Speech perception training as a serious game in the EFL classroom

M.P. Hommel Utrecht Institute of Linguistics OTS, Utrecht
m.p.hommel@uu.nl

Abstract

Speech perception training improves listeners' perception, but it could also significantly improve production. The benefits of perception training are interesting for the English as a foreign language (EFL) classroom, where this tool is currently not being used. This study looks at students' perception and production performance on a pretest and, following perception training sessions, on a posttest and retention test. Results show that perception training improved students' perception and to a lesser extent students' production.

1. Introduction

1.1. Perception training and pronunciation

In the EFL classroom in the Netherlands, little attention is paid to production training and many Dutch students still struggle with the pronunciation of certain English speech sounds [1]. A tool that potentially helps improve students' pronunciation is speech perception training. The intention of this paper is to find out whether this type of training significantly improves students' perception as well as students' pronunciation. Contrary to previous studies, this study will be done in a classroom setting and with many phonemes. Numerous perception training studies have shown that perception training significantly improves perception of the trained stimuli (e.g. [2], [3], [4] [5], [6]). Moreover, many studies have shown that perception training automatically improves speech production (e.g. [2], [3], [5], [6], [7]). Akahane-Yamada et al. [2] claim that "training in speech perception transfered to improvements in the production domain, suggesting a close link between speech perception and production". Neurological studies involving fMRI have shown an "overlap between the cortical areas active during speech production and those active during passive listening to speech" [8, pp. 371]. These scans demonstrate a relationship between perception and production. This evidence suggests that perception training could be a very beneficial tool for language learning in the classroom.

1.2. Gamification in the classroom

The perception training task in this paper will be a type of serious game. Adding game elements to classroom tasks, or gamification, suits the current generation of students. Game-based learning can turn "disconnected, bored learners into engaged participants" [9]. Games involve for example providing feedback, a high level of engagement and competition. Students get constant feedback, and "the more frequent and targeted the feedback, the more effective the learning" [9]. Games can also create a realistic context, rather than abstract information. Understanding the use and purpose of a task makes students more involved. When students are actively involved or engaged, they learn more than simply absorbing passive information. Yet another advantage of a game is that it provides leaners with the freedom to fail. Students can learn a lot from their own mistakes; it promotes trial-and-error learning [9]. Finally, games have a competitive element, either between players or within the player himself.

1.3. A large phoneme inventory for perception training

Nishi and Kewley-Port [4] reviewed earlier training studies and concluded that most of them focused on difficult L2 contrasts (e.g. English /æ-e/ for Dutch L1 listeners) and that none of these studies included more than five vowel pairs. Nishi and Kewley-Port trained one group of listeners with nine vowels (fullset) and one group with only three vowels (subset). The results showed that the fullset vowel training was more effective than the subset vowel training, which suggests that "efficient learning of nonnative vowels requires exposure to a full set of vowel categories" [4, pp. 1506]. Positive results in studies that used a small set of vowel pairs may be skewed as listeners could improve their performance by using vowel-specific cues or strategies "that may fail to generalize to the genuine perceptual learning that is needed to classify the complete set of vowels in the target language" [4, pp. 1507]. Kewley-Port et al. [10] looked at the effects of large training sets using vowels as well as consonants and found that speech perception training significantly improved listeners' vowel and consonant perception. These findings suggest that a large phoneme inventory is most beneficial for language learners, which is why this study uses many phonemes for perception training. The main research question is: does perception training as a serious game improve students' perception and pronunciation?

2. The experiment

2.1. Outline of the experiment

In order to find out whether perception training improves students' perception and production in the classroom, an experiment was conducted. Dutch students who received perception training sessions were compared with a control group regarding their perception and production skills. First, a pretest was conducted testing students' perception. In addition, students' pronunciation was recorded. Next, the test group received perception training sessions. After the training sessions were finished, a posttest was conducted and students' pronunciation was again recorded. Two months after the posttest, a retention test was conducted. The expectation was that the test group would outperform the control group on both perception and production.

2.2. Method of the pretest, training session, posttest and retention test

2.2.1. Design of the tests

SPATS-ESL (Speech Perception Assessment and Training System - English as a Second Language) contains a large inventory of 109 position-dependent phonemes (i.e., a phoneme could be in onset, nucleus or coda position) [11]. This large amount of English phonemes makes it unpractical to use all of them, as it would simply take up too much time in the English classroom; it is recommended to practice with SPATS-ESL for around 25 hours [11]. Another practical issue is that Dutch listeners might already achieve ceiling

10.21437/SLaTE.2017-31

performance in certain L2 vowels or consonants and therefore have little room for improvement. A pilot, the SPATS-ESL quick test, was conducted to solve the issue of manageability by removing phonemes with ceiling performance (i.e. >90% correct) and selecting the more difficult phonemes (≤90% correct). Fifty Dutch students (27 male, 23 female and M = 19 years, SD = 2.1) took the SPATS-ESL quick test. The average score was 83% correctly perceived phonemes (SD = 5.9%). Overall, 38 difficult phonemes (out of 109) remained as items for the pretest and training sessions. Both tasks consist of real words as opposed to non-words to create a realistic context, promoting student engagement.

2.2.2. Procedure of the tests

Eight native speakers with a Southern British English (BrE) accent were recorded in sound proof cabins in the phonetics lab at Utrecht University and the Max Planck Institute for Psycholinguistics in Nijmegen (both in the Netherlands). Audacity [12] software was used for the recordings (sampling frequency of 48 kHz). Because certain speakers had a louder voice and others a softer voice, the recorded data were set to an average intensity of 60 dB. Zep [13] was used to implement the four-interval forced-identification task where students would hear a stimulus and had to choose one of four answers (1 target, 3 foils). The listener's task was to select the answer containing the speech sound that matched the one heard in the stimulus at the indicated position. For example, students would first hear 'push' and could then see four answers: cook, cope, cow and coup. The right answer is 'cook' because it has the /ʊ/-sound in the underlined position as 'push'.

It took a student around 10 minutes to complete the pretest and around 20 minutes to complete one training session. The posttest and retention test were the same as the pretest. The pretest and a training session contained the same target sounds, but the sounds were put in different words and there were different foils. The pretest and a training session also consisted of different speakers to see if listeners would generalise what they learned during the training sessions to novel items and voices. During the pretest, listeners received no feedback. During the training sessions, listeners received feedback on the correctness of each answer. This feedback gave students insight in their mistakes. At the end of either the pretest or a training session, students got to see their average score (in percentage correct). The competitive element was that students could try and improve their own test score in each training session, and compare their score to classmates' scores. Students had five similar training sessions which took place once a week. The pretest took place in September 2016 and the posttest took place two months later. Two months after the posttest, the retention test was conducted to see if any positive results remained.

2.2.3. Stimuli of the pretest and training sessions

The pretest consisted of 38 items (19 fixed items spoken by a male and the other 19 fixed items spoken by a female). Each training session was the same and contained 2 × 38 items (38 items spoken by a male and the same 38 items spoken by a female). The items contained the phoneme (clusters) shown in Table 1.

Table 1: *The targets sounds by phonological position*

Onset	Nucleus	Coda
/j/	/ɒ/	/θ/
/v/	/ʊ/	/g/
/z/	/əʊ/	/t/
/d/	/ʌ/	/dʒ/
/l/	/ɪ/	/z/
/g/	/e/	/s/
/f/	/ɔ/	/f/
/dʒ/	/i/	/tʃ/
/s/	/eɪ/	/k/
/ð/	/u/	/v/
/r/	/æ/	/d/
/θ/		/ʃ/
/b/		/ʒ/
		/ð/

2.2.4. Participants

The participants (initially test group: N = 92, of which 36 male and 56 female; control group: N = 65, of which 28 male and 37 female) were Dutch students attending secondary education (M = 19 years, SD = 1.3). Their average level of English was between upper-intermediate and advanced (CEFR level B2 - C1). Students in the test group received training sessions and students in the control group received regular English classes without training sessions.

Of the initial 92 participants in the test group, 16 were dyslexic, 5 did not have Dutch as their native language and 23 students did not attend all 5 training sessions, leaving 49 students (17 male, 32 female) for data analysis of the training sessions.

Of the 49 students who attended all training sessions, 2 missed the posttest, resulting in N = 47 (17 male, 30 female) for data analysis of the posttest. Of the 65 students in the control group, 5 were dyslexic, 5 did not have Dutch as their native language, 1 missed the pretest, and 6 missed the posttest, resulting in N = 48 (21 male, 27 female) for data analysis of the posttest.

Quite some students missed the retention test, resulting in N = 29 (6 male, 23 female) for the training group and N = 38 (15 male, 23 female) for the control group.

2.3. Method of rating students' production

2.3.1. Design of the production rating

Three native English speakers (BrE) rated the subjects' production. As it was impractical for the native speakers to rate the pre-, post- and retention test production of all 38 items of at least 67 students, only 6 phonemes were chosen for native speakers to listen to, two in each phonological position (see Table 2). These 6 phonemes did not have an equivalent in Dutch, but three (one in each position) showed the most progress in perception for students who received training sessions, and the other three showed no improvement.

Table 2: *The targets sounds for the production rating*

Onset	Nucleus	Coda
/θ/ in 'thanks'	/ɒ/ in 'lot'	/θ/ in 'death'
/ð/ in 'those'	/ʊ/ in 'look'	/ð/ in 'bathe'

2.3.2. Stimuli of the production rating

The pronunciation of twenty speakers in the test group and five speakers in the control group were rated. The retention test was included in the rating of the test group but excluded in the control group [6 stimuli × 3 tests × 20 speakers = 360 and 6 stimuli × 2 tests × 5 speakers = 60 items, which is 420 items in total. Then, recordings with poor quality such as background noise were excluded]. There was a total amount of 372 items for native listeners to rate.

2.3.3. Procedure of the production rating

The method used was a two-interval forced choice task. The spoken target sounds in the pretest were presented in pairs with those from the posttest, pretest with retention test, and posttest with retention test. The forced choice task consisted of two parts, each consisting of 186 items, and listeners could have a break after the first part. Excluding the break, it took them around 60-70 minutes to complete the task.

2.4. Results

2.4.1. Results of the pretest

Nine phonemes already showed ceiling effects (more than 90% correct) on the pretest: onsets /j/, /z/, /g/, /s/, /r/, /b/, nucleus /ɪ/, and codas /ʃ/ and /k/. The two phonemes with the lowest scores were nucleus /ɔ/ (13% correct) and coda /ð/ (6% correct). These two phonemes also scored the worst of the nuclei and codas in a study by Cutler et al. [14]: 22% and 8% correct respectively.

There is a low to medium correlation between pretest score and a student's performance on the Dutch national listening test (*r* = .38, *p* = 0.01). Also, there is a medium correlation between pretest score and a student's performance on the Cambridge English test for schools (*r* = .55, *p* = 0.01).

2.4.2. Results of the training sessions

There was a significant improvement in test score due to test time *F*(4,180) = 36.22, *p* < .01. The Bonferroni post hoc test indicated that the test group significantly improved after the first and second training sessions (*p* < .01), but that the third training session did not differ from the fourth and that the fourth did not differ from the fifth training session. These results suggest that students' perception did not improve significantly after the third training session. The average score on perceiving the 38 target sounds after the first training session was already high (M = 77%, SD = 9.5%), and the score after the fifth training session was significantly higher (M = 86%, SD = 9.4%).

2.4.3. Results of the retention test

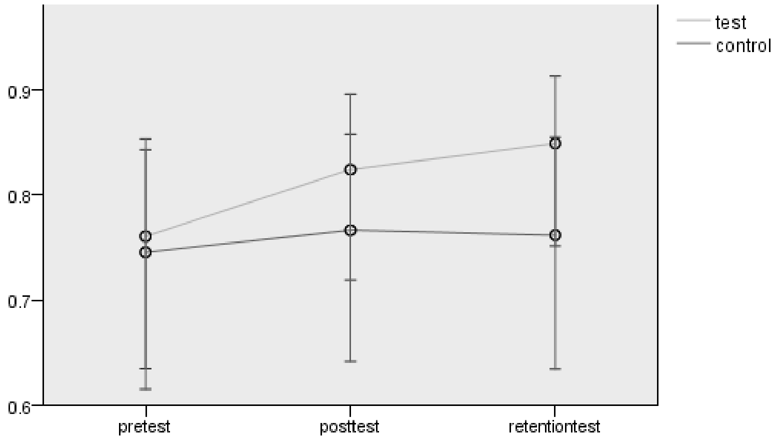
A multinomial logistic regression was carried out to determine the effect of perception training on test score. Time, group, and interaction of time and group as well as level were fixed effects in the model. These effects were all significant: time (*F*(2) = 13.14, *p* < .001), group (*F*(1) = 5.41, *p* = .02), time × group (*F*(2) = 5.93, *p* < .001) and level (*F*(1) = 4.68, *p* = .03). Sex was excluded as fixed effect because it was not significant. Intercepts for subjects and items were random effects in the model. The regression model explained 80.6% of the variation in students' response (*F*(7) = 8.96, *p* < .001). The model showed that the test group outperformed the control group on the posttest and retention test. The odds of the test group having a higher test score than the control group was 1.8 times, so the effect size was relatively small. Moreover, the interaction between test group and pretest was significant, indicating that the test group performed worse on the pretest than on the posttest and retention test. The odds of the test group scoring better on the posttest and retention test than on the pretest was 1.6 times, also indicating a small effect size. In addition, vwo outperformed havo. The odds of vwo having a higher test score than havo was 1.4; this was again a small effect size. The effect sizes were small, but these results suggest that the training sessions improved students' speech perception and that this improvement was retained (see Figure 1).

2.4.4. Results of the production rating

Reliability analysis showed a moderate reliability (α = .46) between the native raters, but reliability would be lower if any native speaker was excluded. The inter-item correlation was low (between .20 and .24) for all listeners, which could mean that listeners were judging different things, but it is more likely that listeners had a hard time hearing a difference between many productions. As the task forced listeners to choose one production over the other, they sometimes had to 'guess' which production was better, hence rendering a relatively low Cronbach's Alpha.

For each production, the raters' response was categorised (production before training sessions was better or production after training sessions was better) and counted. The scores for the control group on the pre- and posttest did not differ, $\chi^2(5) = 10.53$, *p* = n.s, but the scores for the trained group on the pre- and posttest did differ, $\chi^2(5) = 27.25$, *p* < .001. Post-hoc analyses showed that students' production in the test group significantly improved for three out of six phonemes (onsets /θ/, *p* < 0.01 and /ð/, *p* < 0.001 and coda /θ/,

Figure 1: *Test score over time. The test group outperforms the control group.*



p = 0.04).

The scores for the trained group also differed between pretest and retention test, $\chi^2(5) = 15.45$, $p < .001$. Students' production in the test group significantly improved for two out of six phonemes (onset /θ/, $p = .01$ and nucleus /ʊ/, $p = .04$). As expected, the scores for the trained group on the posttest did not differ from those on the retention test, $\chi^2(5) = 8.62$, *p* = n.s, indicating that the improvement in production remained over time. Raters considered some phonemes in students' pronunciation after the training sessions as better than before.

3. Discussion

As mentioned in the introduction, fMRI scans only show a partial overlap between perception and pronunciation. Results here paint a similar picture. Three phonemes that showed progress in perception were the onset /θ/, the nucleus /ɒ/ and the coda /ð/. Three phonemes that showed progress in production were the onsets /θ/ and /ð/, nucleus /ʊ/ and coda /θ/. Only onset /θ/ improved in both perception and production as a result of perception training sessions. Nucleus /ʊ/, onset /ð/ and coda /θ/ only improved in pronunciation. Nucleus /ɒ/ seemed to deteriorate in production in the posttest, but this difference could not be found when production of the pretest and the retention rest were compared. One possible explanation for this finding is hypercorrection of the LOT vowel by Dutch students; the production of this vowel by Dutch speakers is already confused with the STRUT vowel by British English listeners [15], and perhaps hypercorrection made this confusion worse for these listeners.

The experiment conducted here shows that perception training with a large phoneme inventory and in a classroom setting improves perception and to some extent production. Data from this experiment will be scrutinised in order to see how training sessions affected each student individually. The next experiment that will be conducted looks at whether perception training also improves students' word recognition and general listening skills. For future research it would be interesting to examine what the effectiveness is of perception training combined with explicit phonetic instruction on various language skills.

4. Conclusion

Similar to the findings of previous studies, perception training improves perception and to a certain extent production. Students improved their speech perception and generalised what they learned during the training sessions to novel items and voices, and retained their improved perception and production over time. This study shows that perception training also works in a classroom setting. It is plausible that perception training as a serious game improves students' perception and production, which is why it is recommended to start using this tool in the foreign language classroom.

5. Acknowledgements

First and foremost, I would like to thank my supervisors René Kager and Willemijn Heeren. This work is part of the Doctoral Grant for Teachers with project number 023.007.008, which is financed by the Netherlands Organisation for Scientific Research (NWO).

6. References

[1] R. Van den Doel, *How Friendly are the Natives? An Evaluation of Native-speaker Judgements of Foreign-accented British and American English.* Dissertation Netherlands Graduate School of Linguistics. Retrieved 01 June 2015 from http://dspace.library.uu.nl/bitstream/handle/1874/13381/Doel-13-completetext.pdf, 2006.

[2] R. Akahane-Yamada, Y. Tohkura, A.R. Bradlow, and D.B. Pisoni, "Does Training in Speech Perception Modify Speech Production?," Spoken Language *Proc. ICSLP*, vol. 2, pp. 606–609, 1996.

[3] A.R. Bradlow, R. Akahane-Yamada, D.B. Pisoni, and Y.I. Tohkura, "Training Japanese listeners to identify English /r/ and /l/: Long-term Retention of Learning in Perception and Production," *Perception & Psychophysics*, vol. 61, no. 5, pp. 977-985, 1999.

[4] K. Nishi and D. Kewley-Port, "Training Japanese Listeners to Perceive American English Vowels: Influence of Training Sets," *Journal of Speech, Language, and Hearing Research*, vol. 50, pp. 1496-1509, 2007.

[5] B.L. Rochet, "Perception and Production of Second-Language Speech Sounds by Adults," *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, pp. 379-410, 1995.

[6] X. Wang and M.J. Munro, "Computer-Based Training for Learning English Vowel Contrasts," *System*, vol. 32, pp. 539-552, 2004.

[7] T. Lopez-Soto and D. Kewley-Port, "Relation of Perception Training to Production of Codas in English as a Second Language," *Proceedings of Meetings on Acoustics*, vol. 6, pp. 1-15, 2009.

[8] B. Galantucci, C.A. Fowler, and M.T. Turvey, "The Motor Theory of Speech Perception Reviewed," *Psychonomic Bulletin & Review*, vol. 13, no. 3, pp. 361-377, 2006.

[9] K. Kapp, *Games, Gamification and the Quest for Learner Engagement*. Retrieved 25 June 2017 from https://www.td.org/Publications/Magazines/TD/TD-Archive/2012/06/Games-Gamification-and-the-Quest-for-Learner-Engagement, 2012.

[10] D. Kewley-Port, K. Nishi, H. Park, J.D. Miller, and C.S. Watson, "Learn to Listen (L2L): Perception Training System for Learners of English as a Second Language," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2773, 2009.

[11] J.D. Miller, C.S. Watson, D. Kewley-Port, R. Sillings, W.B. Mills, and D.F. Burleson, "SPATS: Speech Perception Assessment and Training System," *J. Acoust. Soc. Am.*, vol. 122, no. 5, 2007.

[12] A. Team, "Audacity(R): Free Audio Editor and Recorder" (Computer program), version 2.0.0 retrieved 22 June 2016 from http://audacity.sourceforge.net/, 2016.

[13] T. Veenker, "The Zep Experiment Control Application" (Computer software), *Utrecht Institute of Linguistics OTS, Utrecht University*. Version 1.12.5 retrieved 13 July 2016 from http://www.beexy.org/zep/, 2016.

[14] A. Cutler, A. Weber, R. Smits, and N. Cooper, "Patterns of English Phoneme Confusions by Native and Non-native Listeners," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3668-3678, 2004.

[15] J. Hillenbrand, L.A. Getty, M.J. Clark, and K. Wheeler, "Acoustic Characteristics of American English Vowels," *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3099-3111, 1995.