A FIRST REPORT ON A PERCEPTUAL TRAINING STUDY USING *PERCY*

Birgitte Poulsen[1], Ocke-Schwen Bohn[1], Christoph Draxler[2]

[1]University of Aarhus, [2]LMU Munich
poulsen@cc.au.dk

Abstract: This paper provides a review of our experiences with using the web-based program PERCY [1] to test and train native speakers of Danish to perceive the English /s-z/ contrast, which Danish does not have. 49 native Danish speakers in two different age groups, 24 seniors and 25 younger adults, participated as either controls or trainees. The trainees conducted 10 web-based training sessions on /zV/ and /sV/ tokens over 3 weeks. All participants were tested for identification accuracy on the initial /zV/, /sV/ tokens, which were trained, as well as on final untrained /Vs/, /Vz/ tokens before and after the 3-week training period. The trainees, but not the controls, were significantly more accurate in perception of /zV, sV/ after the training period.

1 Introduction

Current speech learning models [2], [3] propose that we preserve our speech learning abilities throughout life, and previous training studies, e.g., [4], [5], have found perceptual training to significantly increase perceptual accuracy of a trained contrast. However, research with adults above the age of 40 is very limited [6]. To address this lacuna, we designed a series of studies examining seniors' non-native speech sound learning via web-based training. This paper provides an overview of the procedures and initial results of our first study which focuses on native Danish listeners' ability to produce and perceive the English /s-z/ contrast (Danish has only /s/). The paper includes a review of the experiences with the web-based program PERCY [1] used in testing and training of this contrast. We examined younger (ages 18 – 35 years) as well as older adults (ages 60+) to examine possible differences or similarities in training effects and learning trajectories. This paper reports on the results of the younger group only since analysis of the results of the older group is still pending.

2 Method

2.1 Participants

We recruited participants in two age groups: Seniors (60+) and younger adults (18-35 years). In each age group, participants were randomly assigned to either the experimental (training) sub-group or the control (non-training) group. The senior group (mean age 65.6, range 60-76) had n=19 trainees and n=6 controls (17 f, 8 m). The younger adult group (mean age 23.4, range 20-30) had n=16 trainees and n=9 controls (19 f, 6 m) for a total of 49 participants in the two groups. Average self-reported English proficiency (on a scale of 1 – 5, where 1 is very low and 5 is very high) of the two groups was quite distinct with a much larger spread among the older group (trainees = 3.8, range 2-5, controls = 3.3, range 1-5) and the younger (trainees = 4.5, range 3-5, controls = 4.7, range 3-5).

Because we wanted the trainees to have some "room for improvement" prior to training [4], many potential younger participants, as well as some potential 60+ participants, had to be excluded due to high accurate identification scores at pre-test. Another excluding factor was poor hearing, which, as expected, was mainly relevant for the 60+ group. This group was

initially hard to recruit due to contact channels being web-based. With a more personal recruiting strategy, however, the 60+ group grew to the size of the younger group.

All participants took part in a pre- and a post-test spaced 3 weeks apart. Only the two training groups participated in ten training sessions, each lasting 10-12-minutes, spread evenly over the three weeks. We set up the individual training sessions to accommodate each participant's personal weekly schedule with the condition that the sessions had to be spaced at least two days apart and at most three days apart.

2.2 Procedure

At the initial session in the speech lab at Aarhus University, the participants were presented with a PowerPoint including the instructions, the procedure for a production test, links to three perception tests, and a link to the first training session. After a brief introduction, the PowerPoint provided an explanation of the /s-z/ contrast and participants were instructed on the articulation of English /s/ and /z/. They were asked to touch the front of their throat to feel the absence/presence of vocal fold vibrations during production of /s/ and /z/. Then they completed a delayed-repetition production task consisting of 46 tokens (23 minimal pairs) of real English words with either initial or final /s-z/ (e.g., zinc and sink, and bus and buzz, randomly presented). Their production was recorded for later analysis. This paper will not include any production results as the production test did not involve web-based elements.

Next, three perception tests were administered. The first was designed to acquaint participants with the testing procedure and thus it involved a contrast that was unproblematic for Danish listeners, initial /f-v/, as Danish has both initial /f/ and /ʋ/. This test had 30 /fV-vV/ tokens with V = /ɑ, i, u/, produced by a native English speaker (3 vowels x 5 different tokens x 2 fricatives = 30). The second test examined the perception of final /s-z/ tokens produced by two native English speakers with V = /ɑ, i, u/, 30 tokens each, resulting in 60 /Vs-Vz/ tokens (3 vowels x 5 different tokens x 2 fricatives x 2 speakers = 60). The third tested the perception of initial /s-z/ produced by two native English speakers with V = /ɑ, i, u/, 30 tokens each, resulting in 60 /sV-zV/ tokens (3 vowels x 5 different tokens x 2 fricatives x 2 speakers = 60).

The ten training sessions with feedback trained initial /s-z/. The stimuli for the training sessions were the initial-fricative test tokens presented in two randomizations for a total of 120 trials during each training session. Trainees would hear tokens with either initial /s/ or /z/, then respond by clicking one of two boxes labeled "S" or "Z". After a correct response, the box turned green, (see Figure 1), and after an interval of 0.5 seconds, the next trial was initiated. If the response was incorrect, the box with the incorrect response turned red, and after an interval of 0.5 sec, the box with the correct response turned green, (see Figure 2) the token was played again, and after an interval of 0.5 sec, the next trial was initiated. Training was conducted at home using the web-tool PERCY. By the end of each training session, trainees could see their accuracy rates on the screen. The pre- and post-tests were identical, also done with PERCY, but differed from the training sessions in that no feedback was given for correct or incorrect answers. The pre- and posttest tested the perception of not just the trained initial /s-z/ contrast but also the untrained final /s-z/ contrast to examine whether any training effect would be (positional-) allophone specific or phoneme-general.

Figure 1: The response boxes after a correct answer.

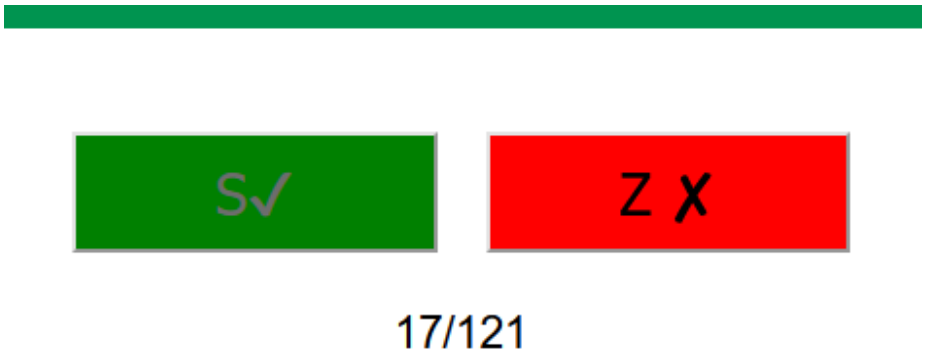

Figure 2: The response boxes after an incorrect answer.

3 Our Experiences using PERCY

The current study consisted of three stages all conducted using the web-tool PERCY: 1) the initial session (including first training), 2) nine evenly spaced sessions of internet-based training at the participant's home, which resulted in a total of ten training sessions per participant, and 3) a posttest, three weeks after pre-test. The control group received no training. The trainees were also invited for a post-posttest eight weeks after the last training session to examine whether perceptual training had had a lasting effect. Because we used PERCY, which allowed each participant to train at home at their own schedule (within some parameters set by us), we were able to include a large number of participants for many training sessions with participants starting the process at different dates throughout the three months of study. Altogether, 76 participants were signed up, and 49 of these completed the program as either trainees or controls.

At the beginning of each test, a username was entered by the participant. Due to the GDPR rules in the EU (General Data Protection Regulation), the participant information had to be anonymized. While we assigned each participant a number at signup to keep track of their appointments and their group status, their access to the online system was done via a code they themselves generated following the guidelines set by the PERCY program. These guidelines included what at first appeared to result in a near perfect anonymization. This involved the first letter from the participant's birthplace, the last letter of her/his first name, the first letter of the mother's first name, the second letter of the father's first name, and the sum of the participant's birthday and month (see Figure 3).

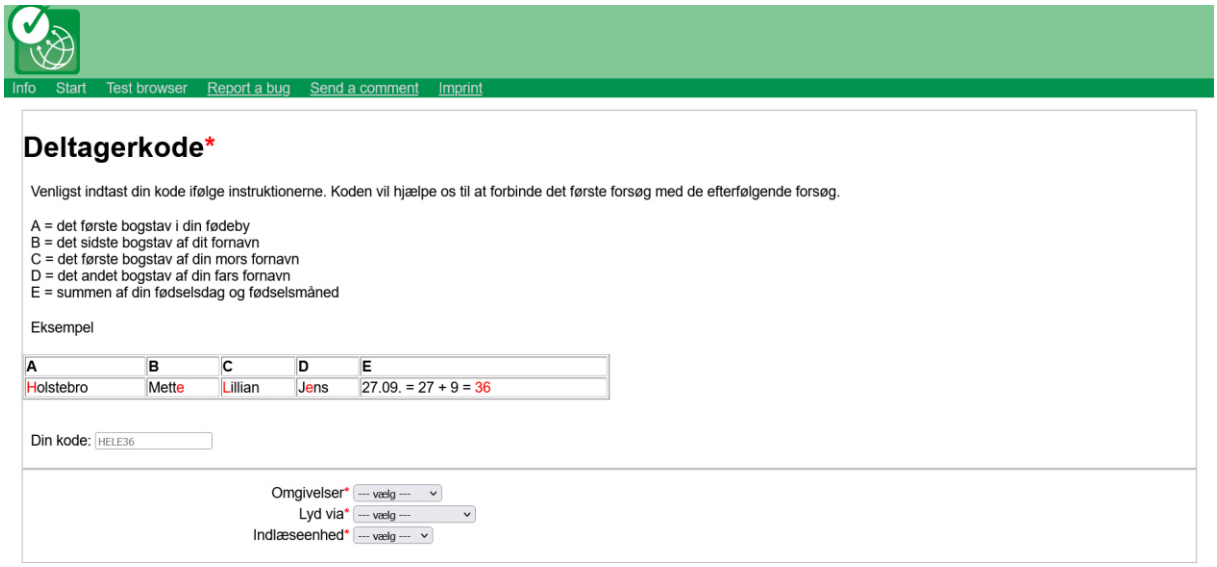

Figure 3: Code generator using personal data to create random anonymous code.

However, this method presented a few challenges: First, some participants were inconsistent in following the guidelines for the generation of their username from one session to the next. This resulted in mismatching usernames between different training or test sessions, an issue we observed in 13 out of the 49 cases. Second, the Danish alphabet contains the letters Æ, Ø, and Å, which are not easily decipherable by R, the tool used for analysis. These letters would occur (with 6 out of the 49 usernames) and had to be manually altered into letters recognizable by R. Third, as many of our participants were born in Aarhus and had a first name ending in <e>, we encountered many very similar usernames, which caused us some confusion at the time of analysis. In future studies, we aim to avoid such complications by assigning the participant a username that only we and they know, write it to them on their schedule information sheet and send it to them by mail.

Once the participants were familiar with the first perception test (identification of /f/ and /v/), they were given the option to advance to the second and third test on their own without interruption. Most were able to do so without error, however, due perhaps to limited computer experience and having to switch back and forth between the PowerPoint and the PERCY testing links, three of the older listeners accidentally completed the final-/s-z/ test twice instead of continuing to the initial-/s-z/ test. This was discovered in time with two of the listeners, who were then assisted in opening the correct link to the initial-/s-z/ test. The third listener, who did not take the initial-/s-z/ test for this reason, was subsequently removed from the experiment records. In future, rather than changing the way we interact with participants to control the process more closely, color coding the test and the matching description in the PowerPoint may be a more intuitive way to prevent this from happening.

We were interested in an immediate access to the accuracy rates for the initial-/s-z/ test. If accuracy rates were too high, there would not be room for improvement and thus, a training effect would not be evident, and we would terminate the process. PERCY made the accuracy rates available and visible on the screen to the participant immediately after completion of the initial-/s-z/ test. In the future, it may benefit the process if the accuracy rates were also visible on the experimenter's screen right away to allow for experimenter evaluation regarding whether to proceed with further participation.

If listeners scored below 90 %, they were included in the rest of the study because this was deemed to leave room for improvement. (A few participants with higher accuracy scores than 90 % were in fact included because we were interested in whether perception training would improve production even for these participants.) Furthermore, as a motivational element in their

training sessions, we wanted trainees to be able to see their accuracy rates upon completion of each training session. This seems to have worked according to intention as trainees reported having kept track of their own progress this way. (After posttest, some trainees in fact reported having obtained higher training scores towards the end of their training than they obtained at their final posttest. This phenomenon may become clearer once all the training reports have been analyzed in depth and at that point, we will examine possible reasons for unexpected declines in accuracy.)

The main overarching focus of the project is to examine whether older and younger adults can learn non-native language sounds and whether any differences exist between the age groups in their learning. Comparing learning trajectories from the training sessions will enable us to analyze possible group differences in perceptual learning between older and younger trainees. This could be done in one go at the end of the four-month-long study by downloading the training data from the first to the last training date from PERCY, as the downloads had to be done by date (see Figure 4). However, perhaps due to the large number of data sets (34 trainees x 10 training sessions = 340), the data could not be downloaded as one set. This problem was solved by downloading the data from two different consecutive timelines and merging the two data sets.

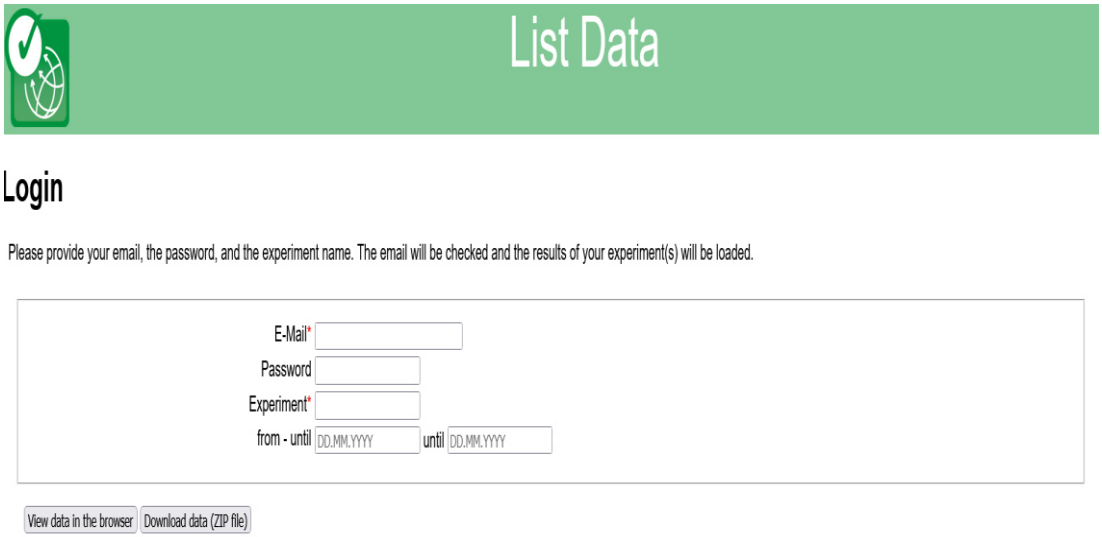

Figure 4: The download page where downloads are available by date.

One feature of the data set download, which was not considered until the data analysis stage, was that each trainee's individual training sessions were identified in columns of session-id and start date (see Figure 5). A column identifying which of the individual trainee's sessions number in his or her 10 sessions would have made learning trajectories simpler to generate. Another aim of downloading training data from PERCY was to be able to register the training intervals and check how often and how many training sessions each trainee had in fact completed before showing up for posttest. Although trainees were responsible for their own diligence in training and thus had a certain measure of freedom in training planning, training had to be done at certain intervals over the course of the three weeks, because it was important that trainees conducted the training according to the same uniform plan to minimize variables. To download only a specific username's training progress, the download had to have been available by username and not as it was, only by date. The problem of whether trainees had complied with the schedule was solved with a clean-up and filter process after download (Figure 5); however, a quick check was not handy at the time. As the time did not allow for an individual check of every trainee before posttest, this introduced some worry about the consistency. As it

turned out in the end, however, very few trainees had to be excluded due to an insufficient number of training sessions.

outdata × importExperimentData-4.R ×

Filter

	experiment	sessionid	startdate	enddate	userid	username	environment	equipment	device
1	Fricatives DK Training	2769745	06.12.2022	06.12.2022	2769744	HEIÃ˜23	studio	headphones	notebook
2	Fricati	2769745	06.12.2022	06.12.2022	2769744	HEIÃ˜23	studio	headphones	notebook
3	Fricati	2769745	06.12.2022	06.12.2022	2769744	HEIÃ˜23	studio	headphones	notebook
4	Fricatives DK Training	2769745	06.12.2022	06.12.2022	2769744	HEIÃ˜23	studio	headphones	notebook
5	Fricatives DK Training	2769745	06.12.2022	06.12.2022	2769744	HEIÃ˜23	studio	headphones	notebook
6	Fricatives DK Training	2769745	06.12.2022	06.12.2022	2769744	HEIÃ˜23	studio	headphones	notebook
7	Fricatives DK Training	2769745	06.12.2022	06.12.2022	2769744	HEIÃ˜23	studio	headphones	notebook
8	Fricatives DK Training	2769745	06.12.2022	06.12.2022	2769744	HEIÃ˜23	studio	headphones	notebook
9	Fricatives DK Training	2769745	06.12.2022	06.12.2022	2769744	HEIÃ˜23	studio	headphones	notebook
10	Fricatives DK Training	2769745	06.12.2022	06.12.2022	2769744	HEIÃ˜23	studio	headphones	notebook

Fricatives DK Training

Figure 5: Data set cleaned in R showing the session-id and the start-date in columns 2 and 3.

Another consistency issue with the web-based training was the possibility of a malfunction without direct technical assistance being available. One specific limitation of PERCY, of which the trainees were informed at the first session, was that it did not work with the Safari browser. This limitation may have been the reason why a few trainees reported difficulties getting the program to run past ten trials on several occasions; most were able to solve the problem themselves the same day or the day after following a consultation with the experiment team. Upon completion of training, these trainees have more than the expected 120 trials on a given date, which could affect the accuracy scores as these trainees would have had slightly more training trials than others.

4 Initial Results

We first compared the identification accuracy of the trainee and the control groups at pretest. For the initial /s-z/ contrast, which would later be trained, the mean percent accuracy was 75.8% (SD=11.6) for the trainee group and 70.7% (SD=16.1) for the control group. A t-test revealed that the difference between the groups was nonsignificant ($t(23) = 0.924$, $p > .3$). Initial results for younger trainees and younger controls for the trained initial /s-z/ are presented in Figure 6. The trainees' mean accuracy increased significantly after training to 92.5% (SD= 9.3) at post-test, $t(15) = 5.870$, $p > .001$). However, the control group's accuracy did not differ significantly over the interval between pre- and post-test (74.2%, SD=17.8 at posttest), $t(8) = 1.631$, $p > .07$.

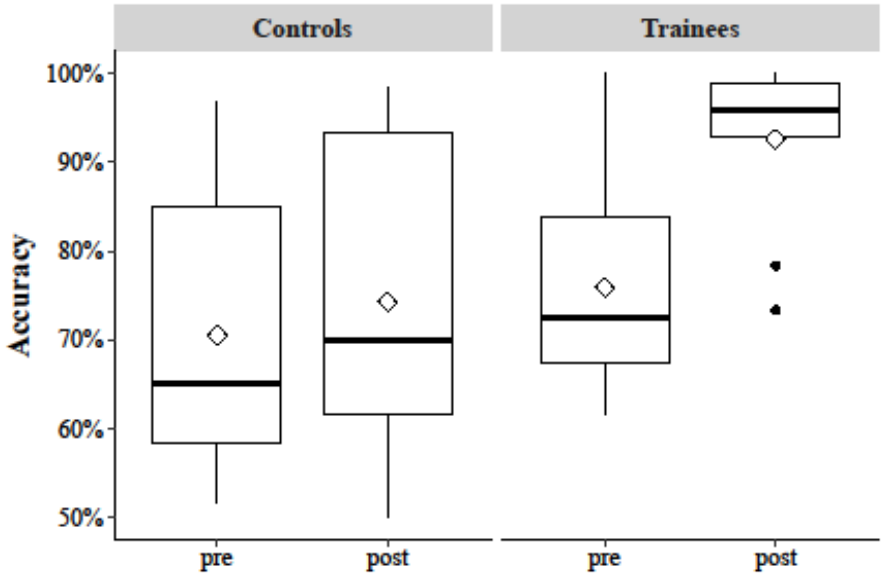

Figure 6 : Accuracy rates of initial /s-z/ for the younger control group and younger training group at pretest and posttest. Diamonds indicate means, bold bars indicate medians.

5 Discussion and Conclusion

The present study examined the effect of internet-based perception training on native Danish speakers' perception of the initial English fricative contrast /s/-/z/ using the online program PERCY. Just as reported in previous training studies, e.g., [4], [5], the results of the present study showed that perceptual training significantly increased perceptual accuracy for the trained contrast, in this case initial English /s/ and /z/. Before pre-test, both the trainee and the control group were informed about the articulatory difference in the implementation of the contrast (presence vs absence of vocal fold vibration). The identification accuracy of the two groups did not differ significantly at pre-test. Trainees, who completed ten sessions of web-based training over a period of three weeks, were significantly more accurate at identifying initial fricatives at post-test than at pre-test. The controls, who had not received any training between testing, did not significantly improve their identification accuracy of initial fricatives /s/ and /z/.

This paper presented our experiences with the methodological aspects of the training study involving PERCY. PERCY enabled us to run perception tests on more than 70 participants and to let 34 of these conduct 10 sessions of perception training from their own homes. Although the vast amount of training data complicated the download as one data set, this was solved by downloading two sets instead. To suit our needs for an immediate ability to check individual training diligence, PERCY might have been set up to returning data sets according to username instead of, as it currently is, according to a time frame delimited by two dates. For future studies, we will know to prioritize this feature.

Perhaps due to the freedom afforded by online training in general rather than directly related to the PERCY program, we did experience many mix-ups with username generation where participants typed in a different username than the one, they used at the first test. Additionally, we should have anticipated the use of Danish letters which are incompatible with R, and we plan to avoid this issue in future studies by assigning usernames which are easy for the participants to remember. Then we administer these in writing at the first on-site test in our lab. For the present, the mistaken usernames can be corrected at the PERCY database.

An additional experience was the occasional participant confusion that occurred when navigating between the PowerPoint and the PERCY links, causing some to repeat a test and

miss the next. In the future, this may be prevented by color coding the backgrounds of the individual tests and matching them to their description in the PowerPoint.

A further feature of PERCY that we found very helpful to both experimenters and participants was the accuracy rates posted at the end of every session. Initially, they aided us in our selection of participants and then they provided motivation to trainees.

6 References

[1] Draxler, C., "Online experiments with the Percy software framework - experiences and some early results," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014, pp. 235–240. Accessed: Dec. 07, 2022. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/564_Paper.pdf

[2] Best, C. T. and M. Tyler, "Nonnative and second-language speech perception," *Lang. Exp. Second Lang. Speech Learn. Honour James Emil Flege*, pp. 13–34, 2007.

[3] Flege, J. E. and O.-S. Bohn, "The revised Speech Learning Model (SLM-r)," in *Second Language Speech Learning.*, pp. 3–83, R. Wayland, Ed. Cambridge, UK: Cambridge University Press., 2021.

[4] Logan, J. S., S. E. Lively, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/: A first report," *J. Acoust. Soc. Am.*, vol. 89, no. 2, pp. 874–886, Feb. 1991.

[5] Trapp, L. and O.-S. Bohn, "Training Danish listeners to identify English word-final /s/ and /z/: Generalization of training and its effect on production accuracy.," in *Proceedings of New Sounds*, pp. 343–350, 2002.

[6] Tamminen, H., T. Kujala, R. Näätänen, and M. S. Peltola, "Aging and non-native speech perception: A phonetic training study," *Neurosci. Lett.*, vol. 740, p. 135430, 2021.