GENERALIZATION OF TRAINING TO AN UNTRAINED PHONETIC ENVIRONMENT: INITIAL /S-Z/ TO FINAL /S-Z/

Birgitte Poulsen, Ocke-Schwen Bohn, and Sidsel Rasmussen

Department of English
Aarhus University, Denmark

ABSTRACT

This study examined whether native speakers of Danish, which has no voiced fricatives, can be trained to perceive the initial English /s-z/ contrast, and whether training on the initial contrast affects the perception of English /s-z/ in final position. 25 native Danish speakers participated as either controls (n=9, no training) or trainees (n=16). The trainees conducted 10 training sessions on /zV/ and /sV/ tokens over three 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 initial /zV, sV/ after the training period. Interestingly, post-training accuracy for untrained /Vs/-/Vz/ also improved for the trainees, and more so for stimuli with voicing cues than for stimuli with vowel duration cues only.

Keywords: Perceptual training of nonnative contrasts, Generalization of training, English /s-z/.

1. INTRODUCTION

Many studies has shown that adults' perception of non-native speech sounds can be improved through perceptual training [1], [2]. In addition, several studies have shown that the efficacy of training extends beyond trained tokens in that learning may generalize to new words spoken by familiar talkers or to novel tokens produced by an unfamiliar talker [3]. However, very few studies have examined whether training for a contrast in one syllable position (e.g., initial) affects perception of this contrast in a different position (e.g., final), but see [4]. Pursuing this question would help clarify whether speech sound learning happens on the level of specific acoustic dimensions, on the level of position-sensitive allophones, or on a more abstract phonemic level.

In initial position, the voicing contrast between English /s/ and /z/ is implemented by the absence (for /s/) or presence (for /z/) of glottal pulsing during the periods of frication. In final position, the nominal "voicing" contrast is produced with presence or absence of glottal pulsing during frication and/or different duration ratios of vowel to consonant: low (short vowel and long fricative) for the voiceless member, and high (long vowel and short fricative) for the voiced member of the contrast. Native English listeners rely mainly on these relational cues for the identification of voicing contrasts of fricatives in syllable-final position [5], but nonnative listeners without native contrasts in this position have been reported to rely more on other cues [6].

The present study examines how perceptual training affects native Danish listeners' perception of the English /s-z/ contrast in the trained initial and the untrained final position. Danish has no voiced fricatives, and previous studies have shown that native Danish speakers assimilate English initial [s] and [z] tokens to the same native category, /s/ [7], [8]. In terms of the Perceptual Assimilation Model (PAM, [9]), this is a Category Goodness assimilation type, in which the goodness-of-fit of English [s] to Danish /s/ is much better than that of English [z].

Previous studies have suggested that experience with a contrast in one phonetic environment (e.g., syllable-initial) can be exploited to perceive the same contrast in a different position [6], [10]. Broersma [6], [10] showed this for Dutch listeners' perception of obstruent voicing contrasts in final position, noting that Dutch has obstruent voicing contrasts in initial and medial, but not final position. Relatedly, Trapp and Bohn [4] reported that training Danish adolescents on the syllable-final English /s-z/ contrast led to improved identification accuracy for this contrast not only in final position, but also in the untrained syllable-initial position. These studies thus address an important issue raised in both the Speech Learning Model [11] and its revised version [12], namely, the level at which speech sounds are perceived and learned: as position-sensitive allophones or at the more abstract level of the phoneme.

The present study examined:

1) To what extent adult L1 Danish speakers' accuracy in identifying the English fricative voicing contrast /s-z/ would improve through internet-based training.
2) Whether any training effect would be allophone specific or phoneme-general, i.e., if trainees rely on trained cues in a different environment.

2. METHOD

2.1. Participants

25 native speakers of Danish (19 f, 6 m, mean age = 23.4 years, range 20-30) participated. None of the participants had spent any extended period in an English-language environment. Self-reported English proficiency (speaking and understanding) averaged 4.6 (range 3-5) on a scale from 1 (very low) to 5 (very high). Participants were randomly assigned to either the experimental (training) group (n = 16) or the control group (n = 9). Participants received an honorarium equivalent to 50 euros (training group) or 25 euros (controls). None of the participants reported hearing impairments.

2.2. Stimuli

The pre- and post- initial-fricative tests consisted of 60 CV tokens produced by two native English speakers: 30 each by a female and a male speaker. Each of the speakers' tokens consisted of 15 /sV/ and 15 /zV/ syllables, with 5 different tokens of V = /ɑ, i, u/ (3 vowels x 5 different tokens x 2 fricative contrasts x 2 speakers = 60). These tokens were selected from the Shannon et al. (1999) corpus [13]. The final-fricative tokens for the pre- and post-training sessions were recorded by two native English speakers (1f, 1m) and validated by three native English speakers. Each of the talkers produced five tokens each of /Vs/ and Vz/ syllables with V = /ɑ, i, u/ for a total of 60 tokens. The stimuli for the training sessions were the initial fricative tokens, presented in two randomizations, for a total of 120 trials.

2.3. Procedure

The current study consisted of three phases all conducted using the web-tool PERCY [14]: 1) a pre-training session (including first training), 2) nine evenly spaced sessions of internet-based training at the participant's home resulting in ten training sessions per participant, and 3) a post-test, three weeks after pre-test. The control group received no training. The pre- and post-tests were identical for the training and control groups, testing 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 allophone-specific or phoneme-general. The pre- and post-training procedures differed on two points: Only the pre-training session contained an explanation and instruction on the articulation of English /s/ and /z/. Here the participants were asked to touch the front of their neck to feel the absence/presence of vocal fold vibrations during production of /s/ and /z/. In addition, only the pre-training session included a familiarization task (identification of initial /fV/ and /vV/) to acquaint participants with the format of the identification tasks. The English initial /f-v/ contrast is unproblematic for native speakers of Danish, who map these fricatives consistently and with high goodness ratings to their native (initial) /f/ and /ʋ/, respectively [7], [8]

The pre- and post-tests consisted of two separate tasks: Identification of final /s/ and /z/, and identification of initial /s/ and /z/. The ten training sessions with feedback trained only initial /s-z/.

At both pre- and post-training sessions participants also took part in a delayed-repetition production task, the results of which are not included in this paper.

Immediately following the identification task on initial /s/-/z/ at our lab, participants assigned to the trainee group completed their first training session. During the following three weeks, the training group completed nine more identical training sessions online at home. Participants were reminded of each training session, and the results obtained via PERCY allowed the experimenters to make sure that participants followed the training schedule. Each session was self-paced and took ca. 10-12 minutes to complete depending on accuracy.

During tests and training, the participants were instructed to listen to the tokens through high quality headphones provided by the experimenters and were asked to identify whether the initial (or, in the case of final fricative, the final) consonant was a /s/ or a /z/. The participants indicated whether they heard a syllable with /s/ or /z/ by clicking on one of two buttons on a computer screen labeled <s> and <z>. During training, immediate feedback was provided after each response: After a correct response, the selected button would turn green, and the next token would be played. After an incorrect response, the button would turn red, the token would be replayed, and, following a short interval, the next token would be played. At the end of each training session, trainees received information on their accuracy rates.

3. RESULTS

The results for the trained initial /s-z/ contrast are presented in Fig. 1, and the results for the untrained final /s-z/ contrast, in Fig. 2.

We first compared the identification accuracy of the trainee and the control group at pre-training. For the initial /s-z/ contrast, which would later be trained, the mean 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 ($t(23) = 0.924$, $p > .3$).

Likewise, the difference in identification accuracy for the final /s-z/ contrast (trainees: 65.3%, SD=14.3; controls: 67.4%, SD=13.9) was also nonsignificant ($t(23) = 0.361$, $p > 0.7$). These results suggest that any difference between the trainee and the control groups at post-training can be attributed to the training regime.

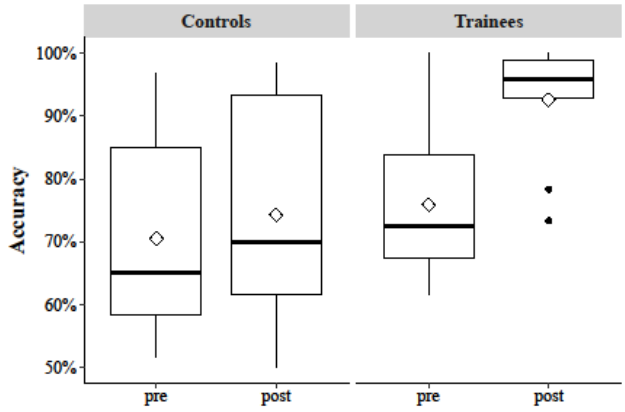

Figure 1: Accuracy rates of initial /s-z/ for control group and training group at pre- and post-training. Diamonds indicate means, bold bars indicate medians.

With respect to the trained initial /s-z/ contrast, the trainees' mean accuracy increased significantly from 75.8% (SD=11.6) at pre-training to 92.5% (SD= 9.3) at post-test, $t(15) = 5.870$, $p < .001$). The control group's accuracy did not change significantly during the interval between pre- and post-training (70.7%, SD=16.1 and 74.2%, SD=17.8), $t(8) = 1.631$, $p > .07$.

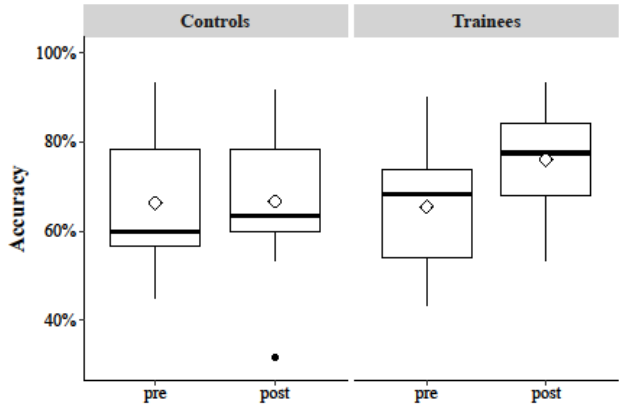

Figure 2: Accuracy rates for final /s-z/ for control group and training group at pre- and post-training. Diamonds indicate means, bold bars indicate medians.

With respect to the untrained final /s-z/ contrast, the group that was trained on the initial fricative contrast improved significantly from an accuracy rate of 65.3% (SD=14.3) to 75.9% (SD=12.0) from pre- to post-training ($t(15) = 3.915$, $p < .01$), whereas the control group's accuracy between the interval of pre- and post-training (67.4.7%, SD=13.9 and 70.7%, SD=13.1) did not change significantly, $t(8) = 0.193$, $p > .1$).

An additional and unexpected finding was that the trainees' overall identification accuracy for final /s/ and /z/ at post-training differed significantly for the two talkers who provided the stimuli, as shown in Fig. 3. At pre-test, the numerical accuracy difference for the tokens from talker LH (mean: 60.6%, SD= 13.3) and talker ZB (mean: 70.2%, SD=17.8) did not differ significantly, $t(29) = 1.727$), $p > .09$). However, after training ZB's tokens were identified significantly more accurately (mean: 82.9%; SD=13.9) than the tokens provided by LH (mean: 69.2% SD=13.5), indicating that something about ZB's tokens signaled the contrast between /s/ and /z/ more clearly than did LH's tokens, $t(29)=2.839$, $p < .01$.

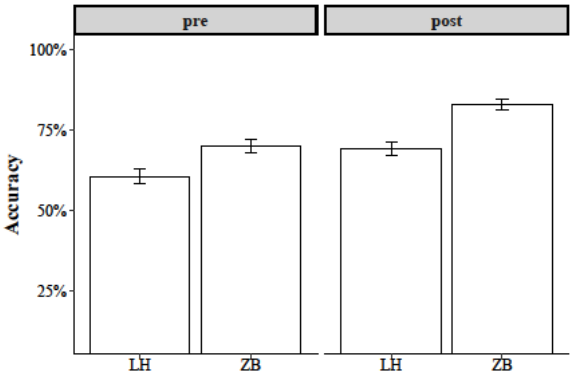

Figure 3: Identification accuracies at pre- and at post-training for syllable-final /s-z/ tokens provided by talkers LH and ZB. Error bars represent +/-1SE.

Acoustic analyses of the final /z/ tokens revealed that the duration ratio of vowel to consonant (V:C) differed significantly for the two talkers: The mean ratio for talker LH was 2.4 (SD=0.34), and the ratio for talker ZB, 2.0 (SD=0.27), $t(28) = 3.6732$, $p < .001$. This was also true for the V:C duration ratios for final /s/ tokens, for which LH's ratio (mean: 0.852, SD= 0.29) was significantly larger than ZB's (Mean: 0.513, SD= 0.08), $t(28) = 4.176$, $p < .001$.

Raphael [5] showed that native English listeners use the V:C ratio as a cue to decide whether the final consonant is voiced or voiceless. This means that LH's V:C ratio would signal the voicing of the final fricative more clearly than ZB's ratio. However, Fig. 3 shows that the nonnative listeners in the present study apparently benefitted less from this cue (duration ratio of V:C) than native speakers would. Further acoustic analyses revealed a systematic

difference between the two talkers' use of voicing in the fricative portions of their syllable final /z/ tokens: Talker LH partly extended voicing from the preceding vowel into the final fricative in only 5 of the 15 tokens, whereas talker ZB consistently extended the voicing from the preceding vowel into all his final /z/ tokens. Fig. 3 shows that the identification accuracy for the tokens for the talker who did not consistently employ partial voicing in the final fricative (LH) was lower at both pre- and post-training than for the talker who did consistently voice (ZB). These findings suggest that the participants in the present study benefited more from the presence of multiple cues (ZB's final voicing as well as a large V:C ratio) than LH's single cue (an even larger V:C ratio).

4. DISCUSSION

The present study examined the effect of internet-based perception training on native Danish speakers' perception of the initial English fricative contrast /s/-/z/. As reported in previous training studies, e.g., [1], [4], the results of the present study showed that perceptual training significantly increased perceptual accuracy for initial English /s/ and /z/. At 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 training over a period of three weeks, were significantly more accurate in identifying initial fricatives at post-training than at pre-training. The controls, who had not received any training between testing, did not significantly improve their identification accuracy of initial fricatives /s/ and /z/.

We also examined whether training of the fricative contrast in one position, syllable-initial, would transfer to an increase in identification accuracy of the contrast in a different position, syllable-final. Interestingly, the results indicate a moderate but significant training effect for the untrained final fricative voicing contrast. A similar albeit reverse effect is reported in [4] where training of final /s-z/ led to an improvement in identification accuracy of initial /s-z/ by adolescent L1 Danish listeners.

Regarding the perception of /s/ and /z/ in final position, we found a numerical difference between the accuracy for the two talkers' tokens at pre-test, and a significant difference at post-test. As expected, see [15], both native talkers' vowel-to-consonant ratio was larger for final /z/ than for final /s/ tokens. However, only one of the talkers, ZB, consistently extended voicing for the preceding vowel into the final fricative portion for /z/, whereas the other talker, LH, did so only for 5 of her 15 tokens. The tokens from the talker who consistently used voicing were more accurately identified by the trainees especially at post-test. This suggests that training on the syllable-initial voicing contrast (where the fricative portions differ with respect to absence vs presence of vocal fold vibration) could have sensitized the trainees to voicing duration during frication. At pre-training, participants were acquainted with the presence of vocal fold vibration for fricative voicing, and in both tests and training tasks, the same two response labels, <s> and <z>, represented the voicing contrast in syllable-initial as well as in syllable-final position. Thus, since the participants were trained to apply vocal fold vibration to the perception of voicing, vocal fold vibration may have been emphasized as a voicing cue at the expense of V:C duration ratio cues [16]. In the present case, this would indicate that speech sound learning happened on the level of specific acoustic dimensions.

Future studies should explore more systematically whether acoustic cues which trainees successfully learn to use in one position can be transferred to a different, untrained position. Addressing this question could contribute to a clarification of what it is that learners attend to and learn in non-native speech learning: Phonemes, position-sensitive allophones, or the specific acoustic cues which are used to signal speech sound differences in different positions.

5. REFERENCES

[1] J. S. Logan, 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.

[2] K. Saito, K. Hanzawa, K. Petrova, M. Kachlicka, Y. Suzukida, and A. Tierney, "Incidental and Multimodal High Variability Phonetic Training: Potential, Limits, and Future Directions," *Lang. Learn.*, vol. 72, no. 4, pp. 1049–1091, Dec. 2022.

[3] S. E. Lively, J. S. Logan, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories," *J. Acoust. Soc. Am.*, vol. 94, no. 3, pp. 1242–1255, Sep. 1993.

[4] L. Trapp 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*, 2002, pp. 343–350.

[5] L. J. Raphael, "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," *J. Acoust. Soc. Am.*, vol. 51, no. 4B, pp. 1296–1303, 1972.

[6] M. Broersma, “Perception of final fricative voicing: Native and nonnative listeners’ use of vowel duration,” *J. Acoust. Soc. Am.*, vol. 127, no. 3, pp. 1636–1644, 2010.
[7] C. S. Horslund and O.-S. Bohn, “Assimilation patterns predict L2 identification accuracy of English initial consonants,” *J. Second Lang. Pronunciation*, vol. 8, no. 2, pp. 218–247, 2022.
[8] O.-S. Bohn and A. A. Ellegaard, “Perceptual assimilation and graded discrimination as predictors of identification accuracy for learners differing in L2 experience: The case of Danish listeners’ perception of Danish fricatives,” in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019, pp. 2070–2074.
[9] C. T. Best and M. Tyler, “Nonnative and second-language speech perception,” *Lang. Exp. Second Lang. Speech Learn. Honour James Emil Flege*, pp. 13–34, 2007.
[10] M. Broersma, “Perception of familiar contrasts in unfamiliar positions,” *J. Acoust. Soc. Am.*, vol. 117, no. 6, pp. 3890–3901, 2005.
[11] J. E. Flege, “Second language speech learning: theory, findings, and problems.,” in *Speech perception and linguistic experience: theoretical and methodological issues*, W. Strange, Ed. Timonium, MD: New York Press, 1995, pp. 229–273.
[12] J. E. Flege and O.-S. Bohn, “The revised Speech Learning Model (SLM-r),” in *Second Language Speech Learning.*, R. Wayland, Ed. Cambridge, UK: Cambridge University Press., 2021, pp. 3–83.
[13] R. V. Shannon, A. Jensvold, M. Padilla, M. E. Robert, and X. Wang, “Consonant recordings for speech testing,” *J. Acoust. Soc. Am.*, vol. 106, no. 6, pp. L71–L74, 1999.
[14] C. Draxler, “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
[15] A. S. House and G. Fairbanks, “The influence of consonant environment upon the secondary acoustical characteristics of vowels,” *J. Acoust. Soc. Am.*, vol. 25, no. 1, pp. 105–113, 1953.
[16] B. Bassetti, P. Escudero, and R. Hayes-Harb, “Second language phonology at the interface between acoustic and orthographic input,” *Appl. Psycholinguist.*, vol. 36, no. 1, Art. no. 1, 2015.