Dimensions of structure and variability in the human vocal tract

Katherine Vaughan-Williams[1], Steven Moran[2,3], Sam Kirkham[1]

[1] *Lancaster University, UK*
[2] *University of Neuchâtel, Switzerland*
[3] *University of Miami, USA*

kpvaughanwilliams@gmail.com, steven.moran@unine.ch, s.kirkham@lancaster.ac.uk

Abstract

A defining characteristic of the human vocal tract is a complex dynamic between structure and variability. Across a population we observe considerable variability in vocal tract dimensions, but variation in one dimension is rarely independent of other dimensions. Are some of these relationships more variable than others, or do there exist invariants in the morphology of the vocal tract? In this study, we report a data-driven investigation into the relationship between vocal tract dimensions based on multi-speaker real-time magnetic resonance imaging data. We discover different sub-populations in the data, which correspond to groups of speakers that share a common relationship between vocal tract parameters. This suggests a range of complex patterns of co-variation in the morphology of the human vocal tract. We conclude by speculating on the possible implications of these results for understanding individual differences in speech production.

Keywords: vocal tract anatomy, magnetic resonance imaging, speaker-specific variation, conditional inference trees

1. Introduction

The human vocal tract exhibits considerable variation between speakers. Most obvious are the changes that accompany child development from birth until adulthood, whereby changes in vocal tract length are largely determined by growth in the pharyngeal regions (Vorperian et al. 2005). But we also observe variation in adult populations, ranging from sexual dimorphism in vocal tract length (Fitch and Giedd 1999) and oral cavity length (Fant 1966) to individual differences in the hard palate (Lammert, Proctor, and Narayanan 2013). In many cases, variation in one dimension is rarely independent of other dimensions. For example, vocal tract dimensions are sometimes correlated with other aspects of the body, such as speaker height and weight (Stone et al. 2018), although in other cases there are no such relationships between speaker weight and vocal tract length (Hatano et al. 2012). In terms of variation within the vocal tract, the length of horizontal vocal tract structures tends to negatively correlate with the length of vertical structures (Honda et al. 1996), yet we also know that scaling is not uniform across different structures.

The fact that the relationship between vocal tract parameters varies between studies has many possible explanations, ranging from measurement technique to data quality to sample size. Aside from these considerations, one possible explanation is that the relationship between vocal tract parameters may be different in different areas of the parameter range. For example, speakers with a longer vocal tract may show a simple linear relationship with palate length, whereas perhaps speakers with a shorter vocal tract show a more complex relationship with palate length that could interact with other factors. This raises a question: do some vocal tract dimensions always scale together uniformly, or do they show a more non-linear relationship in different areas of a parameter range? Such results have implications for patterns of variability in speech production, because anatomical differences place constraints on the use of particular speech production strategies (Fuchs, Winkler, and Perrier 2008; Brunner, Fuchs, and Perrier 2009; Weirich and Fuchs 2013). In order to address this question, we conduct an exploratory study into variation in the morphology of the human vocal tract, with the aim of understanding structured variability in the relationship between vocal tract dimensions using multi-speaker real-time magnetic resonance imaging data.

2. Methods

We use Magnetic Resonance Imaging data of the vocal tract, taken from 69 speakers in the USC Speech MRI Database (Lim et al. 2021). Measurements were extracted by hand from two-dimensional midsagittal images of the vocal tract by the first author. All measurements were based on a single representative rest posture for each speaker and annotations were carried out using ImageJ (Schneider, Rasband, and Eliceiri 2012). The measurements reported in this study are as follows:

1. vocal tract length (mm)
2. palate length (mm)
3. palate height (mm)
4. tongue length (mm)
5. tongue area (mm^2)
6. body height (cm)
7. body weight (kg)
8. body-mass index (kg/m^2)

Our analysis is twofold: (1) what are the primary dimensions of variability? (2) what are the relationships between vocal tract parameters? We address (1) by submitting all measures to Principal Components Analysis, following by k-means clustering, which allows us to observe the ways in which measurements cluster together on a global scale.

The second analysis then aims to better understand the precise relationship between vocal tract parameters. A large number of highly-correlated measurements presents significant problems for modelling using classical parametric statistics, so we instead turn to a class of data-driven machine learning algorithms: conditional inference trees. Conditional inference trees are a class of regression models using binary recursive partitioning. We first test the null hypothesis of independence between the outcome variable and each predictor variable. If the null hypothesis cannot be rejected then the process stops. If the null hypothesis can be rejected then we select the predictor variable that has the strongest association with the outcome variable. We then implement a binary split in the predictor variable that maximises the homogeneity of each group in the binary split, in terms of its relationship with the outcome variable. This process is then repeated recursively until some stopping criterion is achieved, such as a maximum tree depth or minimum node size. The resulting model is a hierarchical tree with the most important predictor at the top and a series of binary splits within this predictor, which continues until all significant predictors have been exhausted. We visualise the models as in Figure 4, where the predictor variables are ordered from top-to-bottom in terms of importance, with the boxplots representing terminal nodes that correspond to the distribution of data points within that combination of variables.

We implement conditional inference trees in R using the `partykit` package (Hothorn and Zeileis 2015). We fitted a conditional inference tree to each variable in the data set as the outcome variable, with all remaining variables as predictor variables. All p-values for the splits were calculated using the Bonferroni method.

3. Results

3.1. PCA

We find that two principal components capture 79.5% of the variance. As shown in Figure 1, these dimensions capture variation across (1) vocal tract length and tongue length/area, and (2) variation in palate height, which is highly independent of the vocal tract/tongue measures. Palate length is equally weighted across both dimensions, showing its interaction with both palate height and vocal tract/tongue length. K-means clustering on these PC values reveals two separable clusters in Figure 2, which highly correlate with speaker sex.

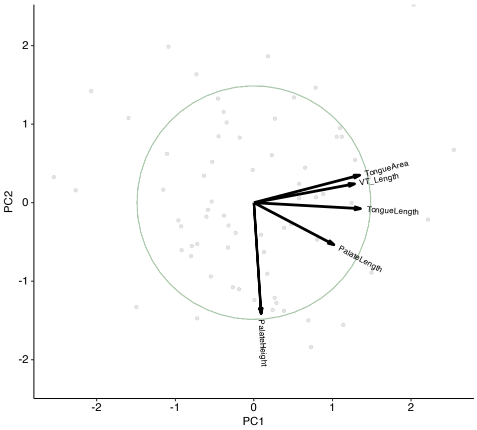

Figure 1: *PCA loadings for PC1 and PC1.*

3.2. Correlation matrix

Before showing the conditional inference trees, we first explore simple pairwise correlations between measurements. Figure 3 shows a correlation matrix for all variables. BMI is unsurprisingly highly correlated with height ($r = 0.82$), given that height is incorporated into the BMI measure. The next strongest correlation is between tongue area and tongue length ($r = 0.79$), which is also unsurprising given the inherent physical relationship between these measures. We also observe moderately strong correlations between tongue area and vocal tract length ($r = 0.75$), height and vocal tract length ($r = 0.74$), and tongue length and vocal tract length ($r = 0.71$). One problem with this analysis is that such variables are likely to be highly correlated with a number of other variables. Our following analysis addresses this point using conditional inference trees, which are well-suited to exposing complex relationships in highly colinear data.

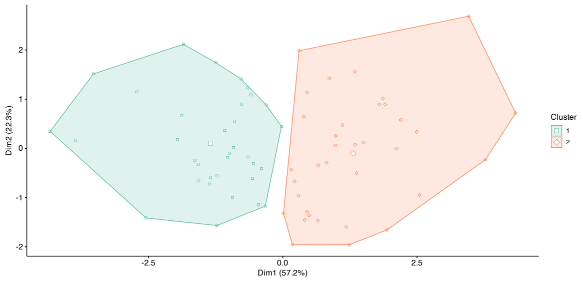

Figure 2: *Cluster plot showing each speaker in two-dimensional PCA space.*

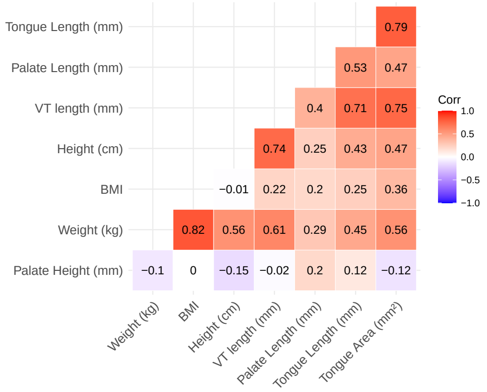

Figure 3: *Correlation matrix for all vocal tract and body measurements in the dataset.*

3.3. Conditional inference trees

The conditional inference trees expose more precise relationships between parameters. We show visualisations for three conditional inference trees that reveal the most interesting relationships and summarise some of the other results in text.

Figure 4 shows a conditional inference tree with vocal tract length as the outcome variable and all other variables as potential predictors. The model finds five distributions in the data, based on the interaction between three predictor variables. Speaker sex is the strongest predictor of vocal tract length, with male speakers having longer vocal tracts than female speakers. Within male speakers, there is one split in the distribution, such that speakers with a smaller tongue area (below or equal to 2937.6 mm^2) are more likely to have a smaller vocal tract. Within female speakers, a similar split occurs, but for tongue

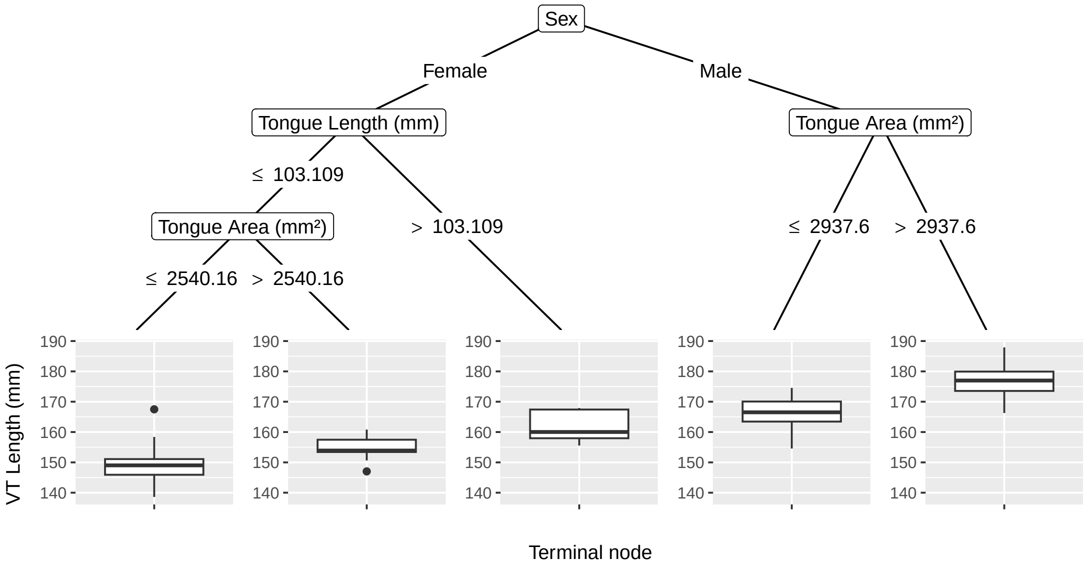

Figure 4: *Conditional inference tree fitted to vocal tract length measurements. Predictors that do not appear on the plot are not significant predictors of vocal tract length in the model.*

length rather than tongue area: speakers with longer tongues (greater than 103.109 mm) have longer vocal tracts. Finally, within female speakers with a shorter tongue, there is a further split based on small differences in tongue area, whereby a larger tongue area correlates with a slightly longer vocal tract. The other variables show no significant association with vocal tract length. This suggests a series of sub-populations in terms of how different measures impact vocal tract length in these data.

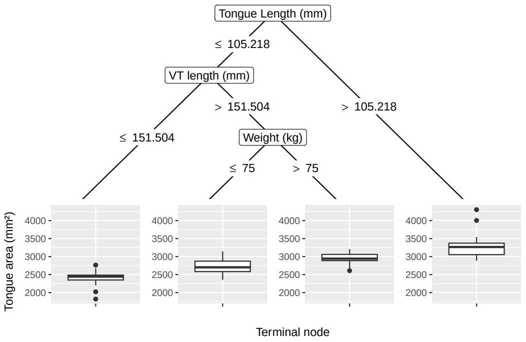

Figure 5: *Conditional inference tree fitted to tongue area measurements. Predictors that do not appear on the plot are not significant predictors of tongue area in the model.*

We fitted a conditional inference tree to tongue length, but found that variation in this measurement was only significantly predicted by variation in tongue area, which is unsurprising as we would expect a strong association between two related measures of the tongue. In the interests of space, we have not included a visualisation of this model. Instead, we show the model visualisation for the predictors of tongue area in Figure 5. This model shows that tongue length is the most important predictor of tongue area, with longer tongues predictably showing a larger area. However, within the lower half of the tongue length range (i.e. below or equal to 105.218 mm) other measures help to explain some of the variation. For example, in speakers with tongue length less than 105.28 mm there is an effect of vocal tract length in the expected direction. But in speakers with a slightly longer vocal tract there is a small difference in tongue area between speakers who weigh more than 75 kg and those that weigh less than 75 kg. This suggests that the relationship between tongue area and weight is a rather complex one and only emerges in a particular area of the range of possible tongue area values in these data. This is the only case where we found a significant relationship between a measurement of the whole body (such as height, weight, BMI) and a measurement of the vocal tract. In all other cases, none of the body measures were significant predictors of variation in vocal tract morphology.

Figure 6 shows a conditional inference tree fitted to palate length. In this case, the only variable that significantly predicts variation in palate length is tongue length. Specifically, speakers with a tongue length greater than 96.524 mm have a significantly longer palate than those with a tongue length below this value. The distributions between these two groups are fairly well separated, suggesting a strong association between tongue length and palate length.

4. Discussion and conclusion

We report a data-driven investigation into patterns of variability in the morphology of the human vocal tract. The most complex relationships are found in explaining the variance in vocal tract length. While the most important predictor is a fairly predictable sex-based difference, we then find that speakers with shorter vocal tracts also have smaller tongues (measured as tongue length in female speakers and tongue area in

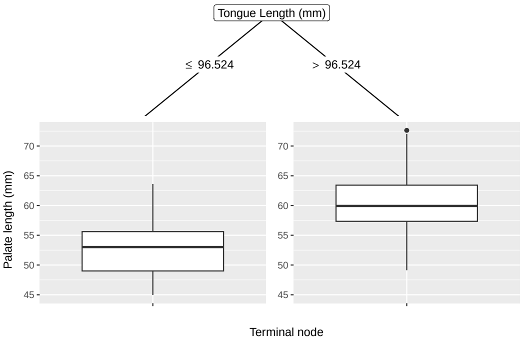

Figure 6: *Conditional inference tree fitted to palate length measurements. Predictors that do not appear on the plot are not significant predictors of palate length in the model.*

male speakers). Within female speakers, there is a sub-grouping of vocal tract length differences within speakers with smaller tongues, whereby those with smaller tongue areas have shorter vocal tracts. We note that these relationships are not uniform across speakers and point toward sub-groupings based on the interactions between vocal tract measurements.

Pairwise correlations showed moderately strong associations between vocal tract length and height, but we do not find this to be a significant predictor in our conditional inference trees, suggesting that this relationship can be captured via other dimensions in the model. In fact, we find relatively few relationships between vocal tract measurements and height/weight/BMI. The only significant effect of such a variable is in the model for tongue area, but the effect is limited. Specifically, the effect of weight on tongue area is only present for speakers with both a tongue length equal to or below 105.218 mm *and* a vocal tract length greater than 151.504 mm. Finally, we observed a simple relationship between tongue length and palate length, where speakers with longer tongues have predictably longer palates.

Overall, these results suggest that the relationship between vocal tract dimensions may vary across different sections of a parameter range, thereby complicating a straightforward scaling between dimensions. In terms of the implications of these results for speech production, it is unknown whether different sub-populations – as represented in the terminal nodes of our conditional inference trees – are likely to show any substantial differences in speech production. One possibility is that the anatomical constraints that characterise different sub-populations could lead to slight differences in articulatory behaviour. Whether such articulatory behaviours are motor equivalent and hence produce similar acoustic outputs is a possibility, but it is also worth investigating whether such anatomical differences underpin any of the observed individual variability in speech. Indeed, this raises the possibility that there could exist different classes of individual speaker variability that correspond with some of the sub-populations reported here.

In summary, this study reports the existence of sub-populations that share a set of relationships between vocal tract dimensions in different regions of the relevant parameter ranges. Future research will investigate whether individual variability in speech production can be grouped into similar classes that correspond to clusters of anatomical variation.

5. Acknowledgements

SK was funded by AHRC grant AH/Y002822/1, and SM was funded by SNSF grant PCEFP1_186841. Thanks to Tiena Danner.

6. References

Brunner, Jana, Susanne Fuchs, and Pascal Perrier (2009). “On the relationship between palate shape and articulatory behavior”. In: *Journal of the Acoustical Society of America* 125.6, pp. 3936–3949.

Fant, Gunnar (1966). “A note on vocal tract size factors and non-uniform F-pattern scalings”. In: *Speech Transmission Laboratory Quarterly Progress and Status Report* 1, pp. 22–30.

Fitch, W. Tecumseh and Jay Giedd (1999). “Morphology and development of the human vocal tract: a study using magnetic resonance imaging”. In: *Journal of the Acoustical Society of America* 106.3, pp. 1512–1522.

Fuchs, Susanne, Ralf Winkler, and Pascal Perrier (2008). “Do speakers’ vocal tract geometries shape their articulatory vowel space?” In: *Proceedings of the International Seminar on Speech Production*, pp. 333–336.

Hatano, Hiroaki, Tatsuya Kitamura, Hironori Takemoto, Parham Mokhtaru, Kiyoshi Honda, and Shinobu Masaki (2012). “Correlation between vocal tract length, body height, formant frequencies, and pitch frequency for the five Japanese vowels uttered by fifteen male speakers”. In: *Proceedings of Interspeech*, pp. 402–405.

Honda, Kiyoshi, Shinji Maeda, Michiko Hashi, Jim S. Dembowski, and John R. Westbury (1996). “Human palate and related structures: their articulatory consequences”. In: *Proceedings of ICSLP ’96* 2, pp. 784–787.

Hothorn, Torsten and Achim Zeileis (2015). “partykit: A modular toolkit for recursive partitioning in R”. In: *The Journal of Machine Learning Research* 16.1, pp. 3905–3909.

Lammert, Adam, Michael I. Proctor, and Shrikanth S. Narayanan (2013). “Morphological variation in the adult hard palate and posterior pharyngeal wall”. In: *Journal of Speech, Language, and Hearing Research* 56.2, pp. 521–530.

Lim, Yongman, Asterios Toutios, Yannick Bliesener, Ye Tian, Sajan Goud Lingala, Colin Vaz, Tanner Sorensen, Miran Oh, Sarah Harper, Weiyi Chen, Yoonjeong Lee, Johannes Töger, Mairym Lloréns Monteserin, Caitlin Smith, Bianca Godinez, Louis Goldstein, Dani Byrd, Krishna S. Nayak, and Shrikanth S. Narayanan (2021). “A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images”. In: *Scientific Data* 8.187, pp. 1–14.

Schneider, Caroline A., Wayne S. Rasband, and Kevin W. Eliceiri (2012). “NIH Image to ImageJ: 25 years of image analysis”. In: *Nature Methods* 9, pp. 671–675.

Stone, Maureen, Jonghye Woo, Junghoon Lee, Tera Poole, Amy Seagraves, Michael Chung, Eric Kim, Emi Z. Murano, Jerry L. Prince, and Silvia S. Blemker (2018). “Structure and variability in human tongue muscle anatomy”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.5, pp. 499–507.

Vorperian, Houri K., Ray D. Kent, Mary J. Lindstrom, Cliff M. Kalina, Lindell R. Gentry, and Brian S. Yandell (2005). “Development of vocal tract length during early childhood: a magnetic resonance imaging study”. In: *Journal of the Acoustical Society of America* 117.1, pp. 338–350.

Weirich, Melanie and Susanne Fuchs (2013). “Palatal morphology can influence speaker-specific realizations of phonemic contrasts”. In: *Journal of Speech, Language, and Hearing Research* 56, S1894–S1908.