HHS Public Access

Author manuscript

Clin Linguist Phon. Author manuscript; available in PMC 2024 February 01.

Published in final edited form as:

Clin Linguist Phon. 2023 February 01; 37(2): 196–222. doi:10.1080/02699206.2022.2039777.

Classification of accurate and misarticulated /ɑr/ for ultrasound biofeedback using tongue part displacement trajectories

Sarah R. Li[a,*], **Sarah Dugan**[b,c], **Jack Masterson**[a], **Hannah Hudepohl**[a], **Colin Annand**[d], **Caroline Spencer**[c], **Renee Seward**[e], **Michael A. Riley**[b], **Suzanne Boyce**[c], **T. Douglas Mast**[a]

[a]Biomedical Engineering, University of Cincinnati, Cincinnati, United States

[b]Rehabilitation, Exercise, and Nutrition Sciences, University of Cincinnati, Cincinnati, United States

[c]Communication Sciences and Disorders, University of Cincinnati, Cincinnati, United States

[d]The Complexity Group, Department of Psychology, University of Cincinnati, Cincinnati, Ohio, USA

[e]Design, University of Cincinnati, Cincinnati, Ohio

Abstract

Ultrasound biofeedback therapy (UBT), which incorporates real-time imaging of tongue articulation, has demonstrated generally positive speech remediation outcomes for individuals with residual speech sound disorder (RSSD). However, UBT requires high attentional demands and may therefore benefit from a simplified display of articulation targets that are easily interpretable and can be compared to real-time articulation. Identifying such targets requires automatic quantification and analysis of movement features relevant to accurate speech production. Our image-analysis program TonguePART automatically quantifies tongue movement as tongue part displacement trajectories from midsagittal ultrasound videos of the tongue, with real-time capability. The present study uses such displacement trajectories to compare accurate and misarticulated American-English rhotic /ɑr/ productions from 40 children, with degree of accuracy determined by auditory perceptual ratings. To identify relevant features of accurate articulation, support vector machine (SVM) classifiers were trained and evaluated on several candidate data representations. Classification accuracy was up to 85%, indicating that quantification of tongue part displacement trajectories captured tongue articulation characteristics that distinguish accurate from misarticulated production of /ɑr/. Regression models for perceptual ratings were also compared. The simplest data representation that retained high predictive ability, demonstrated by high classification accuracy and strong correlation between observed and predicted ratings, was displacements at the midpoint of /r/ relative to /ɑ/ for the tongue dorsum and blade. This indicates that movements of the dorsum and blade are especially relevant to accurate production of /r/, suggesting that a predictive parameter and biofeedback target based on this data representation may be usable for simplified UBT.

* lisr@mail.uc.edu .

Disclosure of interest:
The authors report no declarations of interest.

Keywords

ultrasound; tongue; articulation errors; rhotics; child speech

Introduction

Ultrasound biofeedback therapy

Residual speech sound disorder (RSSD) is characterized as the inability to produce all the speech sounds of a language accurately past 8 or 9 years of age (Shriberg et al., 1994). Individuals with RSSD experience an elevated risk for social and academic challenges (Hitchcock et al., 2015). Treatment for residual speech sound errors can be challenging, and traditional methods using auditory feedback often elicit only limited responses. For this reason, clinicians have explored a number of methods that provide visual feedback about articulation, such as electropalatography (Gibbon & Lee, 2015). Ultrasound imaging is another such method that has particularly promising capabilities, deriving from its ability to non-invasively image in real time the midsagittal tongue surface from much of the root to the blade and sometimes the tip (Preston et al., 2017; Stone, 2005). The real-time imaging capability of ultrasound can be used to deliver speech therapy known as ultrasound biofeedback therapy (UBT), which helps people with RSSD by enabling identification of accurate vs inaccurate articulation models and providing continuous information about articulatory movements of the tongue. UBT of this nature generally has positive outcomes (Sugden et al., 2019).

However, not all speakers respond to UBT, and progress using UBT may take many sessions. One possible reason is that UBT is associated with a high degree of attentional and cognitive load. The biofeedback aspect of UBT involves directing the speaker's attention to the details of rapidly changing, grainy images while ignoring image artifacts. Both clinician and speaker must retain key details of the rapidly changing tongue shapes in memory and identify differences between accurate and inaccurate tongue motion before these memories fade. This interpretation of ultrasound images of the tongue then must be translated into new patterns of tongue movement.

Simplified biofeedback in motor learning

One potential way to overcome those challenges and improve the effectiveness of UBT is to use a simpler form of display for providing feedback regarding production accuracy. This is a common approach in non-speech motor control studies, in which complex movement coordination patterns can be more easily achieved by using simplified visual feedback to represent movement outcomes (e.g. Mechsner et al., 2001). In this approach, the task of movement control is presented to the participant as a task of producing a certain perceptual outcome in the feedback display. For example, in a study by Faugloire et al. (2005), subjects learned to produce specified rhythmic rotations of the lower leg and trunk by watching a real-time feedback display showing a single trace derived from the relative phase of their ankle and hip oscillations. By comparing their own trace to a target trace, subjects using this simplified biofeedback learned to reproduce the specified ankle-hip coordination patterns faster than control subjects. The same approach resulted in long-lasting balance

improvements for stroke patients who performed this task as a balance intervention (Varoqui et al., 2011).

Simplifying feedback in that way also elicits an external focus of attention, known to enhance motor learning by engaging implicit and automatic control processes instead of effortful and conscious motor control processes that are often more error-prone (Wulf & Prinz, 2001; Mechsner, 2004; Wulf et al., 2010; Wulf, 2013). It has been suggested that standard UBT also has the effect of externalizing the focus of attention (Byun & Hitchcock, 2012). However, in most UBT implementations, the speaker responds to explicit instructions about controlling the tongue while watching the ultrasound image, rather than fully directing their attention to an external biofeedback outcome (Bernhardt et al., 2005; Cleland et al., 2018). Thus, even in UBT, speakers are asked to focus on their articulators and their location in vocal tract space. The advantage of providing a simpler feedback display is that it may promote an external focus of attention while reducing cognitive complexity by removing the need to monitor changes across multiple parts of the tongue surface and to decide whether a particular visual pattern of tongue motion represents an accurate production.

Requirements for simplified ultrasound biofeedback

As in the non-speech example above, the ideal is to reduce the coordination patterns of tongue movement to a single, computationally simple parameter that represents production accuracy. Crucially, this method must also be able to distinguish the articulatory features of perceptually accurate speech from misarticulations and to do so in a way that is computationally viable in real time. Identifying parameters that meet these criteria is a necessary first step toward the eventual development of simplified UBT approaches that can maximize the effectiveness of UBT for remediating speech deficits. Accordingly, in the present study we used ultrasound to quantify tongue movements and evaluated the potential for a single, relatively simple tongue movement parameter to discriminate perceptually accurate from inaccurate productions. We focused on productions of American English /r/ (denoted as /ɹ/ in the International Phonetic Alphabet, IPA) because speech sound errors of /r/ make up the lion’s share of errors in RSSD children. These errors are particularly resistant to remediation (Shriberg et al., 1994; Ruscello, 1995), indicating that a method for simplifying tongue movement representations has a substantial requirement to address /r/ errors, which necessitates measuring articulatory features relevant to /r/.

Most scientific investigations of differences between accurate and misarticulated tongue movement have focused on methods evaluating the whole tongue, such as characterizing the overall shape of the tongue contour (Klein et al., 2013; Preston et al., 2019), calculating a global displacement measurement (Bressmann et al., 2016) or developing machine learning methods that use entire B-mode ultrasound images (Ribeiro et al., 2021). This consideration of the whole tongue contour reflects the fact that there are no anatomical distinctions in the structure of the tongue muscle per se, or specific reference points along the vocal tract that can define parts of the tongue (Stone et al., 2004); instead, the parts as described previously (e.g. tongue root and blade) are defined only by how they function for a particular sound or group of sounds. Rather than considering the whole tongue, an alternative method for describing tongue movement is to partition the tongue according to its functionally separate

parts. This method is consistent with the concept of functional segments of the tongue that comprise independent articulators (Green & Wang, 2003; Nguyen et al., 1996; Stone et al., 2004) and is particularly relevant for American English /r/ production.

The many variant tongue shapes for typical /r/ (ranging from retroflex to bunched) all require both an anterior and posterior constriction formed during differentiated, often simultaneous movement of at least two functionally independent parts of the tongue (Adler-Bock et al., 2007; Alwan et al., 1997; Zhang et al., 2005; Campbell et al., 2010; Guenther et al., 1999; Espy-Wilson et al., 2000;Westbury et al., 1998; Dugan et al., 2019a). These tongue parts are the tongue tip/blade or dorsum (for retroflex or bunched /r/, respectively) for the anterior constriction and the tongue root for the posterior constriction (Delattre & Freeman, 1968; Boyce, 2015; Zhou et al., 2008). In contrast, misarticulations produced by RSSD speakers often show simpler tongue shapes (Klein et al., 2013; Preston et al., 2019), implying less differentiated movement patterns. For example, the most common error tongue shapes show a single vocal tract constriction by the tongue dorsum (Boyce et al., 2011) with no sign of tongue root retraction (Klein et al., 2013; Boyce, 2015), resulting in errors perceived as /w/, back vowels, or schwa (Chung et al., 2019).

Accordingly, for a simplified UBT parameter to reliably distinguish between accurate and misarticulated tongue motion for /r/, a minimal requirement is to capture the difference in tongue kinematics for constriction at one vs two vocal tract locations. To be useful for UBT, another requirement is that any such parameter can be efficiently computed from real-time ultrasound image data, so it can be presented to the speaker concurrently with a given production.

Objectives

The goal of this study was to undertake the first steps toward creation of a simplified UBT display by evaluating the predictive ability of tongue movements measured by ultrasound imaging to determine production accuracy for /r/. To address the above requirements identified for simplified UBT, we developed an automatic, computationally simple approach for measurement that partitions the imaged tongue surface contour into parts roughly corresponding to the anatomical root, dorsum and blade, and that quantifies tongue movement as displacement trajectories of those three tongue parts, representing tongue movement as only three displacement values at each frame of a production. Specifically, the automatic separation of the tongue surface contour into three equal-length tongue parts broadly approximates regional functionality of the tongue, addressing the requirement for capturing independent constrictions of different tongue parts in real time. This form of quantification has not generally been done for ultrasound, because of the difficulties of tracking tongue movement with high fidelity and translating a continuous tongue surface contour into quantities for calculation in real time.

To evaluate the predictive ability of these tongue part displacement trajectory measurements, we classified accurate vs misarticulated productions using several data representations of these measurements; we compared classification accuracies in addition to predictions of perceptual ratings using regression models. Of particular interest was the potential to define a simple predictive parameter, based on tongue part movement data, that would

maintain high classification accuracy while accounting for the range of accurate /r/ tongue configurations used by speakers of American English. We focused on the commonly misarticulated postvocalic /ɑr/ (‘are’) context. This choice was based on the fact that tongue shapes for /ɑ/ particularly contrast with both retroflex and bunched tongue shapes and is a starting point for evaluating simple parameters that will distinguish accurate /r/ in different vowel contexts. In a preliminary study (Dugan et al., 2019a), we found that tongue part displacement trajectories were sufficient for distinguishing articulatory movements for /ɑ/ followed by all variants of accurate /r/ in typical vs disordered speech from four child speakers.

In the present study, we analysed tongue motion data measured from a larger population of typically developing (TD) children and children with residual speech sound disorder (RSSD), using our program TonguePART (Dugan et al., 2019a) to calculate tongue part displacement trajectories from ultrasound images of /ɑr/ (‘are’). Production accuracy was determined via auditory perceptual ratings from trained clinicians. While testing several candidate representations of the tongue part displacement data, we used the support vector machine (SVM) machine learning method to classify accurate and misarticulated tongue movements for /ɑr/ and we used support vector regression models to evaluate the ability to predict continuous perceptual ratings. A potential predictive parameter based on these data representations, termed δ, was selected for analysis as a simple parameter potentially suitable to drive a biofeedback display. To compare capability of δ for predicting perceptually rated accuracy of /ɑr/ productions, linear regression related δ to observed perceptual ratings and was compared with the support vector regression for the candidate data representations tested. By evaluating the ability to predict production accuracy with these classification and regression models, we demonstrate the utility of tongue part displacement trajectory measurements, and in particular the parameter δ, as a basis for simplified UBT.

Methods

Participants

Children aged 8–17 years old with TD speech and with RSSD were recruited from the University of Cincinnati clinic and community. The demographic characteristics of the two groups are listed in Table 1. All subjects were native speakers of a rhotic American English dialect.[1] All children in our study had typical orofacial development (absence of cleft lip and palate). All participants had a parent/guardian provide informed consent, under protocols approved by the University of Cincinnati Institutional Review Board (#2013–0985, #2015–4209, #2018–6390). Further constraints (e.g. matching speakers by sex across groups) were not applied because present data suggest that articulatory patterns for /r/ show significant individual variation but are not predictable by gender or dialect (Boyce et al., 2015). The aim was to recruit enough speakers to illustrate the range of individual articulatory patterns.

[1]Criterion for native speakers was exposure to an American English environment from 2 years of age onward. A few participants also had exposure to other languages in their households (including Japanese, Arabic, or Mandarin Chinese) but were primarily in English dominant environments.

The data used in this study represent speaker sets of recorded /ɑr/ productions from 41 sessions. Of these, one speaker contributed data in two separate sessions 4 months apart; tongue movement patterns reflected change in the course of therapy and were notably different between sessions. The remaining sessions were produced by 39 separate speakers.

Data collection

Ultrasound data collection followed procedures used in a previously published paper (Dugan et al., 2019a). A custom head stabilizer was used to hold the ultrasound transducer in a fixed position and to keep the speaker's head position aligned via contact points for the forehead and chin (see Figure 1A and 1B). Midsagittal ultrasound images were recorded from a Siemens Acuson X300 Premium Edition diagnostic ultrasound system with a C6–2 curved linear array transducer, image depth of 8 cm, centre frequency of 4.0 MHz and frame rate of 36 fps. These images were collected as videos at 60 fps, with nearest-neighbour temporal interpolation. Audio was recorded concurrently at 48 kHz with a cardioid condenser USB microphone (Audio-Technica ATR2500-USB). Speakers were instructed to produce 10–22 /ɑr/ ('are') productions in isolation, each repeated after the clinician who recorded the data. Variation in production numbers was due to using slightly different study protocols as well as some environmental disruptions (such as the need for more repetitions when noise was heard from the adjacent hallway).

Tongue part displacement trajectory calculations

Tongue movement in these productions was quantified as tongue part displacement trajectories by TonguePART, an automatic program that tracks the midsagittal tongue surface in ultrasound images with real-time capability and minimal user input, as described previously (Dugan et al., 2019a) and summarized as follows. The basis for this tracking is the large amplitude of ultrasound echoes from the interface between the tongue and air, due to the large acoustic impedance mismatch between the two media. This results in a bright contour at the approximate location of the tongue-air interface, which for simplicity we refer to as the tongue surface.

For any particular series of productions from the same speaker, the TonguePART user provides one spatial calibration point near the tongue surface in the initial ultrasound frame and two threshold values associated with brightness at the ends of the tongue surface. TonguePART performs spatial low-pass filtering to smooth speckle within each frame and then identifies the tongue surface based on brightness maxima within vertical search windows. For initial frames of each production, locations for these search windows are initialized near the user-defined calibration point, progress in the anterior and posterior directions based on Taylor finite difference estimations of brightness maxima positions, and end depending on the user-defined brightness threshold values for each end of the tongue.

The tongue surface contour, determined as the brightness maxima positions within these search windows, is then automatically divided into three regions, with the horizontal span of each region equal to one third of the horizontal span of the tracked contour. The three regions are designated in this analysis as the tongue root, dorsum and blade. Points at 1/3, 1/2 and 2/3 (with regards to the horizontal span) of each region are selected as reference

points. This process of identifying brightness maxima is then repeated on subsequent frames of the production, with search windows instead initialized from the tongue contour of the previous frame analysed. Note that this procedure means that horizontal spans of the tracked contours can differ slightly for each frame, elongating or shortening depending on the shape and extent of the detected tongue surface. Additionally, this method of partitioning the surface tracked by TonguePART means that tongue parts specified may not refer to the anatomical tongue parts defined by articulatory function. For example, the anatomical tongue root is often obscured by the shadow of the hyoid bone; in this case, the tracked surface would not include the shadowed portion of the tongue root, and what TonguePART identifies as the tongue root for any one frame is likely to include the posterior portion of the anatomical tongue dorsum. In the following discussion, unless otherwise indicated, references to root, dorsum and blade as well as tongue part displacements refer to the regions automatically assigned by TonguePART.

For each frame analysed, displacement measurements for each tongue region are calculated as the average of distances between the reference points of the analysed frame and the initial frame of the production. For the dorsum and blade, these distances are measured in the vertical direction with positive displacement corresponding to upward movement, locally constricting the vocal tract. For the root, displacements are measured along a diagonal at 45° to characterize vocal tract restriction in the pharyngeal region, with positive displacement corresponding to movement upward and back. To normalise across speakers with different tongue lengths, all displacement values are divided by a reference distance approximately proportional to the tongue length. This reference distance is defined for each production as the distance between middle reference points of the root and blade regions for the initial frame, which corresponds to the acoustic midpoint of /ɑ/ (Figure 1c). This normalisation accounts for differences in tongue size across the range of speaker ages. Displacements of each region over time are regarded as displacement trajectories.

Productions of /ɑr/ were tracked and displacement trajectories were calculated from video frames corresponding temporally to the acoustic midpoint of /ɑ/ through the end of /r/ in each production. Overall, these displacement trajectories represent time-dependent changes in vocal tract constriction due to movement of the root, dorsum and blade relative to the initial tongue position. Positive displacement values in these trajectories represent narrowing of the vocal tract, while negative values represent widening.

Reliability of tracking

Due to inconsistencies in the brightness of the tongue-air interface as well as bright reflections near the tongue surface in the image (e.g. reflections from tongue musculature or from parasagittal slices adjacent to the imaged plane), tracking errors can occur. False detection of the tongue surface can cause extreme changes in frame-to-frame displacement, compared to the typically smooth movement of tongue parts over time. As a result, productions were automatically excluded if the mean-square difference of frame-to-frame normalised displacement values exceeded a pre-determined threshold of 0.001 (Dugan et al., 2019a). This automatic check removed around 11.4% of the recorded productions. An additional visual check of the tracking was performed to identify other obvious tracking

errors, resulting in the exclusion of about 7.5% of the productions that passed the automatic check. The resulting dataset included 567 productions, excluding about 18% of all recorded productions. Not all speakers were equally well tracked, with tracking errors largely due to poor imaging quality from speaker-specific anatomy (e.g. more fatty tissue, less moisture in the mouth, less smooth tongue surface (Stone, 2005)) and possible minor movement within the head stabilizer. Thus, the percentage of productions excluded from each speaker varied, with 60% of speakers having fewer than 18% of their productions excluded and 10% of speakers having more than 50% of their productions excluded. For each of these speakers, excluded productions and included productions appeared to have similar tongue movement patterns.

Perceptual ratings

Children with RSSD in this study had almost universally received previous therapy for their disorder, and a varying level of accuracy is typical. In the clinical context, it is typical to categorize the outcome of a speech task as a ‘correct’—i.e. ‘accurate’—attempt and an ‘incorrect’—i.e. ‘inaccurate’ or misarticulated—attempt. This study utilized this clinical meaning of the terms ‘correct’ and ‘incorrect’ for perceptual ratings of production accuracy. Previous research has shown that visual analogue scaling (VAS) ratings of production correctness correlate well with the percentage of binary ‘correct’ vs ‘incorrect’ judgments as well as to more continuous acoustic measures of rhoticity (Dugan et al., 2019b; McAllister Byun et al., 2015). Accordingly, we conducted a separate study of trained clinicians’ perceptions using VAS to derive a continuous measure of perceptual accuracy (Munson et al., 2012). However, there can be considerable interrater variation in these ratings, especially for productions with intermediate characteristics (in the middle of the continuum between accurate and misarticulated).

In this study, each production was rated by three clinically trained listeners from the University of Cincinnati Communication Sciences and Disorders Department. A total of 40 listeners participated in the study. Demographically, these raters were all non-Hispanic White, female speakers of Midwestern American English aged in their 30s and 40s. For the collection of ratings, groups of around 100 productions were presented to different groups of three raters, so that the three raters were not the same for all productions. The raters indicated their ratings anywhere on an unmarked line with only the endpoints denoted as ‘incorrect’ and ‘correct’; for analysis, locations on this line were mapped onto a continuous scale from 0 for the misarticulated endpoint and 10 for the accurate endpoint. To evaluate interrater agreement, intraclass correlation coefficient (ICC) values were calculated for these ratings. ICC values can range from 0 to 1 and indicate different reliability levels of collected ratings (Koo & Li, 2016). These ICC estimates and their 95% confidence intervals were calculated using R version 4.1.0 and the package *irr* (Gamer et al., 2019) with a mean-rating, absolute-agreement, one-way random-effects model.

For each production, the observed perceptual rating was defined as the average of three listener ratings. These observed perceptual ratings for all productions were used to determine class labels for training and testing SVM classifiers. If a production had an average perceptual rating greater than or equal to the perceptual rating threshold being

evaluated, the production was considered accurate; otherwise, the production was considered misarticulated. The rating threshold separating the accurate and misarticulated classes was determined by evaluating classification accuracies over the range of possible thresholds. The observed perceptual ratings were also used to test regression models for each classifier, as described below.

Data representations

For the purpose of identifying the most relevant features of tongue part displacement trajectories, we tested several candidate data representations. The aim was to find the representation that had the greatest efficiency for determining the accuracy of productions and met the previously described requirements for future use in simplified visual feedback for UBT. Because accurate production of /r/ requires two constrictions within the vocal tract, data representations based on single tongue parts were not considered. Displacements at the acoustic midpoint frame of /r/ (relative to the acoustic midpoint of /ɑ/) were anticipated as features of interest because midpoint frames potentially best represent the phoneme, as the acoustic midpoint is likely associated with the maximum vocal tract constriction (Boyce & Espy-Wilson, 1997). However, time-dependent displacement trajectories may provide additional information that is useful for distinguishing accurate productions from misarticulations and were therefore included in two of the data representations considered.

In the present study, we focused on five candidate data representations. The first three used displacements of all possible pairs of tongue parts at the midpoint frame of /r/: (1) the dorsum and blade, (2) the dorsum and root and (3) the root and blade. The fourth (4) used displacements of the three tongue parts from all the frames of the /ɑr/ production (i.e. the full trajectories). For the fifth (5), we performed principal component analysis (PCA) to project the tongue part displacement trajectories onto a lower-dimensional subspace in which each dimension is a principal component, emphasizing the dimensions of greatest variance in the trajectories. Thus, PCA produced scores that potentially captured the important properties of the trajectories while reducing both noise and the number of features (i.e. dimensions) (Aggarwal, 2015). We expected the PCA score data representation to have the greatest classification accuracy.

For all data representations, trajectories were interpolated to the same signal length (39 frames, corresponding to one of the longest productions). As an approximation of the midpoint of /r/ for data representations (1) through (3), the selected displacement values were at frame 32 of 39 from these interpolated frames. This timepoint was determined from the average fraction of the duration up to the acoustic midpoint of /r/ compared to the duration through the end of /r/. For data representation (5), we calculated principal components from the 39 time points for each of the three tongue parts (a total of 117 initial dimensions) and selected scores from the first four principal components, which explained 94.22% of the variance in the full dataset.

Classification using support vector machines

We tested each of the five data representations to determine which most accurately distinguished accurate from inaccurate productions. The expectation was that data

representations resulting in classifiers with high accuracies would be suitable for future use to drive ultrasound-based feedback. For example, if the root and blade midpoint displacement data representation resulted in the highest accuracy, that would imply these features best capture the difference between accurate and misarticulated productions. If the full trajectory data representation instead achieved the highest accuracy, important features that distinguish accuracy would apparently be contained in all three tongue parts or in frames other than the midpoint frames, such as movements occurring between the /ɑ/ and /r/ phonemes. Accordingly, for each of the five data representations, separate SVM classifiers were trained to distinguish whether displacement data were from accurate vs misarticulated productions, defined based on perceptual rating thresholds.

Supervised machine learning methods like SVM are typically used to train a classification model on known data to predict classes of new data samples (Hastie et al., 2009). SVMs construct a hyperplane maximizing the distance (margin) between the boundaries of separate classes, allowing for better generalization of patterns from training data to new test data. The advantage of using SVMs is that they are simple linear models yet can handle more complex nonlinear data, which would be necessary if nonlinear patterns relate displacement trajectories to production accuracy. SVMs have previously been used to classify articulator movements, using EMA data to categorize consonants or vowels (Wang et al., 2013). In this study, we classified measured tongue part movements to predict accurate or misarticulated productions, defined based on observed perceptual ratings.

MATLAB's *fitcsvm* function (R2019a, MathWorks, Inc. Natick, MA) with sequential minimal optimization (SMO) was used to train radial basis function (RBF) kernel SVM classifiers (Muller et al., 2001) on the five representations of the displacement trajectory data. SMO is a classic solver used for training SVMs, and the RBF kernel type was selected due to its common use for SVMs and ability to handle nonlinear data (Smola & Schölkopf, 2004; Hastie et al., 2009). Prior to training SVMs, values for two hyperparameters must be optimized: the box constraint (also known as the slack penalty constant) and the kernel scale (also known as sigma). Bayesian optimization was used to lower the time spent tuning SVM hyperparameter values. For each model, MATLAB's Bayesian optimization algorithm, *bayesopt*, was run with at least 150 iterations to find the optimal values for these two hyperparameters. Optimal values were considered those that minimized the objective function, defined in each model as the function that returns the misclassification rate. The use of Bayesian optimization also estimates the classification accuracy as one minus the minimum objective function value (the minimum misclassification rate). Additionally, the perceptual rating threshold used to assign the class of the production as accurate or misarticulated was evaluated as an additional hyperparameter via a broad grid search, which involved selecting thresholds that spanned the range of possible thresholds (0.5 to 9.5 in increments of 0.5) and constructing models for each possibility.

Thus, Bayesian optimization was run for each perceptual rating threshold selected and for each data representation. This step both tuned the classifiers for optimal performance and broadly compared the classification performance obtained using the different data representations. All classification models during this step were trained with 10-fold cross-validation based on the 41 speaker sets in the dataset.

For this cross-validation setup, 10 versions of each model were constructed with productions from four speaker sets (~10% of 41 total speaker sets) excluded from the training fold and used as the test fold to calculate classification accuracy, with one version excluding five speaker sets. The selection of speakers excluded from the training folds was randomized without replacement, meaning each speaker set was included in the test fold for exactly one version. The final optimized classification accuracy for the model was calculated from the test fold results combined from the 10 versions. Thus, each speaker set was tested with a version of the model for which none of the productions from the speaker set had been used in the training fold, assessing the applicability of the classifier on new tongue movement patterns previously unobserved by the model. This separation into 10 folds is a classic choice for balancing model bias and variance (Hastie et al., 2009). The choice of speaker sets for each fold was randomized for each Bayesian optimization iteration.

For a more direct comparison of data representations, optimized SVM classifiers for the three data representations with the highest estimated classification accuracies were then trained by using the respective hyperparameter values that minimized the misclassification rates. These models used the same cross-validation method as used for Bayesian optimization, except that the same randomized choice of speaker sets was used for all three models. Because the models were trained on the same productions, they could be more directly compared.

Predictive parameter δ

The results of testing data representations (1) through (5) suggested that a simple parameter based on the dorsum and blade midpoint data representation could distinguish accurate from misarticulated productions in this study. To evaluate potential use in a simplified UBT display, we compared the predictive ability of a candidate simple parameter, denoted as δ, with other data representations via regression models described in the following section. We defined δ as a linear combination representing the signed distance in dorsum and blade midpoint displacement values from a linear SVM classifier boundary trained on all productions in the dataset. Specifically, when the SVM boundary for the dorsum and blade displacements (Δ_{dorsum} and Δ_{blade}, respectively) is the line defined by the coefficients a and b with the constant bias c,

$$a\,\Delta_{\text{dorsum}} + b\,\Delta_{\text{blade}} + c = 0, \tag{1}$$

the signed distance to the boundary δ can be written as

$$\delta = \frac{a\,\Delta_{\text{dorsum}} + b\,\Delta_{\text{blade}} + c}{\sqrt{a^2 + b^2}}. \tag{2}$$

Regression models

Support vector regression models (Smola & Schölkopf, 2004) relating observed and predicted perceptual ratings were trained on the three data representations with the highest classification accuracies. MATLAB's *fitrsvm* function and similar parameters for training

the SVM classifiers were used in this analysis, with the addition of including the epsilon value (representing allowed error for the prediction) as a hyperparameter optimized with *bayesopt*. In addition to these support vector regression models, a linear regression model related observed perceptual ratings to the predictive parameter δ. To allow for more straightforward interpretation, predicted ratings from all four regression models were limited to the possible range of observed ratings from 0 to 10, by considering predicted ratings less than 0 or greater than 10 as 0 or 10 respectively. The effectiveness of these data representations for assessing misarticulations was evaluated and compared by calculating the correlation and root-mean-square error between observed and predicted ratings. Predicting the output of a continuous variable with regression models, in addition to the classification of accurate vs misarticulated, fits well with our ultimate goal of providing graduated feedback to the user about the accuracy of articulator movements.

Results

Perceptual Ratings

As noted above, our approach was to use results from a perceptual study as 'ground truth' for determining which representations of tongue part movement data worked best to classify accurate vs inaccurate tongue part movement patterns and predict continuous accuracy. The distribution of observed perceptual ratings (averaged from the three raters for each production) is shown in Figure 2, with the left and right histograms indicating the counts of productions from RSSD and TD speakers respectively. Most productions from RSSD speakers had perceptual ratings on the lower end of the scale (0–5), including both sets of productions from the RSSD speaker with differing patterns over the course of therapy. In contrast, most productions from TD speakers were on the higher end of the scale (6–10). Regarding interrater agreement, the ICC value (Koo & Li, 2016) for all rated productions was 0.90 with a 95% confidence interval of 0.89–0.91, indicating good to excellent agreement (values greater than 0.75). However, when computed using only productions with average rating 2–8 (39% of all rated productions), the ICC value was 0.39 with a 95% confidence interval 0.22–0.48, indicating poor interrater agreement (values less than 0.5) for productions of ambiguous accuracy.

Tongue tracking

Example tongue tracking with TonguePART and corresponding tongue part displacement trajectories are shown in Figure 3 for two representative productions of /ɑr/. The frames displayed on the left correspond temporally to the acoustic midpoints of /ɑ/ and /r/. Figure 3a shows a production rated as accurate (average perceptual rating of 10.0) from a TD speaker and 3b shows a production rated as relatively poor (average perceptual rating of 3.5) from a RSSD speaker. The tongue movement pattern shown in 3a (right) with large positive blade and moderate negative dorsum displacement values was a common pattern for the TD speakers in this dataset, and the small magnitude of movement shown in 3b was a common pattern for RSSD speakers in this dataset.

Classification: predicting /ɑr/ as accurate vs misarticulated

Figure 4 shows 2D scatterplots of tongue part displacements between the estimated acoustic midpoints of /ɑ/ and /r/ for each production, with a colour map representing associated perceptual ratings. Displacements of all three tongue parts are typically nearer to zero for misarticulated than accurate productions. These plots include solid black contours indicating the optimized RBF SVM classifiers' decision boundaries for the three pairings of tongue part displacements, with separate contours shown for each cross-validation fold. Cleaner separation between more accurate and more misarticulated productions can be observed in the middle panel, displaying the dorsum and blade displacements. The middle panel also includes a dashed magenta line indicating the decision boundary for a linear SVM classifier trained without cross-validation and used for the predictive δ parameter analysed further.

Figure 5 shows classification accuracies estimated from Bayesian optimization runs, tuning the box-constraint and kernel-scale hyperparameters for each of the five data representations, vs perceptual rating thresholds dividing misarticulated and accurate classes. Thresholds immediately near 0 or 10 were undesirable because most productions would be labelled as correct or misarticulated, respectively, so that the classification model easily learned to predict all productions as correct or misarticulated (i.e. because most productions were labelled with the category predicted for all productions, this resulted in high accuracies). The highest estimated classification accuracy excluding these undesirable thresholds occurred at a perceptual rating threshold of 5.5 for the data representation using scores from PCA; therefore, this rating threshold was used for the rest of the classification analysis. The close classification accuracy at the rating threshold of 5.5 for the full trajectories also implies likely equivalent outcomes for this data representation compared to the PCA score data representation.

Figure 6 shows classification accuracies for RBF SVM classifiers trained on the same cross-validation fold configuration, using the three data representations with the highest accuracies estimated with Bayesian optimization (Figure 5) and the hyperparameters associated with these accuracies (box-constraint and kernel-scale hyperparameters optimized for the rating threshold 5.5). Accuracies are shown for each of the 10 cross-validation folds. Combining the folds, classification accuracies were 88.5%, 87.7% and 89.2% for the dorsum and blade midpoint, scores from PCA and full trajectory data representations, respectively. To compare the simplest and most complex of these data representations, i.e. the dorsum and blade displacement and the full trajectories, respectively, a paired t test was applied to classification accuracies of the different folds and indicated no significant difference (p=0.531, t=−0.651, df=9).

The relative distribution of perceptual ratings for productions misclassified by the final optimized RBF SVM classifiers (Figure 6) is shown by the histogram in Figure 7, with each bin height calculated as the misclassification count for the specified range of ratings divided by the associated total count of productions (Figure 2). The indicated values were calculated as the mean relative proportion for these classifiers using the three data representations, with the error bars indicating the standard deviation. Most of the misclassifications had perceptual ratings near the middle of the scale (3–7). The plotted standard deviations resulted from misclassification counts that only differed by up to three productions within each bin

(appearing larger in some bins due to low counts), indicating that the classifiers from the three data representations performed similarly across the extent of perceptual accuracy.

Regression: predicting /ɑr/ accuracy

Predicted perceptual ratings from cross-validated support vector regression models trained on the dorsum and blade midpoint, PCA score and full trajectory data are plotted vs observed ratings in Figure 8, with lines of best fit determined by linear regression. Pearson's correlation coefficients were 0.788, 0.736 and 0.777 with $p < 0.001$ for all three models, respectively, indicating strong, significant correlations between predicted and observed ratings. Root-mean-square errors (RMSE) of observed vs predicted perceptual ratings for these representations were 2.12, 2.33 and 2.16, respectively.

The predictive parameter δ based on the dorsum and blade midpoint displacement is compared in the lower right panel of Figure 8. The linear SVM classifier decision boundary used to calculate this parameter is shown in the middle panel of Figure 4, with coefficients $a = -14.71$, $b = 13.38$ and $c = -1.59$ as used in Equations (1) and (2). A linear regression of observed vs predicted ratings for this predictive parameter yielded an RMSE of 2.17 and a correlation coefficient of 0.772 with $p < 0.001$.

Discussion

Articulatory features of accurate and misarticulated /r/

The classification methods reported here successfully captured the difference between accurate and misarticulated productions (categorical accuracy) based on ultrasound-tracked displacements of discrete tongue parts. Measurements were performed on a dataset which covered both sexes and a range of ages for child speakers of a rhotic American English dialect. The measured data used to train and validate these models thus included a relatively wide range of articulatory strategies among the 41 speaker sets in the analysed dataset, comprising tongue movement patterns with final tongue shapes that could be characterized either as 'bunched' or 'retroflex' (Dugan et al., 2019a).

High classification accuracies greater than 85% were attained for optimized SVM classifiers employing three data representations. Two of these classifiers used relatively complex data, either the full displacement trajectories of the three tongue parts or PCA scores for the full trajectories. Notably, comparable accuracy greater than 85% was attained using the dorsum and blade midpoint displacement data representation, while the remaining combinations (root/dorsum or root/blade) resulted in lower classification accuracies (Figure 5). In analysis of regression vs perceptual ratings (graduated accuracy), the dorsum and blade midpoint data representation also had correlation coefficient and RMSE values similar to the two, more complex representations. It is notable that the use of PCA, which was expected to reduce noise existing in the full trajectories while retaining information from the tongue root and frames throughout the trajectories, did not appear to substantially improve classification nor regression performance relative to the simple dorsum/blade midpoint displacement data representation.

The observed high classification accuracies and significant correlation between observed and predicted accuracy ratings for the blade vs dorsum data representation make sense if we consider the salient features of accurate articulations and misarticulations in tongue part displacement trajectories measured with TonguePART. (As noted earlier, because TonguePART divides the tracked tongue contour into three partitions of equal spans, automatic assignments of root, dorsum and blade regions by TonguePART may differ from the corresponding anatomical tongue parts.) Most perceptually accurate productions had large positive blade displacement values at the midpoint of /r/, as shown in Figure 3a and Figure 4. This is expected for /ɑr/ productions since /ɑ/ and /r/ have contrasting tongue shapes; the characteristic posture for /ɑ/ involves a lowered tongue dorsum, lowered tongue blade and retracted tongue root, whereas all variants of /r/ involve raising of the tongue blade from the /ɑ/ posture to form the palatal constriction typical for rhotics (Boyce, 2015). TonguePART appeared to capture the presence of this anterior constriction as a positive displacement of the region partitioned as the blade. In addition to the anatomical blade, this region often included portions of the anatomical dorsum, so that some large positive blade displacements represented the constriction formed by the anatomical dorsum in bunched /r/ configurations.

TonguePART appeared to capture the presence of anatomical root constriction—the posterior constriction necessary for /r/—in its measured displacements for the region partitioned as the dorsum. Many accurate productions displayed moderate negative dorsum displacement values or small magnitudes of dorsum displacement. A lowered dorsum is a common characteristic of many accurate /r/ tongue shapes and may follow from the fact that /r/ requires constrictions both at the palate and in the pharynx. Because tongue volume must be maintained, pushing tongue volume into the pharynx while keeping the tongue front raised toward the palate will result in lowering of the dorsum (Alwan et al., 1997; Klein et al., 2013). Typically, this depressed dorsum is achieved by increased grooving along the midline of the tongue dorsum (Boyce, 2015), which is quantified by TonguePART as reduced or negative dorsum displacement. Displacements of the partitioned root were less useful in distinguishing accurate articulation. Negative root displacement indicated tongue root advancement from /ɑ/ characteristic in some misarticulations, such as in humped shapes (Boyce, 2015), but this negative displacement could also indicate accurate posterior constriction because of two possible causes: the typical location for tongue root retraction for /r/ is lower in the vocal tract than for /ɑ/, and the region partitioned as the tongue root may sometimes include some of the anatomical dorsum.

TonguePART also captured relevant patterns in tongue movements for inaccurate production of /r/. One common misarticulation pattern was small or near-zero displacement of all three tongue parts (shown in Figure 3b and in the scatterplots in Figure 4 by the cluster of blue data points near (0, 0)). This pattern is consistent with a percept of a distorted low-back vowel or schwa instead of accurate /r/. The similarity between the initial tongue posture of the low-back vowel /ɑ/ and the following inaccurate /r/ production thus can result in near-zero displacements. The second common /r/ misarticulation pattern was the ‘humped’ error shape. This misarticulation pattern appeared in the TonguePART data as greater positive dorsum displacement values with lower positive blade displacement values. In essence,

blade and dorsum displacements measured by TonguePART were sufficient to distinguish both patterns of common /r/ misarticulations from accurate productions.

Study limitations

Successful classification of accurate vs misarticulated /r/ was obtained despite limitations of TonguePART's tracking approach. Its measured displacements of partitioned tongue regions may not precisely capture full articulatory movements. The automatic partition of the tongue into parts with equal span only provides a broad approximation of regional functionality. Additionally, this tracking methodology depends on the initial image frame used for reference displacement values, so that the classification and regression results may be more variable for speakers with inconsistent /ɑ/ initial shapes. Potential improvements to TonguePART could improve tracking accuracy, while decreasing the fraction of productions excluded. For example, automatically adjusted brightness thresholds may improve tracking for productions with changes in tongue surface brightness over time.

Although classification accuracy was high for the best data representations, and correlation between observed and predicted perceptual ratings in all regression models investigated here was statistically significant ($p<0.001$), regression model predictions had limited precision, indicated by substantial RMSE values (2.12–2.33 on the scale of 0–10) for perceptual ratings as well as the visual spread of points from the regression lines (Figure 8). This is likely not a problem arising with the regression models themselves, but rather with the nature of perceptual ratings. Although overall our study found good to excellent interrater reliability (ICC values > 0.75 for all ratings), there was relatively poor agreement (ICC values < 0.5) for intermediate accuracy ratings (2–8). Supporting this claim, more misclassifications occurred in the middle range of the ratings (Figure 7), for which perceptual ratings are more ambiguous. This variability may be inherent to speech ratings and therefore extant in any analysis with RSSD speakers.

While these tongue part displacement trajectory measurements were intended to account for differences across speakers (e.g. normalisation of displacements by a reference tongue length scale), some individual anatomical differences (e.g. palate shape) may have affected the identified relationship between displacements and production accuracy, potentially lowering classification and regression accuracies despite the wide range of final tongue shapes covered. Ultrasound imaging does not provide direct anatomical measurements of the vocal tract that would allow for further exploration of these differences. However, additional orofacial measurements (e.g. lower facial height measured by calliper) could potentially give insight.

This study employed a head stabilizer for recording ultrasound image data. The use of a head stabilizer is expected to reduce measurement variability, especially given that we see less jaw movement for head-stabilized measurements when compared to handheld use of ultrasound. However, imaging using a head stabilizer may not be ideal for real-world clinical use with children. Thus, a possible limitation for this study is the question of whether these results would differ from handheld measurements. In addition to reduced measurement variability, differences in magnitudes of articulation due to restricted jaw movement may be expected in head-stabilized data. While some studies have compared

handheld and head-stabilized measurements from ultrasound (Ménard et al., 2012; Zharkova et al., 2015), the possible impact on relative measurement of displacement for /ɑr/ is not clear. In our own experience, measured data from handheld vs head-stabilized situations has shown similar tongue part displacements for accurate /ɑr/ productions. Thus, we do not expect our head-stabilized results to greatly differ from handheld measurements or other appropriate imaging configurations, but more data are needed to address this question.

Ultrasound imaging itself has limitations for characterization of tongue motion in speech production. Ultrasound imaging quality was poor for a few speakers, due to speaker-specific anatomy or head movement relative to the probe during imaging despite the head stabilizer. This poor quality caused some of the larger percentages of productions removed for some speakers. Because the visible tongue surface contour was still tracked in images with these problems, the visual check of tracking quality still resulted in the inclusion of these productions. For some productions, different parts of the tongue surface became more visible or less visible over the course of some productions, which in some cases was due to tongue movement in and out of the hyoid bone shadow, affecting partitioning of tongue parts and estimates of their displacement. However, classification accuracies remained high despite these changes in partition, indicating that this effect was often within expected tongue displacement patterns. Difficulty in imaging the tongue tip may also limit differentiation of tongue movement patterns, and ultrasound imaging does not image other regions of the vocal tract like the jaws, palate and lips that may affect production and the resulting perceptual ratings.

Ultrasound imaging quality and reliability could potentially be improved by adjusting transducer placement with a more effective head stabilizer. An ideal head stabilizer design would securely hold the ultrasound array at the midsagittal plane, without causing discomfort to subjects. One possibility may be an array holder integrated into a comfortable headpiece to be worn by the subject, which may be preferable to the fixed stabilizer employed here.

Potential biofeedback using the δ parameter

The observation that dorsum and blade displacements captured accurate /r/ efficiently, without the need for the full, higher-dimensional trajectories from TonguePART, led to definition of the simple δ parameter, i.e. the signed distance of dorsum and blade midpoint displacement values (scatterplot in the middle panel of Figure 4) from the linear SVM boundary (magenta dashed line in Figure 4). Plotting this predictive parameter against perceptual ratings (lower right panel of Figure 8) reveals a general ability to predict a continuous perceptual accuracy value (graduated accuracy), similar to that found using more complex support vector regression models (Figure 8).

Compared to more complex classification methods, a major advantage of the δ parameter as a potential biofeedback target is that it can be evaluated using data recorded from only one time point. For biofeedback, the δ parameter may thus be used to evaluate accuracy of a production frame-by-frame in real time, providing immediate feedback on whether a tongue movement target has been reached. For a real-time biofeedback display, such a single parameter indicating articulatory success or failure is computationally more efficient than

calculations that rely on full trajectories or other higher-dimensional parameters. Moreover, because the δ parameter is based on capturing the articulatory features of accurate /r/, we expect the discriminative ability of this parameter to be maintained for other vowel contexts with /r/, although with different boundaries between accurate and misarticulated productions due to different articulation for vowels other than /ɑ/. Preliminary exploration of a similar simple parameter has demonstrated success in distinguishing accurate vs misarticulated productions for postvocalic and prevocalic /r/ contexts /ir/, /ɑr/, /buər/, /ri/, /rɑ/ and /ru/ (Li et al., 2021). Further testing of this parameter in clinically applicable contexts is currently underway.

For clinical speech therapy, a single articulation target such as $\delta > 0$ using the δ parameter defined here could be presented to subjects in a simplified biofeedback display, portraying the qualities of tongue movement relevant to accurate articulation. For example, a simple display could involve moving a needle on a meter mapped to the δ parameter. Considering that $\delta = 0$ is the linear SVM boundary, a positive needle displacement ($\delta > 0$) would indicate an accurate production based on the measured combination of dorsum and blade displacements. Negative needle displacement would correspond to ineffective coordination of the blade and dorsum. To account for differences in articulatory configuration between speakers, the target parameter value could be individually adjusted by the clinician.

Compared to UBT based on direct observation of B-mode ultrasound imaging, such a simplified display may allow a speaker to better determine whether their productions are accurate or misarticulated, while more easily understanding which attempts are closer to accurate. This understanding of both categorical (accurate vs misarticulated) and graduated (how close to accurate) accuracy may more effectively guide the speaker to learn and practice accurate speech. Although implementation and testing of a new UBT system is beyond the scope of this study, the present results suggest that ultrasound-based motion tracking can classify and characterize speech production accuracy in a manner suitable for use in simplified or gamified real-time biofeedback.

Conclusion

This study compared tongue part displacements for accurate vs misarticulated /ɑr/ productions from children with TD and RSSD speech, by training SVM classification and regression models on different displacement data representations. Results indicate that automatic calculation of tongue part displacements from ultrasound images, which would be necessary for a real-time display showing tongue movement outcomes, captured relevant articulatory features of /r/. Accurate productions of /ɑr/ were characterized by large positive blade displacement and moderate negative dorsum displacement, which may represent the anterior and posterior constrictions for /r/, respectively. Misarticulated productions often had low displacement magnitude, potentially indicating a percept of a distorted low-back vowel. High classification accuracies from the SVM classifiers indicate that tongue part displacement trajectories may be used to distinguish accurate productions from misarticulations. In addition, regression models indicated a general ability to predict a continuous rating of perceptual accuracy.

Results for the data representation based on dorsum and blade midpoint displacements were comparable to more complicated data representations and suggested a linear combination of these values as a predictive parameter. The δ parameter defined here can potentially be used in real time to establish articulation targets in simplified ultrasound biofeedback therapy. Interpreting such a simplified display may reduce the cognitive load of ultrasound biofeedback. Evaluation of the effectiveness of this potential UBT approach is beyond the scope of the current study and will require further exploration.

Acknowledgements

The authors would like to thank Kathryn Eary, Michael Swearengen, Gregory Terrell, Sarah Stack, Maurice Lamb and Sarah Schwab for their contributions to this project, as well as all participants in the study. We are also grateful to Siemens Medical Solutions USA, Inc., for making the Acuson X300 scanner available for this project.

Funding:

This work was funded by the University of Cincinnati Creating Our Third Century funding support and National Institute on Deafness and Other Communication Disorders grants R01 DC017301 (awarded to Suzanne Boyce, Michael A. Riley and T. Douglas Mast) and R01 DC013668 (awarded to Suzanne Boyce and Doug Whalen).

References

Adler-Bock M, Bernhardt BM, Gick B, & Bacsfalvi P. (2007). The use of ultrasound in remediation of North American English /r/ in 2 adolescents. American Journal of Speech-Language Pathology, 16(2), 128–139. 10.1044/1058-0360(2007/017) [PubMed: 17456891]

Aggarwal CC (2015). Data Mining. Springer International Publishing. 10.1007/978-3-319-14142-8

Alwan A, Narayanan S, & Haker K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. The Journal of the Acoustical Society of America, 101(2), 1078–1089. 10.1121/1.417972 [PubMed: 9035399]

Annand CT, Lamb M, Dugan S, Li SR, Woeste HM, Mast TD, Riley MA, Masterson JA, Mahalingam N, Eary KJ, Spencer C, Boyce S, Jackson S, Baxi A, and Seward R. (2019). Using ultrasound imaging to create augmented visual biofeedback for articulatory practice. Proceedings of the 20th Annual Conference of the International Speech Communication Association, 974–975. 10.21437/Interspeech.2019

Bernhardt B, Gick B, Bacsfalvi P, & Adler-Bock M. (2005). Ultrasound in speech therapy with adolescents and adults. Clinical Linguistics & Phonetics, 19(6–7), 605–617. 10.1080/02699200500114028 [PubMed: 16206487]

Boyce S. (2015). The articulatory phonetics of /r/ for residual speech errors. Seminars in Speech and Language, 36(04), 257–270. 10.1055/s-0035-1562909 [PubMed: 26458201]

Boyce S, & Espy-Wilson CY (1997). Coarticulatory stability in American English /r/. The Journal of the Acoustical Society of America, 101(6), 3741–3753. 10.1121/1.418333 [PubMed: 9193061]

Boyce S, Tiede M, Espy-Wilson C, Groves-Wright K. (2015). Diversity of tongue shapes for American English rhotic liquids. Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, Scotland, August 2015.

Boyce SE, Combs S, & Rivera-Campos A. (2011). Acoustic and articulatory characteristics of clinically resistant /r/. The Journal of the Acoustical Society of America, 129(4), 2625–2625. 10.1121/1.3588729

Bressmann T, Harper S, Zhylich I, & Kulkarni GV (2016). Perceptual, durational and tongue displacement measures following articulation therapy for rhotic sound errors. Clinical Linguistics & Phonetics, 30(3–5), 345–362. 10.3109/02699206.2016.1140227 [PubMed: 26979162]

Byun TM, & Hitchcock ER (2012). Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation. American Journal of Speech-Language Pathology, 21(3), 207–221. 10.1044/1058-0360(2012/11-0083) [PubMed: 22442281]

Campbell F, Gick B, Wilson I, & Vatikiotis-Bateson E. (2010). Spatial and temporal properties of gestures in North American English /r/. Language and Speech, 53(1), 49–69. 10.1177/0023830909351209 [PubMed: 20415002]

Chung H, Farr K, & Pollock KE (2019). Rhotic vowel accuracy and error patterns in young children with and without Speech Sound Disorders. Journal of Communication Disorders, 80, 18–34. 10.1016/j.jcomdis.2019.03.003 [PubMed: 31022634]

Cleland J, Wrench A, Lloyd S, & Sugden E. (2018). ULTRAX2020: Ultrasound technology for optimising the treatment of speech disorders: Clinicians' resource manual. Glasgow, Scotland: University of Strathclyde. 10.15129/63372

Delattre P, & Freeman DC (1968). A dialect study of American r's by x-ray motion picture. Linguistics, 6(44), 29–68. 10.1515/ling.1968.6.44.29

Dugan S, Li SR, Masterson J, Woeste H, Mahalingam N, Spencer C, Mast TD, Riley MA, & Boyce SE (2019a). Tongue part movement trajectories for /r/ using ultrasound. Perspectives of the ASHA Special Interest Groups, 4(6), 1644–1652. 10.1044/2019_PERS-19-00064 [PubMed: 32524032]

Dugan SH, Silbert N, McAllister T, Preston JL, Sotto C, & Boyce SE (2019b). Modelling category goodness judgments in children with residual sound errors. Clinical Linguistics & Phonetics, 33(4), 295–315. 10.1080/02699206.2018.1477834 [PubMed: 29792525]

Espy-Wilson CY, Boyce SE, Jackson M, Narayanan S, & Alwan A. (2000). Acoustic modeling of American English /r/. The Journal of the Acoustical Society of America, 108(1), 343–356. 10.1121/1.429469 [PubMed: 10923897]

Faugloire E, Bardy BG, Merhi O, & Stoffregen TA (2005). Exploring coordination dynamics of the postural system with real-time visual feedback. Neuroscience Letters, 374(2), 136–141. 10.1016/j.neulet.2004.10.043 [PubMed: 15644280]

Gamer M, Lemon J, & Singh IFP (2019). irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Gibbon F, & Lee A. (2015). Electropalatography for older children and adults with residual speech errors. Seminars in Speech and Language, 36(04), 271–282. 10.1055/s-0035-1562910 [PubMed: 26458202]

Green JR, & Wang Y-T (2003). Tongue-surface movement patterns during speech and swallowing. The Journal of the Acoustical Society of America, 113(5), 2820–2833. 10.1121/1.1562646 [PubMed: 12765399]

Guenther FH, Espy-Wilson CY, Boyce SE, Matthies ML, Zandipour M, & Perkell JS (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. The Journal of the Acoustical Society of America, 105(5), 2854–2865. 10.1121/1.426900 [PubMed: 10335635]

Hastie T, Tibshirani R, & Friedman J. (2009). The elements of statistical learning. Springer New York. 10.1007/978-0-387-84858-7

Hitchcock E, Harel D, & Byun T. (2015). Social, emotional, and academic impact of residual speech errors in school-aged children: A survey study. Seminars in Speech and Language, 36(04), 283–294. 10.1055/s-0035-1562911 [PubMed: 26458203]

Klein HB, McAllister Byun T, Davidson L, & Grigos MI (2013). A multidimensional investigation of children's /r/ productions: Perceptual, ultrasound, and acoustic measures. American Journal of Speech-Language Pathology, 22(3), 540–553. 10.1044/1058-0360(2013/12-0137) [PubMed: 23813195]

Koo TK, & Li MY (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. 10.1016/j.jcm.2016.02.012 [PubMed: 27330520]

Li SR, Annand CT, Dugan S, Schwab SM, Eary KJ, Swearengen M, Stack S, Boyce S, Riley MA, Mast TD (2021). An automatic, simple ultrasound biofeedback parameter for distinguishing accurate and misarticulated rhotic syllables. Proceedings of the 22nd Annual Conference of the International Speech Communication Association, 636–640. 10.21437/Interspeech.2021-1749

McAllister Byun T, Halpin PF, & Szeredi D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. Journal of Communication Disorders, 53, 70–83. 10.1016/j.jcomdis.2014.11.003 [PubMed: 25578293]

Ménard L, Aubin J, Thibeault M, & Richard G. (2012). Measuring tongue shapes and positions with ultrasound imaging: a validation experiment using an articulatory model. Folia Phoniatrica et Logopaedica, 64(2), 64–72. 10.1159/000331997 [PubMed: 22212175]

Mechsner F. (2004). A psychological approach to human voluntary movements. Journal of Motor Behavior, 36(4), 355–370. 10.1080/00222895.2004.11007993 [PubMed: 15695214]

Mechsner F, Kerzel D, Knoblich G, & Prinz W. (2001). Perceptual basis of bimanual coordination. Nature, 414(6859), 69–73. 10.1038/35102060 [PubMed: 11689944]

Müller KR, Mika S, Rätsch G, Tsuda K, & Schölkopf B. (2001). An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks, 12(2), 181–201. 10.1109/72.914517 [PubMed: 18244377]

Munson B, Schellinger SK, & Carlson KU (2012). Measuring speech-sound learning using visual analog scaling. Perspectives on Language Learning and Education, 19(1), 19. 10.1044/lle19.1.19

Nguyen N, Marchal A, & Content A. (1996). Modeling tongue-palate contact patterns in the production of speech. Journal of Phonetics, 24(1), 77–97. 10.1006/jpho.1996.0006

Preston JL, McAllister Byun T, Boyce SE, Hamilton S, Tiede M, Phillips E, Rivera-Campos A, & Whalen DH (2017). Ultrasound images of the tongue: A tutorial for assessment and remediation of speech sound errors. Journal of Visualized Experiments, (119), Article e55123. 10.3791/55123

Preston JL, McCabe P, Tiede M, & Whalen DH (2019). Tongue shapes for rhotics in school-age children with and without residual speech errors. Clinical Linguistics & Phonetics, 33(4), 334–348. 10.1080/02699206.2018.1517190 [PubMed: 30199271]

Ribeiro MS, Cleland J, Eshky A, Richmond K, & Renals S. (2021). Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors. Speech Communication, 128, 24–34. 10.1016/j.specom.2021.02.001

Ruscello DM (1995). Visual feedback in treatment of Residual Phonological disorders. Journal of Communication Disorders, 28(4), 279–302. 10.1016/0021-9924(95)00058-X [PubMed: 8576411]

Shriberg LD, Gruber FA, & Kwiatkowski J. (1994). Developmental phonological disorders. III: Long-term speech-sound normalization. Journal of Speech and Hearing Research, 37(5), 1151–1177. 10.1044/jshr.3705.1151 [PubMed: 7823558]

Smola AJ, & Schölkopf B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199–222. 10.1023/B:STCO.0000035301.49549.88

Stone M. (2005). A guide to analysing tongue motion from ultrasound images. Clinical Linguistics & Phonetics, 19(6–7), 455–501. 10.1080/02699200500113558 [PubMed: 16206478]

Stone M, Epstein MA, & Iskarous K. (2004). Functional segments in tongue movement. Clinical Linguistics & Phonetics, 18(6–8), 507–521. 10.1080/02699200410003583 [PubMed: 15573487]

Sugden E, Lloyd S, Lam J, & Cleland J. (2019). Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders. International Journal of Language & Communication Disorders, 54(5), 705–728. 10.1111/1460-6984.12478 [PubMed: 31179581]

Varoqui D, Froger J, Pélissier J-Y, & Bardy BG (2011). Effect of coordination biofeedback on (re)learning preferred postural patterns in post-stroke patients. Motor Control, 15(2), 187–205. 10.1123/mcj.15.2.187 [PubMed: 21628724]

Wang J, Green JR, Samal A, & Yunusova Y. (2013). Articulatory distinctiveness of vowels and consonants: A data-driven approach. Journal of Speech, Language, and Hearing Research, 56(5), 1539–1551. 10.1044/1092-4388(2013/12-0030)

Westbury JR, Hashi M, & Lindstrom MJ (1998). Differences among speakers in lingual articulation for American English /ɹ/. Speech Communication, 26(3), 203–226. 10.1016/S0167-6393(98)00058-2

Wulf G. (2013). Attentional focus and motor learning: A review of 15 years. International Review of Sport and Exercise Psychology, 6(1), 77–104. 10.1080/1750984X.2012.723728

Wulf G, & Prinz W. (2001). Directing attention to movement effects enhances learning: A review. Psychonomic Bulletin & Review, 8(4), 648–660. 10.3758/BF03196201 [PubMed: 11848583]

Wulf G, Shea C, & Lewthwaite R. (2010). Motor skill learning and performance: A review of influential factors: Motor skill learning and performance. Medical Education, 44(1), 75–84. 10.1111/j.1365-2923.2009.03421.x [PubMed: 20078758]

Zhang Z, Espy-Wilson C, Boyce S, Tiede M. (2005). Modeling of the front cavity and sublingual space in American English rhotic sounds, Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, Pennsylvania, 893–896.

Zharkova N, Gibbon FE, & Hardcastle WJ (2015). Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation. Clinical Linguistics & Phonetics, 29(4), 249–265. 10.3109/02699206.2015.1007528 [PubMed: 25651199]

Zhou X, Espy-Wilson CY, 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/. Journal of the Acoustical Society of America, 123(6), 17. 10.1121/1.2902168

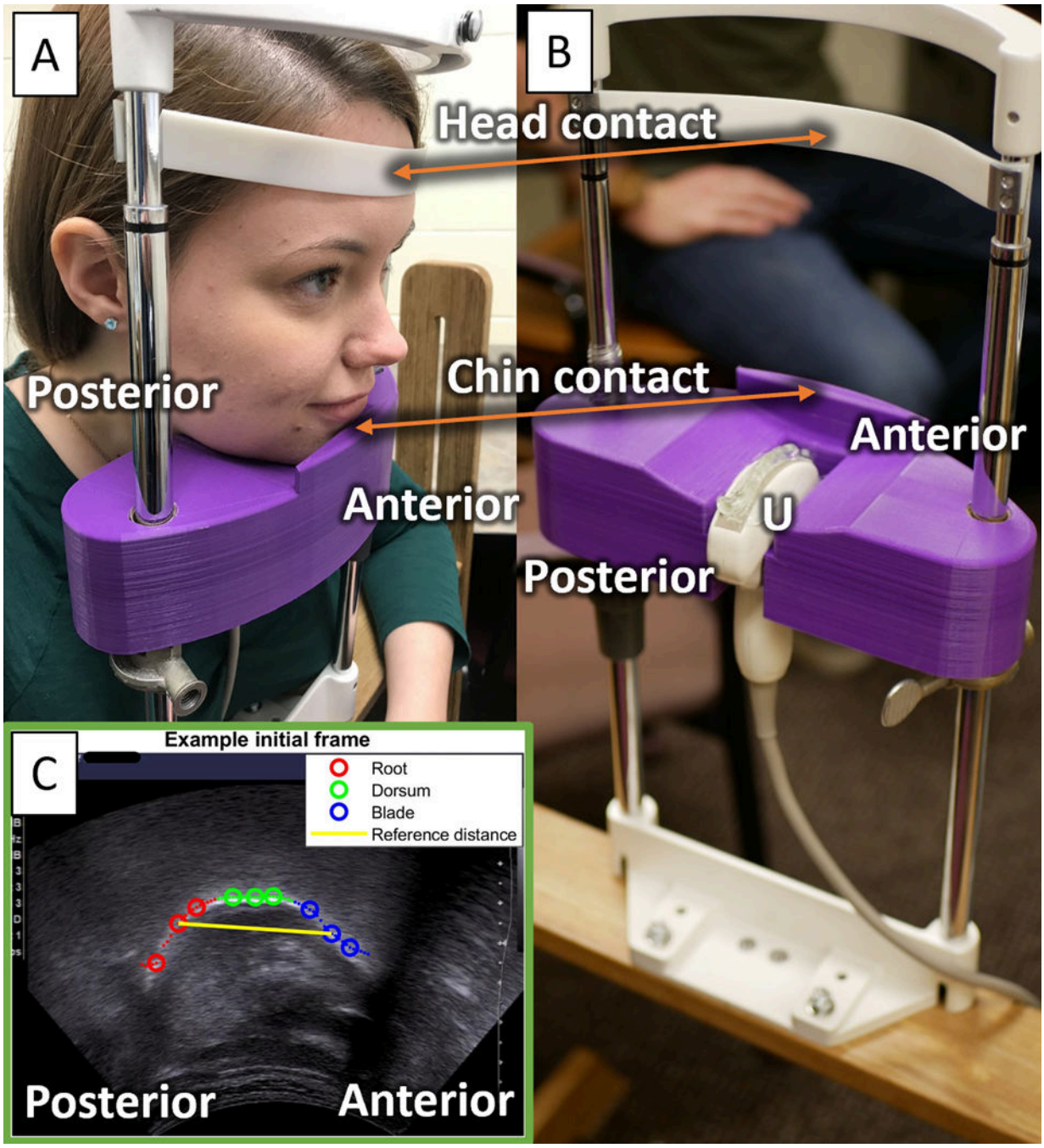

Figure 1:
Ultrasound data collection setup. Panels A and B show the head stabiliser with indicated locations of head and chin contact. Panel A shows a user positioned in the head stabiliser. Panel B shows the ultrasound transducer (U) placement in the head stabiliser. Panel C is an example midsagittal ultrasound image showing the tongue for a typical speaker at the initial video frame (the midpoint of /ɑ/) of tracking with TonguePART. The dotted line represents the automatically detected contour corresponding to the tongue-air interface, coloured red, green and blue for partitions automatically assigned as the tongue root, dorsum

and blade, respectively. The three large circles in each partition represent the reference points, and the length of the yellow line connecting the middle root and blade points is the reference distance used for normalising displacement values. *Panel B photo credit: Corrie Mayer/CEAS Marketing/University of Cincinnati*.

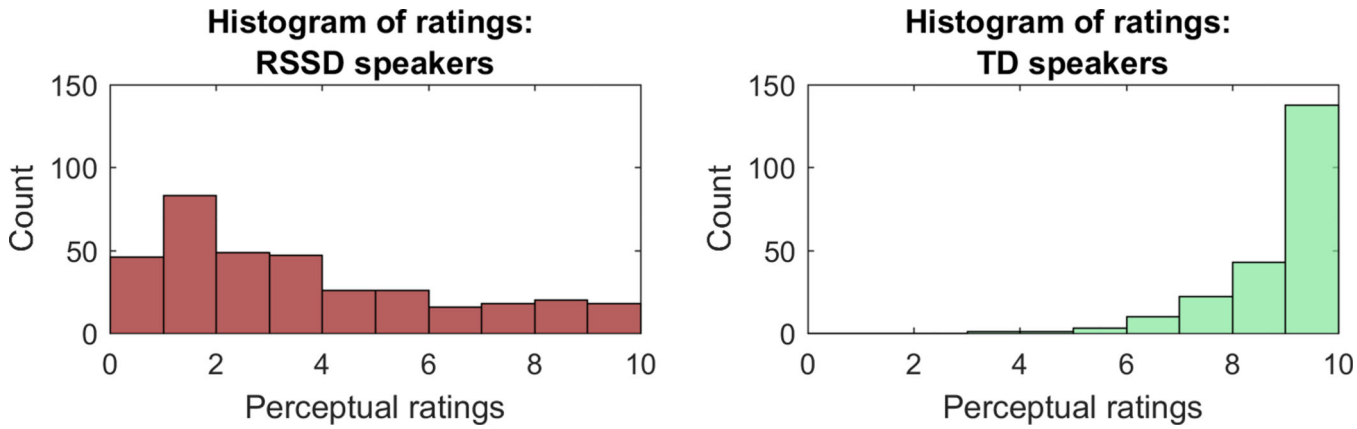

Figure 2:

Histograms of auditory perceptual ratings for productions from RSSD (left) and TD (right) speakers, averaged across the three raters. A rating of 0 was considered most 'misarticulated' and 10 most 'accurate.' Each bin represents the count of productions within the specified range of ratings.

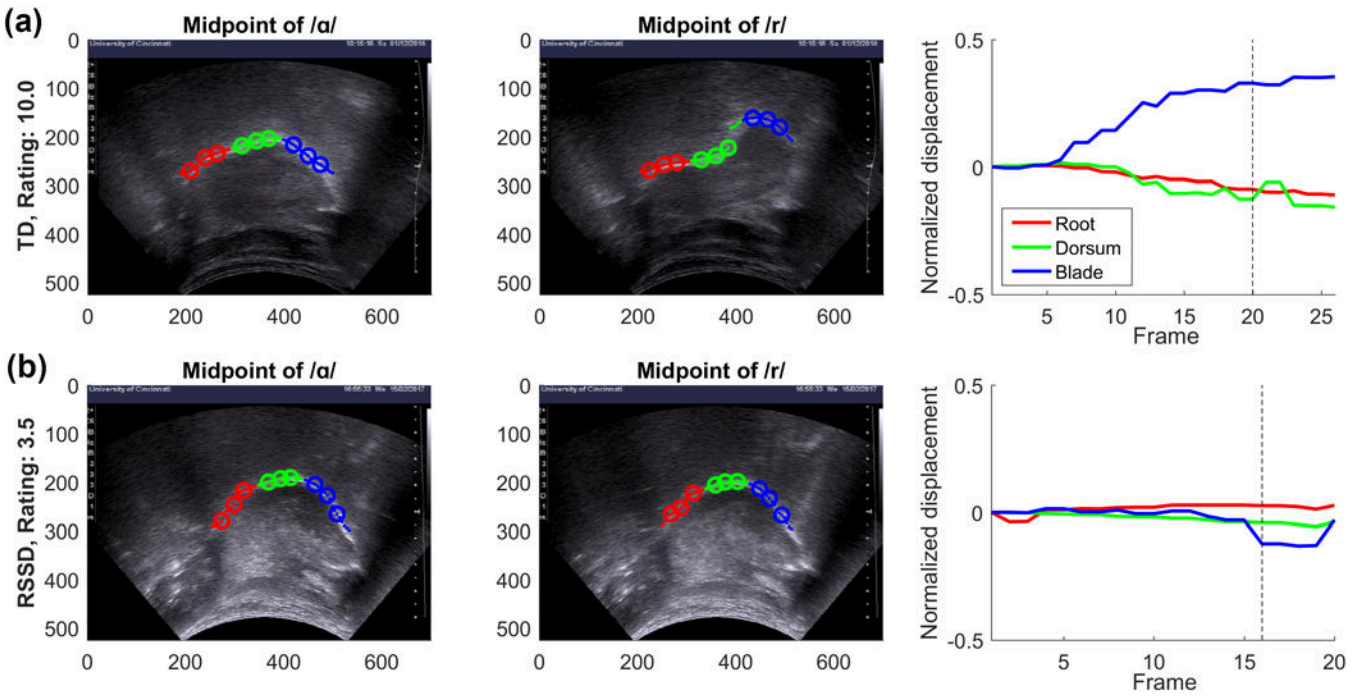

Figure 3:

Example tongue tracking and tongue part displacement trajectories calculated with TonguePART for two productions: (a) a more accurate production with a perceptual rating of 10.0 from a TD speaker; (b) a more misarticulated production with a perceptual rating of 3.5 from an RSSD speaker. Left and middle columns display the ultrasound image frames with the automatically tracked partitions of the tongue (red as the root, green as the dorsum and blue as the blade) at the midpoint of /ɑ/ and of /r/ respectively. Large circular markers indicate the three automatically selected reference points for each tongue part. The right column displays the tongue part displacement trajectories (before interpolation to 39 frames for analysis), with the midpoint of /ɑ/ corresponding to frame 1 on the horizontal axis and the midpoint of /r/ indicated by the dashed line.

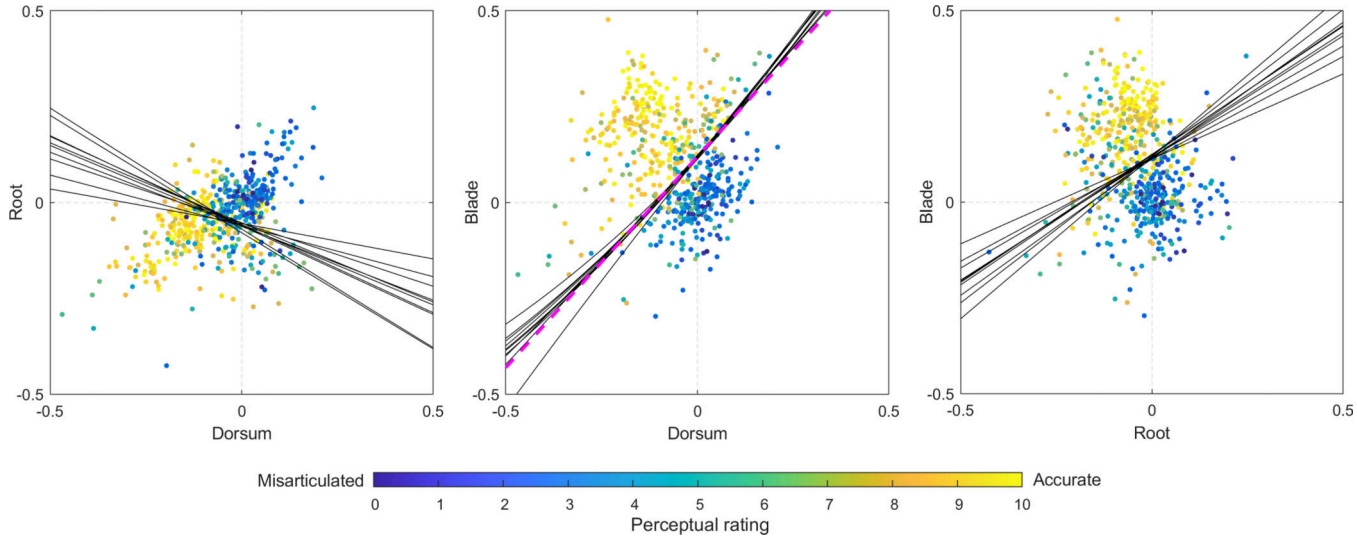

Figure 4:

Tongue part displacement values at the estimated midpoint of /r/ (frame 32 of 39 interpolated frames) and corresponding auditory perceptual ratings, marked as colour shades, for all productions in the dataset (n = 567). On the perceptual rating scale, a rating of 0 was most 'misarticulated' and 10 was most 'accurate'. Each scatterplot displays solid black lines that indicate the decision boundaries of optimized RBF SVM classifiers (Figure 6), trained on 10 cross-validation folds of the respective two tongue part displacement values at the midpoint of /r/ (e.g. dorsum and blade in the middle panel). The middle scatterplot additionally displays a dashed magenta line indicating the linear SVM boundary trained on all data (without cross-validation) and analysed in Figure 8 as a predictive parameter.

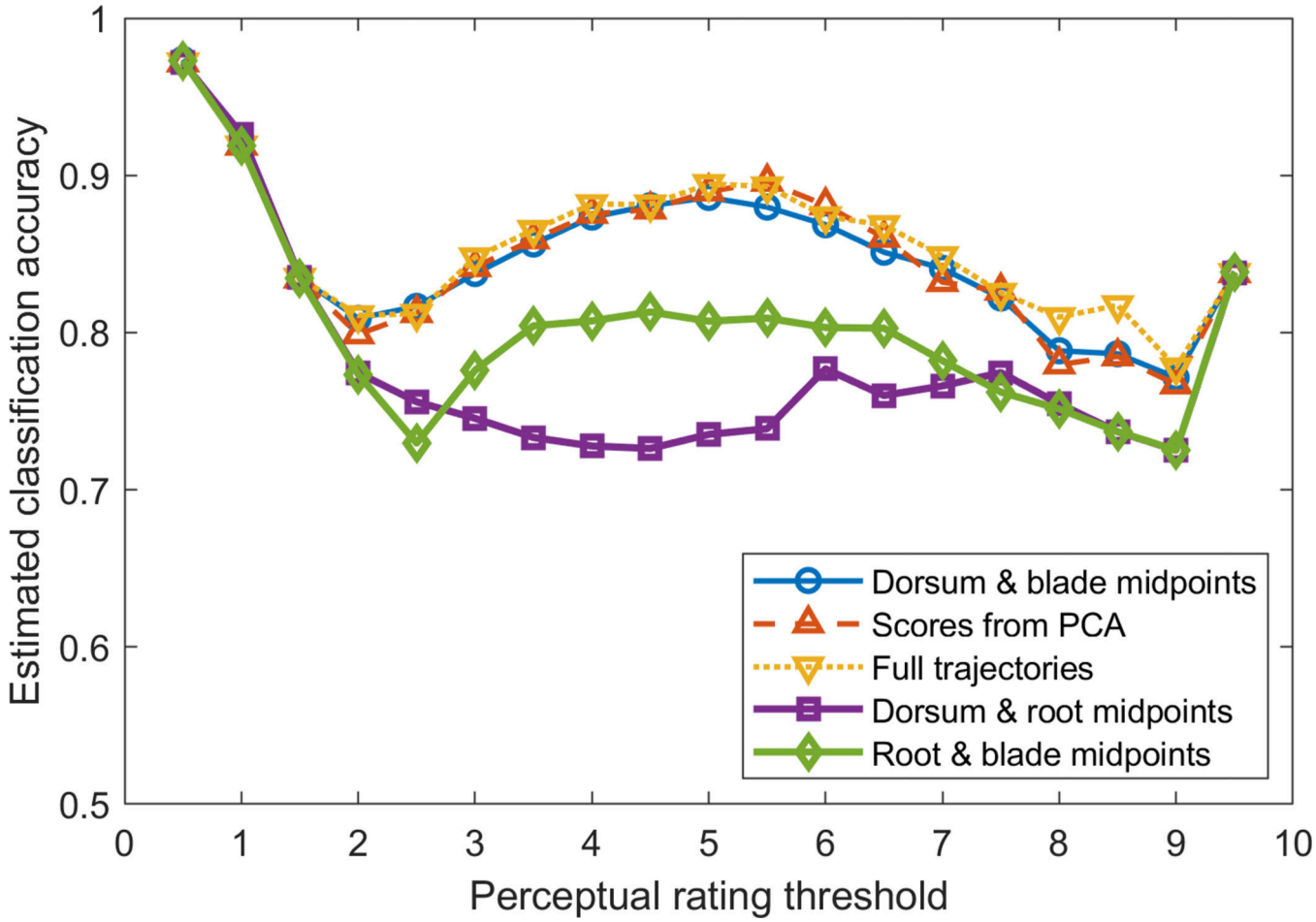

Figure 5:
Estimated classification accuracies resulting from Bayesian optimization of RBF SVM classifiers for the five candidate data representations, tuning the box-constraint and kernel-scale hyperparameters and using different perceptual ratings as the threshold between misarticulated and accurate classes.

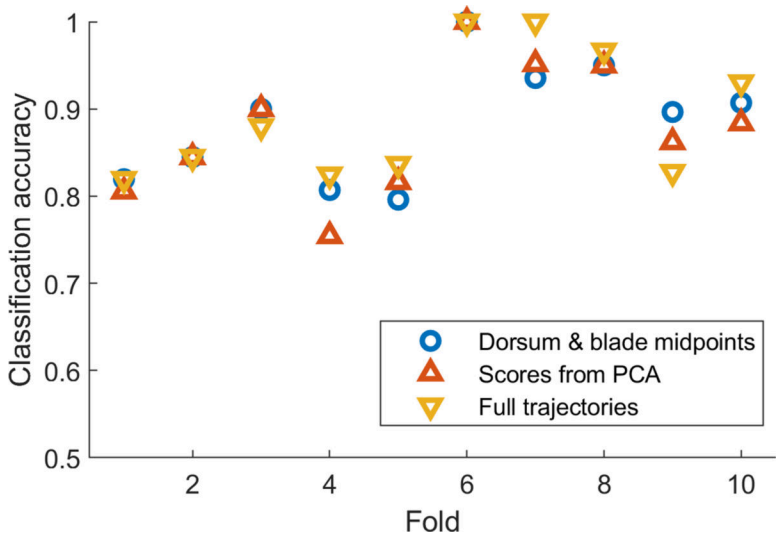

Figure 6:

Classification accuracies from RBF SVM classifiers with the same 10 cross-validation folds, for direct comparison of the three data representations with highest classification accuracies estimated from Bayesian optimization (Figure 5). As indicated by the Bayesian optimization runs, a perceptual rating threshold of 5.5 was used as well as the optimized box-constraint and kernel-scale hyperparameters.

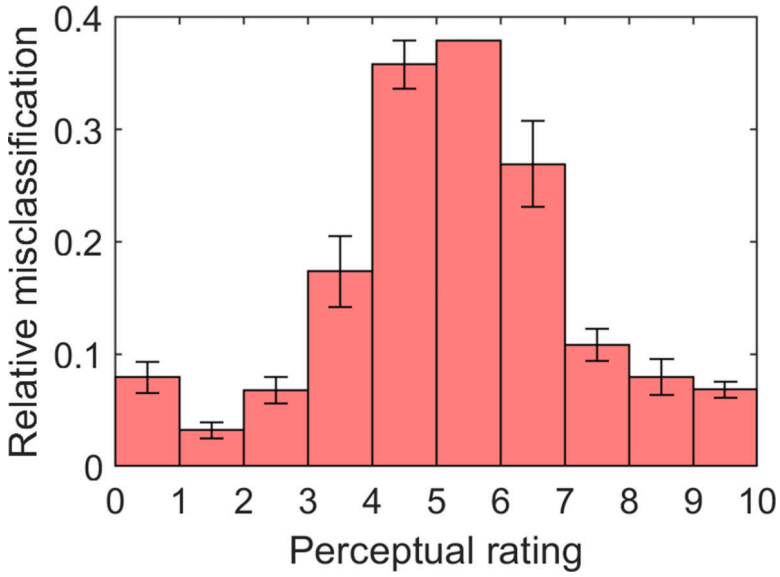

Figure 7:

Histogram of perceptual ratings for misclassifications made by the optimized RBF SVM classifiers (Figure 6). The bin heights represent the misclassification count relative to the count of productions (Figure 2) within each range of perceptual ratings, with the misclassification count averaged across the three different data representations and error bars indicating the standard deviation. For each of the three data representations, misclassification counts were combined from the 10 folds shown in Figure 6 (perceptual rating threshold = 5.5).

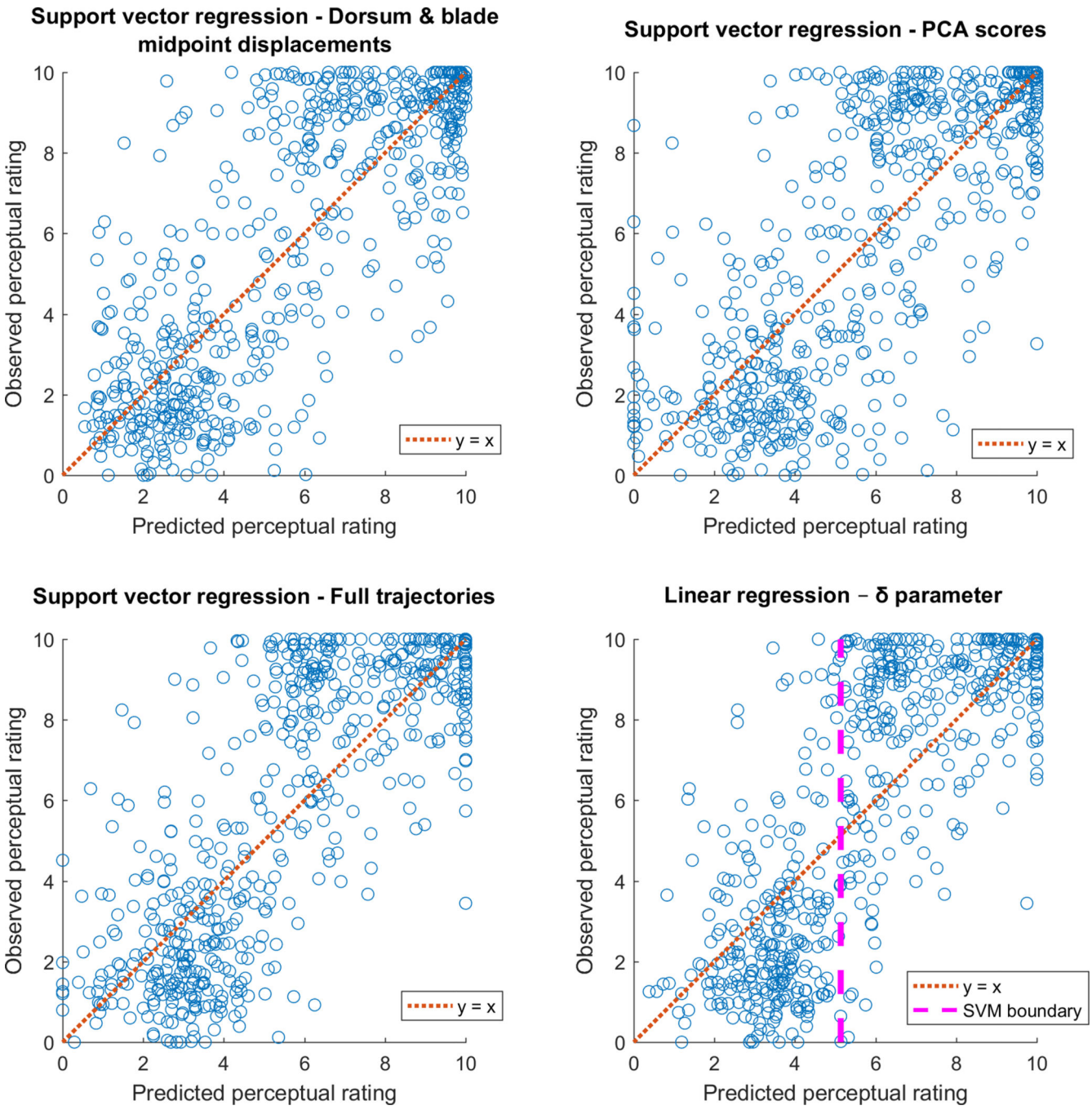

Figure 8:

Predicted perceptual ratings (horizontal axis) compared to observed perceptual ratings (vertical axis) for the support vector regression models and linear regression model trained on all productions. The dotted diagonal lines indicate the line $y = x$. For the lower right panel (linear regression on the δ parameter), the magenta dashed line indicates the predicted perceptual rating (graduated accuracy) for $\delta = 0$, which would occur for productions with

dorsum and blade displacements on the linear SVM boundary in the middle panel of Figure 4.

Table 1:

Demographic characteristics of speakers

	RSSD (*N* = 23)	**TD (*N* = 17)**
Age in years (*M* ± *SD*)	11.35 ± 2.52	12.23 ± 2.58
Number of female speakers	3	10
Number of male speakers	20	7