Two groups of native speakers of Japanese (J) were given perceptual training on the American English (A.E.) vowels /ɪ/ and /iː/. Since J has corresponding vowels which differ primarily in duration while the A.E. vowels differ primarily in vowel quality, the pre-test and post-test use stimuli with the durations manipulated to measure which of these cues subjects used to distinguish between these two vowels. One group of subjects trained on natural stimuli while the other group trained on stimuli with the duration information removed (i.e. all tokens of a particular context of identical length). Results from pre-test show that the subjects do rely on duration as a primary cue to vowel identity. Both training groups categorized both natural and durationally manipulated stimuli better on the post-test than on the pre-test.

1. INTRODUCTION
It has been widely recognized that non-native speakers often have difficulty perceiving phonemic contrasts in a second language (L2) that do not exist in their native language (L1). Best [2] and Flege [4] have claimed the ability to perceive non-native contrasts is at least partially determined by the way that non-native phones are perceptually assimilated to their native phonetic categories. The reason for this is likely that subjects use the acoustic criteria which are important for discriminating phonemes in their L1. While experience with the L2 may improve perception, performance on categorization tests may never reach native levels, and the acoustic cues used in categorization may not be the same as those used by native speakers. For example, Bohn and Flege [3] found that some native speakers of German who had lived in the U.S. for long periods of time still did not discriminate the vowels /ɪ/ and /iː/ in a native English-like manner.

The present study focuses on native Japanese speakers’ ability to correctly categorize the American English vowels /ɪ/ and /iː/. The Japanese vowel system consists of 5 different vowel qualities, each of which can be realized either as a short (1 mora) or a long (2 mora) vowel. Thus the two Japanese vowels corresponding to English /ɪ/ and /iː/, /ɪː/ and /iː/, have nearly identical vowel quality but a contrastive difference in duration [10]. In English, in contrast, the two vowels /ɪ/ and /iː/ have a primary distinction in vowel quality, while vowel length is considered a phonologically redundant feature [5]. If the categorization strategy of the L1 is used for the L2, then Japanese speakers of English may rely on durational cues to discriminate high front vowels in English because Japanese /ɪ/ and /iː/ differ primarily in duration. However, English /ɪ/ and /iː/ differ in both vowel quality and duration, and Americans use vowel quality cues to discriminate the vowels. Since duration is not the primary cue used by Americans, using the Japanese strategy may cause Japanese speakers of English not only to have difficulty discriminating the two vowels, but also to categorize vowels incorrectly when the vowel quality and the (phonologically redundant) vowel duration cues do not agree.

Strange et al. [11] studied Japanese listeners’ perceptual assimilation patterns of the American English vowels to their native vowel categories. They demonstrated that Japanese speakers use can attend to a certain extent to both vowel quality and vowel duration in the assimilation of the vowels. The role of vowel quality cues is somewhat limited; the English vowels /ɪ/ and /iː/ are most often assimilated to either Japanese /iː/ or /iː/. However, duration appears to play a more critical role; when words are presented in isolation (i.e. detailed durational information of the context is not available), both of the English vowels are most frequently assimilated to the short Japanese vowel. In a sentence context, the English tense and lax vowels are more likely to be assimilated to Japanese long and short vowels, respectively. Since both of the English vowels are assimilated to the same Japanese vowel, Best [2] and Flege [4] would predict that native Japanese speakers should have great difficulty in correctly categorizing the two vowels, particularly in the absence of durational cues.

In the present study, the extent to which Japanese speakers of English use vowel quality cues and vowel duration cues was investigated. In addition, the impact of perceptual training for the contrast was investigated. Several studies have shown that non-native perception of these difficult phonemic contrasts can be improved through identification training with immediate feedback (see e.g. [6],[7]). These studies have concentrated on using a variety of naturally produced tokens which contain adequate amounts of variability so that the learner can form a good characterization of each phonetic category. In contrast with previous studies, the current study uses both naturally produced stimuli as well as stimuli which have been manipulated so that only the vowel quality can be used to correctly identify the vowel. This type of training directly addresses the suggestion by Strange et al. [11] that since Japanese subjects consistently perceptually assimilate English /ɪ/ and /iː/ to different Japanese categories only in the sentence condition, training of these subjects on vowel categories in English should use sentence-length materials to help subjects make use of the available durational. However, if native English speakers rely on vowel quality cues, this type of training may reinforce using precisely those cues which native speakers of English do not use.

2. METHOD
produced tokens, which had the normal redundant durational cues found in American English, while a separate group of subjects trained on the same tokens which had been durationally modified to remove durational cues to vowel identity.

2.1. Testing

2.1.1. Test Stimuli. Three native speakers of American English, two females and one male, produced the stimuli used in the test. Since perceptual assimilation of L2 vowels to an L1 is dependent on phonetic environment [12], the stimuli were chosen to maximize the variety of phonetic contexts, and each speaker produced different words. The stimuli, which included words in isolation as well as words in the carrier phrase “Now say X again,” were recorded in a sound-attenuated room at 16kHz. The vowels in the stimuli were then manually labeled, and the durations of the vowels were modified in three different ways using the TD-PSOLA algorithm [9]. Thus, there were four different variants of each stimulus:

- **natural**: length of stimulus not modified
- **long**: length of stimulus vowel doubled
- **short**: length of stimulus vowel halved
- **uniform**: length of stimulus vowel modified to within one pitch period of 140ms

The acoustic characteristics of the natural tokens of the words spoken in isolation are shown in Table 1.

2.1.2. Test Format. In the tests, a two alternative forced choice task was used. For each question, the subjects heard the stimulus, and had to enter on the computer whether the vowel in the word was the same as the vowel in “heat” or the vowel in “bit”. The subjects were instructed to concentrate only on the vowel sound that they heard, rather than using their knowledge of English words to determine the vowel sound. Reference to the actual words that were spoken was avoided to prevent confusion between the vowel sound and the orthography. No feedback was provided during the test.

2.1.3. Test Procedure. The pre-test/post-test consisted of four sections:

1. Natural, long, and short stimuli; words in isolation.
2. Natural, long, and short stimuli; words in carrier phrase.
3. Uniform stimuli only; words in isolation.
4. Natural stimuli only; words in isolation.

For each subject, the order of the questions within each section was randomized, thereby mixing the stimuli produced by different speakers (and of the stimuli of different length types in the first two sections). The test was self-paced; subjects were permitted to listen to a token more than once, although they were discouraged from doing so unless necessary. For sections 1, 2, and 3 of the test, the subjects were notified that some of the tokens may have been modified durationally. The entire testing session lasted approximately 30 minutes.

2.2 Training

The programs for training were written to be accessed through the Internet using a Java applet imbedded in an HTML webpage. The subjects ran the training programs from their home or office, on either Macintoshes or PCs.

2.2.1. Training Stimuli. Words in isolation and in the carrier phrase were recorded by a total of 12 native speakers of American English for the training data. All of the speakers were different from those in the testing. The majority of the tokens were recorded by three primary speakers. The reason for this is to allow the subjects first to learn to discriminate the tokens for a small number of speakers (less variability), and then to gradually introduce the subject to more variability. A total of 446 different stimuli were recorded. In addition to these natural tokens, a full set of durationally resynthesized tokens were also created. These tokens were manipulated so that the length of all of the tokens produced by a given speaker for a given context type (following voiced stop, following voiceless stop, following nasal, etc.) would all be the same length, thereby eliminating the length distinction between the tense and lax vowel.

The total corpus of stimuli was then divided into smaller groups, each consisting of about 30 words. Stimuli within a particular group usually had the same or similar phonetic contexts. Subjects trained on stimuli of one group at a time. The stimuli in the initial stimuli groups generally contained fewer speakers and fewer different words than the more advanced stimuli groups.

2.2.2. Training Format. The training uses a two option categorization task with immediate feedback. For each question, the subjects heard the stimulus, and then had to click on the button labeled with the word (“heat” or “hit”) containing the same vowel as the vowel in the stimulus. When the subject made a choice, he or she was given feedback as to whether the categorization was correct. If the subject had made the correct identification, the subject received positive feedback in the form of a star on the screen, and the identity of the word was also shown. In the case of an incorrect identification, the program
prompted the subject to listen to the stimulus again, and would not proceed to the next question until the subject had correctly answered the question. Subjects were able to listen to the stimulus as many times as they wanted to, and were encouraged to listen to the tokens more than once. In addition, subjects were able to listen to samples of the two words “hit” and “heat” at any point during the training.

2.1.3. Training Procedure. Each of the subjects participated in 10 training sessions of 300 tokens each over the course of two weeks. Each session took approximately 25-30 minutes for most subjects to complete. In order to maximize the subjects’ learning ([11]), the training program recorded the subject’s performance and adapted accordingly. Each subject was trained on words from one stimulus group at a time. The program automatically adjusted to the subject’s performance, increasing the likelihood in future trials of receiving tokens that the subject had trouble with, and decreasing the likelihood of receiving tokens that the subject had little trouble with. When the subject improved on the tokens within a set, the program then reviewed all of the tokens the had heard up to that point before proceeding to the next stimulus group.

2.3 Subjects.

Eight native speakers of Japanese completed the pre-test, training, and post-test of the study. Five of these subjects (three males and two females) trained on the “resynthesized” tokens, and three subjects trained on the “natural” tokens. An additional two subjects took the pre-test and started the training but did not complete it. Prior to taking the pre-test, the subjects were randomly assigned to one of the two groups. The median age of the “resynthesized” group was 31, and they had spent an average of 16 months in the U.S. The median age of the “natural” group was 35, and they had spent an average of 6 months in the U.S.

![Figure 1](image1.png)

**Figure 1:** Accuracy in perception on the *natural* tokens portion of the pre-test and post-test by each subject group.

![Figure 2](image2.png)

**Figure 2:** Accuracy in perception on the *uniform* tokens portion of the pre-test and post-test by each subject group.

of the pre-test and post-test by each subject group.

3. RESULTS

The performance of the two training groups on the pre-test and post-test are shown in Figure 1 (for the section with *natural* tokens) and in Figure 2 (for the section with *uniform* tokens). Single factor analysis of variance (ANOVA) was run on the test type for each of the training group in these two conditions. The improvement for both groups on the *natural* tokens section of the test was significant $[F(1,4) = 9.14, p<.05$ for those trained on natural tokens; $F(1,8) = 5.68, p<.05$ for those trained on resynthesized tokens]. Improvement on the *uniform* tokens was only significant for the group trained on resynthesized tokens $[F(1,4) = 0.29, p=.62$ for those trained on natural tokens; $F(1,8) = 12.14, p<.01$ for those trained on resynthesized tokens]. Due to the small sample sizes and the fact that the two subject groups performed significantly differently on the pre-test, it is not possible to say from this data whether the resynthesized training was more effective than the natural training, but the resynthesized training clearly was successful.

Table 2 contains the performance for the first section of the pre-test and post-test, which includes isolated words of *natural*, *long*, and *short* length. The overall performance on the tests improved for both groups. The performance on long /I/ and short /i/ on the pretest was poor, which was expected since these tokens provided the subjects with contradictory duration and vowel quality cues. For both of these token types, the group training on natural tokens outperformed the group of subjects training on resynthesized tokens.

Figure 3 displays the performance of the two subject groups on the *natural* tokens section of the test. In the pretest, performance by both groups was worst for Speaker 3, who had the least robust durational cues. However, on the post-test both test groups performed well on Speaker 3’s tokens, indicating that

<table>
<thead>
<tr>
<th>Overall</th>
<th>Natural /I/</th>
<th>Long /I/</th>
<th>Natural /i/</th>
<th>Short /i/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>Resyn.</td>
<td>0.59</td>
<td>0.87</td>
<td>0.68</td>
<td>0.91</td>
</tr>
<tr>
<td>Natural</td>
<td>0.75</td>
<td>0.86</td>
<td>0.85</td>
<td>0.87</td>
</tr>
</tbody>
</table>

**Table 2.** Performance on the first section of the pre-test and post-test by subject group

![Figure 3](image3.png)

**Figure 3:** Accuracy in perception on the *natural* tokens portion of the pre-test and post-test by each subject group by subject. For each speaker, natural subject group pre- and post-test are on the left; resynthesized group pre- and post-test are on the right.
they had learned to use the vowel quality cues effectively.

Figure 4 shows the average performance of each of the groups on the pre-test, during each session of the training, and on the post-test. There are two probable reasons that subjects’ performance in the training sessions starts out much higher than performance on the test. First, in the first two sections of the test, the durational cues were modified to deliberately mislead the subjects, while this was not the case in the training. Secondly, the training consisted of similar tokens grouped together, to facilitate the formation of robust phonetic categories according to phonetic environment. Thus, the subjects were able to quickly learn the two vowels in the limited phonetic environments that were included in the initial training. The rather flat performance across the training sessions is also to be expected. The tokens that the subject incorrectly identifies are played more frequently while the correctly identified tokens are played less frequently. Therefore, as the subject progresses through the training, the most difficult tokens are emphasized. In addition, as the subjects moved from one group of stimuli to the next, the groups became more difficult, encompassing more variation in speakers and in contexts within a group.

4. DISCUSSION

The present study duplicates other studies’ results which have shown the effectiveness of intensive computer-mediated training on the formation of new robust phonetic categories. Subjects in both of the two training groups were able to generalize the phonetic categories /i/ and /ɪ/ that they developed in training to new words spoken by new speakers, and they were also able to apply what they had learned (at least to a certain extent) to the tokens which had been durationally manipulated. In addition, the study shows that subjects can be effectively taught the contrast using stimuli which have been durationally manipulated. While the relative effectiveness of training using these resynthesized stimuli and of training using naturally produced stimuli needs to be further examined, the post-test performance of the two training groups on natural and uniform tokens indicates that training focusing only on vowel quality may be just as effective or even more effective than training with both vowel quality and vowel duration. In contrast with the suggestion of Strange et al. [11] that training include robust duration information, optimal training may in fact not include any duration information, thereby forcing the subject to use more native-like acoustic cues.

It should also be noted that perceptual abilities in an L2 are important not only for the non-native speaker’s ability to understand the language, but are also probably important for production in the L2 as well. For example, Lively et al. [8] found that perceptual training of native Japanese speakers on the /r~//l/ contrast in English produced a significant increase in their intelligibility of production of these phonemes. If L2 speakers tend to rely on non-native cues in their categorization of non-native phonemes, as the Japanese speakers in this study appear to do in the pretest, this may be related to a similar non-native use of cues in production, i.e. these subjects may distinguish the two vowels /i/ and /ɪ/ based on duration rather than on vowel quality (Minnick Fox and Maeda [13]). Therefore, the type of training discussed here which focuses on the proper acoustic cues may be essential not only for the subjects to improve their perception but also to achieve more native-like production.

REFERENCES