional grid of stimuli. All other acoustic factors (e.g., vowel formants and transition lengths) were identical for all tokens.

D. Procedure

1. Goodness and identification

At the start of the experiment, subjects completed a short session in which they identified and rated the goodness of individual tokens. For the identification task, subjects judged whether the initial consonant was /r/, /l/, or /w/. After making this identification, subjects were asked to rate whether it was a good example of the category on a scale from 1 (bad) to 7 (good); for example, subjects rated how well a token represented the /r/ category if they had identified the token as /r/.

Each subject completed a practice block of 18 trials with each of the tokens presented once in a random order. After the practice, subjects completed an experimental session of 36 trials (2 blocks of the 18 tokens) with the order of trials randomized within each block.

2. Similarity scaling

After the goodness and identification session, subjects completed a longer session in which they rated the similarity of each pair of the 18 tokens. On each trial, subjects heard two tokens separated by 350 ms and rated the similarity of each pair on an integer scale from 1 (dissimilar) to 7 (similar).

Each subject completed a practice block of 36 trials composed of randomly selected pairs of the 18 tokens. They were instructed to normalize their responses to the range of stimuli in the practice so that the least similar pairs in the set would get the lowest ratings, the most similar pairs would get the highest ratings, and the other pairs would get ratings that reflected their intermediate degree of similarity. After the practice session, subjects completed an experimental session of 306 trials composed of every possible pair of the 18 tokens (tokens were never paired with themselves). Subjects were allowed to take a short break after every 68 trials.

II. RESULTS AND DISCUSSION

A. Identification and goodness

To assess the variability of identifications, the aggregate percentages of /r/, /l/, and /w/ judgments (averaged across stimuli and trials) were calculated for each subject. Histograms of the /r/ and /l/ percentages (Fig. 2) suggested that the distribution of subjects was bimodal, with one group of subjects who made both /r/ and /l/ judgments, and a smaller
group of subjects who made nearly 0% /l/ and 100% /r/ judgments; the percentage of /w/ judgments was low for both groups. To further examine the distribution of responses, one-sample Kolmogorov–Smirnov tests were used to compare the shape of each distribution to a standard normal distribution with the same mean and variance (Lilliefors, 1967; Wilkinson, 1989). The distribution of /r/ identifications, \( D(N=27) = 0.198, \ p < 0.10 \), and the distribution of /l/ identifications, \( D(N=27) = 0.177, \ p < 0.05 \), were both significantly different in shape from the normal distributions, coinciding with the observed distributions of the histograms. Apparently, a subset of the subjects did not think that many of these tokens were acceptable members of the /r/ category.

To examine whether this individual difference influenced the perceptual space underlying these categories, the subjects were divided into two groups for the following statistical analyses: 18 subjects who made more than 25% /r/ identifications (many-/r/ subjects), and nine subjects who made fewer than 25% /r/ identifications (few-/r/ subjects). The 25% criterion for grouping subjects was selected based on visual inspection of the /l/ and /r/ identification histograms displayed in Fig. 2, and on inspection of additional histograms of this data with narrower category widths. Although the histograms seemed bimodal, the two underlying response distributions did not appear to be entirely distinct. The 25% criterion was selected because it was in the middle of the region where the tails of the two response distributions seemed to overlap, and it thus best divided the subjects into distinct groups.

The average identification and goodness ratings for each token (Fig. 3) further show the differences between these two groups of subjects. The many-/r/ subjects heard tokens with low F3 frequencies as good exemplars of /l/, but the few-/r/ subjects mostly heard these tokens as poor exemplars of /l/ (one token was a poor example of /l/). Both groups of subjects heard tokens with high F3 frequencies as good exemplars of /l/.

The best /l/ and /r/ stimulus locations were calculated for each subject by determining the F2 and F3 frequencies of the /l/ and /r/ tokens with the highest goodness ratings (best frequencies were averaged when more than one token received the same highest rating). The means and standard errors of these locations are displayed in Table I. For the many-/r/ subjects, the best /l/ location occurred near the lowest F3 frequency \( (M = 1473 \text{ Hz}) \) and the intermediate F2 frequency \( (M = 977 \text{ Hz}) \). For both groups of subjects, the best /l/ location occurred near the highest F3 frequency \( (M_{\text{many-/r/}} = 3478 \text{ Hz}, M_{\text{few-/r/}} = 3329 \text{ Hz}) \) and the intermediate F2 frequency \( (M_{\text{many-/r/}} = 1011 \text{ Hz}, M_{\text{few-/r/}} = 1086 \text{ Hz}) \). The standard errors for all of these location estimates were substantially smaller than the 200-mel steps between tokens, indicating that these location estimates were quite consistent. Independent-samples t tests revealed that the /l/ location estimates of the two groups were not significantly different in \( F_2, t(25) = -1.77, \ p > 0.05 \), but the best /l/ of the few-/r/ subjects had a significantly lower F3 frequency, \( t(25) = 2.14, \ p < 0.05 \). Although the F3 difference was small (53 mels) compared to the stimulus step size (200 mels), subjects who perceived good /l/ tokens in this stimulus set preferred /l/ tokens that had slightly more extreme F3 frequencies.
B. Similarity

Each subject’s ratings were put into the form of a lower triangular matrix composed of the similarity for each pair of tokens averaged across presentation order. Intersubject correlations were run for each group of subjects to assess the consistency of similarity judgments. For the 18 many-/r/ subjects, the average of the 153 intersubject correlations was $r=0.68$ (df=151) and each was significant at the $p<0.001$ level. For the nine few-/r/ subjects, the average of the 36 intersubject correlations was $r=0.57$ (df=151) and each was significant at the $p<0.001$ level. Thus the similarity ratings were highly consistent among subjects.

The matrices were averaged for the two groups of subjects and were analyzed separately using the Kruskal (1964a, b) MDS algorithm implemented by the SYSTAT computer program (Wilkinson, 1989). The MDS analysis used Kruskal’s stress formula 1, a Euclidean distance metric, and a monotonic regression function. This placed the tokens in a two-dimensional space where the distances between tokens were fit to a monotonic function of the similarity ratings. The solution for the many-/r/ subjects fit the similarity ratings with a stress of 0.05 (accounting for 99% of the variance), and the solution for the few-/r/ subjects fit the similarity ratings with a stress of 0.06 (accounting for 97% of the variance). Figure 3 displays the MDS solutions.

The MDS solutions revealed influences of acoustic distance, category goodness, and phonetic identification on perceptual similarity. For both groups, the ordering of tokens on the horizontal axes corresponded to $F3$ frequency and the vertical ordering of tokens corresponded to $F2$ frequency, demonstrating a strong relationship between similarity and acoustic distance. In addition, both groups of subjects showed clustering of the perceptual space in the $F2$ and $F3$ dimensions near the best exemplars of /l/ (tokens with the highest $F3$ frequencies).

Differences between the solutions for the two groups of subjects were apparent in the middle of the $F3$ series and at the lowest $F3$ frequencies. For the many-/r/ subjects, there was stretching of the perceptual space at the phonetic boundary, but this stretching was less defined for the few-/r/ subjects, coinciding with their lack of a clear boundary. At the lowest $F3$ frequencies, there was strong clustering of the perceptual space in the $F2$ and $F3$ dimensions for the many-/r/ subjects, but the clustering was weaker (especially in the $F2$ dimension) for the few-/r/ subjects; strong clustering of the perceptual space at the lowest $F3$ frequencies seems related to the presence of excellent exemplars of the /r/ category. An independent samples $t$ test evaluated whether this difference in clustering was significant. For each subject, the average similarity rating was calculated for pairs that only included the six tokens with the lowest two $F3$ frequencies (15 pairs of tokens). The similarity ratings for many-/r/ subjects ($M=5.89$) was significantly higher than for few-/r/ subjects ($M=5.35$), $t(25)=-2.30, p<0.05$, supporting the observation that the many-/r/ subjects had greater perceptual clustering near the best exemplars of /r/.

This intersubject difference further supports the link between category representation and the distortion of the perceptual space attributed to the perceptual magnet effect. Although all subjects reported having normal hearing, a third of these subjects did not judge many of the tokens to be acceptable members of the /r/ category, and this difference in categorization led to differences in perceptual distortion. The absence of an excellent /r/ led to less shrinking of the perceptual space at the low $F3$ frequencies and the absence of a sharp category boundary made the stretching of the space less clearly defined.

C. Tests of the relationship between similarity, acoustic differences, goodness, and identification

Additional analyses were conducted to further assess how well acoustic differences, category boundaries, and best exemplars predict similarity. From the earliest experiments on categorical perception, phonetic identification has been used successfully to predict the probability of discriminating speech stimuli (Liberman et al., 1957); the probability of discriminating tokens correlates with the probability of identifying them differently. Through the application of detection theory, identification percentages can also be used to estimate the perceptual distances separating tokens (Macmillan and Creelman, 1991). Within this theoretical framework, the $z$-transformed identification probability for each token, $z(p)$, indicates its location relative to the category boundary. The absolute value of this measure indicates each token’s distance from the boundary in standard-deviation units. The sign of this measure indicates whether each token is within (positive) or out of (negative) the category. For example, $z(p)=0.0$ for tokens that are identified as a member of the category on 50% of trials, $z(p)=2.3$ for tokens that are identified as a member of the category on 99% of trials.

<table>
<thead>
<tr>
<th>Condition</th>
<th>$N$</th>
<th>$M_{mel}(SE_{mel})$</th>
<th>$M_{Hz}$</th>
<th>$M_{mel}(SE_{mel})$</th>
<th>$M_{Hz}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Best /l/ tokens</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many-/r/ subjects</td>
<td>18</td>
<td>977 (28)</td>
<td>968</td>
<td>1306 (27)</td>
<td>1473</td>
</tr>
<tr>
<td>Few-/r/ subjects</td>
<td>9</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>Best /r/ tokens</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many-/r/ subjects</td>
<td>18</td>
<td>1008 (23)</td>
<td>1011</td>
<td>2162 (17)</td>
<td>3478</td>
</tr>
<tr>
<td>Few-/r/ subjects</td>
<td>9</td>
<td>1061 (21)</td>
<td>1086</td>
<td>2114 (15)</td>
<td>3329</td>
</tr>
</tbody>
</table>

### Table I. Average formant frequencies of best /r/ and all /l/ tokens.
and $z(p) = -2.3$ for tokens that are identified as a member of the category on 1% of trials. The perceptual distances between pairs of tokens ($d’$) can then be found by subtracting these location measures; tokens that are at similar locations will have a small $d’$ and tokens that are at dissimilar locations will have a large $d’$. In other words, $d’$ will be greater to the extent that tokens are identified differently. Multiple regression analyses were used to assess the relative contribution of phonetic identification, goodness, and acoustic distance to similarity ratings. The identification judgments were used to estimate perceptual distances by calculating the $z$ transform of the mean /l/ identification percentage for each token and then taking the absolute value of the difference for each pair of tokens. The $z$ transform reaches infinity when percentages equal 0 or 100, so tokens with 0% /l/ identifications were assigned values of 1% and tokens with 100% /l/ identifications were assigned values of 99% (Macmillan and Creelman, 1991). Acoustic distances were estimated by measuring the distances between tokens in the two-dimensional stimulus space displayed in Fig. 1 (calculated by taking the rms of the F2 and F3 mel frequency differences). Goodness was quantified by averaging the goodness ratings for each pair of tokens. The /r/ and /l/ tokens were analyzed separately, so goodness ratings were only averaged for pairs of tokens that listeners judged were members of the same category.

Table II displays the results of the multiple regression analyses. Separate analyses were conducted for (1) the 9 /r/ tokens of the many-/r/ subjects, (2) the 9 /l/ tokens of the many-/l/ subjects, and (3) the 17 /l/ tokens of the few-/l/ subjects. For the /r/ tokens identified by the many-/r/ group, the model accounted for 73% of the variance, and acoustic distance, $t(32)=-6.15$, $p<0.001$, identification distance, $t(32)=-2.88$, $p<0.01$, and goodness, $t(32)=2.47$, $p<0.05$, each significantly contributed to the model. For the 17 /l/ tokens identified by the few-/l/ group, the model accounted for 87% of the variance, and acoustic distance, $t(132)=-14.00$, $p<0.001$, identification distance, $t(132)=-7.31$, $p<0.001$, and goodness, $t(132)=6.17$, $p<0.001$, each significantly contributed to the model.

For all three models, acoustic distance most strongly corresponded to similarity; listeners judged that tokens were less similar to the extent that they had F2 and F3 frequency differences. In addition, similarity was related to identification and goodness; subjects judged that tokens were less similar when they were identified differently, and they thought tokens were more similar when they were good exemplars of the same category. The significant contribution of goodness verifies that the distortion of the perceptual space cannot be fully explained by identification percentages. The identification percentages primarily are a function of F3 frequency, so the distance estimates based on identification best account for perceptual distortion along the F3 dimension; the spacing of tokens along the F2 dimension is poorly determined by identification. In addition, distance estimates based on identification fail to predict the patterns of clustering for tokens that receive 100% /l/ identifications; identification predicts that these tokens should have uniform similarity. Goodness significantly contributes to these models because it accounts for (1) clustering along the F2 dimension of both categories, and (2) clustering for the exemplars that received 100% /r/ identifications.

### III. GENERAL DISCUSSION

The results demonstrate that general auditory sensitivities, categorical perception, and the perceptual magnet effect all contribute to the perception of American English /r/ and /l/ tokens. Phonetic identification accounts for differences in sensitivity that are a function of distances from the /r–l/ boundary; it accounts for the stretching of the perceptual space in the F3 dimension for tokens that receive less than 100% /r/ or /l/ identifications. Goodness better accounts for sensitivity parallel to the /r–l/ boundary (in the F2 dimension), and near tokens that receive 100% /r/ or /l/ identifications. Thus the perceptual distortion attributed to the perceptual magnet effect is in addition to that predicted by traditional categorical perception models (Liberman et al., 1957; Studdert-Kennedy et al., 1970; Repp, 1984). The perceptual space is shrunk near excellent exemplars of /r/ and /l/ and stretched near poor exemplars, and these distortions are independent of stretching at the category boundary. This confirms that the perceptual magnet effect found for vowels (Iverson and Kuhl, 1995; Kuhl et al., 1992; Kuhl, 1991) also influences the perception of /r/ and /l/ consonants.

The results also demonstrate that individual differences in identification and goodness correspond to individual differences in perceptual clustering. Subjects who made many-/r/ identifications showed substantial perceptual clustering for tokens with low F3 frequencies (the best exemplars of the /r/ category); subjects who made few-/l/ identifications showed less perceptual clustering for the same tokens. This further supports the hypothesis that the observed distortions.
of the perceptual space were the result of mental representations for phonetic categories rather than the result of peripheral auditory processing or stimulus artifacts. The perceptual clustering at the lowest F3 frequencies seems dependent on the presence of excellent exemplars of the /l/ category.

One possible cause of these individual differences is that all acoustic parameters other than F2 and F3 were set to the same neutral values for all tokens, and some of these neutral values may have favored /l/ identifications. A likely candidate is F1 transition length. Previous researchers (Polka and Strange, 1985; O’Connor et al., 1957; Dalston, 1975) have suggested that F1 transition length influences the location of /r–l/ category boundaries; stimuli with long F1 transitions sound more like /l/, and stimuli with short F1 transitions sound more like /r/. The F1 transition length was set to 0.65 ms for all tokens in this study, and this length may have promoted /l/ identifications. Although this transition length was acceptable for most subjects, some subjects may have required longer F1 transitions to match their /l/ category representations.

To test this possibility, an additional goodness and identification experiment (detailed in Appendix B) was conducted using tokens with a longer F1 transition more typical of /l/. The F1 transition length was set to 70 ms for all tokens, but in all other respects the tokens were identical to the 18 stimuli of the original experiment. Sixteen subjects identified each token as /l/ or /r/ and /l/ or /r/, and rated goodness on a scale from 1 (bad) to 7 (good). The results did not have the individual differences observed in the original set; the /l/ and /r/ identification histograms appeared normally distributed, and statistical tests demonstrated that they were not significantly different in shape from standard normal distributions. Thus the individual differences of the original set seem attributable to the short F1 transition length. Even for tokens with F2 and F3 frequencies characteristics of /l/, a subset of the subjects needed a longer F1 transition for these tokens to be members of the /l/ category.

The similarity scaling task and MDS technique used in the present experiment are less standard than the discrimination tasks commonly used in speech perception research, but there is little indication that a discrimination experiment would have yielded different results. First, the results from this study are in complete agreement with previous discrimination experiments which have demonstrated that native speakers of English are especially sensitive to acoustic differences at the /r–l/ boundary (Miyawaki et al., 1975; Mackain et al., 1981). Second, recent experiments in our lab (Iverson et al., 1994) have used discrimination tasks to study the influence of linguistic experience on the perception of /l/ and /r/ and /l/, and these discrimination experiments have replicated the perceptual distortions observed in the present MDS solutions. It is important to note that similarity scaling tasks likely place greater demands on memory and attention than do most discrimination tasks, and that these cognitive demands may influence perceptual sensitivity (see related discussion by Macmillan et al., 1988). However, the evidence thus far suggests that the two tasks yield similar perceptual maps for /l/ and /l/.

Although the perceptual magnet effect seems a product of mental representations for phonetic categories, the present results do not reveal the underlying structure of these representations. In the cognitive categorization literature, effects of typicality have been explained by both prototype- and exemplar-based representations (Estes, 1993; Medin and Barsalou, 1987). The perceived goodness of a speech sound could be based on its similarity to an average instance of a category (i.e., a prototype) or on its overall similarity to multiple exemplars of a category stored in memory (Kuhl, 1993a, b). Both of these models are attractive because they indicate that the distribution of ambient speech sounds is sufficient to specify category goodness. Infants seem influenced by the typicality of vowels in their native language prior to the acquisition of word meaning (Kuhl et al., 1992), suggesting that ambient speech specifies typicality without higher level linguistic processing. These models indicate that infants may store whatever speech sounds they hear, and that the most frequent of these stored sounds would then be considered to be excellent exemplars of infants’ native phonetic categories.

One difficulty with prototype- and exemplar-based models is that the best exemplars of phonetic categories tend to have more extreme acoustic values than do average productions. The best stimuli in the present study were at the endpoints of the F3 frequency dimension, but speech produced by female talkers has less extreme average F3 frequencies for /l/ (1839 Hz) and /l/ (3117 Hz) than do these best exemplars (Iverson et al., 1994). Listeners also have been shown to prefer vowel sounds with more extreme formant frequencies than those they normally produce (Johnson et al., 1993). Best exemplars are not always those at the extremes of stimulus sets (e.g., Miller and Volaitis, 1989; Volaitis and Miller, 1992; Wayland et al., 1994; Kuhl, 1991; Iverson and Kuhl, 1995), but they may be more extreme than average productions. In this study, the present experiments demonstrate that the perceptual magnet effect influences the perception of /l/ and /l/ by American listeners. Individual differences in identification and goodness lead to differences in perceptual similarity, supporting the claim that the distortion due to the per-
ceptual magnet effect can be attributed to mental representations for phonetic categories. The perceptual magnet effect accounts for distortion of the perceptual space in addition to that explained by traditional categorical perception models.

**ACKNOWLEDGMENTS**

The research described in this report was supported by an NIH grant to Patricia K. Kuhl (DC 00520), and the first author received additional support from an NIH institutional training grant (HD 07391). We are grateful to Lisa D. Foss, Michelle Neuneker, and Nithya Siva for their assistance in conducting the experimental sessions, and to Michael D. Hall and Raquel Willerman for comments on portions of the manuscript.

**APPENDIX A: STIMULUS PARAMETERS**

Table AI lists the synthesis parameters for the token at the bottom left corner of the stimulus grid (the token with the lowest F2 and F3 frequencies). These parameters controlled the SENSYN (1992) implementation of the Klatt and Klatt (1990) speech synthesizer. All of the transitions between stimulus values were linear.

The parameters for the other stimuli in this set varied from this example stimulus in F2 and F3 frequency for the initial value and the transition, but the stimuli were identical.

---

**TABLE AI. Stimulus parameters for the token with the lowest F3 and F2 frequencies.**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>DU</td>
<td>Duration of the utterance</td>
<td>850 ms (silence was edited after synthesis to create 800-ms long tokens)</td>
</tr>
<tr>
<td>SR</td>
<td>Output sampling rate</td>
<td>10 000 samples/s</td>
</tr>
<tr>
<td>NF</td>
<td>Number of formants</td>
<td>4</td>
</tr>
<tr>
<td>SS</td>
<td>Source switch</td>
<td>natural</td>
</tr>
<tr>
<td>GV</td>
<td>Overall gain scale factor for amplitude of voicing</td>
<td>46 dB</td>
</tr>
<tr>
<td>GH</td>
<td>Overall gain scale factor for amplitude of aspiration</td>
<td>43 dB</td>
</tr>
<tr>
<td>F0</td>
<td>Fundamental frequency</td>
<td>0 Hz (0–20 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 0 to 204 Hz (20–25 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 204 to 186 Hz (25–60 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 186 to 216 Hz (60–260 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 216 to 204 Hz (260–445 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 204 to 137 Hz (445–740 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>137 Hz (740–850 ms)</td>
</tr>
<tr>
<td>AV</td>
<td>Amplitude of voicing</td>
<td>0 dB (0–10 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 0 to 50 dB (10–20 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 50 to 70 dB (20–80 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 70 to 77 dB (80–240 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 77 to 64 dB (240–645 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 64 to 40 dB (645–760 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 40 to 0 dB (760–800 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 dB (800–850 ms)</td>
</tr>
<tr>
<td>OQ</td>
<td>Open quotient</td>
<td>65%</td>
</tr>
<tr>
<td>TL</td>
<td>Extra tilt of voicing spectrum</td>
<td>0 dB (0–535 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 0 to 8 dB (535–850 ms)</td>
</tr>
<tr>
<td>AH</td>
<td>Amplitude of aspiration</td>
<td>Same as AV, although the amplitude of the stimuli was lower than the voicing amplitude due to the differences in the GV and GH parameters.</td>
</tr>
<tr>
<td>F1</td>
<td>Frequency of 1st formant</td>
<td>351 Hz (0–180 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 351 to 796 Hz (180–215 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>796 Hz (215–850 ms)</td>
</tr>
<tr>
<td>B1</td>
<td>Bandwidth of 1st formant</td>
<td>200 Hz</td>
</tr>
<tr>
<td>F2</td>
<td>Frequency of 2nd formant</td>
<td>744 Hz (0–180 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 744 to 1221 Hz (180–215 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1221 Hz (215–850 ms)</td>
</tr>
<tr>
<td>B2</td>
<td>Bandwidth of 2nd formant</td>
<td>100 Hz</td>
</tr>
<tr>
<td>F3</td>
<td>Frequency of 3rd formant</td>
<td>1325 Hz (0–180 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 1325 to 2973 Hz (180–280 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2973 Hz (280–850 ms)</td>
</tr>
<tr>
<td>B3</td>
<td>Bandwidth of 3rd formant</td>
<td>150 Hz</td>
</tr>
<tr>
<td>F4</td>
<td>Frequency of 4th formant</td>
<td>4512 Hz</td>
</tr>
<tr>
<td>B4</td>
<td>Bandwidth of 4th formant</td>
<td>100 Hz (0–150 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transition from 100 to 400 Hz (150–250 ms)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400 Hz (250–850 ms)</td>
</tr>
</tbody>
</table>
in all other respects. The initial $F_1$ frequency was varied on three levels (744, 1003, and 1301 Hz), and the stimuli had linear transitions from these frequencies to the same vowel $F_2$ frequency (1221 Hz). The initial $F_3$ frequency was varied on six levels (1325, 1670, 2067, 2523, 3047, and 3649 Hz), and the stimuli had linear transitions from these frequencies to the same vowel $F_3$ frequency (2973 Hz).

When interpreting these parameters, please note that there were 25 ms of silence at the beginning and end of this stimulus. For example, $F_1$ frequency was at a constant value (351 Hz) for the first 180 ms of the file, but the stimulus was actually silent for the first 25 ms because $F_0$ was set to 0 Hz; thus the stimulus duration of this initial $F_1$ portion was actually 155 ms.

**APPENDIX B: PERCEPTUAL TESTS OF STIMULI WITH LONGER TRANSITIONS**

A short goodness and identification experiment was conducted to assess whether the individual differences in identification can be attributed to the short $F_1$ transition duration (35 ms) used in the original stimulus set. The 18 /rl/ and /l/ tokens were resynthesized with a longer $F_1$ transition duration (70 ms), but in all other respects they were identical to the original set. Sixteen adult members of the University of Washington community participated in this experiment. As in the original experiment, subjects identified whether each token was /rl, l/, or /wl/, and then rated the category goodness of each token on a scale from 1 (bad) to 7 (good). They completed a practice block of 18 trials (each of the 18 tokens presented in a random order), and an experimental session of 36 trials (2 randomized blocks of the 18 tokens).

As in the original experiment, the aggregate percentage of /rl/ /l/, and /wl/ judgments was calculated for each subject to assess the variability of identifications. Histograms of the identification percentages (Fig. B1) suggested that subjects’ /rl/ and /l/ identifications were normally distributed and more homogenous compared to judgments on the original set of tokens. Supporting these observations, one-sample Kolmogorov–Smirnov tests (Lilliefors, 1967; Wilkinson, 1989) determined that the distributions of /rl/, $D(N=16) = 0.150, p > 0.05$, and /l/, $D(N=16) = 0.167, p > 0.05$, identifications were not significantly different in shape from standard normal distributions with the same mean and variance; the identifications of the original stimulus set were significantly different from normal distributions. In addition, Moses tests of dispersion (Moses, 1963; Daniel, 1978) were conducted to compare the variances of the identification distributions for the two stimulus sets. To calculate this statistic, the identification percentages of subjects in each experiment were randomly divided into small subsamples ($n=4$), (2) the sum of squared deviations from the mean was calculated for each subsample, and (3) a Mann–Whitney test (Mann and Whitney, 1947) was conducted to determine whether the subsample deviations differed for the two experiments. These tests revealed that the variances of the /rl/, $U(6,4) = 22, p < 0.05$, and /l/, $U(6,4) = 23, p < 0.05$, distributions were significantly lower in the resynthesized set. Thus the lengthening of the $F_1$ transition resulted in more homogeneous identifications.

Mann–Whitney tests (Mann and Whitney, 1947) were conducted to establish whether the means of the identifications were different. This test is similar to the independent-samples $t$ test, but it does not assume normal distributions. Compared to the original experiment, the /rl/ identifications were significantly more frequent for the longer-transition stimuli, $U(27,16) = 81.5, p < 0.001$, and the /l/ identifications were significantly less frequent, $U(27,16) = 397.0, p < 0.001$. Longer $F_1$ transitions resulted in more /rl/ identifications, confirming previous studies of /r–l/ trading relations (Polka and Strange, 1985). In addition, the /wl/ identifications were significantly more frequent for the longer transition stimuli, $U(27,16) = 111.5, p < 0.001$. 


