CHAPTER III

PERCEPTUAL TRAINING AND EFFECT ON PERCEPTUAL LEARNING

3.1. INTRODUCTION

This chapter reports the design, implementation, and results of the long-term training in the perception mode only. (Production tests will be reported in Chapter 4). It covers the pretest phase, the perceptual training phase, the post-test phase, and the 3-month retention test phase. The general goal of the perceptual training and tests is to investigate whether training is effective in perceptual learning of the three target vowel pairs. The four specific research questions to be addressed in this chapter are:

1) Will training be effective in focusing native Mandarin and Cantonese speakers’ attention to vowel spectral cues in identifying the synthesized target vowel contrasts?

2) Will trainees improve their perceptual accuracy of natural tokens of the three target vowel contrasts in terms of increased identification scores after training?

3) Will perceptual learning be generalized to new words and new speakers?

4) Will the improvement in perception be retained three months after training is completed?

While these four questions focus on the results of the perceptual training, the reasons for using both the fading technique and the high variability method in training will also be discussed.

3.2. PARTICIPANTS

A total of 29 (see Table 3-1) native Mandarin and Cantonese speakers (20 females, 9
males) were recruited from the Simon Fraser University student population through e-mail advertisements and friend referrals. They were all advanced English speakers between the ages of 18 and 42 at the time of the study. The differences in age within this range were not expected to have a large effect on the outcome of the training if they all had perceptual problems with the target vowel contrasts. The other reason to recruit adult speakers of this age range was because they were not likely to fail the hearing screening that was important for the perceptual training.

All 29 participants took the pretest that included the identification of synthesized vowel continua and natural stimuli contrasting the three target vowel pairs. The production test was also conducted at this stage. Five participants who were overqualified for the training were asked to withdraw from the study immediately after the pretest as preliminary analysis of their perception of the synthesized stimuli showed that they did not have obvious perceptual problems with any one of the three vowel contrasts. Detailed information about the preliminary analysis will be discussed later.

Of the 24 remaining participants who had a perceptual problem with at least one of the three target vowel contrasts, 5 served as control subjects and 19 began the long-term perceptual
training. During the training period, 3 more participants dropped out at different stages due to conflict of schedules or lack of interest for the long-term commitment. Sixteen trainees completed the training and took the post-test and the test of generalization immediately after the training. The control subjects did not take the training but took the same post-test and test of generalization at the same time interval. The trained group came back for the retention test three months after the training was completed. The information about the number of participants and their activities at different stages of the study is summarized in the above “Number of Participants and Their Activity Chart”.

Although it would be ideal to have an equal number of participants in the trained and control groups, the size of the control group was much smaller. One reason for the difference in numbers was because most participants were interested in the training and wanted to benefit from it. The other reason was based on the estimation and fear that some trainees might not be able to commit to the end of the long-term training. With the trainees outnumbering the control subjects at the beginning, it was estimated that the number of trainees at the end of the experiment would still remain above 10. Detailed background information about the trained and control participants who completed the study is presented in Table 3-1.

The trained group of 16 (10 females and 6 males) consisted of 13 native Mandarin speakers and three (two female and one male) native Cantonese speakers. Two of the Mandarin speakers were born and raised in Taiwan and the rest were born and raised in Mainland China. Two trainees also spoke a different dialect along with Mandarin as first languages. The three Cantonese speakers were born and raised in Hong Kong. At the beginning of the study, the mean age of the trained group was 26.3 years with a range of 18-
Most of the subjects came to reside in Canada in their 20s and 30s. The mean age of arrival in Canada for the trained group was 24.7 years with a range of 13-34 years. The average length of residence in Canada, or in an English-speaking environment, was 1.7 years (range: 1 month to 6 years). All participants had learned English in their countries of origin before they moved to Canada. Age of learning English (age at which English instruction began at school) varied from 4 -15 years with a mean of 10.9. The three Cantonese speakers from Hong Kong were exposed to English from kindergarten, much earlier than the Mandarin speakers from the mainland and Taiwan, who commonly began to take English courses at elementary or junior high school level. According to self-report, the trainees’ estimated daily use of English outside home ranged from 20%-100% with a mean of 52%. Almost all the participants spoke mainly Mandarin or Cantonese at home and the estimated use of English was only 13% with a range of 0-70%. In fact, 13 out of the 16 reported that they used English at home less than10% of the time. The two subjects who used English at home between 50-70% of the time were sharing living places with native speakers at the time of study.

The control group (4 females and 1 male) consisted of one Cantonese and four Mandarin speakers. The four Mandarin speakers were all born and raised in Mainland China. The Cantonese speaker was born and raised in Hong Kong. The control subjects had backgrounds similar to the trained subjects in estimated use of English and age of learning (see Table 3-1). For example, like the three Cantonese subjects in the trained group, the Cantonese speaker in the control group also had earlier exposure (at age 3) to English than the Mandarin speakers. The mean length of residence in an English-speaking environment was 3.2 years at the time of the study. It happened that three of the five control subjects were
older than the other participants, thus pushing the mean age of the control subjects to 34, somewhat higher than the mean age of the trained group of 26. However, the age of participants in both groups fell within the range of 18-42 at the time of study.

3.3. STIMULI

The stimuli for both testing and training consisted of synthesized and natural minimal pairs containing the three target vowel contrasts /l/-/t/, /u/-/u/, and /e/-/æ/ in mostly CVC syllables. Because the tokens used for tests and training were drawn from the same stimuli database, detailed information about token preparation will be described in the following two sections: preparation of synthesized stimuli and preparation of natural stimuli.

3.3.1. Preparation of Synthesized Stimuli

Synthesized vowel continua that differ in spectral and duration cues step by step are often used in perceptual tests to determine which cues the listeners attend to in distinguishing vowel contrasts. By careful design, synthesized vowel stimuli can be used to direct listeners’ attention away from duration to spectral cues in perceptual training. In the current study, three vowel continua exploiting the duration and spectral differences systematically were generated for the purpose of testing and training.

The English /hid/-/hid/, /hud/-/hud/, and /hed/-/hæd/ continua were generated using a Klatt (1980) synthesizer at 20-kHz sampling rate with 16-bit resolution in the cascade mode. All three continua were synthesized with six duration and six spectral steps producing 36 tokens per continuum. For the duration steps, the six longest vowels were 250 ms and the six shortest were 125 ms with a 25 ms increment between two steps. Formant contour was incorporated by using different values of F1 and F2 at the beginning and end of the vowel portion for each spectral step. The value of F3 was stable (no movement) for each of the six
spectral steps for the /hid/-/hid/, /hud/-/hud/ pairs. The first three formant values are presented in Table 3-2 for /hid/-/hid/, /hud/-/hud/, and /hed/-/hæd/ continua. F4 was fixed at 3300 Hz across all three continua. Vowel formant values were based on Klatt (1980) with adjustment to accommodate the spectral increment along the continuum. F0 was set at 125 Hz at the beginning and dropped gradually to 105 Hz at about the mid point of the vowel and then to 100 Hz toward the end. In order to increase stimulus variability to prevent listeners from adapting to the single F0 contour of the synthesized tokens in identification tasks, especially for listeners who are native speakers of Chinese languages and are sensitive to any pitch changes, three more F0 contours for each vowel continuum were incorporated, generating four continua for each vowel pair. These three extra vowel continua for each pair had exactly the same values in all respects except that they differed in F0.\(^1\) Two of the new continua had the same falling F0 contour of the original continuum but differed in F0 values. They were named “high” and “low” to differentiate them from the original F0 tokens. The “high” F0 contour had a range of values of 135-125-115 Hz, and the “low” had a range of 110-100-90 Hz. The fourth continuum had a rise-fall contour with a rising F0 value from the beginning till the vowel mid point followed by a falling F0 toward the end. F0 had a range of 110-120-105 Hz.

Each stimulus consisted of a 60 ms initial /h/ portion, a vocalic portion, and a 40 ms formant transition from the vowel to the final /d/. The initial /h/ was generated by manipulating the aspiration parameter (AS) of the synthesizer. The formant values of the vowel were set at the beginning of the /h/ but the voicing amplitude (AV) was set to zero.

\(^1\) Token variability is an important factor that affects perceptual learning (Logan et al., 1991).
throughout the /h/ portion and was turned on at the onset of the vowel\textsuperscript{2}. A naturally produced final /d/ was spliced after the formant transition because the attempt to synthesize a natural sounding final /d/ release failed to yield satisfactory results.

During the stimulus synthesis phase, three native English listeners provided feedback in an open-set identification test. They did not know the purpose of the task and were simply told to listen carefully and write down the randomly presented words. The results of this test showed that all the intended /hid/-/hud/, /hod/-/hod/, the spectrally unambiguous stimuli at spectral steps 1 & 2 and steps 5 & 6 in the continua, regardless of the differences in duration, were 100% correctly identified. As expected, the listeners varied in their responses only on the spectrally ambiguous tokens of the /hid/-/hid/, /hud/-/hod/ pairs. For the /hed/-/hæd/ pair, the three listeners did not reach agreement 100% of the time on the intended stimuli (at the two endpoints) in the open-set identification test. In response to the feedback from these listeners, the F1 and F2 value range of the /hed/-/hæd/ continuum was adjusted and then was tested on three different native listeners in another open-set identification task before it was finalized. The three new listeners identified the intended vowels of /hed/-/hæd/, the spectrally unambiguous first and last two spectral steps 100% correctly. All the synthesized words were normalized for peak amplitude using Sound Edit 16 software for playback in the test and training sessions.

\textbf{3.3.2. Preparation of Natural Stimuli}

Six native Canadian English speakers, three males and three females, participated as talkers for the natural stimuli. All speakers were born and raised in one of the two western

\textsuperscript{2} Phonetically, /h/ is a silent vowel with the configuration of the following vowel. Therefore, formant transition was not necessary for the /hV/ portion of the synthesis.
Canadian provinces, British Columbia and Alberta. They were drawn from the Simon Fraser University faculty and student populations at the time of the study. The reading list consisted of a total of 142 words that were mostly minimal pairs containing the target vowels, such as “seat”-“sit”, “bet”-“bat”, and “shooed”-“should”. The majority of the words were CVC, VC, CCVC, and CVCC monosyllabic words with different initial(s) and coda(s). A few minimal pairs with disyllabic words (e.g. beaten/bitten, and kettle/cattle) were also included in the /i/-/u/ and /e/-/æ/ pairs. No disyllabic words were found for the for the back vowel pair /u/-/o/. In fact, some common words like “book” and “boot” were also included without meeting the minimal pair standard that was set strictly for the other two vowel pairs. This is because an extremely limited number of minimal pairs can be found for the /u/-/o/ pair. The complete list of stimulus words is given in Appendix A.

Before recording, the speakers were given some time to read through the list. During the recording phase, they were told to read at normal speed. All the words were produced in a carrier sentence “Now I say ---.” Each speaker read the entire list twice. The readings, therefore, yielded two tokens per word by each speaker. Speakers read the words into a headphone-mounted microphone connected to a Marantz portable cassette recorder (PMD 201) in a sound-treated room. The recordings were then digitized at a sampling rate of 22050kHz with 16-bit resolution on an Apple Macintosh computer using Sound Edit16 software. The target words were separated from the sentence frame, normalized for peak amplitude, and saved for playback.

---

3Research findings indicate that intra-speaker variability in vowel productions is common (Munro et al., 2001). It is, therefore, important to avoid using a single token per speaker in vowel training.
A group of native English speakers screened all the natural tokens in a two-alternative forced choice identification task. All the natural tokens were accurately identified for the /i/-/ɪ/ and /u/-/ʊ/ minimal pairs. For the /e/-/æ/ contrast, the only confusion was found with minimal pairs that had nasal endings, such as “pen”-“pan”, “spend”-“spanned” and so on. As a result, all the /e/-/æ/ minimal pairs with nasal endings were eliminated from training and tests.

3.4. PRETEST

3.4.1. Procedure

The pretest, which lasted about one hour, consisted of a perception test and a reading task providing the production data. Before the test began, the participants filled out a general language background form that requested the information about their L1 background, L2 experience, and their length of residence in an English-speaking environment that was reported earlier. All participants passed a pure tone hearing screen (500-4000 Hz at 20 dB). The participant then proceeded with the production task (to be reported in Chapter 4) followed by the perception test.

The stimuli used for the perception test consisted of both synthesized and naturally-produced minimal pairs from two of the six speakers. The pretest was organized in such a way that each vowel pair, /hid/-/hɪd/, /hud/-/hʊd/, and /hed/-/hæd/ was presented in a separate block. These blocks are referred to as “front”, “back”, and “low” (only /æ/ in /hæd/ is a low vowel but the term was used for convenience) respectively. The synthesized stimuli, which were presented during the first session of each testing block, were the three vowel continua described earlier, each of which had six duration and six spectral steps. To assess the perceptual pattern of the listener when confronted with both clear and ambiguous tokens of
the target vowel pairs, the entire vowel continuum rather than the extreme tokens at the endpoints were included in the pretest. The 36 /hVd/ words per vowel continuum were repeated twice generating 72 tokens for each vowel contrast. For the sake of consistency, only one F0 contour (the original one) of the four available pitch patterns was used across the three vowel pairs.

Each test block also had a second session during which the participants identified minimal pairs produced by speaker 1 (male) and speaker 3 (female) (saving the rest of the speakers’ productions for the training and test of generalization). For the /hid/-/hid/ and /hæd/-/hæd/ pairs, three minimal pairs were included in the test. There were a total of 24 tokens per vowel pair (6 words × 2 speakers × 2 productions). For the /hud/-/hod/ pair, one minimal pair was included in the test unit as there was a limited number of minimal pairs available to cover the tests and training sessions. The natural tokens for /hud/-/hod/, therefore, were 16 (2 words × 2 speakers × 2 productions × 2 repetitions). The complete list of the words used in the perceptual pretest is given in Appendix B.

Individual test sessions for each listener were held in a sound-treated room. Before the listening test, the experimenter explained the identification tasks to the listener carefully to make sure that the participant understood the task thoroughly. Each listener then had a trial session to learn the test procedure before proceeding with the real test sessions. The test tokens were presented as two-way forced choice tasks with no feedback. The keyword labels used for the three target vowel contrasts were heed/hid, who’d/hood, head/had respectively. The order each listener identified the three vowel blocks was counterbalanced. Each stimulus was played only once and the listener identified the token that matched one of the two labels by clicking a button on the computer screen. As soon as the button was clicked, the next
token was played. The listener had control over the pace of the test by either immediately or delaying in clicking the button each time after he heard a stimulus. The test data were collected automatically by the computer and saved for subsequent analysis.

**3.4.2. Participants for Training: The Preliminary Analysis of the Pretest**

As reported earlier, before training began, a preliminary analysis of the pretest results of the synthesized stimuli for each participant was performed as a screening procedure to eliminate participants who did not show perceptual problems from the subsequent perceptual training. The analysis also helped to identify each listener’s perceptual problems. In order to determine which cue(s) the listener responded to in distinguishing the vowel contrasts, each listener’s identification scores of the spectral and duration endpoint stimuli (step 1 & step 2) of the three vowel continua were calculated. If a participant demonstrated a native-like perceptual pattern of responding to vowel spectral cues in distinguishing vowel contrasts, he would identify all the intended “heed”, “who’d”, and “head” tokens, (tokens at spectral step 1 and step 2) 100% of the time correctly. Taking into consideration of the possibility of occasional errors, a range of correct percentage identification scores of 83% -100% (10 our of 12 tokens) was set as standard. Participants who met this standard for all three target vowel contrasts were eliminated from the training. As reported in Section 3.3, five of the participants who showed clear patterns of using spectral cues for all three pairs met this standard and were asked to withdraw from the training experiment before it began. Their identification performance on the pretest was excluded from further analysis.

The twenty-four remaining participants who did not meet this standard either attended to duration cues or did not demonstrate clear perceptual patterns of responding to either spectral or duration cues in their identifications. (More in-depth analysis of individual data
will be presented later in the results section.) Although most of the participants had problems with one or two of the three vowel pairs, very few of them had difficulties with all three pairs. Therefore, each trainee began the training with different problems to focus on.

3.5. PERCEPTUAL TRAINING

3.5.1. Training Stimuli Organization

The perceptual training included both synthesized stimuli and natural tokens produced by four of the six talkers. Three training blocks were developed and named as “front”, “back”, and “low” respectively for the three vowel pairs /hid/-/hɪd/, /hud/-/hʌd/, and /hæd/-/hæd/. In each block, a few sessions with different minimal pairs were developed. The exact number of sessions differed as there were far fewer minimal pairs available for the /u/-/u/ vowel contrast than the /i/-/ɪ/ and /e/-/æ/ pairs. However, each block had the same three sessions of synthesized stimuli at the beginning and the end of the training.

A modified fading technique was adopted for the presentation of the synthesized stimuli. The fading technique (see Chapter 2 for detailed discussion) used by Morosan and Jamieson and others in a number of studies (Morosan & Jamieson, 1989) is believed to be effective in focusing the trainees’ attention on the key acoustic/phonetic properties that signal the differences between the target segments by exaggerating these differences at the beginning. In the process, the differences are gradually narrowed as the listeners gain experience. The reports of success with the fading technique, however, are based mainly on consonant voicing contrasts. Up to now, no record of its use on L2 vowel training has been found. One of the complications in choosing the fading method for such a task is that vowels, especially synthesized vowel tokens, may not be perceived strictly categorically as are consonants (See Chapter 1). For vowels, within-category discrimination is relatively good.
As the goal of the training is to help the subjects to form native-like perceptual categories of the target vowels, not to test their ability in perceiving within-category physical differences of the same vowel category, it is important to choose the appropriate stimuli along the vowel continuum for the purpose of fading. Additionally, as the nature of the identification task requires the predetermination of the categories for the labeling, special precautions have to be taken to prevent unreasonable arbitration of category formation due to training. For example, it may be possible but not appropriate to train the listeners for category distinction using stimuli from spectral step 3 and step 5. For these reasons, only the spectrally unambiguous tokens as identified by the native English speakers in an open-set task were chosen as training stimuli. In the first session of each training block, the spectrally most extreme tokens, those at the endpoints (step 1 and step 6) across all six duration steps were included.

Across the three blocks for the three vowel contrasts, session 1 included synthesized endpoint stimuli with two different F0 contours: the original falling and the rise-fall contours. In session 2, the same endpoint stimuli but with higher (H) and lower (L) F0 values were included (see the description of the F0 variation in the above section). These two sessions form the first phase of the training.

A “fading” session with less extreme stimuli in the spectral domain (step 2 and step 5) across all six duration steps was introduced later in the training process when the subjects advanced to the end after all the other tokens had been introduced. No attempts were made to explore the possibility of using stimuli from spectral step 3 and step 4 for further fading training as these tokens were categorically ambiguous even for the native speakers of English.

During the second phase of training, the modified high variability technique was
introduced using natural stimuli. Minimal pairs with mostly CVC and a few VC syllable structures contrasting the three target vowel pairs produced by four speakers were used to form training sessions for each block. As there were a limited number of minimal pairs available for the /u/-/u/ pair, some common CVC words (non-minimal pairs) were also included for training. The exact number of distinct sessions developed for the /i/-/u/, /u/-/o/, and /e/-/æ/ were 11, 6, and 9 respectively. These minimal pairs in all training sessions are listed in Appendix B.

As with the synthesized stimuli, natural tokens of the three target vowel contrasts formed three separate training blocks. Because three pairs of vowel contrasts were used for training in this study and each had a large number of tokens from multiple speakers, efficient organization for presentation of the stimuli was important. In previous training studies on Japanese /i/-/l/ contrast using natural tokens by multiple speakers, the stimuli were all blocked by talker (Bradlow et al., 1997; Logan, 1991; Logan et al., 1993; Logan & Pruitt, 1995). Each speaker’s tokens were identified separately in different training sessions and were repeated in a similar fashion. This means the trainees never heard the same minimal pairs by different talkers in a single training session. The reasons the researchers blocked the stimuli by talkers were not clearly stated but in the subsequent post-test and test of generalization, new talkers’ productions were measured against those of the familiar talkers.

In the current study, token presentation was blocked by vowels, not by talkers. Listeners would hear and identify the same number of minimal pairs by all four talkers in each session for each target vowel contrast. An important reason for including multiple talkers in each session was that this method might reduce the talker effect in perceptual training. Evidence from previous studies suggests the intelligibility of isolated words can be
affected by the differences in presenting the words produced by an individual speaker and by multiple speakers (Mullennix, Pisoni, & Martin, 1989; Pisoni & Lively, 1995). The results of these studies showed that a change in the talkers’ voices trial by trial affected listeners’ identification performance. In other words, the identification of words produced by a single talker was easier than the same words produced by multiple speakers. In order to help the trainees to adapt to the abstractness of differences between the target vowel contrasts through exposure to multiple talkers in each session, the mixed talker presentation of all minimal pairs was adopted for the presentation of tokens in the current study.

It would be ideal to control the phonetic environments as well as syllable structure differences in the selection of tokens for training, as the effect of phonetic contexts on the target vowels has been repeatedly shown in previous studies (Jenkins, Strange, & Trent, 1999; Munro et al., 2001; Strange, 1992; Strange et al., 1998). Results of training studies on consonant place and voice distinctions also showed that perceptual learning through training was local and improvement on one syllable structure, e.g., word initial or final did not automatically transfer to word medial positions or vice versa (Morosan & Jamieson, 1989; Logan et al., 1991; Rochet, 1995). In the current training design, an effort was made to include tokens in different consonant contexts. Minimal pairs with various syllable structures, such as CCVC or CVCC, as well as disyllabic words that were also recorded were not used in training but saved for the test of generalization.

3.5.2. Training Procedures

Individual training sessions were held in a sound-treated room on a Macintosh computer. The two-alternative forced choice tasks were the same as those in the pretest except that immediate feedback was provided after each trial in the training sessions. The
labels used were heed/hid, who’d/hood, and head/had for the three vowel pairs respectively across all training sessions, the same as those used in the pretest. Every time a trainee identified a token that matched one of the key word labels by clicking the corresponding button, immediate feedback in the form of words “Correct” or “Sorry, that should be xxx” flashed on the screen. The next token was then automatically played back for identification. By the end of the session, the program automatically calculated the percentage of correct identifications of the session and provided the score on the screen. The trainee could then choose to replay the whole session or move on to the next one.

A self-paced, trainee-centered procedure was adopted through the entire course of training. This flexible training approach was designed to meet the different needs of the trainees who had problems with one or two or even all three vowel contrasts under training and therefore progressed at different paces. This flexible training procedure was also adopted to keep the trainees motivated and committed to the end. Although the experimenter was present with each individual trainee throughout the whole training process to provide assistance in opening the training files and give suggestions for the pace of training, it was the trainee’s responsibility to commit to the schedules for each session in the two-month process. Due to differences in each participant’s schedule, beginning levels, and pace of progress, the number of training sessions each trainee completed varied. The average number of times the trainees came to the lab for training was 10 with a range of 6-15. During each training session, the number of blocks a subject completed also differed. Most trainees chose to spend more time on the /hɛd/-/hæd/ contrast because this pair appeared to be more resistant to learning than the other two pairs. In fact, a few trainees did not demonstrate perceptual problems with the /hʌd/-/hud/ or /hɪd/-/hid/ pairs at the beginning in the pretest.
and therefore concentrated only on the /hɛd/-/hæd/ pair in the subsequent training. Detailed information about each participant’s training progress is summarized in Table 3-3. The average number of sessions completed was 18 with a range of 7-44 for the /hɪd/-/hɪd/, 16 with a range of 3-32 for /hʊd/-/hʊd/, and 37 with a range of 21-70 for /hɛd/-/hæd/. Some trainees took a few trial sessions with either the /hʊd/-/hʊd/ or the /hɪd/-/hɪd/ pair but did not proceed with training on these two pairs because they did not have perceptual problems with them as demonstrated in the preliminary analysis of their pretest results. However, all were tested on three target vowel contrasts. It was expected that the training on the whole would have a certain effect on the trainees regardless of how many pairs they concentrated on in the process. It would be interesting to assess the impact of training on one vowel contrast on the learning of another vowel contrast, but it was out of the scope of the current study with only a limited number of trainees.

The trainees recycled the training sessions repeatedly at their own pace. Because of this, the entire course of training differed across trainees, with an average of 8 weeks. This may not be the ideal situation for a study, but it resembles more the real learning situation of ESL learners, who progress at different paces.

**3.6. Post-test and Test of Generalization**

Immediately after the training was completed, the sixteen trainees took the post-test that repeated the tasks of the pretest and a test of generalization. The control group of five also took the same post-test and test of generalization at approximately the same time interval between the pretest and post-test as the trained group. Therefore, the only known difference between the two groups was that the control group did not take the training. The test of generalization consisted of minimal pairs that were new tokens produced by the four familiar
speakers and old and new tokens by two new speakers whose voices were never heard in the
training and pretest. Like those used in the training and pretest, two productions of the same
minimal pairs by each speaker were included in the test of generalization to eliminate the
effect of intra-talker variability. The number of stimuli was 64 for /hid/-/híd/, 34 for /hud/-
/hùd/, and 52 for /hed/-/hæd/. (The difference in number of stimuli in the test of
generalization across the three target contrasts was due to the different number of stimuli
available for each contrast). The complete lists of stimuli for the three sessions of the test of
generalization are also presented in Appendix B. In addition to the CVC, VC syllable
structures that were used in the training, the test of generalization included CCVC, CVCC,
and CVCV disyllabic words that the listeners were never exposed to in the pretest or training
sessions. The test of generalization was presented as the same two-alternative forced choice
identification tasks without feedback.

3.7. RESULTS

The effects of perceptual training were assessed through a detailed examination of
changes in participants’ perceptual patterns for the synthesized stimuli and through a
calculation of changes in their identification scores on the natural stimuli from pretest to
post-test. If the perceptual training was effective, the trained group should demonstrate a
significant increase in identification scores on the natural tokens from pretest to post-test
while the control group should not exhibit such a pattern. The analyses of perceptual patterns
for the synthesized stimuli should demonstrate which cues, spectral, duration, or none
(random), the participants relied on for the vowel distinctions. If the training was effective,
the trained group should demonstrate clear and consistent reliance on the spectral cues for all
three vowel contrasts at post-test. Such a change in perceptual patterns for the synthesized stimuli from pretest to post-test is not expected from the control group.

**3.7.1. Test Results on Synthesized Stimuli**

3.7.1.1. Group Data

Participants’ percentage identification scores on “heed,” “who’d,” and “head” for the three vowel continua were calculated across each of the six spectral and duration steps in the pretest and post-test, and are presented in Figure 3-1, Figure 3-2, and Figure 3-3 respectively. The scores on each spectral step were pooled over the six temporal steps, and the scores at each duration step were pooled over the six spectral steps.

Overall, visual observations of the top panels of Figure 3-1 through Figure 3-3 suggest that both groups responded somewhat to the spectral cues and that performance of the trained and control groups was comparable at pretest. However, in the post test, while the number of “heed”, “who’d”, and “head” responses dropped or increased sharply as the spectral quality changed between step 3 and step 4 for the trained group, a similar change was not observed with the control group. This patterning suggests that training may have been effective in focusing trainees’ attention on spectral cues in the identification of synthesized vowels at post-test.

To quantify the trainees’ responses to spectral cues, average percentage identification scores (hereafter % ID) on “heed”, “who’d”, and “head” at spectral step 1 and step 2 across all six duration steps (the most “heed”, “who’d”, and “head” like tokens) and at spectral step 5 and step 6 across the six duration steps, (the most “hid”, “hood”, and “had”- like tokens) were calculated. By comparing the participants’ % ID scores at the two endpoints at pretest and post-test, the changes in responding to the spectral cues at post-test can be assessed. If
the trainees responded to the spectral cues more strongly as a result of training, their % ID on “heed”, “who’d”, and “head” should increase at spectral step 1 & 2 but decrease sharply at spectral step 5 & 6. Consequently, the % differences between steps 1 & 2 and steps 5 & 6 would increase from pretest to post-test. The % differences were calculated by subtracting the average % ID scores for spectral steps 5 & 6 from the average % ID scores for spectral steps 1 & 2. (Recall that these are the end points for the vowel continua that were all unambiguously identified by the native speakers of English.) The higher the % differences between the two spectral extremes, the more the listeners responded to the spectral cues for labeling the vowel contrast. For example, the value of the difference would be 100% if all the participants identified all the step 1 & 2 tokens as “heed” and step 5 & 6 tokens as “hid”. Therefore, this degree of spectral sensitivity would be reflected by the value of the % differences.

Figure 3-A. Two representative listeners’ percentage identification scores for synthesized heed/hid continuum at pretest and post-test

An example of two individual listeners’ perceptual scores on the heed/hid continuum are plotted and presented in Figure A here for illustration. As shown in Figure A, subject
M03 identified stimuli at steps 1 & 2 as “heed” 100% of the time and stimuli at steps 5 & 6 as “heed” only 8.5% of the time. The % difference between the two endpoints was 93.5% at pretest. Clearly, M03 was responding to the spectral cues for the heed/hid contrast at pretest. The difference increased to 100% at post-test, an 8.5% increase. In contrast, as shown in Figure B, another listener, C01, did not demonstrate a clear response to the spectral cues in her identification of the heed/hid continua in the pretest. Her % ID scores for “heed” were around 50% across all six spectral steps. The difference between the two endpoints was only 4% at pretest. At post-test, however, the % difference increased to 87.5%. Therefore, C01 clearly demonstrated a change in perceptual patterns at post-test. She responded mainly to spectral cues for the heed/hid identification.

This method was used to calculate the group mean data at both pretest and post-test. The percentage differences in ID scores on “heed”, “who’d”, and “head” between spectral steps 1 & 2 and spectral steps 5 & 6 for both the trained and the control groups were averaged and summarized in Table 3-4. At pretest, the mean differences in identification scores between the extreme spectral steps were 50%, 28% and 30% for “heed”, “who’d”, and “head” for the trained group, and 26%, 23%, 30% for the control group. At post-test, they were 90%, 95% and 80% for the trained group but 40%, 4%, and 39% for the control group. The % differences between the pretest and post-test were dramatically different between the trained and control groups.

A two-way ANOVA with group (trained and control) as a between factor, and test (pre and post), and vowel (front, back, low) as repeated measures revealed a significant main effect of group [F(1,19) = 31.609, p < .0001], an effect of test [F(1,19) = 34.601, p < .0001], and an effect of group × test interaction [F(1,19) = 23.667, p < .0001]. (Unless otherwise
indicated, the α level for analysis of variance in this study is set at .01 and .05 for follow up or post hoc tests). The effect of vowel [F(2,38) = .650, p = .5277] was not significant. The vowel × test interaction [F(2,38) = .393, p = .6778], group × vowel interaction [F(2,38) = 3.009, p = .0612], and group × vowel × test interaction [F(2,38) = .784, p = .4637] all failed to reach significance. A test of simple effects on the group × test interaction showed the differences between the trained and control groups at post-test were highly significant [F(1,31) = 53.576, p < .001], but the difference also reached significance at pretest [F(1,31) = 5.485, p = .026]. However, an effect of test was found for the trained group [F(1,19) = 40.426, p < .001] but not for the control group [F(1, 19) = .113, p = .740]. Therefore, only the trained group had substantial changes in responding to spectral cues at post-test.

Participants’ responses to duration cues were more complicated than those observed with the spectral cues. The mean % identification scores for “heed”, “who’d”, and “head” of the trained and control groups in pretest and post-test at each duration step (pooled over six spectral steps) are presented in the bottom panels of Figure 3-1, Figure 3-2, and Figure 3-3 respectively.

Visual observation of Figure 3-1 through Figure 3-3 shows that both the trained and control group demonstrated inconsistent reliance on duration cues across the three vowel continua. Both groups responded to duration cues for the heed/hid contrast at pretest: the longer tokens were identified as “heed” and shorter ones as “hid” along the duration continuum. At post-test, a shift away from duration cues to spectral cues was observed with the trained group, but not with the control group. In general, the trained group did not appear to respond to the duration cues for who’d/hood and head/had contrasts at pretest. The control group appeared to have identified the longer tokens as “hood” and “head” in the pretest, a
reversed use of duration cues. Visual inspection of Figure 3-2 and Figure 3-3 also suggests that the control group showed some unexpected changes with regard to attending to duration cues at post-test. Are these real changes brought about by the familiarization with the test procedures, or are they simply caused by the reversed use of duration cues at post-test? This question will be addressed when detailed analyses of individual data are carried out in the next section.

3.7.1.2. Individual Data

Each trained and control subjects’ % ID scores on “heed”, “who’d”, and “head” were calculated across each of the six spectral and six duration steps in the pretest and post-test (See Appendix C1- C15.) To quantify each individual participant’s perceptual changes in responding to the spectral cues at post-test, the same methods used in analyzing the group data were applied. Each trainee’s % differences between the spectral steps 1 & 2 and steps 5 & 6 from pretest to post-test are summarized in Table 3-4. Considerable individual differences were observed in responding to spectral and duration cues in the identification of synthesized continua at both pretest and post-test across the three vowel contrasts.

Two major perceptual patterns emerged when these individual differences were analyzed. First, across both the trained and the control subjects, the perceptual problems with the heed/hid contrast were all related to duration. Six of the 16 trainees and four of the five control subjects who had problems with the heed/hid contrast responded to the duration cues in the identification task at pretest. They identified long tokens as “heed” and short tokens as “hid”. To illustrate these problems at pretest and changes at post-test, four of the six trainees’ identification scores at six spectral and six duration steps on the heed/hid continuum are presented in Figure 3-4A. (The same figures presenting each individual listener’s scores are
provided in Appendix C). As seen from Figure 3-4A, the trainees’ attention shifted from
duration to spectral cues. In contrast, none of the control subjects showed parallel changes in
responding to duration cues at post-test (see Appendix C-13).

Second, a different pattern was observed with the who’d/hood and head/had than with
heed/hid contrasts at pretest. Among the trained and the control subjects, the perceptual
problems with these two pairs were not always related to duration. For illustration, four
selected trainees’ perceptual ID scores at six spectral and six duration steps on the
who’d/hood contrast are presented in Figure 3-4B. They are typical cases of the two common
perceptual problems observed with this vowel contrast. For example, as seen in Figure 3-4B,
C01 and C03 both had perceptual problems responding to spectral cues for the who’d/hood
distinction. Neither, however, responded to duration cues in their identifications of this
contrast. M05 and M16, who also had problems with the who’d/hood contrast, did appear to
respond to duration cues. While M05 identified longer stimuli as “who’d” and shorter ones as
“hood,” M16 did just the opposite. She identified the shorter tokens as “who’d” and longer
ones as “hood”. More cases of such reversed responses to duration cues were found with
other subjects in the trained and control groups (see Appendix C).

The same pattern was observed with the head/had contrasts. Four representative
trainees’ % ID scores at six spectral and six duration steps on the head/had pair were also
selected for illustration and are presented in Figure 3-4C. For example, both M13 and M08
had problems responding to spectral cues for the head/had contrast, but neither seemed to
rely heavily on the duration cues in the identification task. Two trainees, C01 and M12, who
responded to the duration cues for contrasting the head/had pair, both identified the long
tokens as “head” and short tokens as “had”. Other similar cases were found in both trainees
and control subjects at pretest. Training appeared to be effective in focusing trainees’ attention on spectral cues for the identification of the who’d/hood and head/had contrasts whether the perceptual problems were related to duration or not. As seen in Figure 3-4B and 3-4C, all these selected trainees (and others) who had problems with who’d/hood and head/had contrasts at pretest learned to respond to spectral cues at post-test. The control subjects did not show this change at post-test.

As mentioned earlier, the analysis of the group data showed that the control group did not demonstrate any noticeable improvement in responding to the spectral cues in their identifications of the three target vowel pairs at post-test. Yet, visual observation of Figure 3-2 and Figure 3-3 suggest the control group’s reliance on duration cues in the post-test was less obvious, as seen in the pretest for the who’d/hood and head/had pairs. Analysis of individual data revealed that this change in the use of duration cues was not due to improvement through familiarization with test procedures, as the control group did not show any increased responses to spectral cues at post-test. Rather, the observed changes were related to control subjects’ inconsistencies in responding to the duration cues at pretest and post-test.

For example, for the “who’d” and “hood” contrast, (as seen in Appendix C-14,) two of the five subjects (M18, M24) used reversed duration cues, that is, identified the longer stimuli as “hood” and shorter tokens as “who’d” in the pretest. At post-test, M18 preserved this duration pattern but M24 switched to use the normal, or opposite duration cues, as she identified the longer tokens as “who’d” and shorter ones as “hood”. A third subject (C05) who did not use obvious duration cues in the pretest was found to use reversed duration cues in the post test. The changes in the use of duration cues in the opposite direction affects the
group average at post test as shown in the bottom panel of Figure 3-2 and in Table 3-4. Therefore, the change in the group average seen in Figure 3-2 was probably not an indication of any improvement at all.

A similar pattern was found with the “head” and “had” pair. Four of the five control subjects identified the longer stimuli as “head” and shorter ones as “had” in the pretest (see Appendix C-15). At post-test, one subject (M19) of the four who responded somewhat to duration cues in the pretest did not do so at post-test. Another control subject (M18) reversed the use of duration cues. The shift in responding to duration cues in the opposite direction at pretest and post-test was the major reason for the changes observed in the group data as summarized in Figure 3-2 and Figure 3-3.

Like the trained subjects, the control subjects’ use of temporal cues for the / hud/-/hud/ and / had/-/had/ contrasts was not as consistent as with the / hid/-/hid/. Across three vowel pairs, (as seen in Appendix C-13 through C-15,) none of the five control subjects made any noticeable change in responding to the spectral cues in a systematic way in the post test. Their perceptual problems remained the same in the post-test.

3.7.1.3. Discussion and Interpretation

The findings reported here provide evidence that perceptual training with synthesized stimuli appeared to be effective in improving trainees’ perception of the three target vowel contrasts. At least some trainees shifted attention away from temporal cues to spectral cues in the identification of the three target vowel contrasts, especially the / hid/-/hid/ pair for which the use of duration cues was most persistent. Data from the control group in the post-test contrasted sharply with those of the trained group. First, the control subjects’ reliance on duration cues for the / hid/-/hid/ distinction remained unchanged in the post-test. Second,
while some subjects in both groups demonstrated spectral confusion or showed reversed use of duration cues in identifying the /hud/-/hud/ and /hed/-/hæd/ contrasts in the pretest, the trained subjects’ perceptual change in the post-test was not matched by the control group. Therefore, whether the perceptual problem was related to the use of duration cues or not, the control group did not match the trained group in responding to the spectral cues for the identification of all three target vowel pairs in the post-test. The fact that the change in perceptual patterns was observed with the trained group but not with the control group provides evidence that the perceptual learning was due to the training and not simply exposure to the test stimuli.

Analysis of pretest and post-test results for the synthesized test data also led to the conclusion that not all the perceptual problems observed in the pretest could be fully explained by the use of temporal cues. In fact, subjects who had problems with the /hid/-/híd/ pair (the trained subjects at pretest, and control group at both pretest and post-test) relied heavily on temporal cues for the contrast. They consistently identified long stimuli as “heed” and short tokens as “hid”. This perceptual pattern suggests that native Mandarin and Cantonese speakers who had problems with the /hid/-/híd/ contrasts almost always perceived the /hid/-/híd/ pair as two separate categories that contrasted in duration.

However, the same did not hold for the /hud/-/hud/ and /hed/-/hæd/ pairs, as not all the subjects who had problems with these two pairs relied on duration cues for the contrasts. Some subjects who had problems with the /hud/-/hud/ and /hed/-/hæd/ contrasts demonstrated a random perceptual pattern of confusion over the target contrasts but did not automatically attend to duration cues for the contrasts. Identification of long tokens as “hood” and “head” rather than the expected “who’d” and “had” was not uncommon for those
who actually attended to the duration cues for the /hud/-/hud/ and /hæd/-/hæd/ contrasts. Furthermore, some control subjects even reversed their use of duration cues from the pretest to the post-test. The spectral confusion and the inconsistencies in using the duration cues may suggest that these subjects did not have a two-category distinction for the /hud/-/hud/ and /hæd/-/hæd/ contrasts. It is likely that, for these subjects, the category contrast between /hud/-/hud/, or /hæd/-/hæd/ did not exist and both pairs were treated as single categories.

3.7.2. Test Results on Natural Stimuli

3.7.2.1. Pretest and Post-Test Results

The mean % correct identification scores for the natural tokens in pretest and post-test were compared for the trained and control groups on all three target vowel pairs, and the results are presented in Figure 3-5. Overall, from pretest to post-test, the trained group showed a 14 % increase\(^4\) on the /ɪ/-/ɪ/ pair, a 32% increase on the /ʊ/-/ʊ/ pair, and a 16% increase on the /ɛ/-/æ/ pair. For the control group, there was a 3% increase on the /ɪ/-/ɪ/ pair but 3% decrease on the other two pairs from pretest to post-test.

A two-way ANOVA with group (trained and control) as a between factor, and test (pre and post), and vowel (front, back, low) as repeated measures revealed a significant main effect of group \([\text{F}(1,19) = 9.775, p = .0056]\), an effect of test \([\text{F}(1,19) = 9.462, p = .0062]\), an effect of vowel \([\text{F}(2,19) = 14.642, p < .0001]\), and a significant group \(\times\) test interaction \([\text{F}(1,1) = 11.325, p = .0032]\). The vowel \(\times\) test \([\text{F}(2,1) = .466, p = .6307]\), and group \(\times\) vowel \([\text{F}(2,1) = .2.659, p = .0831]\), and group \(\times\) vowel \(\times\) test \([\text{F}(2,38) = .937 9 = .4007]\) interactions were not significant. A test of simple effects on the group \(\times\) test interaction showed a significant difference between groups at post-test \([\text{F}(1,35) = 20.335, p < .001]\) but not at

---

\(^4\) This refers to the change of % scores from pretest to post-test for convenience in comparisons across groups.
pretest \( [F(1,35) = .293, \ p = .592] \). A significant effect of test was found with the trained group \( [F(1,19) = 14.522, \ p < .001] \), but not with the control group \( [F(1,19) = .009, \ p = .925] \). In other words, the trained group’s accuracy on all three vowel contrasts improved significantly in the post-test while the same did not happen with the control group. Post hoc Tukey tests on the effect of vowel showed a significant difference between /i/-/o/ and /e/-/æ/, and between /i/-/u/ and /u/-/u/ pairs but not between /u/-/o/ and /e/-/æ/ pairs. The /i/-/u/ contrast appeared to be relatively easier to perceive than the other two contrasts. As there was no interaction of vowel and group, differences in performance between vowel contrasts held true for both the trained and control groups.

3.7.2.2. Results of Test of Generalization at Post-Test Phase

The % identification scores on the test of generalization on the three target vowel contrasts (administered at the post-test phase) for both the trained and the control groups are presented in Figure 3-6. A two-way analysis of variance with group as a between factor (trained and control) and vowels (front, back, low) as a within factor showed a significant main effect of group \( [F(1,19) = 44.803, \ p < .0001] \), an effect of vowel \( [F(2, 19) = 10.951, \ p = .0002] \), and a group × vowel interaction \( [F(2, 38) = 9.523, \ p = .0004] \). Tests of simple effects (α level at .05 for follow up test) revealed that the differences between groups were significant on all three vowel contrasts: \( [F(1,56) = 4.181, \ p = .046] \) for /i/-/u/, \( [F(1,56) = 58.817, \ p < .001] \) for /u/-/u/, and \( [F(1,56) = 7.995, \ p = .006] \) for /e/-/æ/ contrast. Therefore, the trained group outperformed the control group in the test of generalization on all three vowel contrasts. The effect of vowel was significant \( [F(2,38) = 4.370, \ p = .020] \) for the trained group and for the control group \( [F(2,38) = 12.07, \ p < .001] \). The performance differences on vowels in the test of generalization were found to be consistent with pretest
and post-test in that both groups identified the /i/-/u/ pair more accurately than the other two pairs.

Summing up, statistical analyses on the perception of natural stimuli showed that while the two groups’ accuracy did not differ significantly at pretest, the trained group was significantly more accurate in perceptual identification of all three target vowel pairs at post-test. Furthermore, the trained group performed significantly better on the test of generalization than the control group. Participants in both groups identified the front vowel pair /i/-/u/ more accurately than the /u/-/u/ and /e/-/æ/ pairs at pretest. Although the overall improvement from pretest to post-test was greater for the /u/-/u/ pair, the difference between vowels was still significant at post-test.

3.7.2.3. Analysis of Speaker and New Word Effect

In the above section, between group comparisons on the test of generalization revealed that the trained group significantly outperformed the control group on all three vowel contrasts. To further investigate whether the trainees’ performance on the new speakers and new tokens was comparable with their performance on familiar speakers and old words, a more detailed analysis on the test of generalization was performed.

The trainees’ average % identification scores on each vowel pair in the test of generalization were broken down to new speakers with new words, new speakers with familiar words, and familiar speakers with new words. The results are presented in Figure 3-7 (Top Panel). As seen from Figure 3-7, the trainees’ mean % ID score on new words with new speakers was actually 6% higher than their ID score on familiar words with new speakers for both the /u/-/u/ pair and the /e/-/æ/ pair. A 2% decrease was seen with the /i/-/u/ pair.

Performance was either the same (for /i/-/u/) or slightly better (for /u/-/u/) or slightly worse
(for /ɛ/-/æ/) across three vowel targets when new speaker/new word and familiar speaker/new word accuracy scores were compared.

A repeated measures ANOVA with vowel (front, back, low) and speaking condition (new speaker/new words, familiar speaker/new words, new speaker/familiar words) as within factors was performed. The results showed that while the effect of vowel was significant \( F(2,15) = 8.308, p = .0013 \), due to overall better performance on the front vowel contrast, the effect of speaking condition \( F(2,15) = 1.710, p = .1980 \) and vowel × speaking condition interaction \( F(2,15) = 1.711, p = .1594 \) did not reach significance. Trainees’ identification of new words and new speakers were as accurate as with familiar speakers and new words. These findings provided evidence that perceptual learning as an effect of training was transferred to new speaker and new words.

3.7.2.4. Individual Trainee Differences

Each trained and control participants’ % correct identification scores on the front, back and low vowel pairs at pretest, post-test, and test of generalization are presented in Table 3-5, Table 3-6, and Table 3-7 respectively. Also presented are each individual’s overall gains for each vowel pair, or the difference in % ID scores between pretest and post-test, which were calculated by subtracting the pretest scores from the corresponding post-test scores. Considerable individual differences in trainees’ perceptual performance on natural tokens were observed at both pretest and post-test as well as in the size of gains on all three target vowel pairs.

As seen from Table 3-5, the size of gain varied from 4% to 34% across the trainees for the /ɪ/-/ʌ/ contrast. The differences in the absolute % increase were due, at least in part, to subjects’ ID scores at pretest. For example, some trainees, such as C02, C03, M05, M07,
M09, and M17 had more “room” for improvement from their relatively low pretest scores and all had an increase of over 20% or 30% in the post-test. In contrast, a ceiling effect may have occurred for M03, M04, M08, M13, M14, and M15 who had no or little room for improvement as all had a high pretest ID scores that were already above 90%. For the control participants, although 3 out of the 5 had room for improvement, an increase was not observed, as was with the trained subjects. One participant, M24, however, had an increase of 17% although her perceptual pattern of heavy reliance on duration cues for the synthesized tokens remained unchanged. As reported earlier, all the trained subjects responded to duration cues for the /ɪ/-/ɪ/ contrast.

For the /ɛ/-/æ/ pair, eight trainees (C01, C02, M04, M05, M08, M09, M17, M20) had an increase of over 20% in identification scores of natural tokens from pretest to post-test. All had either serious or minor problems in the identification of synthesized tokens (either responded to duration cues or showed random identification patterns) at pretest. At post-test, all responded to the spectral cues (see Table 3-4). Three trainees, C03, M03, and M15, who did not demonstrate serious problem with the synthesized tokens, had an increase in perceptual identification scores between 9%-16% for the natural tokens on this pair. The other three, M12, M13, and M16 had a decrease in identification scores for natural tokens although all had learned to respond to spectral cues for the synthesized tokens. A ceiling effect may have occurred for two of them, who had very high pretest scores for the natural stimuli. For the control group, four of the five participants’ % identification scores for natural tokens decreased somewhat in the post-test. Only one had an 11 % increase in scores on natural token identification in the post-test, but her problems with the synthesized tokens remained unchanged. One participant had a decrease of 33% on natural token scores.
The greatest individual differences were observed for the /u/-/u/ pair. While the changes in perceptual patterns with the synthesized stimuli were consistent for all who had problems in the pretest, the % increase in scores on the natural tokens varied considerably across the trained subjects. About half of the trainees had an increase of 28%-100% at post-test. Only one trainee, M07, showed changes with the synthesized stimuli but had a decrease in scores for the natural tokens. This discrepancy was especially noticeable since she had a perfect identification score of 100% in the pretest. Another trainee, M04, who had no problems in the pretest on the synthesized and natural tokens also had a decrease of 21% in the post-test. The dramatic individual differences in performance on the back vowel contrast may be complicated by the possibility that some subjects were able to perceive the differences between vowel categories but were not sure with the key word labels representing the vowels. For example, one trainee’s identification scores went from 0% to 100%. It is possible that they reversed their use of key words for identification. A ceiling effect was observed for four subjects who had an increase of 0-6% to reach the 100% identification score. For the control group, one participant had a decrease of 33% in the natural token identification scores and another had an increase of 11%. For the other three, the pretest and post-test score differences were below 4%. As with the other two vowel pairs, there was no improvement on the synthesized tokens.

Finally, individual performance across the three vowel pairs was examined and compared to evaluate the general effect of perceptual training. As seen from Table 3-4 through Table 3-7, five trainees, C01, C02, C03, M09, and M17 had a steady and consistent change in perceptual patterns on all three target vowel pairs in both synthesized and natural token identification. Their overall size of increase in % ID scores for the natural tokens
ranged between 18% and 48%, averaged across the three target vowel pairs. For these trainees, the perceptual learning appeared to be general and consistent. The other trainees all had substantial gains on mostly one pair (a few on two pairs) and the effect of training was seen on both natural and synthesized stimuli. For these subjects, differences in the effect of training for the three target vowel contrasts were apparently due to different reasons. For example, M03, M08, and M20 had significant gains on the /e/-/æ/ pair primarily because their pretest performance on the other two pairs was almost perfect in identification scores and therefore there was no room for improvement. Similarly, a ceiling effect may be a factor for trainees like M12 and M13 who had significant gains on one or two pairs but limited increases and decreases on the other. However, individual differences in performance on the vowels at post-test were not all due to ceiling effects. For example, M04, M07, and M16 seemed to have differences in learning across the three target contrasts as they had a dramatic increase on at least one pair but also a noticeable decrease in identification scores on another pair in the natural tokens, although all demonstrated changes in perceptual patterns in the identification of the synthesized stimuli.

Could the individual differences in learning be influenced by the differences in the number of training sessions that each trainee had overall and on each target vowel pair? As the self-paced nature of this training experiment was aimed at meeting each trainee’s needs, the amount of training varied considerably, and the differences might be related to the individual differences in learning discussed above. To explore this possibility, each trainee’s total number of training sessions on the /e/-/æ/ pair was compared with their percentage increase in identification scores (post-test minus pretest scores). The results \( r = .019, p = .9463 \) failed to show a significant correlation between the two. A close examination of
individual data led to the same conclusion. For example, M17 had a significant increase of
30% from the pretest to post-test with a total of 23 training sessions on the /ɛ/-/æ/ contrast. In
contrast, M16 had a decrease of 8% in identification scores but had a total of 70 training
sessions. The effect of training and pace of learning differed dramatically between these two
subjects, yet the difference in the number of training sessions did not appear to be an
important factor in these effects.

No correlation tests between the number of training sessions and % increase in
identification scores were performed for the front and back vowel pairs. This was mainly
because some trainees took a few trial sessions on the front and back pairs but did not
proceed with more training, as they did not demonstrate serious problems with these two
pairs. As a result, their trial sessions were not comparable with the multiple sessions of those
who took the actual training. Visual examination of individual differences in training
sessions and the % increase in identification scores for the /i/-/u/ and /u/-/u/ vowel pairs also
suggests that no strict relationship between the two could be established. For example, for the
front vowel pair, the number of training sessions for M17 and M16, and C01 were 18, 32,
and 44 respectively, but the percentage increase was 34%, 9%, and 13%. Other such
mismatches could be found with the /u/-/u/ pair, as seen in Table 3-6.

It is important to point out that the different number of training sessions on each
vowel contrast across the trainees should not be taken at face value. Since the three vowel
contrasts under training have parallel spectral and duration differences, the effect of training
should not be viewed as isolated on each pair. For example, when a trainee learned to ignore
the duration differences in identifying the synthesized stimuli on the heed/hid pair, he may
very naturally apply the strategy to the other two contrasts. Therefore, more intensive
training on one of the three vowel contrasts out of each individual trainee’s needs may also have had some impact on the other two target contrasts for which the training was less intensive.

In summary, although the trained group demonstrated general improvement in the identification of synthesized stimuli by learning to respond to the spectral cues, considerable individual differences were observed for the size of gains in the identification scores of natural tokens from pretest to post-test. While a few trainees had a noticeable increase in identification scores on all three target vowel pairs, many had substantial gains on one or two pairs. One of the causes of the lack of increase in ID scores on all three vowel pairs was an apparent ceiling effect when there was not sufficient room for improvement. Individual differences in % identification scores were not correlated with the number of sessions the participants took.

3.7.3. Results of the Three-Month Test

To examine whether the perceptual learning of non-native vowel contrasts through long-term training was maintained without further training three months later, all 16 trained subjects returned for a retention test in which the same tasks as in the post-test were repeated, including the test of generalization⁵.

The results of the identification of the synthesized stimuli appeared to be comparable with the post-test in that the trained subjects continued to attend to the spectral cues for labeling the three full vowel continua. The average % identifications for “heed”, “who’d”,

---

⁵ The inclusion of the control group for the retention test would be irrelevant as there was nothing to be “retained” for the control group. Even if a “delayed effect” can be argued for the control group 3 months later, which is not likely in the current study, the retention test in this study is aimed only to assess whether the trained group's improvement through training at post-test was retained 3 months later.
and “head” are presented in Figure 3-8, Figure 3-9, and Figure 3-10, along with the pretest and post-test results. The scores on each spectral step were pooled over the six temporal steps, and similarly, scores of each duration step were pooled over the six spectral steps. As seen from these figures, the trainees responded strongly to spectral cues for all three vowel contrasts and the results of the 3-month test were almost identical to those of the post-test for heed/hid and head/had pairs.

To quantify the trainees’ responses to spectral cues in identifying the three vowel contrasts in the 3-month test, the same method for calculating the % differences in responding to spectral cues used at pretest and post-test was applied to measure the changes in the 3-month test, if any, from the post-test. The percentage differences were calculated by subtracting the average % ID scores of spectral steps 5 & 6 from the average % ID scores of spectral steps 1 & 2. The trained group’s average differences in identification scores between the extreme spectral steps were 89%, 90%, and 76% for “heed”, “who’d”, and “head” respectively, a very slight decrease from the 90%, 95% and 80% at post-test. The corresponding % differences were 50%, 28% and 30% at pretest.

To compare the trainees’ performance in responding to spectral cues on the three vowel contrasts at the 3-month test with their performance at pretest and post-test, a repeated measures ANOVA with test (pre, post, and 3-month) and vowel (front, back, low) as factors was performed. The effect of test \( [F(2,30) = 84.8965, p < .001] \) and the effect of vowel \([F(2,30) = 7.796, p = .0019]\) were both significant. The vowel \( \times \) test interaction \([F(4,60) = .275, p = .89232]\) was not significant. A Post hoc Tukey test on the effect of test revealed the differences were significant between pretest and post-test, and between pretest and 3-month test, but not between post-test and 3-month test. The differences on vowel contrasts was
significant between the heed/hid and head/had pair, and between the who’d/hood and head/had pair, but not between the heed/hid and who’d/hood pair. These findings provide evidence that the trained subjects continued to respond to spectral cues in identifying synthesized vowel continua three months after the training was completed. Trainees’ performance in using spectral cues in their identifications of synthesized vowel contrasts were better on the heed/hid and who’d/hood pair pairs than on the head/had contrast across all three tests.

The perceptual identification scores on natural tokens at the 3-month test are presented along with those of pretest and post-test in Figure 3-11. The mean % ID scores were 95%, 83%, and 79% for /ɪ/-/ɨ/, /u/-/o/, and /ɛ/-/æ/ respectively. They dropped slightly from the post-test scores of 97%, 89%, and 84%, but were still substantially higher than the corresponding pretest scores of 83%, 57%, and 68%.

To compare the trained group’s ID scores on natural tokens at the 3-month test with those at pretest and post-test, a repeated measures ANOVA with test (pre, post, and 3-month) and vowel (front, back, low) as factors was performed. The effect of test \( [F(2,15) = 22.329, p < .0001] \), and vowel \( [F(2,15) = 6.705, p = .0039] \) were significant but the vowel \( \times \) test interaction \( [F(4,30) = 1.937, p = .1558] \) was not. A Post hoc Tukey test revealed that the differences were significant between pretest and post-test, and between pretest and 3-month test, but not significant between post-test and 3-month test. For the vowels, the differences were significant between /ɪ/-/ɨ/ and /u/-/o/, and /ɪ/-/ɨ/ and /ɛ/-/æ/ pairs. As there was no interaction between test and vowel, the improvement on all three target vowel pairs was significant from pretest to post-test and was retained at the 3-month test.
In the 3-month test, the trainees also took the same test of generalization with stimuli by new talkers and new stimuli by old talkers as they did at the post-test phase. To examine whether the trainees retained their performance on the test of generalization at the 3-month test phase, the results of the two tests of generalization were submitted to a one-way repeated measures ANOVA with vowel (front, back, low) and test (gen 1 (at post-test), gen 2 (3-month test)) as two within factors. The effect of vowel \( [F(2,15) = 6.397, p = .0049] \) was significant but the effect of test was not \( [F(1,15) = .403, p = .5353] \), nor was the interaction of vowel and test \( [F(2,30) = .268, p = .535] \). A Post hoc Tukey test revealed that performance on vowels was significantly different only between the /ɪ/-/ɨ/ and /ɛ/-/æ/ pairs. Trainees continued to perform better on the /ɪ/-/ɨ/ contrast than on /ɛ/-/æ/.

As with the first test of generalization done at the post-test phase, the trainees’ performance on the test of generalization in the 3-month test was also broken down to new speakers with new words, new speakers with familiar words, and familiar speakers with new words. The results are presented in the bottom panel of Figure 3-7. The effect of word and speaker was not observed at all for the back vowel pair. Only slight differences in speaker and word were seen with the front and low vowel pairs. A one-way repeated measures ANOVA with vowel (front, back, low) and speaking condition (new speaker/new words, familiar speaker/new words, new speaker/familiar words) as within factors revealed that neither the effect of vowel \( [F(2,15) = 2.772, p = .0786] \), nor speaking condition \( [F(2,15) = .001, p = .99921] \), nor the vowel × speaking condition interaction \( [f(4,30) = .994, p = .4450] \) was significant. The results of the comparison showed that the learning was effectively transferred to the new words and new speakers, as shown in the earlier test of generalization, and was retained three months later.
Summing up, a comparison of the results of the 3-month test with those of the pretest and post-test on the three vowel contrasts established that the effect of perceptual training was still comparable with that of the post-test three months later. The results of the generalization test provide further evidence that perceptual learning of the three target vowels through training was generalized to new talkers and new tokens and this generalization was maintained three months after the training was completed.

3.8. DISCUSSION AND INTERPRETATION

3.8.1. Summary of Perceptual Learning

Training appeared to be effective in modifying native Mandarin and Cantonese speakers’ perceptions of the three target vowel contrasts. Referring back to the research questions raised at the beginning of this chapter, the results of the perceptual tests provide positive answers to each one of the four specific research questions. First, trainees who either relied heavily on the temporal cues or were confused with the spectral cues in identifying the synthesized stimuli in the pretest learned to respond to spectral cues in the post-test. The same change was not observed with the control group. Therefore, training appeared to be effective in modifying native Mandarin and Cantonese speakers’ perceptual patterns. Second, the trained group’s perceptual identification scores on natural tokens increased significantly (14 % for the /i/-/u/ pair, 32% for the /u/-/u/, and 16% for the /e/-/æ/ pair) from pretest to post-test. The significant improvement in identifying natural tokens was not matched by the control group. Third, in the test of generalization, trainees identified the new words by new talkers as well as familiar and new words by familiar talkers. Therefore, perceptual learning was generalized to new talkers and new words. Fourth, the trained groups’ perceptual performance on both the synthesized and natural tokens on all three pairs of vowel contrasts
at the 3-month test was comparable with their performance at the post-test. Perceptual learning was retained three months after training was completed.

The results of the current study also demonstrated that not all the perceptual problems observed with Mandarin and Cantonese speakers could be fully explained by the use of temporal cues. In fact, participants who had problems with the /hid/-/híd/ pair almost all relied heavily on temporal cues for the contrast. They consistently identified long stimuli as “heed” and short tokens as “hid”. However, the same strategy was not always used for the identification of /hud/-/hud/ and /héd/-/hæd/ pairs. Instead of relying on duration cues, participants who had problems with these two pairs often demonstrated confusion in responding to either spectral or duration cues in identifying the synthesized /hud/-/hud/ and /héd/-/hæd/ stimuli. This random identification pattern suggests that they did not have a clear two-category distinction for /hud/-/hud/ and for /héd/-/hæd/ contrasts. Even those who responded to duration cues for /hud/-/hud/ and /héd/-/hæd/ did not always identify the long tokens as the expected “who’d” and “had”. The fact that some control participants switched direction in the use of duration cues in the post-test as compared with the pretest provided evidence that the use of duration cues was not consistent in the identification of /hud/-/hud/ and /héd/-/hæd/. The findings are in agreement with an earlier study (Wang & Munro, 1999) in which Mandarin speakers were found consistently relying on temporal cues for /hid/-/híd/ but not for /hud/-/hud/ distinctions. With the additional findings on the /héd/-/hæd/ contrasts, the current study provided evidence that native Mandarin listeners’ persistent reliance on duration cues for English vowel contrasts was limited to the /i/-/i/ pair only. Depending on their perceptual problems, trainees had to learn to shift from the use of temporal cues to the
relevant spectral cues to distinguish the /ɪ/-/u/ contrast, and to develop phonetic categories for the /u/-/ʊ/ and /ɛ/-/æ/ contrasts.

The results of the perceptual tests also showed that both the trained and control groups identified the natural tokens of the /ɪ/-/u/ contrast better than the /u/-/ʊ/ and /ɛ/-/æ/ contrasts at pretest and post-test. In particular, the /ɛ/-/æ/ contrast appeared to be more resistant to learning, and the trainees spent considerably more training time on this pair than the other two pairs. One reason for the difficulties in learning the /ɛ/-/æ/ contrast might be the fact that the spectrally very close /ɛ/-/æ/ contrast is highly confusing. The difficulties might also be exacerbated by the mixed presentation of the minimal pairs by multiple talkers.

The findings provided evidence that both native Mandarin and native Cantonese speakers had perceptual problems with English vowel contrasts. As the goal of the current study was not to compare Mandarin and Cantonese speakers’ differences in their problems on the three target vowel contrasts in relation to their L1 experience, the number of speakers from each L1 background was not controlled for such a purpose. However, the analysis of individual data suggested that Mandarin and Cantonese trainees’ problems were comparable regardless of their L1 experience. For example, although Cantonese makes use of duration differences in its vowel system and Mandarin does not, the Cantonese trainees in the current study did not appear to rely more on duration cues than the Mandarin subjects in the identification of the synthesized vowel stimuli. (Only one of the three Cantonese trainees relied systematically on duration cues for two (/ɪ/-/u/ and /ɛ/-/æ/) of the three vowel contrasts.) Similarly, although Cantonese has an allophonic /ɪ/ and /ʊ/ and Mandarin doesn’t, the Cantonese subjects’ problems with the /ɪ/-/ɪ/ and /ʊ/-/ʊ/ contrasts were comparable to those of the Mandarin subjects. As discussed in Chapter 2, Cantonese vowel quality changes
are affected by syllable structure differences. The lowered quality of /i/ and /u/ to /ɪ/ and /ʊ/
in closed syllables might account for the Cantonese subjects’ perceptual difficulties of the
English /i/-/ɪ/ and /u/-/ʊ/ contrasts. However, no matter at what level problems in perception
of the non-native vowels occur, the difficulties in learning the contrasts remain similar. The
learning of L2 vowel contrasts at the phonetic level appears to require the same training for
Cantonese speakers as for Mandarin speakers.

3.8.2. Evaluation of Training Techniques Used

Several observations about methodologies used in the current training study on non-
native vowel contrasts warrant discussion. First, the fading technique that has been used
repeatedly in training on consonant contrasts appeared to be applicable for vowel training.
Second, the mixed presentation of natural tokens by multiple speakers in training might have
been effective in reducing the talker effect that is commonly seen in other studies. Third, the
combined use of synthesized stimuli with natural words at different stages in training non-
native vowel contrasts exploited the advantages of both stimulus types. These methodologies
will be discussed in connection with the results in the following sections.

3.8.2.1. Combined Use of Synthesized and Natural Stimuli in Training

In the current study, the fading technique was modified and used for training non-
native vowel contrasts. Synthesized stimuli at the end points of a continuum were included in
the initial training sessions followed by a fading session with phonetically less extreme but
still categorically unambiguous tokens. This technique is not much different from what has
been commonly used in training on consonant contrasts, except that its use with vowels
requires some special precautions. This is because within-category distinction is relatively
good for vowels, especially synthesized vowel tokens. Therefore, in the current study, only
categorically unambiguous but phonetically less extreme synthesized tokens were included in the fading session.

In addition to the spectral and duration differences in the synthesized stimuli, the end point vowel prototypes and the fading sessions were presented with different F0 values and pitch contours to distract the trainees’ attention. With immediate feedback trial by trial, the listeners were forced to ignore the temporal cues and F0 variations and to attend to spectral cues only in labeling the vowels. Some trainees reported that they were constantly distracted by the change in F0 patterns in the training but learned to ignore the irrelevant information to concentrate on the spectral cues through training with immediate feedback. Different F0 patterns incorporated in the stimuli appeared to be useful in increasing the variability of the synthesized stimuli that usually do not represent the full array of speaker information in natural tokens. The change in F0 values and contours trial by trial might be especially distracting to native Mandarin and Cantonese speakers who are sensitive to tonal changes because of their L1 experience.

Although the use of synthesized stimuli for training on L2 consonant voicing contrasts appeared to be sufficient for perceptual learning to be transferred to natural talkers (Jamieson & Morosan, 1986; 1989), similar findings with L2 vowel contrasts have not been reported in any previous work. There is also evidence that training with synthesized stimuli on non-native consonant place contrast was not sufficient for transfer to natural stimuli (Strange & Dittman, 1984). It would be ideal to control the results of the current study by a third group of subjects to be trained on synthesized stimuli only and tested on the natural tokens. This would allow the comparison of the effect of training with two different types of stimuli. However, the trainees’ problems with the natural stimuli observed both at pretest,
during training, and even at post-test suggest that subsequent training with natural tokens was necessary for a stronger effect on perceptual learning. Besides, as the goal of the study was to promote learning through training rather than comparing the effectiveness of token types for training, both synthesized and natural tokens were used to maximize the effectiveness of training.

3.8.2.2. Mixed-Talker Presentation of Training and Test Tokens

The high-variability identification paradigm in which natural stimuli produced by multiple talkers were used for training has been proved to be effective in training Japanese speakers to identify the English /ɪ/-/ɻ/ contrasts. Despite the general success of this methodology in training, the talker effect was reported to be persistent during training and in tests of generalization (Bradlow, 1997; Logan et al., 1991). One of the reasons they offered for this phenomenon was that talker-specific information was acquired along with the target contrast in the training process (Lively et al. 1994). Given that the Japanese trainees were exposed to multiple talkers throughout the training process, the fact that their performance level still varied as a function of talker raises questions about the methodology used in training, if the intrinsic intelligibility of the talkers was controlled as they reported. A talker effect might be, at least in part, due to the fact that the training stimuli were blocked and presented by individual talkers. As the listeners heard only one talker’s words repeatedly throughout a training session, they may have automatically become adjusted to the talker-specific information and may have encoded the information along with the key phonetic cues that signal the target consonantal contrasts.

In the current study, the mixed-talker presentation of highly variable natural words produced by multiple talkers in each training session increased the variability of talkers. Of
course, any claim about the advantage of mixed-talker presentation of test and training tokens over the individual talker presentation must be based on a controlled test with different training groups. As the current study was not intended to test these differences in token presentation methods, no comparisons could be made. However, it is speculated that the change of voices from trial to trial might have discouraged the listeners from attending to talker-specific information. Instead, the subjects had to focus on the phonetic cues to the target contrasts at each trial that was accompanied by different voice information, forcing them to extract the abstract, key acoustic/phonetic information from each token presented. Even if talker information was acquired along with the acoustic information intended for the training, the exposure to multiple voices in each training session appeared to have reduced the talker effect. The results of the test of generalization demonstrated that familiar talkers were not at any advantage when their productions of minimal pairs were presented with those of the new talkers. The same level of performance on the new and old talkers was observed for all three vowel contrasts, including the highly confusing /e/-/æ/ distinction and was maintained three months after training stopped. Future training studies on reducing talker effects might take into consideration these token presentation techniques for better results.

The combined use of synthesized and natural tokens that are presented in a way that increases the variability of both stimulus types has not been explored in previous studies. The current study has provided empirical data for exploiting the effectiveness of both stimuli types. The results suggest that listener’s problems with the “weighting” of cues in perceiving English vowel contrasts could be effectively addressed by synthesized vowel tokens at the initial stage of training. The richness of the natural token stimuli from multiple talkers in subsequent training increased the uncertainty and may have facilitated learning through
increased tolerance of phonetic variability. This stage of training was important to help the subjects to develop phonemic categories for the target vowel pairs that were abstract enough for them to ignore the sub-phonemic details and idiosyncratic talker characteristics in real speech perception.

3.8.3. Perceptual Learning and Speech Learning Model

Finally, the results of perceptual learning through long-term training seem to support the claim of Flege’s Speech Learning Model that perceptual categories can be established if listeners can perceive the differences between the target vowel contrasts. In addition, phonetic categories can also be modified if non-native speakers learn to shift from attending to duration to relying on the relevant spectral cues for non-native vowel contrasts through training.

Perceptual training appeared to be successful in the perceptual mode. Will the perceptual learning of these vowel contrasts affect the trainees’ production? This question leads to another major research goal of the current study: the effect of perceptual learning on production. Chapter 4 will report and discuss the results of the production test.