Production inconsistencies delay adaptation to foreign accents

Ann-Kathrin Grohe[1], *Gregory J. Poarch*[2], *Adriana Hanulíková*[3], & *Andrea Weber*[1]

[1] English Linguistics, University of Tübingen, Germany
[2] English Linguistics, University of Münster, Germany
[3] German Linguistics, University of Freiburg, Germany

ann-kathrin.grohe@uni-tuebingen.de, poarch@wwu.de,
adriana.hanulikova@germanistik.uni-freiburg.de, andrea.weber@uni-tuebingen.de

10.21437/Interspeech.2015-627

Abstract

The effects of production inconsistencies and speaker's accented production preferences on speech comprehension were investigated in an eyetracking experiment. Using the visual world paradigm, native speakers of German with L2 English listened to single English words produced by a German speaker that had their *th* either pronounced canonically or substituted with an /s/ or a /t/. Looks to the target word were most likely for the canonical pronunciation and did not differ between the substitutes. However, target looks increased for items with *th* substitutions in the course of the experiment, indicating slow adaptation to inconsistently foreign accented speech.

Index Terms: *th* substitutions, foreign accent, spoken word recognition, speech comprehension, visual world paradigm, eye-tracking

1. Introduction

Speech is highly variable and second language (L2) speech very often deviates extensively from the canonical pronunciation pattern of native speakers. This deviation is known as foreign accent that many speakers display when producing their L2. Recent research has shown that speakers can rapidly overcome initial processing difficulties and adapt to foreign-accented speech (e.g., [1]-[4]). Furthermore, foreign-accented speech is not always more difficult to understand than native speech. In an L2, for example, we can understand L2 speakers that match our native language background equally well or even better than native speakers of that language [5].

Effects of experience with a foreign accent on the comprehension of that accent were explored by Hanulíková and Weber [6]. They looked at foreign accents in L2 listening with an on-line paradigm. Using the printed word visual world paradigm, they compared Dutch to German participants' eye-movement patterns in response to three different pronunciation variants of English words with /θ/. Eye-tracking studies employ an on-line technique that provides a detailed picture of the time pattern of speech processing by tracking participants' eye movements to displayed objects while they are listening to speech. The authors presented an English target word with initial /θ/ (e.g. theft) on the screen together with a competitor word (e.g., left) that mismatched in onset with the target and two further distractor words (e.g., kiss; mask). Simultaneously with the printed word presentation, the target word was presented auditorily, but the /θ/ was substituted with /t/, /s/, or /f/ (e.g., theft pronounced as teft, seft, or feft). All three substitutions occur in Dutch- and German-accented English, with /f/ being perceptually the closest match, and, as confirmed in a subsequent production experiment, /s/ the most frequently chosen substitution by German speakers, and /t/ the preferred substitution of Dutch speakers. Target words with all three substitutions led to correct identification of the intended target word (i.e., visual fixation of the target word), but the ease of mapping from the acoustic signal to the target word depended on the listeners' native language background and not on the perceptual similarity of the substitution. That is, Germans displayed looking preferences for s-variants, whereas Dutch participants did so for t-variants. Critically, this pattern corresponded with the production preferences of each group.

However, this study included a speaker that consistently produced substitutions for one sound and never produced the canonical form. It has previously been shown that inconsistencies in a foreign accent, thus when a speaker intermixes accented tokens with the canonical pronunciation, can affect adaptation to that accent. Witteman, Weber, and McQueen [7] reported a study in which participants' processing of foreign-accented speech differed depending on whether the heard speaker consistently produced a foreign accent or whether the speaker inconsistently intermixed accented and nativelike pronunciation. In a cross-modal priming experiment, one group of native Dutch participants listened to German-accented Dutch words only (consistent accent group). The second group of participants (also L1 Dutch) listened to the same words that included German-accented as well as non-accented pronunciation (inconsistent accent group). Whereas the consistent accent group showed facilitatory priming effects from the beginning of the experiment, such priming effects only occurred for the inconsistent accent group from the second half of the experiment onward. Inconsistency in foreign-accented speech initially slows down processing for L1 listeners; however, they are still able to adapt to the foreign accent.

In their production experiment, Hanulíková and Weber [6] found that even though Germans tended to substitute /θ/ with /s/ in their L2 English, nevertheless 51% of their /θ/ occurrences were pronounced canonically. An inconsistent foreign accent therefore reflects more natural, real-life settings than a speaker who never produces the canonical pronunciation. We were therefore interested in whether inconsistencies in foreign accents also affect the relationship between own production preferences and the proportion of target looks for words with *th* substitutions found in [6].

Do listeners' looking patterns still correspond to their production preferences if they know that the talker is able to produce the canonical pronunciation? Or does the same pattern

apply for L2 listeners as for the L1 listeners in [7] and do inconsistencies delay the effects in [6]? In a visual-world paradigm eye-tracking experiment, we investigated the role of inconsistencies and own production preferences in processing foreign accented speech for L2 listeners.

2. Experiment

We tested second language learners' processing of spoken English *th* words that either matched their own accent, that were produced with a different accent, or were produced with the canonical pronunciation. The procedure used was that by [6], but instead of presenting items that had their /θ/ replaced with an /f/, we included the canonical pronunciation of *th*. Including the canonical form creates the condition of a speaker displaying pronunciation inconsistencies. We are therefore able to test the effects of the interplay of inconsistent foreign accent production with own accent production preferences on the comprehension of accented speech.

2.1. Participants

Fifty-one monolingual students with German as their L1 from the University of Tübingen (mean age = 25.6, *SD* = 6.7; number of females = 39, males = 12) participated in the experiment for a small reimbursement. The participants were highly proficient in their L2 English, did not suffer from any hearing impairments, and reported normal or corrected-to-normal vision.

2.2. Material

We used the same word quadruplets as in [6]. They used 33 English target words with an initial voiceless dental fricative that was either replaced with an /f/, (e.g., feft for theft), an /s/ (seft for theft), or with a /t/ (teft for theft). For our experiment, we included the canonical pronunciation of the voiceless dental fricative /θ/ instead of presenting f-substitutes. This allowed us to test for the effects of inconsistent foreign accent production. As in [6], we paired each target word with a competitor (e.g., left for theft) and two unrelated distractors. Neither the competitor nor the target matched the auditory signal; however, the mismatch is typical for the target but not for the competitor (left cannot become seft or teft). Competitors were included to test whether the substitution items are correctly interpreted as the target. The entire stimuli set consisted of the target word quadruplets and 60 filler and three practice quadruplets. A detailed description of the filler quadruplets can be found in [6]. None of the filler and practice quadruplets included any visually presented word with *th*, or any of the *th* pronunciation variants. The visual stimuli for each single trial were based on these quadruplets. They consisted of the four printed words in black Times New Roman in font size 34 against a white background. Each word was centered on one of four positions on the computer display (192 × 256 pixels, 192 × 768 pixels, 576 × 256 pixels, and 576 × 768 pixels). The positions of the target, competitor, and the two distractors were randomized across trials.

The target words were embedded in the English carrier sentence '*Now you will hear'* and recorded by a native German male speaker who was highly proficient in L2 English, but still had a noticeable German accent. We made digital recordings with an Olympus LS-11 sound recorder (sampling rate 44.1 kHz, 16bit resolution). The sentences were recorded in one session and the resulting file was cut into single sound files using the software Praat [8]. Subsequently, the onset of the target words were labeled, applying the same procedure as [6]. For the fricatives, they determined the onset by the onset of frication, and for the voiceless /t/ by the onset of closure of /t/. This was defined by the ceasing point of the vowel period of the preceding word. As there was a short break between the carrier sentence and the target word, we had to select a different method for /t/ onsets. In contrast to [6], we measured the voice onset times (VOT) of each initial /t/ in the respective targets, determined the burst of sound, and subtracted the VOT from the time of the burst. As can be seen in [9], VOT and closure duration tend to have the same length.

2.3. Design and Procedure

The recording of the target words with their carrier sentences were paired with the respective printed word quadruplet. We used the same lists as in [6], which means there was a different list containing all target words and filler items for each participant. Before each experimental trial, there was at least one filler. Participants were tested individually in a soundproof room with dimmed lights. Wearing closed headphones, they were placed in front of a computer monitor that was connected to an SMI EyeLink 1000 eye tracker (SR Research Ltd., Canada). They were calibrated for their dominant eye and eye movements were recorded with a sampling rate of 1000 Hz. First, they saw written instructions on the screen, which indicated that on each trial, they would first see four printed words on the screen and then, auditorily, a sentence would be played to them. On some trials, the last word of the sentence would be displayed on the screen, whereas on other trials it would not be displayed. Their task was to look at the words on the screen without performing an explicit action. Each trial started with a fixation cross displayed for 1,000 ms that was followed by a display with four printed words. 900 ms after the onset of the display, the auditory sentences were presented over headphones. The following trial started 2,500 ms after the onset of the auditory sentence. An automatic drift correction in the calibration of the eyetracker was initiated by a small fixation circle every six trials. After the eyetracking, the participants performed the LexTale test [10], measuring their English proficiency and the LEAP-Q [11], a questionnaire assessing the participants' language experience and proficiency.

2.4. Coding and Analysis

The data from each participant's dominant eye was used to determine the coordinates and timing of fixations. Only fixations that fell within a cell of one of the four interest areas – target, competitor, and two distractors – were analyzed. Saccades were not added to fixation times. We then analyzed the fixations for the four interest areas in 20-ms steps in a time window from 0 to 1,000 ms after target word onset. Our dependent variable 'target' indicated whether in the respective 20-ms step a participant fixated the target or not.

Generalized linear mixed effects models were used to analyze the eye movement data in the target trials. These models account for the binary nature of our dependent variable by converting the data to log odds (cf. [12]) resulting in a continuous variable. The model used had as fixed effect the type of *th* pronunciation (canonical vs. t- vs. s-substitute), subjects and items were included as crossed random factors. By-timebin (20-ms time steps) and by-list-position random slopes for each subjects and items were also added. Statistical

analyses were conducted for 200-ms windows starting 200 ms after target word onset and ending at 900 ms after target word onset. The onset of 200 ms was chosen because programming an eye movement usually takes about 180 to 200 ms (e.g., [13]). The offset of 900 ms was the same as selected by [6]. The time windows were then shifted by 100 ms (e.g., 200-400, 300-500, 400-600 ms ...), cf. [14]. If consecutive time windows resulted in similar significance values, we re-analyzed the merged, larger time windows. Only the results of these final, accumulated time windows will be reported below. Initially, the entire data set was analyzed and then the data was split into three groups depending on the target item's list position. This allowed testing for effects of adaptation during the experiment.

2.5. Results

2.5.1. Complete item list

For the complete item list, two final time windows, 200-500 ms and 500-900 ms (see Figure 1), were chosen based on the statistical analysis. In the first time window (200-500 ms), there was no effect of *th* substitution. Items that had their */θ/* replaced with /t/ or /s/ elicited the same amount of target fixations as items with the canonical pronunciation (all p-values > .60). For the second time window (500-900 ms), there were significantly more target fixations if *th* was pronounced canonically than if it had been replaced with /s/ (ß = 0.64, SE = 0.17, z = 3.7, p < .001) or /t/ (ß = 0.56, SE = 0.17, z = 3.4, p < .001).

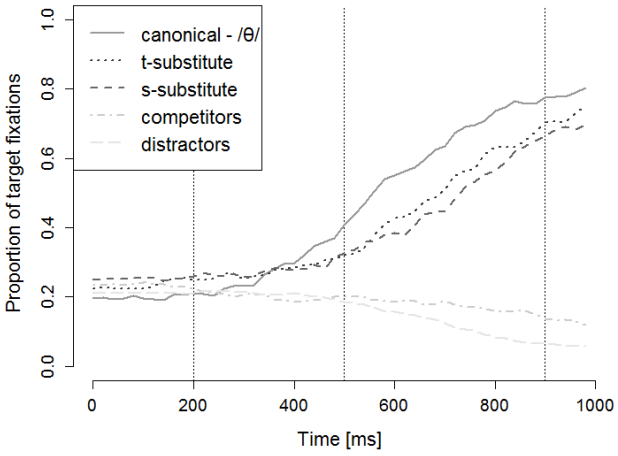

Figure 1: *Proportions of target fixations for the canonical-, s- and t-condition as well as competitor and distractor fixations summarized for all three conditions. The vertical lines divide the two final time windows (200-500 ms and 500-900 ms). Competitor and distractor fixations indicate that all three th-variations were recognized as the intended target.*

So far, the results suggest that the canonical form receives more target looks than the accented items, but only from 500 ms on. Thus, the awareness that the speaker is able to pronounce the canonical /θ/ may elicit adaptations in the course of the experiment. Therefore, we tested for potential effects of adaptation during the experiment by splitting it into three different subsets. The first subset (n = 11) included all targets from the first third, the second subset (n = 11) those targets from the second third, and the third subset (n = 11) the targets from the last third of the experimental list. As each participant had a different list, targets varied per participant. Target fixations between items from the first third and the last third of the experiment were compared.

2.5.2. Split by list position

The same procedure was employed for subset analysis as for the complete data set. Statistical analysis of 200 ms time windows resulted in larger, final time windows that will be reported in the following.

The data including items from the first third of the list were split into a 200-500 ms and a 500-900 ms time window. From 200-500 ms, t-items had significantly more target fixations than canonical items (ß = 0.90, SE = 0.44, z = 2.0, p = .04). For s-items, there was no significant difference to canonical items (p = .17) or t-items (p = .34). From 500-900 ms canonical items resulted in more target fixations than both s-items (ß = 0.86, SE = 0.34, z = 2.5, p = .01) and t-items (ß = 0.83, SE = 0.33, z = 2.5, p = .01).

Analyses of items in the third part of the experimental list resulted in three final time windows. First, from 200-400 ms there were no significant effects (all p-values > .10). Second, from 400-600 ms canonical items had significantly more target fixations than t-items (ß = 0.85, SE = 0.33, z = 2.5, p = .01). The difference between t-items and s-items was not significant (p > .10). Third, from 600-900 ms the advantage of canonical items dropped and their amount of target fixations did not differ any more significantly from the other groups (all p-values > .20).

We eventually compared the number of target fixations for canonical items and accented items in the first third of experimental trials with that of target fixations in the last third. It was possible to merge both t-items and s-items into one group because there was no significant difference between those groups either in the first or in the last third of the experiment. The time window 600-900 ms was selected because this is where the difference between the canonical and the accented targets is not significant in the last third, but it still is in the first third of the experiment. This allows conclusions about whether the number of fixations for accented items increased or whether fixations for canonical items decreased in the course of the experiment. A model with an interaction of trial position (first vs. third third) with condition (canonical vs. substitution) in the fixed part was run. Subjects and items were included as random factors with by-timebin random slopes for both subjects and items. Contrary to the previous models reported above, by-list-position random slopes were not included because this factor was tested in the fixed part of the model. Adding random slopes for by-list-position would account for individual item and subject variation depending on the item's list position and therefore cancel out exactly those potential effects on which our analysis is focused.

As shown in Figure 2, canonical *th* items tended to have less target fixations in the first than in the third third (ß = 0.11, SE = 0.06, z = 1.8, p = .07), but this comparison did not reach significance. Substitution items (t- and s-items) had significantly more target fixations in the third than in the first third (ß = 0.28, SE = 0.08, z = 3.6, p < .001). Whereas target fixations for the substitution items increased drastically in the course of the experiment, they only dropped slightly for canonical items.

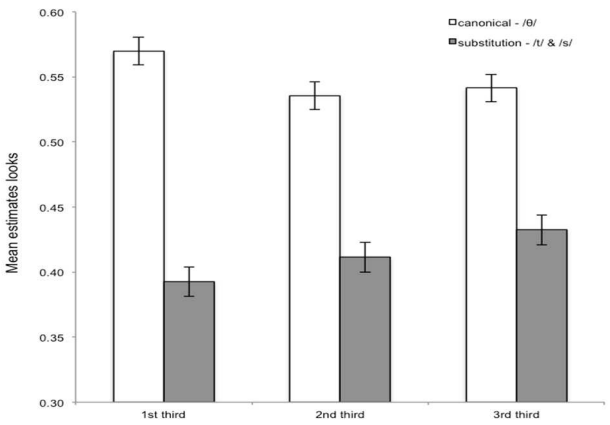

Figure 2: *Probabilities of target fixations for canonical items vs. items with th-substitutions (t- and s-items) in the first, second, and third third of the list for the time window 600-900ms. Error bars are ±1 standard error of the mean.*

3. Discussion

In the present study, the canonical and substitution items resulted in increased target fixations in the time course of each item. The *th* substitutes /s/ and /t/ were still interpreted as *th* even if the canonical /θ/ drew significantly more looks than both substitutions.

Nevertheless, there was a significant preference for canonical items over both s-substitutes and t-substitutes. Contrary to the finding reported by Hanulíková and Weber [6], whose German participants displayed a preference for s-substitutes over t-substitutes in line with the participants' production preferences, the German participants in the present study showed no such preference. This finding may stem from the inclusion of the canonical /θ/ instead of the perceptually closest *th* substitution /f/, inducing an inconsistency in the speaker's pronunciation and clearly indicating to the participants that the speaker was indeed able to produce the canonical form. The role of inconsistencies is highlighted by the analysis of the first vs. third third of the list. During the time course of the experiment, target looks for substitution items increased while those for canonical items decreased slightly. This increase in target looks to substitution items is construed as evidence of successful adaptation to the accent during the experiment. The fact that this adaptation occurs rather late in the experiment and comparatively later than that found in previous studies [1]-[4] is in line with L1 listeners' performance tested in [7] and can be explained by the presence of pronunciation inconsistencies.

The similar preferences for /s/ and /t/ of the German listeners in the present study could also be explained by (1) the duration of pauses between the carrier sentence '*Now you will hear'* and the target word and/or (2) the acoustic properties of /θ/ and /t/ of our L2 speaker. First, pauses before t-items (mean: 308 ms) were longer than before s-items (mean: 260 ms; mean before canonical: 246 ms). Thus, participants spent more time with the visual display before t-items than before s-items, giving them more opportunity to look at the printed words before the actual target was played to them through the headphones. While [6] found no effects of orthography in their study, the longer pauses and thus the increased time for participants to observe the display may have induced such effects. Second, the acoustic properties of /θ/ and /t/ provide information on the similarity between these sounds. They indicate whether our L2 speaker's /t/ is acoustically very similar to his /θ/, which would result in high perceptual similarity and fewer discrimination cues between both sounds. As in [15], amplitude, duration, center of gravity of /θ/ vs. /t/ were measured and compared to L1 speech. A second speaker with American English as L1 recorded the experimental sentences including canonical items vs. t-substitutes. The data were analyzed with linear mixed effects models with an interaction of condition (t-substitute vs. canonical) with speaker (L1 vs. L2) in the fixed part and item as random intercept. Duration analyses resulted in significantly longer durations for [θ] than for [t] for both speakers, but with a higher significance for the L1 than the L2 speaker (L1: ß = 0.14, SE = 0.01, t = 12.5, p < .001; L2: ß = 0.03, SE = 0.01, t = 2.7, p = .01) (see Figure 3 for an example).

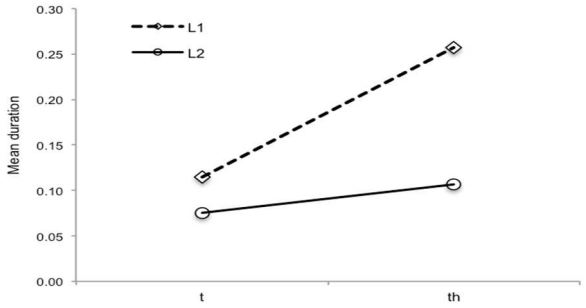

Figure 3: *Mean duration for [t] and [θ] as produced by the L1 speaker vs. the L2 speaker.*

Similar patterns were borne out in the analyses of amplitude and center of gravity. The amplitude was significantly higher for [t] than for [θ] in L1 speech, (ß = 0.02, SE = 0.002, t = 8.8, p < .001), while there was only a marginally significant trend for the L2 speaker (ß = 0.004, SE = 0.002, t = 1.9, p = .06). Finally, center of gravity was significantly higher for [θ] than for [t] in L1 speech (ß = 2507.6, SE = 442.9, t = 5.9, p < .001), while this difference was less marked for the L2 speaker (ß = 1181.6, SE = 422.9, t = 2.8, p = .008). For all three phonetic properties, [t-θ] differences were present in both speakers and in a similar direction; however, the differences between [t] and [θ] were less prominent for the L2 speaker. His [t]-pronunciation was much more similar to [θ] than that of the L1 speaker. This explains why t-substitutes tended to elicit a similar number of target fixations as did s-substitutes.

4. Conclusions

Adaptation to a foreign accent is initially inhibited if both canonical and non-canonical pronunciation forms are present, inducing talker inconsistencies in foreign-accented and native-like pronunciation. However, this initial inhibition in adaptation can be overcome. Adaptation then takes place later compared to when no pronunciation inconsistencies are evident. High acoustic similarity between the intended sound and a substitute enhances the adaptation for that substitute.

5. Acknowledgements

We thank Silvia Dulau and Verena Haug for data collection.

6. References

[1] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[2] C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented English," *J. Acoust. Soc. Am,* vol. 116, no. 6, pp. 3647–3658, 2004.

[3] M. J. Witteman, A. Weber, and J. M. McQueen, "Rapid and long-lasting adaptation to foreign-accented speech," *J. Acoust. Soc. Am*, vol. 128, p. 2486, 2010.

[4] M. J. Witteman, A. Weber, and J. McQueen, "Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation," *Atten Percept Psychophys*, vol. 75, no. 3, pp. 537–556, 2013.

[5] T. Bent and A. R. Bradlow, "The interlanguage speech intelligibility benefit," *J. Acoust. Soc. Am,* vol. 114, no. 3, pp. 1600–1610, 2003.

[6] A. Hanulíková and A. Weber, "Sink positive: Linguistic experience with th substitutions influences nonnative word recognition," *Atten Percept Psychophys*, vol. 74, no. 3, pp. 613–629, 2012.

[7] M. J. Witteman, A. Weber, and J. M. McQueen, "Tolerance for inconsistency in foreign-accented speech," *Psychonomic bulletin & review*, vol. 21, no. 2, pp. 512–519, 2014.

[8] P. Boersma and D. Weenink, "Praat software," *Amsterdam: University*, 2006.

[9] Y. Yao, "Closure duration and VOT of word-initial voiceless plosives in English in spontaneous connected speech," *UC Berkeley Phonology Lab Annual Report*, pp. 183–225, 2007.

[10] K. Lemhöfer and M. Broersma, "Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English," *Behav Res*, vol. 44, no. 2, pp. 325–343, 2012.

[11] V. Marian, H. K. Blumenfeld, and M. Kaushanskaya, "The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals," *Journal of Speech, Language and Hearing Research*, vol. 50, no. 4, p. 940, 2007.

[12] D. J. Barr, "Analyzing 'visual world' eyetracking data using multilevel logistic regression," *Journal of Memory and Language*, vol. 59, no. 4, pp. 457–474, 2008.

[13] G. T. M. Altmann and Y. Kamide, "Now you see it, now you don't: Mediating the mapping between language and the visual world," *The interface of language, vision, and action: Eye movements and the visual world*, pp. 347–386, 2004.

[14] K. Poellmann, H. Mitterer, and J. M. McQueen, "Use what you can: storage, abstraction processes, and perceptual adjustments help listeners recognize reduced forms," *Frontiers in psychology*, vol. 5, 2014.

[15] A. Hanulíková and A. Weber, "Production of English interdental fricatives by Dutch, German, and English speakers," in *New sounds: Proceedings of the Sixth International Symposium on the Acquisition of Second Language Speech*, 2010.