DIFFERENTIATING NORMAL VS MISARTICULATED TONGUE TRAJECTORIES FROM ULTRASOUND FOR FAST AUTOMATIC ARTICULATORY BIOFEEDBACK

Sarah R. Li, Hannah M. Woeste, Sarah Dugan, T. Douglas Mast, Michael A. Riley, Colin Annand, Jack Masterson, Neeraja Mahalingam, Kathryn Eary, Caroline Spencer, and Suzanne Boyce

University of Cincinnati
lisr@mail.uc.edu, woestehm@mail.uc.edu, hamilsm@ucmail.uc.edu, masttd@ucmail.uc.edu, rileym@ucmail.uc.edu, annandct@mail.uc.edu, masterjk@mail.uc.edu, mahalina@mail.uc.edu, earykn@mail.uc.edu, spenceco@mail.uc.edu, boycese@ucmail.uc.edu

ABSTRACT

Ultrasound images of the tongue surface have been increasingly used for immediate visual feedback while a speaker is working to change pronunciation. However, the tongue surface contour is complex and changes rapidly, meaning that speakers find it difficult to compare undesired vs. desired tongue shapes under realistic speaking conditions. For biofeedback in speech therapy, it is important to identify deviation from the desired tongue trajectory in real time. We present results from a preliminary study characterizing differences between misarticulated and normally articulated ("accurate") American English /ɑɹ/ using an efficient, automatic tongue surface tracking method that separately characterizes motion of the blade, dorsum and root. Results from principal component analysis of trajectory data from child speakers show that misarticulated trajectories are distinctly different from accurate trajectories. A statistical model for classifying accurate vs. misarticulated productions is discussed.

Keywords: Ultrasound, Articulation, Feedback, Speech Disorder, Automatic Tracking.

1. INTRODUCTION

Cross-linguistically, rhotic speech sounds are more difficult for children to acquire and more likely than many other sounds to be a feature of speech sound disorders [1]. In the U.S., 1-2% of children reach college age without acquiring the ability to produce /ɹ/ [2]. In addition, many non-native speakers struggle to acquire a native-like pattern of /ɹ/ production. This difficulty is particularly problematic for speakers (or prospective speakers) of American English (AM), because an inability to produce /ɹ/ is perceived as a marker of immaturity [3]. Misarticulated /ɹ/ is generally transcribed as a [w] in syllable onset positions but as a schwa or back vowel [ɑ], [ʊ], or [ɐ] in nucleus or postvocalic rime positions. (Slashes around phonetic symbols indicate the phoneme attempted. Brackets indicate the actual production.)

For children and adults working to acquire a normally articulated production of /ɹ/, ultrasound images of the tongue surface are increasingly being used as immediate visual feedback. However, the tongue surface shape for /ɹ/ is a complex curve that changes in real time during speech, and speakers find it challenging to understand how to coordinate tongue movements over time. Further, there are alternative potential tongue shapes used by typical speakers, categorized as bunched vs. retroflex variants [4,5]. Rapid detection of differences between desired and undesired movements in real time would improve the immediacy and efficiency of visual feedback and potentially make it easier for speakers to learn new productions.

As a first step toward this end, we describe a fast, automatic method of tracking the tongue surface in real time based on ultrasound imaging. All variants of /ɹ/ involve independent motion along different vectors by regions of the tongue corresponding to the blade, dorsum, and root [4,5,6], so our system tracks each separately. The focus of our approach on tongue regions rather than whole contours, and on the classification of trajectories into "accurate" vs. "misarticulated" categories, is somewhat different than seen in alternative approaches to tongue contour tracking [6,7,8]. Our ultimate aim is transformation of tongue motion into gradient real-time feedback under realistic speaking conditions. Because misarticulations can be perceived as /ɑ/, we have focused our preliminary work on /ɑɹ/ productions by children with and without a residual speech sound disorder (RSSD) diagnosis. An important test of our system is whether our data can appropriately distinguish between accurate and misarticulated attempts at /ɹ/ using only time-dependent tongue part displacements.

2. METHODOLOGY

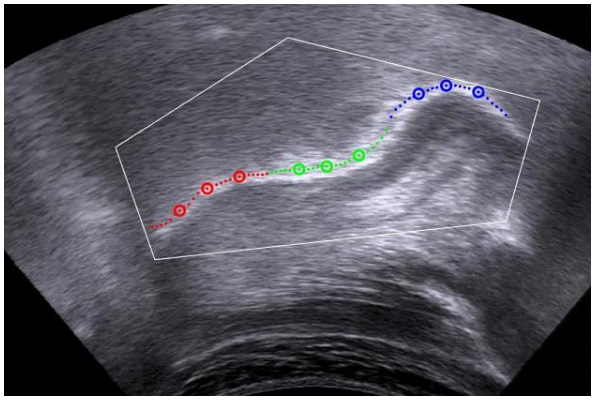
Study participants were 35 speakers of a rhotic American English dialect, aged 8 to 17. Of these, 12

showed consistently accurate production of /ɹ/, while 23 had been diagnosed with RSSD involving consistent inaccurate production. (RSSD children had a history of therapy and some were capable of accurate production on occasion.) Midsagittal tongue images were captured with a Siemens Acuson X300 Premium Edition Diagnostic Ultrasound System with a C6-2 curved linear array transducer (8 cm depth, 4.0 MHz frequency) at 36 frames per second (fps). Images were recorded as digital video at 60 fps with 1024×768 resolution along with audio sampled at 48 kHz; audio and image data were extracted from these video files. Children recorded 15-20 productions of /ɑɹ/ while seated using a custom-built head-stabilizing device. Productions were rated for accuracy by trained listeners on a 10-point continuous scale with 0 considered most "misarticulated" and 10 most "accurate." Three of the children were also recorded producing sustained /ɑ/ and /ɹ/ in a Philips Magnetic Resonance Imaging (MRI) scanner.

For processing purposes, each production of /ɑɹ/ was defined as extending from the acoustic midpoint of /ɑ/ to the end of /ɹ/, defined manually using Praat software. The image frames spanning these two time points were automatically selected via a graphical user interface in MATLAB. Each image frame was cropped to isolate the ultrasound image and a user-defined region of interest (ROI) was drawn around the tongue. A smoothing filter was applied within this region and an initial point on the tongue surface was found by locating the point of maximum brightness within the ROI. The tongue surface was then mapped by first estimating surface points via second-order Taylor series approximations along the anterior and posterior directions, then searching for local brightness maxima within vertical windows centered on each estimated point. This process was repeated in both directions until the maxima fell below a user-defined threshold.

Three reference points each for the blade, dorsum, and root regions of the tongue surface were identified automatically, with each region defined as one-third of the visible horizontal span of the tongue. Time-dependent displacements for each region, defined as the average displacement of the three reference points relative to their positions at the acoustic midpoint of /ɑ/, were computed for the entire production, with positive displacements corresponding to local narrowing of the vocal tract. Displacements were normalized via dividing by the distance between the central reference points from the blade and root regions. For analysis, all productions were linearly interpolated to a duration of 39 frames. Note also that 60-fps acquisition of the 36-fps scanner output is equivalent to nearest-neighbor interpolation of the video data in time. Effects of temporal interpolation on trajectory data were slight relative to the overall uncertainty of our tracking method. An example ultrasound image with reference points for each region is shown in Figure 1. Productions with obviously incorrect tracking were excluded (4% of analyzed productions).

Figure 1: Midsagittal ultrasound image of the tongue at /ɹ/ midpoint for an accurate production, with surface identified by our automatic image processing program.



To characterize differences between movement for accurate vs. misarticulated productions, trajectories of the root, dorsum, and blade regions were analyzed by building "prototypical" trajectories using principal component analysis (PCA).

PCA is a dimension-reduction technique that identifies covariation among a set of input variables (time-dependent trajectories of tongue region displacements, in this case) and maps the original input variables into a new space whose orthogonal dimensions (i.e., principal components) represent the primary directions of variation in the data. Covariation among the input data typically means that the majority of the variance in the original data set can be captured by fewer principal components than the number of original input variables. The space defined by the principal components additionally represents a Principal Component Model (PCM)—that is, a "prototypical" pattern of variation in the data. In this case, our PCMs are a model set of tongue region trajectories identified over a collection of productions from multiple participants. For creation of these models, only principal components explaining greater than 5% of the variance were retained.

These PCMs can then be used to determine whether any individual production is consistent or inconsistent with a prototypical production. We used 85 productions by children with RSSD with low average perceptual ratings (less than 7) and 52 productions by TD children to create respective "misarticulated" and "accurate" PCMs. We then fit a new set of 232 misarticulated and 104 accurate productions (not used in PCM creation) to each model to determine how well the identified prototypes

captured either misarticulated or accurate productions. Although the RSSD children used to create the PCMs all had low perceptual ratings, there were 60 misarticulated productions in the test set from several RSSD children with relatively high perceptual ratings, defined as an average rating of 7 or greater. These are distinguished in Figure 5.

3. RESULTS

Representative trajectories for accurate and misarticulated /ɑɹ/ are shown in Figure 2. It should be noted that typical speakers are known to employ different but perceptually equivalent articulatory variants for /ɹ/ [6,7]. These variants form a continuum but are often loosely categorized into tip-up "retroflex" and dorsum-up "bunched" shapes. Accurate productions generally showed significant displacements of all three tongue regions, regardless of the variant used.

In contrast, RSSD children's misarticulated productions show little total movement from /ɑ/ to /ɹ/, with blade and dorsum regions in particular showing a reduced degree of movement. This pattern fits with previous observations about RSSD articulation of /ɹ/ regardless of phonetic context.

To further elucidate the differences between accurate" and "misarticulated" trajectories, we turned to data from those children who recorded speech with both midsagittal ultrasound for dynamic productions of /ɑɹ/ plus midsagittal magnetic resonance images for sustained /ɑ/ and /ɹ/. Comparisons between the tongue shapes for the two sustained sounds are not necessarily exact representations of the beginning and end points of the dynamic productions seen on ultrasound, but they provide a rough estimate of vocal tract shape at the start and end of the dynamic production. Figure 3 shows examples from one typically-speaking child who used a retroflex tongue configuration for sustained /ɹ/, and one child from the RSSD group who produced misarticulated versions of /ɹ/ sounding like a somewhat rounded [ə].

Figure 2: Trajectories from acoustic midpoint of /ɑ/ to end of /ɹ/ for an "accurate" production of /ɑɹ/ on the left and a "misarticulated" production of /ɑɹ/ on the right. The time axis indicates the frame number after interpolation of all productions to the same duration.

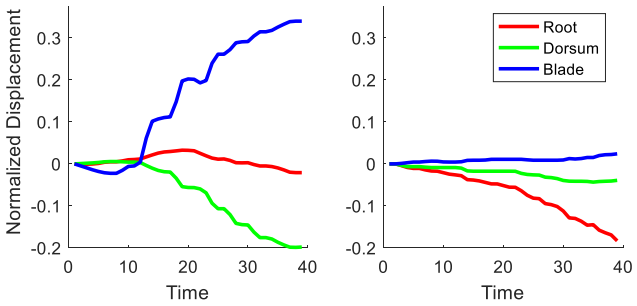


Figure 3 also shows an example from an RSSD child who learned to produce an "accurate" /ɹ/ under sustained conditions in the MR session, but was unable to do so under previously recorded dynamic ultrasound conditions. This child's MR images are included in Figure 3 to illustrate a bunched configuration for "accurate" /ɹ/. His ultrasound images were rated as "misarticulated" and treated as such in analyses. The similarity of the /ɹ/ production to /ɑ/ for the RSSD misarticulated production in Figure 3 is consistent with the observation of relatively smaller tongue displacements for children with RSSD, as also seen in Figures 2 and 4.

Figure 4 shows tongue trajectories from the TD and RSSD groups (solid lines; TD top row, RSSD bottom row) together with predicted trajectories based on the PCM fit for each group (dotted lines; TD PCM left column, RSSD PCM right column). These data illustrate that trajectories of the TD group can be accurately mapped by the model made from a smaller set of TD data, and that RSSD trajectories can be accurately mapped by the model from a smaller set of RSSD data. The difference between the predicted and actual trajectories is much greater when the

Figure 3: Midsagittal MR images from sustained production of /ɑ/ (left) and /ɹ / (right). The top 2 images are from an RSSD child whose /ɹ/ sounded like /ə/. The middle 2 images are from an RSSD child who produced an accurate (bunched) /ɹ/ after weeks of instruction with ultrasound feedback. The bottom 2 images are from a typically speaking child with accurate (retroflex) /ɹ/.

production is fit to the PCM of the opposite group (TD production fit to the RSSD PCM, and vice versa). In particular, the RSSD PCM is unable to capture the rapid changes in blade and dorsum displacement typical of TD productions. This discrepancy is indicated by the sharp curves of the TD trajectory shown in Figure 4: the prediction by the RSSD PCM has slopes less steep than both the observed trajectory and the trajectory predicted by the TD PCM. In addition, due to the trend of lower overall displacements for RSSD productions, the magnitude of the difference between predicted and observed trajectories is smaller for RSSD than TD productions.

These trends are also apparent in the residuals, defined as the mean Euclidean distance between the observed and predicted trajectories. Figure 5 shows that the residuals from fitting individual productions to the PCMs exhibit group-associated clustering. The majority of RSSD productions showed small residuals when fit to the TD PCM and even smaller residuals when fit to RSSD PCM, while the residuals for TD productions are much greater when fit to the RSSD PCM. Note that many of the RSSD productions that overlap the TD cluster are from speakers with high average perceptual ratings, indicating a range of similarity to TD accurate productions. A decision-tree classification model fit to 70% of the data shown in Figure 5 and tested on the remaining 30% has a misclassification rate of 13% (12% for productions from the RSSD group and 16% for the TD group). When productions from RSSD speakers with high perceptual ratings are removed, this drops to a misclassification rate of 7.3% (5.2% for productions from the RSSD group and 12.5% for the TD group).

The PCA results are also consistent with the observation that the trajectories of TD speakers feature distinct patterns of motion associated with the different articulatory /ɹ/ variants in Figure 3. On the other hand, the trajectories of RSSD misarticulations are more varied, both within and across speakers, and do not feature general patterns other than the trend of lower displacements. Figure 5 suggests that PCM residuals emphasize the difference between TD and RSSD speakers despite these variations.

4. CONCLUSION

Rapid movements of the tongue can be difficult for users of ultrasound biofeedback to identify and interpret. The results of this study suggest that characteristic movement trajectories of regions of the tongue extracted from ultrasound are significantly different for misarticulated and accurate productions of /ɹ/ in the context of a closely related sound /ɑ/, and that data of this nature can be collected and analyzed quickly enough to provide definitive feedback. In particular, the movement of the tongue blade appears to be a differentiating factor between accurate and misarticulated productions of /ɑɹ/. We conclude that real-time identification of misarticulated vs. accurate production is viable and holds promise for improving the efficacy of ultrasound feedback in speech therapy. Future steps toward this goal include determination of the most predictive combinations of tongue region movements and extension of this tracking approach to a wider range of phonetic contexts.

Figure 4: Selected trajectories for the TD group (top) and RSSD group (bottom). The dotted lines show trajectories predicted from PCMs fitted to the TD group (left) and RSSD group (right), while solid lines are observed trajectories. The vertical scale of the bottom row (RSSD productions) has been expanded to show detail. Displacements shown are normalized.

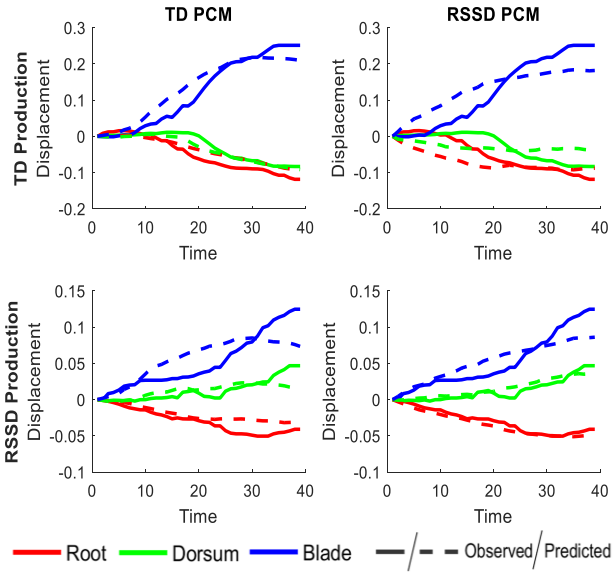
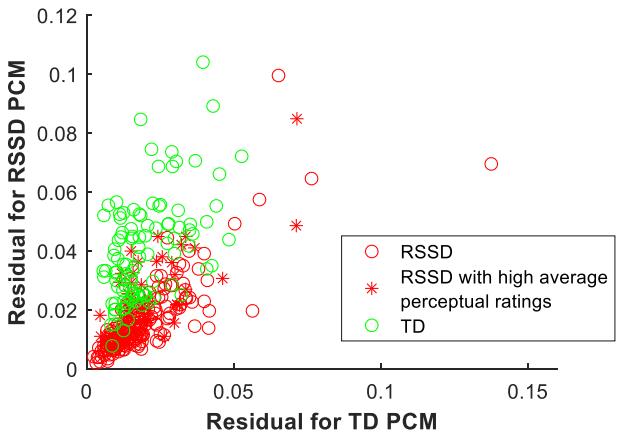


Figure 5: Residuals from the PCM models. Each point represents the mean Euclidean distance between observed and predicted trajectories for a single production.



5. ACKNOWLEDGEMENTS

This work was funded by University of Cincinnati Creating our Third Century Funding Support and by NIH/NIDCD grants 1R01DC013668-01A1 and 1R01DC017301-01. We gratefully acknowledge the contributions of Maurice Lamb and all the speakers who participated in this study.

6. REFERENCES

[1] Boyce, S. Hamilton, S. Rivera Campos, A. 2016. Acquiring rhoticity across languages: An ultrasound study of differentiating tongue movements. *Clin. Linguist. Phon.,* 30, 174-201.

[2] Culton, G. L. 1986. Speech disorders among college freshmen: a 13-year survey. *J. Speech Hear. Disord.*, 51(1), 3-7.

[3] Hitchcock, E. R, Harel, D., McAllister Byun, T.M. 2015. Social, emotional, and academic impact of residual speech errors in school-aged children: a survey study. *Semin Speech Lang*., 36(4), 283-294.

[4] Delattre, P. C., & Freeman, D. C. 1968. A dialect study of American r's by x-ray motion picture. *Linguistics*, 44, 29-68.

[5] Zhou, X., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C., & Choe, A. 2008. A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. *J. Acoust. Soc. Am.*, 123, 4466-4481.

[6] Zharkova, N., Gibbon, F. E., Hardcastle, W. J. 2015. Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation. *Clin. Ling. Phon.*, 29, 249-265.

[7] Tiede, M. K. 2018. GetContours. Version: 1.3. https://github.com/mktiede/GetContours.

[8] Akgul, Y. S., Kambhamettu, C., Stone, M. L. 1999. Automatic extraction and tracking of the tongue contours. *IEEE Trans. Med. Imaging*, 18, 1035-1045.