INTERSPEECH 2009 BRIGHTON

An Articulatory Analysis of Phonological Transfer Using Real-Time MRI

Joseph Tepperman, Erik Bresch, Yoon-Chul Kim,
Sungbok Lee, Louis Goldstein, and Shrikanth Narayanan

University of Southern California, Los Angeles, USA
{tepperma@,bresch@,yoonckim@,sungbokl@,louisgol@,shri@sipi.}usc.edu

Abstract

Phonological transfer is the influence of a first language on phonological variations made when speaking a second language. With automatic pronunciation assessment applications in mind, this study intends to uncover evidence of phonological transfer in terms of articulation. Real-time MRI videos from three German speakers of English and three native English speakers are compared to uncover the influence of German consonants on close English consonants not found in German. Results show that nonnative speakers demonstrate the effects of L1 transfer through the absence of articulatory contrasts seen in native speakers, while still maintaining minimal articulatory contrasts that are necessary for automatic detection of pronunciation errors, encouraging the further use of articulatory models for speech error characterization and detection.

Index Terms: real-time MRI, nonnative speech, articulation, phonological transfer

1. Introduction

It is widely known that learners of a foreign language will produce speech variants that reflect the phonology of their native language [6]. The systematic substitution, deletion, or insertion of phonemes when speaking a second language (L2), as predicted by the phonological rules of the speaker's first language (L1), is called "phonological transfer". Though not all errors in foreign speech can be attributed to transfer from the speaker's L1, phonological speech errors are exemplified by (but not limited to) transfer in the form of the substitution of a close L1 phoneme for a target L2 phoneme nonexistent in the speaker's L1. Such a substitution may be "close" to the target either in terms of perception (acoustic similarity) or production (articulatory similarity), or both, depending on the speaker's L1 phoneme set.

In our previous work in [9], we found that an articulatory representation of speech could provide additional discriminatory information to a phonetic representation when automatically detecting these kinds of close segment-level errors produced by English learners. Though we had no true articulatory data, we could infer the true articulation from the known phoneme sequence, and acoustic models trained to represent these inferred articulations were not redundant, even when used alongside phoneme models trained on the same acoustics. Our interpretation was that representing these close errors in terms of articulation had more explanatory power than simply conceiving of them as full-on substitutions - the difference between the target and its substitution existed perhaps on only a subset of the articulatory organs that produced the acoustics. For example, a German speaker's common error of substituting /d/ for /ð/ in English demonstrates articulatory contrast only in terms of the tongue tip - the rest of the vocal tract should be identical in both cases, ceteris paribus. This is a partial explanation for the improvement seen when using articulatory models.

However, several details about nonnative articulation remained obscure after those experiments. What was the nature of these substitutions? Did they show the influence of transfer, or were these articulations seemingly unrelated to the speaker's L1? Did their true articulations follow our phoneme-derived expectations, or were they somewhere in between the L2 target and the L1 substitution? The answers to these questions will help us to better model articulation in foreign-accented speech errors, and to determine what pseudo-articulatory models can really capture. Real-time Magnetic Resonance Imaging (MRI) of the vocal tract [7] has recently proven to be a useful tool for anaylzing variation in fricative production [2], emotional speech [5], and articulatory coordination of nasals [4]; we intend to use real-time MRI to shed light on nonnative production.

The main question we want to address in this study is, does nonnative speech show evidence of phonological transfer from the speaker's L1? If so, in what ways, and along what articulatory dimensions? More specifically, our questions of the real-time MRI data are the following:

1. When prompted to produce a target phoneme outside their L1 set, do nonnative speakers employ an articulation indistinguishable from the one they use when prompted to produce the cohort "close" phoneme in their L1 set?

2. Do nonnative speakers demonstrate more variability in their articulation of out-of-L1 targets than for their closest in-L1 counterparts?

If the answer to Question 1 is yes, then that means that articulatory models would be of no help in discriminating close phoneme-level errors, and that the pseudo-articulatory models that worked so well in [8, 9] probably did not represent the true articulation. However, we would expect nonnative articulation of out-of-set phonemes to be similar to close in-set targets at least for some vocal tract organs, otherwise a difference in articulation is not the cause of a listener's perception of a foreign accent. We might expect the answer to Question 2 to be yes, since an increase in articulatory variability for an out-of-set phoneme would reflect a speaker's unfamiliarity with producing it. Either way, the presence of a diversity of variability in production, if seen, will have to be incorporated into articulatory models of speech. In this study we will compare articulatory contrasts between close phonemes for both native English and German speakers in the hopes of illuminating some of these issues.

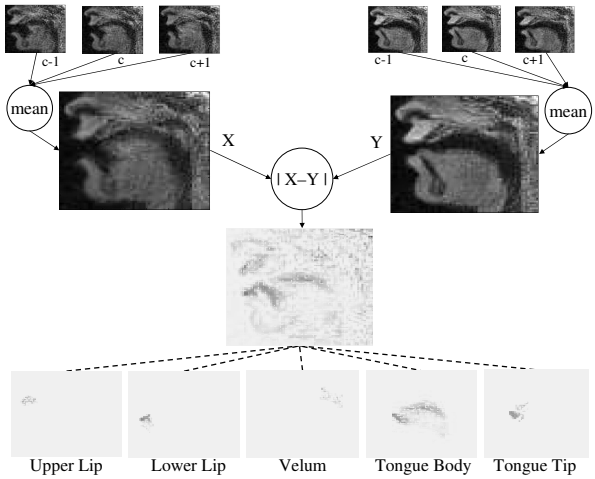

Figure 1: Illustration of finding the pixel-by-pixel difference between tokens of /d/ and /ð/ (X and Y, respectively), including masked versions of the five organs of interest.

2. Corpus

2.1. Speakers and Stimuli

This study used MRI and synchronized audio data from 3 native speakers of German and, for comparison, 3 native speakers of American English. The Germans were advanced learners of English who had been living in Los Angeles for more than 6 months. All but one of them (native English speaker H5) were male. All subjects were asked to read two standard English texts for phonetic elicitation: The Rainbow Passage, and The North Wind and the Sun. Additionally, the German speakers read a phonetically-balanced German translation of each passage. All readings were repeated once, for a total of 31.3 minutes of speech for the English natives and 43.1 minutes for the German natives.

We chose to look only at two well-documented phonological errors made by German speakers of English that are usually explained as L1 transfer: the substitution of /v/ for /w/, and the substitution of /d/ for /ð/. They represent contrasts of different articulatory organs - /v/ and /w/ are expected to differ in the tongue body, lips, and velum, while /d/ and /ð/ should differ only in the tongue tip - and so would yield complementary results. Contexts in which /d/ becomes dental (i.e. before dental consonants) were not included. For each German speaker, we divided all tokens into three categories: phonemes outside their L1 (*OUT*), phonemes present in both languages and elicited in their L2 (*IN-L2*), and phonemes present in both that are elicited in their L1 (*IN-L1*). Though the English speakers had all phonemes in their L1, for simplicity we gave them the same category names. In the stimuli, including repetitions, per speaker there were roughly 35 tokens of /w/, 45 of German /v/, and 30 of English /v/; the stimuli also had roughly 65 tokens each of /ð/ and English /d/, and about 115 tokens of German /d/.

2.2. Imaging and Image Tracking

By "real-time" MRI we mean the generation of MR scans at a sufficiently high frame rate to capture dynamic vocal tract shaping in the midsagittal view of the upper airway during natural speech production [7]. The computational demands of MRI and the scientific demands of speech analysis necessitate a tradeoff in favor of finer temporal resolution at the expense of image resolution - this compromise results in the generation a series of 68 x 68 pixel images reconstructed at a rate of 22.4 frames per second. All MR data used in this study was acquired on a GE Signa 1.5-T scanner using fast gradient echo pulse sequences and a 13-interleaf spiral acquisition technique. With a custom-made multichannel upper airway receiver coil, the images were reconstructed using standard sliding window gridding and inverse Fourier transform techniques.

The contours of the vocal tract articulators were automatically segmented in 2-D space based on an anatomical object model's fit to the data in the spatial frequency domain [1]; the fit of the model to the image was optimized using a hierarchical and anatomically-informed version of gradient descent. Based on this segmentation, we derived polygonal masks enclosed by these contours and representing the regions occupied by each of 5 articulators: the upper lip (UL), lower lip (LL), velum (VE), tongue body (TB), and tongue tip (TT). The boundaries of all but the tongue tip were defined by the segmentation algorithm's object model, and the tongue tip was defined as the polygon enclosed by the leftmost four points of the tongue body contour (out of 11 points used in the object model of the tongue body).

2.3. Audio Processing and Alignment

Simultaneous audio recordings during each MRI scan were collected using a fiber optical microphone, at a sampling rate of 20 kHz. This audio was made useable by way of a noise-canceling filter based on a model of the MRI scanner's gradient noise and pulse sequence [3]. In the absence of phoneme-level transcripts, we used forced alignment of the expected sequence of phonemes to generate segmentation times. Phoneme-level acoustic models were Hidden Markov Models trained on 39-dimensional MFCC feature vectors, with 3 hidden states and 8 Gaussian mixtures per state. The window length was a standard 25 msec and the frame rate was shorter than usual (5 msec) so as not to miss any rapid changes in articulation. Training and alignment of these models were done using an iterative bootstrap procedure like that described in [10].

These automatic segmentations were potentially inaccurate if the speaker paused at an unexpected place while reading the stimuli. In those cases, the alignment would include the pause as part of an abnormally long segmentation for the preceding phoneme. To eliminate these alignment errors, we removed all outlier phonemes with a duration more than 2 standard deviations from the mean for that target phoneme. This eliminated no more than about 10% of the tokens for each speaker.

3. Experimental Methods

3.1. Overview

Because of a wide variety of vocal tract shapes and no standard way to warp one onto another, we restricted these experiments to intra-speaker comparisons between pairs of phonemic MRI tokens. To obtain a representative MRI image of each phoneme's articulation, the token of one phoneme was defined as the mean of the center frame, c, of that phoneme (determined from the automatic segmentation times) and one frame on either side of it, frames $c + 1$ and $c - 1$. When the middle of the segmentation boundaries fell between two MRI frames, we rounded down to the nearest frame, since the characteristic articulatory closures typically occurred more toward the side of the start boundary. Most phonemes examined here (/v/, /w/, /d/, and /ð/) were automatically segmented as being between 2 and

Table 1: Results of two-tailed t-test comparing the mean difference between any two *IN-L2* tokens and the mean difference between any *IN-L2* token and any *OUT* token. Sample means were determined to be equal or unequal (denoted by $=$ and $\neq$) on the 95% level.

		English speakers			*German speakers*		
		J2	*G1*	*H5*	*JS*	*DH*	*CW*
/d/ : /ð/	*overall*	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$
	upper lip	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$
	lower lip	$\neq$	$\neq$	$\neq$	$\neq$	$=$	$=$
	velum	$\neq$	$\neq$	$\neq$	$\neq$	$=$	$\neq$
	tongue body	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$
	tongue tip	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$
/v/ : /w/	*overall*	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$
	upper lip	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$
	lower lip	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$
	velum	$\neq$	$=$	$\neq$	$=$	$\neq$	$=$
	tongue body	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$	$=$
	tongue tip	$\neq$	$\neq$	$\neq$	$\neq$	$=$	$=$

4 MRI frames in length.

Once mean tokens of two productions are obtained, they could be compared by summing the absolute values of their pixel-by-pixel differences. Difference values for each of the organs of interest (UL, LL, VE, TB, and TT) were localized using the polygonal masks derived from the image tracking, as explained in Section 2.2. A mask for one token was defined as the union of the three masks from frames $c-1$, c, and $c+1$. Similarly, to capture all regions of potential difference, the mask for the difference of two tokens was defined as the union of both tokens' masks. Since the sizes of the masks might vary from one token to another and from one frame to another, the sum of all pixel-by-pixel differences for each masked image was normalized by dividing by the number of pixels in the mask. Figure 1 depicts the generation of two tokens, the calculation of their difference frame, and the masked versions of that difference frame for each organ.

3.2. Place of Articulation

The *Place of Articulation* experiments were designed to answer Question 1 posed in Section 1: do nonnative speakers produce their *OUT* tokens with place of articulation identical to their *IN-L2* tokens, and, in the case of the German speakers, to their *IN-L1* tokens? For all speakers separately, we calculated pixel-by-pixel differences between every *IN* token and every other *IN* token, and between every *OUT* token and every *IN* token. If the mean of the differences between all *IN-L2* tokens and *OUT* tokens is equal to the mean of the differences within all pairs of *IN-L2* tokens, then that suggests that the *OUT* tokens are articulated just like the *IN-L2* tokens, and the same goes for the *IN-L1* tokens. This difference in means was assessed statistically using a two-sample t-test. Over all speakers and masks, and for both phoneme contrasts, Tables 1 and 2 show the mean differences that were statistically equal or unequal on the 95% confidence level.

3.3. Articulatory Variability

The *Articulatory Variability* experiments were intended to answer Question 2 from Section 1: do nonnative speakers produce their *OUT* tokens with more variability than their *IN-L2* and *IN-L1* tokens? We measured pixel-by-pixel differences between every pair of *IN-L1* tokens, and similarly for every pair of *IN-L2* tokens and *OUT* tokens. The means of each set's pairwise differences were then compared using a one-tailed t-test. Table 3 indicates for which organs and speakers the pairwise mean of differences for the *OUT* tokens was statistically greater than, less than, or equal to the *IN-L2* mean pairwise difference, on the 95% confidence level. Table 4 displays the same thing, but for *OUT* vs. *IN-L1* sets.

Table 2: Results of two-tailed t-test comparing the mean difference between any two *IN-L1* tokens and the mean difference between any *IN-L1* token and any *OUT* token. Sample means were determined to be equal or unequal (denoted by $=$ and $\neq$) on the 95% level.

	/d/ : /ð/			/v/ : /w/		
	JS	*DH*	*CW*	*JS*	*DH*	*CW*
overall	$\neq$	$=$	$\neq$	$\neq$	$\neq$	$\neq$
upper lip	$\neq$	$=$	$=$	$\neq$	$\neq$	$\neq$
lower lip	$\neq$	$=$	$\neq$	$\neq$	$\neq$	$\neq$
velum	$=$	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$
tongue body	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$
tongue tip	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$	$\neq$

4. Results and Discussion

Table 1 demonstrates that in general, for native English speakers, the mean difference between any *OUT* token and any *IN-L2* token was not equal to the mean difference between any two *IN-L2* tokens, with 95% confidence. This indicates that native speakers use statistically different places of articulation to distinguish /ð/ from /d/, and /w/ from /v/, as we would expect. This finding is encouraging for articulatory modeling because it suggests that these two pairs of close phonemes are potentially separable in articulatory space using statistical models. The results for the German speakers are quite different - we see more of a tendency for the mean difference between *OUT* and *IN-L2* tokens to be equal to the mean difference between any pair of *IN-L2* tokens, indicating that in some sense the *OUT* phonemes are being articulated like a typical *IN-L2* phoneme - this implies phonological transfer from the L2 articulation of the closest L1 phoneme. However, as with the native English speech, the places of articulation did differ on enough dimensions for them to be potentially separable with some set of articulatory models. These equal means were not seen across the board but specifically for the velum and lower lip in the /d/ : /ð/ contrast, and for the velum and tongue in the /v/ : /w/ contrast. The velum and tongue should contrast between tokens of /v/ and /w/, but only the tongue tip of /ð/ should show the transfer from /d/. It is possible that the image segmentation algorithm could not always distinguish the lower lip from the tongue tip for dental consonants, where they would be likely to come into contact (see token Y in Figure 1 for an example).

Similarly, in Table 2 the mean difference between any L2 /ð/ and any L1 /d/ is found in some cases to be equal to the mean difference between any two L1 /d/ tokens, specifically for the lips and velum. This suggests phonological transfer from the closest phoneme not only as produced in the speaker's L2 (according to Table 1) but in their L1 articulation as well. The same was not true for L2 /w/ and L1 /v/ - these were observed to be separable in articulatory space in all dimensions. The implication is that L1 /v/ is articulated significantly differently from L2 /v/, which is closer to /w/ than the L1 version. This may be evidence for L1 attrition for certain consonants but not others, or it may indicate a non-uniform transfer within a phoneme set due to reasons such as markedness [6] - a larger study would be

Table 3: Results of one-tailed t-test of the alternative hypothesis that the mean difference between any two *OUT* tokens is greater than or less than the mean difference between any two *IN-L2* tokens. Inequalities are given on the 95% level.

		English speakers			*German speakers*		
		J2	*G1*	*H5*	*JS*	*DH*	*CW*
/ɖ/ : /ð/	*overall*	>	>	=	=	>	>
	upper lip	>	>	<	=	>	>
	lower lip	<	<	<	<	<	>
	velum	=	>	>	>	>	>
	tongue body	<	=	<	<	<	>
	tongue tip	<	<	<	<	<	>
/ʋ/ : /w/	*overall*	>	>	>	<	>	>
	upper lip	>	>	>	<	>	>
	lower lip	>	>	>	=	=	>
	velum	>	<	=	>	>	<
	tongue body	=	=	>	<	<	<
	tongue tip	<	<	>	=	<	<

necessary to know for sure.

There was a tendency for differences between *OUT* pairs and differences between *IN-L2* pairs to have unequal means with 95% confidence, as Table 3 shows, but this was true for native speakers as well as nonnative ones. Furthermore, the set with the greater mean difference sometimes flipped depending on the articulatory organ, though overall *OUT* seemed to have more pairwise variability than *IN-L2*. The English speakers generally agreed with one another about as much as they agreed with the German speakers, or as much as the Germans agreed among themselves. The main conclusion to gather from Table 3 is that this difference in variability of articulation between close phonemes is present in both native and nonnative speech, but the direction of the difference may be organ- or even speaker-dependent. Whatever the cause of this variation, articulatory models of speech will have to account for it.

Table 4 tells a somewhat different story. For two of the German speakers (JS and DH) the *OUT* set overall showed less mean pairwise difference than the *IN-L1* set, indicating that the L1 articulations were actually more variable than those of the *OUT* tokens - the opposite of the L2 versions. One interpretation is that speakers show more versatility of articulation - and therefore more appropriate variability according to context - in their L1 than in an L2, and out-of-L1 articulations are somewhat static in comparison. As with the results in Tables 1 and 2, this shows that the native German speakers articulate /ʋ/ and /ɖ/ differently depending on whether they are using them in German or English. The third German speaker, CW, showed the opposite effect from the other two in most cases, but more data is needed to determine if CW is an outlier or within the expected range of German speakers. At any rate, speaker-dependent variability is observed here.

Table 4: Results of one-tailed t-test of the alternative hypothesis that the mean difference between any two *OUT* tokens is greater than or less than the mean difference between any two *IN-L1* tokens. Inequalities are given on the 95% level.

	/ɖ/ : /ð/			/ʋ/ : /w/		
	JS	*DH*	*CW*	*JS*	*DH*	*CW*
overall	<	<	>	<	<	>
upper lip	<	<	<	<	<	>
lower lip	<	<	>	<	<	>
velum	<	<	<	<	<	>
tongue body	<	<	>	<	>	<
tongue tip	<	<	>	>	<	<

5. Conclusion

In summary, articulatory evidence of German phonological transfer can be found in German-accented English speech where it isn't seen in native English speech. Native speakers tend to produce their close pairs of contrasting phonemes with more contrast than nonnative speakers do, but both populations produce close phonemes through contrasts in articulation on some level. This validates past studies in pseudo-articulatory modeling of speech [8, 9] that have shown improved discrimination between close phonemes through an articulatory representation, even without any real articulatory data, though it does imply that any mapping from phonemes to expected articulations in nonnative speech does need to account for the possibility of L1 transfer. Pairwise variability between phonemes can be rather phoneme-, speaker-, and organ-dependent, but we do see that, in nonnative English, phonemes not found in German are in many cases produced with more variability than their closest substitutions that are in German, and articulatory models will have to capture this variability. Much work remains in analysis of intra-phoneme articulatory dynamics, and in similar studies with speakers who are native to other languages.

6. References

[1] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Trans. Med. Imaging*, 28(3):323-338, March 2009.

[2] E. Bresch, D. Riggs, L. Goldstein, D. Byrd, S. Lee, and S. Narayanan, "An analysis of vocal tract shaping in English sibilant fricatives using real-time magnetic resonance imaging," in *Proc. of Interspeech*, Brisbane, 2008.

[3] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during real-time magnetic resonance imaging scans," *J. Acoust. Soc. Amer.*, 120(4):1791-1794, Oct. 2006.

[4] D. Byrd, S. Tobin, E. Bresch, and S. Narayanan, "Timing effects of syllable structure and stress on nasals: a real-time MRI examination," *Journal of Phonetics*, 37:97-110, 2009.

[5] S. Lee, E. Bresch, and S. Narayanan, "An exploratory study of emotional speech production using functional data analysis techniques," in *Proc. of 7th International Seminar On Speech Production*, Ubatuba, Brazil, 2006.

[6] R. C. Major, *Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology*. Mahwah: Lawrence Erlbaum, 2001.

[7] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *J. Acoust. Soc. Amer.*, 115:1771-1776, 2004.

[8] M. Richardson, J. Bilmes, and C. Diorio, Hidden-Articulator Markov Models for speech recognition, *Speech Communications*, vol. 41, no. 2, October 2003.

[9] J. Tepperman and S. Narayanan, "Using articulatory representations to detect segmental errors in nonnative pronunciation," in *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):8-22, Jan. 2008.

[10] S. Young et al. *The HTK Book.* [Online]. Available: http://htk.eng.cam.ac.uk/, 2002.