INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Database of volumetric and real-time vocal tract MRI for speech science

Tanner Sorensen, Zisis Skordilis, Asterios Toutios, Yoon-Chul Kim,[*] Yinghua Zhu,[†] Jangwon Kim,[‡] Adam Lammert,[§] Vikram Ramanarayanan,[¶] Louis Goldstein, Dani Byrd, Krishna Nayak, Shrikanth Narayanan

University of Southern California, Los Angeles, CA, USA

tsorense@usc.edu

Abstract

We present the USC Speech and Vocal Tract Morphology MRI Database, a 17-speaker magnetic resonance imaging database for speech research. The database consists of real-time magnetic resonance images (rtMRI) of dynamic vocal tract shaping, denoised audio recorded simultaneously with rtMRI, and 3D volumetric MRI of vocal tract shapes during sustained speech sounds. We acquired 2D real-time MRI of vocal tract shaping during consonant-vowel-consonant sequences, vowel-consonant-vowel sequences, read passages, and spontaneous speech. We acquired 3D volumetric MRI of the full set of vowels and continuant consonants of American English. Each 3D volumetric MRI was acquired in one 7-second scan in which the participant sustained the sound. This is the first database to combine rtMRI of dynamic vocal tract shaping and 3D volumetric MRI of the entire vocal tract. The database provides a unique resource with which to examine the relationship between vocal tract morphology and vocal tract function. The USC Speech and Vocal Tract Morphology MRI Database is provided free for research use at `http://sail.usc.edu/span/morphdb`.

Index Terms: speech production, speech corpora, magnetic resonance imaging, multi-modal database, large-scale phonetic tools

1. Introduction

The articulatory speech data-sets that are readily available to the research community have been consistently well-utilized in pursuit of addressing fundamental questions about speech production [1, 2, 3]. Until relatively recently, speech articulatory data had been difficult to obtain and generally lacking. A growing number of resources has begun to reverse this problem, but many still tend to focus on targeted laboratory speech (e.g., simple syllables or isolated phonemes) or only on read speech. Here, we present the USC Speech and Vocal Tract Morphology MRI Database, a new database for the community that captures a wide variety of dynamic speech tasks in conjunction with detailed structural parameters and also non-speech articulations, all with an eye toward understanding and explaining speech and speaker variability.

Magnetic resonance imaging (MRI) is a flexible technology for speech research. Rapid imaging methods have achieved a balance among the competing factors of temporal resolution, spatial resolution, and signal-to-noise ratio that allows for flexible characterization of vocal tract morphology and function using a suite of complementary MRI methods [4, 5]. Real-time MRI (rtMRI) characterizes the dynamic shaping of the vocal tract during speech in any scan plane(s) of interest with no need for repeated scans [6]. 3D volumetric MRI characterizes the entire vocal tract with high spatial resolution during sustained sounds in as little as 7 s [7]. Together, these scans characterize the function and morphology of the vocal tract with high temporal (rtMRI) and spatial resolution (3D volumetric MRI). The USC Speech and Vocal Tract Morphology MRI Database provides rtMRI of dynamic vocal tract shaping, denoised audio recorded simultaneously with rtMRI, and 3D volumetric MRI of a comprehensive set of the American English continuant sounds. The USC Speech and Vocal Tract Morphology MRI Database is provided free for research use at the project page: `http://sail.usc.edu/span/morphdb`.

2. Database acquisition

2.1. Experiments

Seventeen (8 m, 9 f) speakers of American English participated. None of the participants spoke a language other than English fluently, nor had any lived outside the United States for a significant amount of time. See Table 1 for participant age and state of origin. The parents of each participant were native speakers of American English. None of the speakers reported abnormal hearing or speech pathology.

Each speaker participated in two sessions on different days. One session was for acquiring rtMRI data-sets; the other session was for acquiring 3D volumetric MRI data-sets. The experimenter explained the nature of the experiment and the experiment protocol to the participant before each scan. The participant lay on the scanner table in a supine position. The head was fixed in place by foam pads inserted between each temple and the receiver coil. The participant read visual stimuli from a back-projection screen from inside the scanner bore without moving the head. The speech corpus captured 3D MRI of sustained continuant sounds (see Table 2) and rtMRI videos of iso-

Table 1: *Participant characteristics*

ID	age	state of origin	ID	age	state of origin
F1	25	CA	M1	33	WI
F2	25	NY	M2	27	VA
F3	26	CA	M3	28	WI
F4	25	DC	M4	20	CA
F5	28	SC	M5	38	DC
F6	31	HI	M6	24	NJ
F7	64	MN	M7	33	TX
F8	26	TX	M8	26	IA
F9	22	RI			

[*] now at Samsung Medical Center
[†] now at Google, Inc.
[‡] now at Canary Speech, LLC
[§] now at MIT Lincoln Laboratory
[¶] now at Educational Testing Service R&D

http://dx.doi.org/10.21437/Interspeech.2017-608

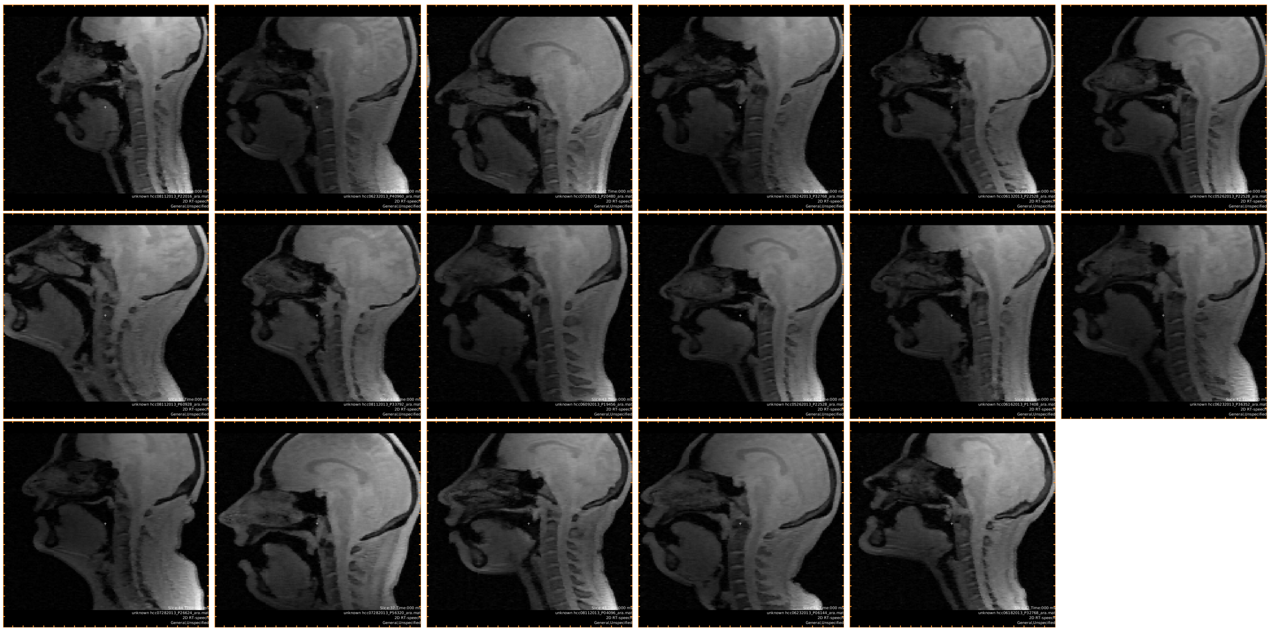


Figure 1: *Slices of 3D volumetric MRI showing inter-speaker variation in the midsagittal vocal tract shape for American English [ɹ] from the 17 speakers. Each 3D volumetric MRI was acquired in one 7-second scan in which the participant sustained the sound.*

lated consonant-vowel-consonant utterances, vowel-consonant-vowel utterances, passages (neutral, fast, clear, whispered, yelling), and spontaneous speech (see Table 3). After completing a session, the participant was paid for their participation in the study. The USC Institutional Review Board approved the data collection procedures.

2.2. 3D volumetric MRI acquisition

The 3D volumetric MRI sequence captured the 3D volume of the upper airway in 7 s [7, 8]. Participants did not report experiencing difficulty sustaining the continuant phonemes of English for 7 s.

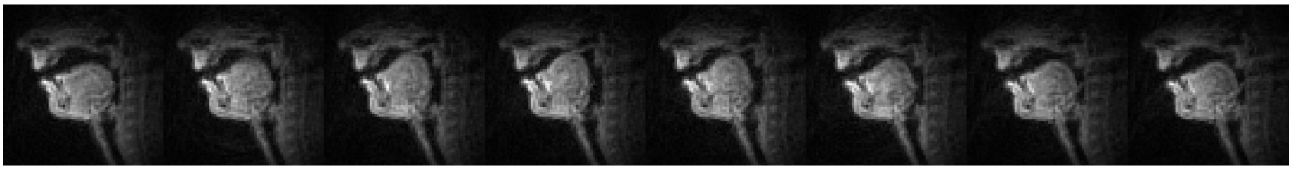
Data were acquired on a $3.0\,\mathrm{T}$ Signa Excite HD MRI scanner (GE Healthcare, Waukesha, WI) with gradients capable of $40\,\mathrm{mT/m}$ amplitudes and $150\,\mathrm{mT/m/ms}$ slew rates. A body coil was used for RF transmission, and an 8-channel neurovascular array coil was used for signal reception. Only the 4 superior elements were used for reconstruction. The vocal tract region of interest (ROI) was imaged using a midsagittal slice with $8\,\mathrm{cm}$ thickness in the right-left (R-L) direction. The readout direction was superior-inferior (S-I), and the phase encode directions were anterior-posterior (A-P) and right-left (R-L). A gradient echo (GRE) sequence was used with TE=$2.3\,\mathrm{ms}$, TR=$4.7\,\mathrm{ms}$, $10°$ flip angle, $\pm 125\,\mathrm{kHz}$ receiver bandwidth ($4\,\mu\mathrm{s}$ sampling rate), NEX=1, $1.33\,\mathrm{mm} \times 1.33\,\mathrm{mm} \times 1.33\,\mathrm{mm}$ spatial resolution, and $20\,\mathrm{cm} \times 24\,\mathrm{cm} \times 8\,\mathrm{cm}$ FOV. Additional technical specifications for the 3D volumetric MRI acquisition and reconstruction are reported in [8]. Figure 1 presents 17 speakers producing American English [ɹ], showing midsagittal slices of the 3D volumetric image.

2.3. Real-time MRI acquisition

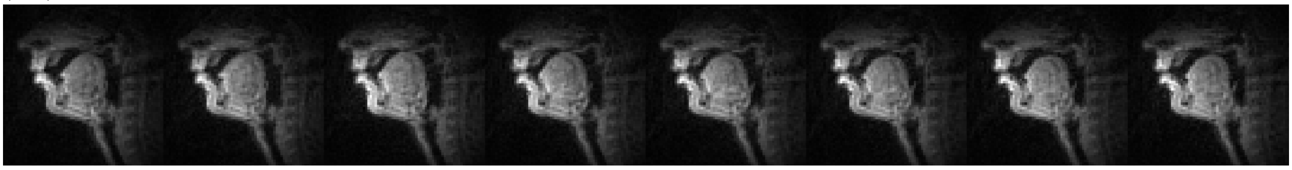
Data were acquired on a Signa Excite HD $1.5\,\mathrm{T}$ scanner (GE Healthcare, Waukesha WI) with gradients capable of $40\,\mathrm{mT/m}$ amplitude and $150\,\mathrm{mT/m/ms}$ slew rate. A body coil was used for radio frequency (RF) signal transmission. A custom upper airway receiver coil array was used for RF signal reception. This 4-channel array included two anterior coil elements and two coil elements posterior to the head and neck. Only the two anterior coils were used for data acquisition because the posterior coils of this hardware were shown to result in aliasing artifacts. The rtMRI acquisition protocol was based on a spiral fast gradient echo sequence. Thirteen interleaved spirals together formed a single image. Each spiral was acquired over $6.164\,\mathrm{ms}$ (repetition time, TR, which includes slice excitation, readout, and gradient spoiler), and thus every image comprises information spanning 13×6.164 = $80.132\,\mathrm{ms}$. A sliding window technique was used to allow for view sharing and thus to increase frame rate [9, 10]. The TR-increment for view sharing was 7 acquisitions, which resulted in the generation of an MRI video with frame rate $1/(7 \times \mathrm{TR}) = 1/(7 \times 6.164\,\mathrm{ms}) = 23.18\,\mathrm{frames/s}$. The imaging sequence had $15°$ flip angle, $\pm 125\,\mathrm{kHz}$ receiver bandwidth, one $5\,\mathrm{mm}$ midsagittal slice, $2.9\,\mathrm{mm}^2/\mathrm{pixel}$ in-plane spatial resolution, and $200\,\mathrm{mm} \times 200\,\mathrm{mm}$ FOV. Scan plane localization of the midsagittal slice was performed using RTHawk (HeartVista, Inc., Los Altos, CA), a custom real-time imaging platform [11]. Additional technical specifications for the rtMRI acquisition and reconstruction were reported in [3]. Figure 2 exemplifies the rtMRI videos for three vowel-consonant-vowel sequences from a single speaker.

2.4. Audio acquisition

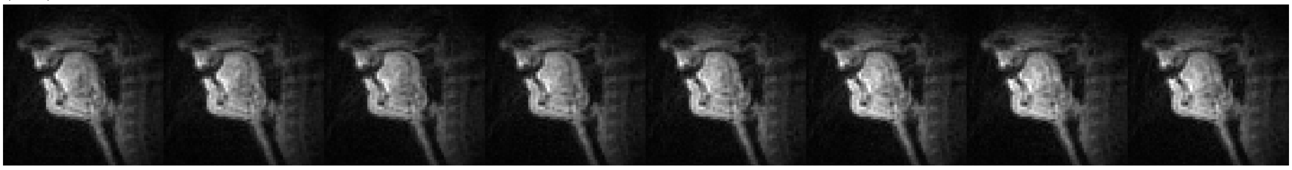
Audio was recorded concurrently with MRI acquisition at a sampling frequency of $100\,\mathrm{kHz}$ inside the MRI scanner bore using a fiber-optic microphone (Optoacoustics Ltd., Moshav Mazor, Israel) and a custom recording and synchronization setup [12]. Synchronization with the video signal was controlled through the use of an audio sample clock derived from the scanner's $10\,\mathrm{MHz}$ master clock and triggered using the scanner RF master-exciter unblank signal. A post-processing



(ɑkɑ)



(uku)



(iki)

Figure 2: *Frames of rtMRI videos for speaker M3 producing [ɑkɑ], [uku], and [iki]. Time progresses from left to right. Coarticulation affects the place of articulation for [k], yielding an anterior closure for [iki] and a posterior closure for [uku] and [ɑkɑ].*

class	sustained sounds
morphological indicators	breathing, hold breath, clench teeth, tongue out, tongue back, tongue tip up
vowel	biːt, bɪt, beit, bɛt, bæt, pɑːt, bʌt, bɔːt, boʊt, buːt, pʊt, bɹ̩d, æbʌt[a]
consonant	ɑfɑ, ɑvɑ, ɑθɑ, aða, ɑsɑ, ɑzɑ, ɑʃɑ, ɑʒɑ, ɑhɑ, ama, ana, aŋa, ala, aɹa

[a] [ʌ] was the sustained and imaged vowel

Table 2: *Speech materials for 3D volumetric MRI.*

step down-sampled the audio to 20 kHz and enhanced the recorded speech using customized de-noising methods (see [12] for more detail). This attenuated the loud scanner noise in the audio recording.

3. Potential research and development use

As Figure 1 illustrates, speakers have diverse vocal tract morphology, which can bring about uniquely individual speech patterns. Differences in craniofacial morphology (often osteological) have long been measured for the purpose of understanding their clinical significance with regard to, for instance, mastication [13, 14], swallowing [15], sleep apnea [16], and development patterns [17]. A growing body of work has looked at the significance of morphological variation to speech production. Previous work with other MRI data-sets has studied speech-relevant structural diversity displayed in terms of the size, shape and relative proportions of the hard palate and posterior pharyngeal wall, aiming to characterize such differences [18], and also to examine how they relate to speaker-specific articulatory and acoustic patterns [19], and to explore the possibility of predicting them automatically from the acoustic signal [20].

Our initial motivation for developing the USC Speech and Vocal Tract Morphology MRI Database was to study how individual differences in vocal tract morphology are reflected in the acoustic speech signal and what articulatory strategies are adopted in the presence of morphological differences to produce speech sounds. The USC Speech and Vocal Tract Morphology MRI Database has already been used to quantify differences among speakers in how much individual articulators (e.g., jaw versus tongue, jaw versus lips) contribute to linguistically relevant constrictions in the vocal tract [21] and to examine the acoustic effects of the shaping of the epilarynx across speakers [22]. Such studies underscore the potential of the database to help illuminate how and to what degree vocal tract morphology may shape speech articulation and speech signal properties within and across talkers.

4. Author contributions

AL, AT, DB, KN, LG, SN, VR designed the experiments. AL, AT, JK, VR, YK, YZ collected data. AT, TS, ZS prepared the database. TS prepared the manuscript. The USC Speech and Vocal Tract Morphology MRI Database is freely provided for research use at `http://sail.usc.edu/span/morphdb`.

5. Acknowledgements

Work supported by NIH (R01DC007124) and NSF (1514544).

6. References

[1] J. Westbury, P. Milenkovic, G. Weismer, and R. Kent, "X-ray microbeam speech production database," *The Journal of the Acoustical Society of America*, vol. 88, no. S1, pp. S56–S56, 1990. [Online]. Available: http://dx.doi.org/10.1121/1.2029064

[2] A. A. Wrench, "A new resource for production modelling in speech technology," in *Proceedings of the Workshop on Innovations in Speech Processing, Stratford-upon-Avon*, UK, 2001.

[3] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014. [Online]. Available: http://dx.doi.org/10.1121/1.4890284

Table 3: *speech materials for rtMRI*

repetitions	task	materials
3	CVC	sɑs, sus, sis, zɑz, zuz, ziz, ʃɑʃ, ʃuʃ, ʃiʃ, θɑθ, θuθ, θiθ fɑf, fuf, fif, vɑv, vuv, viv, lɑl, lul, lil, ɹ ɑɹ , ɹ uɹ , ɹ iɹ
3	VCV	ɑpɑ, upu, ipi, ɑtɑ, utu, iti, ɑkɑ, uku, iki, ɑbɑ, ubu, ibi, ɑdɑ, udu, idi, ɑgɑ, ugu, igi ɑθɑ, uthu, ithi, ɑsɑ, usu, isi, ɑʃɑ, uʃu, iʃi, ɑmɑ, umu, imi, ɑnɑ, unu. ini, ɑlɑ, ulu, ili ɑfɑ, ufu, ifi, ɑmɑ, umu, imi, ɑɹ ɑ, uɹ u, iɹ i, ɑhɑ, uhu, ihi, ɑwɑ, uwu, iwi, ɑjɑ, uju, iji
2 × neutral 2 × fast 2 × clear 2 × yell 2 × whisper	Rainbow Passage	When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long, round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.
2	Grandfather Passage	You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an old black frock coat, usually several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a bit. Twice each day, he plays skillfully and with zest upon a small organ. Except in the winter, when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, “Banana oil!” Grandfather likes to be modern in his language.
2	North Wind and the Sun Passage	The North Wind and the Sun were disputing which was the stronger when a traveler came along, wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew, the more closely did the traveler pull his cloak around him, and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveler took off his cloak, and so the North Wind was obliged to confess that the Sun was the stronger of the two.
2	sentences	She had your dark suit in greasy wash water all year. Don’t ask me to carry an oily rag like that. The girl was thirsty and drank some juice, followed by a coke. Your good pants look great. However, your ripped pants look like a cheap version of a K-Mart special. Is that an oil stain on them?
1	spontaneous speech	What is your favorite music? How do you like LA? What is your favorite movie? What are the best places you have been to? What is your favorite restaurant?
1	picture description	5 pictures
1	singing	highest note lowest note
1	miscellaneous	trace palate with tongue tip open mouth wide swallow vowel triangle (i.e., [i]-[a]-[u]-[i])

[4] S. G. Lingala, A. Toutios, J. Töger, Y. Lim, Y. Zhu, Y.-C. Kim, C. Vaz, S. S. Narayanan, and K. S. Nayak, "State-of-the-art mri protocol for comprehensive assessment of vocal tract structure and function," in *Interspeech 2016*, 2016, pp. 475–479. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-559

[5] 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. [Online]. Available: http://dx.doi.org/10.1016/j.ejmp.2014.05.001

[6] A. Toutios and S. S. Narayanan, "Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research," *APSIPA Transactions on Signal and Information Processing*, vol. 5, p. e6, 2016. [Online]. Available: http://dx.doi.org/10.1017/ATSIP.2016.5

[7] Y.-C. Kim, S. S. Narayanan, and K. S. Nayak, "Accelerated three-dimensional upper airway mri using compressed sensing," *Magnetic Resonance in Medicine*, vol. 61, no. 6, pp. 1434–1440, 2009. [Online]. Available: http://dx.doi.org/10.1002/mrm.21953

[8] ——, "Accelerated 3d mri of vocal tract shaping using compressed sensing and parallel imaging," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 389–392. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2009.4959602

[9] ——, "Flexible retrospective selection of temporal resolution in real-time speech mri using a golden-ratio spiral view order," *Magnetic resonance in medicine*, vol. 65, no. 5, pp. 1365–1371, 2011. [Online]. Available: http://dx.doi.org/10.1002/mrm.22714

[10] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004. [Online]. Available: http://dx.doi.org/10.1121/1.1652588

[11] J. M. Santos, G. A. Wright, and J. M. Pauly, "Flexible real-time magnetic resonance imaging framework," in *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 1. IEEE, 2004, pp. 1048–1051. [Online]. Available: http://dx.doi.org/10.1109/IEMBS.2004.1403343

[12] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, 2006. [Online]. Available: http://dx.doi.org/10.1121/1.2335423

[13] V. Toro-Ibacache, V. Z. Muñoz, and P. OHiggins, "The relationship between skull morphology, masticatory muscle force and cranial skeletal deformation during biting," *Annals of Anatomy-Anatomischer Anzeiger*, vol. 203, pp. 59–68, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.aanat.2015.03.002

[14] B. R. Chrcanovic, M. H. N. G. Abreu, and A. L. N. Custódio, "Morphological variation in dentate and edentulous human mandibles," *Surgical and radiologic anatomy*, vol. 33, no. 3, pp. 203–213, 2011. [Online]. Available: http://dx.doi.org/10.1007/s00276-010-0731-4

[15] N. Fakhry, L. Puymerail, J. Michel, L. Santini, C. Lebreton-Chakour, D. Robert, A. Giovanni, P. Adalian, and P. Dessi, "Analysis of hyoid bone using 3d geometric morphometrics: an anatomical study and discussion of potential clinical implications," *Dysphagia*, vol. 28, no. 3, pp. 435–445, 2013. [Online]. Available: http://dx.doi.org/10.1007/s00455-013-9457-x

[16] R. J. Schwab, M. Pasirstein, R. Pierson, A. Mackley, R. Hachadoorian, R. Arens, G. Maislin, and A. I. Pack, "Identification of upper airway anatomic risk factors for obstructive sleep apnea with volumetric magnetic resonance imaging," *American journal of respiratory and critical care medicine*, vol. 168, no. 5, pp. 522–530, 2003. [Online]. Available: http://dx.doi.org/10.1164/rccm.200208-866OC

[17] H. K. Vorperian, S. Wang, M. K. Chung, E. M. Schimek, R. B. Durtschi, R. D. Kent, A. J. Ziegert, and L. R. Gentry, "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study a," *The Journal of the Acoustical Society of America*, vol. 125, no. 3, pp. 1666–1678, 2009. [Online]. Available: http://dx.doi.org/10.1121/1.3075589

[18] A. Lammert, M. Proctor, and S. Narayanan, "Morphological variation in the adult hard palate and posterior pharyngeal wall," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 2, pp. 521–530, 2013. [Online]. Available: http://dx.doi.org/10.1044/1092-4388(2012/12-0059)

[19] ——, "Interspeaker variability in hard palate morphology and vowel production," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 6, pp. S1924–S1933, 2013. [Online]. Available: http://dx.doi.org/10.1044/1092-4388(2013/12-0211)

[20] M. Li, A. Lammert, J. Kim, P. Ghosh, and S. Narayanan, "Automatic classification of palatal and pharyngeal wall shape categories from speech acoustics and inverted articulatory signals," in *ISCA Workshop on Speech Production in Automatic Speech Recognition*, Lyon, France, August 2013.

[21] T. Sorensen, A. Toutios, L. Goldstein, and S. S. Narayanan, "Characterizing vocal tract dynamics across speakers using real-time mri," in *Interspeech 2016*, 2016, pp. 465–469. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-583

[22] E. Godoy, A. Dumas, J. Melot, N. Malyska, and T. F. Quatieri, "Relating estimated cyclic spectral peak frequency to measured epilarynx length using magnetic resonance imaging," in *Interspeech 2016*, 2016, pp. 948–952. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1362