A Real-Time MRI Study of Articulatory Setting in Second Language Speech

Andrés Benítez[†]*, Vikram Ramanarayanan*[‡]*, Louis Goldstein*[†] *and Shrikanth Narayanan*[†‡]

[†]Department of Linguistics, University of Southern California, USA
[‡]Ming Hsieh Department of Electrical Engineering, University of Southern California, USA
a.benitez@usc.edu, vramanar@usc.edu, louisgol@usc.edu, shri@sipi.usc.edu

Abstract

Previous work has shown that languages differ in their articulatory setting, the postural configuration that the vocal tract articulators tend to adopt when they are not engaged in any active speech gesture, and that this posture might be specified as part of the phonological knowledge speakers have of the language. This study tests whether the articulatory setting of a language can be acquired by non-native speakers. Three native speakers of German who had learned English as a second language were imaged using real-time MRI of the vocal tract while reading passages in German and English, and features that capture vocal tract posture were extracted from the inter-speech pauses in their native and non-native languages. Results show that the speakers exhibit distinct inter-speech postures in each language, with a lower and more retracted tongue in English, consistent with classic descriptions of the differences between the German and the English articulatory settings. This supports the view that non-native speakers may acquire relevant features of the articulatory setting of a second language, and also lends further support to the idea that articulatory setting is part of a speaker's phonological competence in a language.

Index Terms: articulatory setting, speech production, second language speech acquisition, real-time MRI.

1. Introduction

Difficulties in second-language (L2) speech production are partly due to cross-linguistic differences in phonemic inventories, phonetic realizations and prosodic patterns of the native and the target languages [1, 2]. However, recent work suggests that languages also differ in their *articulatory setting* [3, 4], the set of postures that the vocal tract articulators tend to adopt when they are not engaged in any active speech gesture [5, 6, 7]. Articulatory setting (also known as phonetic setting, organic basis of articulation or voice quality setting) reflects a global configuration of the speech apparatus that is believed to provide a common, long-term quality to speech produced by speakers of the same language. For example, native speakers of a given language may have a tendency to maintain their lips protruded throughout speech, or to keep the tongue slightly retracted into the pharynx. If articulatory setting is part of the knowledge that native speakers have of their language, then this is another aspect of the target language that L2 speakers need to learn in order to achieve native-like proficiency. This study investigates whether L2 speakers can acquire the articulatory setting of a second language.

Notions of language-specific articulatory settings have been present in the phonetic literature for at least several decades [5, 6, 8], but have only recently been experimentally validated through imaging of the inter-speech posture (ISP), the static position adopted by the articulators during pauses at overt syntactic junctures [3, 4]. For example, Gick *et al.* [3] found differences in the ISPs of 5 French and 5 English speakers; compared to the French ISP, the English ISP had a narrower pharynx, a higher tongue tip and tongue body, a more protruded upper lip, and a less protruded lower lip. Noting that the variability of the ISP was very similar to that of an actual speech target (namely, /i/), Gick *et al.* [3] further proposed that the ISP might be specified in the same way as actual speech targets of the language.

If articulatory setting is indeed part of the sound system of a language, it is reasonable to wonder to what extent it might be learned by L2 speakers. Recent work using a combination of ultrasound and flesh-point tracking [4] shows that even some early bilinguals (who have learned two languages before puberty and use them on a regular basis) may not employ distinct language-specific ISPs, which causes them to be perceived as non-native speakers of one or both of their languages. Therefore, it is possible that one of the reasons why adult L2 learners typically produce foreign-accented speech even after many years of exposure to L2 is that they never acquire the articulatory setting of the target language. To test this, our study examines the ISPs of a group of L2 speakers in their native and non-native languages. If the speakers have acquired even in part the articulatory setting of their non-native language, we expect to find language-specific ISPs that are distinct in a manner consistent with the available descriptions of the languages' articulatory settings.

In particular, we analyze the ISPs of three native speakers of German who had learned English as a second language. Descriptions of the differences between the German and the English articulatory setting are only available through classic impressionistic comparisons, which consistently noted that the default posture of the tongue in English is wider or lower and more retracted than in German [9, 10, 11]. More recent, although limited, empirical data seems to corroborate the characterization of the tongue's preferred position in English as retracted, at least when compared to the posture in French [3]. Therefore, if the L2 speakers in our study have learned aspects of the English articulatory setting, we may observe a more retracted tongue in the ISPs in English. Additionally, we analyze the absolute rest postures of the same speakers before speaking in English and in German. This posture is adopted simply for respiration when speakers are presumably not in speech mode, and so we do not expect to observe differences between the postures adopted before speaking English and those adopted before speaking German.

To observe ISPs in the course of fluent speech, we employ real-time magnetic resonance imaging (rtMRI) of the vocal tract [12]. This imaging technique allows for a complete midsaggital examination of posture along the entire vocal tract that is unavailable with other methods used for observing ISPs, like ultrasound or flesh-point tracking techniques (see [13] for a review).

2. Method

2.1. Subjects and stimuli

Three male native speakers of German were imaged, using a custom MRI protocol developed for the examination of speech production [12], while reading *The Rainbow Passage* and *The North Wind and the Sun* passage in English and in German. The participants were high proficiency non-native speakers of English, enrolled in the University of Southern California at the time. None of them reported any history of speech or language disorders.

Each speaker lay supine in the MR scanner, with his head padded with foam rubber to minimize movement. A mirror placed over the subject's face allowed him to see out of the scanner and read the stimuli projected onto a screen. The participants read each passage (divided into two chunks of similar length) first in English and then in German, and then again in the same order, for a total of 16 recordings per participant (2 passages by 2 parts by 2 languages by 2 repetitions).

2.2. Data acquisition

Midsagittal real-time MR images of the vocal tract were acquired on a GE Signa 1.5T scanner, using a 13 interleaf spiral gradient echo pulse sequence. The slice thickness was approximately 3 mm, and the image resolution in the sagittal plane was 68 by 68 pixels (covering approximately 3 mm per pixel). The images were reconstructed as 22.4 FPS video using a sliding window technique. Field-of-view (FOV), similar to a zoom factor, was set depending on the subject's head size. Audio was recorded at 20 kHz simultaneously with the image acquisition, and later noise-reduced and video-synchronized [14]. The resulting AV recordings allow for a complete midsagittal view of the subject's vocal tract, including glottis, pharynx, and oral and nasal cavities.

2.3. Contour and feature extraction

In order to capture the time-varying vocal tract outline and the position of the articulators during the MR images, we employed an algorithm that automatically extracts air-tissue boundaries of the articulatory structures by hierarchically optimizing the observed image data fit to an anatomically informed object model using a gradient descent procedure [15].

Following this contour extraction, we employed a procedure specifically designed to automatically extract features that capture vocal tract posture [7]. This algorithm first uses the previously extracted contours to calculate cross-distances in all frames for lip aperture (LA), velic aperture (VA), tongue tip constriction degree (TTCD), tongue dorsum constriction degree (TDCD), and tongue root constriction degree (TRCD). The cross-distance for LA is computed as the minimum distance between the upper and lower lip contours. In the same way, the VA cross-distance is taken as the minimum distance between the velum and the pharyngeal wall. For tongue-related cross-distances, which have no obvious boundary landmarks, we used locations along the palate and pharyngeal wall where the vocal tract can be maximally (and ideally, completely) constricted. For example, TTCD was calculated as the distance between the tongue tip and the average coordinate point of contact on the palate during the coronal stops /t, d/ in the German recordings. Similarly, for TDCD, we obtained the mean point of contact on the palate during dorsal stops /k, g/ and calculated its distance to the tongue dorsum. In the case of TRCD, we used instances of the low back vowel /a:/, where the tongue was maximally (although not completely) constricted against the pharyngeal wall. Note that the points of constriction used for calculating tongue-related cross-distances were selected for each speaker from the German recordings. This was to ensure meaningful comparisons both across subjects and across languages, as the points of constriction can be expected to be maximally consistent within and between speakers in their native language. Finally, the lowermost boundary of the vocal tract was found as the minimum distance between the root of the epiglottis and the pharyngeal wall contour.

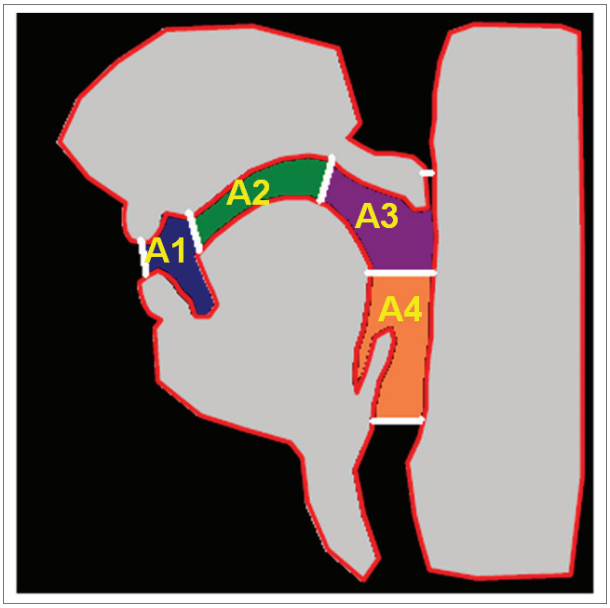

Figure 1: *A schematic depicting the concept of vocal tract area descriptors or VTADs (adapted from [15]). These VTADs are bounded by cross-distances (depicted by white lines), and are, in order, from lips to glottis: lip aperture, tongue tip constriction degree, tongue dorsum constriction degree, velic aperture, tongue root constriction degree, and the epiglottal-pharyngeal wall cross-distance.*

After computing these cross-distances, the algorithm proceeds to partition the vocal tract airway into four areas (A1, A2, A3, and A4) referred to as *vocal tract area descriptors* (VTADs). VTADs have been shown to capture vocal tract posture (and thus articulatory setting) in a way that is robust to head movement. [7]. As shown in Figure 1, A1 is the area bound by LA and TTCD cross-distances; A2 is the area between TTCD and TDCD; A3 is the area between TDCD and TRCD; and A4 is the area below TRCD. Given a lower signal-to-noise ratio in the area of the pharynx, we used the sum of A3 and A4 areas (henceforth referred to as A3-4) as a more robust measure of shape of the pharyngeal region. Additionally, the jaw angle (JA) was computed as the obtuse angle between linear regression lines fitted to the pharyngeal wall and chin contours. In order to account for speaker-specific traits and allow for meaningful comparisons, each variable (VTADs, cross-distances and JA) was normalized per speaker by its range such that the transformed variable took values between 0 and 1.

2.4. Frames of interest

For the purposes of this study, two types of vocal tract postures are of relevance: inter-speech postures (ISPs) and abso-

lute rest postures. For ISPs, we selected frames from the rtMRI recordings corresponding to grammatical inter-speech pauses [16], manually identified in the audio track as periods of silence between overt syntactic constituents, where speakers were expected to adopt the ISP. For absolute rest posture, the first three and last three frames of each recording were selected. These frames corresponded to the moments immediately after turning on the scanner and before turning it off for each recording, when the speaker was presumably not engaged in any speech activity. Frames where the speaker was visibly yawning, swallowing or biting his lip or tongue were manually excluded from further analysis. The previously computed VTADs A1, A2 and A3-4, cross-distances and JA values were extracted from all selected ISP and absolute rest frames, and collected for subsequent statistical analysis as dependent variables.

2.5. Statistical analysis

R software was used to conduct all statistical analyses. To minimize the potential effects of non-normality or unequal variance, robust statistical techniques and measures were adopted (see [17] for details). Planned linear contrasts were performed using a percentile bootstrap analysis based on the 20% trimmed means. Familywise Type I error rate was controlled at α = 0.05 using Rom's method [18].

Bootstrap analyses were chosen because they avoid the assumptions of normality and equal variance that can greatly reduce power when using traditional ANOVAs and *t*-tests [17]. Instead, they generate a bootstrap distribution by repeatedly sampling with replacement from the original dataset, and confidence intervals can then be computed by using the percentiles of the estimated sampling distribution. The trimmed mean was chosen as a measure of central tendency in the analyses because it has been shown to maintain high power when testing from both normal and non-normal distributions [19]. The 20% cut-off, which removes the lowest 20% and the highest 20% of observations, has proven to be a good default in most situations [19].

For each speaker, the comparisons of interest were the differences in trimmed means of VTADs (A1, A2 and A3-4), cross-distances (LA, TTCD, TDCD and TRCD) and JA. For each of these dependent variables, we performed the following planned comparisons based on *a priori* predictions: *(1)* German ISP and English ISP, *(2)* German ISP and German Rest, *(3)* English ISP and English rest, and *(4)* German rest and English rest. Note that the percentile bootstrap method used for the comparisons does not require that an *omnibus* test be performed and rejected before testing specific hypotheses, and that doing so would result in a decrease in power.

3. Results

Figure 2 summarizes the results. Most relevantly, the comparisons of inter-speech postures in German and English showed that there were significant differences ($p < .05$) in the postures adopted by all speakers between the two languages. VTAD A1 was less constricted in English for speakers 1 and 3, as was A2 for speakers 1 and 2, most likely indicating an overall lower tongue position in English. Notably, all speakers had a more constricted A3-4 in English than in German, suggesting a narrower pharynx and thus a more retracted tongue root.

The comparisons of cross-distances shed further light into how to interpret the VTADs results. All speakers had a lower tongue tip in English (i.e. larger TTCD), but no differences in lip aperture (LA) between languages, which indicates that A1 was less constricted in English due to differences in TTCD rather than in LA. Two of the speakers (2 and 3) also had a lower tongue dorsum in English. Notably, the results in TRCD were consistent with those obtained in area A3-4 for all speakers, indicating that the tongue root was more constricted against the pharyngeal wall. Finally, JA was different only in the case of Speaker 2, who had a lower jaw during English ISPs.

Further, the comparisons of German ISP and German rest posture, and of English ISP and English rest posture, revealed that absolute rest postures were significantly different ($p < .05$) to ISPs for all speakers in almost all VTADs and cross-distances, and in JA. During absolute rest postures, the cross-distances were considerably smaller and the jaw angle wider, and VTADs were more constricted, indicating an overall much more closed vocal tract than during ISPs. The variability of rest postures was also higher than that of ISPs. Finally, the comparisons of absolute rest posture in German and English showed that these were not different across languages in any areas for any speaker ($p > .05$).

Taken together, the results indicate that the three speakers patterned very similarly. In all cases, they had a more retracted tongue position during IPSs in English than in German, and this was clearly captured both in area A3-4 and the TRCD cross-distance. For two of the speakers, this tongue retraction was accompanied by a lower tongue body, while one speaker maintained the tongue body height despite the retraction. Additionally, all speakers had a lower tongue tip in English ISPs.

4. Discussion

The results demonstrated three main findings. First, the ISPs adopted by the L2 speakers in the study varied between their native and their non-native languages in significant ways. First, all three speakers consistently adopted a more retracted tongue posture, accompanied by a lower tongue tip, during ISPs in English than in German. Two of the speakers also had a lower tongue dorsum in English. This overall tongue configuration is largely in line with the classic descriptions mentioned earlier, particularly in regards to the tongue being more retracted and lower in English than in German. The fact that all speakers patterned in a very similar way suggests that they may have learned at least some aspects of the English articulatory setting.

Second, ISPs in both languages were significantly different from the absolute rest postures that the speakers adopted presumably only for respiration. This finding replicates results that show that the vocal tract is considerably more constricted during absolute rest postures than during ISPs, and that the former are more variable, possibly reflecting a lesser degree of active control [7]. Finally, as predicted, the postures adopted during absolute rest before speaking in English and those adopted before speaking in German did not differ. This suggests that, unlike ISPs, absolute rest postures do not capture language-specific traits. We note, however, that the very small sample sizes of rest postures only allow for detecting somewhat large differences as statistically significant. While highly unexpected, if there are in fact small differences between the absolute rest postures adopted before speaking one or the other language, these may have gone undetected in the statistical tests. Although it is very unlikely that absolute rest postures capture any language-specific characteristics, a much larger sample size would be required in order to confirm this with a higher degree of certainty.

In conclusion, the results of this study support the view that proficient non-native speakers may acquire relevant features of

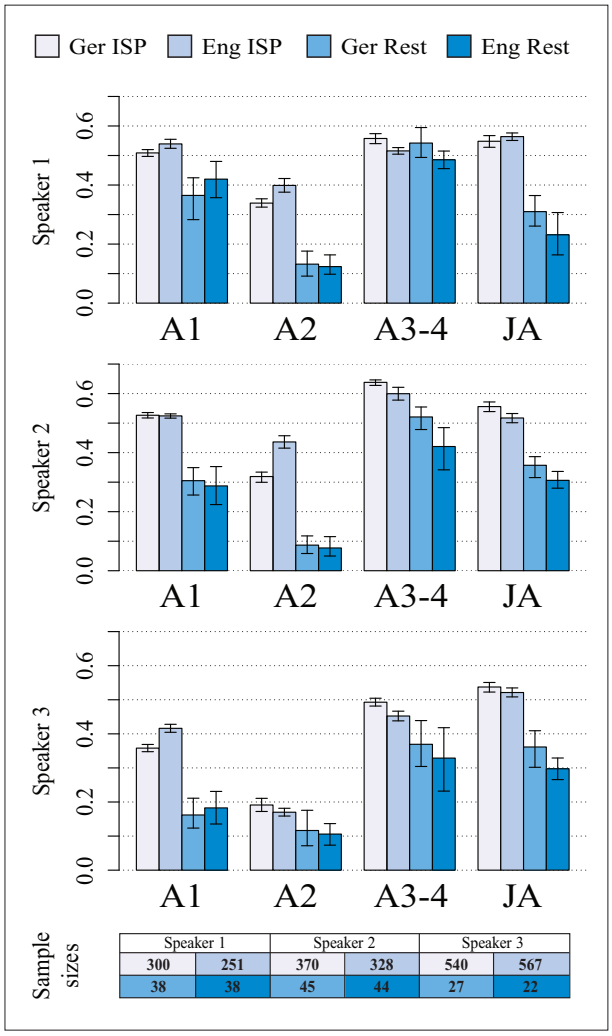

Figure 2: ***(Top)*** *20% trimmed means and 95% confidence intervals of A1, A2, A3-4 and JA for each speaker and posture category. Lower values indicate a smaller (i.e., more constricted) area for A1, A2 and A3-4, and a wider angle for JA. For clarity of display, A3-4 values shown here are scaled down by two.* ***(Bottom)*** *Sample sizes (i.e., frames extracted from rtMRI video).*

the articulatory setting of a second language. It is, of course, difficult to confirm this without a larger sample size of both native and non-native speakers of the same language. While the results do not necessarily mean that the speakers in this study were adopting the same ISPs as native English speakers, two of our findings do provide some initial support for the idea. First, all three speakers patterned in very similar ways; second, the pattern was largely consistent with classic descriptions of how the English and German articulatory setting differ. The former might conceivably be a first-language effect, and so work is underway to examine whether non-native speakers of English from other language backgrounds behave similarly. The latter, on the other hand, will need to be backed up by modern instrumental studies that more carefully describe what articulatory settings look like in each language. In any case, the different ISPs adopted for each language lend further support to the idea that articulatory setting is part of a speaker's phonological competence in a language.

Additionally, we have shown that real-time magnetic resonance imaging of the vocal tract is a useful tool for observing language-specific articulatory settings in the area of L2 speech acquisition. Many interesting issues remain for future study. For example, despite extensive literature advocating to incorporate articulatory setting into L2 instruction, it is unclear how a non-native articulatory setting relates to degree of foreign accent in the target language. Relatedly, it has recently been suggested that ISPs are more mechanically advantageous than absolute rest and other speech postures with respect to articulation [20]. In this line, it would be important to investigate whether non-native articulatory settings are less dynamically advantageous than their native counterparts, as this could explain some of the difficulties faced by non-native speakers in producing fluent, unaccented speech in their second language.

5. Acknowledgements

We acknowledge useful comments and suggestions from Rachel Walker and audiences at USC Speech Production and Articulation kNowledge Group (SPAN) and PhonLunch/Phonetics Lab meetings. This research was supported by a Fulbright Scholarship to the first author from the U.S.-Spain Fulbright Commission and NIH Grant DC007124.

6. References

[1] C. T. Best, "A Direct Realist View of Cross-Language Speech Perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Speech Research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–206.

[2] J. E. Flege, "Second Language Speech Learning: Theory, Findings, and Problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 233–277.

[3] B. Gick, I. Wilson, K. Koch, and C. Cook, "Language-specific articulatory settings: evidence from inter-utterance rest position." *Phonetica*, vol. 61, no. 4, pp. 220–33, 2004.

[4] I. Wilson and B. Gick, "Bilinguals use language-specific articulatory settings," *Journal of Speech, Language, and Hearing Research*, 2013.

[5] B. Honikman, "Articulatory settings," in *In Honour of Daniel Jones*, N. S. D. Abercrombie, D.B. Fry, P.A.D. MacCarthy and J. Trim, Eds. London: Longman, 1964, pp. 73–84.

[6] J. Laver, "The Concept of Articulatory Settings: An Historical Survey," *Historiographia Linguistica*, vol. 5, no. 1-2, pp. 1–14, 1978.

[7] V. Ramanarayanan, L. Goldstein, D. Byrd, and S. S. Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging." *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 510–9, Jul. 2013.

[8] J. H. Esling and R. F. Wong, "Voice Quality Settings and the Teaching of Pronunciation," *TESOL Quarterly*, vol. 17, no. 1, pp. 89–95, 1983.

[9] W. Viëtor, *Elemente der Phonetik und Orthoepie des Deutschen, Englischen, und Französischen mit Rücksicht auf die Bedürfnisse der Lehrpraxis*. Heilbronn: Henninger, 1887.

[10] H. Sweet, *A Primer of Phonetics*. Oxford: Clarendon Press, 1892.

[11] T. Diekhoff, *The German Language: Outlines of its Development*. New York, NY: Oxford University Press, 1914.

[12] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production." *The Journal of the Acoustical Society of America*, vol. 115, pp. 1771–1776, 2004.

[13] I. Mennen, J. M. Scobbie, E. de Leeuw, S. Schaeffler, and F. Schaeffler, "Measuring language-specific phonetic settings," *Second Language Research*, vol. 26, no. 1, pp. 13–41, Mar. 2010.

[14] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *The Journal of the Acoustical Society of America*, vol. 120, pp. 1791–1794, 2006.

[15] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images." *IEEE transactions on medical imaging*, vol. 28, pp. 323–338, 2009.

[16] V. Ramanarayanan, E. Bresch, D. Byrd, L. Goldstein, and S. S. Narayanan, "Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation." *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. EL160–5, Nov. 2009.

[17] R. R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, 3rd ed. San Diego, CA: Academic Press, 2012.

[18] D. M. Rom, "A sequentially rejective test procedure based on a modified Bonferroni inequality," *Biometrika*, vol. 77, no. 3, pp. 663–665, Sep. 1990.

[19] R. R. Wilcox and H. J. Keselman, "Modern Robust Data Analysis Methods: Measures of Central Tendency," *Psychological Methods*, vol. 8, no. 3, pp. 254–274, Sep. 2003.

[20] V. Ramanarayanan, A. Lammert, L. Goldstein, and S. Narayanan, "Articulatory Settings Facilitate Mechanically Advantageous Motor Control of Vocal Tract Articulators," in *Proceedings of Interspeech 2013*, Lyon, France, 2013.