NON-NATIVE PRODUCTION TRAINING WITH AN ACOUSTIC MODEL AND ORTHOGRAPHIC OR TRANSCRIPTION CUES

Kimmo U. Peltola[1,2], Henna Tamminen[1,2], Paavo Alku[3], Maija S. Peltola[1,2]

[1]Department of Phonetics, University of Turku, Finland
[2]Learning, Age and Bilingualism laboratory (LAB-lab), University of Turku, Finland
[3]Department of Signal Processing and Acoustics, Aalto University, Finland
kimmo.peltola@utu.fi

ABSTRACT

The perception and production of non-native speech sounds is the key to learning a new language. The differences between the native and the target language sound systems cause learning problems, but orthographic conventions may also affect the learning process. We tested whether a misleading orthography in contrast to phonemic transcription affects the manner in which native Finns learn to produce a non-native speech sound embedded in a pseudo word context. After the two day training protocol, the subjects who, in addition to the acoustic stimulation, were exposed to transcription cues, significantly changed their non-native productions according to the target. In contrast, the subjects who trained with the orthographic stimuli, changed their productions away from the acoustic target and towards the visual one. This result suggests that visual information is of crucial importance in learning to modulate articulation according to the target language model.

Keywords: production, training, orthography, transcription.

1. INTRODUCTION

The acquisition of a foreign language demands that new, non-native speech sound contrasts are perceptually categorised accurately and produced appropriately. The native speech sound categories function as a perceptual filter through which non-native speech sounds are perceived even at a preattentive level of processing [16]. Also, the production motor commands develop in accordance with the maternal system [11]. According to the Speech Learning Model [7] the most problematic categories to acquire are those that resemble the native ones but are still distinct. The Perceptual Assimilation Model [2] points towards the same analysis by suggesting that the most problematic difference between sound systems is a situation where two target language categories are assimilated into one category in the native system. Problems in discrimination and identification inevitably result in production errors, since it has been shown that perception needs to be modified before the production patterns can start to develop [8].

Despite the strong transfer from the mother tongue [13, 5], earlier research has shown that language learners may overcome the obstacles and acquire new perceptual and productional patterns. Winkler et al. [23] showed that new memory traces for non-native categories are formed for adult immigrants exposed to a new language in a natural setting and Peltola et al. [18] revealed that these memory traces are also formed for children attending an early foreign language immersion programme. Flege et al. [10] showed that a low age of exposure to a foreign language results in a less foreign accent in the production of non-native speech and also that a continued use of the native language may result in the preservation of a non-native accent [9]. Training studies have also shown that even adult learners can learn to perceive and to produce non-native sound contrasts [4].

Speech perception seems to be bimodal [20] in the way that the visual input may affect the interpretation of the acoustic input and this may be seen e.g. in the famous McGurk Effect [15]. Therefore, in addition to the acoustic and phonological relations between the native and the target language, orthographic conventions may also affect the learning process in literate learners. Peltola [17] suggested that some perceptual categorisation errors in the foreign language vowel perception resulted from the transparency of the native orthographical system and the misleading production hints provided by the non-native non-transparent orthography. Also, Lintunen [14] showed that training on transcription skills may improve the production of target language speech sounds. Therefore, it seems that the written form of language could affect the manner in which non-native speech is perceived and thus the effect ought to be seen also in the production of foreign speech sounds.

In this experiment the aim was to see how visual orthographic and transcriptional cues affect the production when learners are trained with the acoustically equal and theoretically difficult

contrast. It should be carefully noted that the learners' native language has a transparent writing system and a near-phonemic correlation between phonemes and letters. Therefore, the learners link each acoustic speech signal directly to one specific grapheme. This may affect the interpretation of the acoustic stimuli so that it is perceived not on the basis of its acoustic qualities, but on the visual presentation. The hypothesis was that the phonemic writing system of the native language will affect the production learning of a new non-native speech sound so that the primary cue may in fact be the visual one.

2. METHODS

2.1. Subjects

Twenty monolingual Finnish-speaking young adults participated as test subjects. None of them had studied any Nordic languages at the university level, nor lived in Nordic countries other than Finland. Also, none of the subjects used any Nordic languages (other than Finnish) in their daily lives. They participated in this study voluntarily and a written consent was obtained from each participant. None reported of any hearing deficits. The study was carried out with the permission of the Ethics Committee of the University of Turku, Finland.

Subjects were divided into two groups. First group, "Transcription Instructions", consisted of ten subjects (aged 21–34 years, mean 25.8, 5 females). The second group, "Orthographic Instructions", also had ten subjects (aged 20–28 years, mean 24, 5 females).

2.2. Stimuli

The stimuli represented a theoretically extremely difficult contrast for Finnish learners /ʉ/ – /y/. The target word was a pseudo word /tʉ:ti/ where the first syllable vowel is a rounded closed central vowel, a category non-existent in the Finnish vowel system. On the other hand, the non-target word /ty:ti/ contained the Finnish rounded closed front vowel. Therefore, the contrast is perceptually difficult for Finns, since the target vowel is similar [7] to the Finnish vowel category /y/, or both stimulus vowels may be said to assimilate into the Finnish /y/ category [2]. The stimuli were created using a semisynthetic method, where the formant structure of the vowel of interest is synthesised and the glottal pulse excitation is from a naturally uttered word, for more details see [1, 21]. In this manner, the stimuli sound natural, but the acoustic characteristics can be controlled. Therefore, the only difference between the target and the non-target words was on the quality of the first syllable vowel. The target vowel had the values 338 Hz for F1 and 1258 Hz for F2, while the non-target F1 and F2 were 269 Hz and 1866 Hz, respectively. Both in the training as well as the recording sessions, the inter-stimulus interval (ISI) was 3 s. The visual cues appeared on a screen (presented using PowerPoint) at the same pace and in synchrony with the auditory stimuli. However, in the transcription protocol the slide show contained the transcribed word form /tʉ:ti/ while in the orthographic version the visual cue was 'tuuti', showing the vowel grapheme used for denoting the Swedish central vowel in most cases. Therefore, in relation to the target acoustic stimulus, the orthographical visual cue was misleading in denoting the Finnish closed rounded back vowel /u/.

2.3. Procedure

The experiment was carried out in a sound attenued laboratory on two consecutive days. On the first day prior to the experiment, the subjects filled in a questionnaire where their linguistic and educational backgrounds as well as their current health were carefully checked. After that they had an opportunity to adjust the sound volume and familiarise themselves with the experimental protocol by listening to the target- and non-target –stimuli three times in turns. The acoustic stimuli were presented using Sanako Headset SLH-07 and the acoustic outputs were registered with Sanako Lab 100 – software, while the visual cues were shown in a PowerPoint slide show.

On the first day the actual experiment started with a recording block (baseline). After that there was a training session, then again recording and finally another training session. The second day started with the training session followed by a recording and training, and ending with the final recording. Altogether, the protocol consisted of four recordings and intermediate trainings.

In the recording and training sessions subjects repeated in turns the target- and the non-target – stimuli. In the recording session both stimuli were repeated 10 times each and in the training session both stimuli were repeated 30 times in turns. While both groups received an identical amount of perceptual input, the PowerPoint slide show was different: the Transcription Instructions group saw phonemically transcribed presentations of the acoustic signal, while the Orthographic Instructions group was presented with conventional written forms of the stimuli.

2.4. Analysis

The acoustic data were analysed using Praat software (5.3.56) [3]. We obtained the values for the fundamental frequency, and the first and the second formant (F1, F2) from the steady-state phase using Linear Predictive Coding (LPC) Burg algorithm. The F1 and F2 values were subjected to a statistical analysis of variance (ANOVA) using IBM SPSS Statistics (version 22). Also standard deviations were calculated for each F1 and F2 value of each produced word, and these data were also statistically analysed. Further post hoc tests were performed when required.

3. RESULTS

To begin with, we analysed the formant values by subjecting the whole data into an omnibus ANOVA analysis (Group(2) x Session(4) x Word(2) x Measure(2)). The aim was to see whether this type of an analysis would show overall differences between the groups and whether the training protocol affected production performance. The analysis showed the significant main effects of Word ($F(1,18) = 188.91$, $p<0.001$) and Measure ($F(1,18) = 1677.16$, $p<0.001$) revealing that the subjects were able to produce the target word systematically differently from the non-target word and that they maintained the clear distinction between the two formants. In addition, there was an interaction between Word and Measure ($F(3,16) = 211.34$, $p<0.001$) indicating that the two words differed from each other in relation to the formant values. More importantly, the analysis indicated that the productions were differently affected by training (Word x Session interaction $F(3,16) = 6.35$, $p = 0.005$) and that this different kind of a change was found centered in one of the formants (Word x Session x Measure $F(3,16) = 3.40$, $p = 0.005$). Further post hoc tests performed in order to see, which word was affected by the training and which formant was decisive, showed that the main difference was in the F2 value (Group(2) x Session(4) x Word(2): Word x Session $F(3,16) = 6.58$, $p = 0.003$). In addition, further tests (Group (2) x session (4)) performed on the target word F2 values showed that it changed during the training protocol (main effect of session $F(3,17) = 5.21$, $p = 0.010$).

Similar analyses were performed on the standard deviation data, which revealed the main effects of Word ($F(1,18) = 23.93$, $p<0.001$) and Measure ($1,18) = 68.59$, $p<0.001$) as well as a Word x Measure interaction ($F(1,18) = 18.52$, $p<0.001$) suggesting more hesitation in the production of the non-native sound, and naturally more significant deviation in the values of F2. More importantly, the main effect of Session ($F(3,16) = 3.28$, $p = 0.048$) showed a general reduction in the standard deviations as a result of training. No other analyses reached significance.

Table 1. Average Hz and standard deviation values for F1 and F2 in the four recording sessions.

Session	Tran-scription Hz	Ortho-graphic Hz	Tran-scription stdev	Ortho-graphic stdev
1. F1	426	405	30	15
1. F2	1423	1383	182	138
2. F1	425	422	22	17
2. F2	1339	1198	161	125
3. F1	429	423	19	14
3. F2	1352	1156	119	104
4. F1	431	425	19	15
4. F2	1397	1163	119	94

Despite the fact that the general omnibus ANOVA failed to locate overall significant differences between the groups, we decided to perform further tests. This was justified by the clear differences between the groups in the mean formant values visible in Table 1.The values suggest that the Transcription Instructions group (Group 1) shows development of the crucial F2 value towards the F2 values of the provided acoustic model, while the productions of the Orthographic Instructions group (Group 2) appear to result in more /u/ like values. The Group 1 data analysis (Session(2) x Word(2) x Measure(2)) revealed the significant main effects of Word ($F(1,9) = 66.24$, $p<0.001$) and Measure ($F(1.9) = 1046.49$, $p<0.001$) as well as the Word x Measure interaction ($F(1,9) = 90.42$, $p<0.001$) suggesting that the words were separated by the significant F2 value. More importantly, the analysis revealed the main effect of Session ($F(3,7) = 4.52$, $p = 0.046$) showing a change as a function of training. The same analysis of Group 2 data showed the main effect of Word ($F(1,9) = 130.218$, $p<0.001$) and most significantly, the interaction between Word, Session and Measure ($F(3,7) = 4.61$, $p = 0.044$) showing that the values for F2 in the target word changed significantly away from the acoustic model. Furthermore, we performed One-Way ANOVAs to see, whether the F2 values were different in the two groups within any session. The analysis revealed a significant Group difference in the F2 values of the target word vowel ($F(1,18) = 4.51$, $p = 0.048$) and no other differences were found. This reveals that the Groups did not differ at the baseline registration, but a significant difference in the productions was

located in the relevant F2 value of the non-native vowel after training.

Altogether, the analyses showed that training has an effect on the F2 values of the target vowel and that the two types of trainings alter the F2 values differently.

4. DISCUSSION AND CONCLUSIONS

Since non-native speech production is simultaneously very demanding for learners as well as an obligatory and crucially significant part of oral proficiency, there is a high demand for training methods for helping to overcome the acquisition problems. Previous studies have shown that various types of trainings seem to result in less accented target language productions and the role of transcription as a tool for improving non-native production has also been suggested [14]. In addition, the orthographical systems of the native and non-native language may affect the learning process, when the learners become explicitly aware of the target language sound system on the basis of the orthographical system [17, 6]. Our results show that, on the whole, training with both acoustic and visual cues results in production changes, since acoustic values and standard deviations changed as a function of training. Most importantly, the changes are clear in the relevant acoustic parameter F2. More interestingly, the findings show that transcription cues may help the subjects to modulate their productions towards the acoustic goal, while the orthographical visual cues alter the production towards the visual cue. Thus it seems that visual cues are of high significance and they may even be of more importance than acoustic models.

The basic finding that production training with audio-visual cues results in production changes does not in fact imply that visual cues are mandatory, and in relation with previous findings by Tamminen et al. [22] and Iverson et al. 2011 [12], it may well be that production training with mere acoustic stimulation would also result in learning. Irrespective of this, visual cues are clearly of high significance, since they seem to have a more powerful role than acoustic information, if the two are in conflict. This is suggested by the finding that Orthographic Instructions led to changes of production, but not towards the acoustic model, but instead to the opposite direction in accordance with the visual model. The role of Transcription Instruction may either be that it supports the production change in the right direction, or at least it does not interfere with the process of learning to pronounce according to the acoustic model.

In terms of theories of production learning and the connection between speech perception and production, the current results suggest, firstly, that new production templates, or articulation patterns, may evolve during a short training and secondly, that in the interaction between the auditory and visual input, the visual cues may be more significant. According to the Template theory on how speech articulation patterns develop [19], our results suggest that the acoustic model towards which the learner is striving, may be achieved. More interestingly, this acoustic model may be of secondary importance, if the learner is provided with visual cues that are in contradiction with the acoustic model. This suggests that, when learning a new non-native articulatory configuration, the visual cues may inhibit the acoustic cues from being the primary targets.

In conclusion, our present results suggest that new production patterns can be acquired by training to produce the non-native speech items with audio-visual cues. More importantly, it seems that visual information may affect the outcome of production learning even more strongly than the auditory information. This may, in fact, be a significant factor in second language classroom teaching, where the new language is often learned through the written form.

6. ACKNOWLEDGMENT

I would like to thank Utuling doctoral program for funding this project and Sanako Corp. for sponsoring Learning, Age and Bilingualism laboratory. We also warmly thank Pekka Lintunen (PhD) for his valuable comments for this article.

7. REFERENCES

[1] Alku, P., Tiitinen, H., Näätänen, R. 1999. A method for generating natural-sounding speech stimuli for cognitive brain research. *Clinical Neurophysiology* 110, 1329–1333.

[2] Best, C. T., Strange, W. 1992. Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics* 20, 305–330.

[3] Boersma, P., Weenink, D. 2013. Praat: doing Phonetics by Computer. www.fon.hum.uva.nl/praat/

[4] Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., Tohkura, Y. 1999. Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics* 61 (5), 977–985.

[5] Ellis, R. 1985. *Understanding Second Language Acquisition.* Oxford: Oxford University Press.

[6] Erdener, V. D., Burnham, D. K. 2005. The role of audiovisual speech and orthographic information in

nonnative speech production. *Language Learning* 55:2, 191–228.
[7] Flege, J. E. 1987. The production of ”new” and ”similar” phones in a foreign language: evidence of speech perception. *Journal of Phonetics* 15, 47–65.
[8] Flege, J. E. 1993. Production and perception of a novel, second-language phonetic contrast. *J. Acoust. Soc. Am.* 93, 1589–1608.
[9] Flege, J. E., Frieda, E. M. 1997. Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics* 25, 169–186.
[10] Flege, J. E., MacKay, I. R. A., Meador, D. 1999. *J. Acoust. Soc. Am.* 106 (5), 2973–2987.
[11] Guenther, F. H., Vladusich, T. 2012. A neural theory of speech acquisition and production. *Journal of Neurolinguistics* 25, 408–422.
[12] Iverson, P., Pinet, M., Evans, B. G. 2011. Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics* 33 (01), 145–160.
[13] Lado, R. 1957. *Linguistics across Cultures*. Michigan: University of Michigan Press.
[14] Lintunen, P. 2004. *Pronunciation and Phonemic Transcription: A study of advanced Finnish learners of English*. Turku: Anglicana Turkuensia 24.
[15] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264 (5588), 746–748.
[16] Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., Sinkkonen, J., Alho, K. 1997. Language specific phoneme representations revealed by electric and magnetic responses. *Nature* 385, 432–434.
[17] Peltola, M. S. 2004. Extensive language learning affects the perception of vowels. In: Peltola, M. S., Tuomainen, J. (eds), *Studies in Speech communication.* Turku: Digipaino, 43–60.
[18] Peltola, M. S., Kuntola, M., Tamminen, H., Hämäläinen, H., Aaltonen, O. 2005. Early exposure to non-native language alters preattentive vowel discrimination. *Neuroscience Letters* 388, 121–125.
[19] Perkell, J. S., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., Guiod, P. 1997. Speech motor control: acoustic goals, saturation effects, auditory feedback and internal models. *Speech communication* 22, 227–250.
[20] Stevens, K. N. 1989. On the quantal nature of speech. *Journal of Phonetics* 17(1/2), 3–45.
[21] Taimi, L., Jähi, K., Alku, P., Peltola, M.S. 2014. Children learning a non-native vowel – the effect of a two-day production training. *Journal of Language Teaching and Research* Vol.5, No. 6, 1229–1235.
[22] Tamminen, H., Peltola, M.S., Kujala, T., Näätänen, R. 2015. Phonetic training and non-native speech perception – new memory traces evolve in just three days as indexed by the mismatch negativity (MMN) and behavioral methods. *Int. J. Psychophysiol.* Accepted.
[23] Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., Czigler, I., Csépe, V., Näätänen, R. 1999. Brain responses reveal the learning and of foreign language phonemes. *Psychophysiology* 26, 638–642.