Articulatory differences between L1 and L2 speakers of English

Martijn Wieling[1,2], Pauline Veenstra[1], Patti Adank[3] and Mark Tiede[2]

[1]University of Groningen, The Netherlands
[2]Haskins Laboratories, USA
[3]University College London, UK

m.b.wieling@rug.nl, pauline.veenstra@gmail.com, p.adank@ucl.ac.uk, tiede@haskins.yale.edu

Abstract

In this study, we investigate differences between native English speakers and the English pronunciation of Dutch and German speakers. We focus on the articulatory trajectories obtained using electromagnetic articulography and particularly investigate two sound contrasts: /t/-/θ/ and /s/-/ʃ/. Our results show that while German speakers make both sound contrasts adequately, the Dutch speakers do not distinguish them clearly. To further evaluate these results, both a human Dutch listener as well as an automatic speech recognition (ASR) system classified the pronounced words on the basis of the acoustic recording. Both classifications lined up with the articulatory results. For Dutch speakers, /θ/-words (and /s/-words) were more frequently recognized as /t/-words (and /ʃ/-words). However, the intended utterance was still recognized in the majority of cases for the Dutch speakers. The perceptual results therefore do not support a complete merger of the sounds in Dutch.

Keywords: Second language acquisition, English, Articulography

1. Introduction

Second language learners (L2) learners typically have a clear accent, especially when L2 learning begins later in life (Flege et al. 1995). Speech learning models, such as Flege's Speech Learning Model (SLM; Flege 1995) or Best's Perceptual Assimilation Model (Best 1995) explain these L2 pronunciation difficulties by considering the phonetic similarity of the speaker's L1 and L2. Sound segments in the L2 that are very similar to those in the L1 (and map to the same category) are predicted to be harder to learn than those which are not (as these map to a new sound category).

In this study we focus on two English sounds, the dental fricative /θ/ and the voiceless palato-alveolar sibilant fricative /ʃ/. For both sounds, we assess if Dutch and German L2 speakers of English distinguish these sounds correctly from similar sounds (/s/ vs. /ʃ/, and /t/ vs. /θ/). The sound /θ/ is not included in the phonemic inventory of both languages. The sound /ʃ/ does not occur in the phonemic inventory of Dutch, and can be seen as an allophone of /s/ (though note that loan words from English, such as 'match' do contain the sound). By contrast, the sound /ʃ/ does occur in the phonemic inventory of German

Instead of studying the acoustic differences, here we focus on the underlying articulatory trajectories. Particularly, we will investigate the movement of the tongue during the pronunciation of the two sound contrasts. There are relatively few studies which have investigated L2 (second language) pronunciation differences from this perspective. A notable example is Nissen et al. (2007), who investigated differences between native Korean and Spanish speakers with respect to their L2 English production. In their study, however, they compared the L1 and L2 pronunciations of the Korean and Spanish speakers, rather than including a native English speaker group. In this study, we focus on English pronunciations (i.e. articulation) of Dutch and German speakers, and compare these to the pronunciations of a group of native English speakers.[1]

2. Data collection

For a total of 69 speakers: 22 native English speakers (mean age: 25, 14 women), 20 native Dutch speakers (mean age: 21, 8 women) and 27 native German speakers (mean age: 23, 16 women) we collected articulatory data when speaking English. Initially, we included 71 speakers, but one Dutch speaker did not finish the English-speaking part of the experiment, and the data from one German speaker was excluded as the articulatory and acoustic data were not correctly synchronized. Before participating, the nature of the experiment was explained and each participant signed an informed consent form. Data for the English speakers was collected at the University College London, while the data for the Dutch speakers was collected at the University of Groningen. The data for the German speakers, finally, was collected both at the University of Groningen (10 speakers) and the University of Tübingen (17 speakers). Participants were reimbursed for their time, either monetarily (£15 or €15) or via course credit. Ethical approval was obtained before the experiment from the UCL Ethics Committee and the Ethics Committee Psychology Groningen.

Articulatory data was collected using a portable 16-channel 100 Hz NDI Wave device. The data was corrected for head movement via five sensors attached to the head (i.e. two sensors attached to the left and right mastoid; all sensors were glued using Cyano Veneer Fast dental glue), a reference sensor and a normal sensor attached to the forehead,[2] and a sensor attached to the upper incisor), and rotated relative to the maxillary occlusal plane using a separate biteplate recording (with three sensors attached). The remaining sensors were attached to the midline of the tongue (3), the lips (3) and the jaw (1). The three tongue sensors were positioned as follows: one sensor as far back as possible (T3), one at about 0.5 – 1 cm behind the tongue tip (T1), and one positioned in between the other two sensors (T2). The three lip sensors were attached to the vermillion border at the center of the upper lip as well as the lower lip, and in the right corner of the mouth. Finally, to

[1] This study extends that of Wieling et al. (2015) by including a German speaker group, perceptual results, and employing a more sophisticated analysis.

[2] There appeared to be a synchronization issue between the two system control units (SCUs). For this reason, the participants were asked to nod their head three times at the start of each recording. The two sensors attached to the forehead were connected to different SCUs and used to correct for the synchronization problem.

measure jaw movement, a single sensor was attached to the lower incisor. The NDI Wavefront software was used to record the articulatory data and synchronize it to the simultaneously collected audio (recorded using an Audio-Technica AT875R microphone, which was connected to the control laptop via a Roland Quad-Capture USB Audio interface).

Our material consisted of several hundred words, but here we focus on 10 sets of minimal pairs in English involving /t/ vs. /θ/ (e.g., “fate”-“faith”, “team”-“theme”), and 11 sets of minimal pairs in English involving /s/ vs. /ʃ/ (e.g., “seat”-“sheet”, “lease”-“leash”). Table 1 shows the full list of minimal pairs for both contrasts. Each word (pronounced twice; the word order was random for every speaker) was preceded and followed by a schwa in order to generate a neutral articulatory context around the pronunciation of each individual word. After the data collection, the words were segmented on the basis of the articulatory trajectories. Particularly, we extracted the articulatory positions from the gestural onset of the initial sound to the gestural offset of the final sound using MView (Tiede 2005). For this study, we focus on the anterior-posterior position of the tongue sensor closest to the tongue tip (T1).

For each speaker the anterior-posterior position of the T1 sensor was normalized by subtracting the mean position and dividing by the standard deviation (on the basis of all data collected for a speaker, i.e. hundreds of words).

Table 1: *List of /t/-/θ/ and /s/-/ʃ/ minimal pairs.*

/t/ – /θ/	/s/ – /ʃ/
team – theme	crust – crushed
tank – thank	fist – fished
tick – thick	lease – leash
ties – thighs	plus – plush
tongs – thongs	mess – mesh
fate – faith	rust – rushed
fort – forth	save – shave
kit – kith	seat – sheet
mitt – myth	self – shelf
tent – tenth	sign – shine
	sun – shun

3. Analysis

We analyzed the word-based tongue sensor trajectories using generalized additive modeling (Wood 2017; see Tomaschek et al. 2013, Tomaschek et al. 2013, and Wieling et al. 2016 for applications involving articulatory data), which is a non-linear mixed-effects regression approach. In particular this approach (implemented in the *mgcv* R package) is able to model the non-linear trajectories of the T1 sensor over time, while taking into account a non-linear random-effects structure (i.e. incorporating the dependency structure of the data: each speaker pronounces multiple words). Furthermore, the approach is able to correct for autocorrelation in the residuals of the model. Specifically, when analyzing smooth trajectories, autocorrelation is a large problem and if unaccounted for, the result will be overconfident (i.e. too low) *p*-values. A useful tutorial about how to create a generalized additive model, while also discussing the autocorrelation problem, is provided by Winter & Wieling (2016) as well as Wieling (submitted).

3.1. Replication

The results of the analysis may be inspected and replicated by downloading the data and analysis via the paper package available at http://www.let.rug.nl/wieling/ISSP2017.

4. Results

4.1. Articulatory results

For the /t/-/θ/ contrast, Figure 1 shows that the English and German speakers clearly distinguish /t/ from /θ/, both when the contrast occurs at the start of the word, as well as when it occurs at the end of the word. For these two groups of speakers, the pronunciation of /θ/ is more anterior than the pronunciation of /t/ at the appropriate position in the word. (Note that the difference for the English speakers for words where the contrast was located at the end of the word was non-significant: *p* = .053.) By contrast, the Dutch speakers do not show a significant difference between the two sounds (both *p*’s > 0.2).

For the /s/-/ʃ/ sound contrast the results were similar. Both English and German speakers show a significant difference in the anterior position of the tongue tip sensor (again, the contrast located at the end of the word for the English speakers was not significant: *p* = .097), with a more posterior position for the /ʃ/-words than for the /s/-words at the appropriate position in the word (see Figure 2). The difference for the Dutch speakers was not significant (both *p*’s > 0.4).

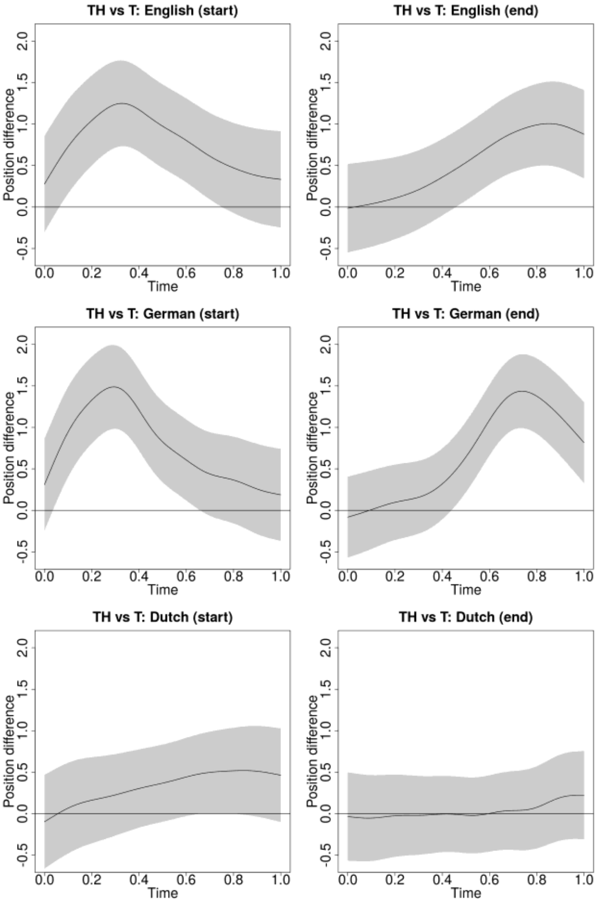
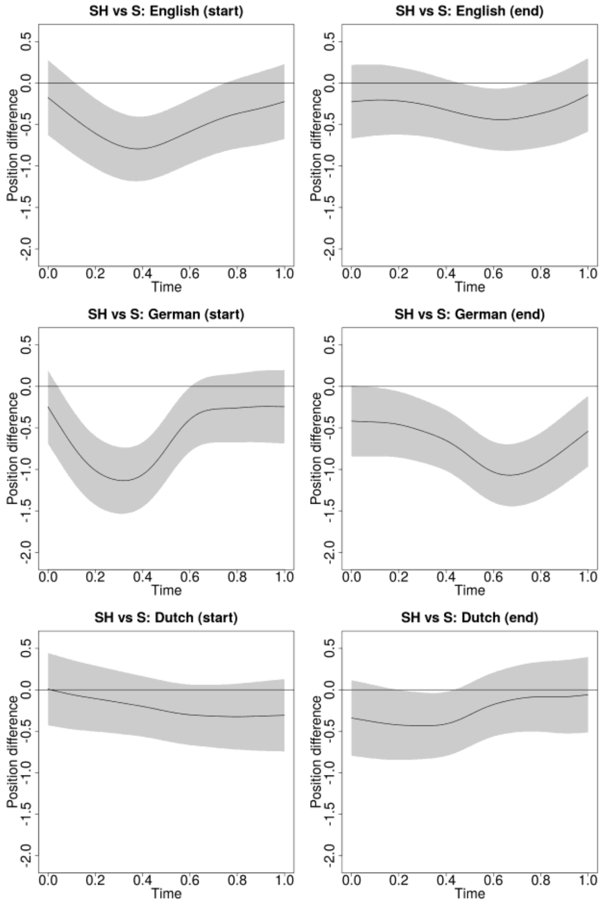


Figure 1: *Normalized position differences between /θ/ and /t/ over normalized time, separately for the three languages and both contrast locations. The shaded band indicates the 95% pointwise confidence interval.*

4.2. Perceptual results

To evaluate these production differences, we asked a Dutch L2 speaker of English to listen to all auditory recordings of the word pronunciations and identify which word was pronounced. For each pronunciation the listener could choose from several alternatives. For the /ʃ/ and /s/-words, there were two alternatives, namely the /s/- and /ʃ/-word (e.g., the listener had to select either ‘crust’ or ‘crushed’). For the /t/ and /θ/-words, there were three alternatives, the /t/ and /θ/-words plus the alternative where an /s/ was used instead (e.g., ‘team’, ‘theme’, ‘seam’).

Table 2 shows the relative frequency of the three alternatives for the words where the speaker intended to produce the /θ/-words (i.e. how often the /θ/-word was confused with an /s/-word or a /t/-word; the total number of utterances was 1245). Similarly, Tables 3 and 4 show how often /s/-words (1440 utterances) and /ʃ/-words (1260 utterances) were identified as /s/ or /ʃ/.


SH vs S: English (start)

SH vs S: English (end)

SH vs S: German (start)

SH vs S: German (end)

SH vs S: Dutch (start)

SH vs S: Dutch (end)

Figure 2: *Normalized position differences between /ʃ/ and /s/ over normalized time, for the three languages separately. The shaded band indicates the 95% pointwise confidence interval.*

Table 2: Perceptual confusion of /θ/-words.

	/θ/	/t/	/s/
English speakers	0.88	0.03	0.09
German speakers	0.86	0.07	0.07
Dutch speakers	0.69	0.17	0.14

Table 3: Perceptual confusion of /s/-words.

	/s/	/ʃ/
English speakers	0.90	0.10
German speakers	0.95	0.05
Dutch speakers	0.81	0.19

Table 4: Perceptual confusion of /ʃ/-words.

	/ʃ/	/s/
English speakers	0.97	0.03
German speakers	0.94	0.06
Dutch speakers	0.96	0.04

Table 5: ASR confusion of /θ/-words.

	/θ/	/t/	/s/
English speakers	0.33	0.06	0.61
German speakers	0.31	0.07	0.62
Dutch speakers	0.19	0.14	0.67

Table 6: ASR confusion of /s/-words.

	/s/	/ʃ/
English speakers	0.93	0.07
German speakers	0.98	0.02
Dutch speakers	0.84	0.16

Table 7: ASR confusion of /ʃ/-words.

	/ʃ/	/s/
English speakers	0.86	0.14
German speakers	0.82	0.18
Dutch speakers	0.91	0.09

Linear mixed-effects regression models with a random intercept for speaker and a single fixed-effect predictor distinguishing the three language groups showed that the /θ/-words were significantly (p < .001) less frequently identified as such for Dutch speakers than for English speakers (German speakers and English speakers did not differ significantly). For the /s/-words, these were significantly (p = .02) less frequently identified as /s/ (i.e. more frequently identified as /ʃ/) for the Dutch speakers compared to the English speakers. The pattern was inverse for German speakers (p = .04) for whom the /s/ was more often identified as /s/ than for the English speakers. Finally, there was no significant difference (p’s > .20) in the detection of /ʃ/-words between the different speaker groups.

As the listener was not a native English speaker, but rather a native speaker of Dutch, this almost certainly will have influenced the perceptual results. Given the large number of utterances (almost 4000), we opted against asking another listener to judge the speech samples, but instead we used an automatic speech recognition (ASR) system (i.e. the Google Cloud Speech API with the language set to British English) to obtain the automatically detected pronunciations. As the acoustic data only consisted of single-word pronunciations, we facilitated the ASR system by setting the dictionary of words to those listed in Table 1 plus the /s/-alternatives to the /t/ and /θ/-words. Tables 5 to 7 show the results and clearly reveal that the performance of the ASR system for the /θ/-words is much lower than the human performance. For the other two sounds performance is relatively similar. Importantly, however, the pattern with respect to the three languages is similar to that observed in Tables 2 to 4, and is also reflected by linear mixed-effects regression models (with a by-speaker random intercept and a single fixed-effect predictor distinguishing the three language groups). Specifically, both the /θ/-words and /ʃ/-words were significantly (p < .001) less well recognized when pronounced by the Dutch speakers than the English and German speakers.

5. Discussion and conclusion

In this study, we have investigated the articulation for two sets of minimal pairs: one set contrasting /t/ from /θ/, and another contrasting /s/ from /ʃ/. In particular we have contrasted two groups of non-native (i.e. Dutch and German) speakers of English to a group of native English speakers. Besides obtaining articulatory data, we have also collected perceptual data. An important characteristic of our study is its large sample size. We have included almost 70 speakers, which, to our knowledge, is the largest sample size in a study employing electromagnetic articulography.

In the context of Flege's Speech Learning Model, our articulatory results for /t/ and /θ/ suggest that these sounds have merged for Dutch L2 speakers of English. While the perceptual results also show an increased confusion between those sounds for Dutch speakers (more so than for the English and German speakers; see Table 2) which is in line with earlier studies of Hanuliková & Weber (2012) and Wester et al. (2007), it is important to note that the pronunciation of the /θ/-words can still be distinguished reasonably well from /t/-words (at a much higher level than chance). In the majority of cases, Dutch speakers are perceived as (correctly) producing a fricative, despite this not being apparent in the anterior-posterior position of the T1 sensor. Of course, this is not completely surprising, as the difference between the two sounds also involves the height of the tongue (i.e. the T1 sensor would be expected to have a more inferior position for /θ/-words than for /t/-words), and we have ignored this dimension here.

With respect to the /s/ and /ʃ/ contrast, the articulatory results again suggest a merger. Here, the perceptual results are also insightful and reveal that the Dutch /s/ is often confused with /ʃ/, even from the perspective of a Dutch listener. A more retracted articulation of /s/ is indeed characteristic of the Dutch language (Collins & Mees 1984) and this clearly affects the English pronunciation. Similar to the /t/-/θ contrast, the /s/ and /ʃ/ can be distinguished correctly much more often than chance, and this again indicates that there is no complete merger from a perceptual perspective.

Given that both /s/ and the /ʃ// are present in the phonemic inventory of German, it is not unexpected that the German speakers contrast them clearly. The German L2 speakers also show the contrast between /t/ and /θ/, despite the /θ/ not being present in the phonemic inventory of German. While this may be due to German speakers confusing the /θ/ more frequently with /s/ than with /t/, our perceptual results do not support this explanation (and therefore contrast with earlier findings of Hanuliková & Weber 2012). For the German speakers, the perceptual results generally support the pattern observed in the articulation. Thus, the German speakers distinguish the two sound contrasts (at least) as well as the native English speakers.

While an articulatory investigation of the pronunciations in a second language is certainly useful, the absence of a clear articulatory difference contrasting a series of minimal pairs for a single (well-chosen) sensor in a single dimension is insufficient evidence for concluding that two different (L2) sounds have merged for second language learners. Consequently, either obtaining perceptual data (as we have done here) to supplement the articulatory (and acoustic) data, or obtaining a more detailed view of the articulatory differences (i.e. considering more sensors in multiple dimensions) is essential.

6. Acknowledgements

This work is part of the research program "Improving speech learning models and English pronunciation with articulography", which was financed by the Netherlands Organisation for Scientific Research (NWO) via a Veni grant awarded to Martijn Wieling. We thank Josine Rawee for the classification of the English pronunciations, and we thank Harald Baayen and the members of his lab for facilitating the data collection for this experiment in Tübingen.

7. References

Best, C. T. (1995). A direct realist view of cross-language speech perception. In: Strange, W. (ed), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, Timonium: York Press, pp. 171-204.

Collins, B., & Mees, I. M. (1984). *The sounds of English and Dutch*. Brill.

Flege, J. (1995). Second-language speech learning: Theory, findings, and problems. In: Strange, W. (ed), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, Timonium: York Press, pp. 233-277.

Flege, J., Munro, M., & MacKay, I. (1995). Factors affecting strength of perceived foreign accent in a second language. *J. Acoust. Soc. Am.* 97, 3125-3134.

Hanulíková, A., & Weber, A. (2012). Sink positive: Linguistic experience with th substitutions influences nonnative word recognition. *Attention, Perception, & Psychophysics*, *74*(3), 613-629.

Nissen, S. L., Dromey, C., & Wheeler, C. (2007). First and second language tongue movements in Spanish and Korean bilingual speakers. *Phonetica*, *64*(4), 201-216.

Tiede, M. (2005). MVIEW: software for visualization and analysis of concurrently recorded movement data. New Haven, CT: Haskins Laboratories.

Tomaschek, F., Wieling, M., Arnold, D., & Baayen, R. H. (2013). Word frequency, vowel length and vowel quality in speech production: an EMA study of the importance of experience. *Proceedings of Interspeech*, pp. 1302-1306.

Tomaschek, F., Tucker, B.V., Wieling, M., & Baayen, R. H. (2014). Vowel articulation affected by word frequency. *Proceedings of the 10th International Seminar on Speech Production*, Cologne, May 5-8, pp. 429-432.

Wester, F., Gilbers, D., & Lowie, W. (2007). Substitution of dental fricatives in English by Dutch L2 speakers. *Language Sciences*, *29*(2), 477-491.

Wieling, M. (submitted). Generalized additive modeling to analyze dynamic phonetic data: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. Available at: http://martijnwieling.nl/files/GAM-tutorial-Wieling.pdf.

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S.N., & Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics*, *59*, 122-143.

Wieling, M., Veenstra, P., Weber, A., Adank, P., & Tiede, M. (2015). Comparing L1 and L2 speakers using articulography. *Proceedings of the 18th International Congress of Phonetic Sciences* (Vol. 18). International Phonetic Association.

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, *1*(1), 7-18.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.