Interspeech 2024
1-5 September 2024, Kos, Greece

Analysis of articulatory setting for L1 and L2 English speakers using MRI data

Kevin Huang[1], *Jack Goldberg*[1], *Louis Goldstein*[1], *Shrikanth Narayanan*[1]

[1]University of Southern California, USA

kevinyhu@usc.edu

Abstract

This paper investigates the extent to which the geographical region (country) where a speaker acquired their English language affects the articulatory setting in their speech. To obtain accurate measurements for evaluating articulatory setting, we utilized a large real-time MRI corpus of vocal tract articulation. The corpus was obtained from speakers from a variety of linguistic backgrounds producing continuous English speech. We use an automated pipeline to process and extract articulatory positional information from the MRI video data. This data is used to draw comparisons between English language speakers from the United States and speakers who acquired their English in India, Korea, and China. Analysis of the speaker groups reveals statistically significant articulatory setting posture differences in multiple places of articulation.

Index Terms: Articulatory setting, phonetics, language variation, real-time MRI

1. Introduction

Articulatory setting (AS) is the speech-ready posture taken by speech articulators of a speaker within approximately 200 ms before speaking. AS is generally believed to be consistent across speakers of a given dialect of a language. Despite the relatively sparse existence of rigorous measurements of AS, the idea of AS has persisted in the linguistics literature for well over 50 years. Under the assumption that different articulatory postures are used by speakers of different languages, the concept of AS has been used by instructors as a tool to teach different languages.

It was not until recently that measurement techniques have been able to accurately capture the articulator positions necessary to analyze AS. Previous works have utilized X-ray [1], electromagnetic articulography (EMA) [2, 3], and real-time magnetic resonance imaging (RT-MRI) [4] to capture articulatory data. In particular, RT-MRI offers a full midsagittal view of the vocal tract at high spatial resolution, with more recent developments allowing higher temporal resolution.

This paper aims to leverage recent advances in RT-MRI data collection to investigate AS. Specifically, this study investigates L1 speakers of American English as well as L1 speakers of Indian English. On top of those groups, we investigate L2 speakers of English from China and Korea. We will demonstrate that articulatory settings can differ statistically not only between speakers of two languages but also among speakers of the same language from different linguistic backgrounds.

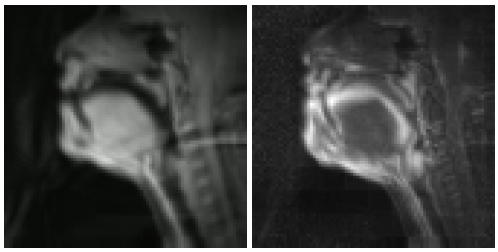

Figure 1: *Mean (left) and standard deviation (right) images of the articulatory setting frames from the MRI data of one speaker.*

2. Background

The term articulatory setting was coined by Beastrice Honikman in 1964 in her article titled "Articulatory Settings" [5]. While Honikman's definition was slightly broader than today's, it laid the groundwork for our current understanding of AS.

In 2004, Gick et al. provided evidence that different languages have statistically different AS [1], while also establishing the modern definition of AS as a single ready stance just before an utterance. They accomplished this by analyzing X-ray data from five English and five French speakers. For each speaker, the average minimum pixel distance between two articulators was measured in aggregate over multiple articulator pairs. Between the French and English speakers, statistically significant differences for many places of articulation were observed.

AS was additionally shown to differ between different accents of the same language. Swiecinski demonstrated this by examining the AS of four Polish speakers of English, where two have strong accents and the other two exhibited near-native English proficiency [2]. AS positions were measured using EMA, and statistical differences were found to support their hypothesis that AS is related to a speaker's accentedness. Wieling et al. demonstrated similar results over 34 speakers of two different dutch dialects via EMA sensors and mixed regression analysis [3].

Multiple theories have been formed as to why different languages and dialects affect AS in different ways. It is unclear, for example, which way causation between the segments of a language and that language's articulatory setting goes. It could be the case that AS is a motor goal in itself, influencing the phonetic structure of the language without changing the specific phonemes. Dart's observation that French coronal stops tend to be produced with more dental articulation than English coronal stops supports this hypothesis, with the differences seemingly

10.21437/Interspeech.2024-2175

caused by differences in the AS position of the tongue [6]. On the other hand, it could be the case that a given AS is merely an emergent property of learning all of the phonetic units in a language, simply an automatic average of some sort over the phones. Ramanarayanan et al. investigated the second hypothesis in more depth through the lens of mechanical advantage, hypothesizing AS offers a mechanically advantageous position to facilitate efficient articulatory control [7]. In an extremely simplified example, a language only containing the consonants [p, b, t, d, s] may have an AS where the tongue tip is curved up near the alveolar ridge and tongue back is relatively far forward, but a language with emphasis on [k, g, x, ɣ] may have an AS with a comparatively low tongue tip and high tongue back.

3. Methods

3.1. Data

An investigation of AS requires articulatory data of speech, in this case, we use data provided by MRI imaging. MRI is used to study articulation in speech due to its ability to obtain comparatively high resolution images of the *entire midsagittal* vocal tract in real-time in a non-invasive manner. We used the publicly-available USC 75-Speaker Speech MRI Database [8] as at the time of writing, this dataset is the largest corpus of MRI speech samples from speakers of varied demographics. Of the 32 half-minute videos recorded for each speaker, we only analyzed the 22 containing continuous speech. All of the speech is in English, however many of the speakers are L2 English speakers from countries outside of the United States. This study focused on four groups: L1 American English Speakers from the USA, L1 Indian English speakers from India, and L2 English speakers from China and Korea. Any participants that did not meet these criteria were not included in the statistical analysis. This resulted in 44 speakers from the USA, 9 speakers from India, 7 speakers from Korea, and 4 speakers from China.

3.2. Preliminary Data Analysis

Prior to applying an automatic pipeline to all speakers, we conducted a preliminary manual analysis over several speakers. For this analysis, we utilized human reviewed transcripts and alignments to determine the location and nature of the inter-speech pauses. While not perfectly representative of AS, inter-speech pauses have been used as a proxy for AS as in [1, 4]. We also determined the locations of irregularities such as resting, swallowing, yawning, and other non-linguistic tasks. Identifying the key articulation characteristics of these non-linguistic tasks with the hand-aligned data allowed us to better remove the undesirable actions in the fully automated pipeline. From this analysis, only frames within 200 ms prior to a word being spoken were taken to be representative of AS as they did not involve breathing, swallowing, or other non-linguistic artifacts. The analysis also showed that any constrictions smaller than 0.1 pixels should be excluded, as such constrictions were consistently markers of non-linguistic tasks while not speaking. These restrictions on frames of interest are also largely mirrored in [4], where only pauses longer than 170 ms were considered, and speech ready frames were taken to be those in a window of 100-200 ms before the start of an utterance.

3.3. Frame Selection

Frames considered to be within the AS posture were selected using a multi-step pipeline. First, text transcriptions were obtained using the whisper-large-v3 [9] ASR model. Half of the videos were readings of passages commonly used in linguistic studies, however automatic transcriptions were still used to account for occasions the passage is not read correctly. These transcriptions were used with the Montreal Forced Aligner to obtain timings for each start and stop timings for each word. As per the prior manual analysis of the inter-speech pauses, only pauses between words longer than 200 ms were considered, and only the frames occurring in a window of 200 ms before words were extracted as a single pause. The number of pauses obtained for each demographic is displayed in Table 1.

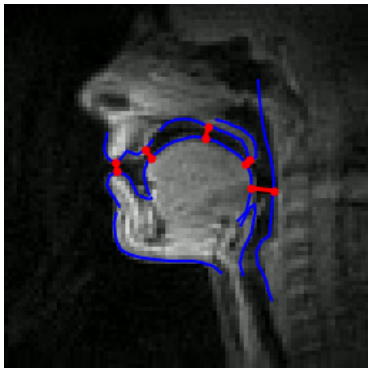

Figure 2: *Image segmentation of one MRI frame with constrictions marked with red segments. From left to right, the constrictions are bilabial, coronal, palatal, velar, and pharyngeal.*

Table 1: *Number of samples per country of origin*

Country	Subjects	Pauses
USA	44	7037
India	9	1640
Korea	7	789
China	4	1096

3.4. Feature Extraction

While articulatory features can be examined in a variety of ways, for this study we decided to examine constriction task variables as in [10]. The constriction task variables chosen and their corresponding articulators are described in Table 2. To obtain these variables in a fully automatic manner for each frame, we applied contour segmentation as described in [11] to obtain lists of points, $\{p_{a,1}, p_{a,2}, ...p_{a,m}\}$, corresponding to the boundary of each articulator a. We obtained more points via linear interpolation between adjacent points, and defined the constriction task variable for constriction g between articulator a_1 and a_2 for frame i as the minimum pairwise distance between the points of each articulator as demonstrated in Equation (1).

$$v_{g,i} = \min_{j,k}(||p_{a_1,j} - p_{a_2,k}||) \quad (1)$$

Figure 2 illustrates an example of the location and width of each constriction as a result of the previous steps. We excluded constriction distances smaller than 0.1 pixels, as we observed constrictions lower than this threshold primarily occurred in irregular articulatory positions such as rest and swallowing.

To account for speaker specific attributes such as head shape and vocal tract geometry, the values were normalized by the mean and standard deviation, $\bar{v}_g$ and σ_{v_g}, of the values occurring in the speech regions (frames corresponding to words

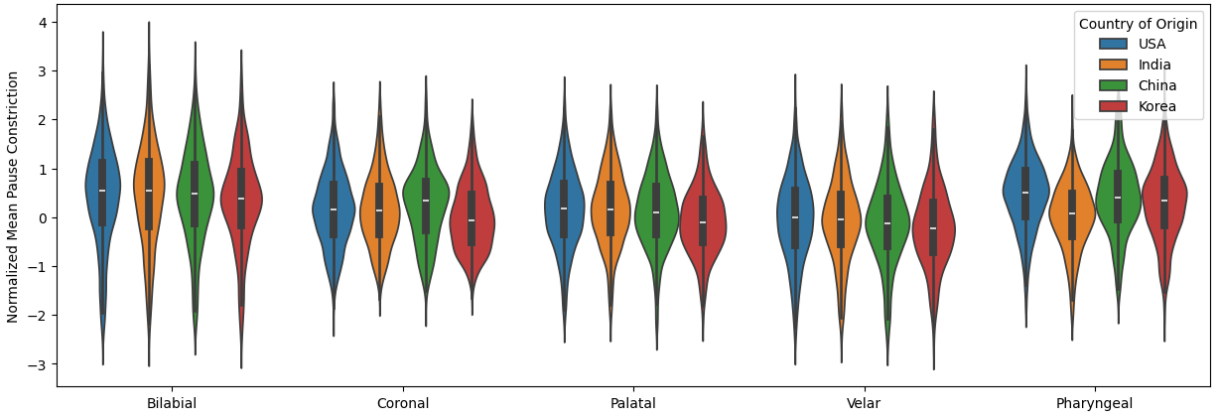

Figure 3: *Violin plots showing the distributions of the normalized constriction $\hat{v}_{g,i}$ for each location of articulation, grouped by the subject's country of English acquisition.*

Table 2: *Articulators corresponding to each constriction.*

Constriction (g)	**Articulator 1** (a_1)	**Articulator 1** (a_2)
Bilabial	Upper Lip	Lower Lip
Coronal	Anterior ¼ of Hard Palate	Tongue
Palatal	Posterior ¾ of Hard Palate	Tongue
Velar	Velum	Tongue
Pharyngeal	Pharynx	Tongue

found by the aligner in the frame selection process).

$$\hat{v}_{g,i} = \frac{v_{g,i} - \bar{v}_g}{\sigma_{v_g}} \tag{2}$$

Without normalization, we observed significantly smaller constriction distances for the Korean speakers of English. For each pause between words, we utilize the mean value of the normalized $\hat{v}_{g,i}$ for frames i occurring within the pause.

4. Results

Table 3 shows the statistics of the normalized constrictions $\hat{v}_{g,i}$ for each place of articulation grouped by country of language acquisition, and the distribution of the data is visualized in Figure 3. To lead our discussion, we ran a multiple linear regression for each of the five constriction locations on which there was data. The data were filtered to remove any data points beyond $\pm$3SD from the mean, which may be a result of faults in the automated pipeline. In each regression, the four countries of speakers were modeled on a single categorical variable with four categories. Age and sex of participants were also added as controls in order to avoid type 1 error. The results of this statistical test are shown in Table 4. While not every country of language acquisition was statistically significant from the grand mean on every place of constriction, the large majority were. Age and sex also turned out to be statistically significant predictors in many, although not all, of the places of constriction.

5. Discussion

Since AS varies across many places of articulation, five articulator locations were included in the study. Each of these locations warranted their own multiple linear regression. This discussion will reference both trends across these five regressions as well as specifics in each predictor variable.

Age was not included in this regression for its predictive power, rather as another effect to assist in eliminating type 1 error. That being said, in every place of articulation barring bilabial, it turned out to have a statistically significant effect on the constriction distance, even if that effect was consistently the smallest effect on the mean, between negative 0.002 and 0.007. A similar story can be told for sex, which was included without knowing whether or not it would be statistically significant. In every place other than bilabial and pharyngeal locations, it did have a significant impact with $p < 0.001$. The generalization made for age magnitude and sign cannot be extended here, as the effect for sex was sometimes positive, sometimes negative, and of various magnitudes.

The primary variable of interest was the country where the language was first acquired. Taken in turn, the constriction data taken from Chinese speakers tended to be closest to the mean, only showing a significant difference from the mean in the pharyngeal and velar locations.

The data taken from the Indian speakers of English was the only country data to be consistently significant in all five places of articulation. The magnitude was also consistently large compared to its peers, while the sign oscillated between positive and negative. The pharynx in particular had a coefficient of -2.44, meaning that the estimated mean pharynx constriction for the Indian group was more than two standard deviations away from the grand mean of all groups. This was a considerably higher deviation than the next highest deviation from the grand mean, and the only coefficient to breach even .5 SD from the grand mean. This leads to the conclusion that the pharyngeal region of Indian English speakers in our dataset has the most divergent overall articulatory setting among the four groups.

The American English speakers showed statistically significant variation from the grand mean in the velar, pharyngeal and palatal place. They also demonstrated a consistently positive coefficient in the places of significance. This can be interpreted as the tongue dorsum being maintained in a lowered posture than what is represented in the grand mean across the four speaker groups. The Korean speakers' data were statistically significantly different from the mean in each case except for the pharyngeal location. In the cases where it was significant, there was also a consistently negative coefficient. This means that the English speakers from Korea had an AS with a higher tongue position and more constricted lips. As the differences are largest in the coronal and palatal regions, less in the velar and non-significant in the pharyngeal region, this suggests that the Korean speakers are maintaining a higher jaw posture

Table 3: *Mean and standard deviations of the normalized constriction $\hat{v}_{g,i}$ for each location of articulation and country of origin.*

Country of Origin	Bilabial	Coronal	Palatal	Velar	Pharyngeal
USA	0.446 ± 1.023	0.158 ± 0.767	0.158 ± 0.805	-0.034 ± 0.868	0.467 ± 0.742
India	0.452 ± 1.053	0.155 ± 0.739	0.165 ± 0.772	-0.073 ± 0.850	0.020 ± 0.693
China	0.436 ± 0.973	0.256 ± 0.751	0.119 ± 0.756	-0.130 ± 0.837	0.401 ± 0.754
Korea	0.337 ± 0.946	0.003 ± 0.675	-0.085 ± 0.692	-0.212 ± 0.817	0.284 ± 0.764

Table 4: *Results of multiple linear regression analysis for various places of articulation over demographic categories.*

		Est.	SE	t-value	p
Bilabial	Intercept	0.485	0.043	11.248	$< .001$
	China	-0.006	0.03	-0.199	0.842
	India	0.067	0.024	2.761	0.006
	Korea	-0.077	0.026	-2.988	0.003
	USA	0.016	0.017	0.942	0.346
	Sex	-0.05	0.022	-2.349	0.019
	Age	-0.002	0.002	-1.55	0.121
$R^2 = 0.002$		Res SE $= 1.021$		$DF = 10346$	$n = 10352$
Coronal	Intercept	0.31	0.032	9.739	$< .001$
	China	0.05	0.022	2.269	0.023
	India	0.072	0.018	4.036	$< .001$
	Korea	-0.114	0.019	-5.977	$< .001$
	USA	-0.008	0.012	-0.673	0.5
	Sex	-0.156	0.016	-9.727	$< .001$
	Age	-0.006	0.001	-5.543	$< .001$
$R^2 = 0.021$		Res SE $= 0.748$		$DF = 10127$	$n = 10133$
Palatal	Intercept	0.206	0.033	6.191	$< .001$
	China	-0.03	0.023	-1.316	0.188
	India	0.137	0.019	7.327	$< .001$
	Korea	-0.152	0.02	-7.705	$< .001$
	USA	0.045	0.013	3.558	$< .001$
	Sex	-0.162	0.017	-9.736	$< .001$
	Age	-0.004	0.001	-3.445	$< .001$
$R^2 = 0.021$		Res SE $= 0.786$		$DF = 10312$	$n = 10318$
Velar	Intercept	0.038	0.036	1.053	0.292
	China	-0.097	0.025	-3.898	$< .001$
	India	0.112	0.02	5.533	$< .001$
	Korea	-0.076	0.021	-3.561	$< .001$
	USA	0.061	0.014	4.431	$< .001$
	Sex	-0.227	0.018	-12.572	$< .001$
	Age	-0.005	0.001	-4.257	$< .001$
$R^2 = 0.026$		Res SE $= 0.849$		$DF = 10227$	$n = 10233$
Pharyngeal	Intercept	0.484	0.032	15.363	$< .001$
	China	0.085	0.022	3.903	$< .001$
	India	-2.44	0.018	-13.799	$< .001$
	Korea	-0.007	0.019	-0.389	0.697
	USA	0.167	0.012	13.841	$< .001$
	Sex	0.011	0.016	0.666	0.505
	Age	-0.007	0.001	-6.337	$< .001$
$R^2 = 0.046$		Res SE $= 0.743$		$DF = 10258$	$n = 10264$

than speakers from other regions.

Addressing the places of articulation themselves, the bilabial regression had extremely high p values comparatively, with only three out of the six coefficients being $p < 0.05$. The coefficients were also extremely low, none higher than 0.07. As such, the lip constriction distance changed the least between the speaker groups. The other locations, all of which involve the tongue, were similar to one another, with more regions than not having statistically significant means, and a wide coefficient values.

While this initial analysis has provided a preliminary glimpse into the role of linguistic background in AS, it is limited in several ways. The primary limiting factor is the lack of cross-linguistic understanding of articulatory setting. There has been little to no work on the AS of languages other than English and French. The corpus that this study utilized was only spoken English, even though English was the L2 of many of the speakers. Data containing recordings of the same speaker speaking multiple language would aid further in the understanding of how L1 and L2 influence AS.

Another limitation of this study was the data-processing pipeline. Being composed of many automated steps, errors may arise in the articulatory measurements from unreliable image segmentation, and poor transcription alignment can lead to incorrectly detected pauses. While an automated pipeline enables the large-scale analysis of data across many speakers required to thoroughly study AS, it is important to scrutinize the errors that can arise in each step. A more specific methodological limitation was that jaw posture was not analyzed. Differences in the position of the jaw might be the sources of the differences across the measured constriction locations.

Finally, we used data from inter-speech pauses, ignoring whether they are grammatical, disfluent, planning-related or other. Work has shown that inter-speech pauses are different for different kinds of inter-speech pauses [4, 12]. What this means for being able to use inter-speech pauses to index articulator setting per se has not been fully addressed in those papers or in this preliminary work.

6. Conclusion

This paper has demonstrated a pipeline to analyze Articulatory Setting (AS) in speech using a large corpus of real-time MRI video data, and used that pipeline to observe how the L2/L1 language background affects AS. There are still many open questions about L2/L1 language interaction in AS as well as dialectal differences within one language's AS. From the relatively small group sizes of participants included in this study, there are strong indications that Standard American English and Indian English have differing articulatory settings.

Finally, there are questions yet to be answered about the relationship between AS and phonology. For example, Whether there is a relationship between a language's AS and its phonological inventory and or phonotactics still remains to be explored with the backing of the both physical and statistical mechanisms we now have. There is also the interaction between perceived accent and AS, to what degree can variance of AS from the language's standard be predicted from the strength of a speaker's accent in said L2 and vice-versa. We hope that this study has contributed in laying the groundwork for a broader investigation of AS and that the field will begin to answer many of the questions above and more in the coming years.

7. References

[1] 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. 220–233, 2004.

[2] R. Swiecinski, *An EMA study of articulatory settings in Polish speakers of English*. Springer Science & Business Media, 2012.

[3] M. Wieling and M. Tiede, "Quantitative identification of dialect-specific articulatory settings," *The Journal of the Acoustical Society of America*, vol. 142, no. 1, pp. 389–394, 2017.

[4] 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. 510–519, 2013.

[5] B. Honikman, *Articulatory settings*. na, 1964.

[6] S. N. Dart, "Comparing French and English coronal consonant articulation," *Journal of Phonetics*, vol. 26, no. 1, pp. 71–94, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447097900607

[7] V. Ramanarayanan, A. Lammert, L. Goldstein, and S. Narayanan, "Are articulatory settings mechanically advantageous for speech motor control?" *PloS one*, vol. 9, no. 8, p. e104168, 2014.

[8] Y. Lim, A. Toutios, Y. Bliesener, Y. Tian, S. G. Lingala, C. Vaz, T. Sorensen, M. Oh, S. Harper, W. Chen *et al.*, "A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images," *Scientific data*, vol. 8, no. 1, p. 187, 2021.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[10] T. Sorensen, A. Toutios, L. Goldstein, and S. Narayanan, "Task-dependence of articulator synergies," *Journal of the Acoustical Society of America*, vol. 145, pp. 1504–1520, 03 2019.

[11] R. Jain, B. Yu, P. Wu, T. Prabhune, and G. Anumanchipalli, "Multimodal segmentation for vocal tract modeling," 2024.

[12] J. Krivokapić, W. Styler, and D. Byrd, "The role of speech planning in the articulation of pauses," *The Journal of the Acoustical Society of America*, vol. 151, no. 1, pp. 402–413, 2022.