The Effects of Perceptual and/or Productive Training on the Perception and Production of English Vowels /i/ and /ɪ:/ by Cantonese ESL learners

Janice Wing Sze Wong
The Chinese University of Hong Kong
jwong_aletheia@cuhk.edu.hk

Abstract
This study investigates and compares the effects of different training paradigms in both the perception and production of English /i/ and /ɪ:/ by Cantonese-speaking ESL learners. A total of 39 subjects participated in this study, in which 9 were trained under the High Variability Phonetic Training (HVPT) approach (H group), 11 under explicit articulation training (P group), 9 under a combination of both (HP group) whereas 10 as the control group (C group). Only the H and HP groups, the two groups receiving the HVPT, showed significant perceptual learning and generalization to new words and new speakers. The P group which received productive training only had no significant learning in the perceptual domain. Yet, robust improvement in production was observed in all three treatment groups, with the HP group significantly outperforming the other two training groups. The HP group was also the only one group that transferred the learning at word level to sentence level during a passage reading task. The results suggest that perceptual learning can be transferred to the production domain, but mere productive training barely improves perception. Training effectiveness can also be enhanced when training in both domains are provided.

Index Terms: High Variability Phonetic Training, explicit articulation training, second language acquisition, perception and production of non-native contrast

1. Introduction
The perception and production of non-native contrasts has always been a challenge to second language (L2) learners as they are becoming a more language-specific perceiver [1][2][3]. Different phonetic training paradigms have been devised by researchers to examine the effects of training on the modification of perception and/or production of non-native contrasts, which are usually perceived and produced by L2 learners according to their L1 categories [4][5]. Among all, the High Variability Phonetic Training (HVPT), which involves the use of natural training stimuli with various phonetic contexts produced by multiple speakers in an identification task with immediate feedback, has received particular attention due to its effectiveness in perceptual learning of different consonantal and vowel contrasts [6][7]. More research interest has also been drawn as perceptual gain in the HVPT was found to promote production improvement [8].

However, previous studies such as [8]-[14] were limited to investigating the effects of either perceptual or productive training alone on L2 speakers’ phonology, hence provided limited evidence on how perceptual learning transfers to production, or productive learning to perception modification. Particularly few studies (cf. [12]) assessing the effects of production training on the two domains have been found. The present research thus compared the training effects of the HVPT, an explicit articulatory training, and a combination of the two, in order to preliminarily garner some insights in how to optimize the training effects with different training paradigms. L2 teachers and learners can also benefit from the present research if an effective and efficient combination of training approach is found. The present study has chosen to investigate a commonly-confused vowel pair /i/-/ɪ:/ in terms of both the perception and production, among Hong Kong Cantonese learners of English [15][16][17].

2. Methodology

2.1. Participants
Thirty-nine secondary school students aged around 16 to 17 who had Cantonese as their L1 and English as the L2 were recruited to participate in the present experiment. They were divided into different groups: 9 (6 females and 3 males) were trained under the HVPT approach (H); 9 (7 females and 2 males) were trained under both the HVPT and an explicit production training (HP); 11 (6 females and 5 males) were given only the explicit production training (P); and 10 (6 females and 4 males) were the control group receiving no training (C). All the subjects started learning English as an L2 at the age of 3.21 (SD = .56) for an average of 13.1 years (SD = .98). They had not resided in any English-speaking countries. None reported hearing or speaking impairment.

A total of 8 native English speakers (6 females and 3 males) were also invited to produce stimuli for the tests and training. Their ages ranged from 20 to 40, with four speaking General American accent, three Southern English and one Canadian English. The use of a variety of accents was to increase the stimulus variability.

2.2. Design
The present study adopted a pretest-treatment-posttest design, in which three groups of subjects were trained under different paradigms whereas the control group received no training. All subjects participated in these three phases:
PHASE 1. Pretest Phase, which included one production pretest and one perception pretest;
PHASE 2. Treatment Phase:
a. High Variability Phonetic Training, with a total of 10 training sessions (only the H and HP groups participated) in which they attended two sessions a day, each lasted for 10 minutes;
b. Explicit Articulation Training, with a total of 4 sessions, each lasted for 30 minutes (only HP and P groups participated) in which they attended one session a day;
PHASE 3. Posttest Phase, with one production posttest, one perception post-test, one production Test of Contextualization (TC) and three perception Tests of Generalization (TG1, TG2 and TG3).

2.3. Setting and Apparatus
All the subjects completed the tests and training sessions in a language laboratory. They used a computer program designed...
by the researcher to complete all perceptual training and test sessions. The subjects listened to audio speech tokens presented by the program and data were saved into a Microsoft Access database. In the explicit production training sessions, videos in which a native speaker demonstrated the articulation of vowels were given to the subjects to watch before practicing the target vowels with the researcher. They recorded the production test tokens using Adobe Audition 1.5.

2.4. Stimulus Materials

Six of the eight native English speakers produced stimuli used in both the perceptual pre/posttest and the HVPT training. All of them produced 20 minimal word pairs, i.e. a total of 40 tokens. All words were CVC monosyllabic words with different onsets and codas. One of the speakers, i.e. a familiar subject to the speakers, recorded another new word list of 20 minimal word pairs with /ɪ/ and /iː/ for the use in TG2 (new words by a familiar speaker). Another speaker whose voice had not appeared in the training or tests recorded the same set of tokens for the use in TG3 (familiar words by a new speaker). The last speaker who had not recorded any tokens for the training stimuli or the tests, i.e. a new speaker, recorded another new list with 20 /iː/ minimal word pairs for the use in TG1 (new words by a new speaker). All the minimal pairs in TGS were with various consonant-vowel-consonant contexts and syllable structures (mono-, di- and poly-syllabic) with a view to testing the transfer of learning under various conditions.

Each speaker read the tokens at least three times to avoid the use of a single token per speaker for the stimuli. All the stimuli were made by the native speakers reading into a headset-mounted microphone with Adobe Audition 1.5 software for digitization.

2.5. Procedure

2.5.1. Pretest Phase

All groups had to participate in both the production and perception tests in the first phase. The production pretest was administered first to avoid subjects’ cueing or being exposed to the items which would appear later in the perception pretest.

- **Production Pretest**: The subjects were given a word list of 20 words (with 10 /ɪ/ and 10 /iː/) and had to record all the words which would appear either in the perception pretest or the training. To ensure authentic performance and that the subjects could produce also other segments apart from the vowel, before the pretest, the subjects could hear the pronunciations of the words produced by a native speaker who had not been involved in the study. The instructions for this production pretest were offered to the subjects in the form of five practice trials and they had to produce them with natural loudness and speaking rate. They were not provided with any audio prompts or instructions during the recording. They could also pause and resume during the recording based on their own pace. The test took less than five minutes to complete.

- **Perception Pretest**: The subjects could get access to the computer program to complete the identification test which included 50 questions (40 words with either /ɪ/ or /iː/ with 10 distractors). They did 10 practice trials before the test, which were not analyzed. Each stimulus could be played by the subjects as many times as they needed before they choose the answer from three choices with conventional English orthography, or a blank for a free answer, in which they could type their own word. The frequency of occurrence of the correct answer that appeared in the four serial positions, i.e., word 1, word 2, word 3, free answer, were equal; thus the chance level set was 25%. This design was an attempt to avoid using simply two choices with 50% of chance level. The program was designed to limit the subjects from not answering one question before moving onto another. The whole test could be completed within 10 minutes.

2.5.2. Treatment Phase

There were two types of training in this phase:

- **High Variability Phonetic Training**: A total of 40 stimuli (20 /iː/ and 20 /ɪ/) produced by six different native English speakers, all randomized in terms of speakers and word order in each session, were presented to the subjects. The subjects were trained on a two-alternative forced choice paradigm which directed their attention to identifying the target word and raised the training effect, unlike the 4 choices used in the pretest. The stimuli were one of the counterparts in a minimal word pair contrasting the two vowels, /iː/ and /ɪ/ (e.g., among “bead” and “bid,” only one of them was chosen for one test item). During training, immediate feedback was given; at the end of each session, their total scores were also shown.

- **Explicit Articulation Training**: The subjects were provided with videos in which a native speaker demonstrated the articulation of the vowels. The subjects had to watch the video first and read after the video host. A word list with 20 different words displaying the vowel contrast was given to the subjects and they pronounced after the researcher at least three times and corrective feedback was given. Articulation information of the vowel pairs, i.e. the tongue height/vowel openness, tongue frontness/backness as well as the length of the vowels was also emphasized in the each session explicitly with pictures as illustrations and mirrors to help them better articulate.

2.5.3. Posttest Phase

The Posttest Phase involved two production posttests (posttest and TC) that were completed before the four perception posttests (posttest, TG1, TG2 and TG3).

- **Production Posttest**: the same as the Production Pretest

- **Test of Contextualization**: All the subjects were given a 200-word passage which included 30 content words with /iː/ and /ɪ/. Subjects were asked to record the whole passage naturally, at their own pace and loudness.

- **Perception Posttest**: the same as the Perception Pretest

- **Test of Generalization 1**: The subjects heard 40 tokens (with 20 /iː/ and 20 /ɪ/) spoken by a new speaker whose voice was not heard in any of the training stimuli or the tests. The procedures were similar to those administered in the Perception Pretest, and subjects were also given four choices to choose from.

- **Test of Generalization 2**: The subjects had to listen to 40 new words (with 20 /iː/ and 20 /ɪ/) spoken by a familiar speaker, who had been one of the speakers in the training stimuli. Procedures were the same as those in TG1.
Test of Generalization 3: The subjects revisited 40 familiar words which they had come across in the perception training sessions, but the words were produced by a new speaker. Again, the procedures were the same as those in TG1 and TG2.

2.6. Evaluation of Production Data

The production scores were evaluated by directly counting the number of accurate productions of the target vowels. The productions of the subjects were transcribed twice by a phonetically-trained researcher for whom Cantonese was the L1 and English the L2. The intra-rater reliability obtained was 93.27% (α = .898). Another researcher who had English as L1 also transcribed the data phonetically. During transcription, they transcribed phonetically the word they heard, which was not limited to only the target vowels. The reliability check was done without referring to any completed transcriptions and it was 92.62%. The inter-rater reliability was α = .821. A follow-up acoustic analysis on half of the transcriptions, by checking the F1, F2, F3 values and the vowel durations, was conducted by a third phonetically-trained researcher to confirm that the transcriptions aligned with the acoustic measures and were reliable. The acoustic analysis results were consistent with the transcription.

3. Results

3.1. Perceptual Performance

The following figure displays the overall results of the perceptual identification performance of the H, HP, P and control groups.

![Figure 1: Mean percentages of correct identification for three trained groups and control subjects in the pretest, the posttest, and three Tests of Generalization (TG1, TG2 & TG3).](image)

Their identification scores were submitted to a three-factor repeated measures analysis of variance (ANOVA) using Group (H, HP, P, C), Test (pretest, posttest), and Vowel (/iː/, /ɪ/) as factors. Significant main effects of Group \(F(3, 35) = 11.66, p < .001\), and Test \(F(1, 35) = 40.87, p < .001\) were observed. Group × Test interaction was also significant \(F(3, 35) = 8.49, p < .001\). Post-hoc pairwise comparisons (Bonferroni) on the Group × Test interaction showed a significant difference between groups in the posttest \(F(3, 35) = 15.26, p < .001\), but not in the pretest \(p = .53\). In the posttest, only the H \(F(3, 35) = 43.72, p < .001\) and HP \(F(3, 35) = 16.68, p < .001\) groups demonstrated significant improvement from the pretest to posttest, but the P \(p = .20\) and C \(p = .65\) groups did not. Moreover, the H group outperformed both the P \(p = .003\) and C \(p < .001\) groups by 16.72% and 25.58% respectively with significance while the HP group also demonstrated significantly better learning than both the P \(p = .009\) and C \(p < .001\) groups by 15.08% and 23.92% respectively. It reveals that after the perceptual training, the two perceptually trained groups performed significantly better than the other two groups, with no significant differences between the performance of the H and HP group. Yet, no significant main effect of Vowel was observed, indicating that the learning progress of the two vowels were similar. The above results suggested that only the two groups receiving perceptual training could gain in the perceptual identification of the target vowels, whereas explicit production training alone did not benefit the perception.

Concerning the results in the generalization tests, for TG1 (new words produced by a new speaker), a two-way ANOVA revealed the main effect of Group \(F(3, 35) = 21.19, p < .001\) was robust. The H group was performing significantly better than the P \(p < .001\) and C \(p < .001\) groups by 23.59% and 28.72% respectively. However, the performance of the H and HP groups were not significantly different. In TG2 (new words produced by a familiar speaker), another two-way ANOVA revealed a robust main effect of group \(F(3, 35) = 5.81, p = .002\) whereas vowel was not significant \(p = .82\). Under this condition, only the H group performed significantly better than the P \(p = .018\) and C groups \(p = .018\) whereas other groups showed no significant differences with each other. With regard to TG3 (familiar words produced by a new speaker), only the main effect of group \(F(3, 35) = 5.77, p = .003\) was significant. This time, the H and HP groups performed significantly better than the C group only (H vs. C: \(p = .011\); HP vs. C: \(p = .013\)), but not the P group (H vs. P: \(p = .128\); HP vs. P: \(p = .146\)). Perceptual learning was shown to be able to generalize to new words and new tokens, and the H group was consistently performing the best among all groups in the generalization tests.

3.2. Transfer of perceptual learning to production

One of the aims of the present study was to investigate how the different training paradigms affected learning in both perception and production domains. This is particularly interesting to also look at how perceptual learning can be transferred to the production domain, or to what extent explicit articulation training can help improve the production accuracy of the target vowel pair. Figure 2 below shows the production performance of all four groups.

![Figure 2: Percentage of target production performance for three trained groups and control subjects in the pretest, the posttest and Test of Contextualization (TC).](image)

A three-factor repeated measures ANOVA with Group (H, HP, P, C), Test (pretest, posttest) and Vowel (/iː/, /ɪ/) as factors was conducted. Significant main effects of Group \(F(3, 35) = 5.51, p = .003\), Test \(F(3, 35) = 141.0, p < .001\) and Vowel \(F(3, 35) = 23.14, p < .001\) were found. Group × Test interaction was also robust \(F(3, 35) = 21.54, p < .001\).
Simple effects analyses of the Group × Test interaction revealed that the difference between groups was significant in the posttest \(F(3,35) = 12.9, p < .001\), but not in the pretest \(p = .48\). The effect of Test was also significant among the three trained groups \(p < .001\) but not in the control group \(p < .72\) indicating that all three trained groups improved their production accuracy of the vowel pair after treatment \(H: 36.11\%; HP: 66.11\%; P: 29.09\%\). Furthermore, in the posttest, the H group performed significantly better than the control group for 43.61% \(p < .001\); the HP group had more robust results than both the P \(p = .018\) for 32.12% and C \(p < .001\) for 61.94%; the P group also did significantly better than the C group for 29.77% \(p = .027\). These results all indicated that both the perceptual and productive training benefited the production learning of the two vowels.

Test of Contextualization aimed to further study the transfer of production learning in the word level to the sentence level. A two-way repeated measures of ANOVA showed that the main effect of Group \(F(3,35) = 3.90, p = .017\) was significant, indicating that there were differences in the performance among groups; yet, the main effect of Vowel and other interactions were not robust. Even though the H and P groups outperformed the control group by 20.69% and 22.61% respectively, their differences were not significant. Only the HP group demonstrated significantly better results than the control group by 34.03% \(p = .013\). These figures indicated that the learning effect on production generalized to contextualized speech only among the group receiving both perceptual and productive training.

4. Discussion

The main purpose of the present study was to compare the effectiveness of the HVPT, the explicit production training and a combination of the two training paradigms in modifying both the perception and production of a non-native vowel contrast /i/-/ɪ/ among Cantonese ESL learners.

In general, the above results confirm that the HVPT was a useful paradigm in training the perception of /i/-/ɪ/ contrast in English, as shown by the performance of the H group. Perceptual learning can also be transferred to the production domain. Generalization to the perception of new words produced by new or familiar speakers was also observed. These results support previous studies’ findings, e.g. [7], [8] as mentioned earlier, that the HVPT, being a solely perceptual training program, is effective in promoting the re-categorization of more near native-like phonological patterns due to the use of high stimulus variability in terms of speaker variation and phonetic contexts. Exposure to highly variable stimuli (cf. training with tokens with low variability, see [18]-[20] for reference) is necessary for the subjects to form robust phonetic representations by learning which acoustic cues are relevant to a specific sound; that is, to let the subjects selectively attend to a wider range of acoustic dimensions and weightings so as to develop more language-specific phonetic categories. The fact that learners can generalize the learning may also be attributed to the use of high stimulus variability as they had to focus on the criterial properties and acoustic cues in common of the vowels produced by the different speakers, in which they also had to strive to ignore the between-speaker variability that might exhibit obstacles to the perceptual learning. The above findings are in line with earlier studies showing the effective and successful adoption of the HVPT in the modification of different sound segments, although how much stimulus variability is playing a role in learning needs further comparison studies with a training paradigm with lower stimulus variability. What cues the learners actually rely on before and after training should also be one focus of follow-up studies.

One interesting finding in the present study goes to the failure in perceptual learning of the P group. Meanwhile, the HP group being not the best among all groups in perceptual tests (the H group performed consistently better than the HP group in the perceptual posttest and all three Tests of Generalizations; the HP group performed the best only in the production tasks), is also not in line with a common logic that receiving more training leads to better learning. The failure in perceptual learning of the P group leads to the speculation that productive training alone may not facilitate or even hinder category reformation; instead, directly training in the perceptual domain benefits the formation. Even though during the productive training, a certain amount of perceptual training was given (e.g. the researcher cannot avoid pronouncing the vowels to the subjects during training; or when learners were watching the videos clips, they were still exposed perceptually to the target vowel pair), with only very limited perceptual input and training solely in how to articulate, subjects could not exercise their abilities to map the acoustic output into their own phonetic space, leading to the insignificant increase in their perceptual performance. It means that without direct perception training, productive training alone cannot modify the perception of the target vowels. The results of the HP group further suggest that when learners were trained in both domains in the same time frame (productive training was provided right after 2 sessions of perceptual training, but not after all perceptual training sessions were finished), if they are yet to successfully perceive the vowel contrast, i.e. in the process of category formation, additional explicit instruction on production will not give extra benefits to perceptual learning; hence learners receiving both perceptual and productive training may not be learning so attentively or focused as the H group did in terms of perceptual category formation. The above findings suggest that given sufficient perceptual input, learners were able to improve both the identification and the output of the sound segments, i.e. learning in the perceptual modality can benefit production, but not vice versa. However, L2 learners may still benefit from the above training paradigms as they successful help them improve the perception and/or the production of a certain non-native contrast to a certain extent.

5. Conclusion

The High Variability Phonetic Training, a perception training paradigm, has proven successful for improving both the perception and production of a non-native contrast among Cantonese speakers of English in the present study. The results are in line with some previous studies that stimulus variability helps the learning of a non-native contrast. However, this study also indicates that productive training could only help improve the production of the target sounds but not the perception, i.e. learning merely in the perceptual domain can be transferred to the production domain, but not vice versa. Yet, a more thorough understanding of the intricate link between speech perception and production certainly requires more in-depth investigations. Other variables and factors such as training intensity, sentence-level training or training in the suprasegmental aspect are all relevant in the research agenda of assessing phonetic training and raising specific pedagogical implications.
6. References


