Original Paper

Phonetica 2007;■:1–23
DOI: 10.1159/000010■■■■

Received: August 21, 2004
Accepted: April 28, 2006

Effects of Acoustic Variability in the Perceptual Learning of Non-Native-Accented Speech Sounds

Travis Wade Allard Jongman Joan Sereno

Linguistics Department, University of Kansas, Lawrence, Kans., USA

Abstract

This study addressed whether acoustic variability and category overlap in non-native speech contribute to difficulty in its recognition, and more generally whether the benefits of exposure to acoustic variability during categorization training are stable across differences in category confusability. Three experiments considered a set of Spanish-accented English productions. The set was seen to pose learning and recognition difficulty (experiment 1) and was more variable and confusable than a parallel set of native productions (experiment 2). A training study (experiment 3) probed the relative contributions of category central tendency and variability to difficulty in vowel identification using derived inventories in which these dimensions were manipulated based on the results of experiments 1 and 2. Training and test difficulty related straightforwardly to category confusability but not to location in the vowel space. Benefits of high-variability exposure also varied across vowel categories, and seemed to be diminished for highly confusable vowels. Overall, variability was implicated in perception and learning difficulty in ways that warrant further investigation.

Introduction

Speech produced by non-native users of a language can differ markedly from native speech, due to a variety of factors, including the interaction of speakers' native and target phonetic inventories and the time course of the language acquisition process [e.g., Best, 1995; Flege, 1995]. As a result, non-native-accented speech can be difficult to understand by both native speakers and machines. Less well understood is whether and how human listeners can improve their recognition of non-native sounds as a result of training and exposure. A few studies have explored this issue, with mixed results. The pattern emerging thus far seems to be that listeners can adapt to productions of individual non-native speakers rapidly and robustly as a result of exposure, but that training effects generalizable across speakers of a particular linguistic background are more elusive, and perhaps limited to sentential stimuli. For example, Clarke [2000,

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0031–8388/07/0■■■–00■
$23.50/0
Accessible online at:
www.karger.com/journals/pho

Travis Wade
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Azenbergstraße 12
D–70174 Stuttgart (Germany)
Tel. ■■■
Fax ■■■
E-Mail travis.wade@ims.uni-stuttgart.de

2002] found that listeners become faster at recognizing words produced by a single accented speaker after hearing just a few sentences, but also that participants exposed to native English and either Spanish- or Chinese-accented speech over 3 days showed advantages attending to previously encountered speakers but *not* new speakers of the same L1. Weil [2001] found that transfer of learning to other non-native speakers of the same L1 may depend on the particular task used for testing: participants trained using English word, sentence, and prose stimuli recorded by a native speaker of Marathi generalized to better recognize speech produced by another Marathi speaker only for certain sentence materials, and only speaker-specific learning effects occurred for word-level stimuli. Bradlow and Bent [2003] recently replicated some of these results, observing intelligibility advantages on English sentences produced by a novel Chinese native speaker for listeners who underwent training transcribing sentences produced by other Chinese speakers.

Considering these results and the methods of training used in the studies, it seems that non-native-accented speech may represent an interesting testing ground for ideas concerning phonetic acquisition in general. Like many phonetic training and categorization studies in recent years, the experiments discussed above incorporated, to differing extents, a 'high variability (HV) training paradigm' in which listeners are exposed to targeted sounds in a maximal variety of speaker, word, and utterance contexts. The popularity of this type of training derives from the finding that, while increased acoustic variability can cause some difficulty in identifying sounds [e.g., Mullennix et al., 1989], second-language learners acquire segmental contrasts better and retain them longer when they are heard in as large a variety of contexts as possible [e.g., Lively et al., 1993; Bradlow et al., 1997; Wang et al., 1999, 2001, 2003a]. The underlying assumption is that learning new sounds – or adapting to new variants of sounds – requires exposure to sufficient variability, perhaps because the sounds are stored in memory in an exemplar fashion [e.g. Nosofsky, 1986; Pisoni, 1997]. Notably, however, previous demonstrations of the importance of HV have generally involved sound categories that were produced by native speakers and learned by non-natives. Variability may relate in a more complicated way to non-native productions, since in addition to deviating from native category tendencies, non-native speech may be inherently more variable acoustically. It has been noted that, due to factors including non-constant proficiency across speakers and the acquisition process itself, there exists a greater range of acoustic distortion or variability in non-native productions than is found in native productions of speech sounds. This distortion is commonly observed to contribute to the poor performance of automatic speech recognizers on non-native speech in general [Van Compernolle, 2001] and has recently come under some scrutiny with respect to human perception [Nissen et al., 2004]. The present study is designed to investigate whether and how this additional variability may become an issue with human perceivers, and also whether the HV training assumptions hold up for highly variable, confusable categories.

Nygaard and Pisoni [1998] postulate that listening to speech produced by talkers of different accents is merely an extreme example of what occurs routinely as we encounter unfamiliar talkers. That is, while non-native-accented speech is difficult because non-native categories differ acoustically from native categories, exposure to proper levels of variability along the relevant dimensions should be beneficial in learning to recognize the sounds. However, this assessment may underestimate the way that the additional sources of variability in non-native speech may interact with perception

and learning. Differences between native talkers are likely to be governed by somewhat discrete physiological and sociolinguistic sources [e.g. Ladefoged and Broadbent, 1957]. Non-native speech, on the other hand, would introduce in addition to these issues more continuous variables related to proficiency and the time course of learning. Another possibility, then, is that, since this additional variability in non-native productions could lead to increased category overlap and confusability, learning (some) non-native sounds may be inherently more difficult and might proceed along different lines.

To explore these possibilities, this study was designed as a first step in characterizing the effects of variability in non-native productions on (1) the distribution of acoustic cues to non-native category identity and (2) the perceptual adaptation of native listeners to these sounds, in particular on the benefits of HV exposure. Our investigation centered about a moderately large set of word-level English productions by a group of Spanish speakers. We first estimated the overall recognition and learning difficulty posed by this set in a 3-day HV training study (experiment 1) and characterized its acoustic variability by comparison with a parallel set of native productions (experiment 2). With this information as a starting point, experiment 3 was designed to observe the effects of non-native variability patterns on non-native category learning. Considering previous findings and assumptions, we adopted as working hypotheses that (1) recognition and learning difficulty would primarily relate to the magnitude of differences between non-native and typical native productions, and that (2) exposure to more acoustic variability in training would result in robustly better learning, regardless of the specific challenges presented by a particular category or inventory.

Experiment 1

To gauge the overall difficulty and learnability associated with the inventory of non-native productions we selected for study, a training experiment exposed native English speakers unfamiliar with Spanish-accented speech sounds to this set over 3 days.

Method

Stimuli

The stimuli considered in experiments 1–3 were isolated productions of monosyllabic English words, taken from the 20 PB word lists designed by Egan [1948] to represent common usage and to be equal in range and degree of difficulty and in phonetic content. Each of 6 adult native speakers (3 men, 3 women) of Latin American varieties of Spanish read a unique set of three of the 50-word lists, for a total of 900 different words. Speakers' language backgrounds and length of English study are given in table 1. The speakers were judged by the experimenters and by an additional native Spanish speaker to encompass a fairly wide range of English proficiency but to each have conspicuously non-native English pronunciation. Subjectively, the accent type was judged to be similar across all speakers. Speakers were recruited from the University of Kansas community and paid $8.00 for their participation.

Each speaker was first given a list of 150 words to study and encouraged to query the experimenter regarding the meaning or pronunciation of unfamiliar items. They then read the words twice, beginning with 5 filler items and with a short break between the two repetitions. Words were presented in random order on a laptop computer screen (SuperLab, 1999, Cedrus Corp.) at a constant rate of 3 s/word. Recording took place in the University of Kansas anechoic chamber; participants were seated at a desk facing an Electrovoice RE-20 microphone. A Fostex D-5 DAT recorder was used, and productions were later digitized at 22.05 kHz by computer using Praat (1991–2002, P. Boersma and D. Weenink). For training and testing, only one token of the two repetitions was used.

Table 1. Language background of non-native speakers

Speaker	Native country	Years in English-speaking environment	Gender
1	Argentina	3	M
2	Guatemala	2	M
3	Costa Rica	2	M
4	Paraguay	8	F
5	Costa Rica	4	F
6	Mexico	0.33	F

Pre- and post-test stimuli were taken from the productions of 2 speakers (S6 and S3), whose productions were judged to be close to the average and representative of the set in terms of accentedness and recognition difficulty. Pre-test items were 50 words (one production of one PB list) chosen arbitrarily from speaker S6, and post-test items were another PB list from S6 as well as one from S3. The second list from speaker S6 was included at post-test to provide for the most direct comparison with the pre-test, while the list from novel speaker S3 was included to ensure that speaker-specific learning did not obscure more general training effects. That is, we considered it possible (although unlikely) that adaptation to the pre-test speaker's accent patterns might carry over to post-test for both trainee and control subjects, but assumed that any learning generalizable to a new, previously unencountered speaker would be limited to trainees. Each training session was comprised of one production of four PB lists, one from each of the 4 remaining speakers (S1, S2, S4, and S5), resulting in three unique 200-item sessions balanced for expected word difficulty and for number of tokens from a given speaker. No word was repeated within or across training or testing sessions. Composition and order of sessions (their combination of word lists and the day of training on which they were presented) was held constant across trainees.

Participants

Participants were 31 college-age native English speakers reporting normal hearing and little or no experience with the Spanish language or Spanish-accented English. Fifteen participants were arbitrarily selected as trainees, and the remaining 16 served as controls. One control was discarded for failure to complete the training. Participants were recruited from introductory Psychology or Linguistics classes at the University of Kansas and received course credit for their participation.

Procedure

Testing and training sessions were administered by computer. Stimulus presentation order was randomized, across speakers, within each session. Participants first heard a production over Sony MDR-7502 dynamic stereo headphones, after which they were instructed to type the English word they perceived, followed by the SPACE bar. In training sessions, participant responses were followed by feedback, in which information about the accuracy of the response, the typed participant's response, and the intended word appeared on the screen accompanied by a bell (correct) or buzz (incorrect) sound, immediately after a response. After 1,500 ms, an additional repetition of the same token was played as this information remained on the screen, providing maximal feedback. Finally, after an additional 1,000 ms the screen was cleared and the next token was played. In testing, no feedback was given, and termination of a response was simply followed by the next word after a 1,500-ms pause. As far as possible, typed responses homophonous with intended words within General American English were treated as correct for purposes of both feedback and scoring; misspelled responses were simply treated as incorrect. Participants were encouraged to type carefully rather than quickly.

All tests were administered in sound-attenuated rooms in the Kansas University Phonetics and Psycholinguistics Laboratory. Both trainee and control participants completed pre- and post-tests, trainees on the days immediately preceding and following the 3 consecutive days of training, and controls in sessions 5–7 days apart.

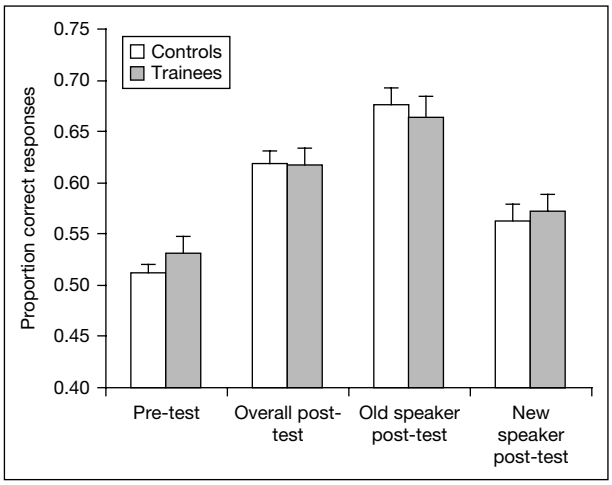

Fig. 1. Overall proportion correct responses for trainee and control participants at pre-test and post-test, including old speaker and new speaker test results. Error bars show standard error of the mean.

Results

Overall Performance and Training Effects

Pre- and post-test data for the 30 control and trainee participants are given in figure 1. Recognition accuracy was quite low overall, in the neighborhood of 60% across tests, speakers and listeners. A 2 (Training) × 2 (Test) ANOVA revealed a main effect of Test [$F(1, 28) = 102$, $p < 0.001$], such that participants performed significantly better on words in the post-test condition. This overall Test effect can be attributed to list effects or task familiarity. No effect was observed for Training [$F(1, 28) = 1.08$, $p = 0.308$], nor was there a reliable Test × Training interaction [$F(1, 28) = 0.27$, $p = 0.608$]. In fact, the nonsignificant difference in performance across participant groups was in a direction opposite that predicted by a training effect: controls performed slightly better at post-test, and improved more over pre-test, than trainees. For the post-test, performance for the pre-test speaker was contrasted to the novel speaker. The new speaker was more difficult for both control and trainees at post-test [$F(1, 14) = 33.03$, $p < 0.001$], but again no overall training effect ($p = 0.941$) or Speaker × Training interaction ($p = 0.553$) was observed.

These results suggest that the 3-day HV training program had little effect on listeners' ability to recognize isolated Spanish-accented monosyllable words, either for a previously heard or a new speaker. While it is difficult to infer from this that training had *no* effect, or that the non-native sounds were in some sense unlearnable, with the present 15-listener samples power for the cross-groups contrast would reach 0.8 at approximately effect size $d = 0.9$. Using the trainees' overall post-test standard deviation (SD) of 3.7%, this suggests that a training effect of about 3.3% would probably have been detected. Since most previous HV training studies [e.g. Logan et al., 1991; Lively et al., 1993; Bradlow et al., 1997; Bradlow and Bent, 2003] have observed larger effects (often 10% or more), it seems fair to characterize our set of non-native productions as comparatively difficult to learn.

While they may not have learned much about Spanish-accented sounds in general, experiment 1 trainees did seem to learn something about the speakers they heard during

training. Listeners' identification accuracy on the 4 speakers encountered over the course of the three training sessions increased each day (day 1, 58.4%; day 2, 59.6%; day 3, 60.5%); a repeated-measures ANOVA demonstrated that overall performance on day 3 reliably (although by only 2.1%) exceeded performance on day 1 [$F(1, 14) = 5.714$, $p = 0.031$]. The present study cannot conclusively rule out the possibility that this effect was due to differences in list difficulty rather than learning, since the composition of each training session was held constant over participants. However, each session included three word sets that were normalized for phonetic content [Egan, 1948] and produced by the same 4 speakers, so it seems unlikely that an accuracy trend over the three sessions – in precisely the order and direction predicted by a learning effect – would arise purely by chance from list composition. Thus, although there was no detectable learning of accent-level characteristics in a comparison of pre-test to post-test, training did seem to give participants some advantage in comprehending individual non-native speakers.

Phonetic Sources of Difficulty

Manual examination of participant responses revealed that most of the errors (at least 95%) were due to genuine misidentification of words and not to spelling or typing mistakes. Considering accuracy across word stimuli, a normal distribution of response scores was seen such that few words were either always or never identified correctly, while most fell somewhere between these extremes. Given the variability in phonetic content within and across speakers' word lists and the variety of recognition mistakes that were made, it was difficult to assign blame for the low accuracy on a particular set of sounds or contrasts. Individual speech sounds usually occurred in a given position few times over the course of training, so that specific differences in phonetic content across words were convolved with differences in word frequency, neighborhood density, and order of appearance in training or test.

One informally observed trend, however, was that a disproportionately large number of errors seemed to result from misidentification of vowels. Since responses were constrained to be monosyllable English words, many errors were not attributable to simple phoneme substitutions; as a result, it was impossible to quantify this observation or make any clear comparison involving specific misidentifications. However, since all words were monosyllables so that different vowel nuclei never co-occurred within an item, it was possible to compare performance crudely *across* vowel categories by simply counting errors resulting from words containing each category. Figure 2 shows the accuracy (averaged over tests and training) of responses to words containing the eight vowels [i], [I], [ɛ], [æ], [ɑ], [ʌ], [ʊ], and [u]. (Stimuli containing these vowels comprised a subset of about 70% of the total number of productions; words with diphthong or rhotacized nuclei were excluded from the present analysis since (1) they resulted in fewer overall errors and (2) their dynamic patterns introduced additional sources of acoustic variability that, while interesting, tended to obscure lower-order patterns in an analysis of vowel space (experiment 2) that seemed to characterize the production of these eight commonly misperceived categories.) In figure 2, some tendencies can be seen that might relate to differences in non-native production, in particular the very low performance on words containing [i] and [ʌ] nuclei. Some of the other vowels seemed to cause fewer difficulty, although the lack of power for the more rarely occurring categories (particularly [ʊ]) is still another cause for caution in interpreting these patterns. In any case, with this data as a starting point, experiments 2 and 3 were designed to observe more precisely the role of acoustic variability in the production and perception of these sounds.

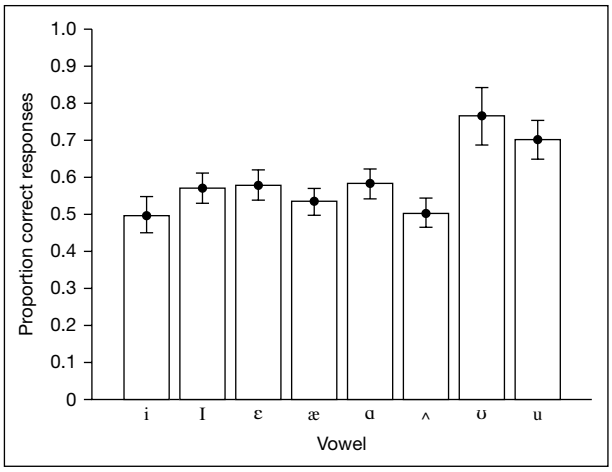

Fig. 2. Proportion correct responses to words containing the eight vowels [i], [I], [ɛ], [æ], [ɑ], [ʌ], [ʊ], and [u]. Error bars show standard error of the mean.

Discussion

Experiment 1 demonstrated that the multispeaker set of non-native productions we selected for study presented both recognition and learning challenges for native listeners. On average, the words were identified with about 60% accuracy, and 3 days of HV training resulted in little or no accent-level (though perhaps some speaker-level) improvement. This is in line with our expectations regarding difficulty with non-native speech and suggests that this set of productions may provide an informative base for testing the effects of variability on this difficulty. In order to measure and characterize the variability that the set contained, it was compared acoustically with a parallel set of productions taken from native speakers in experiment 2.

Experiment 2

To determine what effect the acoustic variability in experiment 1 speakers' vowel productions had on their recognition and learning difficulty, it was first necessary to characterize this variability by comparing the non-native productions acoustically with a parallel set of native-speaker productions of the same words. Since, as discussed above, vowel identification was the source of a number of errors in experiment 1, this investigation focused on non-native speakers' use of the English vowel space.

Method

Stimuli

Sounds examined consisted of the eight vowel segments [i], [I], [ɛ], [æ], [ɑ], [ʌ], [ʊ], and [u] of the monosyllable PB words produced for use as stimuli in experiment 1, and those from an identical set of words elicited from 6 adult native speakers (3 men, 3 women, with speaker-list assignments matched for gender with non-natives) of General American English. All participants were recruited from the University of Kansas in the same manner as those used in experiment 1. Speakers received course credit for their participation. 2,456 total productions were examined, with individual vowels represented as follows [see also Egan, 1948]: [i] occurred 264 times, with a range of 18–26 productions

per subject; [I], 404 total (range 26–48); [ɛ], 304 (14–38); [æ], 480 (32–48); [ɑ], 404 (24–42); [ʌ], 380 (24–40); [ʊ], 52 (2–6), and [u], 380 (24–40).

Procedure

Words were segmented manually to identify vowel portions, which were analyzed using the Praat program. The beginning of the vowel was judged to be the point at which a clear F1 appeared and the end of the vowel was denoted by a simultaneous cessation of the formants or a rapid shift in the formant pattern. (A more detailed description can be found in Wade [2003].) Within vowels, f0 was detected using an autocorrelation method [Boersma, 1993]. The first three formant resonances were estimated using the Burg algorithm, with a 50-ms Gaussian window at 10-ms intervals to calculate a maximum of five formants with a ceiling of 5,500 Hz for female speakers and 5,000 Hz for males. 50 of the productions were then selected at random across speakers, vowels, and native languages, with the restriction that each of the eight vowels was chosen at least once, and the first three formants were measured by a trained phonetician at vowel midpoint from spectrograms of these sounds. Observed values correlated highly with their automatically detected counterparts ($r = 0.981$, $p < 0.001$) and differed on average by only 59.6 Hz or 9%, suggesting that this particular method of deriving formant frequencies for analysis was satisfactory for the set of productions observed.

Vowel measurements were represented as points in a two-dimensional vowel space based on Miller's [1989] formulations of height and backness, as this method was observed to generalize effectively across speakers' widely differing f0 ranges. Under this formulation, each speaker's formant pattern was scaled to a sensory reference (SR) related to his or her average fundamental frequency:

$$SR = 168\ (GM\ f0/168\ )^{1/3}$$

From this reference, a given vowel's height was represented as a normalized version of the first formant resonance:

$$y = \log\ (\ F1/SR\ )$$

Similarly, backness was assumed to be a measure of the distance between the first two formants:

$$z = \log\ (\ F2/F1)$$

All graphical and numerical data presented below represent values in these two (height and backness) dimensions.

Results

Overall Variability

Figure 3 shows observed distributions of eight English vowels [i], [I], [ɛ], [æ], [ɑ], [ʌ], [ʊ], and [u] at vowel midpoint as produced by native and non-native speakers. Visual examination of these data suggests that while native vowels for the most part tend to cluster in regular, elliptical patterns, many of their non-native counterparts disperse to form amorphous, heavily overlapping categories. As one way of quantifying this observation, SDs calculated from the observed height and backness values for each vowel – combined across speakers – were taken as measures of the variability of the vowel. The 16 SD values (one for each vowel in each dimension) were compared pairwise across language groups. Non-native speakers were consistently more variable in their productions than native speakers; specifically, by a factor of about one third. Across vowels, non-native variation (mean SD value, 0.0995) was significantly greater than native variation (0.0775) [$t(15) = 3.78$, $p = 0.002$]. Examination of individual values, shown in table 2, showed that this general pattern held true in both dimensions for every vowel, except for the vowel [æ]. For each of the seven remaining vowels [i], [I], [ɛ], [ɑ], [ʌ], [ʊ], and [u], non-native productions were more variable in height and backness than native productions. For [æ], much of the native variability appeared to be

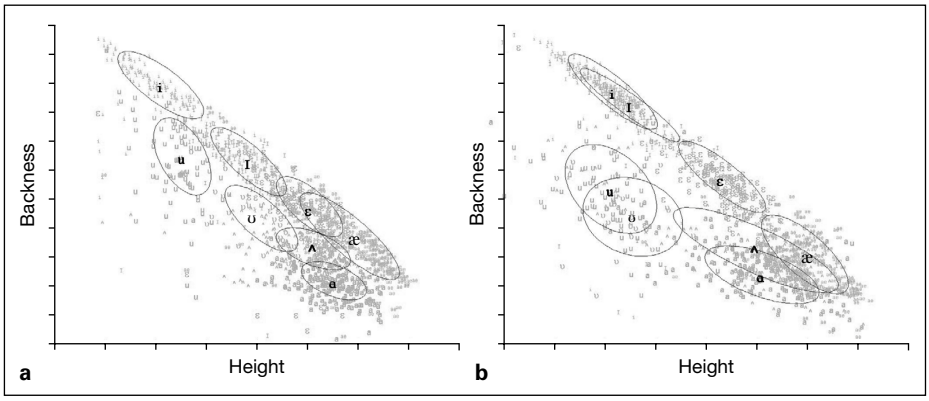

Fig. 3. Individual plots for native (**a**) and non-native (**b**) speakers for the English vowels in this study. Ellipses represent equal-likelihood contours after the removal of a total of 65 total outlier values, at an arbitrary constant level selected to demonstrate orientation and variability of categories. Dimensions are y (height) and z (backness) specifications [Miller, 1989] (see text).

Table 2. Production data from experiment 2

	Mean height	Height SD	Mean backness	Backness SD	Height-backness correlation	Overall variability ($\times 10^{-4}$)	Category confusability
Native productions							
i	0.3005	0.0719	0.8803	0.0988	−0.784	0.1945	0.1426
I	0.4732	0.0633	0.6181	0.1002	−0.814	0.1357	0.3274
ɛ	0.5931	0.0554	0.4631	0.0878	−0.72	0.1139	0.3757
æ	0.6716	0.0833	0.3618	0.1274	−0.846	0.3202	0.4729
ɑ	0.6405	0.0544	0.2095	0.0558	−0.521	0.0671	0.1782
ʌ	0.6006	0.0619	0.3191	0.0615	−0.563	0.0990	0.4806
ʊ	0.4981	0.0609	0.4227	0.0969	−0.726	0.1647	0.3788
u	0.343	0.0477	0.6353	0.1125	−0.491	0.2185	0.1985
Average	0.515	0.062	0.489	0.093	−0.683	0.164	0.3193
Non-native productions							
i	0.3058	0.0747	0.86	0.1101	−0.874	0.1597	0.4665
I	0.3394	0.0842	0.8142	0.1078	−0.94	0.0959	0.4471
ɛ	0.5209	0.0732	0.5666	0.105	−0.851	0.1629	0.2237
æ	0.6844	0.0727	0.2998	0.1151	−0.748	0.3084	0.2545
ɑ	0.5996	0.0959	0.2294	0.0848	−0.61	0.4153	0.3144
ʌ	0.5875	0.1381	0.3149	0.1244	−0.822	0.9572	0.7161
ʊ	0.345	0.0835	0.4292	0.1153	−0.187	0.8945	0.3962
u	0.3018	0.0768	0.5251	0.1298	−0.397	0.8371	0.4058
Average	0.461	0.087	0.505	0.112	−0.679	0.479	0.403

Height and backness [Miller, 1989] (see text) means, SDs, and correlation values as well as derived variability measures (overall variability and category confusability) are presented. Overall variability is the product of the eigenvalues of the relevant covariance matrix, and category confusability is the complement of ideal performance given the vowel distributions (see text).

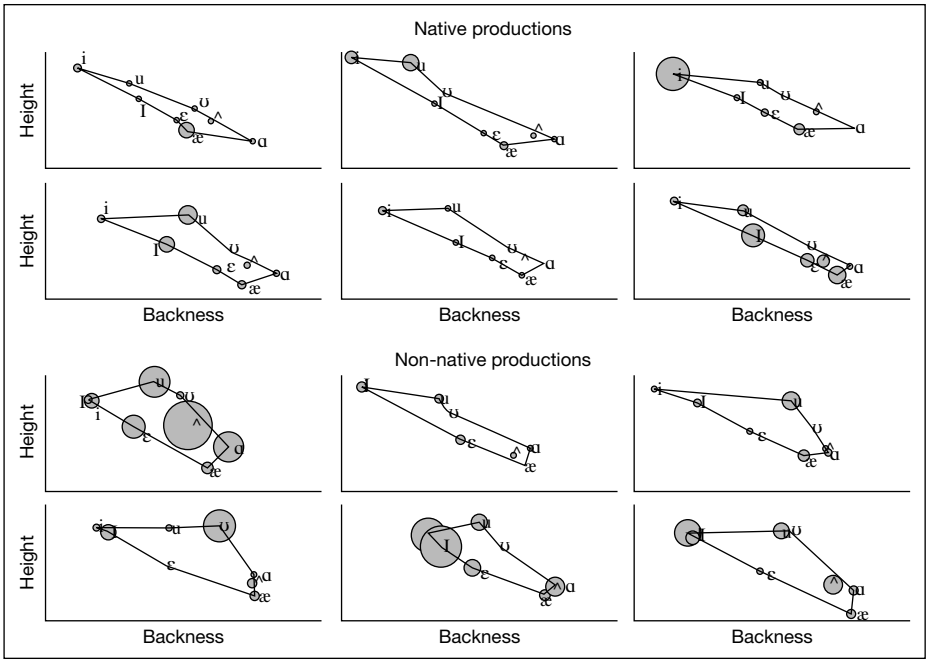

Fig. 4. Vowel location and variability across speakers. Marker size is proportional to summed height and backness SDs; variability is represented only for plots represented by 5 or more productions, and the vowel [ʌ] is left out for clarity.

phonologically conditioned, as native Eng lish speakers tended to produce a substantially raised [æ] preceding certain alveolar consonants.

Height-backness correlation values for each vowel are also given in table 2. Although these two features are often assumed to function independently in linguistic systems, their values were highly correlated in the two-dimensional space considered here, with most categories arranged elliptically along a low back–high front axis. Therefore, in comparing variability across language backgrounds it was important to consider the overall sizes of vowel category distributions as well as the separate height and backness values. Table 2 also gives overall variability, the relative areas of the ellipses shown for each vowel in figure 3 (the exact value given is the product of the eigenvalues of the relevant covariance matrix). Again, on average non-native categories are much more variable than native categories. However, as is evident in figure 3, this difference is mostly due to the four back vowels [ɑ], [ʌ], [ʊ], and [u]; front vowels involve approximately equal absolute variability across groups.

Within-Speaker and Across-Speaker Variability

Since each speaker in a language group produced a unique set of words, it was difficult to determine whether the overall increase in variability in non-native vowels was due primarily to within- or across-speaker factors. Considering the possible changes within and differences across learners that might result from the acquisition process [e.g. Best, 1995; Flege, 1995], we predicted that both factors might play a role. Figure 4 shows the mean locations and variability for each vowel as produced by each of the 12 speakers. It is

difficult to discern any consistent effects of word list, gender, specific language background, or time spent studying English in figure 4 (table 1), nor can a disproportionate amount of the variability in either group be attributed to a subset of speakers. Instead, it appears that non-native productions simply involved both more within-speaker and more cross-speaker variability in general than native productions.

Category Locations and Confusability

Differences were also observed in absolute location of the native and non-native vowels. For example, examination of figure 3 reveals that while the Spanish-accented [I] category was significantly lower ($p = 0.003$) and further back ($p = 0.014$) than the neighboring [i], the two categories are positioned in much greater proximity than their native-produced counterparts, both occurring toward the lower-back range of the native [i]. This indicates that the tense distinction between [i] and [I] is almost (but not quite) neutralized across non-native speakers, contributing to the overall confusability of the two non-native categories. A similar pattern can be seen with the parallel pair of back vowels [u] and [ʊ] and, to a lesser extent, the three low vowels [æ], [ɑ], and [ʌ]. Figure 3 also shows larger-scale differences in the overall central tendencies of the non-native vowel categories compared to the native produced ones. Notably, non-native vowels are on average further back. In particular, [ɑ], [ʊ], and [u] categories are much further back in non-native productions. This could be due to a number of factors; it has been observed [Bradlow, 1995] that (Madrid) Spanish speakers produce vowels with lower F2 values than English speakers, perhaps due to a language-specific base of articulation, and also that these three vowels are in general fronted in recent American English [e.g., Hillenbrand et al., 1995].

To determine the effects of the variability patterns discussed above on vowels in these different locations, and to predict the confusability of vowel categories across language groups, discriminant analysis was employed, as follows. For each production set, the vowel space was divided by optimal decision bounds into regions corresponding to the maximum probability distribution function value across the eight observed vowel distributions (defined by mean, SD, and correlation values in table 2). Based on these same distributions, the ideal recognition performance on each vowel was calculated as the proportion of its productions that would be correctly classified. Table 2 gives the differences of these ideal scores from perfect performance, a measure proportional to category confusability. On average, non-native categories are about 10% more confusable than native categories. Across vowels, this pattern holds for all but the [æ] and neighboring [ɛ] categories, for which the native productions were more confusable. As mentioned above, this probably resulted from phonologically conditioned variability in [æ] productions.

Judging by the far-from-perfect ideal observer classification (less than 70% for native categories), some of the observed variability for all vowel categories in both groups was undoubtedly due to coarticulatory influences from the consonant contexts in which they were produced. However, since the different vowels were produced in similar contexts [Egan, 1948], and since native and non-native sets involved the same two productions of the same 900 word contexts, the finding that non-native productions were comparatively more variable and confusable is of primary significance.

Relation to Experiment 1 Performance

Comparing figures 3, 4 and table 2 with the experiment 1 error patterns in figure 2, while it is difficult to draw firm conclusions for the same reasons listed above, there is

at least one striking similarity. The vowels [i] and [ʌ], for which the most errors were seen in experiment 1, were the two most confusable vowels in terms of discriminant classification, and also showed the largest increases in confusability from native to non-native productions. This is consistent with the notion that category variability contributed to the difficulty of experiment 1 vowel identification and learning. It is inconsistent with the assumption that absolute deviation from native mean category locations was the main contributor to difficulty, since as shown in table 2 these two non-native vowels were in fact the *closest* to their native counterparts.

Discussion

The non-native productions were characterized by a robust increase over native productions in the variability with which they made use of the English vowel space. This was true in terms of both height and backness for every vowel except [æ]. Differences in the absolute locations of categories across native and non-native productions were observed as well, with the means of certain non-native vowel pairs (i-I, u-ʊ, and æ-ɑ-ʌ) located closer together. Using discriminant analysis, non-native vowel categories (except for the vowels æ and ɛ) were found to be more confusable (about 10%) than native vowel categories. Overall, non-native productions showed more vowel variability and often overlapped adjacent vowel categories, and the most heavily overlapping categories were precisely the categories that seemed to cause the most errors in experiment 1. These observations present the possibility that variability, rather than deviation from native-produced categories, is to blame for difficulty in their learning. Experiment 3 was designed as a training experiment to test this possibility in a more controlled manner.

Experiment 3

To investigate the effects of the vowel location and variability patterns reported in experiment 2 on perceptual learning of the sounds – and on the benefits of HV training in driving this learning – a training study exposed listeners to artificially derived vowel inventories in which these variables were manipulated explicitly. In this study, all word, token, speaker, and accent variability observed in productions was modeled as a single source and explicitly controlled in a limited set of production stimuli from a single artificial 'speaker'. Three levels of variability (Minimal, Native, and Non-native) and two sets of category means (Native and Non-native) were created based on values observed in experiment 2. For each of the six possible variability-location pairings, an inventory of vowel distributions was derived and placed in a single consonant context, and learning and recognition patterns were compared across inventories. With respect to the results of experiment 2, this method does involve an unnatural simplification in its representation of variability, since potentially informative speaker and phonetic context cues to vowel identity were removed. Thus, experiment 3 was not intended as an absolute measurement of the learnability of Spanish-accented English vowels, but as an investigation of the effects of vowel location and vowel variability – of the types that may occur in non-native as opposed to native speech – on learnability. Again, we roughly adopted the predictions that (1) recognition and learning difficulty would primarily relate to deviation from native

category tendencies and not to variability and that (2) exposure to more acoustic variability in training would consistently result in better learning.

Method

Stimuli

All training and testing stimuli consisted of monosyllable words containing the eight vowels [i], [I], [ɛ], [æ], [ɑ], [ʌ], [ʊ], and [u] in the constant CVC context [h-V-d]. All stimuli were primarily synthetic, based on values obtained in experiment 2. Within-category variation involved *only* the first two steady-state formant frequencies of the vowel, so that tokens were manipulated within the height-backness space for which production variability was analyzed in experiment 2, while all temporal and higher-order spectral characteristics were held constant at reasonable values. In particular, vowel length was withheld as a secondary cue for identification; this was necessary to ensure that learning differences related only to use of vowel space.

Six distribution types were defined, based on the locations of categories in the vowel space and the absolute degree of variability in their occurrence in this space during training. Mean and variability values are all given in table 2. For location, mean F1 and F2 values, and their correlation coefficients, were based on observed values from experiment 2 for either native English speakers or non-native Spanish-accented English speakers. For variability, F1 and F2 SD values were based on observed values from experiment 2 for either (1) the observed native speaker value for the vowel, (2) the observed non-native value, or (3) a constant, arbitrarily low value (0.01). Within these groups, height and backness values of individual vowels used to train participants were randomly taken from normal distributions specified by one of the two appropriate mean/correlation designations and one of three possible SD values. The only exception was the vowel [æ], for which, as discussed above, more overall variability in production was actually observed for native speakers than for non-natives. Since including these contradictorily high values in the native-like condition might confound any overall effects of the 'elevated' non-native variability, and leaving the vowel out of the study altogether would require an artificially incomplete, potentially confusing set of vowels for training, it was decided simply to hold height and backness SDs for the vowel [æ] in all six conditions at the constant value of 0.01.

During training, the likelihood of a given vowel in a given condition occurring at any point in the possible vowel space was determined simply by the relevant bivariate normal probability density function. Representations of the resulting distribution inventories for the six distributions (native and non-native means with minimal, native, or non-native variability) are shown in figure 5.

H_d word stimuli containing these vowels were constructed dynamically during training and testing. A typical, clear token of the word *had* was first selected from the productions of non-native speaker S6, and from this production was constructed a basis for all generated words. Offline, the word was first segmented into /h/, /æ/, and /d/ portions. The LPC residue of the /h/ sound was adopted as a fricative portion that did not reflect any detectable vocal tract resonances and was therefore not specific to effects of a following vowel. A pulse train was created using the pitch and intensity patterns derived from the produced vowel, and this sound was concatenated with the fricative to create a neutral /hV/ section that resembled the relevant segments from the original production except that it carried no vowel information. For each word that appeared during testing or training, the Praat program was used to create a formant resonator based on (1) F1 and F2 values calculated from a (height, backness) pair randomly selected from the relevant distribution, (2) the average observed F3 value for the vowel in the appropriate language condition, and (3) values for F4–F8 that were derived from direct measurement of their corresponding center frequencies in the original /had/ production and held constant over all words. F1, F2, and F3 were held constant at their randomly derived values from the beginning of a word until the beginning of a transition to the final /d/, after which they linearly approached stop closure values of 350, 1,600, and 2,800 Hz, respectively. F4–F8 values were held constant at 4,800, 5,200, 6,600, 7,700, and 8,250 Hz. Formant bandwidths were 500 Hz during the initial fricative and decreased linearly during the 50-ms preceding vowel onset to 50 Hz (F1), 100 Hz (F2), 200 Hz (F3), 300 Hz (F4), or 400 Hz (F5–8). The stored neutral /hV/ segment was filtered according to these parameters, and the resulting sound was finally concatenated to the /d/ segment extracted from the original production described above, resulting in a complete *heed*, *hid*, *head*, *had*, *hod*, *hud*, *hood*, or *who'd* token.

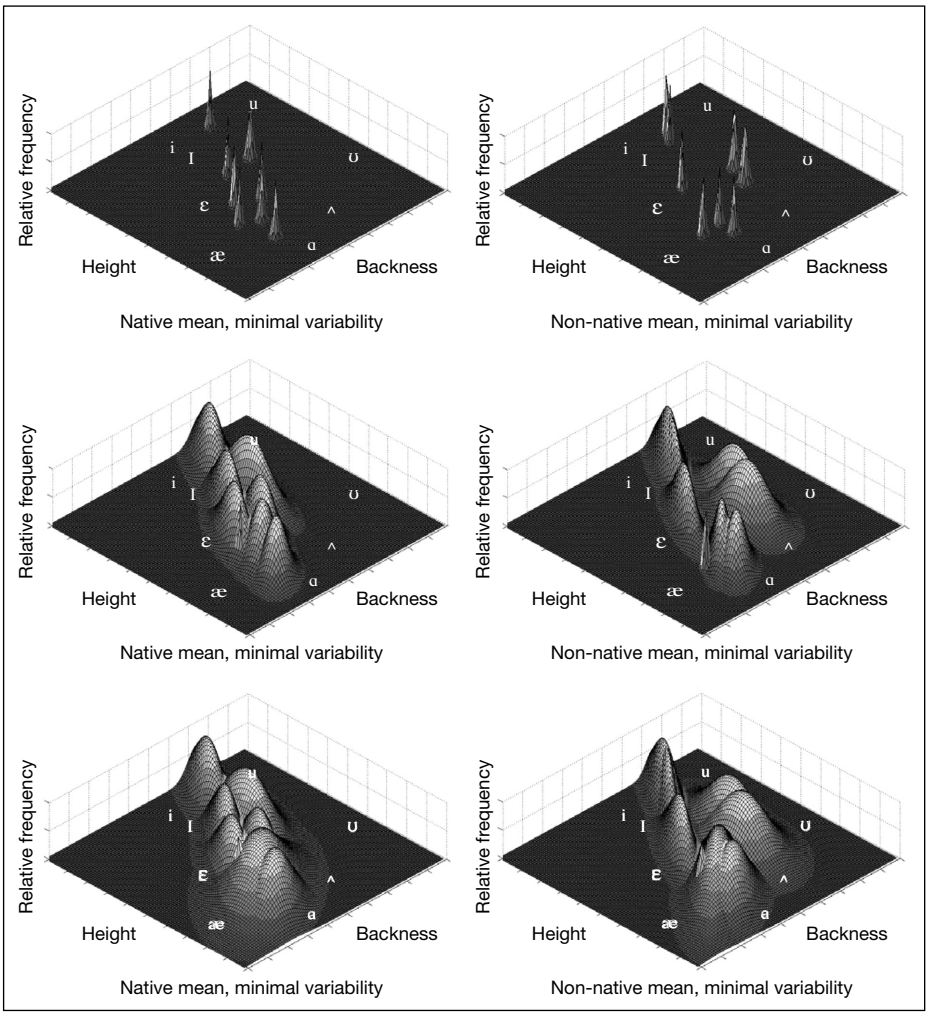

Fig. 5. Training distributions (normalized probability distribution functions) of vowel inventories used in experiment 3 across mean (Native, Non-native) and variability (Minimal, Native, Non-native) conditions.

Participants

Participants were 72 adult native speakers of English with no known hearing or speaking impairments, recruited from the University of Kansas community. Most participants received course credit for their participation. Twelve participants were arbitrarily assigned to training on each of the six distributions.

Procedure

Training and testing sessions were administered by computer. Training consisted of a single session involving an identification-with-feedback task that continued until participants met a learning criterion (see below), and an identification test session immediately followed.

Training stimuli for each participant were *h_d* words generated from vowel loci taken from one of the six distributions (Native or Non-native Mean × Minimal, Native or Non-native Variability) described above. Word order was randomized separately across participants with the condition that all vowels always occurred in equal proportions, and individual (height, backness) points were taken at random from a given distribution, separately across participants, so participants within a group encountered the same distributions but not the same exact stimuli.

During training, participants first heard a word over Sony MDR-7502 dynamic stereo headphones and were instructed to choose (mouse-click the button) from a set of eight possible words (*heed*, *hid*, *head*, *had*, *hod*, *hud*, *hood*, or *who'd*) appearing on a computer screen. No time limit was set for responding to individual stimuli, and participants were encouraged to take short breaks if they experienced any stress or fatigue from the task. After a participant chose a word, the button corresponding to the correct word was highlighted visually, accompanied by a bell (correct) or buzz (incorrect) sound to indicate the accuracy of the response. After 900 ms, an additional repetition of the same token was played as the correct word remained highlighted. Following an additional 1,000 ms, the screen returned to its original configuration and the next sound was generated and presented.

This procedure continued until a participant had (1) answered correctly to 30 out of any 50 consecutive presented words and (2) answered correctly to each of the eight possible words at least once. This 60% accuracy criterion was selected to match participants' typical post-test performance on the natural non-native-produced stimuli in the training study in experiment 1. To ensure some learning momentum, participants were informed of the criterion beforehand and allowed to monitor their progress (the number correct out of the most recent 50 items), though they were advised generally to ignore this information and to listen closely to each sound. If a participant did not meet the criterion within 1,000 items, training was terminated and the participant's results were discarded.

After training and a short break, participants were given a post-test in which stimuli were 10 tokens of each possible word (*heed*, *hid*, *head*, *had*, *hod*, *hud*, *hood*, or *who'd*) derived from a distribution with the vowel mean condition (Native or Non-native) on which the participant was trained, and non-native variability in all cases. These 80 productions were also randomly derived from the relevant distribution, and presented in random order. Participants were instructed to click the appropriate button for each word as in training, but no feedback was given. The next stimulus was presented 1,000 ms after a response.

Results

One participant in the (Non-native Mean, Non-native Variability) condition failed to reach the training criterion after exposure to 1,000 items, and the results of 2 additional participants were rendered unusable due to errors in test administration. These results were discarded and 3 additional participants were recruited so that comparable data from 12 participants was elicited for each of the six training conditions. On average, participants heard 195 stimuli (SD 193.3), or about 24 exemplars of each of the eight vowels before reaching criterion. Individual sessions ranged from 50 (the minimum possible) to 867 stimuli and took from less than 10 min to nearly 1 h.

Training Difficulty

Number of errors to criterion was taken as a measure of the difficulty associated with each set of vowel distributions used in training. Figure 6a shows errors across training conditions. A 3 (Variability) × 2 (Mean) univariate ANOVA revealed a main effect of Variability [$F(2, 71) = 24.4$; $p < 0.001$], such that more variable training sets generally resulted in more errors. This was in line with various previous observations [e.g. Creelman, 1957; Mullennix et al., 1988; Goldinger et al., 1991] that increasing stimulus variability translates predictably to difficulty in perceptual learning. More interesting was the observed main effect of category Mean value [$F(1, 71) = 14.3$;

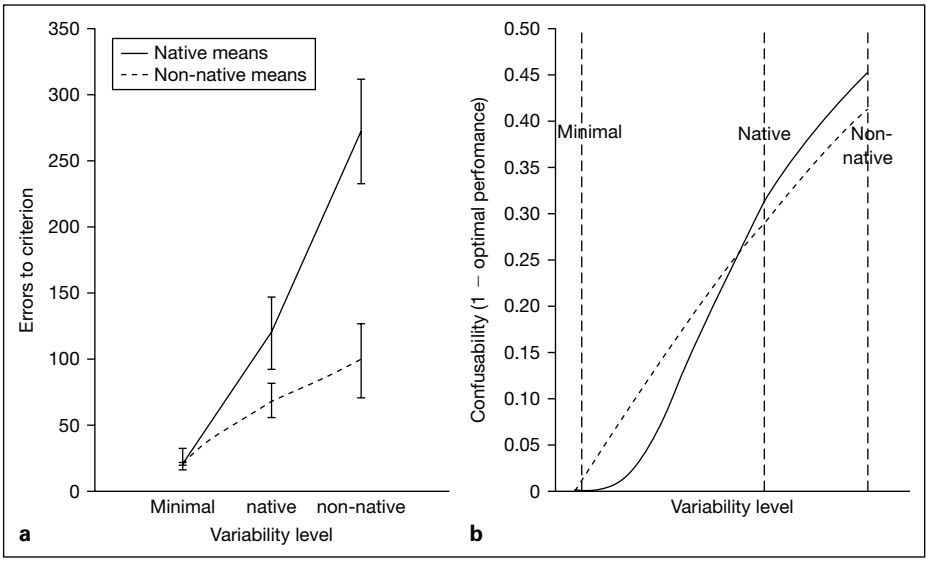

Fig. 6. **a** Summary of errors to criterion in experiment 3; error bars show the standard error of the mean. **b** Approximate average confusability (see table 2 for calculation) over differences in absolute variability. Mean locations and height-backness correlation coefficients are held constant for the two conditions, and, for simplicity variability is derived by multiplying the average of the two height and backness SDs for each vowel by a constant from 0 to 1.5.

$p < 0.001$] and the Mean × Variability interaction [$F(2, 71) = 7.8$; $p = 0.001$]. Distributions based on non-native means were actually easier to learn than those based on native-produced means, and increasingly so as variability increased from minimal to non-native levels. This is at first difficult to reconcile with the fact that, as observed in experiment 2, native categories were inherently less variable and less confusable than non-native categories, as well as occupying (presumably) more familiar central locations. However, as shown in figure 3, despite a few very confusable non-native distinctions the native vowel space was more compact overall, so that increased variability led to great overlap for the native distributions. This can be seen in figure 6b, where the average inventory confusability measure shown in the last column of table 2 is derived for different absolute variability levels. As shown here, predicted confusability demonstrates the same interaction as the observed errors: an advantage for the native means at lower variability levels gives way to advantages for the non-native inventory with higher variability. This immediately contradicts the assumption that the difficulties posed by a non-native accent are merely a product of its deviation from the standard (native) pronunciation. Rather, overall category overlap and confusability seem to be the driving factor in learning difficulty. It should also be noted that the (Native Mean, Native Variability) and (Non-native Mean, Non-native Variability) conditions, corresponding most closely to the two sets observed in experiment 2, did not differ significantly in errors to criterion [$F(1, 22) = 0.282$; $p = 0.601$].

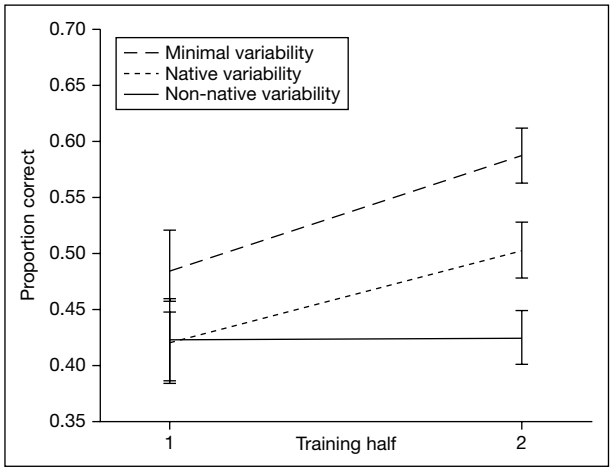

Fig. 7. Accuracy over the course of training in experiment 3; error bars show the standard error of the mean.

Learning Effects

As a rough indicator of improvement over the course of training, each participant's exposure session was divided into halves and overall accuracy was compared across the two sub-sessions. A 3 (Variability) × 2 (Mean) × 2 (Half) mixed model ANOVA revealed a within-subjects effect of Half [$F(1, 66) = 13.21$, $p = 0.001$] and the Half × Mean [$F(1, 66) = 4.38$, $p = 0.04$] and Half × Variability [$F(2, 66) = 3.36$, $p = 0.041$] interactions but no Half × Mean × Variability [$F(2, 66) = 1.87$, $p = 0.162$] interaction. Between-subjects main effects of Variability [$F(2, 66) = 10.51$, $p < 0.001$] and Mean [$F(1, 66) = 7.71$, $p = 0.007$] were observed, but not the Variability × Mean interaction [$F(2, 66) = 1.16$, $p = 0.319$]. Accuracy was better overall in the second half and, in line with difficulty (errors to criterion), for non-native mean categories and at lower variability levels.

Figure 7 shows improvement patterns over the three variability levels, collapsed over mean conditions. To interpret the Half × Variability interaction, separate repeated measures ANOVA across variability levels revealed that participants encountering native [$F(1, 23) = 8.18$; $p = 0.013$] and minimal [$F(1, 23) = 6.77$; $p = 0.016$] variability levels clearly improved from the first half of training to the second. However, participants dealing with non-native variability showed no such improvement [$F < 1$]. Thus, listeners in the non-native variability conditions were not able to adapt to these speakers' use of the vowel space, at least over the course of the relatively short HV training exposure employed here. This is generally consistent with the results of experiment 1, where listeners experienced difficulty perceiving and adapting to a similarly variable set of vowel productions. However, as demonstrated in the next section, post-test scores indicated that all participant groups in experiment 3 did acquire some knowledge of at least some of the vowel categories.

Post-Test Performance

Participants' post-test sensitivity for each vowel [hits (correct responses to the vowel) − false alarms (identification of the vowel in response to any other vowel)] they learned was compared across training conditions. An 8 (Vowel) × 2 (Mean condition) × 3 (Variability condition) mixed-model ANOVA revealed a within-subjects

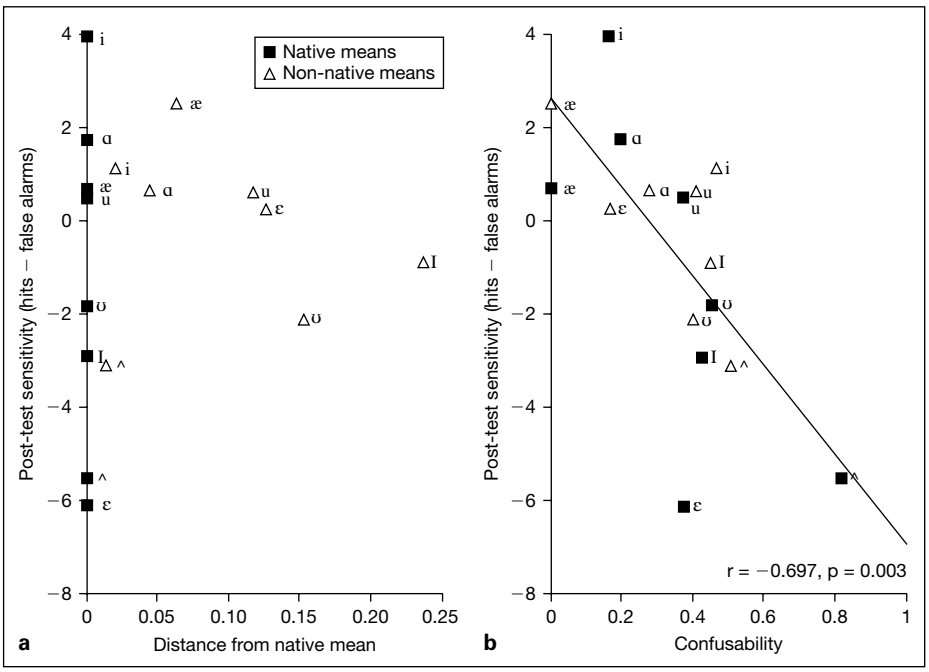

Fig. 8. Observed post-test sensitivity (hits − false alarms) plotted against distance from native mean (**a**) and estimated confusability (**b**).

effect of Vowel [$F(7, 462) = 59.78$, $p < 0.001$] and the Vowel × Mean [$F(7, 462) = 19.2$, $p < 0.001$] and Vowel × Variability × Mean [$F(14, 462) = 2.09$, $p = 0.011$] interactions as well as a between-subjects effect of Mean [$F(1, 66) = 14.06$, $p < 0.001$; all other $p > 0.05$]. Again, the native mean vowels were recognized less accurately overall, with some vowels presenting more difficulty than others. Figure 8 shows average post-test sensitivity for each vowel, plotted against two possible sources of recognition difficulty. As shown in figure 8a, the distance of a vowel category mean from the observed native mean location did not correlate well with post-test sensitivity ($|r| < 0.1$), whether all possible vowels (i.e. including the Native Mean vowels clustered on the y axis) or only the non-native-mean categories were considered. Along with the training error and improvement patterns discussed above, this suggests that absolute deviation from a typical native category was not a major factor in predicting either recognition difficulty or learnability.

Similar analyses compared post-test sensitivity with height, backness, and overall vowel variability (comparison can be made from table 2 and fig. 8). While this analysis suggested some rough trends (sensitivity generally seemed to be inversely related to variability), no reliable relationships were detectable. Absolute variability, then, did not seem to predict difficulty directly, either.

Figure 8b, however, demonstrates that the result of increased variability, category confusability (as derived from the discriminant analysis above and listed in table 2),

was a clear, linear predictor of post-test performance ($r = -0.697$, $p = 0.003$). Categories that were easily separable in the vowel space were universally recognized better at post-test. Thus, once again, variability and the resulting category confusability – and not simply deviation from a native mean – seemed to drive recognition difficulty.

Regarding the benefit of HV exposure across contrasts of different types and difficulty, the Vowel × Variability × Mean interaction in post-test sensitivity (again, hits − false alarms for each vowel category) indicated that – contradicting our predictions – exposure to variability in training did not have the same effect for all categories in all circumstances. To explore possible causes of the interaction, separate 2 (Mean) × 3 (Variability) univariate ANOVAs were calculated for each of the eight vowel categories. The only vowel for which a reliable effect related to training variability was observed with Bonferroni alpha correction ($p = 0.0063$) was the [i] vowel. Main effects of Mean [$F(1, 66) = 147.34$; $p < 0.001$], Variability [$F(2, 66) = 5.29$; $p = 0.007$] and a Mean × Variability interaction [$F(2, 66) = 6.65$; $p = 0.002$] were all observed for [i] category sensitivity. Examining the effect of training variability on each mean condition separately, a one-way ANOVA revealed effects of Variability on [i] category sensitivity in both non-native [$F(2, 33) = 4.34$; $p = 0.021$] and native [$F(2, 33) = 8.35$; $p = 0.001$] mean conditions. Tukey HSD post-hoc comparisons, however, revealed that the effect was in the opposite direction across language-mean groups: in non-native-mean tests, participants trained with minimal variability outperformed native and non-native variability participants, while no differences were observed between the latter two groups. In native-mean tests, on the other hand, participants trained with non-native variability were most sensitive to [i], outperforming native and non-native groups, with the latter two again equivalent. When tested on native mean categories, then, variability enhanced performance; when tested on non-native means, however, minimal variability was preferable.

Qualitatively, this observation seems best explainable based on the absolute positions of the [i] vowel in each of the two distribution types. Examination of figure 5 reveals that the [i] category is far and away the most isolated, least confusable vowel in the native inventory. Conversely, the non-native [i] distribution almost completely overlaps that of its nearest neighbor, [I], comprising the most difficult distinction in the set. Perhaps, then, the location and orientation of the native [i] are such that it is readily learnable under normal circumstances even when variability is increased unnaturally – with HV exposure aiding learning in the same way it has been observed to for native language categories [e.g., Lively et al., 1993; Bradlow et al., 1997; Wang et al., 1999]. The non-native [i], on the other hand, may be oriented such that listeners become aware of the dimensions of the [i]-[I] contrast only when exposed to prototypical, minimally variable exemplars during training. This interpretation suggests that HV exposure might provide less advantage, and perhaps some disadvantage, for increasingly difficult, confusable distinctions. To explore this possibility, we compared the effects of HV exposure across differences in category confusability, as follows. For each vowel, the relative degree of HV benefit was estimated as the slope of the regression line relating post-test sensitivity with averaged inventory variability (e.g. table 2, column 6). Figure 9 shows this measure plotted against category confusability (e.g. table 2, column 7) for each vowel. As shown in figure 9, HV benefit tends to vary inversely with category confusability, although the correlation does not reach significance ($r = -0.324$, $p = 0.22$) due to a few outlier vowels and the small set of data points. Implications of this observation are discussed further in the following sections.

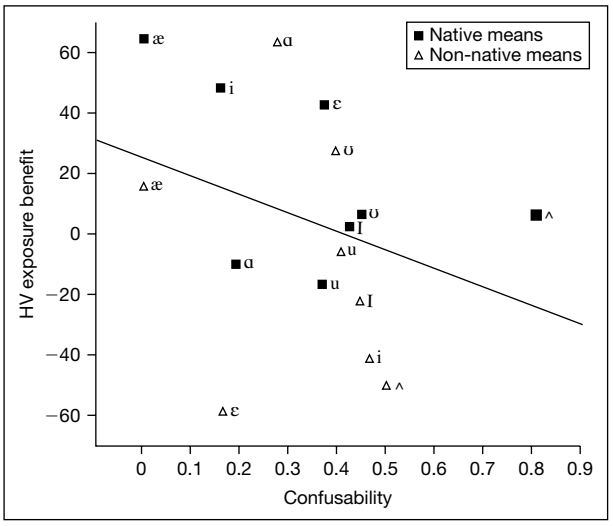

Fig. 9. HV exposure benefit (see text) plotted against estimated category confusability.

Discussion

Experiment 3 introduced a training paradigm that shed some light on the effects of HV exposure on increasingly variable, overlapping categories, and more specifically on whether increased variability and category confusability contribute to difficulty adapting to non-native accented speech. During training, participants had more difficulty as absolute variability increased, particularly with the overall more tightly grouped native-mean categories. This is not to say that non-native sounds are in some sense inherently positioned to provide for greater variability; certainly the particular trend observed here is specific to Spanish-accented English, and is perhaps due to base-of-articulation differences or native back-vowel fronting as discussed with regard to experiment 2. It does suggest, however, that it was the confusion and increased category overlap caused by non-native variability, and not simply non-native speakers' absolute deviation from native-like sounds, that was the primary cause of difficulty.

This finding was borne out further in post-test sensitivity to the eight vowels appearing in each inventory. Post-test identification errors seemed to have little to do with vowels' deviation from native mean height-backness locations, but were closely related to category overlap and confusability. More confusable categories robustly resulted in poorer recognition.

Post-test scores also revealed interesting differences in the effects of HV exposure on categories of differing confusability. Overall, there was a conspicuous lack of HV exposure advantages. This was surprising given our working hypotheses but difficult to interpret on a global level. The only vowel for which the effects were unambiguous was the high front vowel [i]; the isolated native [i] showed the canonical HV benefits, while the highly confusable non-native [i] was learned best under minimal exposure. In line with this observation, the relative benefits of HV exposure tended to vary inversely with category confusability, although not reliably. The observed Vowel × Variability × Mean

interaction, however, indicates at least that HV exposure is not of uniform benefit across the challenges presented by different categories.

General Discussion

This study was designed to investigate whether and how the acoustic variability associated with non-native productions of phonetic categories contributes to difficulty identifying and adapting to non-native accented productions. This question was addressed with the additional, more general goal of monitoring the benefits of HV exposure across differences in the inherent confusability of categories and distinctions to be learned.

To this end, we examined a set of monosyllable word-level productions of Spanish-accented English speakers. Experiment 1 measured the recognition challenges posed by exposing native English speakers to the accented words in a multisession training study. Overall recognition accuracy of listeners was quite low (around 60% overall), and detectable learning effects only took place for the speakers used in training and did not generalize to others, indicating that the set indeed posed some difficulty for native listeners. Examination of training and testing error patterns suggested that vowels, particularly [i] and [ʌ], were particular sources of difficulty. Experiment 2 measured the types and degree of vowel category variability and confusability of the non-native productions by comparing them to a parallel set of native productions with respect to height and backness. Non-native productions were robustly more variable in each of these dimensions. As a result of this variability and some differences in the absolute location, non-native categories were on average more confusable in terms of ideal discriminant performance. Mirroring experiment 1 results, [i] and [ʌ] were the most confusable non-native vowels.

Experiment 3 incorporated these findings in a training study designed to directly study the effects of the observed category location, variability, and confusability on recognition, and on the benefit of HV exposure in their learning. The primary finding in experiment 3 training was that identification and learning difficulty were mostly driven by variability and confusability and not by category location. Participants in non-native variability conditions made more errors and showed less improvement over the course of exposure, especially in native mean conditions, where a tighter overall inventory grouping led to more confusability at higher variability levels. Similarly, post-test sensitivity across vowel categories was clearly related to category confusability.

Experiment 3 post-test performance also revealed an interesting difference in the benefits of HV exposure across categories. For the most easily distinguished vowel, the native-mean [i], participants demonstrated something like the canonical HV effect; i.e., groups trained with non-native variability showed greater sensitivity to the vowel than those trained with less variability. For the least easily distinguished vowel, coincidentally the non-native-mean [i], however, the opposite effect was observed; i.e., groups trained with only prototypical, minimal variability vowels outperformed other groups at post-test. An overall comparison of the benefits of HV exposure across categories was consistent with the notion that relative HV advantages may decline for very difficult, confusable categories. Generally speaking, such a trend is not predicted by exemplar memory models [e.g. Nosofsky, 1986, 1991; Goldinger, 1996, 1998] which are often used to explain the HV phenomenon in phonetic acquisition. That is to say, all of

the categories tested in experiment 3 took the form of distributions that were predictably arranged in bivariate normal patterns in the height-backness space. Since no two categories overlapped *completely*, all distinctions could be made at better than chance, so that optimal encoding of the relevant distribution information in an exemplar fashion should have resulted from exposure to the full range of variability (as opposed to only prototype exposure, which conveyed no information other than the mean category location). While we will not suggest a detailed alternative based on the preliminary data reported here, one possibility is that, in addition to retaining information about lawful sources of variability (perhaps in the form of exemplars), listeners also take into account some more abstract information. In Klatt's [1979] LAFS (Lexical Access from Spectra), for example, it is suggested that specific information (regarding at least the properties of individual speakers' productions) may be stored alongside abstract templates for prototypical male and female talkers. It is likely that template information is more easily gleaned from very good exemplars of a category than from a wide range of variability, and possible that its advantages might surface only in very difficult classification tasks; this or some similar scenario would explain the present results. A related possibility is simply that different types of information are extracted by learners for different types of contrasts. Reporting on English speakers' acquisition of German nonlow vowels, for example, Kingston [2003] observes that for certain contrasts the introduction of natural, irrelevant acoustic variation hinders category or feature learning, while for others it is beneficial. These data suggest that for easily defined features such as [high] abstract values may be learned, while more polymorphous features like [tense] may be represented by exemplar sets. The results of the present experiments may represent a parallel pattern whereby some measure of variability or confusability dictates the nature of learning. Acoustically isolated categories encompassing large domains of legal variability (e.g., the native [i]) may be best represented by exemplar sets and learning may benefit from HV. Those occurring in close proximity to neighboring categories and involving boundary locations at critical locations in areas of high production density (e.g., the nonnative [i]) may instead involve abstract values and learning may benefit from low variability. Further study involving a larger number of categories from a number of different languages will be required to address these possibilities more thoroughly.

Acknowledgments

The research reported here is based on the first author's doctoral dissertation and was supported in part by a grant from the University of Kansas General Research Fund to the second author. Portions of the work were presented at the 15th International Congress of Phonetic Sciences, Barcelona, Spain, 2003.

References

Best, C.T.: A direct realist view of cross-language speech perception; in Strange, *Speech perception and linguistic experience: issues in cross-language research* (York Press, Baltimore 1995).

Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc. Inst. of Phonet. Sci. *17:* 97–110 (1993).

Bradlow, A.: A comparative acoustic study of English and Spanish vowels. J. acoust. Soc. Am. *97:* 1916–1924 (1995).

Bradlow, A.; Bent, T.: Listener adaptation to foreign-accented English. Proc. 15th Int. Congr. Phonet. Sci., 2003.

Bradlow, A.; Pisoni, D.; Akahane-Yamada, R.; Tohkura Y.: Training Japanese listeners to identify English /r/ and /l/. IV. Some effects of perceptual learning on speech production. J. acoust. Soc. Am. *101:* 2299–22310 (1997).

Clarke, C.: Perceptual adjustments to foreign accented English. Research on Spoken Language Processing Progress Report No. 24: Indiana University (2000).
Clarke, C.: Perceptual adjustment to foreign-accented English with short-term exposure. Proc. 7th Int. Conf. on Spoken Lang. Processing, 2002.
Creelman, J.: Case of the unknown talker. J. acoust. Soc. Am. *29:* 655 (1957).
Egan, J.: Articulation testing methods. Laryngoscope *58:* 955–991 (1948).
Flege, J.: Second language learning; theory, findings, and problems; in Strange, Speech perception and linguistic experience: issues in cross-language research (York Press, Baltimore 1995).
Goldinger, S.; Pisoni, D.; Logan, J.: On the nature of talker variability effects on recall of spoken word lists. J. exp. Psychol. *17:* 152–162 (1991).
Goldinger, S.: Words and voices: episodic traces in spoken word identification and recognition memory. J. exp. Psychol. *22:* 1166–1183 (1996).
Goldinger, S.: Echoes of echoes? An episodic theory of lexical access. Psychol. Rev. *105:* 251–279 (1998).
Hillenbrand, J.L.; Getty, L.A., Clark, M.; Wheeler, K.: Acoustic characteristics of American English vowels. J. acoust. Soc. Am. *97:* 3099–3111 (1995).
Kingston, J.: Learning foreign vowels. Lang. Speech *46:* 295–349 (2003).
Klatt, D.: Speech perception: a model of acoustic-phonetic analysis and lexical access. J. Phonet. *7:* 279–312 (1979).
Ladefoged, P.; Broadbent, D.E.: Information conveyed by vowels. J. acoust. Soc. Am. *24:* 98–104 (1956).
Lively, S.; Logan, J.; Pisoni, D.: Training Japanese listeners to identify English /r/ and /l/. II. The role of phonetic environment and talker variability in learning new perceptual categories. J. acoust. Soc. Am. *94:* 1242–1255 (1993).
Lively, S.; Pisoni, D.; Yamada, R.; Tohkura, Y.; Yamada, T.: Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. J. acoust. Soc. Am. *96:* 2076–2087 (1994).
Logan, J.; Lively, S.; Pisoni, D.: Training Japanese listeners to identify English /r/ and /l/: a first report. J. acoust. Soc. Am. *89:* 874–886 (1991).
Martin, C.; Mullennix, J.; Pisoni, D.; Summers, W.: Effects of talker variability on recall of spoken word lists. J. exp. Psychol. *15:* 676–684 (1989).
Miller, J.: Auditory-perceptual interpretation of the vowel. J. acoust. Soc. Am. *85:* 2114–2134 (1989).
Mullennix, J.; Pisoni, D.; Martin, C.: Some effects of talker variability on spoken word recognition. J. acoust. Soc. Am. *85:* 365–378 (1989).
Mullennix, J.; Pisoni, D.: Stimulus variability and processing dependencies in speech perception. Percep. Psychophys. *47:* 379–390 (1990).
Mullennix, J.: On the nature of perceptual adjustments to voice; in Johnson, Mullennix, *Talker Variability in Speech Processing* (Academic Press, San Diego 1997).
Nissen, S.L.; Smith, B.L.; Bradlow, A.; Bent, T.: Accuracy and variability in vowel targets produced by native and non-native speakers of English. J. acoust. Soc. Am. *116:* 2604 (2004).
Nosofsky, R.: Attention, similarity, and the identification-categorization relationship. J. exp. Psychol. *115:* 39–57 (1986).
Nosofsky, R.: Typicality in logically defined categories: exemplar-similarity versus rule instantiation. Memory Cognition *19:* 131–150 (1991).
Nygaard, L.; Pisoni, D.: Talker-specific learning in speech perception. Percep. Psychophys. *60:* 355–376 (1998).
Nygaard, L.; Sommers, M.; Pisoni, D.: Speech perception as a talker-contingent process. Psychol. Sci. *5:* 42–46 (1994).
Pisoni, D.: Some thoughts on 'normalization' in speech perception; in Johnson, Mullennix, *Talker variability in speech processing* (Academic Press, San Diego 1997).
Van Compernolle, D.: Recognizing speech of goats, wolves, sheep and . . . non-natives. Speech Commun. *35:* 71–79 (2001).
Wade, T.: *Acoustic variability and perceptual learning of nonnative-accented speech sounds*; PhD diss. University of Kansas (unpublished, 2003).
Wang, Y.; Jongman, A.; Sereno J.: Dichotic perception of Mandarin tones by Chinese and American Listeners. Brain Lang. *78:* 332–348 (2001).
Wang, Y.; Jongman, A.; Sereno, J.: Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. J. acoust. Soc. Am. *113:* 1033–1044 (2003a).
Wang, Y.; Sereno, J.; Jongman, A.; Hirsch, J.: fMRI evidence for cortical modification during learning of Mandarin lexical tone. J. cogn. Neurosci. *15:* 1019–1027 (2003b).
Wang, Y.; Spence, M.; Jongman, A.; Sereno, J.: Training American listeners to perceive Mandarin tones. J. acoust. Soc. Am. *106:* 3649–3658 (1999).
Weil, S.: Foreign accented speech: adaptation and generalization; MA thesis Ohio State University (unpublished, 2001).