Published in final edited form as:
Clin Linguist Phon. 2016 ; 30(3-5): 328–344. doi:10.3109/02699206.2015.1099164.

Methods for Quantifying Tongue Shape and Complexity using Ultrasound Imaging

Katherine M. Dawson[*,1,2], **Mark K. Tiede**[2], and **D. H. Whalen**[1,2,3]

[1] City University of New York Graduate Center, USA

[2] Haskins Laboratories, New Haven, CT, USA

[3] Yale University, New Haven, CT, USA

Abstract

Quantification of tongue shape is potentially useful for indexing articulatory strategies arising from intervention, therapy and development. Tongue shape complexity is a parameter that can be used to reflect regional functional independence of the tongue musculature. This paper considers three different shape quantification methods – based on Procrustes analysis, curvature inflections, and Fourier coefficients – and uses a linear discriminant analysis to test how well each method is able to classify tongue shapes from different phonemes. Test data are taken from six native speakers of American English producing 15 phoneme types. Results classify tongue shapes accurately when combined across quantification methods. These methods hold promise for extending the use of ultrasound in clinical assessments of speech deficits.

Keywords

ultrasound; tongue shape; speech production measurement

Introduction

The shaping of the tongue during speech is a finely controlled motor activity involving the coordination of a complex array of intrinsic and extrinsic agonist-antagonist muscles (Takemoto, 2001). However, measuring and quantifying tongue shape has proved challenging, as the tongue, a muscular hydrostat, is a difficult structure to image. Nonetheless, such information is of potential use not just in fundamental research, but also for increasing understanding of motor control in normal and disordered speech.

A number of studies have suggested that speech acquisition disorders may be characterized by reduced regional functional independence of the tongue (Gibbon, 1999; Hardcastle, Barry, & Clark, 1987; Kent, Netsell, & Bauer, 1975), but quantitative measures of such

[*] Correspondence concerning this article should be addressed to: Katherine Dawson, Graduate Program in Speech-Language-Hearing Sciences, City University of New York, 365 5th Avenue, New York City, NY, 10016. Phone: 212-817-8842, kdawson2@gradcenter.cuny.edu.

Declaration of Interest
The authors report no conflicts of interest

impairments are lacking. One possibility for characterizing functional independence is tongue shape complexity. Complexity is a multi-faceted concept, but the operational definition within this work is the number (and extent) of inflections of the mid-sagittal tongue shape. The tongue shapes involved in the production of the phonemes /ɹ/ and /l/ (for example) are more complex than those involved in the production of vowels due to their multiple points of vocal tract constriction (Gick et al., 2008; Lawson, Stuart-Smith, Scobbie, Yaeger-Dror, & Maclagan, 2011; Zhou et al., 2008). Motor control of complex tongue shapes likely employs more degrees of freedom than are involved in producing simpler shapes (see figure 1 for ultrasound images of the vowel /ɑ/ and the rhotic consonant /ɹ/).

The aim of this study is to compare analysis methods for quantifying the complexity of mid-sagittal tongue shapes obtained from ultrasound data. Objective measures of complexity for providing the "ground truth" of our measures are difficult to find: although complexity can potentially be indexed motorically (cortical activation, number of muscle activations; Stavness, Gick, Derrick, & Fels, 2012), mathematically (Grassberger, 1986a, 1986b, 2012) and linguistically (Chitoran & Cohn, 2009), no definitions or scales exist for this particular purpose. Explicit discussions of articulatory complexity can be found in Kent (1992) and Stokes and Surendran (2005). The method we explore is to define, *a priori*, tongue shapes that represent high, medium and low complexity classes in American English, to see how well our chosen quantification methods (outlined in the data analysis section) perform in classifying across these broad groups. Based on our experience with mid-sagittal tongue shapes as seen in X-ray, MRI and ultrasound images, we divided the tongue shapes obtained in this study into three roughly equal groups. The low complexity group consists of /ɑ/, /æ/, /ɪ/, /ʌ/ and /ɛ/; the medium complexity group consists of /u/, /ɡ/, /w/, /j/; the high complexity group consists of /d/, /z/, /l/, /ɹ/, /θ/, /ʒ/. The rationale for these groupings is that unrounded vowels tend to have a single lingual constriction, requiring little in the way of secondary shaping. Sounds requiring additional tongue shaping, either driven by bracing as with /j/ or dorsal constriction as with /g/, are included in the medium group. The high complexity group includes tongue shapes requiring more than one constriction (/l/ and /ɹ/) as well as tongue shapes made with a tongue tip constriction. The rationale for this is that phonemes produced with tongue tip involvement are assumed to require a more fine-grained level of motor control than those only using the tongue body, and additionally require more in the way of shaping of the posterior tongue to support the tongue tip constriction (Bresch et al., 2008; Narayanan & Alwan, 1996). Fricatives are included in this group due to the precise nature of the constriction involved in producing frication.

Ultrasound data is suited to the goal of shape analysis, as it provides an image of most of the tongue surface, rather than a few locations as in point parameterization techniques such as electromagnetic articulography (EMA) or X-ray microbeam methods. It is also a safe, noninvasive and relatively easy technique to perform, and is hence suitable for use with clinical populations and children, who are often of interest in studying speech motor control. Ultrasound is increasingly being used for clinical remediation purposes, specifically to provide biofeedback for speech disorders (Bernhardt, Gick, Bacsfalvi, & Adler-bock, 2005; Bernhardt, Bacsfalvi, Gick, Radanov, & Williams, 2005; Hitchcock & McAllister Byun, 2015; McAllister Byun & Hitchcock, 2012). Clinical use of ultrasound often does not require any quantification measures, as the intent is to provide real-time information on tongue movement and shape. However, for research purposes and to help us understand the nature of tongue function, it is useful to apply measures that tell us something about the mathematical attributes of the tongue shape, either intrinsically or with regard to a reference shape. Existing research has attempted to index the tongue shape and/or contact pattern differences exhibited by certain clinical populations. For example, Zharkova (2013) developed methods to quantify abnormal articulatory patterns in cleft palate speakers and Klein, McAllister Byun, Davidson, & Grigos (2013) applied quantitative measures to analyse ultrasound images of children undergoing therapy for /ɹ/ misarticulations. The aim of the current study is to adapt existing methods for the purpose of quantifying tongue shape complexity, without reference to a single clinical population.

In this study, ultrasound images of the tongue were analysed using three distinct methods, and the output (metrics) associated with each method were compared using linear discriminant analysis to classify shape type and complexity. In this we employ a methodology similar to that of Wang, Green, Samal and Yunusova (2013) who used Procrustes analysis and support vector machine classification to quantify articulatory distinctiveness among vowels and consonants using EMA data collected from ten typical speakers. However, the current study does not use any temporal or kinematic information and relies solely on the midpoint or maximal constriction image of each target sound.

The quantification of ultrasound data of the tongue presents a number of challenges. First, images obtained from ultrasound are of varying quality, which can make the tongue surface difficult to identify reliably. Optimizing tongue contour extraction is beyond the scope of this paper, but it is important to take this issue into account during data analysis, in that poor image quality may result in inaccurate contour identification. Also, unless the locations of the head and jaw are stabilized or tracked relative to probe position, the position of the tongue relative to palatal hard structures is impossible to determine accurately. However, measurements of shape are well-suited to ultrasound data, as shape-related variables can still be measured reliably in the absence of objective spatial information (Ménard, Aubin, Thibeault, & Richard, 2012; Stone, 2005).

Previous Approaches to Quantification

A number of methods of quantifying tongue image data exist in the linguistic and clinical speech science literature, some based on shape and some on tongue position. One of these is the smoothing spline (SS) ANOVA (Davidson, 2006). The SS ANOVA evaluates the variance in the data (in the form of sets of tongue surface contours) that is accounted for by membership in a particular group (e.g. onset /z/ versus coda /z/). Confidence intervals around group splines demonstrate locations at which the tongue contours can be considered statistically different or the same at the chosen confidence level. This method requires that head and jaw movement relative to the probe is constrained or corrected over comparable groups. If this is not the case, the SS ANOVA is inappropriate.

Another method, one used to describe tongue shape directly, focuses on fitting a triangle to the tongue contour and derives the extent and location of maximal inflection using the properties of the triangle (Aubin & Ménard, 2006). This method is simple to apply and is useful for describing vowel shapes. However, for tongue shapes with multiple inflections it is problematic, as multiple triangles are necessary and the demarcation of where they are applied may not be clear.

A number of investigators have used methods based on factor analysis or principle component analysis (PCA) to quantify tongue position (Harshman, Ladefoged, & Goldstein, 1977; Hoole et al., 2000; Slud, Stone, Smith, & Goldstein Jr, 2002). PCA and factor analysis provide dimensional reduction of shapes to a sparse set of salient features, typically related to properties of the tongue such as vertical and horizontal position in the mouth (tongue height and backness). Harshman et al. (1977) found that two factors accounted well for the variation in their tongue shape data taken from X-ray images of vowel productions. A potential problem with a PCA analysis is that the results are not always readily interpretable. When considering tongue shapes from the perspective of clinical disorders of motor control, it is preferable to have interpretable factors so that a remediation strategy can be suggested. However, PCA does provide a useful index of data complexity, by indicating what subset of the total number of potential components are sufficient to account for some threshold percentage of the data variability. To this end, we have applied a PCA to a subset of our data in the current study, in order to compare the efficiency of our proposed analysis methods to this well-known approach.

Another set of methods that have been used previously in ultrasound studies of tongue shape are based on polynomial fitting (Morrish, Stone, Sonies, Kurtz, & Shawker, 1984; Morrish, Stone, Shawker, & Sonies, 1985). A polynomial fit can potentially be used to index complexity, as a more complicated shape requires a higher order polynomial (as in Morrish et al.,1985). The derivatives of fitted polynomials can also be used to index the curvature of the shape, with higher curvature values indicating higher complexity. A recently published method for evaluating curvature based on polynomial fitting is the Curvature Index (CI) metric developed by Stolar and Gick (2013). This method fits a 7th order polynomial to the entire tongue surface (represented by the discrete contour points extracted from the ultrasound image). The integral of the absolute curvature (the reciprocal of the radius of curvature) is then taken with respect to the horizontal offset along the curve. Stolar and Gick (2013) found higher CI values for ‘complex’ tongue shapes (particularly liquids) compared to other tongue shapes.

A modified version of this method is included in the current analysis, together with a method based on the Procrustes fit between each tongue shape and a resting tongue shape, and also a method based on transforming tongue shapes into the spatial frequency domain using a discrete Fourier transform. These analysis methods will be described in more detail in the data analysis section below, but were all chosen for their potential to give information not just in terms of distinguishing shapes of different types (although this is an important consideration), but also to provide a quantifiable index as to how complex a given tongue shape may be to produce. In addition, these analysis methods share the useful property of being relatively insensitive to probe rotation and its effects on tongue surface orientation.

We expect that each of the three methods used for shape analysis in this study will yield information on different characteristics of the shapes (e.g. presence of single vs. multiple inflections, extent of curvature of inflections, etc.) and that their predictive power in terms of classification can be assessed by including the resulting metrics in a linear discriminant analysis. We also hypothesize that the qualitative patterns seen among participants will be similar, although quantitatively the metric values are likely to differ considerably, due to inter-individual variations in physiology and articulatory strategy.

Method

Participants

The participants in this study were six native speakers of American English (three female and three male). They were between 24 and 45 years of age, with no reported history of neurological disorders or disease, or any speech, language or hearing difficulties. The study was approved by the institutional review board (IRB) of the City University of New York Graduate Center. Participants provided informed consent, and were compensated for their time.

Stimuli

Stimuli consisted of symmetrical CVC / VCV utterances. The stimuli used in this study were: /ɑ/, /ɪ/, /ɛ/, /æ/, /ʌ/, /u/ in a /bVb/ context, and /ɡ/, /d/, /z/, /ʒ/, /l/, /ɹ/, /j/, /w/, /θ/ in a /ɑCɑ/ context. Representative images of the resting (pre-phonatory) tongue shape were also collected for each participant. These images were taken when the tongue lay flat in the mouth, with no palate contact, and were used as a ‘baseline’ shape for comparison with tongue shapes produced during speech in the Procrustes analysis method.

Procedure

Ultrasound images of the tongue were obtained with an Ultrasonix SonixTouch system (Richmond, BC, Canada) using a C9-5/10 microconvex transducer. Ultrasound data were collected at a frame rate of 60Hz. During the experimental session the participant was seated with their head unrestrained. The reason no head restraint or correction was used is that the analysis methods under development are intended to be suitable for use on data taken from clinical populations and children, for whom restraint and head correction procedures are not always possible or desirable. However, the consideration remains that if the participant is not reasonably still, the areas of the tongue imaged (including the endpoints) will vary, which may affect the analyses. For a comparison of head-stabilized versus non-stabilized data see Zharkova, Gibbon, & Hardcastle (2015).

The ultrasound probe was held using an adjustable, heavy-duty metal stand. The probe arm of the stand was spring-loaded, so that it moved in the vertical plane with the motion of the jaw. The probe was positioned in a mid-sagittal alignment under the participant's chin. The experimental task was to produce the stimuli displayed on a large computer monitor positioned approximately 1m in front of the participant. The stimuli were presented using Presentation software (“Presentation,” 2013). This programmable software was used to send a triggering signal to the ultrasound system to initiate video capture synchronously with presentation of each stimulus, while simultaneously recording audio data from a microphone connected to the computer. The ultrasound machine produced an audible sound at the end of trials, which was recorded as part of the acoustic data. To verify synchronization, the experimenter reviewed the ultrasound data by hand to ensure that the areas of maximal excursion of the tongue in producing constrictions were aligned with acoustic landmarks.

Participants were asked to rehearse the list of stimuli prior to the ultrasound session, in order to correct any errors in pronunciation and stress pattern (stress was elicited on the second vowel of the VCV utterances). A short familiarization period then preceded the main task, predominantly to acquaint the participant with speaking whilst in the experimental setup. If the participant expressed discomfort during the familiarization period, adjustments were made to the probe and / or seating position prior to initiation of the main task.

The stimuli were presented in a randomised order, which was the same for each participant. The participant was asked to repeat the word displayed on the screen, at a normal rate, until the end of the trial (indicated by a beep sound produced by the ultrasound). Each trial lasted 8 seconds, during which most participants produced at least six tokens of the same stimulus item. The stimulus list was repeated twice (with a different order the second time, again consistent across participants), yielding two trials (an average of 12 repetitions) per stimulus item. Participants were instructed to maintain contact with the ultrasound probe under their chin and to move as little as possible aside from speaking.

Data Analysis

Analysis was performed on frames from the midpoint or maximal constriction of the target sounds. For the vowels, the selection was based on the acoustic midpoint of the vowel. For the consonants, the acoustic data was used to identify the constriction, and the ultrasound frame containing the largest displacement from the probe was selected.

Six tokens of each stimulus, for each participant, were used for analysis. Tokens were selected using the following process: the first and last token from each trial were excluded, as were any tokens where the line indicating the air tissue interface between the tongue and the oral cavity was unclear. Images were also excluded if the utterance included any disfluencies or swallows. Following this, the first, second and third remaining images from the first trial and from the second trial were selected, yielding six images per stimulus item. One shape from the pre-phonatory state was also extracted.

The next step in the analysis was to fit a contour to the tongue shape. The contour consisted of a series of *x-y* coordinate points that represent the tongue edge and were extracted for further analysis. For the contour fitting procedure, we used a custom interactive MATLAB procedure (‘GetContours’, written in-house at Haskins laboratories) to position anchor points controlling the position of a cubic spline fit to the tongue surface, similar in functionality to Edgetrak (Li, Kambhamettu, & Stone, 2005). One hundred equally-spaced points defining each contour were automatically exported.

Quantification Methods Used in this Study

Three shape analysis methods were tested here: modified curvature index (MCI), Procrustes analysis and the discrete Fourier transform (henceforth DFT). Each was applied using custom procedures written in Python (“Python Language” 2013). These procedures, including test data and worked examples, are freely available at: https://github.com/kdawson2/tshape_analysis.

Modified curvature index (MCI)—This method is an extension of the curvature index (CI) of Stolar and Gick (2013). The original (1) and modified (2) equations for this method are shown below, where κ is curvature, a and b are the start and end of the tongue contour on the *x*-axis and α and β are the start and end of tongue contour along the arc length.

Equation 1 - CI (Stolar & Gick, 2013):

$$\int_a^b |\kappa| \quad dx$$

Equation 2 - MCI (current study):

$$\int_\alpha^\beta |\kappa| \quad ds$$

The first modification is that we integrated curvature values with respect to the arc length of the tongue (the length of the curve if it were stretched out into a straight line – *s* in equation 2), rather than the *x*-axis. This is because the *x*-axis integration introduces differences in the metric if the ultrasound probe is rotated. In our original efforts in extending this method, a piecewise polynomial was used. The primary use of polynomial fitting is to extract curvature values that are smoother and less noisy than the raw data. However, higher-order polynomials can introduce unwanted effects at the boundaries of the shape (Runge's phenomenon), and so in the current iteration of this technique polynomial fitting was not used at all. Instead, we used central differencing of the shape coordinate points (that is, half the difference at each point between the preceding and following points) to compute the derivatives and a low pass (5th-order) Butterworth filter to remove noise. This gives very similar answers to the original method but with less computational burden and represents a more explicit, and therefore controllable, filtering process.

Procrustes analysis—This method is based on the Procrustes distance between tongue shapes during speech and the resting (pre-phonatory) shape of the tongue. Procrustes analysis is a form of statistical shape analysis (Goodall, 1991) that attempts to superimpose two shapes and computes a metric to describe any remaining difference between them. The superimposition of two shapes in Procrustes analysis involves translation, rotation and scaling (compression or expansion) of the shape, in order to minimize the sum of squared differences (SSD). The metric is the remaining SSD between all the points on the two shapes. The output is a single number, with a higher number representing a greater difference between two shapes than a lower number.

Discrete Fourier transform (DFT)—The third method included here is based on a DFT of the tongue shapes, where the shapes are transformed into the spatial frequency domain. For this we adapted a method described in Liljencrants (1971). The differences between Liljencrant's method and our own are that the Liljencrants method was based on X-ray images and a fixed coordinate system, and the function transformed was the tongue shape relative to this coordinate system. The *x* and *y* axes of this coordinate system were defined relative to hard structures in the vocal tract. In our analysis, we do not have any static reference points for a coordinate system. Hence, we transform the tangent angle values for each point on the shape as a function of arc length. The DFT procedure can be conceptualized as obtaining a measure of correlation between a function (in this case the tangent angle) and sine and cosine waves of increasing frequency, the first coefficient relating to waves with wavelength equal to the full arc length, and higher coefficients relating to multiples of this frequency.

The output of the DFT analysis for our purposes is the value of a given shape provided by the first three Fourier coefficients. The first coefficient of the Fourier transform (C1) corresponds to the largest-scale features of the shape. The higher coefficients reflect smaller scale features. Very high coefficients represent small variations in the shape, often introduced during the contour fitting procedure (i.e. noise). Hence, coefficients above the third (C3) are not included in this analysis. Each coefficient has a real and an imaginary part, and the location of a data point on these real and imaginary axes can be described in radial coordinates, giving the corresponding phase and magnitude of the coefficient.

Assessment via Linear Discriminant Analysis

A linear discriminant analysis was performed to examine how well the methods differentiated the complexity groups. Linear discriminant analysis (LDA) is a classification technique for multivariate data. The aim is to find a set of transformed variables, made up of a linear combination of the original variables, which maximizes the between-class separation while simultaneously minimizing the within-class variance. These linear combinations are known as discriminant functions. The number of useful discriminant functions that can be found for a given data set depends on the number of variables and the number of classes. A classifier is then built using the values of the discriminant functions as new variables for each observation (e.g. Burns & Burns, 2008; Welling, 2005). The method used here was the ‘lda’ function within the MASS package of R (Ripley, Hornik, Gebhardt, & Firth, 2013).

The success of the classification can be most simply shown in the apparent error rate, which is the result of applying the classifier to the training data. However, this is likely to be an optimistic estimate of the error. A better measure can be obtained by separating the data into training and test sets, but this is problematic when the data set is relatively small. A good compromise, and the method used here, is a ‘leave-one-out’ cross-validation, where the classifier is constructed with one data point (i.e. the values for one tongue shape) missing, and then the missing data point is used to judge the success or failure of the classification (Burns & Burns, 2008). This is done iteratively until all data points have been tested. We used the LDA to quantify classification rates for each measure among three complexity groups: low (ɑ/, /ɪ/, /ɛ/, /æ/, /ʌ/) medium (/u/, /ɡ/, /w/, /j/) and high (/d/, /z/, /l/, /ɹ/, /θ/, /ʒ/). We also used the LDA to show classification rates among individual phoneme types, for all participants.

Results

The analysis of the tongue shapes resulted in a single value for the MCI and for the Procrustes metrics, and six values (real and imaginary parts of three coefficients) for the DFT method. These numbers can be plotted in a pairwise fashion on a graph, to see how they cluster into shape categories. As an illustration, graphs are shown for a single participant, and results of the shape complexity and phoneme classification (LDA) are shown subsequently for all participants.

Figure 2 shows the values for each tongue shape using the MCI (*x* axis) and Procrustes metric (*y* axis). These metrics are combined on to one graph for ease of interpretation (and to see the extent to which they correlate). This is not essential, and the metrics can be used in isolation. Each point on the graph represents a single tongue shape; the phoneme type is indicated by the symbol. The figure also includes one standard deviation confidence ellipses, drawn around the phoneme classes. Although individual phonemes are not well separated, certain clusters represent common classes of sounds. For example, the vowels are largely clustered in the bottom left of the figure, indicating that they are similar to each other but different from the consonants. The vowel /u/ looks more like a consonant in this regard, due to its constriction near the soft palate.

Figure 3 shows the values for each tongue shape using the real (*x* axis) and imaginary (*y* axis) parts of the DFT first coefficient (C1). The DFT first coefficient separates the phoneme classes into relatively discrete groups.

Figures 4 and 5 show the LDA classification results for our *a priori* defined complexity groups, for the same participant as shown in the previous figures. Figure 4 shows the LDA classifications using the Procrustes metric and the MCI. Figure 5 shows the classifications using the DFT first coefficient (C1, real and imaginary parts). The symbols in both graphs indicate the phoneme type. Symbols in black are correctly classified; symbols in red are incorrectly classified.

For the classification using the LDA, C1 provides the best level of separation among the groups. This is apparent from visual inspection of the graphs, and is confirmed by the discriminant analysis in the next section, in that the DFT C1 misclassifies fewer data points than the MCI and Procrustes methods. The Procrustes metric and MCI provide a good level of separation between the low versus the medium and the high groups, but a considerable number of misclassifications among the medium and high groups. Table 1 shows the classification results across metrics and across participants (as well as for the metrics and the participants individually). The ‘ALL’ column shows the success rate of each metric across all participants; the ‘ALL’ row shows the success rate within each participant across all metrics. The first DFT Coefficient (C1, imaginary part) is the best single parameter classifier for delineating these complexity groups, with a success rate of 0.77 across all participants, rising to 0.81 when combined with the C1 real part. The Procrustes and MCI metrics are less successful, with a combined rate of 0.61. The Procrustes metric appears more successful than the MCI, and in fact the combined success is almost identical to using the Procrustes metric alone. Classification success rates within participants are considerably higher than those across all participants. Rates within participants using all metrics range from 0.87 to 1.0. When using all participants, after classifying using C1 (real and imaginary), addition of more metrics (including higher coefficients) does not improve classification substantially.

Table 2a shows classification rates from the LDA by phoneme type rather than complexity group, across all participants and using all metrics. The rows represent the target phoneme and the columns the predicted phoneme. Thus, the diagonal values (in bold) indicate for each phoneme how many times it was classified as itself (i.e. correctly) out of a possible 36 instances. The shaded cells represent the complexity classes. Table 2a also shows the error pattern, i.e. the pattern of misclassifications. The most common errors were within complexity classes – this is shown more succinctly in table 2b. Outside the complexity classes the most common misclassifications were /j/ being classified as /ɪ/, /ʒ/ as /æ/ (and vice versa), /w/ as /ɑ/ and /ʌ/ and /u/ as /ɪ/. Other than /ʒ/ as /æ/, these confusions make sense articulatorily. The highest classification rates were for /ɹ/ (33/36), /ɪ/ (29/36) and /l/ (28/36).

Assessment via Principal Component Analysis

We performed a PCA on data from one speaker to provide a basis for comparison between a standard approach for quantifying variance and the methods discussed here (a single speaker was used to avoid the additional complexity of cross-speaker variability). The dataset was the same set of contour shapes analysed above, converted to curvature as for the MCI method, consisting of six repetitions of 15 distinct utterances. Results showed that seven principal components were necessary to account for 81% of the variance. A plot projecting PC1 vs. PC3 (figure 6), the pairing showing best separation, suggests that PCA is less parsimonious and less effective in separating phonemes than the DFT approach; nor does it form a continuum of responses along which complexity can be indexed, as with the MCI and Procrustes approaches.

Inter-rater reliability

To assess inter-rater reliability of the methods, all the data for one participant (90 tongue shapes) was tracked by two different experimenters. We estimated inter-rater reliability using the R^2 correlation coefficient of the data from both raters (see table 3). The R^2 values indicate that the MCI seems to be the method most sensitive to noise introduced by having different experimenters track the same tongue shapes. The other metrics are more resistant to this, particularly the DFT first coefficient.

Discussion

The aim of this study was to assess three types of shape analysis methods, in terms of their success and suitability for quantifying tongue shape complexity and for identifying phonemes by tongue shape. We applied the three methods – the DFT, Procrustes Analysis and MCI to ultrasound data of mid-sagittal tongue contours taken from six typical speakers across 15 American English phoneme types. The success of the methods was quantified using a linear discriminant analysis to classify across phoneme type and into three pre-defined levels of complexity – high, medium and low.

The DFT first coefficient was the best classifier, both of complexity groups and (as indicated by figure 3) of phoneme types. One drawback of this method is that it does not form a continuum that can be easily interpreted in terms of complexity. The low complexity group borders the high complexity group, and ideally this would not be the case as low can potentially be confused with high, rather than with just medium. The DFT is also not easily interpretable from a motor control point of view without a coordinate system constructed relative to vocal tract anatomy. In Liljencrants (1971), where such a coordinate system did exist, the magnitude of a data point (distance from *0,0* origin) in a coefficient corresponded to degree of constriction and the phase to the place of constriction. Inspection of figure 3 does show a rough trend for magnitude to be related to degree of inflection, but only for the tongue body, as the first coefficient of the Fourier transform reflects the most salient, or largest-scale features of the shape (i.e. shapes with a single, extreme inflection such as /u/, /g/ /j/ and /w/ have high magnitude values). It is also possible that the low magnitude values for the more complex shapes mean that they have more energy distributed in higher coefficients (i.e. cannot be described well using a low frequency sinusoid). Inclusion of higher coefficients (C2 and C3) did not greatly improve the classification rates, at least in this analysis.

The results for the Procrustes and MCI methods indicate that they are potentially useful, as they form a continuum, where low complexity shapes are less likely to be confused with high. Measures of curvature and distance from a resting shape also relate more intuitively to the concept of complexity than the shape parameters indexed by the DFT. However in terms of classification they do not perform as well as the DFT method. The MCI has a lower level of consistency across experimenters than the other methods, which implies that it is more sensitive to noise introduced by the contour tracking process. The LDA analysis using all three methods was slightly more successful than the one using only the DFT, but the amount of information gained is presumably not justified by the added degrees of freedom in the input.

Across all methods, the within-participant classification rates were higher than across participants. This was expected and most likely due to differences in vocal tract morphology among participants and also differences in mid-sagittal tongue shape formation strategy during speech. At this point therefore, these methods are most valid for within-person comparisons, such as tracking articulation over time for a single participant. Future work will include exploration into ways to make across-person comparisons more useful.

Caveats to this work include the fact that some tongue shapes are often less reliably imaged than others. Although this was not a particular issue in the current study, a general point is that if most of the complexity for a shape is located in areas that often cannot be successfully imaged (such as the tongue tip), then the analysis may give an erroneous indication of complexity and shape type. The measures for this study were also taken from repeated tokens of simple utterances. Analysing tongue shapes from running speech would be a natural extension of this work, but one that would likely produce even more complicated results.

Another key point is that complexity of lingual articulation does not just relate to mid-sagittal tongue shape. Timing, precision of constriction, coronal tongue shape, tongue bracing and involvement of other articulators are all parameters than can feed into complexity. However, measurement of these additional parameters, as well as techniques that involve imaging hard structures, often involves an increase in the response burden of the participant and also the instrumentation and analysis burden on the researcher. For clinical purposes, ultrasound and the use of shape-related parameters constitute a relatively practical experimental paradigm. The trade-off is that shape measures are somewhat less informative than those that include positional or hard-structure related data. However, the results of this study demonstrate that shape measures can (or can be adapted to) provide useful information on complexity-related control of the tongue, which can potentially be informative from a clinical perspective.

Acknowledgements

This work was supported by National Institutes of Health Grant DC-002717 to Haskins Laboratories. We thank Hannah King for technical assistance and two anonymous reviewers for helpful comments.

References

Aubin J, Ménard L. Compensation for a labial perturbation : An acoustic and articulatory study of child and adult French speakers. 7th International Seminar on Speech Production. 2006:209–216.

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

Bernhardt B, Bacsfalvi P, Gick B, Radanov B, Williams R. Exploring the use of electropalatography and ultrasound in speech habilitation. Journal of Speech-Language Pathology and Audiology. 2005; 29(4):169–182.

Bresch E, Riggs D, Goldstein L, Byrd D, Lee S, Narayanan S. An analysis of vocal tract shaping in English sibilant fricatives using real-time magnetic resonance imaging. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2008:2823–2826.

Burns, RP.; Burns, R. Chapter 25: discriminant analysis.. In: Sage. , editor. Business research methods and statistics using SPSS. London: 2008. p. 589-608.

Chitoran, I.; Cohn, AC. Complexity in phonetics and phonology: gradience, categoriality, and naturalness.. In: Pellegrino, F.; Marsico, E.; Chitoran, I.; Coupé, C., editors. approaches to phonological complexity. Mouton de Gruyter; New York: 2009. p. 23-46.

Davidson L. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. The Journal of the Acoustical Society of America. 2006; 120(1):407–415. [PubMed: 16875236]

Gibbon FE. Undifferentiated lingual gestures in children with articulation / phonological disorders. Journal of Speech, Language and Hearing Research. 1999; 42(2):382–397.

Gick B, Bacsfalvi P, Bernhardt BM, Oh S, Stolar S, Wilson I. A motor differentiation model for liquid substitutions: English /r/ variants in normal and disordered acquisition. In Proceedings of Meetings on Acoustics. 2008; 1:1–9.

Goodall C. Procrustes Methods in the Statistical Analysis of Shape. Journal of the Royal Statistical Society. 1991; 53(2):285–339.

Grassberger P. How to measure self-generated complexity. Physica A: Statistical Mechanics and Its Applications. 1986a; 140(1):319–325.

Grassberger P. Toward a quantitative theory of self-generated complexity. International Journal of Theoretical Physics. 1986b; 25(9):907–938.

Grassberger P. Randomness, information, and complexity. arXiv Preprint arXiv. 2012; 1208.3459

Hardcastle WJ, Barry RM, Clark CJ. An instrumental phonetic study of lingual activity in articulation-disordered children. Journal of Speech, Language, and Hearing Research. 1987; 30(2):171–184.

Harshman R, Ladefoged P, Goldstein L. Factor analysis of tongue shapes. Journal of the Acoustical Society of America. 1977; 62(3):693–707. [PubMed: 903511]

Hitchcock ER, McAllister Byun T. Enhancing generalisation in biofeedback intervention using the challenge point framework: A case study. Clinical Linguistics & Phonetics. 2015; 29(1):59–75. [PubMed: 25216375]

Hoole P, Wismueller A, Leinsinger G, Kroos C, Geumann A, Inoue M. Analysis of the tongue configuration in multi-speaker, multi-volume MRI data. Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling. 2000:157–160.

Seeon, Kloster; Germany; Kent, RD. The biology of phonological development.. In: Ferguson, C.; Menn, L.; Stoel-Gammon, C., editors. Phonological development: Models, research, implications. York Press; Timonium, MD: 1992. p. 65-90.

Kent RD, Netsell R, Bauer LL. Cineradiography assessment of articulatory mobility in the dysarthrias. Journal of Speech and Hearing Disorders. 1975; 40(4):467–480. [PubMed: 1234962]

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

Lawson, E.; Stuart-Smith, J.; Scobbie, J.; Yaeger-Dror, M.; Maclagan, M. Liquids.. In: Di Paolo, M.; Yaeger-Dror, M., editors. Sociophonetics: a student's guide. Routledge; London: 2011. p. 72-86.

Li M, Kambhamettu C, Stone M. Automatic contour tracking in ultrasound images. Clinical Linguistics & Phonetics. 2005; 19(6-7):545–554. [PubMed: 16206482]

Liljencrants J. Fourier series description of the tongue profile. Speech Transmission Laboratory-Quarterly Progress Status Reports. 1971; 12(4):9–18.

McAllister Byun T, Hitchcock ER. Investigating the Use of Traditional and Spectral Biofeedback Approaches to Intervention for /r/ Misarticulation. American Journal of Speech-Language Pathology. 2012; 21(3):207–222. [PubMed: 22442281]

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

Morrish KA, Stone M, Shawker TH, Sonies BC. Distinguisability of tongue shape during vowel production. Journal of Phonetics. 1985; 13(2):189–203.

Morrish KA, Stone M, Sonies BC, Kurtz D, Shawker T. Characterization of tongue shape. Ultrasonic Imaging. 1984; 6(1):37–47. [PubMed: 6540910]

Narayanan S, Alwan A. Imaging applications in speech production research. Proc. SPIE 2709, Medical Imaging 1996: Physiology and Function from Multidimensional Images. 1996; 2709:120–131.

Presentation. Neurobehavioral systems; Albany, CA: 2013. Retrieved from www.neurobs.com

Python Language. Python Software Foundation; 2013. Retrieved from www.python.org

Ripley, B.; Hornik, K.; Gebhardt, A.; Firth, D. Package “MASS.”. 2013. Retrieved from http://cran.r-project.org/web/packages/ MASS/MASS.pdf

Slud E, Stone M, Smith P, Goldstein M Jr. Principal Components Representation of the Two-Dimensional Coronal Tongue Surface. Phonetica. 2002; 59(2-3):108–133. [PubMed: 12232463]

Stavness I, Gick B, Derrick D, Fels S. Biomechanical modeling of English /r/ variants. The Journal of the Acoustical Society of America. 2012; 131(5):EL355–60. [PubMed: 22559452]

Stokes SF, Surendran D. Articulatory Complexity, Ambient Frequency, and Functional Load as Predictors of Consonant Development in Children. Journal of Speech, Language and Hearing Research. 2005; 48(3):577–592.

Stolar S, Gick B. An index for quantifying tongue curvature. Canadian Acoustics. 2013; 41(1):11–15.

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

Takemoto H. Morphological Analyses of the Human Tongue Musculature for Three- Domensional Modeling. Journal of Speech, Language and Hearing Research. 2001; 44(1):95–107.

Wang J, Green JR, Samal A, Yunusova Y. Articulatory Distinctiveness of Vowels and Consonants: A Data-Driven Approach. Journal of Speech, Language and Hearing Research. 2013; 56(5):1539–1552.

Welling, M. Fisher linear discriminant analysis. Toronto: 2005. Retrieved from http://www.cs.huji.ac.il.ezproxy.gc.cuny.edu/~csip/Fisher-LDA.pdf

Zharkova N. Using ultrasound to quantify tongue shape and movement characteristics. The Cleft Palate-Craniofacial Journal. 2013; 50(1):76–81. [PubMed: 22117937]

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

Zhou X, Espy-Wilson CY, Boyce S, Tiede M, Holland C, Choe A. A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. The Journal of the Acoustical Society of America. 2008; 123(6):4466–4481. [PubMed: 18537397]

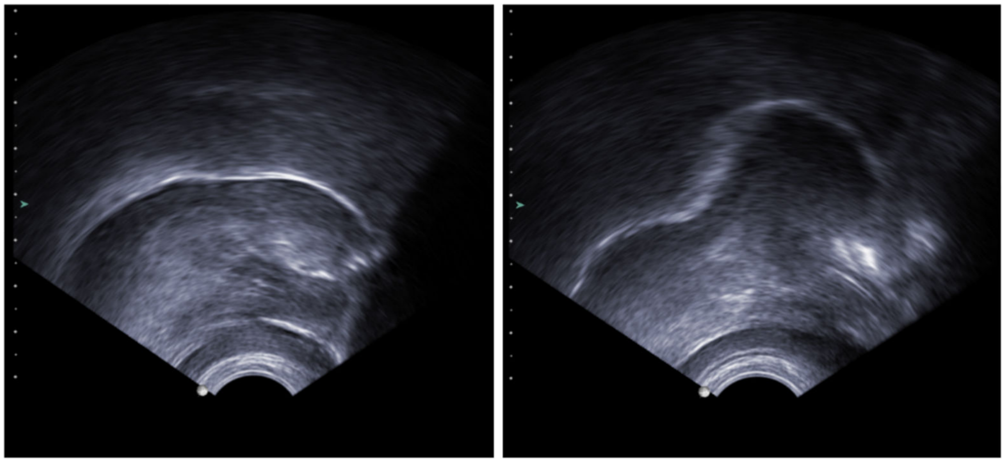

Figure 1.

Ultrasound images in the mid-sagittal plane of the tongue during the midpoint of the vowel /ɑ/ (left panel) and the maximal constriction of the consonant /ɹ/ (right panel) produced by an American English speaker. The tongue tip is to the right in both images.

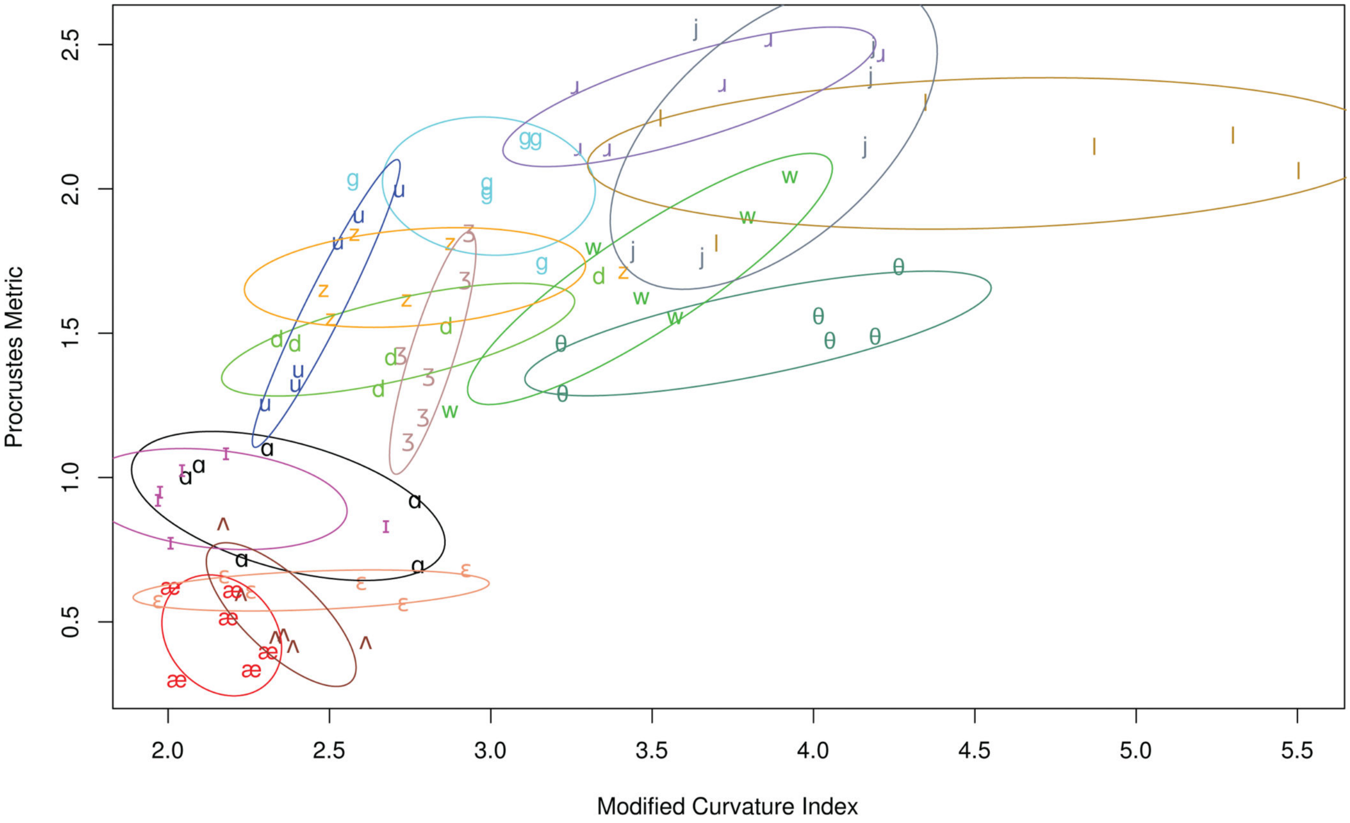

Figure 2.

MCI and Procrustes metric results for participant 01_FC (female typical speaker). The MCI is plotted along the *x* axis and Procrustes metric along the *y* axis. Phoneme types are indicated by the symbol. One standard deviation confidence ellipses surround the phoneme classes.

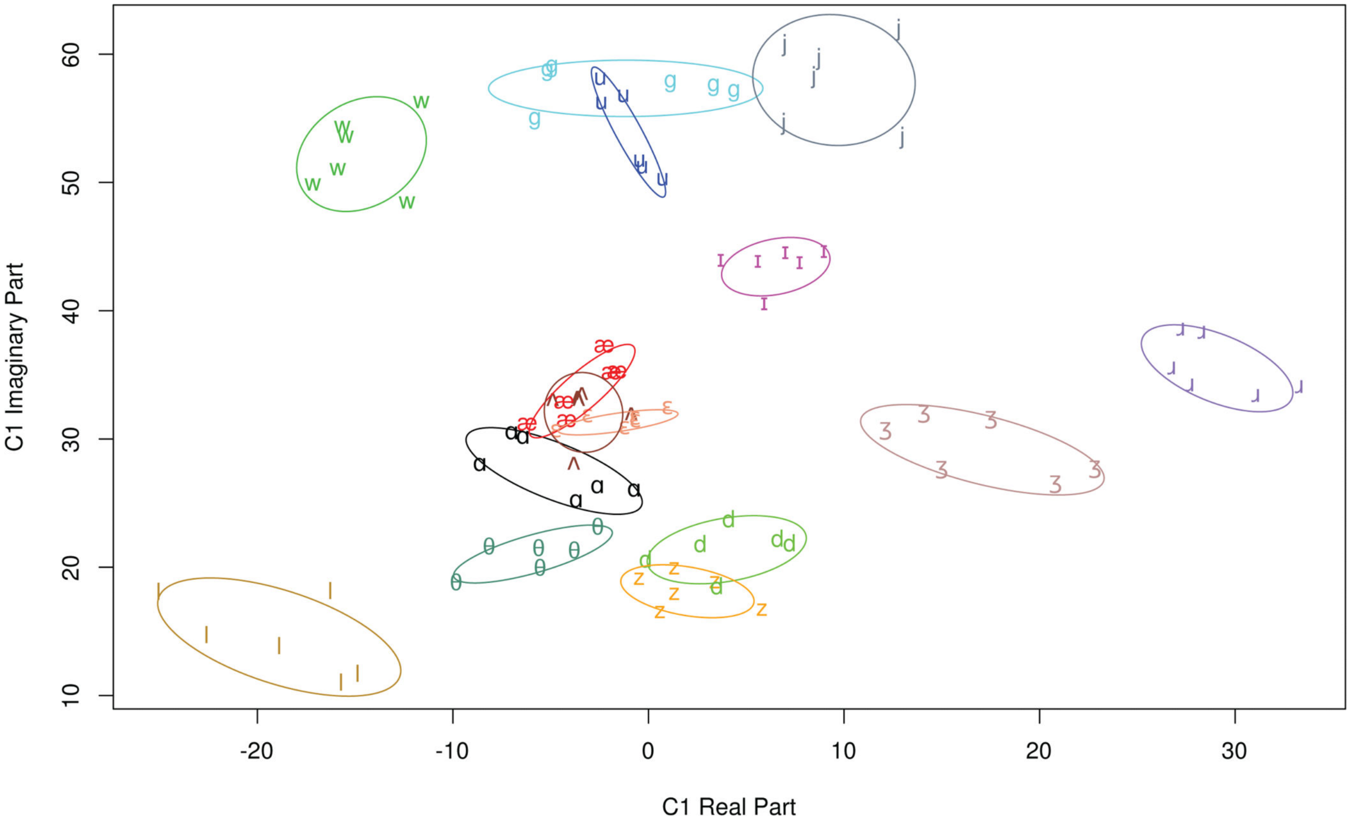

Figure 3.

Fourier transform results for 01_FC (female typical speaker). This graph shows all the data points for the first Fourier coefficient values (real part on the *x* axis, imaginary part on the *y* axis). Phoneme types are indicated by the symbol. One standard deviation confidence ellipses surround the phoneme classes.

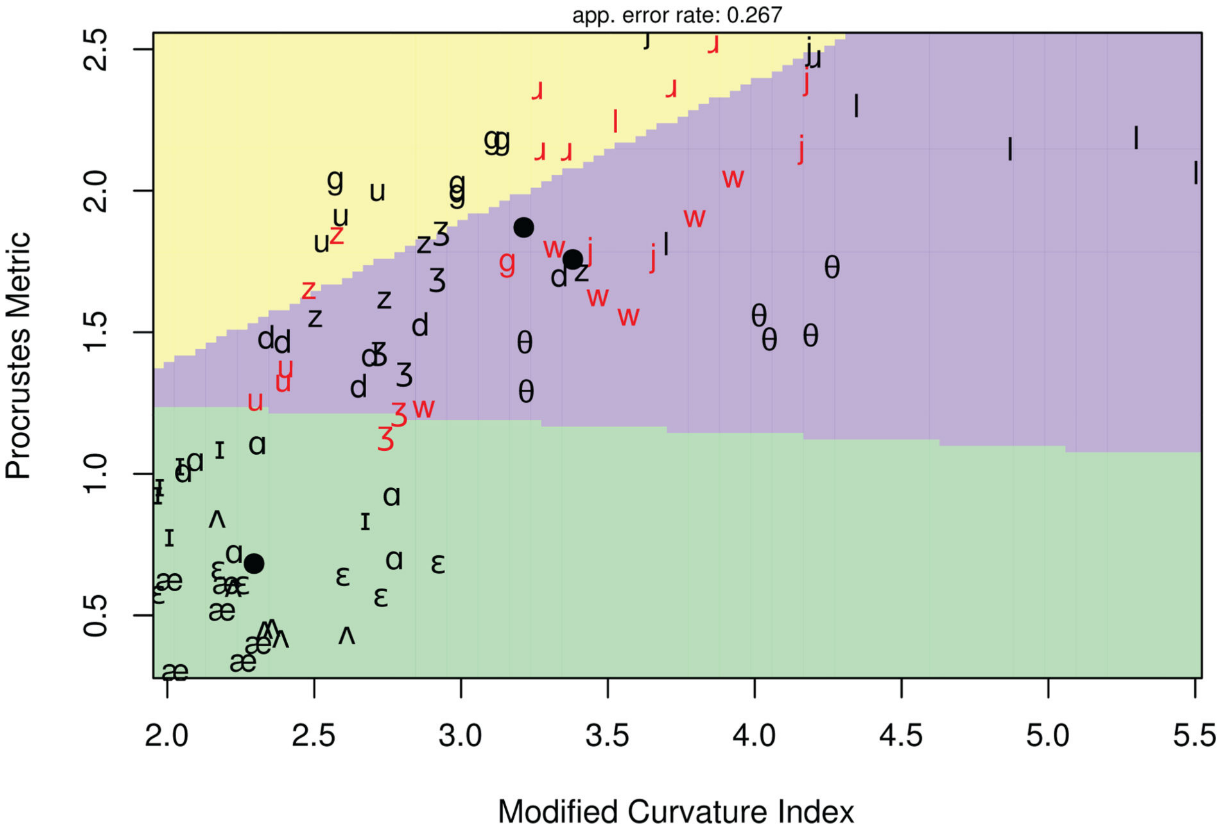

Figure 4.

LDA results for low, medium and high complexity classes for the Procrustes metric and MCI. Data are for 01_FC. Shaded areas represent complexity groupings: purple (medial shaded area) = high, yellow (uppermost shaded triangle) = medium, green (lowermost shaded area) = low. Symbols represent phoneme classes. Black = correctly classified; red = incorrectly classified.

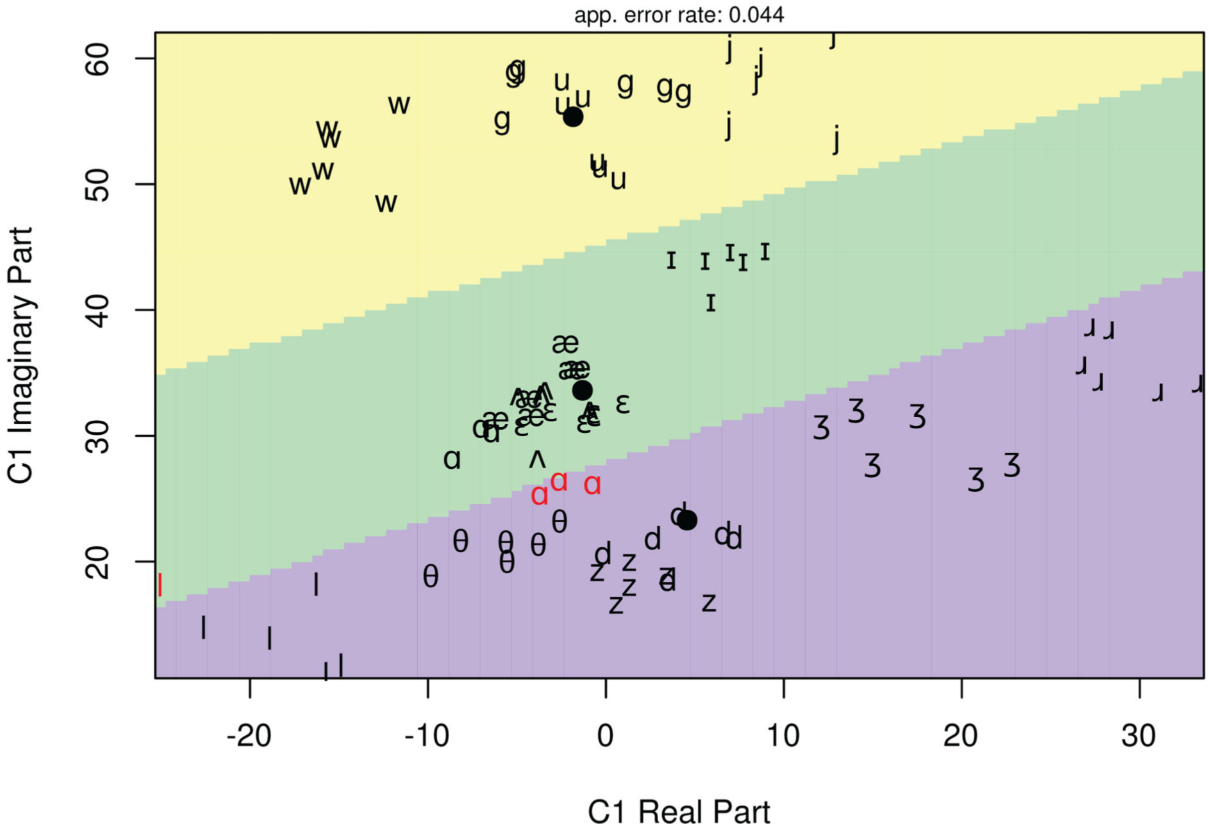

Figure 5.

LDA results for low, medium and high complexity classes for the Fourier transform first coefficient (C1) real and imaginary parts. Data are for 01_FC. Shaded areas represent complexity groupings: purple (lowermost shaded triangle) = high, yellow (uppermost shaded triangle) = medium, green (medial shaded area) = low. Symbols represent phoneme types. Black = correctly classified; red = incorrectly classified.

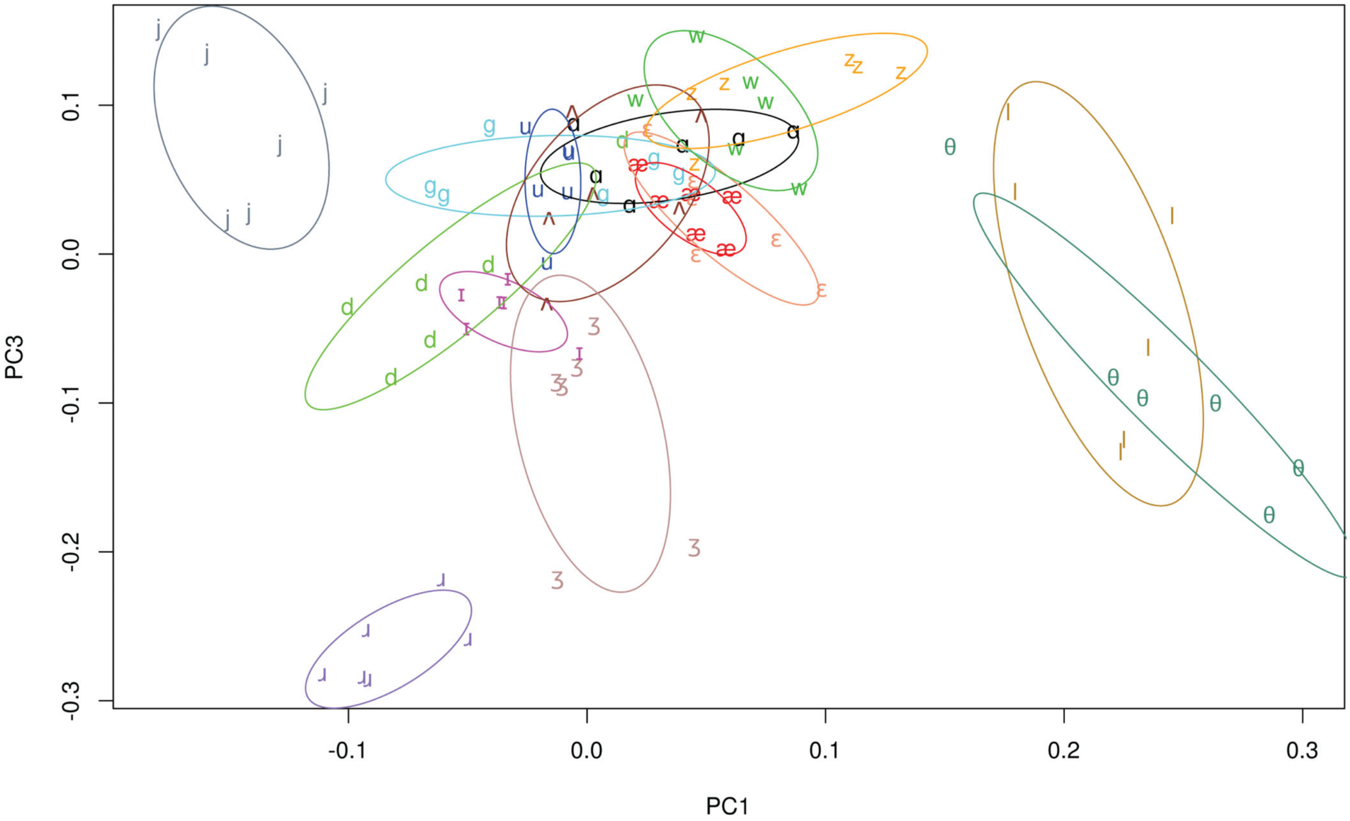

Figure 6.

PCA results (PC1 vs. PC3) of curvature values for participant 01_FC (female typical speaker). PC1 is plotted along the *x* axis and PC3 along the *y* axis. Phoneme types are indicated by the symbol. One standard deviation confidence ellipses surround the phoneme classes.

Table 1

LDA classification rates by metric and participant for L, M and H complexity groups

	Participant						
Metric ↓	**01_FC**	**02_FC**	**03_FC**	**05_MC**	**06_MC**	**07_MC**	**ALL**
Procrustes	0.67	0.73	0.58	0.61	0.72	0.62	0.62
MCI	0.59	0.52	0.53	0.53	0.61	0.48	0.56
C1 real	0.43	0.32	0.41	0.34	0.44	0.36	0.40
C1 imag	0.84	0.90	0.80	0.62	0.86	0.81	0.77
C2 real	0.34	0.46	0.50	0.33	0.34	0.36	0.36
C2 imag	0.46	0.50	0.49	0.47	0.51	0.31	0.39
C3 real	0.40	0.49	0.29	0.50	0.63	0.30	0.44
C3 imag	0.41	0.69	0.33	0.51	0.37	0.54	0.41
Procrustes + MCI	0.71	0.79	0.62	0.70	0.72	0.58	0.61
C1	0.96	0.89	0.83	0.86	0.87	0.80	0.81
C1 + C2 + C3	0.99	0.92	0.93	0.93	0.88	0.86	0.81
ALL	1	0.92	0.96	0.97	0.91	0.87	0.83

Table 2a

Confusion matrix for LDA classification success rates by phoneme type for all metrics and all participants

		Predicted														
		ɑ	**æ**	**ɪ**	**ʌ**	**ɛ**	**u**	**g**	**w**	**j**	**d**	**z**	**ʒ**	**θ**	**l**	**ɹ**
Actual	**ɑ**	**10**	2	2	12	5	0	0	3	0	0	2	0	0	0	0
	æ	0	**11**	4	0	14	0	0	0	0	0	0	6	1	0	0
	ɪ	0	4	**29**	0	0	2	0	1	0	0	0	0	0	0	0
	ʌ	15	2	0	**8**	5	0	0	1	0	0	4	1	0	0	0
	ɛ	0	15	0	2	**14**	0	0	2	0	1	2	0	0	0	0
	u	0	0	5	0	0	**21**	7	3	0	0	0	0	0	0	0
	g	0	0	0	0	0	11	**18**	1	6	0	0	0	0	0	0
	w	7	0	2	5	1	6	0	**15**	0	0	0	0	0	0	0
	j	0	0	8	0	0	0	6	0	**18**	0	0	0	0	0	4
	d	1	0	0	0	0	0	0	0	0	**8**	11	8	3	5	0
	z	4	0	0	0	0	0	0	1	0	7	**18**	6	0	0	0
	ʒ	0	5	1	2	0	0	0	0	0	0	6	**21**	0	0	1
	θ	0	0	0	1	0	0	0	0	0	5	0	0	**26**	4	0
	l	0	0	0	0	0	0	0	0	0	2	1	0	5	**28**	0
	ɹ	0	0	0	0	0	0	0	0	0	0	0	3	0	0	**33**

Table 2b

Confusion matrix for LDA complexity groups: all metrics and all participants

		Predicted		
		high	medium	low
actual	high	**185**	0	31
	medium	1	**105**	38
	low	15	9	**156**

Table 3

Inter-rater R^2 correlations for MCI, Procrustes metric and Fourier C1

Metric	MCI	Procrustes metric	C1 imaginary part	C1 real part
R^2 value	0.59	0.88	0.97	0.95