Received: 18 June 2018 | Revised: 25 September 2018 | Accepted: 26 September 2018

DOI: 10.1002/mrm.27570

RAPID COMMUNICATION

Magnetic Resonance in Medicine

3D dynamic MRI of the vocal tract during natural speech

Yongwan Lim[1] | **Yinghua Zhu**[1] | **Sajan Goud Lingala**[3] | **Dani Byrd**[2] | **Shrikanth Narayanan**[1] | **Krishna Shrinivas Nayak**[1]

[1]Ming Hsieh Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, California

[2]Department of Linguistics, Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, California

[3]Department of Biomedical Engineering, College of Engineering, University of Iowa, Iowa City, Iowa

Correspondence
Yongwan Lim, 3740 McClintock Avenue, EEB 400, University of Southern California, Los Angeles, CA, 90089-2564.
Email: yongwanl@usc.edu

Funding Information
National Institutes of Health, Grant/Award Number: R01DC007124; National Science Foundation, Grant/Award Number: 1514544

Purpose: To develop and evaluate a technique for 3D dynamic MRI of the full vocal tract at high temporal resolution during natural speech.

Methods: We demonstrate 2.4 × 2.4 × 5.8 mm^3 spatial resolution, 61-ms temporal resolution, and a 200 × 200 × 70 mm^3 FOV. The proposed method uses 3D gradient-echo imaging with a custom upper-airway coil, a minimum-phase slab excitation, stack-of-spirals readout, pseudo golden-angle view order in k_x-k_y, linear Cartesian order along k_z, and spatiotemporal finite difference constrained reconstruction, with 13-fold acceleration. This technique is evaluated using in vivo vocal tract airway data from 2 healthy subjects acquired at 1.5T scanner, 1 with synchronized audio, with 2 tasks during production of natural speech, and via comparison with interleaved multislice 2D dynamic MRI.

Results: This technique captured known dynamics of vocal tract articulators during natural speech tasks including tongue gestures during the production of consonants "s" and "l" and of consonant–vowel syllables, and was additionally consistent with 2D dynamic MRI. Coordination of lingual (tongue) movements for consonants is demonstrated via volume-of-interest analysis. Vocal tract area function dynamics revealed critical lingual constriction events along the length of the vocal tract for consonants and vowels.

Conclusion: We demonstrate feasibility of 3D dynamic MRI of the full vocal tract, with spatiotemporal resolution adequate to visualize lingual movements for consonants and vocal tact shaping during natural productions of consonant–vowel syllables, without requiring multiple repetitions.

KEYWORDS
dynamic MRI, dynamic speech imaging, golden-angle stack-of-spirals, lateral production, rapid vocal tract shaping, speech articulation

A preliminary version of this manuscript was presented at the 2018 Annual Meeting of ISMRM, Abstract 3541.

1 | INTRODUCTION

Dynamic MRI has emerged as a powerful tool for speech production research[1-3] because of its numerous advantages over other imaging and movement tracking modalities such as x-ray microbeam,[4] electromagnetic articulography,[5] and ultrasound.[6] Speech scientists seek a comprehensive understanding of human vocal tract shaping and its dynamics and now can use dynamic MRI techniques to extract linguistically meaningful patterns in airway constriction dynamics during natural speech.

Generally, dynamic MRI techniques have been limited to one mid-sagittal imaging plane or a few 2D imaging planes.[7-15] This has nevertheless provided good utility to speech scientists due to the fact that important information about "place of articulation," which is critical in linguistic contrasts, can be obtained from constriction details in the mid-sagittal plane (e.g., in phonemes /p/, /t/, and /k/ with constrictions that are located at the lips, alveolar ridge, and velar region). However, vocal tract shaping during speech is enormously complex in geometry and in temporal structuring and cannot be fully understood from mid-sagittal constriction posture along the vocal tract.[16] For example, articulation of English fricative /s/ and lateral approximant /l/ both involve constriction of the tongue tip at the alveolar ridge, but the production of these sounds differ in that [s] has the tongue sides braced and air directed centrally along a groove, while [l] has (1 or both) tongue sides lowered, allowing for lateral airflow channels.[17] Detailed and direct 3D information about airway shape and spatiotemporal dynamics is essential to understanding speech production control and to relating articulation to speech acoustics. In the past, however, shaping imaging for speech has only been available indirectly from mid-sagittal 2D dynamic MRI after transformation to 3D or in static volume from 2D multi-planar imaging or in 3D for non-natural and/or sustained phonation.[18-21]

Recently, several research groups have demonstrated dynamic 3D MRI of the vocal tract.[22-24] Burdumy et al.[22] proposed an imaging method with 200 × 200 × 62 mm^3 spatial coverage using variable density and stack-of-stars radial sampling patterns and measured dynamic modification of articulators during singing and speech tasks. With temporal resolution of 1.3 s, this approach was restricted to relatively slow speech tasks. Fu et al.[24] proposed an imaging method using a combination of 3D cones sampling for a navigator acquisition and Cartesian sampling for image encoding. The method achieved full vocal tract coverage with a high frame rate (166 fps) by using a partially separable model (low-rank constraints) during reconstruction. This approach inherently requires long acquisition times, potentially resulting in several repetitions of speech tasks, and reconstruction performance may depend on a reliable estimation of temporal basis from the navigator.[1] These constraints may limit its application to natural speech tasks. A review of current state-of-the-art MRI protocols for speech production study can be found in Lingala et al.[1]

To address the unmet need for full vocal tract 3D dynamic MRI at high temporal resolution during natural speech, without requiring multiple repetitions of a speech task, we have developed a new technique that achieves 2.4 × 2.4 × 5.8 mm^3 spatial resolution and 61-ms temporal resolution over a 200 × 200 × 70 mm^3 FOV, using parallel imaging and simple spatiotemporal constraints previously validated in the context of 2D dynamic MRI[25] (and used in >50 cases).[10,26] We extend a 2D spiral gradient sequence[10] to 3D by incorporating a slab excitation and adding phase-encoding along the k_z direction and use spatiotemporal finite difference (FD) constrained reconstruction with an empirically optimized penalty.

2 | METHODS

2.1 | Data sampling

Our method uses a pseudo-golden angle (GA) stack-of-spirals sampling pattern. Spiral trajectories balance trade-offs among temporal resolution, spatial resolution, and SNR and have been shown to be robust in speech MRI acquisition.[1,7,10,27,28] Pseudo-GA increment has previously been used in the context of 2D spiral dynamic MRI[10,27] and provides a nearly uniform sampling pattern that allows more reduced side-lobe energies of point spread function and retrospective temporal resolution selection.[27] Most importantly, the pseudo-GA increment (compared to true GA) allows for high quality audio recording because the gradient waveforms and corresponding acoustic noise are periodic. The 2D spiral sequence can be converted to 3D stack-of-spirals sequence by adding phase-encoding lines along the k_z direction. We leverage the pseudo-GA spiral sampling in the k_x-k_y plane.

Figure 1 illustrates the data sampling scheme. A pseudo-GA spiral sampling is used in the k_x-k_y plane and Cartesian sampling is used along the k_z direction. Each spiral is acquired for all k_z phase encodes (linear order) before moving to the next spiral, with a GA increment, $\theta_{GA} = 2\pi \times 2/(\sqrt{5} + 1)$. The spiral angle is reset after N interleaves (e.g., after 12*N TRs with 12 phase encoding lines as illustrated in Figure 1), where N is a periodicity of the pseudo-GA.[10] We use $N = 34$ in this work.

2.2 | Image reconstruction

3D reconstruction is performed slice-by-slice, after inverse Fourier transforming data collected within a temporal window (12 TRs) along the fully sampled k_z direction, as illustrated in Figure 1. Note that it is also possible to perform a full 3D reconstruction instead, but given a very large data set (e.g., 630 samples × 800 spirals × 8 channels × 12 k_z),

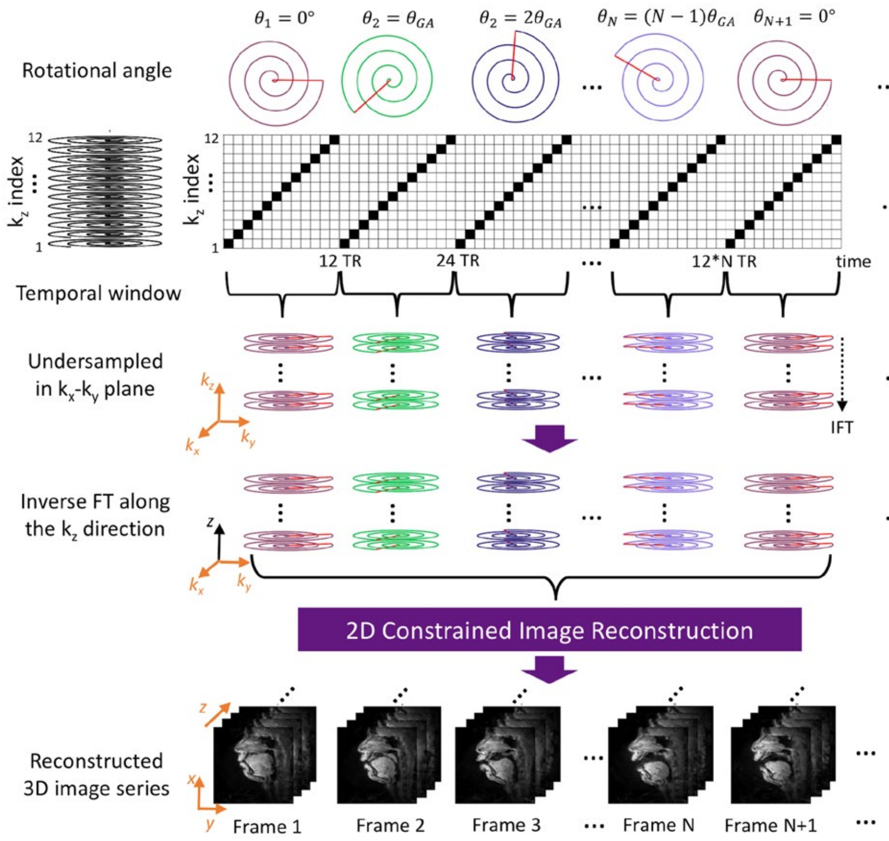
$\theta_1 = 0°$
$\theta_2 = \theta_{GA}$
$\theta_2 = 2\theta_{GA}$
$\theta_N = (N-1)\theta_{GA}$
$\theta_{N+1} = 0°$
Rotational angle
k_z index
12
1
12 TR
24 TR
12*N TR
time
Temporal window
Undersampled in k_x-k_y plane
k_z
k_x
k_y
IFT
Inverse FT along the k_z direction
z
2D Constrained Image Reconstruction
Reconstructed 3D image series
x
y
Frame 1
Frame 2
Frame 3
Frame N
Frame N+1

FIGURE 1 An example of a pseudo golden angle stack-of-spirals sampling scheme for 3D dynamic MRI. Spiral interleave with a rotation angle is acquired for all k_z phase encodes while the k_z step is sequentially increased. After acquiring all of the k_z steps, the rotation angle of spirals is increased by the golden angle, $\theta_{GA} = 2\pi \times 2/(\sqrt{5}+1)$. The spiral angle is reset after N interleaves. Inverse Fourier transform is applied to the data collected within a temporal window along the (fully sampled) k_z direction. Then 2D constrained reconstruction is performed slice-by-slice to form a 3D image series

decoupling the reconstruction into 2D problems is more computationally efficient and practical. We use a sparse SENSE-based parallel imaging and compressed sensing approach with spatiotemporal first-order FD constraints.[25] Regularization parameters for spatial and temporal sparsity (λ_s and λ_t, respectively) are empirically chosen by visual assessment and once calibrated, are held constant for all studies. Coil sensitivity maps are assumed to be time-invariant and are estimated from time-averaged 3D data from each coil by using ESPIRiT.[29] We perform the reconstruction using the Berkeley Advanced Reconstruction Toolbox.[30]

In this report, the full data collection window for each clip (11–25 s) was reconstructed in a single step. Shorter time segments can also be reconstructed, and we report the impact of this segment duration on image quality in the Supporting Information and Supporting Information Video S1.

2.3 | 3D dynamic MRI acquisition

3D slab excitation is achieved by using a minimum phase RF pulse designed with the Shinnar–LeRoux RF design tool software package.[31] The pulse excites a mid-sagittal slab with 5-cm thickness using a flip angle (FA) of 5° and a time-bandwidth product of 16, and stop-band and pass-band ripples of 0.5% and 1%, respectively. The benefit of using the minimum phase pulse is that it can provide a sharp slice profile (higher time-bandwidth product) for a given specification and allows for shorter TE because it has an asymmetric pulse shape and requires a short refocusing gradient.

3D data acquisition was performed using a stack-of-spirals spoiled gradient echo readout with imaging parameters presented in Table 1.

2.4 | 2D multislice dynamic MRI acquisition

For comparison, we also perform 2D pseudo-GA dynamic MRI with 2 or 3 interleaved slices—1 mid-sagittal and 1 or 2 oblique slices—relevant to the speech task,[10] using a previously published approach.[9] The GA increment for the 2- and 3-slice sequence occurred every 2 and 3 TRs, respectively. The periodicity of the pseudo-GA (34 interleaves) was the same as in the 3D sequence. Imaging parameters

TABLE 1 Acquisition parameters for 2D multislice and 3D dynamic MRI protocols

	2D Multislice		3D
FOV (mm^3)	200 × 200 × 6		200 × 200 × 70
FA (°)	15		5
TR (ms)	6.004		5.048
TE (ms)	0.8		0.68
Spatial resolution (mm^3)	2.4 × 2.4 × 6		2.4 × 2.4 × 5.8
Slices (N)	2	3	12 (no. of k_z encodes)
Temporal resolution (ms/frame)	12	18	61
BW (kHz)	±125		
The periodicity of pseudo-GA	34 interleaves		
Interleaves for Nyquist sampling (N)	13 in the k_x-k_y plane		
Acceleration factor for reconstruction	13		

used are listed in Table 1. We reconstruct the dynamic image slice-by-slice by using the sparse SENSE-based reconstruction described in "Image Reconstruction" section with 1 spiral interleave per frame with a reduction factor of 13, which corresponds to temporal resolution of 12 ms per frame and 18 ms per frame for 2- and 3-slice, respectively.

2.5 | In vivo speech experiments

All experiments were performed on a commercial 1.5T scanner (Signa Excite, GE Healthcare, Waukesha, WI) using a real-time interactive imaging platform (RT-Hawk, Heart Vista, Los Altos, CA)[32] with a gradient strength of 40 mT × m^{-1} and a maximum slew rate of 150 mT × m^{-1} × ms^{-1}. A body coil was used for RF transmission, and a custom 8-channel upper airway coil[10] was used for signal reception. The imaging protocol was approved by our Institutional Board. Two healthy adult volunteers were scanned, after providing written informed consent.

All the stimuli were read in the scanner using a mirror projector setup for presentation.[26] Speaker 1 (female American English speaker) was scanned with both the 3D and 2D 3-slice sequences with plane locations as shown in Figure 2A. The stimuli for speaker 1 are listed in Table 2 and were each spoken twice at a natural speech rate. These stimuli deployed two sounds [s] and [l] with contrasting lingual actions: [s] involves tongue sides up and braced and the tongue surface grooved for central airflow, whereas [l] involves tongue sides low, allowing lateral airflow. The stimuli placed [s] and [l] temporally "closer together" or "farther apart" in both orders (i.e., [s] preceding [l] and [l] preceding [s]) creating a direction of lingual action of the tongue sides going from up to down or down to up, respectively. Speaker 2 (male native Korean speaker producing English as a second language) was scanned with both the 3D and 2D 2-slice (1 mid-sagittal and 1axial plane at the level of mid-pharyngeal airway) sequences. This speaker read the English stimuli: "/loo/-/lee/-/la/-/za/-/na/-/za/" repeated twice at a natural rate to produce alternating consonant and vowel sounds. These consonant–vowel syllables use consonants ([l], [z], [n]) made with the tongue tip and relatively extreme vowel postures ("ee" [i], "ah" [a], "oo" [u]) made with the tongue body high and front, low and back, and high and back, respectively.

For speaker 1, audio was recorded inside the scanner simultaneously with data acquisition using a commercial fiber optic microphone (Optoacoustics, Yehuda, Israel) and a custom recording setup.[33] The recorded speech was then enhanced using a dictionary learning-based acoustic denoising method[34] and was synchronized with the reconstructed dynamic images to aid linguistic analysis.

2.6 | Data analysis

2.6.1 | VOI analysis for identifying tongue actions for [l] and [s]

Actions of the tongue tip, sides, and rear (dorsum) are critical in the production of [s] and/or [l], and therefore form the basis of our derived data analysis. In analogy with established region-of-interest analyses,[35-37] volumes-of-interests (VOIs) were designated around 3 vocal tract locations—the tongue tip (TT), dorsum (TD), and tongue sides (TS)—by manually drawing 2D regions-of-interest in the mid-sagittal and axial image planes and extending those regions to adjacent parallel image planes as shown in Figure 3A. Mean pixel intensity was calculated within each of VOI over time. Lingual tissue moving into and out of these VOIs allows the identification of 3 critical lingual gestures for these sounds: a tongue tip raising gesture, a tongue dorsum backing gesture, and a tongue lateral lowering or dipping gesture. Specifically, the actions of these articulatory gestures are expected to reflect:

1. Between /l/ and /s/, the temporal lag or offset between the 2 segmental articulations, which should accord with the phonological "temporal distance" between the target consonants, as organized in Table 2.
2. Within /l/, the relative coordination of the lingual gestures within the articulation of [l]. In particular, this should accord with prior data from other techniques regarding the internal temporal organization of tongue tip and dorsum gestures for [l].[38,39]

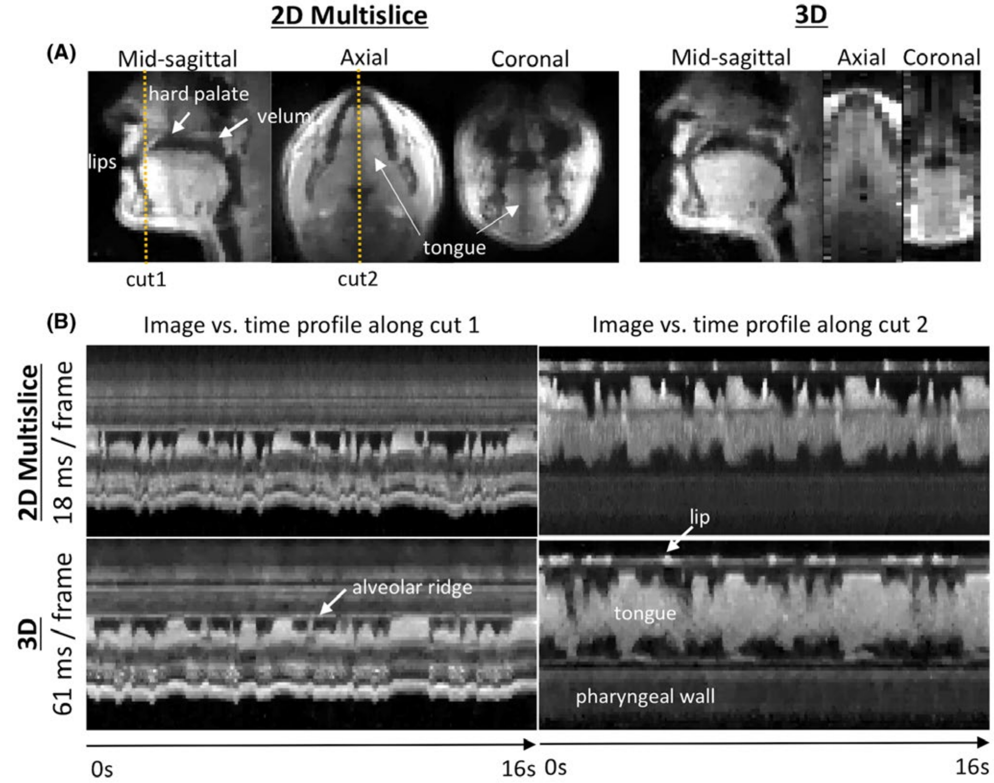

FIGURE 2 Reconstructed images from both 2D multislice and 3D dynamic MRI for speaker 1. (A) Three orthogonal planes (from the left: mid-sagittal, axial, and coronal slices) at consonant /s/ from 2D multislice and 3D dynamic MRI. For comparison purpose, 3 slices are extracted from 3D that would be aligned with those obtained from 2D multislice imaging. See Table 1 for acquisition parameters for both the protocols. (B) Illustration of the tongue movements for speech tasks DU 2–5 listed in Table 2. Two intensity versus time profiles corresponding to the cuts marked by the dot lines in the images in (A) are shown. Both use the same regularization parameters ($\lambda_t = 0.02$ and $\lambda_s = 0.01$)

2.6.2 | Measurement of vocal tract area function

The vocal tract area function is defined as the cross-sectional area of the airway as a function of distance from the glottis and is an important measurement in the study of the relation between vocal tract shaping and acoustics. We tested the ability of 3D dynamic MRI to estimate dynamics of vocal tract area function (using speaker 2's data). From the mid-sagittal plane, we obtained grid lines that were perpendicular to the airway centerline obtained from an airway boundary segmentation method[40] and extracted angled slices along the grid lines through the 3D volume (61 slices with 2-mm increments). From each of the angled slices, we estimated the airway area [cm^2] encompassed by articulator boundaries from a region growing method,[21] applied in this case to the dynamic data. Region growing was performed for each of the angled slices at every time frame independently with seed points automatically chosen as the intersection of the airway centerline from the mid-sagittal plane, and the angled slices. Note that the teeth are not visible in this imaging modality and therefore are not reflected in the area function. The resulting error is temporally constant and appears only at the mouth termination region. Subject-specific dental correction could be performed during post-processing, using additional data that captures the geometry of the teeth.[18,41]

3 | RESULTS

Figure 2 shows representative reconstruction results from 2D multislice and 3D methods for speaker 1's utterances DU 2–5 (see Table 2). The tongue shape at onset of /s/ in the syllable "*sap*" is shown in 3 different views in Figure 2A; constriction of the tongue tip and grooving of the medial tongue surface are clearly observed in the sagittal and coronal slices, respectively, from both results. Figure 2B compares temporal tongue tip dynamics from the 3D result with that from the 2D multislice. The 3D result shares a similar temporal pattern with the tongue tip motion with the 2D multislice result, although it exhibits a slight temporal blurring around the tongue tip compared with its 2D counterpart. Overall, the 3D result provides adequate quality to discern tongue tip actions for articulation of these consonants in this natural speech task.

Figure 3C shows mean pixel intensity curves calculated from 3 VOIs (TT, TD, and TS) over time for stimuli UD

TABLE 2 The stimuli for speaker 1

	Stimuli		"Temporal" distance between [s] and [l]
UD1	Type "a slab," Abigail	[.sl]	Adjacent in same syllable (cluster)
UD2	Type "pass lab," Abigail	[s.l]	Adjacent across a word boundary
UD3	Type "a Sal," Abigail	[.sVl.]	Vowel intervening (in monosyllable)
UD4	Type "a say lab," Abigail	[.sV.l]	Vowel intervening (in disyllable)
UD5	Type "a sap lab," Abigail	[sVC.lV]	Vowel + consonant intervening
DU1[a]	—	[.ls]	same as UD1
DU2	Type "pall sap," Abigail	[l.s]	same as UD2
DU3	Type "alas," Abigail	[.lVs.]	same as UD3
DU4	Type "a lay sap," Abigail	[.lV.s]	same as UD4
DU5	Type "a lab sap," Abigail	[lVC.sV]	same as UD5

Abbreviations: UD and DU, directions of movements; UD, sides up (groove) to sides down (lateral); DU, the reverse of UD; ., a syllable and/or word boundary; V, a stressed vowel; C, a consonant.

[a]DU1 (the word-initial cluster [ls]) does not exist in English.

1, 3, 5, and DU 2, 3, 5. The temporal positions of /s/ and /l/ are measured at their TT mean intensity peaks (i.e., the maximum constriction) as annotated on the time functions in Figure 3C. It is clearly apparent that the articulation /s/ and /l/ are temporally close in "*a slab*" and "*pall sap*" and become farther away from each other as other vowel and consonant segments intervene between the 2 target consonants. This pattern is consistent with the phonological "temporal" organization of the stimuli as listed in Table 2.

For /s/ the tongue tip raising motion is the sole critical articulation apparent, whereas for /l/ co-articulation of tongue tip raising, dorsum backing (higher signal in TD), and sides lowering (lower signal in TS) is observed. Interestingly, depending on the position of /l/ in the syllable, distinct spatiotemporal characteristics are observed for the gestures of /l/. In a syllable-final /l/ (e.g., "*a sal*" and "*pall sap*"), the tongue dorsum backing is extended for a longer period of time and is more spatially extreme than in a syllable-initial /l/ (e.g., "*a sap lab*" and "*a lab sap*"). Similarly, the tongue sides are lowered more in a syllable-final /l/ than in a syllable-initial /l/ as indicated with up-down arrows in Figure 3C. The word-internal, intervocalic ambisyllabic /l/ in "*alas*" shows an intermediate behavior in this regard. In terms of [l]'s internal gestural coordination, its 3 lingual gestures begin almost simultaneously in syllable-initial position, whereas in syllable-final position the tongue dorsum backing and tongue sides lowering start earlier than the tongue tip raising gesture, leading to a timing lag. This syllable-driven coordination asymmetry has previously been observed for tongue tip-dorsum coordination using point tracking kinematic data on [l][38,39]; the proposed protocol not only replicates this finding but also provides new quantitative evidence of a parallel coordination asymmetry involving the tongue sides.

Figure 4 shows a direct comparison of tongue shape for /l/ versus /s/ in the phrase "*pall sap*" for speaker 1. For both segments, constriction of the tongue tip at the alveolar ridge can be observed in the mid-sagittal images. For /l/, side channels are visible in the axial and coronal slices, as well as tongue body retraction in the coronal slices, all of which funnel air laterally along the tongue sides, whereas for /s/, the tongue is grooved mid-sagittally as shown in the coronal slices, channeling the airstream anteriorly toward the front teeth.

Figure 5 shows vocal tract area function dynamics for the utterances of speaker 2. Critical constriction events are visible along the length of the vocal tract. Specifically, when consonants /l/, /z/, and /n/ are articulated (e.g., frames 12, 27, 39, 52, 65, 79 shown in Supporting Information Video S4), the relatively rapid tongue tip constrictions used to create these consonants are clearly shown in the area function dynamics (grid line 3). And, when the vowel /ee/ is articulated (frames 31–34 and 117–122), vocalic tongue body constrictions are observable in the palatal region (grid lines 4–7), as is pharyngeal volume expansion (grid lines 13–15) associated with / ee/'s tongue body fronting.

4 | DISCUSSION

We have demonstrated a dynamic 3D imaging technique that provides complete spatial coverage of the human vocal tract, with spatiotemporal resolution adequate to visualize lingual tongue movements occurring during natural speech without the need for task repetition and with results comparable to interleaved multislice 2D dynamic MRI. Based on data obtained using this proposed technique, we developed a VOI analysis to characterize the coordination of tongue gestures for consonants /l/ and /s/. Earlier point-tracking techniques have established that coordination of the tongue tip and dorsum gestures for American English /l/ varies as a function of syllable position.[38,39,42] To our knowledge, the work presented here provides for the first time quantitative imaging data on the magnitude and duration of tongue side movement and on its relative timing variation with respect to the other lingual gestures comprising /l/. Additionally, this technique has allowed us to quantify dynamic vocal tract area functions during natural productions of consonant–vowel syllables

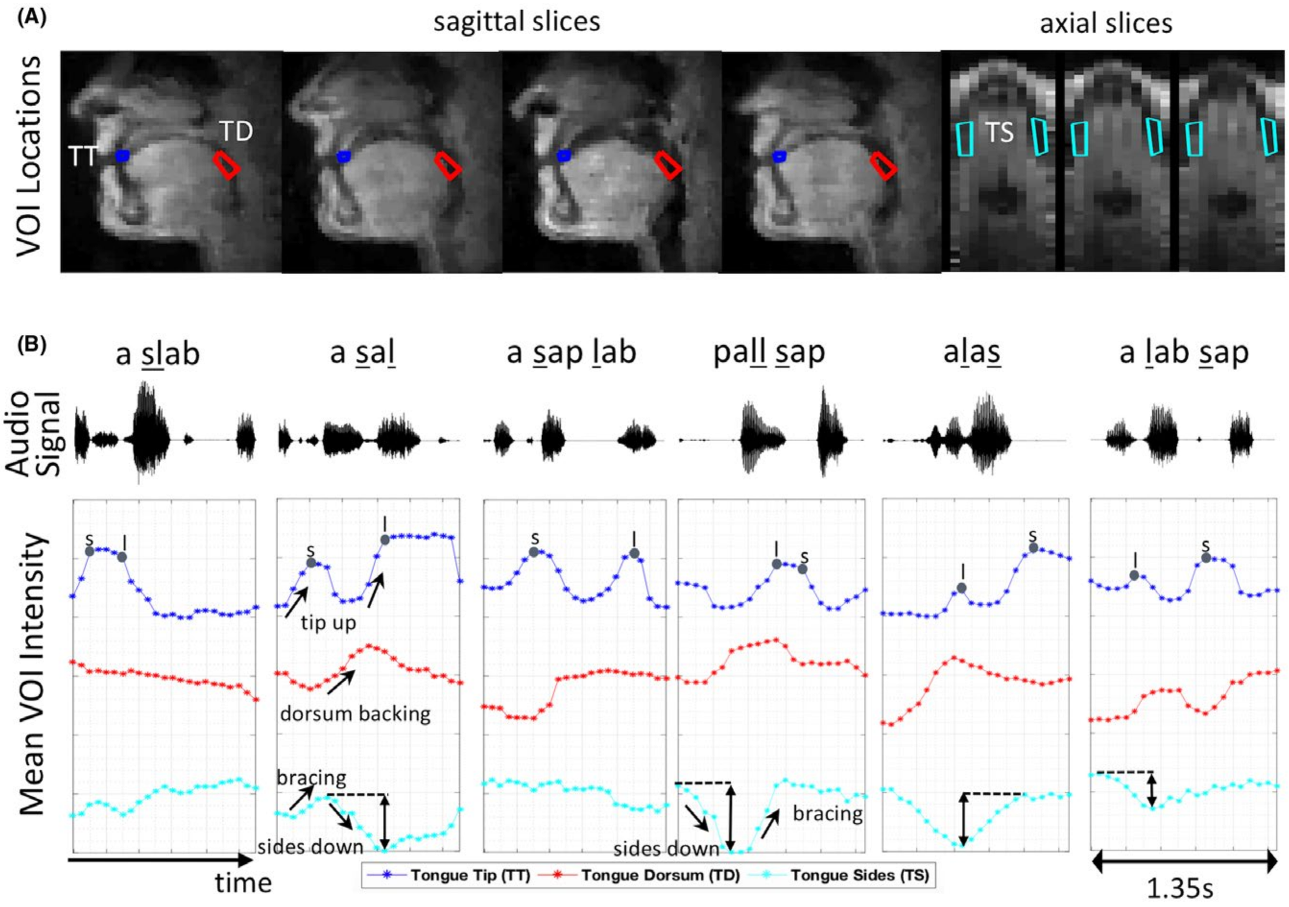

FIGURE 3 VOI analysis for identifying tongue action for [l] and [s]. (A) Placement of VOIs at the tongue tip (blue), back (red), and sides (cyan) overlaid on sagittal and axial images. Illustration of (B) the synchronized denoised audio signals and (C) mean intensity for 3 VOI locations over time for different stimuli. Mean intensity over time was calculated within each of VOIs shown in (A). Each time window corresponds to 1.35 s with a temporal resolution of 61 ms

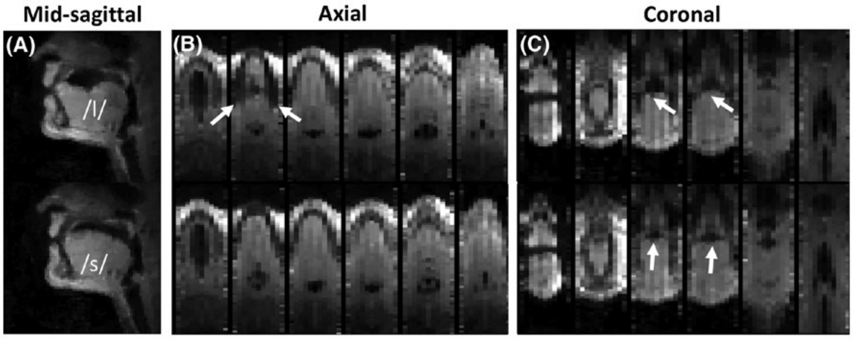

FIGURE 4 Comparison of the vocal tract shape between /l/ and /s/ in the context of "*pall sap*" for speaker 1. (A) Mid-sagittal, (B) axial, and (C) coronal views. For both (top) /l/ and (bottom) /s/, mid-sagittal images show constriction of the tongue tip at the alveolar ridge. For /l/, side channels are shown in the axial and coronal slices, as well as the retraction of the tongue rear in the coronal slices; whereas for /s/, grooving of the tongue is shown in the coronal slices. See also Supporting Information Videos S2 and S3

having varied consonants articulated with the tongue tip and vowels with varied tongue postures. These area functions show a conservation relation between the changes in area function at different parts of the vocal tract, which is expected to be the case.[43]

Validation of our proposed technique is challenging because vocal tract shaping during speech, unlike cardiac or respiratory motion, is not cyclic, and intra-speaker variability makes it difficult to compare the results between methods in a reproducible way (although see Fu et al.[24]). Even

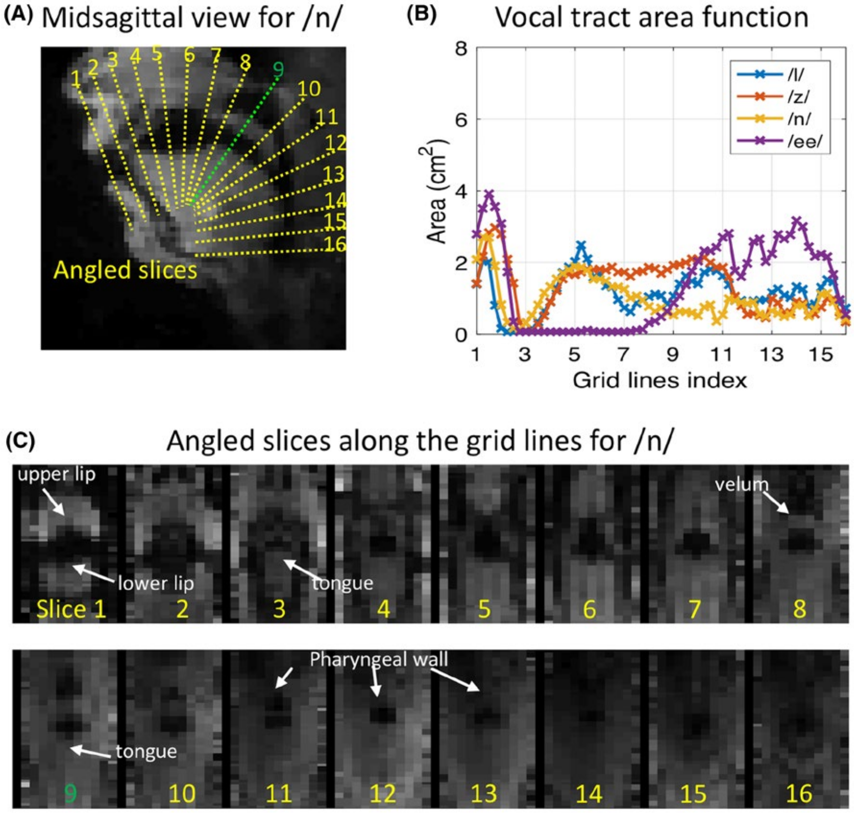

FIGURE 5 Illustration of the capability of estimation of vocal tract area function from 3D dynamic MRI for the "*na*" utterance of speaker 2. (A) Image at the midsagittal plane for /n/ in *"na"* from dynamic 3D. Grid lines that are perpendicular to the airway centerline are chosen to obtain angled slices shown in (C) (only 16 of the 61 gridlines are shown here). (B) The vocal tract area functions for /l/, /z/, /n/, and /ee/ estimated from the 61 angled slices. See also Supporting Information Video S4

after acoustic alignment, the quality of retrospective CINE 3D MRI of the vocal tract is poor.[44] In this work, we use the multislice 2D dynamic MRI as an image quality reference because it provides the current best data quality for natural speech in our experience. However, this 2D method lacks information beyond the acquired (usually mid-sagittal) slices, and this method is applied during a separate production of the speech task. Further validation may be possible with numerical 3D vocal tract phantoms that allow realistic simulation of fluent speech, repeated speech utterances, and flexibility of varying speech rate, or with simultaneous acquisition of MRI with another modality such as optical endoscopy.

This work should be considered an initial demonstration of feasibility. The parameters chosen, specifically the spatial and temporal resolutions, may not be optimal for all speech tasks or regions of interest. Higher spatial resolution in the slice (left–right) direction may be needed to precisely measure the vocal tract area function or to precisely identify the borders of smaller articulators such as the velum. Higher temporal resolution may be desirable for study of rapid speech tasks such as alveolar trills taking place on the order of tens of milliseconds.[1]

In addition, the RF pulse used for 3D slab excitation may be further improved. The slab thickness was designed to be 2 cm thinner than FOV along the slice direction. This margin along with a high TBW allows the avoidance of aliasing in the slice direction caused by a transition in the slab profile and/or shifts by resonant offsets that can be up to ±625 Hz at 1.5T at air–tissue interfaces. It is possible that the margin can be reduced. Likewise, there may be room to reduce the TBW and/or use variable-rate selective excitation pulse,[45] which would allow for shorter pulse duration and shorter TR.

Speech production experiments require that the scan operator be able to monitor the articulatory movements to identify when there might be a substantial unexpected change in head positioning and to identify when speech utterances have been performed correctly per instructions. In the proposed method, these requirements are fulfilled by lower-quality zero-filled linear reconstructions with mediocre temporal resolution (303 ms/frame) and low reconstruction latency (<10 ms/frame), which were not shown here. Detailed linguistic analysis and computational modeling of speech MRI is almost always performed off-line,[46] permitting a high-quality and high-latency reconstruction before processing. The constrained reconstruction temporal window is the only fundamental limit on latency for the proposed method. We found that adequate image quality can be achieved with a temporal window of ≥16 frames (976 ms) (see Supporting Information Video S1). This indicates that the ultimate minimum latency of the proposed method is ~1 s. In the future, this could make it possible to perform real-time analysis with an overall latency of a few seconds.

5 | CONCLUSION

We demonstrated a technique for 3D dynamic imaging of the full vocal tract at high temporal resolution during natural speech. The proposed method uses a minimum-phase 3D slab excitation, pseudo GA stack-of-spirals, and spatiotemporal finite difference constrained reconstruction and achieves 2.4 × 2.4 × 5.8 mm^3 spatial resolution and 61-ms temporal resolution over a 200 × 200 × 70 mm^3 FOV. This technique is evaluated through in vivo imaging of natural speech production from 2 subjects with synchronized audio and via comparison with interleaved multislice 2D dynamic MRI. This promising tool for speech science for the first time enables a quantitative identification of spatial and temporal coordination of important tongue gestures coproduced on and off the midline in the articulation of consonants/l/ and /s/ via VOI analysis and allows a direct assessment of vocal tract area function dynamics during natural speaking of utterances.

ACKNOWLEDGMENTS

This work was supported by NIH Grant R01DC007124 and NSF Grant 1514544. We acknowledge the support and collaboration of the Speech Production and Articulation kNowledge (SPAN) group at the University of Southern California, Los Angeles, CA.

ORCID

Yongwan Lim http://orcid.org/0000-0003-0070-0034
Dani Byrd http://orcid.org/0000-0003-3319-5871
Shrikanth Narayanan http://orcid.org/0000-0002-1052-6204
Krishna Shrinivas Nayak http://orcid.org/0000-0001-5735-3550

REFERENCES

1. Lingala SG, Sutton BP, Miquel ME, Nayak KS. Recommendations for real-time speech MRI. *J Magn Reson Imag*. 2016;43:28–44.
2. Bresch E, Kim YC, Nayak KS, Byrd D, Narayanan SS. Seeing speech: capturing vocal tract shaping using real-time magnetic resonance imaging. *IEEE Signal Process Mag*. 2008;25:123–129.
3. Scott AD, Wylezinska M, Birch MJ, Miquel ME. Speech MRI: morphology and function. *Med Phys*. 2014;30:604–618.
4. Westbury JR. The significance and measurement of head position during speech production experiments using the x-ray microbeam system. *J Acoust Soc Am*. 1991;89:1782–1791.
5. Perkell JS, Cohen MH, Svirsky MA, Matthies ML, Garabieta I, Jackson MT. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *J Acoust Soc Am*. 1992;92:3078–3096.
6. Denby B, Stone M. Speech synthesis from real time ultrasound images of the tongue. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, 2004. pp. I-685-8.
7. Narayanan SS, Nayak KS, Lee S, Sethy A, Byrd D. An approach to real-time magnetic resonance imaging for speech production. *J Acoust Soc Am*. 2004;115:1771–1776.
8. Sutton BP, Conway CA, Bae Y, Seethamraju R, Kuehn DP. Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3 T. *J Magn Reson Imag*. 2010;32:1228–1237.
9. Kim YC, Proctor MI, Narayanan SS, Nayak KS. Improved imaging of lingual articulation using real-time multislice MRI. *J Magn Reson Imag*. 2012;35:943–948.
10. Lingala SG, Zhu Y, Kim YC, Toutios A, Narayanan SS, Nayak KS. A fast and flexible MRI system for the study of dynamic vocal tract shaping. *Magn Reson Med*. 2017;77:112–125.
11. Lingala SG, Zhu Y, Lim Y, et al. Feasibility of through-time spiral generalized autocalibrating partial parallel acquisition for low latency accelerated real-time MRI of speech. *Magn Reson Med*. 2017;78:2275–2282.
12. Niebergall A, Zhang S, Kunay E, et al. Real-time MRI of speaking at a resolution of 33 ms: undersampled radial FLASH with nonlinear inverse reconstruction. *Magn Reson Med*. 2013;69:477–485.
13. Scott AD, Boubertakh R, Birch MJ, Miquel ME. Towards clinical assessment of velopharyngeal closure using MRI: evaluation of real-time MRI sequences at 1.5 and 3 T. *Br J Radiol*. 2012;85:e1083–e1092.
14. Kim YC, Hayes CE, Narayanan SS, Nayak KS. Novel 16-channel receive coil array for accelerated upper airway MRI at 3 Tesla. *Magn Reson Med*. 2011;65:1711–1717.
15. Fu M, Barlaz MS, Shosted RK, Liang ZP, Sutton BP. High-resolution dynamic speech imaging with deformation estimation. *Conf Proc IEEE Eng Med Biol Soc*. 2015;1568–1571.
16. Narayanan SS, Byrd D, Kaun A. Geometry, kinematics, and acoustics of Tamil liquid consonants. *J Acoust Soc Am*. 1999;106:1993–2007.
17. Stone M. Toward a model of three-dimensional tongue movement. *J Phon*. 1991;19:309–320.
18. Story BH, Titze IR, Hoffman EA. Vocal tract area functions from magnetic resonance imaging. *J Acoust Soc Am*. 1996;100:537–554.
19. Martins P, Carbone I, Pinto A, Silva A, Teixeira A. European Portuguese MRI based speech production studies. *Speech Commun*. 2008;50:925–952.
20. Kim Y, Kim J, Proctor M, Toutios A, Nayak K, Lee S, Narayanan S. Toward automatic vocal tract area function estimation from accelerated three-dimensional magnetic resonance imaging. In Proceedings of the Interspeech 2013 Workshop on Speech Production in Automatic Speech Recognition, Lyon, France, 2013. pp. 2–5.
21. Skordilis ZI, Toutios A, Toger J, Narayanan SS. Estimation of vocal tract area function from volumetric Magnetic Resonance Imaging. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, 2017. pp. 924–928.
22. Burdumy M, Traser L, Burk F, et al. One-second MRI of a three-dimensional vocal tract to measure dynamic articulator modifications. *J Magn Reson Imag*. 2017;46:94–101.
23. Zhu Y, Toutios A, Narayanan SS, Nayak KS. Faster 3D vocal tract real-time MRI using constrained reconstruction. In Proceedings of the Annual Conference of Interspeech, Lyon, France, 2013. pp. 1292–1296.

24. Fu M, Barlaz MS, Holtrop JL, et al. High-frame-rate full-vocal-tract 3D dynamic speech imaging. *Magn Reson Med*. 2017;77:1619–1629.
25. Chen J, Lingala SG, Lim Y, Toutios A, Narayanan SS, Nayak KS. Task-based optimization of regularization in highly accelerated speech RT-MRI. In Proceedings of the 25th Annual Meeting of ISMRM, Honolulu, HI, 2017. p. 1409.
26. Lingala SG, Toutios A, Toger J, et al. State-of-the-art MRI protocol for comprehensive assessment of vocal tract structure and function. In Proceedings of the Annual Conference of Interspeech, San Francisco, CA, 2016. pp. 475–479.
27. Kim YC, Narayanan SS, Nayak KS. Flexible retrospective selection of temporal resolution in real-time speech MRI using a golden-ratio spiral view order. *Magn Reson Med*. 2011;65:1365–1371.
28. Lim Y, Lingala SG, Narayanan SS, Nayak KS. Dynamic off-resonance correction for spiral real-time MRI of speech. *Magn Reson Med*. 2018. doi:https://doi.org/10.1002/mrm.27373.
29. Uecker M, Lai P, Murphy MJ, et al. ESPIRiT - an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA. *Magn Reson Med*. 2014;71:990–1001.
30. Tamir JI, Ong F, Cheng JY, Uecker M, Lustig M. Generalized magnetic resonance image reconstruction using the Berkeley advanced reconstruction toolbox. ISMRM Workshop on Data Sampling and Image Reconstruction, Sedona, AZ, 2016.
31. Pauly J, Roux PL, Nishimura D, Macovski A. Parameter relations for the Shinnar-Le Roux selective excitation pulse design algorithm. *IEEE Trans Med Imaging*. 1991;10:53–65.
32. Santos JM, Wright GA, Pauly JM. Flexible real-time magnetic resonance imaging framework. *Conf Proc IEEE Eng Med Biol Soc*. 2004;2:1048–1051.
33. Bresch E, Nielsen J, Nayak KS, Narayanan SS. Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *J Acoust Soc Am*. 2006;120:1791–1794.
34. Vaz C, Ramanarayanan V, Narayanan SS. Acoustic denoising using dictionary learning with spectral and temporal regularization. *IEEE/ACM Trans Audio Speech Lang Process*. 2018;26:967–980.
35. Proctor MI, Lammert A, Katsamanis A, Goldstein L, Hagedorn C, Narayanan SS. Direct estimation of articulatory kinematics from real-time magnetic resonance image sequences. In Proceedings of the Annual Conference of Interspeech, Florence, Italy, 2011. pp. 281–284.
36. Lammert A, Ramanarayanan V, Proctor M, Narayanan SS. Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis. In Proceedings of the Annual Conference of Interspeech, Lyon, France, 2013. pp. 959–962.
37. Töger J, Sorensen T, Somandepalli K, et al. Test–retest repeatability of human speech biomarkers from static and real-time dynamic magnetic resonance imaging. *J Acoust Soc Am*. 2017;141:3323–3336.
38. Delattre P. Consonant gemination in four languages: an acoustic, perceptual, and radiographic study part I. *IRAL Int Rev Appl Linguist Lang Teach*. 1971;9:31–52.
39. Sproat R, Fujimura O. Allophonic variation in English /l/ and its implications for phonetic implementation. *J Phon*. 1993;21:291–311.
40. Kim J, Kumar N, Lee S, Narayanan SS. Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. In Proceedings of the 10th International Seminar on Speech Production (ISSP), Cologne, Germany, 2014. pp. 222–225.
41. Traser L, Birkholz P, Flügge TV, et al. Relevance of the implementation of teeth in three-dimensional vocal tract models. *J Speech Lang and Hear Res*. 2017;60:2379.
42. Narayanan SS, Alwan AA, Haker K. Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals. *J Acoust Soc Am*. 1997;101:1064–1077.
43. Iskarous K. Patterns of tongue movement. *J Phon*. 2005;33:363–381.
44. Zhu Y, Kim YC, Proctor MI, Narayanan SS, Nayak KS. Dynamic 3-D visualization of vocal tract shaping during speech. *IEEE Trans Med Imaging*. 2013;32:838–848.
45. Hargreaves BA, Cunningham CH, Nishimura DG, Conolly SM. Variable-rate selective excitation for rapid MRI sequences. *Magn Reson Med*. 2004;52:590–597.
46. Ramanarayanan V, Tilsen S, Proctor M, et al. Analysis of speech production real-time MRI. *Comput Speech Lang*. 2018;52:1–22.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

VIDEO S1 Movie display of comparison of mid-sagittal plane images reconstructed with various time segment durations. From the top left to the bottom right in a Raster order, window sizes are 402 (full length of the dynamic data), 256, 128, 64, 32, 16, 8, 4 frames. This corresponds to 24.5, 15.6, 7.8, 3.9, 1.95, 0.98, 0.49, and 0.24 s, respectively. Difference images (downsized, superimposed on the bottom left side of each panel, and amplified by a factor of 15 for better visualization) are generated by subtracting each of the reconstructed images from the image shown at the top left panel

VIDEO S2 Movie display of 3D dynamic MRI of the vocal tract with a synchronized audio shown in Figure 4. This video shows reformatted image planes of dynamic vocal tract for speaker 1's utterances UD 1–5 listed in Table 2

VIDEO S3 Movie display of 3D dynamic MRI of the vocal tract with a synchronized audio shown in Figure 4. This video shows reformatted image planes of dynamic vocal tract for speaker 1's utterances DU 1–5 listed in Table 2

VIDEO S4 Movie display of the vocal tract area function dynamics estimated from 3D dynamic MRI shown in Figure 5. This video shows the area function dynamics for the utterance of "/loo/-/lee/-/la/-/za/-/na/-/za/" of speaker 2 spoken twice at a natural rate

How to cite this article: Lim Y, Zhu Y, Lingala SG, Byrd D, Narayanan S, Nayak KS. 3D dynamic MRI of the vocal tract during natural speech. *Magn Reson Med*. 2019;81:1511–1520. https://doi.org/10.1002/mrm.27570