Acoustic Analysis of Vowel Production Using Magnetic Resonance Imaging

Tharinda Piyadasa[1], Michael Proctor[2], Amelia Gully[3], Yaoyao Yue[1], Kirrie Ballard[4], Naeim Sanaei[5], Sheryl Foster[4, 5], Tünde Szalay[2], David Waddington[4], Craig Jin[1]

[1]School of Electrical and Computer Engineering, University of Sydney, Australia
[2]Department of Linguistics, Macquarie University, Australia
[3]Department of Language and Linguistic Science, University of York, UK
[4]Faculty of Medicine and Health, University of Sydney, Australia
[5]Radiology Department, Westmead Hospital, Australia
tharinda.piyadasa@sydney.edu.au

Abstract

Details of individual speaker vocal tract configurations remain understudied due to the limitations of most instrumental phonetic methods. Midsagittal articulation of /iː-ɑː-ɔː-uː-ɜː/ by a speaker of Southern Standard British English was captured using real-time Magnetic Resonance Imaging. Three-dimensional tract configurations during production of the same vowels were acquired using high-resolution volumetric imaging. Acoustic models derived from imaging data were compared with reference acoustic recordings. Models demonstrate particular sensitivity to palatal and velar tract geometry; details of pharyngeal structures had less influence on acoustic responses. These data demonstrate the importance of multimodal data in acoustic characterization of individual speaker vowels.

Index Terms: vowels, MRI, English, vocal tract, area function, acoustic modeling

1. Introduction

Understanding how the vocal tract is configured during vowel production has been a central concern of speech science and a foundational topic informing models of speech production [1, 2, 3, 4]. Magnetic resonance imaging (MRI) has advanced the study of vowel production by allowing safe, flexible and accurate imaging of soft tissue [5, 6, 7]. Vocal tract geometries have been resolved in detail by orienting image planes perpendicular to the axis of airway [8, 9, 10, 11], and volumetric imaging techniques have provided comprehensive coverage of the whole upper airway at increasing spatial resolutions [12, 13, 14]. These methods have revealed the complex geometries involved in vowel production and how they vary between speakers, languages, and allophones [15, 16, 17], informing more detailed vocal tract representations beyond idealized tube models [18, 19, 20, 21, 22]. These data are advancing our understanding of the complex relationships between vocal tract morphology, articulation, and acoustics [23, 24, 25], but many aspects of vocal tract shaping and its acoustic consequences are still imperfectly understood.

High resolution volumetric imaging of the vocal tract can be achieved using multi-second acquisition times, but because participants must sustain vowels in these studies, the resulting postures are static, and may be hyperarticulated. Real-time MRI (rtMRI) allows imaging of the upper airway during connected speech produced with more natural prosody [26, 27], which is important for phonetic characterization of vowels [28, 29, 30]. rtMRI has provided insights into the dynamics of vowel articulation in French, Portuguese, English, and other languages [31, 32, 33, 34]. Note that for speech MRI studies, the supine participant posture may affect vocal tract shape [35], and the loud noise means that Lombard speech is usually captured [36].

Volumetric and real-time imaging offer important complementary insights into vowel production, but reconciling data from different modalities creates additional challenges. Companion volumetric and real-time MRI data have previously been combined to examine vocal tract shaping in Swedish and French vowels [31, 16]. Articulatory data obtained using different sensing methods can be assessed by comparing acoustic responses of models derived from the corresponding vocal tract configurations. Acoustic responses have been estimated from MRI data of Czech, Finish and English vowels using Finite Element [37, 38, 39, 40], finite-difference time-domain [20], and 3D digital waveguide methods [41, 42]. While these techniques rely on multi-dimensional representations of the vocal tract, acoustic responses can be estimated from 1D tract models [43, 44], allowing for direct comparison of models based on articulatory data captured during vowel production with acoustic recordings of the corresponding vowels produced by the same speaker.

The goal of this study is to examine details of vowel production in a speaker of British English in new detail using volumetric MRI, real-time MRI, and acoustic recordings. We explore the acoustic responses of vocal tract configurations by synthesizing vowels from area functions derived from MRI data, and validate these acoustic models against out-of-scanner reference recordings of vowels produced by the same speaker. By comparing acoustic outputs of vocal tract models derived from each dataset, we aim to assess the relative advantages of each imaging modality for acoustic modelling of vowels, and methods for extracting vocal tract representations appropriate for vowel models from each type of data. Finally, we assess the impact on acoustic modelling of different methods of representation of data obtained from each imaging modality, and examine how different approaches to segmentation of vocal tract boundaries affect acoustic responses of tract models.

2. Methods

Data were collected during the pilot phase of a larger project examining development of speech motor control in adolescents. An adult female L1 speaker of Standard Southern British English produced vowels in a series of speech tasks recorded out of and inside an MRI scanner. Vowels /iː-ɑː-ɔː-uː-ɜː/ were elicited

in monosyllabic words *"beet', 'Bart', 'bought', 'boot', 'Bert"*. Each token was recorded once in a quiet room with a Glottal Enterprises EG2-PCX2 digital speech recorder to familiarize the participant with the experimental materials. The same utterances were later recorded five times during a real-time MRI scan, and additionally as sustained productions during a volumetric MRI scan. A total of (5 words) × (1 pre-scan + 5 rtMRI + 1 volumetric MRI) = 35 vowel exemplars were included in the analysis.

2.1. Vocal Tract Imaging

MRI data were acquired at Westmead Hospital on a Siemens Magnetom Prisma 3T scanner with a 64-channel head/neck receiver array coil. The speaker's upper airway was imaged while lying supine. Data were acquired from an 8 mm slice aligned with the mid-sagittal plane, over a 280 × 280 mm field of view, using a 2D RF-spoiled, radially-encoded FLASH sequence [45]. Audio was recorded concurrently in-scanner at 16 kHz using an Opto-acoustics FOMRI-III ceramic noise-canceling microphone designed for MRI environments [46]. rtMRI data were reconstructed in Matlab into midsagittal videos with a pixel resolution of 0.97 mm^2, encoded as 72 frames per second MP4 files. Audio and video were time-aligned during postprocessing and video reconstruction.

3D configuration of the vocal tract during sustained (7.6 s) vowel production was captured using volumetric imaging of the upper airway. Data were acquired using a T1-weighted fast 3D gradient-echo sequence, with a spatial resolution of 160 × 160 × 32 px over a 256 × 256 × 64 mm field of view centred on the pharynx. These data provide detailed imaging of the entire upper airway, extending vertically from the upper trachea to the nasal cavities and sagittally from cheek to cheek, with a voxel resolution of 1.6 × 1.6 × 2.0 mm.

2.2. Vocal tract segmentation

Volumetric data (DICOM format), were processed using ITK-SNAP [47], an open-source tool for 3D segmentation of medical images. Contrast was enhanced, and the Snake tool was used for semi-automatic segmentation of vocal tract boundaries, from which 3D tract outlines were extracted (Fig. 1).

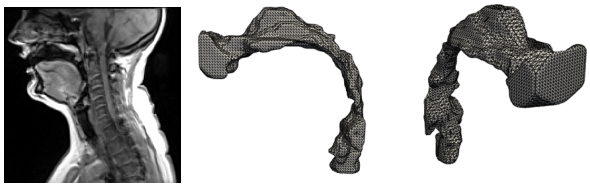

Figure 1: *Vocal tract configuration, sustained [aː].* Midsagittal slice and 3D volume extracted from segmented volumetric data

rtMRI data were analyzed using *inspect_rtMRI*, a Matlab-based graphical interface for inspection and semi-automatic segmentation of rtMRI data [48]. Midsagittal vocal tract boundaries were located in image frames corresponding to articulatory target postures for each vowel, and area functions were extracted at 7.7 mm intervals, from glottis to labial midpoint. Midsagittal slices were extracted from 3D vocal tract models, and an additional set of vocal tract area functions were calculated using the same method, to obtain a second set of high resolution midsagittal vocal tract representations for each vowel (Fig. 2).

2.3. Modelling Acoustic Responses

Each vocal tract (VT) was modeled as a series of concatenated cylindrical tubes with cross-sectional areas (CSA) derived from area functions, normalized by π. Tube models were downsampled through linear interpolation:

$$N = \text{round}\left(\frac{\text{len}(\text{VT}) \times \frac{Fs}{2} \times 4}{c}\right)$$

where *Fs* = 16 kHz, c = 350 m/s (speed of sound in moist air at body temperature 37°C [2]).

Reflection coefficients were calculated for each segment junction within the vocal tract model to simulate acoustic impedance mismatches [49]. The number of reflection coefficients corresponds to the length of the interpolated tube model, where the coefficients were derived using the formula:

$$r = \frac{A_{i+1} - A_i}{A_{i+1} + A_i}$$

where A_i and A_{i+1} are the CSAs of adjacent tube sections. Additionally, a finite lossy tube was modeled by appending a reflection coefficient to represent the mouth's impedance, with the value set to 0.71 [50]. A Rosenberg glottal pulse [51] was generated and processed through the vocal tract filter designed by converting reflection coefficients into filter coefficients using Durbin's recursion:

$$a_{k+1}[n] = a_k[n] + r_{k+1} \times a_k[k-n] \quad \text{for } n = 0, 1, \ldots, k$$

where $a_k[n]$ are the filter coefficients at recursion step k, r_{k+1} is the reflection coefficient at the $k+1$-th junction, and n indexes the coefficients in the filter.

The output speech signal was generated by convolving the glottal pulse with the acoustic filter coefficients, followed by amplitude normalization.

Acoustic properties of synthesized and recorded speech were compared using formant frequencies. F1, F2, F3 were tracked over speech intervals containing target vowels following the approach proposed in [52], using 20 ms Hamming analysis windows, 50% overlap, *max_F34cutoff* = 4500 Hz, and a pre-emphasis filter factor of 0.98.

3. Results and Discussion

Formant frequencies for vowels produced by the participant in reference (out-of-scanner) recordings were first compared to mean values (Table 1) reported for female speakers in Standard Southern British English [53]. Formants generally align closely with SSBE means; the participant's /iː/ is more fronted, and /ɔː/ is lower. Overall, formant values for short pronunciations are closer to SSBE means compared to sustained pronunciations, which may be attributed to the effects of hyperarticulation in sustained vowels. In particular, the large difference (30%) in F2 values for /uː/ shows that the sustained vowel was produced with a backed, more peripheral articulation.

Compared to out-of-scanner recordings, the in-scanner recordings typically exhibit larger F1 and F2 values ($\geq$ 4% difference). This is in line with the findings of [36], where it was established that increases in F1 and F2 occur due to scanner noise, and additional F1 increases can be attributed to the supine posture of the subject (Table 2). These effects were found to be subject-dependent. In this case, the scanner environment caused the tongue to be positioned higher and more

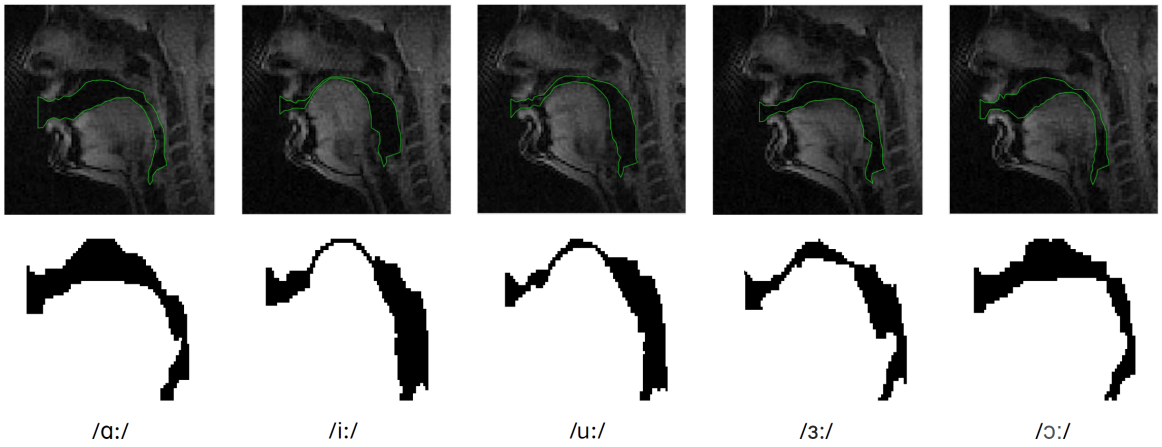

Figure 2: *Vocal tract segmentations used to calculate area functions*: Top: vowel target frames in rtMRI data; Bottom: midsagital sections from 3D tract volumes of sustained vowel postures. L-to-R: [ɑː-iː-uː-ɜː-ɔː]

Table 1. *Comparison of participant reference vowel formants with mean SSBE female vowel formant frequencies (Hz) [53]*

		F1	F2	F3
/ɑː/	SSBE Mean (F)	**910**	**1316**	**2841**
	Out-scanner (short)	-74	-152	+238
	Out-scanner (sustained)	-171	-40	-32
/iː/	SSBE Mean (F)	**303**	**2654**	**3203**
	Out-scanner (short)	+35	+144	-67
	Out-scanner (sustained)	-12	+202	+90
/uː/	SSBE Mean (F)	**328**	**1437**	**2674**
	Out-scanner (short)	+73	+60	+59
	Out-scanner (sustained)	+68	-436	+261
/ɜː/	SSBE Mean (F)	**606**	**1695**	**2839**
	Out-scanner (short)	-31	-112	+214
	Out-scanner (sustained)	-57	+36	+148
/ɔː/	SSBE Mean (F)	**389**	**888**	**2796**
	Out-scanner (short)	+18	-128	-796
	Out-scanner (sustained)	+110	-123	+354

Table 2. *Comparison of Formant Values Between In-Scanner Recordings and 1D Acoustic Model (rtMRI) with Out-Scanner Recordings (Short)*

		F1	F2	F3
/ɑː/	Out-scanner (short)	**836**	**1164**	**3079**
	In-scanner	+36	+72	-331
	1D acoustic model (rtMRI)	-80	+432	-214
/iː/	Out-scanner (short)	**338**	**2798**	**3136**
	In-scanner	+95	-492	-449
	1D acoustic model (rtMRI)	-44	-599	-546
/uː/	Out-scanner (short)	**401**	**1497**	**2733**
	In-scanner	+28	+243	-187
	1D acoustic model (rtMRI)	-37	+354	-53
/ɜː/	Out-scanner (short)	**575**	**1583**	**3053**
	In-scanner	+362	+58	-212
	1D acoustic model (rtMRI)	-72	+61	-234
/ɔː/	Out-scanner (short)	**407**	**760**	**2000**
	In-scanner	+208	+112	+576
	1D acoustic model (rtMRI)	+63	+424	+710

forward, reflecting a more constrained vocal tract shape during in-scanner recordings. In contrast, the 1D acoustic models based on rtMRI typically exhibit lower F1 values which may arise from the simplifications in the modeling process that fail to fully capture the open vocal tract configuration. Additionally, the acoustic model tends to have higher F3 values compared to in-scanner recordings, indicating differences in the back cavity configuration. This difference can be attributed to the back cavity segmentation being influenced by the presence of soft tissue, particularly around the epiglottis area.

Formants from the 1D acoustic models based on midsagittal volumetric images generally align with out-of-scanner sustained vowel recordings, though there are some notable discrepancies (Table 3). For instance, /uː/ and /ɔː/ show considerable differences in F2 values, with the 1D acoustic model having much higher values ($\geq$ 50% difference) compared to the corresponding out-of-scanner recordings. However, it should be noted that when compared to SSBE mean values and out-of-scanner values for short utterances, the out-of-scanner sustained values are much lower (Table 1). Also, the large difference in F3 values for /ɜː/ and /ɔː/ (16% and 22% difference respectively) suggest variations in the pharyngeal cavity shape, as from Figure 2

Overall, the F1 values for both 1D acoustic models are close to the out-of-scanner values, indicating a reasonable approximation of vertical tongue positions. However, F2 and F3 values exhibit greater deviations. Furthermore, the models often show smaller formant values compared to out-of-scanner recordings, which may be due to the lack of lip radiation effects in the synthesized speech.

3.1. Refinements in 3D midsagittal slices

Several adjustments were made to the 3D midsagittal segmentations to observe the accuracy of our acoustic modeling. These adjustments involved refining soft tissue boundaries in the regions around the hard palate, velar constrictions, and epiglottis. The changes were prompted by initial observations that revealed anatomical inaccuracies in the 3D segmentations, such as an unusually large cavity at the hard palate in /ɑː/ and /ɔː/, likely due to the relatively smaller amounts of soft tissue affecting the

Table 3. *Comparison of Formant Values Between 1D Acoustic Model (Volumetric) with Out-Scanner Recordings (Sustained)*

		F1	F2	F3
/ɑː/	Out-scanner (sustained)	**739**	**1276**	**2809**
	1D acoustic model (volumetric)	-65	+138	-187
/iː/	Out-scanner (sustained)	**291**	**2856**	**3293**
	1D acoustic model (volumetric)	+51	-630	-68
/uː/	Out-scanner (sustained)	**396**	**1001**	**2935**
	1D acoustic model (volumetric)	0	+825	+265
/ɜː/	Out-scanner (sustained)	**549**	**1731**	**2987**
	1D acoustic model (volumetric)	-97	-76	-476
/ɔː/	Out-scanner (sustained)	**499**	**765**	**3150**
	1D acoustic model (volumetric)	+54	+379	-708

resolution of the upper airway boundary. All adjustments were made through manual post-processing of the initial segmentations located using ITK-SNAP (Sec. 2.2).

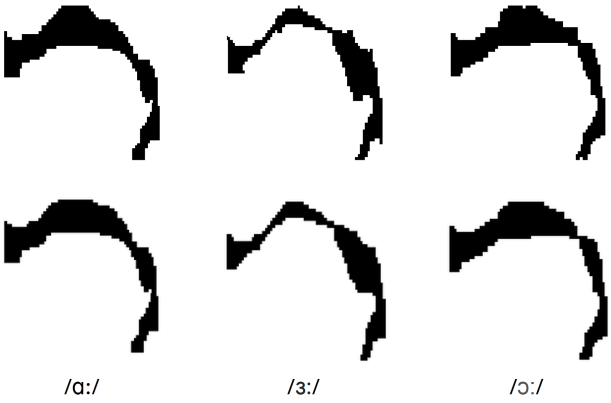

Figure 3: *Refinements in 3D midsagittal slices*: Top: original midsagittal sections; Bottom: midsagital sections after manual post-processing. L-to-R: [ɑː-ɜː-ɔː]

The adjustments made to the 3D midsagittal segmentations led to closer alignment of the formant values with the out-of-scanner recordings (Table 4). Refining the hard palate in /ɑː/ and /ɔː/ generally improved the alignment of F1 and F2 values by reducing exaggerated resonances caused by an initially larger hard palate. A similar improvement was observed when the velar constriction was increased in /ɔː/. This was expected, as increased velar constriction lengthens the front cavity of the vocal tract, thereby reducing F1 and F2 values.

The impact of the epiglottis definition in /ɑː/ and /ɜː/ was investigated to determine whether simplifying the epiglottis representation, as seen in rtMRI area functions, would improve the formant values (Table 4). However, this adjustment did not noticeably improve the formant values for either vowel. While the epiglottis influences the shape of the pharyngeal cavity, its impact on F1 and F2 is less pronounced compared to velar and palatal constrictions.

4. Conclusions

A method for determining detailed vocal tract configurations associated with vowel production has been proposed and validated using an acoustic synthesis framework. The impact of different tissue segmentation strategies has been assessed using 1D area functions extracted from rtMRI images and midsagittal slices of 3D data, validating these against both in-scanner and out-of-scanner acoustic recordings. The analysis showed that both tube models show variations in formant frequencies compared to out-of-scanner recordings. Overall, acoustic models based on midsagittal slices of 3D volumetric data more accurately represent natural speech formants compared to the models based on rtMRI images, indicating the importance of a better representation of the vocal tract geometry. The findings suggest that further improvements should include comprehensive vocal tract models considering lip radiation and dental structures. Therefore, future work will focus on using complete 3D vocal tract models to obtain acoustic responses and incorporating dental features and refined lip radiation models to improve vocal tract modeling.

Table 4. *Comparison of Formant Values for Original and Refined Midsagittal Slices of 3D Volumetric Representations of Vowels /ɑː/ /ɔː/, and /ɜː/*

		F1	F2	F3
/ɑː/	Out-scanner (sustained)	**739**	**1276**	**2809**
	1D acoustic model (Volumetric)	-65	+138	-187
	Refined hard palate	-111	+123	+22
	Non-refined epiglottis	-128	+176	-198
/ɔː/	Out-scanner (sustained)	**499**	**765**	**3150**
	1D acoustic model (Volumetric)	+54	+379	-708
	Refined hard palate	+28	+366	-921
	Increased constriction	+10	-29	-934
/ɜː/	Out-scanner (sustained)	**549**	**1731**	**2987**
	1D acoustic model (Volumetric)	-97	-76	-476
	Non-refined epiglottis	-134	-76	-685

5. Acknowledgements

Supported by Australian Research Council Discovery Grant DP220102933.

6. References

[1] T. Chiba and M. Kajiyama, *The vowel – its nature and structure*. Tokyo: Kaseikan, 1941.

[2] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Springer-Verlag, Berlin, 1972.

[3] G. Fant, *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. s'Gravenhage: Mouton, 1960.

[4] K. N. Stevens, *Acoustic Phonetics*. Cambridge: MIT Press, 2000.

[5] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *JASA*, vol. 90, no. 2, pp. 799–828, 1991.

[6] C. A. Moore, "The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images," *JSLHR*, vol. 35, no. 5, pp. 1009–1023, 1992.

[7] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *JASA*, vol. 100, no. 1, pp. 537–554, 1996.

[8] M. Matsamura, "Measurement of three-dimensional shapes of vocal tract and nasal cavity using magnetic resonance imaging," in *Proc. ICLSP*, 1992, pp. 779–782.

[9] D. Demolin, T. Metens, and A. Soquet, "Three-dimensional Measurement of the Vocal Tract by MRI," *ICLSP*, pp. 272–275, 1996.

[10] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images," *JPhon*, vol. 30, no. 3, pp. 533–553, 2002.

[11] P. Badin, G. Bailly, M. Raybaudi, and C. Segebarth, "A three-dimensional linear articulatory model based on MRI data," in *Proc. 3rd ETRW on Speech Synthesis*, 1998.

[12] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions for an adult female speaker based on volumetric imaging," *JASA*, vol. 104, no. 1, pp. 471–487, 1998.

[13] K. C. Welch, G. D. Foster, C. T. Ritter, T. A. Wadden, R. Arens, G. Maislin, and R. J. Schwab, "A novel volumetric magnetic resonance imaging paradigm to study upper airway anatomy," *Sleep*, vol. 25, no. 5, pp. 532–542, 2002.

[14] P. Badin and A. Serrurier, "Three-dimensional modeling of speech organs: Articulatory data and models," in *Tech. Comm. Psychological and Physiological Acoustics*, vol. 36, no. 5. Acoust. Soc. Japan, 2006, pp. 421–426.

[15] M. K. Tiede, "An MRI-based study of pharyngeal volume contrasts in Akan and English," *J. Phon.*, vol. 24, no. 4, pp. 399–421, 1996.

[16] O. Engwall, V. Delvaux, and T. Metens, "Interspeaker variation in the articulation of nasal vowels," *Proc. 7th ISSP*, pp. 3–10, 2006.

[17] Y. Wang, J. Dang, X. Chen, J. Wei, H. Wang, and K. Honda, "An MRI-based acoustic study of Mandarin vowels," in *Interspeech*, 2013, pp. 568–571.

[18] P. Mokhtari, T. Kitamura, H. Takemoto, and K. Honda, "Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients," *JPhon*, vol. 35, no. 1, pp. 20–39, 2007.

[19] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *JASA*, vol. 5, pp. 3231–3254, 2005.

[20] H. Takemoto, P. Mokhtari, and T. Kitamura, "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method," *JASA*, vol. 128, no. 6, pp. 3724–3738, 2010.

[21] S. Stone, M. Marxen, and P. Birkholz, "Construction and evaluation of a parametric one-dimensional vocal tract model," *IEEE/ACM Trans. ASLP*, vol. 26, no. 8, pp. 1381–1392, 2018.

[22] 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," *Scientific data*, vol. 7, no. 1, p. 255, 2020.

[23] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoustical science and technology*, vol. 26, no. 1, pp. 16–26, 2005.

[24] A. Lammert, M. I. Proctor, A. Katsamanis, and S. S. Narayanan, "Morphological variation in the adult vocal tract: a modeling study of its potential acoustic impact," in *Interspeech*, Florence, Italy, 27-31 Aug. 2011, pp. 2813–2816.

[25] A. J. Gully, "Quantifying vocal tract shape variation and its acoustic impact: A geometric morphometric approach." in *Interspeech*, 2021, pp. 3999–4003.

[26] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *JASA*, vol. 115, no. 4, pp. 1771–1776, 2004.

[27] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K. S. Nayak, and S. Narayanan, "Analysis of speech production real-time MRI," *Comput Speech Lang*, vol. 52, pp. 1 – 22, 2018.

[28] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *JASA*, vol. 97, no. 5, pp. 3099–3111, 1995.

[29] M. Hashi, J. R. Westbury, and K. Honda, "Vowel posture normalization," *JASA*, vol. 104, no. 4, pp. 2426–2437, 1998.

[30] G. Morrison and P. Assmann, *Vowel Inherent Spectral Change*. Berlin: Springer, 2012.

[31] D. Demolin, S. Hassid, T. Metens, and A. Soquet, "Real-time MRI and articulatory coordination in speech," *Comptes Rendus Biologies*, vol. 325, no. 4, pp. 547–556, 2002.

[32] A. Teixeira, P. Martins, C. Oliveira, C. Ferreira, A. Silva, and R. Shosted, "Real-Time MRI for Portuguese," in *Computational Processing of the Portuguese Language*, H. Caseli, Ed. Berlin: Springer, 2012, pp. 306–317.

[33] M. Proctor, C. Lo, and S. Narayanan, "Articulation of English Vowels in Running Speech: a Real-time MRI Study," in *Proc. ICPhS*, Glasgow, 10-14 Aug. 2015.

[34] C. Carignan, R. K. Shosted, M. Fu, Z.-P. Liang, and B. P. Sutton, "A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French," *JPhon*, vol. 50, pp. 34–51, 2015.

[35] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech MRI: morphology and function," *Physica Medica*, vol. 30, no. 6, pp. 604–618, 2014.

[36] A. J. Gully, P. Foulkes, P. French, P. Harrison, and V. Hughes, "The Lombard effect in MRI noise," in *Proc. ICPhS*, 2019, pp. 5–9.

[37] P. Kršek, "Design of FE models of vocal tract for Czech vowels," in *Proc. Interaction and Feedbacks*, 2000, pp. 103–110.

[38] K. Dedouch, J. Horácek, T. Vampola, J. Švec, P. Kršek, and R. Havlık, "Acoustic modal analysis of male vocal tract for Czech vowels," *Interaction and Feedbacks*, pp. 13–19, 2002.

[39] D. Aalto, O. Aaltonen, R.-P. Happonen, P. Jääsaari, A. Kivelä, J. Kuortti, J.-M. Luukinen, J. Malinen, T. Murtola, R. Parkkola *et al.*, "Large scale data acquisition of simultaneous mri and speech," *Applied Acoustics*, vol. 83, pp. 64–75, 2014.

[40] M. Arnela, R. Blandin, S. Dabbaghchian, O. Guasch, F. Alías, X. Pelorson, A. Van Hirtum, and O. Engwall, "Influence of lips on the production of vowels based on finite element simulations and experiments," *JASA*, vol. 139, no. 5, pp. 2852–2859, 2016.

[41] M. Speed, D. Murphy, and D. Howard, "Modeling the vocal tract transfer function using a 3d digital waveguide mesh," *IEEE/ACM Trans. ASLP*, vol. 22, no. 2, pp. 453–464, 2014.

[42] A. J. Gully, H. Daffern, and D. T. Murphy, "Diphthong synthesis using the dynamic 3D digital waveguide mesh," *IEEE/ACM Trans. ASLP*, vol. 26, no. 2, pp. 243–255, 2018.

[43] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *The Bell System Technical Journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[44] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Comm.*, vol. 1, no. 3-4, pp. 199–229, 1982.

[45] A. J. Kennerley, D. A. Mitchell, A. Sebald, and I. Watson, "Real-time magnetic resonance imaging: mechanics of oral and facial function," *Br J Oral Max Surg*, vol. 60, no. 5, pp. 596–603, 2022.

[46] Optoacoustics Ltd., "FOMRI-II version 2.2," 2007.

[47] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[48] M. I. Proctor, D. Bone, and S. S. Narayanan, "Rapid semi-automatic segmentation of real-time Magnetic Resonance Images for parametric vocal tract analysis," in *Interspeech*, Makuhari, 26-30 Sept. 2010, pp. 1576–1579.

[49] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, A. V. Oppenheim, Ed. Prentice-Hall, 1978.

[50] A. Beköz, "Modeling of plosive to vowel transitions," Master's thesis, Middle East Technical University, 2007.

[51] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *JASA*, vol. 49, no. 2B, pp. 583–590, 1971.

[52] T. M. Nearey, P. F. Assmann, and J. M. Hillenbrand, "Evaluation of a strategy for automatic formant tracking," *JASA*, vol. 112, no. 5_Supplement, pp. 2323–2323, 2002.

[53] D. Deterding, "The formants of monophthong vowels in Standard Southern British English pronunciation," *JIPA*, vol. 27, no. 1-2, pp. 47–55, 1997.