CHAPTER II

REVIEW OF PERCEPTUAL TRAINING ON NON-NATIVE SEGMENTAL CONTRASTS

In Chapter 1, the causes of difficulties in cross-linguistic speech perception and production were explored from several different perspectives. Despite these difficulties, research has shown that adult L2 learners’ perceptual patterns can be modified through immersion in the ambient language environment or through intensive laboratory training (Bradlow et al., 1997, Lively, Logan, & Pisoni, 1993; Logan et al., 1991, 1993; Logan & Pruitt, 1995). This raises the issue of how malleable adult L2 learners can be when learning the sound system of a second language.

A number of important questions concerning the modifications (of perceptual patterns) in L2 speech learning need to be addressed. First, to what extent can adult L2 learners’ perceptual patterns be modified through training? Second, what are the constraints that cause difficulties for such modifications? Third, what methodologies are the most efficient? Fourth, to what extent will the effect of perceptual learning be retained and to what extent will it be transferred to production?

Some of these questions have been explored in training studies over the last two decades. In particular, researchers have explored the effect of perceptual training on adult speakers’ perception of non-native contrasts under laboratory conditions. In this chapter, a number of training studies will be reviewed. The findings of the studies will be discussed along with the methodologies used in the training. A general overview of training methodologies and assessment commonly used in laboratory training studies will be provided in the following section.
2.1. METHODOLOGIES AND ASSESSMENT OF TRAINING

The use of appropriate methodologies is essential when using perceptual training to modify the perception of non-native contrasts. The choice of appropriate tasks to meet the goals of the training can be complicated by interactions among the stimulus variables, task variables and subject characteristics. The following sections introduce training methodologies commonly used in training studies. Issues covered are the types of training tasks, the selection of stimuli for training and testing, and the assessment of the effect of training.

2.1.1. Identification vs. Discrimination Tasks

In training studies, a fundamental distinction is made between discrimination and identification tasks. In a discrimination task, the listeners must determine whether the stimuli presented are the same or different. Discrimination tasks include three basic forms: the category change paradigm, the AX (same/different) paradigm, and the ABX (oddity) paradigm. In a category change paradigm, often used in infant speech perception, category A stimuli are usually presented repeatedly till, at a certain time, they are replaced with category B stimuli. The listener is expected to signal the change of category as soon as category B is introduced. In the AX or the same/different paradigm, the task involves the presentation of two tokens, one after the other in a single trial, and the listener has to decide whether the two tokens are the same or different. A variation on the AX task is the categorical AX task in which the stimuli in the “same” pairs are physically different tokens drawn from the same phonemic category, while stimuli in the “different” pairs are drawn from distinct phonemic categories (Logan & Pruitt, 1995). Another form of discrimination task is the ABX or oddity task in which the X is the same category as either A or B. Three tokens are presented in a
single trial and the listeners must decide whether the third token is the same as the first token (A) or the second token (B).

In an identification task, the target sounds are presented one by one in a forced-choice paradigm. Listeners identify a stimulus presented to them and label it among the choices they are given. Immediate feedback is provided after the listeners label each stimulus. Two common identification procedures in an identification task are the fading technique and the high variability method.

In an identification task using the fading technique, which was invented and used by Jamieson and Morosan (1986), listeners begin by identifying the most extreme tokens at the ends of a synthetic continuum. They then progress gradually along the continuum to the less extreme tokens. The crucial phonetic/acoustic properties signalling the category differences are most clearly presented (sometimes exaggerated) at the endpoints of the synthetic continuum. Towards the centre of the continuum, the phonetic/acoustic properties signalling the category differences fade away gradually and become ambiguous. The fading technique is aimed at helping the listener to attend to the relevant phonetic/acoustic properties for category distinctions, beginning by presenting the most easily perceived phonetic/acoustic differences and ending with phonetic/acoustic differences near the categorical boundary.

The high variability method is often used to train Japanese speakers on the English /u/-/l/ contrast (Bradlow et al., 1997; Logan et al. 1991; Lively et al. 1993). It emphasizes variability in training stimuli in terms of differences due to different speakers and different phonetic environments. Listeners are exposed to natural stimuli produced by multiple talkers with minimal pairs contrasting the target sounds in all possible phonetic environments.
In general, discrimination tasks have not been used in recent training studies, perhaps because the results of earlier studies suggested that these tasks might not be optimal for inducing changes in phonetic categorization (Logan & Pruitt, 1995). It has also been suggested that discrimination tasks in general may be relatively ineffective in training listeners to perceive novel phonetic categories. The reason for this is that they tend to focus listeners’ attention on within-category phonetic/acoustic differences and fail to concentrate on the core acoustic properties that are essential for category formation (Jamieson & Morosan, 1986, 1989).

In contrast, identification tasks are widely believed to be more efficient at improving, in a linguistically significant way, adult learners’ perception of non-native phonemic contrasts (Jamieson & Morosan, 1986, 1989; Lively et al., 1994; Logan et al., 1991; Logan & Pruitt, 1995). Identification tasks, especially with token and speaker variability, are believed to force listeners to focus on the phonetic cues that underlie categorical distinctions and to enhance long-term recall (Logan et al., 1991). These claims are based on the results of individual studies with different target contrasts. Identification procedures have dominated recent training studies.

One study (Flege, 1995b), however, directly compared discrimination and identification procedures in a controlled manner. Two random groups of native Mandarin speakers received training on English /t/-/d/ stops in word-final position. One group received two-alternative forced choice identification training while the other group received AX same/different discrimination training. Both groups improved in their performance on the trained target contrast. The findings of this study challenged the general claim that identification training is superior to same/different discrimination training (Flege, 1995b).
However, Flege pointed out that there was still evidence that identification training yielded greater non-statistically-significant gains than did the same/different task. He also found that the subjects who received identification training responded more positively toward the training. He suggested that this might be the reason why identification training was favoured in general. Further studies need to be carried out to determine which method is more beneficial, especially studies involving different targets, such as place of articulation contrasts that are believed to be more resistant to learning (Strange, 1992). So far, identification training remains the preferred procedure, despite limited, if any, evidence of its superiority.

2.1.2. Stimuli

Synthetic and naturally produced tokens are both commonly used in training studies. One clear advantage of synthetic stimuli is that they allow for the manipulation of key acoustic differences. For example, it is possible using synthesized stimuli to exaggerate or shrink the differences along a continuum. However, a disadvantage of synthetic stimuli is that they do not represent the full range of acoustic properties that specify phonetic categories. As the ultimate goal of perceptual learning is to enable learners to perceive the target contrasts in real speech, not synthesized talk, success in training that uses synthesized stimuli entails finding ways to ensure that newly-acquired perceptual patterns are transferred to natural tokens. Using natural stimuli would prevent this problem, especially when multiple speakers are included in order to present sufficient talker variability. However, the choice between the two types of stimuli really depends on the nature of the training tasks. For instance, the fading technique would require the use of synthesized stimuli in order to control the phonetic properties. On the other hand, for high variability training, multiple talkers’
natural tokens make it possible to represent the required range of differences. To address both needs, a combination of both stimulus types may be a suitable approach.

2.1.3. Duration of Training

Training in L2 speech learning can be either long term or short term depending on the type of contrasts and goal of training. According to Logan & Pruitt (1995), short-term training refers to training that does not extend beyond one session on one day. It has been used with some success in training on stop voicing and vowel duration contrasts (Pisoni, Aslin, Perey, & Hennessy, 1982; Wang & Munro, 1999).

Training that occurs over several days or weeks is generally considered long term. In previous work, the duration of long-term training has varied from as little as six sessions (Rochet, 1995) to 45 sessions (Yamada, 1993). The duration of most long-term training studies to be discussed in the following sections of this chapter falls between these extremes. There is no fixed standard for the number of sessions or hours of training for a long-term training design. Moreover, very few studies have set any criteria as justification for the termination of training. Many studies seem to have had a predetermined number of sessions of training, and post-tests have been conducted as soon as the training sessions end. Factors like the subjects’ rate of progress, their level of interest and motivation, and the resources to maintain the subjects have all played a part in determining the actual length of training.

2.1.4. Assessment

To assess the change in subjects’ performance as a function of training, almost all training studies have adopted the pretest - training - post-test design. A simple comparison of the performance in the pretest and post-test, in terms of percentage increase in accuracy or reduced response time, has been carried out after training. In addition, in order to rule out the
possibility of improvement due to practice with the test, many studies have also included a control group that participates only in the pretest and post-test but not the training. Any gain by the control group can be assumed to be due to familiarity with the test tokens, and not to the effect of training. Additionally, it is common to include tests of generalization to assess whether perceptual learning has transferred to new talkers, new stimuli, new phonetic environments, and new syllabic positions in which the target sounds occur. As a rule, these generalization tests are not a part of the pretest and are always administered along with the post-test (Logan & Pruitt, 1995; Bradlow et al., 1997). Some researchers have also assessed long-term retention by calling back participants and administrating the post and generalization tests three and/or six months after the end of training. In some recent studies, (Rochet, 1995; Bradlow et al., 1997), tests of generalization have included an assessment of production improvement in the post-test. In these cases, the subjects provided spoken samples both before and after the perceptual training, along with the baseline perceptual data. Any improvement in the post-test in the production mode on the target sounds would be taken as an effect of perceptual learning generalized to production.

2.2. TRAINING STUDIES

In this section, perceptual training studies addressing different cross-linguistic phonetic problems will be reviewed. These include early training studies on VOT series, early studies on place of articulation contrasts, more recent studies on voice and place of articulation contrasts with more refined methodologies such as the fading technique and high variability paradigm, and a limited number of production training studies. This review is aimed at establishing the achievements that have been made in this area and identifying the unexplored issues that motivate the current study.
2.2.1. Early Studies on VOT Series

Early attempts at modifying adult learners’ perceptions of non-native contrasts focused mainly on the voicing distinctions in VOT phenomena. In one early study, Pisoni et al. (1982) investigated whether adults whose L1 stops made use of a two-way voicing distinction could be trained to perceive a 3-way distinction. In a series of experiments, monolingual English speakers of American English were exposed to synthesized /ba - pa - pʰa/ stimuli in an ABX task with feedback (training) and an identification task without feedback. The majority of the naïve subjects could identify an additional perceptual category in voicing after a short period of training.

McClaskey, Pisoni, and Carrell (1983) replicated and extended the Pisoni et al. 1982 study by showing that listeners were able to transfer the perceptual learning of a third voicing category to a new place of articulation without further training. Carney, Widin, and Viemeister (1977) found that after a relatively short period of training with an AB discrimination task, an ABX task, and an identification task, trainees showed excellent within-category discrimination in all three stop categories. Further training led to improvement in perceptual ability at both intra- and inter-category boundaries at arbitrary values predetermined by the experimenters. As a result, the authors argued that the perception of stop consonants does not need to be categorical. Although this study was not aimed at exploring the perceptual improvement in L2 speech contrasts in a linguistically meaningful way (perceptual sensitivity to intra-category boundaries), the results indicate a strong potential for the ability to modify adults’ perceptual capabilities under laboratory conditions.


2.2.2. Early Studies on Place of Articulation Contrasts

Attempts to train adult listeners to perceive other types of contrasts have also been reported. In one of these studies, Tees & Werker (1984) tested and trained native English speakers on the Hindi dental voiceless-aspirated and breathy-voiced (/ȟa/- /ďa/) contrast, and the retroflex (/t̃a/) and dental (/t̄a/) place contrast. When multiple natural CV tokens were presented to the listeners using the category change paradigm, the place distinction was found to be significantly more resistant to perceptual learning and required more training trials than the voice distinction.

The majority training studies on consonant contrasts have focused on native Japanese speakers’ perception of the English /l/ and /r/ distinction. One early study by Strange & Dittmann (1984) adopted the same/different (AB) discrimination procedure with immediate feedback using a synthesized “rock”-“lock” continuum for training, and used both synthesized and natural stimuli for testing. Learning was transferred to the untrained “rake”-“lake” series and also to more demanding identification and ABX tasks. However, learning did not extend to naturally produced words. The authors concluded that modification of perception of some phonetic contrasts in adulthood was “slow and effortful.”

The limited success of Tees and Werker’s (1984) study on place of articulation, and the lack of transfer to natural tokens of the Strange & Dittman (1984) study was later often attributed to inappropriate training methods (Logan & Pruitt, 1995). Specifically, paradigm change and AB discrimination tasks with feedback may not be the most efficient procedures for helping listeners to concentrate on the key acoustic information that distinguishes the two trained phonemes (especially if the contrast entails place distinctions that appear to be resistant to learning).
To summarize, although training with discrimination tasks along with some identification tasks without feedback were found to be sufficient to modify English listeners’ perceptions of a third VOT category (Pisoni et al., 1982), training with Paradigm Change or AB discrimination tasks were not very effective in modifying perception of place of articulation contrasts (Tees & Werker, 1984). The latter techniques did not appear to maximize the effect in terms of transfer to untrained stimuli and naturally-produced words. Nevertheless, these early studies have provided important empirical data on the effectiveness of laboratory training to modify adult learners’ perceptions of consonant distinctions. In addition, they suggest that the most potent training methodology depends on the type of the contrast to be learned and the aim of the study (including the specific goals of the training tasks).

2.2.3. Fading Technique: Studies on Voice Contrasts

Jamieson and Morosan carried out several studies to train Canadian francophone adult speakers on the English /θ/-/ð/ contrast. In the first study of this series (Jamieson & Morosan, 1986), they adopted a fading technique. Only synthesized stimuli were included for training, while both synthesized and natural tokens were included in the tests. In most cases, the trainees improved in their perceptions of the pair of tokens that reflected the categorical or phonemic differences. No differences in the ability of the trainees to perceive within category phonetic/acoustic differences were observed. Thus, the fading technique, with carefully selected synthesized tokens, appeared to be effective in focusing the listeners’ attention on the linguistically significant aspects of the voicing contrast. The most significant result of this study, the authors claimed, was that perceptual learning transferred to the
identification of natural tokens that were not included in the training. Therefore, the training task was successful in improving discrimination in a linguistically meaningful way.

Morosan & Jamieson (1989) replicated and extended their 1986 study to determine whether learning would transfer to new voices and new phonetic contexts that were not included in the training. The pretest and post-test included both synthesized tokens and natural stimuli from multiple speakers producing /o/ and /ð/ in different phonetic contexts. The tests also included the /ð/ and /d/ contrast, another contrast which is problematic for francophones that was not included in the training. The training procedures were the same as those in Jamieson and Morosan (1986). The success of the earlier study was replicated; however, learning did not generalize to different phonetic environments, such as word medial and word final positions for the /o/ and /ð/ contrast. There was also no improvement evident for the identification of /ð/ and /d/. These findings suggest that when the target contrast appears in a different syllabic environment, it exhibits phonetic differences that need to be learned separately. Similarly, the perceptual learning of one contrast, /o/ and /ð/, did not help the listeners with the other new contrast, /ð/ and /d/. Learning is therefore local and context-dependent. However, transfer to new voices was obtained.

Rochet (1995) used a modified fading technique that was similar to that of Jamieson and Morosan (1996) to train native Mandarin speakers on French voiced and voiceless stops (distinguished by different VOT values). Subjects participated in six 30-minute training sessions in which they identified synthesized French labial stops (/pu/-/bu/) and were given feedback. The trainees improved significantly in the post-test, and their identification patterns approximated those of native French speakers. Although learning did not transfer to intervocalic positions, the effect of training was extended to different vowel contexts and to
two other places of articulation not included in the training. The perceptual learning also
generalized to natural tokens not used in training. Furthermore, the improvement in
perception was accompanied by better production. The generalization of perceptual learning
to production will be discussed in later sections.

It is important to point out that, like the Morosan and Jamieson (1989) study,
Rochet’s (1995) study showed that training on the target contrast in word initial position was
not automatically transferred to word medial or intervocalic positions. Although there was
evidence of transfer of learning to different places of articulation and to a different vowel
contrast in the same CV position, the acoustic cues to voicing in initial and intervocalic
positions appear to be sufficiently different that they need to be trained separately.

In another study, Jamieson and Morosan (1989) compared the effect of training using
the fading technique with the effect of training using prototypical stimuli. In prototypical
stimuli training, good exemplars of synthesized stimuli representing prototypical /θ/ and /ð/
were used. They were presented in CV syllables in an identification task with immediate
feedback. In the post-test, the listeners showed improvement in their ability to identify both
the synthesized and the natural tokens. Prototype training was as effective as the fading
technique in improving the listeners’ identifications of natural stimuli not used in the test.
However, with synthetic stimuli, training with prototype stimuli produced less improvement
than the fading technique in terms of identification scores. Similarly, in the discrimination
test the prototype training was not as effective as the fading technique with the synthesized
stimuli. This was especially true for those tokens that had reduced frication durations. The
authors concluded that the crucial issue lies in the specification of the prototype that would
be ideal for the phonemic category under study. Even native speakers of the same linguistic
community may not have the same location for such a prototype. The fading technique was therefore favored over the prototype training.

In summary, Jamieson and Morosan’s series of studies have provided important empirical data regarding the effectiveness of perceptual training procedures. More importantly, they have provided insights into the relative effectiveness of different training techniques. For example, the use of the fading technique appears to be effective in focusing listeners’ attention on the most relevant acoustic/phonetic information for category formation. However, training using synthesized stimuli only may still not be “ideal” for perceptual learning of the target contrast across all phonetic/syllabic conditions. There is evidence that perceptual learning is local and context-dependent and the treatment of the target contrast may not automatically transfer to the same contrast in different phonetic positions. It is, therefore, advisable to include in training multiple tokens of all the relevant phonetic contrasts in all environments. It is also important to point out that these studies dealt with voicing contrasts only. These are believed to be relatively less resistant to training than place of articulation contrasts. Future training studies adopting the same training method need to explore the question of how place of articulation contrasts are best trained.

2.2.4. High Variability Training on the English /l/ and /l/ Contrast

During the last decade, Lively, Logan, Pisoni, Yamada and colleagues have conducted numerous training studies on Japanese speakers’ perceptual learning of the English /l/ and /l/ contrast (Bradlow et al. 1997; Logan et al., 1991; Lively et al., 1993; Lively et al., 1994). Their work has used a high-variability technique emphasizing the variability of the training stimuli in terms of speaker differences and phonetic environments. In the first of the series (Logan et al., 1991), six native Japanese speakers were trained to
identify English /ʌ/ and /l/ in different phonetic environments. Minimal pairs produced by
native English speakers were blocked by speaker and presented for identification with
immediate feedback. While identification improved from the pretest to post-test in general,
there was an effect of phonetic environment and an effect of talker in the test of
generalization. Familiar talkers’ productions were more accurately identified than those of
new talkers. The authors concluded that “listeners encoded detailed talker-specific
information and apparently stored this information in long-term memory” (p.881).

Despite the general success of the Logan et al. study (1991) with the high variability
training procedure, its experiment design and methodology were criticized on several
grounds by Pruitt (1993). Among other problems, such as lack of control group and
overstatement of the results of the test of generalization, Pruitt pointed out that the
experimental methodology provided no clear evidence on which type of variability is helpful,
variability in the talkers, the phonetic context, the stimulus tokens, or the syllabic
environment.

Although Logan et al. (1993) argued that it was not their intention to investigate the
specific contribution of each source of variability in their study, they did consider the effects
of talker and phonetic environment separately in subsequent studies. In fact in a later study,
(Lively et al., 1993), separate experiments were carried out to compare the results of two
different procedures. In the first experiment, multiple speakers’ productions of minimal pairs
with target sounds in three different phonetic environments (initial singleton, initial cluster,
and word medial) that posed the most difficulties were included in the training. In the second
experiment, minimal pairs produced in five phonetic environments by a single talker were
used. The purpose of the design was to separate the two kinds of variability, variability due to
multiple speakers and variability due to phonetic environment, and to measure their contributions to perceptual learning. In terms of overall improvement at post-test and the generalization to new talkers and new tokens, high variability training with tokens from multiple talkers was found to be more efficient than training with a single talker but more phonetic environments. Both groups showed the effect of phonetic contexts as the performance in different phonetic contexts varied considerably.

The third study of this series (Lively et al., 1994) tested the long-term retention of perceptual learning by administering two retention tests three and six months after the training was completed. Significant improvement at the post-test was retained at the 3-month test and was still partially retained in the 6-month test. However, as in the previous two studies, there were effects of talker and phonetic environment.

Despite the general success of the high variability training on the English /u/ and /l/ contrast, it is important to note that the effects of talker and phonetic environments were persistent across all studies. Trainees identified the tokens of familiar talkers more accurately than those produced by new talkers. The fact that the listeners were exposed to a single speaker rather than multiple speakers during each training session may have played a role in the talker effect observed in all these training studies.

The high variability procedure used in the Japanese /u/ - /l/ studies was also adopted in training studies focusing on supra-segmental contrasts. For example, in a recent study, Wang, Spence, Jongman, and Sereno (1999) used natural tokens produced by multiple speakers to train native English listeners on Mandarin lexical tone differences. The trainees’ identification scores increased significantly after two weeks of training, and perceptual learning was generalized to novel words produced by new speakers. More importantly, the
retention test three months later indicated that the effect of training was maintained without any further training.

To summarize, identification training using high variability techniques has proved to be effective on non-native contrasts at both segmental and supra-segmental levels. Training with stimuli that represents the full range of acoustic phonetic cues that characterize the target contrast helps the listeners to form categories and overcome idiosyncratic differences. The talker effect that has appeared consistently across studies needs to be addressed in future research.

2.2.5. Effect of Perceptual Training on Production

The link between perception and production in L2 speech learning has been investigated in a number of studies (Bohn & Flege, 1990, 1992; Bohn, 1997; Flege et al., 1997; Sheldon & Strange, 1982; Wang, 1997). However, these studies focused on the speakers’ performance in both modes at a certain point in time and therefore explored a correlational relationship between the phenomena in questions (Bradlow et al., 1997). Direct assessments of better production as a function of perceptual learning of a non-native contrast through laboratory training have been very limited and have been reported only recently. One of the series of studies on perceptual training of the Japanese /a/ - /l/ contrast investigated changes in production after 3-4 weeks of perceptual training (Bradlow et al., 1997). The design of the study was the familiar pretest, post-test, and test of generalization, using high variability natural tokens. After 45 sessions of training, subjects were tested on the /a/ - /l/ contrast in both the perception and the production modes. The trainees’ productions of the /a/ - /l/ contrast were better identified than their productions at pretest. As the training was
provided in perception only, the authors claimed that the effect of training was transferred to production.

Rochet (1995) also reported that perceptual learning was transferred to production in his training study of native Mandarin speakers’ perceptions of French stops. After six 30-minute perceptual training sessions in which they identified a synthesized French /pu/-/bu/ series with feedback, the Mandarin trainees improved in both perception and production of the voiceless stops. Acoustic measurements of the Mandarin subjects’ productions showed a change in VOT values that approximated native French productions.

An effect of perceptual training on production was also reported in studies of children with phonological impairment (Rvachew & Jamieson, 1995). For example, in the treatment problems that preschool children had with English /ʃ/, Rvachew (1994) incorporated perceptual training in addition to the traditional production therapy. Two groups of children received relevant perceptual training on /ʃ/ while a third group received the same amount of training on words not related to the target /ʃ/ sound. In the post-test, those who received the relevant training demonstrated a greater ability to articulate the target sound in comparison with those who received the irrelevant training.

In another study, Jamieson & Rvachew (1992) examined the effectiveness of perceptual training on the production of English fricatives by children with functional articulation disorders. Several synthetically produced English fricative stimuli designed to direct the children’s attention to the critical differences between the target phoneme and their incorrect substitutions were used as training tokens (/s/ for /ʃ/, etc.). The stimuli were presented in an identification task, a modified fading technique was used and feedback was given. Children with both production and perception difficulties showed significant gains in
production following perception training. However, perception training did not improve speech production either in children who showed normal perception ability during the pretest or in those who showed abnormal perception during the pretest but who failed to learn the perceptual distinction. The authors concluded that for children who have both abnormal perception and production, an appropriately designed program of perception training could quickly improve speech production. Perceptual training had no clear effect on production if the problem did not lie with perception.

Studies of phonologically impaired children and children with functional articulation disorders provide further evidence for the effect of perceptual training on production. However, like the other training studies, this research has focused exclusively on the modification of consonant contrasts. Comparable work on vowel contrasts has rarely been reported. Recently, Wang & Munro (1999) trained Mandarin listeners to identify English /hid/-/hīd/ and /hud/-/hūd/ pairs in a single session. The synthesized stimuli that systematically manipulated the temporal and spectral differences in the target contrasts were presented along with immediate feedback. Although training appeared to shift listeners’ attention from temporal to spectral cues in their perception of the synthesized vowels, the study did not test whether the effect of training was transferred to natural tokens or to the production mode.

In one recent study Yamada, Strange, Pruitt, & Masuda (1998) investigated the effect of L2 vowel perception training on vowel production. After 36 sessions of identification training on the American English vowels /a/ /æ/ /ʌ/ in CVC syllables over a period of nine days, the native Japanese speakers’ vowel production scores (as assessed by native speakers of English) showed no improvement. Acoustic analysis of vowel formant frequencies showed
large differences between the Japanese and native English speakers’ productions. However, in a related study, Strange and Yamada (1997) found that native Japanese speakers’ perceptions of the same English vowels in CVC syllables improved significantly (from 42% to 85%) after perceptual training. The findings suggest that although there was greater improvement in perceptual learning of the vowel contrast than the contrast between /I/ and /l/, vowel production ability did not improve as substantially as it did for /I/ and /l/. The authors concluded that “vowel distinction in production is more difficult to acquire than for consonants because their articulatory control is continuous while the articulatory control of consonants is somewhat discrete” (Yamada et al., 1999, p 2972).

Although most production problems are often assumed to be related to perceptual problems, the observation with children who had functional articulation disorders gives rise to another important question: Will perceptual training be helpful to those who have production problems without functional articulation disorders but no serious perceptual problem? Or, is there such a thing as a “pure” production problem that has no perceptual basis? No convincing empirical data have been presented so far, at least from training studies. Obviously, it would be helpful to examine production training and assess its effect on perceptual learning.

2.2.6. Production Training

Studies of production training in L2 speech learning are extremely rare. Leather (1997) carried out a parallel perceptual and production training study on native Dutch speakers learning the four citation-form lexical tones of standard Chinese (Mandarin). One group received perceptual training followed by a production test, while the other group had production training but was tested on perception. Both groups demonstrated improvement in
perception and production regardless of the mode in which they received training. Leather concluded that learners did not need to be trained in production to be able to produce the tonal contrasts or in perception to be able to perceive them.

2.2.7. Summary of Previous Studies and Problems for Future Studies

Summing up the results of training studies discussed in the above sections, although evidence on vowel contrasts is still very limited, there is substantial evidence that well-designed training in the laboratory is generally effective in improving listeners’ perceptual learning of non-native contrasts at the segmental and supra-segmental levels. The size of the gains in perceptual identification scores vary from about 5%-25% across studies. Some studies have shown that learning can transfer to different places of articulation and to new talkers that were not directly dealt with in the training. A few studies have also provided evidence for long-term retention of the training results three and six months after the training. Furthermore, perceptual learning can generalize to better production of the target contrast without production treatment. Similarly, training on production can be effective in the perceptual learning of contrasts. In addition, a number of techniques have been tested, improved and assessed to maximize training effects.

Despite the successes with the training studies in general, a number of problems and limitations can be found in the study of modification of non-native contrasts. One of the limitations is that transfer of learning appeared not to extend to new syllabic positions. Perceptual learning appeared to be local and required treatment at a context-dependent level. A similar limitation has been found with transfer of learning to new talkers. In a number of studies (Logan et al. 1991; Lively et al., 1993; Lively et al., 1994) the trainees performed better in identifying the target contrasts produced by familiar speakers even months after the
training was completed. This problem appeared to persist even when trainees were exposed to multiple speakers (the high variability training paradigm). Part of the problem might be related to the way the training tokens were presented in the training sessions. (The trainees were exposed to a single speaker’s tokens in each training session.) Future studies may expose listeners to the tokens of multiple speakers in each training session to maximize the talker variability during the training process. The effect of talker, the effect of phonetic environment, and other problems observed in the reported training studies need to be addressed in future studies.

Another limitation of the previous studies is that the research has concentrated almost exclusively on consonant contrasts. Vowels continue to be overlooked, especially in long-term training studies. This asymmetry in L2 training studies limits the claims that can be made about training, especially in terms of the long term retention of the learning results and the transfer of perceptual learning to production of L2 vowels. For example, so far, there have been no reports on the success of perceptual learning of L2 vowel contrasts being transferred to production or being retained months after the training is completed.

2.3. THE CURRENT RESEARCH: BACKGROUND

This thesis examines native Mandarin and Cantonese speakers’ perceptual learning of the English /i/-/I/, /u/-/ɔ/, and /e/-/æ/ contrasts as a result of long-term training under laboratory conditions. These three pairs of vowels were chosen for training because they pose serious problems for Mandarin and Cantonese speakers as indicated by previous studies (Bohn, 1995; Flege et al., 1997; Munro et al., 2001; Rogers, 1997; Wang, 1997). The other reason for choosing these vowel pairs is that they are structurally parallel in terms of spectral
and duration differences. The treatment of the problems in one pair may be beneficial for the other two pairs.

Wang (1997, 1998) found that native Mandarin speakers demonstrated persistent problems in distinguishing certain English vowel contrasts in both perception and production. In particular, acoustic analysis of Mandarin speakers’ English vowel productions showed that they often substitute Mandarin vowels for English vowels that do not have counterparts in the Mandarin system (Wang & Munro, 1998). The findings of these studies suggest that Mandarin speakers’ perception and production of some L2 vowels can be strongly influenced by their L1 vowel system. Native Cantonese speakers have also demonstrated serious problems in perception and production of English vowel contrasts that do not exist in Cantonese system (Hung, 2000). As discussed in Chapter 1, adult learners of a second language have difficulties in acquiring the sounds of the target language often because they perceive the sounds of L2 in terms of L1 phonology. In order to gain a better understanding of Mandarin and Cantonese speakers’ perceptions of English vowels, a brief phonetic description of Mandarin and Cantonese vowel systems will be provided here.

It is well known that a direct comparison of vowel systems across languages at the phonemic level is insufficient or even misleading for predicting which particular segments are problematic for the learners of the target language. (Bohn & Flege, 1992; Brière, 1966; Flege, 1995b; Munro et al., 2001; Odlin, 1989; Rochet, 1995). As L2 learners are sensitive to the differences at positional or phonetic levels (Flege, 1995a), detailed cross-language descriptions and analysis at the acoustic/phonetic level is required. For example, although both English and Portuguese have /i/ and /u/ and both lack the /y/ category in the high vowel space, while English speakers generally assimilate French /y/ to /u/, Portuguese speakers are
more likely to assimilate it to /i/. The differences in the acoustic/phonetic properties of /i/ and /u/ as perceived by Portuguese and English listeners explain the different assimilation patterns (Rochet, 1995).

For a better understanding of cross-linguistic perceptual patterns, some recent researchers have adopted the method of direct measurement by mapping the L2 segments directly onto the L1 system through cross-language labeling of speech sounds. This is done by having L2 listeners identify and label L2 sounds in terms of L1 sounds. (Baker & Trofimovich, 2000; Flege et al., 1997; Munro et al., 2001; Strange et al., 1998). Direct measurement is found to be especially useful for assessing the cross-language assimilation of positional variants of the target contrasts. Although the current study is not intended to carry out a detailed analysis of the cross-linguistic vowel contrasts and the assimilation patterns, it is important to note that there are different ways of comparing the phonetic similarities and differences between L1 and L2 sounds. The decision on the three pairs of vowels as targets for training was based on the results of previous studies. A brief description of the phonetic comparisons of the relevant vowels of English and Mandarin, Cantonese will be given as background in the following section.

2.3.1. Mandarin Vowels

According to the general descriptions of the Mandarin vowel system in the literature (Cheng, 1966; Dow, 1972; Howie, 1976; Svantesson, 1984), Mandarin has five vowel phonemes /i y u e a/ with various surface forms. Mandarin [u] is said to be phonetically different from its English counterpart in that it is higher and more posterior in the vowel space than the English [u] (Howie, 1976; Norman, 1988). Acoustic studies of Mandarin
vowels have provided empirical data for this claim. (Flege, 1995a; Howie, 1976; Svantesson, 1984; Wang, 1997).

The /ɪ ʊ ɛ æ/ categories that all occur as stressed monophthongs in English do not have counterparts in the Mandarin system. Although the symbols [ɪ] and [ʊ] are sometimes used in diphthongs such as in [æt] or [ɑu], they are actually glides indicating the direction of the movement (Dow, 1972). The [ɪ] and [ʊ] symbols are, therefore, often replaceable by glides [j] and [w] respectively. Although [ɛ] occurs as a surface form of /ɛ/ after the glides [j] and [t] as in [je] “leaf” and [te] “moon” in diphthongs, it is never realized as a stressed monophthong. There is no obvious [æ] category in Mandarin. Compared with the English /ɪ ʊ ɛ æ/ which all occur as stressed monophthongs, Mandarin lacks such counterparts in its vowel system.

How would this mismatch of the two vowel systems affect Mandarin speakers’ learning of English /ɪ ʊ ɛ æ/ and other vowels that are not included in the current study? As discussed earlier, in a series of recent studies on native Mandarin speakers’ production and perception of English vowels, Wang (1997,1998) found that English /ɪ ʊ ɛ æ/ caused serious problems for native speakers of Mandarin in both perception and production. They systematically substituted /i u/ for the /ɪ ʊ/ categories as identified by the native English listeners. Moreover, Mandarin speakers’ /ɛ æ/ productions were often confused with each other by native English listeners. Acoustic analysis showed that there was spectral overlapping between /i/ and /ɪ/, /u/ and /ʊ/, and /ɛ/ and /æ/. It is obvious that Mandarin speakers often have problems distinguishing these vowel pairs in their productions.

It is also reported that Mandarin listeners may use temporal cues to distinguish English /i/-/ɪ/ (Bohn, 1995; Wang & Munro, 1999), but not /u/-/ʊ/ (Wang & Munro, 1999).
The strategies used for perception of these vowel pairs were also found in their production. As observed in a production test, Mandarin speakers’ duration ratio for English /i/-/i/ was found to be significantly greater than the ratios for /u/-/u/ and /e/-/æ/ pairs (Wang, 1997). As length is not used in the Mandarin system to contrast vowels, Mandarin speakers’ reliance on duration cues for distinguishing the English /i/-/i/ contrast cannot be explained as an effect of transfer (Bohn, 1995). The inconsistent use of temporal cues in identifying the structurally similar English tense and lax vowel pairs /i/-/i/ and /u/-/u/ that pose difficulties for Mandarin speakers calls for further investigation.

2.3.2. Cantonese Vowels

According to general phonetic descriptions, Hong Kong Cantonese has 13 vowel surface forms [i: y: u: i ø e o e: ø: a a:] (Lee, 1983). In addition, vowel length plays a distinctive role in Cantonese (Lee, 1983; Ramsey, 1987). Vowel length differences are also characterized by syllable structure differences, as short vowels can occur only in diphthongs (e.g.: [a] in [gaj] “chicken”) or in closed syllables (e.g. [i] in [sik]). The short [i] and [u] are said to be in complementary distribution with the long vowels [i:] and [u:], as they are predictable in terms of syllable ending conditions and tonal differences (Lee, 1983). Both [i] and [u] occur in closed syllables only, as in [sit] “tongue” and [fuk] “luck”. Lee’s acoustic study showed that Cantonese [i] and [u] were lower in quality and, at least for some speakers, there was spectral overlap between the Cantonese [i] and [e:], and [u] and [æ:]. This might also be explained by the fact that Cantonese [e] and [æ] are both raised to [ɛ] and [ɔ] (Zee, 1999). According to Zee, another important fact about Cantonese vowels is that vowel quality changes as a function of syllable structure differences. For example, the tense vowels [i: u:] and some other categories are lowered in closed syllables with plosive and nasal
endings. Furthermore, vowel length differences are also related to vowel quality. For example, the short Cantonese [i] in [sik] “eat” can sometimes be lowered to such an extent that it is not heard as [i] as in American English “sick” but sounds more like French [e] as in [ete] (Ramsey, 1987).

Because of the complications of vowel duration differences and vowel quality change in closed syllables in the Cantonese system, it would be difficult to predict how native Cantonese speakers would assimilate the English vowels to the Cantonese system. Similarly, it is difficult to predict how the Cantonese [i ə u u] would be mapped onto the English system. For example, will the lowered Cantonese [i u] in closed syllables sound like [i u] or [i u] to the native English listeners? In a recent cross-linguistic perception study in which Cantonese [i ə u u] in CVC words were identified by native English listeners using English vowels, Munro et al. (2001) found that these categories are mapped differently onto the English system depending on the syllable codas. For example, while the Cantonese [u] in [ut] endings were identified 96% of the time as English /u/ by native English listeners, [u] in open syllable was heard 75% of the time as its English counterpart. Cantonese [u] in [uk] endings were usually mapped onto English /ow/ (96%). Similarly, while the Cantonese [i] in [it] was heard 96% of the time as English /i/, [i] in [ik] endings was mapped onto the English counterpart only 26% of the time. These findings for [u u] assimilation seem to be in agreement with the description of lowered [u u] in closed syllables, and there is certainly a mismatch of Cantonese [i ə u u] onto the English counterparts.

Based on the above descriptions and findings, it might be predicted that native Cantonese speakers, who are sensitive to the lowering of vowels in closed syllables in their L1, will have problems with English /i u e æ/ categories in the closed syllables. This was
partly confirmed by the results of the same Munro et al. (2001) study in which native Cantonese speakers’ productions of English high vowels /ɪ ʊ ɛ ə/ were also assessed by trained native English listeners. For example, Cantonese speakers’ /u/ in the word “food” and “suitcase” was consistently identified as the English /ʊ/. However, /i/ in the word “read” was almost always heard as /ɪ/, not /u/. It was not clear whether vowel length played a part or not.

However, the substitution of /ʊ/ for /u/ found in the Munro et al. (2001) study may indicate that Cantonese speakers are not aware of the English /i/-/ɪ/ and /u/-/ʊ/ differences in all phonetic environments or even in closed syllables. In a recent study of Hong Kong English phoneme structures, Hung (2000) found a significant degree of overlapping of F1 and F2 values of the /i/-/ɪ/, /u/-/ʊ/ and /ɛ/-/æ/ pairs in the closed syllables with both voiceless and voiced final stops as produced by 15 Cantonese speakers. The data suggest that Hong Kong English speakers do not have a tense-lax distinction in producing these English vowel pairs. A perception test on the same vowel pairs produced by a native English speaker administered to another group of the Cantonese listeners yielded similar results, i.e. that /ɪ/-/ɪ/ and /ɛ/-/æ/ were not clearly distinguished by the Cantonese listeners. Subjects relied on the release of the final stop, rather than the quality of the vowel when identifying the /ʊ/-/ʊ/ vowel contrast. Hung concluded that Hong Kong English does not make a tense-lax vowel distinction and proposed a single vowel /i/-/u/, and /ɛ/ for each of the English /i/-/ɪ/, /u/-/ʊ/ and /ɛ/-/æ/ contrasts. Hung’s findings suggest that native Cantonese speakers will have difficulties learning the three target vowel pairs in the current study (Hung, 2001).

In summary, Mandarin lacks the /ɪ ʊ ɛ ə/ categories in its vowel system as monophthongs and Cantonese [i i u ʊ ɛ] differ in quality and duration systematically as a function of syllable structure and are mapped differently onto the English counterparts. It is
important to point out that while length plays an important role in the Cantonese vowel system, the Mandarin system does not contrast its vowels by duration. Although native Mandarin speakers use duration cues to contrast English tense-lax vowels, such reliance has so far been found to be limited to the /i/-/I/ pair. The complexity of Cantonese vowel quality change in different syllable structures and the concomitant duration differences make it difficult to predict which cues, temporal or spectral, native Cantonese speakers would use to contrast English /i/-/I/, /u/-/u/ and /e/-/æ/ pairs in perception and production.

2.3.3. Research Questions and Design of the Current Study

One general goal of the current study is to examine the effect of perceptual training on the perceptual learning of L2 vowel contrasts. The other goal is to investigate the effect of perceptual training on production. A group of native Mandarin and Cantonese speakers who are highly proficient in English will receive long-term perceptual training on three English vowel contrasts: /i/-/I/, /u/-/u/, and /e/-/æ/. A comparison group with the same L1 background will participate in the tests without the training. Two different training techniques: fading with synthetic stimuli and high variability method with natural tokens produced by multiple speakers will be used for the long-term training. While treatment with synthetic stimuli exploiting the spectral and temporal differences is aimed at shifting listeners’ attention from temporal to spectral cues in identifying English vowels, training with a large number of natural tokens is intended to expose the listeners to highly variable but more real speech from multiple native speakers of English.

Five specific research questions to be addressed are: 1) Will training be effective in modifying native Mandarin and Cantonese speakers’ perceptual patterns by focusing their attention on 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? 5) Will the perceptual learning be transferred to production in terms of increased production accuracy scores and more native-like vowel duration ratios in their productions of the target vowel pairs?

The study has five phases: a pretest phase, a perceptual training phase, a post-test phase, a 3-month retention test phase, and a production assessment phase. During the pretest phase, both perception and production data from the trained and control groups will be collected as pre-treatment baseline information. During the perceptual training phase, subjects from the treatment group only will participate in a 6-8 week perceptual training. In the post-test phase, both the trained and the control groups will repeat the pretest tasks. In addition, a test of generalization (in the perception mode only) will be included to assess whether perceptual learning is transferred to new talkers and new words. During the retention test phase, the trained group only will repeat the post-test tasks three months after training is completed to assess whether the effect of training is retained. Finally, during the production evaluation phase, to assess whether there is any improvement in the trainees’ productions of the target vowel contrasts after training, both the trained and the control subjects’ productions of target vowels in the pretest and post-test will be identified by native English speakers in an intelligibility test. Acoustic measurement of vowel durations will also be carried out to assess whether there is any change in the trainees’ productions of vowel duration differences between the tense and lax vowels.