INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Quantifying vocal tract shape variation and its acoustic impact: a geometric morphometric approach

Amelia J. Gully

Department of Language and Linguistic Science, University of York, York, UK
amelia.gully@york.ac.uk

Abstract

The shape of the vocal tract varies considerably between individuals. The relationship between detailed variation in vocal tract shape and the acoustics of speech is not yet well understood, despite its potential for increasing understanding in the fields of voice biometrics, forensic speech science, and personalised speech synthesis. One reason that this topic has not yet been extensively explored is that 3D vocal tract shape is difficult to quantify robustly. Geometric morphometrics is a technique developed in evolutionary biology for statistically valid quantification and comparison of anatomical shapes. This study makes use of 3D magnetic resonance imaging data of the vocal tracts of eight individuals, and accompanying audio recordings, combined with geometric morphometric techniques to determine whether the method offers useful information for speech science. The results suggest a linear relationship between the shapes of the vocal tract and output spectra, and there is evidence of possible sexual dimorphism and allometry (a systematic variation of shape with size) in the vocal tract, although due to the limited sample size the results did not reach statistical significance. The results suggest that geometric morphometrics can provide useful information about the vocal tract, and justify further study using this technique.

Index Terms: vocal tract shape, inter-speaker variation, geometric morphometrics

1. Introduction

The human vocal tract has a highly complex 3D shape which is extremely variable between individuals [1]. Source-filter theory [2] indicates that the shape of the vocal tract is an essential contributor to the speech signal, and the shape of an individual's vocal tract at any time is governed by its anatomical *structure*, the speaker's habitual articulatory *setting*, and moment-to-moment articulatory *strategy* [3]. Although anatomical structure may have the least impact on the speech signal of these three factors, it is also the most robust and will remain constant across all speech events. Therefore, if such anatomical shape differences can be measured and their acoustic impact quantified, it would be of considerable value for the study of speaker identity and the field of acoustic phonetics in general.

At present, no robust method has been presented to quantify the shape of the whole vocal tract and compare between individuals, despite advances in magnetic resonance imagine (MRI) and data sharing making 3D vocal tract data more widely available [4]. Variation in vocal tract shape between individuals has historically been considered in one dimension only, by measuring the length of the vocal tract (e.g. [3]). Classical morphometric methods—i.e. making a set of measurements representing distances between anatomical features—have been used for the vocal tract (e.g. [5]), but such measurements are difficult to incorporate into a robust statistical framework as they are not independent. Furthermore, the interaction between the shape and size of the vocal tract mean that detailed shape variations are often 'lost in the noise' given the larger effect of vocal tract size, and prevent robust comparisons of shape being made between subjects with different size vocal tracts.

Some studies have considered the shape of the vocal tract, or certain aspects of it, in two dimensions. For example, [6] measured the variation in the shape of the hard palate midline between five individuals, and [7] attempted to quantify the shape of the entire midsagittal vocal tract outline for nine individuals. Both studies made use of principal component analysis to quantify the variation and found considerable variation between individuals; [7] also linked this back to acoustics by considering the shape of the resulting vowel space. Other studies have used simulation to control the majority of the vocal tract in order to determine the acoustic effect of varying one aspect of vocal tract shape, such as the piriform fossae [8]. However, without a method for normalising for the effect of differences in the rest of the vocal tract, inter-speaker comparisons are difficult. Considering only a single anatomical feature at a time also means that possible multi-way interactions between different anatomical features in the vocal tract may exist which have not yet been identified. Furthermore, it is possible that the shape of some vocal tract features may co-vary systematically with the shape of other tract features, or with tract size. Unless the shape of the entire tract is considered holistically, such possible relationships remain undetectable.

1.1. Geometric morphometrics

Unlike classical morphometric approaches, which produce a series of individual measurements whose relationships to each other are not defined, geometric morphometrics (GMM) treats the entire shape as a single entity [9]. In GMM, an n-dimensional shape is defined as a set of landmarks, whose semantic meaning is identical across all subjects (for example, landmark number 21 might correspond to the position of inner corner of the left eye for all subjects). Since the whole set of landmarks is used throughout all subsequent analysis, spatial relationships between them are preserved. Analysis techniques such as principal component analysis can be applied to the shapes in order to quantify the dimensions of variation in the population, and specialist statistical procedures can be used to determine the relationship of shape to other quantities of interest. See [9] for more details on GMM.

1.2. This study

The aim of this study is to determine whether GMM provides useful information for the analysis and comparison of vocal tract shape between individuals, and the correspondence between vocal tract shape and acoustic output, rather than technical considerations such as MRI resolution or audio quality. As

http://dx.doi.org/10.21437/Interspeech.2021-1400

such the current study is deliberately limited to participants for whom 3D MRI data with the same acquisition parameters are available along with contemporaneous clean audio recordings (unaffected by MRI noise or noise cancellation). This limits the study to eight subjects, four male and four female. Additionally the study is limited to a single vowel per subject in order to control for the wide variation in articulation of different utterances expected between participants. As a result there is likely to be insufficient data to draw strong conclusions, but the study can be considered a success if the technique reveals meaningful relationships worthy of further investigation. A comprehensive study expanding the dataset to more individuals, MRI protocols, and utterances is underway; this will follow in a later publication.

2. Method

2.1. Data collection and landmarking

Volumetric MRI data were collected for eight participants over the course of several projects, four males (M1–4) and four females (F1–4), with a voxel resolution of $0.75 \times 0.75 \times 1$mm; see [10] for more details about the participants and complete details of the MRI scan protocol. Participants were required to hold each articulation for total of 16s in order to complete each 3D scan. Data were collected for a number of vowels, but this study will consider only [ə] as the 'neutral' vowel is expected to represent an approximately average vocal tract position for each participant.

Immediately following the MRI scan, participants repeated the experiment in an anechoic chamber under "MRI-like" conditions [11] in order to capture clean audio. Audio was recorded at a sample rate of 48kHz using a DPA 4066 head-mounted microphone placed approximately 4cm away from the corner of the participant's lips. A long-term average spectrum was then calculated from a visually-determined steady-state portion of the vowel with approximately 0.8s duration.

Following data capture, MorphoDig [12] was used to place landmarks on the MRI volume data. A total of 81 3D landmarks were selected per subject; these include 41 landmarks on the midsagittal plane (including those defined by [1] for maximum compatibility), and an additional 40 off-midsagittal landmarks at locations including the lips, hard palate, and larynx. Of the 81 landmarks selected, 47 are 'semi-landmarks' representing outlines or curves (such as the tongue mid-line) rather than distinct anatomical points; these are treated differently in the GMM procedure. Please see the supplementary materials [10] for a full list of landmark locations. Note that for one subject (F1) it was necessary to re-slice the MRI data to obtain a true midsagittal slice due to misalignment in the MRI scanner.

Landmarking on the speech spectrum was intentionally simpler, given the lower dimensionality of the data and the exploratory nature of this study. The long-term average spectra were normalised to 0dB peak, and 2D landmarks selected in MATLAB corresponding to the frequency and magnitudes of the first four formants and the spectral notch following the fourth formant which was present for all eight subjects. This resulted in five landmarks per subject which generally followed the downward trend of the spectral slope. It is acknowledged that the shape of the spectrum will necessarily include contributions from the source as well as the filter; nevertheless this is intended as a starting point to determine whether it is possible to establish relationships between vocal tract shape and acoustic output. Future studies with multiple vowels will be able to make use of the shape of the vowel space in place of the shape of the spectrum, thus removing the impact of the voice source.

2.2. Geometric morphometrics

The R [13] package *geomorph* [14] was used for geometric morphometric analysis. The first step in the GMM procedure is generalised Procrustes analysis (GPA), wherein each set of landmarks is scaled to a unit centroid size, translated, and iteratively rotated to minimise the Procrustes distance between shapes [9]. By scaling shapes in this way, GMM makes a clear distinction between *shape* and *size*. During landmarking, some points can be denoted as 'semi-landmarks' rather than true landmarks, in order to capture curves or outlines that may not be semantically equivalent across participants (for example, a set of points denoting the curve of the tongue mid-line may not each be located on identical anatomical locations across participants, but taken together the set of points denotes the entire tongue curve). During GPA alignment, these semi-landmarks are allowed to slide along the curve in order to minimise Procrustes distance between landmarks while preserving the overall curve shape [15].

Following GPA, each shape exists as a single point in shape space; a curved domain with the mean shape at the centre [9]. Since most statistical processes assume a linear domain, these points are then projected back to a flat domain tangential to the shape space. Variation in vocal tract shape among the population can then be quantified using a principal component (PC) analysis on the points in this tangent space. Any point in the space can be mapped back to a set of landmarks—even points with no associated shape in the original dataset—so it is possible to determine not only the principal axes of variation but to map these back to shapes so that they can be easily interpreted. The nature of PC analysis means that each PC will be orthogonal to the others, providing useful information about which aspects of the vocal tract shape co-vary and which are independent. Note however that the PCs are only relevant to the original population of shapes and cannot be used to generalise about shapes outside of this population.

Geomorph also offers a wide range of statistical processes adapted for use with shape data, making use of residual randomisation permutation procedures [16] in order to counteract the singular covariance matrices produced as a consequence of the GPA procedure. In this study, analysis of variance (ANOVA) methods are used to determine whether there is a statistically significant relationship between shape and other factors (participant sex, and vocal tract size as determined by the original centroid size prior to the GPA scaling process). Additionally, the two-block partial least squares (PLS) method is used to determine the degree of correlation between two sets of landmarks, or a set of landmarks and any other set of features.

It should be noted that GMM is commonly performed on tens, if not hundreds, of samples. Therefore, it is not expected that results in this study will reach significance given a population size of only eight. Nevertheless, if the process is sufficient to reveal variation in the population which corresponds to established acoustic phonetic theory, the method will be deemed worthy of further investigation.

3. Results and discussion

All figures in this section use the midsagittal landmarks only, for clarity of presentation. Interactive 3D versions of all plots in this section, and additional materials, are available in the supplementary materials [10].

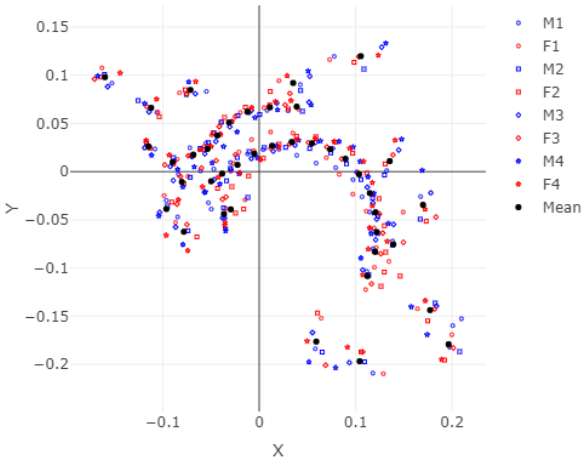

Figure 1: *GPA alignment of vocal tract landmarks for all eight subjects. Axis scales are arbitrary following GPA scaling. Landmarks have been rotated to display the tip of the nose as the left-most landmark, with laryngeal landmarks in the lower right hand corner. Midsagittal landmarks only, see [10] for an interactive 3D plot of all landmarks.*

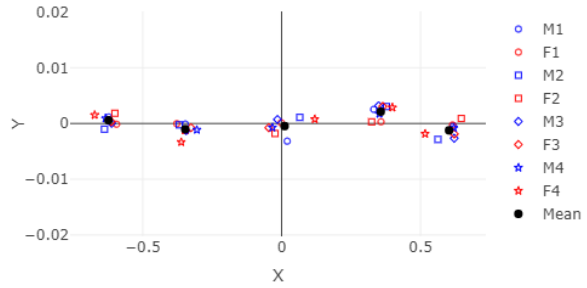

Figure 2: *GPA alignment of spectral landmarks for all eight subjects, aligned along first principal axis. Axis scales are arbitrary following GPA scaling.*

3.1. Overall findings

The initial GPA alignment of vocal tract landmarks for all subjects can be seen in Fig. 1. A principal component (PC) analysis in shape space indicates that the first PC accounts for 45.57% of shape variation, with the remaining seven PCs accounting for 19.14%, 11.49%, 7.59%, 6.36%, 5.36%, 4.48%, and <0.001% of variation respectively. The GPA alignment of the corresponding spectral landmarks are presented in Fig. 2. Spectral landmarks corresponding to the first two formants exhibit less variability between subjects than higher-frequency landmarks, which is expected due to the phonetic constraints placed upon them for intelligibility. For the spectral landmarks the first three PCs account for over 99.9% of the variation in shape, with the first PC accounting for 85.86% alone. As expected, PC1 appears to relate mostly to frequency associated with landmarks, while PC2 appears to be related to spectral slope; see [10] for additional figures illustrating this relationship.

Fig. 3 illustrates the vocal tract shapes corresponding to the maximum and minimum of the PC1 axis (similar plots for all PCs are available in the supplementary materials [10]). It can be seen from Fig. 3 that multiple vocal tract shape features appear to co-vary along this first principal axis of shape variation. The tongue shape, relative height of the tongue-epiglottis join, relative anteriority of the tongue tip, and angle and relative anteriority of the glottis—along with hard palate doming and lower lip anteriority as revealed by the 3D landmarks—co-vary in this dataset. Co-variation of multiple different vocal tract features is also observed for all other PCs. Although it is not possible to generalise from eight participants, this result suggests that different—and apparently unrelated—aspects of vocal tract shape may co-vary systematically. Further study on a larger dataset is necessary to explore this finding further.

A two-block PLS analysis reveals a weak linear relationship between the shape of the vocal tract (as determined by the anatomical landmarks), and the shape of the resulting audio spectrum (as determined by spectral landmarks), with a correlation coefficient of 0.712; however this does not reach statistical significance. In the case of [ə], the spectral landmarks (at least the first four which correspond to formant peaks) are expected to be approximately equally spaced in frequency, presenting perhaps the simplest case for such a comparison, and an analysis incorporating different vowels would provide more information. A more comprehensive set of spectral landmarks—perhaps incorporating sliding semi-landmarks—would also be beneficial. Despite these limitations, the presence of a correlation between vocal tract and spectrum shape suggests this this is a promising avenue for further study.

3.2. Sexual dimorphism and allometry

An ANOVA on vocal tract shapes reveals that for this population, 21.32% of shape variation is attributable to subject sex, 11.93% of variation in shape is attributable to vocal tract size, and 19.13% of variation is attributable to the interaction between sex and size, although due to the small population these results did not reach significance. The results indicate that over half of the vocal tract shape variation in the population under study is attributable to sex, vocal tract size, or the interaction between these two variables; however there remains 47.62% of shape variation that is not attributable to either of these factors and may be due to individual differences, as well as factors not considered in this study such as participant height and weight.

A similar ANOVA using the spectral landmarks reveals that 16.41% of the variation in spectral shape is attributable to sex; again these results do not reach significance. It is unsurprising that some differences in the spectrum will be due to the participant's sex. Further study with a larger participant group—and possibly a larger number of spectral landmarks—is necessary in order to examine the relationship between spectral shape and other factors in more detail.

Further exploring the difference in vocal tract shape with sex, Fig. 4 illustrates the mean vocal tract shapes for female and male subjects. It is clear from the figure that the mean shapes of the two populations are quite similar, suggesting that when individual effects are averaged out and size is compensated for, the vocal tract shape of males and females exhibit only small differences in the production of [ə]. This supports the suggestion above that idiosyncratic differences in anatomy have a greater impact on vocal tract shape than sex alone. However, small differences in shape can still cause substantial changes in acoustics [17], and such differences may have a larger impact on vowels with a narrower airway constriction. Fig. 4 suggests some small differences around the tongue root and epilarynx, and the 3D landmarks reveal differences in hard palate doming, particularly in the anterior region. However, a larger population and

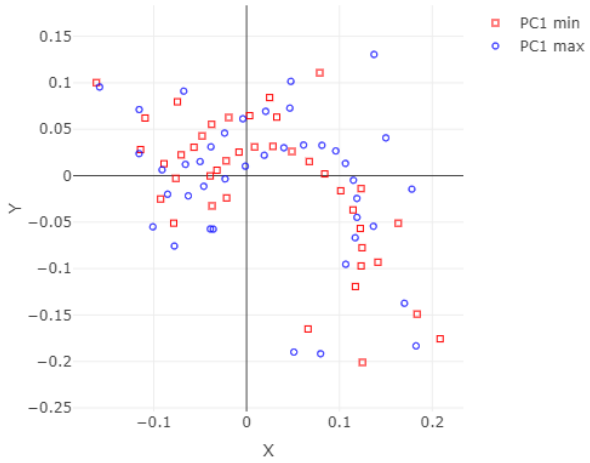

Figure 3: *Vocal tract shapes at the maximum and minimum of the PC1 axis. Axis scales are arbitrary following GPA scaling. Landmarks have been rotated to display the tip of the nose as the left-most landmark, with laryngeal landmarks in the lower right hand corner. Midsagittal landmarks only, see [10] for interactive 3D plots of all landmarks and for all PCs.*

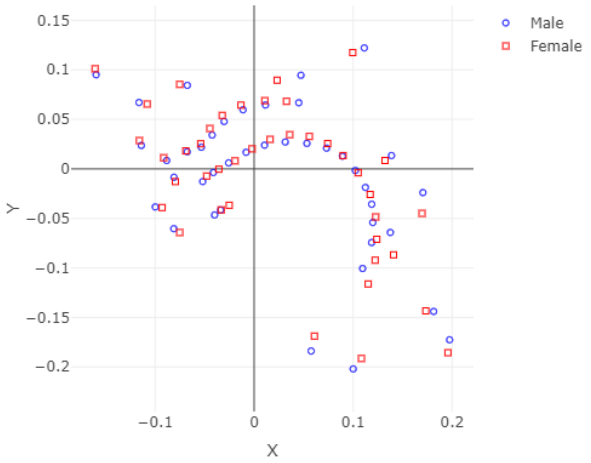

Figure 4: *Mean vocal tract shapes for males and females. Axis scales are arbitrary following GPA scaling. Landmarks have been rotated to display the tip of the nose as the left-most landmark, with laryngeal landmarks in the lower right hand corner. Midsagittal landmarks only, see [10] for an interactive 3D plot of all landmarks.*

a greater range of utterances is required before any conclusions can be drawn.

Finally, considering the variation of shape with size, a two-block PLS analysis illustrates that vocal tract shape and size exhibit a weakly linear relationship, with a correlation coefficient of 0.774, although again the results did not reach statistical significance. These results provide further support to the ANOVA results and suggest allometry in vocal tract shape, providing additional motivation to extend the proposed GMM techniques to a larger set of vocal tract data.

3.3. Future work

The most pressing avenue for further study is to expand this work to a larger dataset in order to draw statistically robust conclusions. The recent release of 75 subjects' worth of MRI data [18], in addition to existing datasets ([4] and references therein), make this a possibility. This larger population size would also make it possible to consider the correlation between other features—such as participant height, weight, or age (where available)—and vocal tract shape, using mixed-effects modelling. It is also necessary to expand the dataset to (at least) other vowels, requiring the development of statistically valid approaches for comparing within, as well as between, individuals.

Questions remain about whether the landmarks selected here are sufficient to capture vocal tract shape. It should be noted that 3D vocal tract data is captured while the subject is supine, subjected to loud noise, and articulating for an unnaturally long time, meaning that this data is not necessarily representative of vocal tract shape in running speech [11, 19]. It would be of interest to determine how well certain landmark positions can be predicted from others—and whether midsagittal landmarks are sufficient to describe 3D shape—to reduce data dimensionality, ease the burden of manual landmarking, and perhaps eventually permit the use of 2D real-time MRI of running speech to inform dynamic vocal tract shape models.

Finally, the appropriateness of the spectral landmarks needs to be considered. It may be useful to incorporate spectral semi-landmarks, since it is likely that the most spectral variation will occur away from formant locations, in areas which are not critical for intelligibility. It may be more useful to consider the shape of the vowel space for each participant, which would be possible with the incorporation of additional vowels in the dataset. Alternatively, correlations between vocal tract shape and more abstract measures of spectral shape, such as MFCCs, may be examined.

At present, no distinction is made between structure, setting, and strategy and their relative contributions to vocal tract shape. However, the author has recently collected MRI data of phoneticians using different articulatory settings, and the results of the current study suggest that GMM is an appropriate framework for analysis of these data. Furthermore, the GPA procedure provides a means of normalising vocal tract shape, permitting the study of subject-specific articulations, which is of considerable value for the field of phonetics.

4. Conclusions

The aim of this paper was to determine whether geometric morphometric analysis can provide useful information about the shape of the vocal tract and its impact on speech acoustics. The results reveal a weakly linear relationship between vocal tract shape and spectrum shape, despite a limited number of participants and a highly simplified representation of the speech spectrum. Additionally, results suggest that some apparently unrelated aspects of vocal tract shape may co-vary, as well as suggesting an effect of sex on vocal tract shape (independent of size), and a systematic relationship between vocal tract shape and size. All these results are worthy of further study with a larger dataset.

5. Acknowledgements

This work is funded by the British Academy, grant reference PF19/100024. The author would like to thank all the subjects for their participation.

6. References

[1] M. Eslami, C. Neuschaefer-Rube, and A. Serrurier, "Automatic vocal tract landmark localization from midsagittal MRI data," *Nature Scientific Reports*, vol. 10, no. 1, p. 1468, 2020.

[2] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton & Co. N.V., 1960.

[3] P. Mokhtari, F. Clermont, and K. Tanaka, "Toward an acoustic-articulatory model of inter-speaker variability," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 158–161.

[4] P. Birkholz, S. Kürbis, S. Stone, P. Häsner, R. Blandin, and M. Fleischer, "Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties," *Nature Scientific Data*, vol. 7, no. 1, p. 255, 2020.

[5] M. Leppävuori, E. Lammentausta, A. Peuna, M. K. Bode, J. Jokelainen, J. Ojala, and M. T. Nieminen, "Characterizing vocal tract dimensions in the vocal modes using magnetic resonance imaging," *Journal of Voice [in press]*, 2020.

[6] A. Lammert, M. Proctor, and S. Narayanan, "Interspeaker variability in hard palate morphology and vowel production," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 6, pp. 1924–1933, 2013. [Online]. Available: http://dx.doi.org/10.1044/1092-4388(2013/12-0211)

[7] S. Fuchs, R. Winkler, and P. Perrier, "Do speakers' vocal tract geometries shape their articulatory vowel space?" in *Proceedings of the 8th International Seminar on Speech Production*, Strasbourg, France, 2008, pp. 333–336.

[8] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoust. Sci. Tech.*, vol. 26, no. 1, pp. 16–26, 2005.

[9] D. C. Adams and E. Otárola-Castillo, "geomorph: an R package for the collection and analysis of geometric morphometric shape data," *Methods in Ecology and Evolution*, vol. 4, no. 4, pp. 393–399, 2013.

[10] A. J. Gully, "Supplementary materials for "Quantifying vocal tract shape variation and its acoustic impact: a geometric morphometric approach", Interspeech 2021," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4946782

[11] A. J. Gully, P. Foulkes, P. French, P. Harrison, and V. Hughes, "The Lombard effect in MRI noise," in *Proceedings of the International Congress of Phonetic Sciences 2019*, Melbourne, Australia, 2019, pp. 800–804.

[12] R. Lebrun, "MorphoDig, an open-source 3D freeware dedicated to biology," in *IPC5*, Paris, France, 2018.

[13] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[14] D. C. Adams, M. L. Collyer, and A. Kaliontzopoulou, *Geomorph: Software for geometric morphometric analyses. R package version 3.2.1*, 2020. [Online]. Available: https://cran.r-project.org/package=geomorph

[15] F. J. Bookstein, "Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape," *Medical Image Analysis*, vol. 1, no. 3, pp. 255–243, 1997.

[16] M. L. Collyer and D. C. Adams, "RRPP: An R package for fitting linear models to high-dimensional data using residual randomization," *Methods in Ecology and Evolution*, vol. 9, no. 2, pp. 1772–1779, 2018.

[17] B. H. Story, "A comparison of vocal tract perturbation patterns based on statistical and acoustic considerations," *J. Acoust. Soc. Am.*, vol. 122, no. 4, pp. EL107–EL114, 2007.

[18] Y. L. et al., "A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images," 2021. [Online]. Available: arXiv:2102.07896

[19] A. Tsukanova, I. K. Douros, A. Shimorina, and Y. Laprie, "Can static vocal tract positions represent articulatory targets in continuous speech? matching static MRI captures against real-time MRI for the French language," in *Proceedings of the International Congress of Phonetic Sciences 2019*, Melbourne, Australia, 2019, pp. 2801–2805.