Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling

Paul Iverson and Patricia K. Kuhl
Department of Speech and Hearing Sciences and Virginia Merrill Bloedel Hearing Research Center, University of Washington, Seattle, Washington 98195

(Received 23 September 1993; revised 26 July 1994; accepted 27 July 1994)

Recent experiments have demonstrated that the category goodness of speech sounds strongly influences perception in both adults and infants [Kuhl, Percept. Psychophys. **50**, 93–107 (1991); Kuhl *et al.*, Science **255**, 606–608 (1992)]. Stimuli judged as exceptionally good instances of phonetic categories (prototypes) make neighboring tokens in the vowel space seem more similar, exhibiting a *perceptual magnet effect*. Three experiments further examined the perceptual magnet effect in adults. Experiment 1 collected goodness and identification judgments for 13 variants of the vowel /i/. Experiment 2 used signal detection theory to assess the discrimination of these tokens using a bias-free measure (d'). Experiment 3 employed multidimensional scaling (MDS) to geometrically model the distortion of the perceptual space due to the magnet effect. The results demonstrated a strong relationship between category goodness and discrimination. Vowel tokens receiving high goodness ratings in experiment 1 were more difficult to discriminate in experiment 2 and were more tightly clustered in the MDS solutions of experiment 3. These findings support the existence of a perceptual magnet effect, and may help explain some aspects of first language learning in infants and second language learning in adults.

PACS numbers: 43.71.An, 43.71.Es

One of the challenges of psychological science is to understand how categories are mentally represented (Estes, 1993). Recently, the categorization and representation of phonetic information have received a great deal of attention as new experiments uncover perceptual effects related to exceptionally good exemplars of phonetic categories (prototypes) (Miller and Volaitis, 1989; Volaitis and Miller, 1992; Kuhl, 1991, 1992, 1993a, b; Kuhl *et al.*, 1992). In the psychological literature on categorization, prototypes play a central role (Rosch, 1973; Posner and Keele, 1968). Peoples' abilities to readily identify, classify, and remember a category's best instances are well documented (Garner, 1974; Goldman and Homa, 1977; Mervis and Rosch, 1981; Rosch, 1975, 1977). Studies on the perception of prototypes are thus promising for examining the organization of phonetic categories.

Experiments by Kuhl and her colleagues (1991; Kuhl *et al.*, 1992) have demonstrated that the category goodness of speech sounds strongly influences perception. Listeners exhibit relatively poor discrimination in the region of prototypic exemplars of phonetic categories. Kuhl (1991) synthesized many exemplars of the vowel /i/, as in the word "peep," and had adult American subjects rate the perceived goodness of each vowel on a scale from 1 ("poor") to 7 ("excellent"). Subjects consistently gave highest goodness ratings to tokens in a particular region of the vowel space. On the basis of these goodness ratings, an excellent exemplar, the *prototype* (P), and a poor exemplar, the *nonprototype* (NP), were selected. Thirty-two variants of P and NP (Fig. 1) were synthesized by manipulating the first and second formants of the stimuli while holding the third, fourth, and fifth formants constant. In discrimination experiments, American adults and 6-month-old infants exhibited a *perceptual magnet effect*; P was more difficult to discriminate from its variants than NP was from its variants (Grieser and Kuhl, 1989; Kuhl, 1991; Kuhl *et al.*, 1992). In studies using identical procedures, Rhesus monkeys did not exhibit the perceptual magnet effect (Kuhl, 1991).

Additionally, Kuhl *et al.* (1992) demonstrated that the magnet effect is influenced by exposure to language early in life. American and Swedish 6-month-old infants were tested with the American English /i/ prototype used in the Kuhl (1991) study and with a Swedish front-rounded /y/ prototype. American infants had reduced discrimination for variants of the American English /i/, and Swedish infants had reduced discrimination for variants of the Swedish /y/. The results demonstrated that both groups of infants exhibited the magnet effect for their native-language vowel sound; both groups of infants treated the foreign-language vowel like a nonprototype.

This previous work reveals four points about the perceptual magnet effect. The results demonstrate that the magnet effect is present in 6-month-old infants, the magnet effect is sensitive to early linguistic experience, monkeys tested with the same technique fail to provide evidence of a perceptual magnet effect, and the magnet effect is associated with reduced discrimination sensitivity around a phonetic prototype. This suggests that the perceptual space underlying a phonetic category is distorted such that the perceptual distance around a prototype is reduced.

The goal of the three experiments reported here was to map the perceptual magnet effect with finer detail than had previous studies. Specifically, we directly assessed two implications of the magnet effect: (a) That the area around a phonetic prototype is associated with reduced sensitivity, and (b) that the perceptual space underlying a prototype is dis-

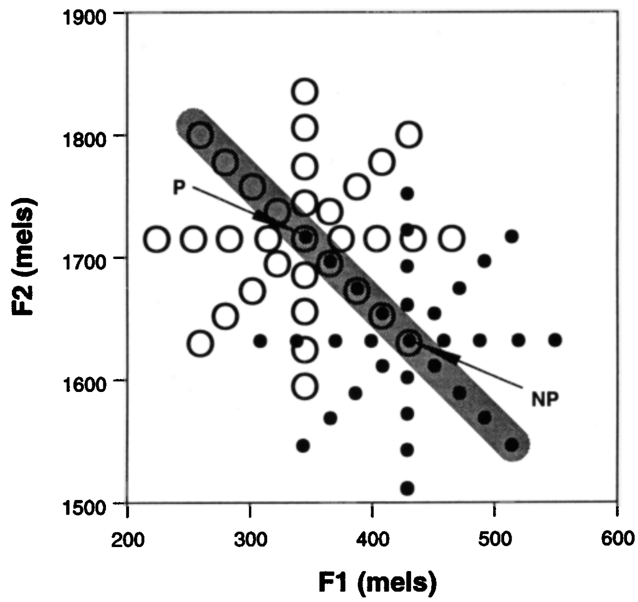

FIG. 1. The Kuhl (1991) prototype /i/ (P) and its 32 variants (open circles) and the nonprototype /i/ (NP) and its 32 variants (filled circles). Adult and infant listeners were worse at discriminating P from its variants than they were at discriminating NP from its variants. The shaded area highlights the set of 13 stimuli used in the present experiments.

torted. For these experiments, 13 tokens (highlighted in Fig. 1) were selected from the set employed by Kuhl (1991). Experiment 1 assessed perceived goodness and phonetic identification for these tokens. Experiment 2 employed a discrimination task and applied signal detection theory to measure sensitivity using a bias-free measure (Green and Swets, 1966; Macmillan *et al.*, 1977; Macmillan and Creelman, 1991). Finally, experiment 3 employed multidimensional scaling (MDS) to model reaction times in a discrimination task (a measure of similarity) and map the distortion of the perceptual space due to the magnet effect.

I. EXPERIMENT 1

This experiment examined the category goodness of the tokens used in all three experiments. Subjects rated the goodness of the tokens on a numerical scale, as in previous studies (Grieser and Kuhl, 1989; Kuhl, 1991). Additionally, subjects identified each token as /i/ (as in the word "he") or /e/ (as in the word "hay"). In Kuhl (1991) subjects did not identify the phonetic category of the tokens. Earlier experiments had indicated that all stimuli were considered members of the /i/ category, but subsequent tests have suggested that subjects may not categorize all of the stimuli as /i/. The goals of this experiment were to collect identification judgments, isolate the location in the vowel space associated with the highest goodness ratings, and measure the accuracy of this location estimate. These results were used for comparison with the results of experiments 2 and 3.

A. Method

1. Subjects

Ten adult members of the University of Washington community participated in this experiment. All were native English speakers with normal hearing, and five subjects had some training in phonetics. Subjects received course credit for participating in this half-hour experiment.

2. Apparatus

The stimuli were presented by a Data Translation DT2821 digital audio board controlled by an NEC 386 microcomputer. The vowels were played to subjects using the right-ear speaker of a pair of Telephonics TDH-39P headphones while subjects sat in a sound-treated booth. Subjects' responses were entered and recorded using the computer that controlled the presentation of stimuli.

3. Stimuli

The stimuli were 13 /i/ tokens from the original set used by Kuhl (1991) (highlighted in Fig. 1). The tokens included the Kuhl (1991) P and NP, and all tokens fell on a single vector through the acoustic space.

The five-formant tokens were synthesized using Klatt's (1980) cascade-parallel speech synthesizer on an NEC 386 microcomputer. The values of the first two formants varied as displayed in Fig. 1, and the third through fifth formants were the same for all stimuli. The frequencies of $F1$ and $F2$ were set so that the tokens fell on a single vector with a 30-mel distance (Stevens *et al.*, 1937) between neighboring stimuli. Fant (1973) has argued that the mel scale is appropriate for vowel stimuli because difference limens for the first three formants (reported by Flanagan, 1957) are similar when measured in mels, and the mel scale corresponds to excitation patterns on the basilar membrane. The mel scale is linear at low frequencies and logarithmic at high frequencies.

The frequency of $F1$ varied from 197 to 429 Hz, $F2$ varied from 1925 to 2489 Hz, and the third through fifth formants were 3010, 3300, and 3850 Hz, respectively for all tokens. The bandwidths of the five formants were 53, 77, 111, 175, and 281 Hz, respectively for all tokens. The $F0$ for all tokens rose from 112 to 130 Hz over the first 100 ms and dropped to 92 Hz over the remaining portion of each vowel. Each vowel was 435 ms long. The stimuli were equalized in intensity, and played to subjects at a comfortable level. Careful listening after the intensity equalization demonstrated that the vowels were equally loud.

4. Procedure

On each trial, subjects identified and rated the goodness of a single token. For the identification task, subjects judged whether the token sounded like the vowel in the word "he" (/i/) or the vowel in the word "hay" (/e/). These two response categories were used because subjects in a pilot experiment with a large number of response categories selected either /i/ or /e/ on 94% of the trials. After making their identification judgments, subjects were asked to rate how good an example the token was of that category on a scale from 1 ("bad") to 7 ("good"). Subjects rated how well the vowel represented the /i/ category if they had made an /i/ identification, and how well the vowel represented the /e/ category

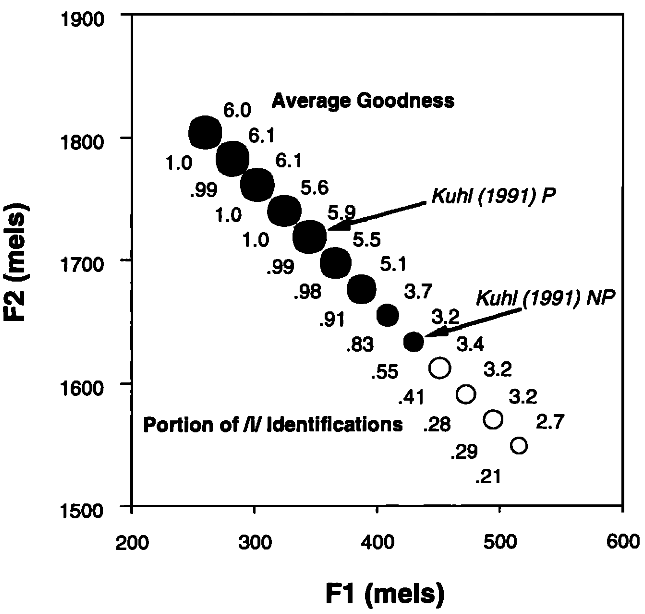

FIG. 2. Average goodness ratings and identification judgments for the tokens in experiment 1. Filled circles indicate greater than 50% /i/ identifications; open circles indicate greater than 50% /e/ identifications. Larger circles indicate high goodness ratings; smaller circles indicate low goodness ratings. The displayed goodness ratings for the /i/ tokens are based only on trials where the tokens were identified as /i/, and the goodness ratings for the /e/ tokens are based only on trials where the tokens were identified as /e/.

if they had made an /e/ identification. Subjects were allowed to hear each token as many times as they needed to make their judgments.

Each subject completed a practice block of 13 trials with each of the 13 tokens presented once in random order. After the practice, subjects completed an experimental session of 104 trials (8 blocks of the 13 tokens) with the order of trials randomized within each block.

B. Results and discussion

Figure 2 displays the average identification and goodness ratings. Nine tokens were identified as /i/ ("he") greater than 50% of the time, and the four tokens on the right end of the stimulus series were more often identified as /e/ ("hay"). To assess the variability of the identification ratings, the percentage of /i/ identifications was calculated separately for each block of trials for each subject. These measures were analyzed using an ANOVA with block and subject as independent factors. There was a significant effect of subject, $F(9{,}63)=37.568$, $p<0.001$, indicating that subjects differed in their overall percentage of /i/ identifications. The percentage of /i/ judgments for different subjects ranged from 47% to 100%, and the average was 73%. There was also a significant effect of block, $F(7{,}63)=2.627$, $p<0.05$; the frequency of /i/ identifications rose by about 10% from the start to the end of the experiment. Identification responses were influenced both by subject differences and presentation order.

It is surprising that some of the tokens in this set received less than 50% /i/ identifications considering that all of the subjects in Kuhl (1991) informally reported that they heard all of these tokens as /i/. Three factors might explain this difference. First, the overall percentage of /i/ identifications varied with subject in experiment 1; it is possible that the group of subjects in Kuhl (1991) happened to include more subjects who identify all of these tokens as /i/. Second, identification responses could vary with the phonetic training of the listeners; Kuhl's (1991) subjects all had some phonetic training. Finally, it is possible that identification responses are influenced by the stimulus set. Previous research has suggested that the identification of vowels near boundaries can be influenced by context (Fry *et al.*, 1962; Eimas, 1963; Stevens and Öhman, 1969; Nearey, 1989), so the restricted range of stimuli in this set may have influenced the location of the /i-e/ boundary.

The goodness results of experiment 1 roughly corresponded to those of Kuhl (1991); stimuli near the Kuhl (1991) P received consistently higher goodness ratings than stimuli near the Kuhl (1991) NP. However, the best stimulus location for experiment 1 was 75 mels to the left of the Kuhl (1991) P. To assess the accuracy of this location estimate, the location of the prototype was estimated for each block of stimuli for each subject. The prototype location was defined as the location of the token with the highest goodness rating, or as the location between tokens when multiple tokens received the same highest rating. An ANOVA was run on these prototype location estimates with block and subject as independent factors. There was no significant effect of subject, $F(9{,}63)=0.706$, $p>0.05$, indicating that all subjects had a similar prototype location. Additionally, there was no significant influence of block, $F(7{,}63)=0.971$, $p>0.05$, indicating that the prototype location did not vary systematically across the eight blocks of trials. In contrast to identification judgments, the estimate of the prototype location was consistent across subjects and trials.

To further measure the accuracy of the prototype location estimates for individual subjects, the standard error (SE) of the location estimate for each subject was calculated using the estimate for each block of trials. The average SE for individual subjects$=$10.9 mels, so the prototype estimate for the average subject was ± 25.8 mels, $t(7)=2.365$, $p<0.05$, of the true prototype. Considering that the tokens were 30 mels apart and that the perceptual difference between neighboring tokens was small, the prototype location estimates appear to be quite reliable. The good accuracy of the prototype location estimate suggests that the differences between the prototype locations of Kuhl (1991) and experiment 1 may not be due to random subject differences. It appears that subjects consistently judge that certain stimuli are the best exemplars, but the location of the best stimuli may be influenced by the stimulus set.

II. EXPERIMENT 2

This experiment was designed for analysis within the framework of signal detection theory (Green and Swets, 1966) as extended to the same–different discrimination paradigm commonly used in speech perception research (Macmillan and Creelman, 1991). In this experiment, subjects heard two stimuli on each trial and judged whether they were same or different. A comparison of the percentage of hits ("different" responses when the stimuli were different) and false alarms ("different" responses when the stimuli were the same) allowed for orthogonal measurement of the sensitivity (d') and response bias for each pair of tokens (Macmillan and Creelman, 1991).

In Kuhl's (1991) tests of the magnet effect, a *go/no-go* psychophysical technique was selected so that adults, 6-month-old infants, and Rhesus monkeys could be tested with only minor adjustments to procedures. With this technique, subjects monitored a constantly repeating background stimulus for a change in the sound. In tests on all three populations, there were 50% *test* trials and 50% *control* trials. On test trials the stimulus was changed during a 4.5-s observation interval and subjects were rewarded for detecting the change (adults and infants with a visual stimulus; monkeys with applesauce). On control trials the stimulus was not changed and subjects were monitored to assess the probability of false positive responses. This procedure measured the four outcomes of discrimination experiments (hits, correct rejections, misses, and false alarms) and allowed calculation of an overall percentage correct measure based both on test and control trials (hits+correct rejections/2). However, the decision criteria of this technique has not been studied systematically, making it difficult to reliably estimate sensitivity and response bias within the context of signal detection theory.

The present experiment was designed to evaluate the results of Kuhl (1991) using a bias-free discrimination measure, and to further test the relationship between goodness and discrimination. In separate blocks of trials, listeners discriminated tokens from the Kuhl (1991) P and NP. A comparison of the P and NP blocks paralleled Kuhl's (1991) test of the magnet effect. However, experiment 1 demonstrated that the NP stimulus is near a category boundary for these tokens, so a comparison of the P and NP blocks would not examine whether the perceptual space is distorted *within* the /i/ category. To more effectively examine discrimination within the category, experiment 2 compared discrimination for tokens to the left of P and to the right of P. Experiment 1 demonstrated that tokens immediately to the left of P were better exemplars than those to the right, suggesting that discrimination should be worse to the left even though tokens on both sides of P are within the /i/ category.

A. Method

1. Subjects

Eleven members of the University of Washington community participated in this experiment. One subject was dropped from the subsequent analysis because he performed at chance. All subjects were native English speakers with normal hearing, and none had training in phonetics. Subjects received course credit for participating in this half-hour experiment.

2. Apparatus

A PDP 11/73 computer controlled the presentation of stimuli and recorded subjects' responses. As in experiment 1, the stimuli were played to subjects using the right-ear speaker of a pair of Telephonics TDH-39P headphones.

3. Stimuli

The tokens were 11 of those used in experiment 1. All but the tokens at each end of the experiment 1 stimulus vector were used in this experiment.

4. Procedure

On each trial, subjects heard two stimuli with a 250-ms ISI. Subjects initiated a trial by pressing a response key, and they continued to hold down the key during the presentation of stimuli. Subjects were instructed to lift the response key when they thought the stimuli were different. If they thought the stimuli were the same, they continued pressing the key until the computer signaled that the trial was over (2 s after the start of the second token).

Subjects completed 2 experimental blocks of trials. In one block, subjects heard a random ordering of trials composed of the Kuhl (1991) P stimulus paired with the tokens 30, 60, and 90 mels away in both directions on the stimulus series. In another block, subjects heard a random ordering of trials composed of the Kuhl (1991) NP stimulus paired with the tokens 30, 60, and 90 mels away in both directions on the stimulus series. Half of the subjects completed the P block first, and half completed the NP block first.

There were 120 *same* trials and 120 *different* trials in each block. The P or NP stimulus (depending on the block) was presented on 60 same trials and each of the other six stimuli were presented on ten same trials. There were 20 different trials for each of the six pairs of stimuli. On half of the different trials, the P or NP stimulus (depending on the block) was presented first in the pair; on the other trials it was presented second.

Before the experimental trials, subjects completed a short practice session. The trials were the same as in the experimental blocks, except that subjects received feedback after each trial. Each subject heard 12 control trials and 12 change trials presented in a random order. Half of the subjects heard the P stimulus for the practice block, and half of the subjects heard the NP stimulus for the practice block.

B. Results and discussion

Measures of d' (Kaplan *et al.*, 1978; Macmillan and Creelman, 1991) were calculated for each interval for each subject; the average d' scores are displayed in Fig. 3 and Table I. These d' measures were analyzed using paired t-tests. Subjects were significantly worse, $t(9)=-3.07$, $p<0.05$, at discriminating stimuli from P (mean $d'=2.75$) than from NP (mean $d'=3.46$). This replicates the basic finding of Kuhl (1991) using a bias-free measure, and further verifies that subjects have greater sensitivity to acoustic differences near the category boundary (NP) than within the vowel category (P). Additionally, discrimination was significantly lower, $t(9)=-2.96$, $p<0.05$, to the left of P (mean $d'=2.27$) than to the right of P (mean $d'=3.22$). Experiment 1 had indicated that all of these stimuli were identified as /i/ on at least 83% of the trials, but tokens to the left of P received higher goodness ratings than those to the right. This confirms that category goodness influences discrimination *within* the category. There was no significant difference, $t(9)$

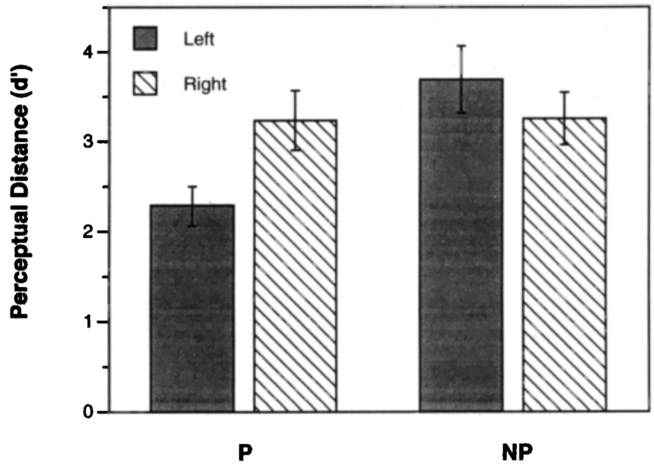

FIG. 3. Average d' scores for experiment 2. Error bars indicate ± 1 se. Subjects were worse at discriminating stimuli from P than from NP. Additionally, subjects were worse at discriminating stimuli to the left of P than they were at discriminating stimuli to the right of P, coinciding with the goodness judgments of experiment 1. No significant differences were found for stimuli to the left and right of NP.

$=-1.67$, $p>0.05$, between discrimination to the left of NP (mean $d'=3.68$) and to the right of NP (mean $d'=3.24$).

Although these results agree with the main findings of Kuhl (1991), Grieser and Kuhl (1989) and Kuhl (1991) additionally found discrimination differences for the tokens on the stimulus vector between P and NP. Infants (though not adults) were worse at discriminating these "shared vector" tokens from P than from NP. In the present study, the discrimination of stimuli to the right of P (mean $d'=3.22$) was worse than the discrimination of stimuli to the left of NP (mean $d'=3.68$), but this difference only approached significance, $t(9)=-1.79$, $p=0.11$. The distortion of the perceptual space between P and NP may be less reliable for adult listeners than for infants.

III. EXPERIMENT 3

Experiment 3 employed multidimensional scaling (Shepard 1962a, b) to map the correspondence between discrimination and goodness with finer detail. MDS has been effective at modeling the perceptual similarity of vowels in previous experiments (Pols *et al.*, 1969; Singh and Woods, 1971; Shepard, 1972; Terbeek, 1977; Fox, 1982, 1983, 1985; Kewley-Port and Atal, 1989). The perceptual similarity of vowel pairs can be assessed using a variety of psychological measures including confusions in identification (more frequent confusions for similar vowels), similarity ratings on numerical scales, and reaction-time measures in discrimination tasks (longer RTs for more similar vowels). MDS is used to assign each vowel to a point in a geometric space so that distances in the space correspond to perceived similarity; vowels placed close in MDS solutions are more similar than vowels far apart. Mapping vowel similarity to a geometric space reveals relationships among the vowels that would not be readily apparent from raw similarity measures.

TABLE I. Average perceptual distance (d') for each interval in experiment 2.

Interval location	Interval length 30 mels	60 mels	90 mels
Left of P	1.32	2.18	3.32
Right of P	1.47	3.81	4.39
Left of NP	2.30	4.09	4.66
Right of NP	2.27	3.64	4.59

Most previous experiments have employed vowels from different phonetic categories, and they have demonstrated a high correspondence between acoustic differences and distances in MDS solutions (Pols *et al.*, 1969; Shepard, 1972; Terbeek, 1977; Fox, 1982, 1983, 1985). Dimensions in MDS spaces have corresponded quite closely to acoustic measurements of vowels ($F1$, $F2$, and $F3$) so that MDS solutions tend to match traditional acoustic maps of vowel inventories. Additionally, Kewley-Port and Atal (1989) demonstrated that subphonemic differences among vowels are effectively mapped using MDS. Acoustic differences among stimuli of the same phonetic category were represented in MDS solutions as well as acoustic differences among stimuli of different phonetic categories.

The experiment reported here also examined the similarity of vowels from the same phonetic category. Experiment 2 and previous studies (Kuhl, 1991; Kuhl *et al.*, 1992) had suggested that good stimuli (prototypes) make neighboring stimuli seem more similar, effectively shrinking the perceptual space underlying prototypes. The main goal of this experiment was to map this distortion of the perceptual space in finer detail than in previous studies. The experiment used MDS to model subjects' reaction times when discriminating pairs of vowel tokens used in the previous two experiments. A related technique has been effectively used by Nosofsky (1984, 1986) to examine distortions of similarity in classification tasks. Additionally, this experiment varied ISI in an initial attempt to evaluate the influence of memory on the magnet effect.

Modeling RTs allows for interesting comparisons to the accuracy measure of experiment 2. Pisoni and Tash (1974) found that within-category acoustic differences influence RTs more than discrimination accuracy. When subjects failed to detect a difference between acoustically different tokens of the same phonetic category (/ba/ or /pa/), they still took longer to respond than they had on trials with identical tokens. This suggests that RTs are more sensitive to acoustic differences than are percentage correct measures, so it is possible that an assessment of the perceptual magnet effect using RTs may differ from one using discrimination accuracy.

The MDS design also differs from experiment 2 and previous studies (Kuhl, 1991; Kuhl *et al.*, 1992) because each token is presented the same number of times throughout the experiment and subjects hear all possible pairs of tokens. Previous investigations of the perceptual magnet effect presented P and NP more frequently than their variants, and the P and NP trials were presented in separate blocks. Frequently presented tokens can distort perception through adaptation (Miller *et al.*, 1983; Samuel, 1982), but the MDS design examines the distortion of the perceptual space in the absence of presentation frequency differences.

A. Method

1. Subjects

Eighteen adult members of the University of Washington community participated in this experiment. All were native English speakers with normal hearing, and none had training in phonetics. Subjects received course credit for participating in this 1-h experiment.

2. Stimuli

The stimuli were the same as in experiment 1.

3. Apparatus

The apparatus was the same as in experiment 2.

4. Procedure

Subjects heard all possible pairs of the 13 tokens at 3 different ISIs (25, 250, and 2500 ms) and judged whether the tokens in each pair were the same or different. As in experiment 2, listeners initiated a trial by pressing a response key, and they continued to press the key during the presentation of stimuli. If they thought the tokens were different, they immediately stopped pressing the key. If they thought the tokens were the same, they continued to press the key until the computer signaled that the trial was over (2 s after the onset of the second token). Their response for each trial (same or different) and RT (on trials with different responses) were recorded.

The trials were blocked by ISI, with the order of blocks counterbalanced between subjects. In each experimental block, subjects heard 52 *same* trials and 156 *different* trials in a random order. The 52 same trials consisted of 4 trials for each of the 13 tokens. The 156 different trials consisted of each possible pair of the 13 tokens.

Before the experimental trials, subjects completed a short practice session. The trials were the same as in the experimental session except that subjects received feedback after each trial. There were ten same trials and 20 different trials chosen randomly from the trials used in the experimental session. Each subject heard one ISI for the entire practice session, with the ISI counterbalanced between subjects.

B. Results and discussion

Each subject's responses were put into the form of three triangular matrices (one matrix for each ISI). Each triangular matrix was a list of the log RT for the discrimination of each pair of tokens averaged across presentation order. Responses to same trials were not considered in the analyses (subjects false-alarmed on 31% of these trials). Subjects correctly detected that the stimuli were different on 77% of the trials. The average percentage of errors on different trials was significantly correlated with the average RTs for the 25 ms ISI trials, $r=-0.678$ ($df=76$), $p<0.001$, the 250 ms ISI trials, $r=-0.767$ ($df=76$), $p<0.001$, and the 2500 ms ISI trials, $r=-0.784$, ($df=76$) $p<0.001$. Intersubject correlations examined the consistency of reaction times among subjects for all three matrices. Of the 153 intersubject correlations, the average was $r=0.55$ ($df=232$), and each was significant at the $p<0.001$ level. Thus reaction times were highly consistent among subjects.

The log RTs were averaged across subjects to form three triangular matrices (one for each ISI), and these matrices were analyzed separately using the Kruskal (1964a, b) MDS algorithm implemented by the SYSTAT computer program (Wilkinson, 1989). The duration of the trial (2000 ms) was used as the reaction time on trials when subjects failed to detect a difference between tokens. This was necessary to ensure that the average RTs for the most similar tokens were not based only on the responses of subjects who were most sensitive to acoustic differences. Preliminary analyses demonstrated that the pattern of results was similar when a 2000 ms RT was not used for these trials, but analyses using this correction accounted for more variance. The MDS analyses used Kruskal's stress formula 1, a Euclidian distance metric, and a linear regression function. This placed the tokens in a one-dimensional space where the distances between tokens were fit to a linear function of the log RTs. A linear function was used instead of a more traditional monotonic function to avoid degenerate solutions (artificially strong clustering caused by some nonlinear distance functions), and because preliminary analyses indicated that there was a strong linear relationship between acoustic distance and log RT. The MDS solutions modeled the responses with a stress of 0.298 ($R^2=0.724$) for 25 ms ISI, a stress of 0.242 ($R^2=0.812$) for 250 ms ISI, and a stress of 0.147 ($R^2=0.937$) for 2500 ms ISI. Figure 4 displays the MDS solutions.

The acoustic locations of stimuli on the single vector roughly corresponded to their locations in the one-dimensional MDS solutions, and the perceptual magnet effect was apparent at all three ISIs. The good /i/ tokens were clustered more tightly than the poor /i/ tokens, supporting both the findings of Kuhl (1991) and experiment 2. The prototype acted like a perceptual magnet by drawing tokens toward the prototype in the perceptual space. Additionally, tokens most often identified as /e/ in experiment 1 were clustered more tightly than the poor /i/ tokens, although the clustering was not as tight as for the good /i/ tokens. This suggests that tokens on the right end of the stimulus series may have been approaching a location in vowel space for good /e/ tokens, even though none of the tokens in the experiment were excellent exemplars of that category. The MDS solutions were similar for all three ISIs, although there appears to be somewhat stronger clustering at the longest ISI (2500 ms) for the good /i/ tokens and the /e/ tokens.

An analysis of covariance further supported the perceptual magnet effect. The three triangular matrices of reaction times were the dependent variable, with pairs that had a token most often identified as /e/ in experiment 1 excluded from this analysis. The midpoint on the stimulus series for each pair of stimuli was calculated, and the distance (in mels) from this midpoint to the best stimulus in experiment 1 was an independent continuous variable. In addition, ISI was coded as a three-level (25, 250, and 2500 ms) categorical variable, and the distance between each pair of stimuli (in mels) was a continuous variable. The solution accounted for a substantial portion of variance, $R^2=0.801$. There was a

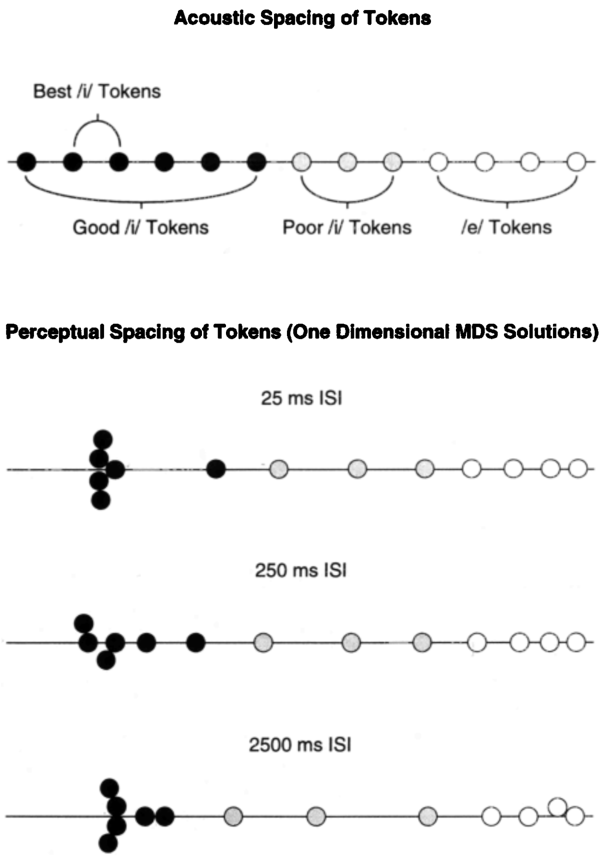

FIG. 4. Acoustic and perceptual spacing of tokens in experiment 3. The best /i/ tokens (as rated in experiment 1) are black (average goodness=6.1), the good /i/ tokens (average goodness >5.3) are dark gray, and the poor /i/ tokens (average goodness <5.3) are light gray. The unfilled circles represent the tokens that received more than 50% /e/ identifications in experiment 1. Horizontal positions of tokens in the perceptual graphs correspond to values in the one-dimensional MDS solutions, but vertical displacement was sometimes necessary to prevent overlap of dots that were close in the MDS solutions. Although the tokens equally divided the acoustic space, the perceptual space was distorted. Perceptual space was *shrunk* in the region of best instances of /i/ and *stretched* in the region of the poorest instances. This occurred at all three interstimulus intervals (ISI).

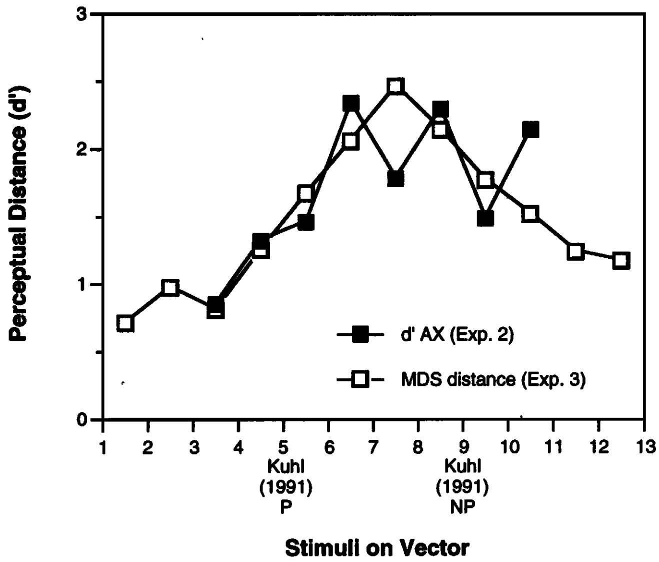

FIG. 5. Perceptual distance between adjacent tokens in experiments 2 and 3; data points are plotted midway between tokens. The distances between tokens in experiment 2 are in d'. The MDS distances of experiment 3 are from the 250 ISI condition, and have been scaled for this graph so that the mean and SD of the distances match those of experiment 2. The results of these experiments are quite similar; both demonstrate relatively poor discrimination near the best /i/ tokens and relatively good discrimination for the worst /i/ tokens.

significant influence of acoustic distance from the midpoint of each pair to the best stimulus location, $F(1,101)=19.46$. $p<0.001$. This further demonstrates that tokens acoustically proximate to the best stimulus were perceptually clustered. There was also a significant main effect of the acoustic distance between tokens, $F(1,101)=218.25$, $p<0.001$, strongly supporting a relationship between RTs and acoustic differences. Additionally, there was a significant influence of ISI, $F(2,101)=22.55$, $p<0.001$, demonstrating that subjects were slower with the 2500 ms ISI (average $\log \mathrm{RT}=2.984$) than they were with the 250 ms ISI (average $\log \mathrm{RT}=2.903$) or the 25 ms ISI (average $\log \mathrm{RT}=2.841$). There was no significant interaction between ISI and distance from the prototype, $F(2,101)=0.81$, $p>0.05$.

The perceptual distance between adjacent tokens was calculated from the results of experiments 2 and 3 to allow for a more direct comparison of the two experiments. The perceptual distance between adjacent tokens in experiment 2 can be obtained by subtraction because d' measured perceptual distance on a single dimension for this stimulus set (Macmillan and Creelman, 1991). For example, subtracting the d' value for the 30 mel interval to the left of P from the value for the 60 mel interval to the left of P estimates the perceptual distance separating the 30- and 60-mel tokens. Unfortunately, the 90-mel intervals for both the P and NP blocks approached a ceiling in discrimination performance, so a calculation using these intervals would underestimate the perceptual distance between adjacent tokens. As a result, the estimates of perceptual distance were based only on the 30- and 60-mel intervals from experiment 2. The distance between tokens in experiment 3 was obtained by finding the distance between adjacent tokens in the 1-dimensional solution for the 250 ms ISI trials (the ISI used in experiment 2). The results are displayed in Fig. 5 with the MDS distances scaled to match the mean and standard deviation of the experiment 2 results.

The results from experiments 2 and 3 were significantly correlated, $r=0.720$ ($df=6$), $p<0.05$, demonstrating that the perceptual distance estimates for the two experiments were similar. Although this comparison indicates that the overall patterns of results were related, the absolute magnitude of the perceptual distances in each experiment cannot be directly compared. The MDS technique does not provide an absolute measure of perceptual distance because MDS solutions are normalized based on the distribution of tokens. Thus it cannot be determined whether the magnitude of the distortion due to the magnet effect was the same in both experiments. However, the results suggest that the *relative* perceptual distortion along the stimulus series was measured similarly by the two experimental techniques.

IV. GENERAL DISCUSSION

The three experiments reported here provide additional support for Kuhl's (1991) finding of a perceptual magnet effect for the best instances of phonetic categories. Experiment 1 demonstrated that all vowels are not perceived to be equally good exemplars of their phonetic category. Subjects consistently judged that certain tokens were best exemplars. Experiment 2 extended Kuhl's (1991) tests by showing that subjects have reduced sensitivity near prototypic stimuli independent of response bias. Subjects had worse discrimination within the /i/ category than at the boundary, and discrimination within the category was worst for the best exemplars. Experiment 3 used multidimensional scaling to map the distortion of the perceptual space in the region of the prototype. Perceptual distances appeared to be shrunk in the region where the best instances occur, and stretched in the region where the worst instances occur. The results demonstrate a perceptual magnet effect. Speech perception is not equivalent within phonetic categories (Grieser and Kuhl, 1989; Kuhl, 1991) and the best instances of the category are associated with reduced discrimination and perceptual clustering.

The results also suggest that the perceptual magnet effect may be influenced by experimental context. The location of the best stimuli and highest perceptual clustering in these experiments was at a more extreme location in the vowel space than for Kuhl (1991). It is uncertain exactly what influenced this difference in prototype locations. Although the location of the best stimulus seems determined by long-term exposure to language (Kuhl *et al.*, 1992), it may also be influenced by the experimental design or by the set of stimuli used in particular experiments.

Having established the perceptual magnet effect at an empirical level, it is useful to relate it to previous findings in speech perception. The perceptual magnet effect can be distinguished from the well-known finding that there is relatively good discrimination for vowels at phonetic boundaries and poor discrimination within vowel categories, as shown in studies of categorical perception (Stevens and Öhman, 1969; Pisoni, 1973, 1975; Repp *et al.*, 1979). The perceptual magnet effect and categorical perception are similarly marked by discrimination peaks at identification boundaries. However, categorical perception predicts that the discrimination of identically labeled stimuli should be equally poor, and the perceptual magnet effect predicts that the discrimination of identically labeled stimuli should be influenced by category goodness. The results fit the latter prediction: Discrimination is worst for good exemplars and best for poor exemplars, even for tokens that are consistently identified as members of a single category.

It is also of theoretical interest to relate the perceptual magnet effect to the data and theorizing of Macmillan *et al.* (1988). These authors have extended the intensity discrimination model of Durlach and Braida (1969) to suggest that listeners may use *perceptual anchors* (stimuli that are easy to label consistently) for discriminating speech stimuli. When there is a substantial degree of stimulus variability, the authors argue that listeners may store an estimate of how distant each token is from a particular easy-to-remember stimulus instead of encoding every acoustic detail of each token. This strategy leads to increased discrimination near perceptual anchors and poor discrimination far from perceptual anchors, because the distance estimates become more variable as distances increase. Macmillan *et al.* (1988) have proposed that listeners in vowel perception experiments may have perceptual anchors located at phonetic boundaries since they tend to coincide with peaks in discrimination. The predictions made by this theory for the /i/ category are isomorphic with those of the perceptual magnet effect because the category's best instances are located far from the boundary region. Thus both theories predict that discrimination should be low near the best instances of the vowel category and should increase toward vowel boundaries. Perceptual magnets and perceptual anchors may be similar constructs.

The present results suggest that the perceptual magnet effect may be insensitive to experimental manipulations that influence the encoding of tokens in memory. Differences in ISI did not reliably influence the magnet effect in experiment 3, and the pattern of results in experiments 2 and 3 were similar even though the MDS task employed a much wider range of stimulus pairs. However, the parallels to Macmillan *et al.*'s (1988) findings suggest that the perceptual magnet effect may be diminished in tasks where a much narrower range of stimuli is employed. For example, when subjects only compare two tokens within each block (a fixed discrimination task), the peaks in discrimination at the boundaries tend to diminish (at least for vowels). This finding led Macmillan *et al.* (1988) to conclude that the peaks in discrimination seen in categorical perception studies using vowels are not based on sensory factors. It remains for future experiments to determine whether the perceptual magnet effect is diminished when fixed discrimination procedures are used.

Kuhl (1992, 1993a, b) has incorporated the perceptual magnet effect into the native language magnet (NLM) model of speech perception. The implications of the model extend both to infants learning their first language and to adults learning a second language. Regarding infants, previous results demonstrate that the perceptual magnet effect is influenced by exposure to a particular language in early infancy (Kuhl *et al.*, 1992). This suggests that the perceptual magnet effect structures infants' perception of language prior to the acquisition of word meaning. The shrinking and stretching of the underlying perceptual space could facilitate the acquisition of words from a particular language: Perceptual shrinking around the prototype would effectively reduce infants' sensitivities to the acoustic differences that *were not* phonemically relevant, and perceptual stretching around nonprototypes would highlight the differences that *were* phonemically relevant. This would occur prior to the time that the child's understanding of phonological contrast develops and could assist in the development of a language-specific phonology.

The implication for adults' perception of speech is equally strong, and from the perspective of second-language learning, less helpful. The NLM model predicts that the restructuring of perceptual space that assists infants in the acquisition of phonology can adversely affect adults' subsequent ability to perceive foreign-language distinctions. Adult listeners have demonstrated a difficulty in perceiving some of the phonetic contrasts that are not used in their native language; the case of Japanese listeners' difficulty with the English /r-l/ contrast has been well documented (Goto, 1971; Logan *et al.*, 1991; Miyawaki *et al.*, 1975; Strange and Dittmann, 1984). NLM argues that exposure to a primary language distorts the underlying perceptual space by reducing sensitivity near phonetic prototypes, and that these perceptual effects can be difficult to alter (Kuhl and Iverson, in press). The model predicts that adults learning a second language would find it difficult to perceive a phonetic contrast from a new language when the sounds are proximate to a native-language prototype (see Best, 1993, for a related argument). Phonetic prototypes of Japanese listeners may thus interfere with their perception of the English /r-l/ contrast.

Although the data presented here support the existence of the perceptual magnet effect, they do not go further to reveal the underlying category representation that is hypothesized as the cause of this effect (Kuhl, 1992, 1993a, b). In the literature on cognitive categories, two models of category representation are currently being compared (Estes, 1993). In one case, the category representation is thought to be an abstract statistical summary of all the exemplars a person has experienced; in the other, category representation consists of individual instances of the category stored in memory. Effects of typicality can be explained by either type of representation (Estes, 1993; Medin and Barsalou, 1987), and this is true for the magnet effect as well. Concerning speech, therefore, we do not take a position as to whether phonetic category information is stored in terms of an abstract summary or as individual instances. The perceptual magnet effect, as tested thus far, does not distinguish the two alternatives in this debate (see Kuhl, 1993a,b for further discussion). We also note that the human perceptual system may have access to both types of information (see, e.g., Knowlton and Squire, 1993).

The three studies described here strongly support Kuhl's (1991) perceptual magnet effect for the prototypes of phonetic categories. The present results extend Kuhl's findings by measuring the magnet effect using a bias-free estimate of sensitivity and geometrically modeling the effect using multidimensional scaling. The experiments suggest that the best instances of vowel categories are associated with decreased sensitivity and a warping of the perceptual space underlying the category. Further study is needed to examine the role of experimental context on this effect and extend these findings to additional phonetic categories. Toward that end, two new studies in our laboratory provide preliminary evidence that the perceptual magnet effect will also characterize the perception of consonant stimuli (Iverson and Kuhl, 1994; Davis and Kuhl, 1994).

ACKNOWLEDGMENTS

The research described in this report was supported by a grant to Patricia K. Kuhl by NIH (DC 00520). The data were presented at the 125th meeting of the Acoustical Society of America, May, 1993. We are grateful to Neil Macmillan and Winifred Strange for their comments on a previous version of this manuscript, and Charmaine Morton for her assistance in conducting the experimental sessions.

Best, C. T. (1993). "Language-specific developmental changes in non-native speech perception: A window on early phonological development," in *Developmental neurocognition: Speech and Face Processing in the First Year of Life*, edited by B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, and J. Morton (Kluwer Academic, Boston), pp. 289–304.

Davis, K., and Kuhl, P. K. (1994). "Tests of the perceptual magnet effect for American English /k/ and /g/," J. Acoust. Soc. Am. **95**, 2976 (Pt. 2).

Durlach, N. I., and Braida, L. D. (1969). "Intensity perception. I. Preliminary theory of intensity resolution," J. Acoust. Soc. Am. **46**, 372–383.

Eimas, P. D. (1963). "The relation between identification and discrimination along speech and non-speech continua," Lang. Speech **6**, 206–217.

Estes, W. K. (1993). "Concepts, categories, and psychological science," Psychol. Sci. **4**, 143–153.

Fant, G. (1973). *Speech Sounds and Features* (MIT, Cambridge, MA).

Flanagan, J. L. (1957). "Estimates of the maximum precision necessary in quantizing certain "dimensions" of vowel sounds," J. Acoust. Soc. Am. **29**, 533–534.

Fox, R. A. (1982). "Individual variation in the perception of vowels: Implications for a perception-production link," Phonetica **39**, 1–22.

Fox, R. A. (1983). "Perceptual structure of monophthongs and diphthongs in English," Lang. Speech **26**, 21–60.

Fox, R. A. (1985). "Auditory contrast and speaker quality variation in vowel perception," J. Acoust. Soc. Am. **77**, 1552–1559.

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. (1962). "The identification and discrimination of synthetic vowels," Lang. Speech **5**, 171–189.

Garner, W. R. (1974). *The Processing of Information and Structure* (Erlbaum, Potomac, MD).

Goldman, D., and Homa, D. (1977). "Integrative and metric properties of abstracted information as a function of category discriminability, instance variability, and experience," J. Exp. Psychol.: Hum. Learn. Mem. **3**, 375–385.

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

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Grieser, D., and Kuhl, P. K. (1989). "Categorization of speech by infants: Support for speech-sound prototypes," Dev. Psychol. **25**, 577–588.

Iverson, P., and Kuhl, P. K. (1994). "Tests of the perceptual magnet effect for American English /r/ and /l/," J. Acoust. Soc. Am. **95**, 2976 (Pt. 2).

Kaplan, H. L., Macmillan, N. A., and Creelman, C. D. (1978). "Tables of d' for variable-standard discrimination paradigms," Behav. Res. Methods Instrum. **10**, 796–813.

Kewley-Port, D., and Atal, B. S. (1989). "Perceptual differences between vowels located in a limited phonetic space," J. Acoust. Soc. Am. **85**, 1726–1740.

Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Knowlton, B. J., and Squire, L. R. (1993). "The learning of categories: Parallel brain systems for item memory and category knowledge," Science **262**, 1747–1749.

Kruskal, J. B. (1964a). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," Psychometrika **29**, 1–27.

Kruskal, J. B. (1964b). "Nonmetric multidimensional scaling: a numerical method," Psychometrika **29**, 115–129.

Kuhl, P. K. (1991). "Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not," Percept. Psychophys. **50**, 93–107.

Kuhl, P. K. (1992). "Infants' perception and representation of speech: Development of a new theory," in *Proceedings of the International Conference on Spoken Language Processing*, edited by J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, and G. E. Wiebe (University of Alberta, Edmonton, Alberta), pp. 449–456.

Kuhl, P. K. (1993a). "Infant speech perception: A window on psycholinguistic development," Int. J. Psycholing. **9**, 33–56.

Kuhl, P. K. (1993b). "Innate predispositions and the effects of experience in speech perception: The native language magnet theory," in *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*, edited by B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, and J. Morton (Kluwer Academic, Boston), pp. 259–274.

Kuhl, P. K., and Iverson, P. (in press). "Linguistic experience and the "perceptual magnet effect"," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research*, edited by W. Strange (York, Timonium, MD).

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (**1992**). "Linguistic experience alters phonetic perception in infants by 6 months of age," Science **255**, 606–608.

Logan, J. S., Lively, S. E., and Pisoni, D. B. (**1991**). "Training Japanese listeners to identify English /r/ and /l/: A first report," J. Acoust. Soc. Am. **89**, 874–886.

Macmillan, N. A., and Creelman, C. D. (**1991**). *Detection Theory: A User's Guide* (Cambridge U.P., New York).

Macmillan, N. A., Goldberg, R. F., and Braida, L. D. (**1988**). "Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua," J. Acoust. Soc. Am. **84**, 1262–1280.

Macmillan, N. A., Kaplan, H. L., and Creelman, C. D. (**1977**). "The psychophysics of categorical perception," Psychol. Rev. **84**, 452–471.

Medin, D. L., and Barsalou, L. W. (**1987**). "Categorization processes and categorical perception," in *Categorical Perception: The Groundwork of Cognition*, edited by S. Harnad (Cambridge U.P., New York), pp. 455–490.

Mervis, C. B., and Rosch, E. (**1981**). "Categorization of natural objects," Ann. Rev. Psychol. **32**, 89–115.

Miller, J. L., Connine, C. M., Schermer, T. M., and Kluender, K. R. (**1983**). "A possible auditory basis for internal structure of phonetic categories," J. Acoust. Soc. Am. **73**, 2124–2133.

Miller, J. L., and Volaitis, L. E. (**1989**). "Effect of speaking rate on the perceptual structure of a phonetic category," Percept. Psychophys. **46**, 505–512.

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

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.

Nosofsky, R. M. (**1984**). "Choice, similarity, and the context theory of classification," J. Exp. Psychol.: Learn. Mem. Cog. **10**, 104–114.

Nosofsky, R. M. (**1986**). "Attention, similarity, and the identification-categorization relationship," J. Exp. Psychol.: Gen. **115**, 39–57.

Pisoni, D. B. (**1973**). "Auditory and phonetic memory codes in the discrimination of consonants and vowels," Percept. Psychophys. **13**, 253–260.

Pisoni, D. B. (**1975**). "Auditory short-term memory and vowel perception," Mem. Cognit. **3**, 7–18.

Pisoni, D. B., and Tash, J. (**1974**). "Reaction times to comparisons within and across phonetic categories," J. Acoust. Soc. Am. **15**, 285–290.

Pols, L. C. W., van der Kamp, L. J. T., and Plomp, R. (**1969**). "Perceptual and physical space of vowel sounds," J. Acoust. Soc. Am. **46**, 458–467.

Posner, M. I., and Keele, S. W. (**1968**). "On the genesis of abstract ideas," J. Exp. Psychol. **77**, 353–363.

Repp, B. H., Healy, A. F., and Crowder, R. G. (**1979**). "Categories and context in the perception of isolated steady-state vowels," J. Exp. Psychol.: Hum. Percept. Perform. **5**, 129–145.

Rosch, E. (**1975**). "Cognitive reference points," Cognit. Psychol. **7**, 532–547.

Rosch, E. H. (**1973**). "On the internal structure of perceptual and semantic categories," in *Cognitive Development and the Acquisition of Language*, edited by T. E. Moore (Academic, New York), pp. 111–144.

Rosch, E. H. (**1977**). "Human categorization," in *Studies in Cross-Cultural Psychology, Vol. 1*, edited by N. Warren (Academic, San Francisco), pp. 1–49.

Samuel, A. G. (**1982**). "Phonetic prototypes," Percept. Psychophys. **31**, 307–314.

Shepard, R. N. (**1962a**). "The analysis of proximities: Multidimensional scaling with an unknown distance function. I.," Psychometrika **27**, 125–140.

Shepard, R. N. (**1962b**). "The analysis of proximities: Multidimensional scaling with an unknown distance function. II.," Psychometrika **27**, 219–246.

Shepard, R. N. (**1972**). "Psychological representation of speech sounds," in *Human Communication: A Unified View*, edited by E. E. David and P. B. Denes (McGraw-Hill, New York), pp. 67–113.

Singh, S., and Woods, D. R. (**1971**). "Perceptual structure of 12 American English vowels," J. Acoust. Soc. Am. **49**, 1861–1866.

Stevens, K. N., and Öhman, S. E. G. (**1969**). "Crosslanguage study of vowel perception," Lang. Speech **12**, 1–23.

Stevens, S. S., Volkmann, J., and Newman, E. B. (**1937**). "A scale for the measurement of the psychological magnitude pitch," J. Acoust. Soc. Am. **8**, 185–190.

Strange, W., and Dittmann, S. (**1984**). "Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English," Percept. Psychophys. **36**, 131–145.

Terbeek, D. (**1977**). "A cross-language multi-dimensional scaling study of vowel perception," Working Papers in Phonetics **37**, UCLA, Los Angeles, CA.

Volaitis, L. E., and Miller, J. L. (**1992**). "Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories," J. Acoust. Soc. Am. **92**, 723–735.

Wilkinson, L. (**1989**). *SYSTAT: The System for Statistics* (SYSTAT, Evansville, IL).