

# Improved 3D real-time MRI of speech production

Ziwei Zhao<sup>1</sup>  | Yongwan Lim<sup>1</sup>  | Dani Byrd<sup>2</sup>  | Shrikanth Narayanan<sup>1,2</sup>  |  
Krishna S. Nayak<sup>1</sup> 

<sup>1</sup>Ming Hsieh Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA

<sup>2</sup>Department of Linguistics, Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, USA

## Correspondence

Ziwei Zhao, Ming Hsieh Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California, 3740 McClintock Avenue, EEB 400, Los Angeles, CA 90089-2564, USA.  
Email: ziweiz@usc.edu

## Funding information

National Science Foundation, Grant/Award Number: 1514544; National Institutes of Health, Grant/Award Number: R01-DC007124

**Purpose:** To provide 3D real-time MRI of speech production with improved spatio-temporal sharpness using randomized, variable-density, stack-of-spiral sampling combined with a 3D spatio-temporally constrained reconstruction.

**Methods:** We evaluated five candidate  $(k, t)$  sampling strategies using a previously proposed gradient-echo stack-of-spiral sequence and a 3D constrained reconstruction with spatial and temporal penalties. Regularization parameters were chosen by expert readers based on qualitative assessment. We experimentally determined the effect of spiral angle increment and  $k_z$  temporal order. The strategy yielding highest image quality was chosen as the proposed method. We evaluated the proposed and original 3D real-time MRI methods in 2 healthy subjects performing speech production tasks that invoke rapid movements of articulators seen in multiple planes, using interleaved 2D real-time MRI as the reference. We quantitatively evaluated tongue boundary sharpness in three locations at two speech rates.

**Results:** The proposed data-sampling scheme uses a golden-angle spiral increment in the  $k_x-k_y$  plane and variable-density, randomized encoding along  $k_z$ . It provided a statistically significant improvement in tongue boundary sharpness score ( $P < .001$ ) in the blade, body, and root of the tongue during normal and 1.5-times speeded speech. Qualitative improvements were substantial during natural speech tasks of alternating high, low tongue postures during vowels. The proposed method was also able to capture complex tongue shapes during fast alveolar consonant segments. Furthermore, the proposed scheme allows flexible retrospective selection of temporal resolution.

**Conclusion:** We have demonstrated improved 3D real-time MRI of speech production using randomized, variable-density, stack-of-spiral sampling with a 3D spatio-temporally constrained reconstruction.

## KEYWORDS

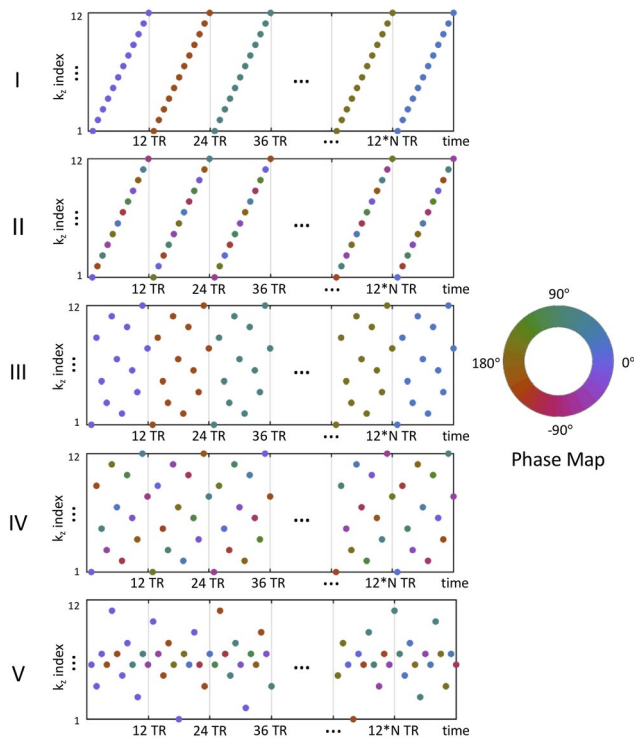
3D real-time MRI, golden angle spiral, speech production, stack-of-spiral sampling, variable-density sampling

Ziwei Zhao and Yongwan Lim contributed equally to this work.

## 1 | INTRODUCTION

Real-time MRI (RT-MRI) has emerged as one of the most powerful tools for studying human speech production. It has allowed researchers to demonstrate both functional and morphological aspects of the vocal tract during speech<sup>1</sup> and to understand the complex spatiotemporal coordination of upper airway structures in motion.<sup>2,3</sup> Some challenges persist, such as choosing an appropriate trade-off between spatial and temporal resolution or between SNR and sampling artifacts. Speech production RT-MRI has benefited from early adoption of technology for rapid and SNR-efficient imaging, such as custom RF coils,<sup>4,5</sup> parallel imaging,<sup>6,7</sup> advanced data sampling,<sup>8,9</sup> constrained reconstruction,<sup>8,10</sup> and automated postprocessing.<sup>11-13</sup> These technologies have continued to provide new insights to linguistics research.

Spatiotemporal requirements vary with speech task, and recommendations have been made based on consensus of linguistics experts (see Figure 1 of Lingala et al<sup>3</sup>). For



**FIGURE 1** Data-sampling patterns. Five sampling strategies are compared. Cases I-V,  $k_z$  versus time plots. Each dot represents one spiral arm; the color represents its initial angle in the  $k_x-k_y$  plane (see color disc). Case I,  $k_z$  is in a linear order, and the golden-angle (GA) increment is applied for every 12 TRs (full  $k_z$  encoding). Case II, The linear order along fully sampled  $k_z$  is maintained while the GA in the  $k_x-k_y$  plane is increased for each TR. Case III, Constant angle remained in the  $k_x-k_y$  plane for 12 TRs; a bit-reversed interleaved temporal ordering is obtained in the  $k_z$ . Case IV, golden-angle increment is applied for each TR, while the temporal order of the  $k_z$  is bit-reversed. Case V,  $k_z$  is acquired in a randomized and variable-density fashion, and the GA increment is applied every 1 TR

visualizing tongue movements, it is recommended to achieve 50-100 ms (10-20 frames per second [fps]) time resolution and  $3.5\text{-mm}^2$  spatial resolution. For fast articulatory movements, such as consonant constriction and coarticulation events, it is recommended to achieve 70 ms (14 fps) time resolution. For this reason, speech RT-MRI protocols are often tailored to the specific speech task.

Two-dimensional midsagittal RT-MRI has had a particularly profound effect on our understanding of speech production.<sup>4,8,9,14,15</sup> It is now a standard and mature method. A typical setup can achieve  $2.4 \times 2.4\text{ mm}^2$  spatial resolution and 12-ms temporal resolution (83 fps).<sup>4</sup> However, there are some limitations. Visualization is limited to a single imaging plane (typically midsagittal), which is inadequate for some articulations<sup>6,16</sup> such as /l/, which have critical maneuvers off the midsagittal plane. Asymmetric behavior of the tongue body, which is only recognized from axial or coronal views, and volumetric properties critical for generating the acoustic resonance structure of speech have also generated researcher interest. These shapes carry important functionality for understanding speech acoustics. Interleaved multiplanar imaging<sup>14</sup> has been achieved with 18-36-ms temporal resolution (28-55 fps), and this provides additional but still incomplete information.

For these reasons, 3D RT-MRI is desirable. The key challenge is the spatial versus temporal resolution tradeoff. Several advanced sampling and reconstruction approaches have been explored to improve on this tradeoff.<sup>17-19</sup> In particular, Lim et al<sup>19</sup> used 3D stack-of-spiral sampling with spatio-temporally constrained reconstruction to achieve full vocal tract imaging with  $2.4 \times 2.4 \times 5.8\text{ mm}^3$  spatial resolution and 61-ms temporal resolution (16 fps). Lim et al also demonstrated the value of 3D volume information to visualize tongue grooving and doming during consonants /s/ and /l/. This did not require any repetition of the utterances. Finer temporal resolution is desirable but would require trading off either in-plane spatial resolution or slice coverage. Improvements in temporal resolution and spatial coverage stand to allow more complete characterization of complex vocal tract geometries. One example is the rhotic alveolar consonant (eg, “r”), which can involve the creation of a sublingual cavity during tongue tip retraction. Another example are lateral consonants (eg, “l”), which involve stretching and narrowing of the tongue, such that the sides of the tongue become lower to allow lateral airflow.

In this study, we explore improvements in achievable spatio-temporal resolution for 3D RT-MRI, based on the method demonstrated by Lim et al.<sup>19</sup> This work used spiral sampling along the  $k_x-k_y$  (sagittal) plane, and linear phase-encode order along  $k_z$  (left-right). Our hypothesis is that the spatio-temporal resolution can be further improved by altering the data-sampling approach.

Rotated spiral interleaves<sup>20,21</sup> and randomized sampling<sup>22,23</sup> in  $(k, t)$  space with variable density<sup>24-26</sup> have previously proved beneficial in volumetric imaging for different

applications. For example, Deng et al<sup>21</sup> used rotated stack-of-spiral sampling to mitigate aliasing artifacts. Nayak et al<sup>23</sup> used random sampling in  $k$ -space to reduce the coherence of undersampling artifacts. Zhou et al<sup>27</sup> explored different angulation strategies in stack-of-stars sampling to experimentally optimize for constrained reconstruction. Liao et al<sup>25</sup> used increased sampling density in spiral trajectories to mitigate motion artifacts in cine MRI. Lee et al<sup>26</sup> investigated applying variable-density stack-of-spiral sampling to significantly reduce the scan time. The relative contribution of these different ideas can best be determined through experiments. To our knowledge, the application of a similar advanced design of 3D RT-MRI to human speech production is original, and investigation of the effects of the aforementioned sampling strategies on the detection of articulatory movements is valuable.

## 2 | METHODS

### 2.1 | Experimental methods

Experiments were performed on a commercial 1.5T scanner (Signa Excite HD; GE Healthcare, Waukesha, WI) with gradients of 40-mT/m strength and 150-mT/m/ms maximum slew rate per axis. Experiments used a body coil for RF transmission and a custom eight-channel upper airway coil<sup>5</sup> for signal reception. We use a real-time interactive imaging platform (RT Hawk; Heart Vista, Los Altos, CA)<sup>28</sup> to perform acquisition and on-the-fly reconstruction. This platform uses a separate computer (Linux laptop) for simple reconstruction (gridding with view sharing) that is used to ensure correct scan prescription and compliance with the stimuli. The proposed reconstruction was performed offline, as described in Section 2.3. We used a spoiled gradient-echo stack-of-spiral trajectory with imaging parameters: spatial resolution =  $2.4 \times 2.4 \times 5.8 \text{ mm}^3$ , FOV =  $200 \times 200 \times 70 \text{ mm}^3$ , TR/TE = 5.05/0.68 ms, and readout duration = 2.52 ms. Spiral readout was used due to its scan efficiency and prior successful applications to speech RT-MRI.<sup>3,4,9,29</sup> Shimming is critical, and our protocol uses autocalibration, followed by manual adjustment of the center frequency by the scan operator to minimize visible blurring of the tongue boundary. The imaging protocol was approved by the University of Southern California's institutional review board, and all subjects provided written, informed consent. Subjects, while being imaged, read stimuli presented by a mirror projector setup.<sup>4</sup>

### 2.2 | Data sampling

Figure 1 illustrates sampling patterns. We consider five different  $(k, t)$  data-sampling patterns (cases I-V), all of which are based on the stack-of-spiral sequence. Case I is

identical to Lim et al,<sup>19</sup> and is taken as the baseline method. The selection of sampling patterns enables pair-wise comparison that elucidates the benefits of each specific data-sampling change.

Case I uses a linear temporal order along  $k_z$  with a constant spiral angle in the  $k_x-k_y$  plane. After  $k_z$  is fully sampled, the spiral angle is incremented by the golden angle (GA),  $\theta_{\text{GA}} = 2\pi \times 2/(\sqrt{5} + 1)$ .<sup>4,14,15</sup> The increment angle is reset after  $N$  interleaves. We use  $N = 34$  in this work. Case II uses the same linear temporal order along  $k_z$  as in case I, but the spiral angle is incremented by GA with the  $k_z$ . Case III introduces the bit-reversed temporal order<sup>30</sup> along  $k_z$  while using the same constant spiral-angle strategy as in case I. Case IV adopts the bit-reversed temporal order along  $k_z$  but with the GA increment in  $k_x-k_y$ . Case V uses GA increment in  $k_x-k_y$  and introduces variable-density random temporal order along  $k_z$ . Bit-reversed temporal order<sup>30</sup> maximizes the time between acquiring adjacent indices to mitigate motion artifacts. Although this is a coherent order, it provides similar benefits to random temporal order. Variable density produces more frequent sampling of low spatial frequencies of  $k_z$ ; specifically, the density along  $k_z$  is implemented as

$$p(k_z) = \frac{1}{|\alpha k_z|}, \quad (1)$$

where  $k_z$  is normalized from  $-0.5$  to  $0.5$ . The scale factor is  $\alpha = \sum_{n=0}^{\frac{N_z}{2}-1} \frac{4(N_z-1)}{1+2n}$ , where  $N_z$  is the number of  $k_z$  steps (slices). We use  $N_z = 12$  for all experiments.

### 2.3 | Image reconstruction

Image reconstruction was performed by solving the following constrained optimization:

$$\arg \min_{f(\mathbf{r}, t)} \|A(f) - \mathbf{b}\|_2^2 + \lambda_s \|TV_s(f)\|_1 + \lambda_t \|FD_t(f)\|_1, \quad (2)$$

where  $f(\mathbf{r}, t)$  represents the dynamic images to be reconstructed;  $A$  is the encoding function (including coil sensitivity and non-Cartesian Fourier transform);  $\mathbf{b}$  is multicoil  $(k, t)$  space measurement data;  $\mathbf{r} \in (x, y, z)$  represents spatial coordinates;  $t$  denotes time;  $TV_s$  represents isotropic 3D spatial total variation;  $FD_t$  represents the first-order temporal finite difference; and  $\lambda_s$  and  $\lambda_t$  denote the corresponding spatial and temporal regularization parameters, respectively.

This optimization problem was solved using the alternating direction method of multipliers algorithm,<sup>31</sup> as implemented in the Berkeley Advanced Reconstruction Toolbox.<sup>32</sup> Coil sensitivity maps were generated using Efficient iTerative Self-consistent Parallel Imaging Reconstruction (ESPIRiT)<sup>33</sup> and were assumed to be

time-invariant. Reconstruction is fully 3D in this work, which allows for greater flexibility in data sampling, but carries substantially higher memory requirements. During the experiment, we acquired 3888–8244 spiral interleaves (TRs) (19.7–41.9 seconds) in each continuous acquisition, depending on the stimuli. This corresponds to 324–687 time frames when reconstructing 12 TRs per frame. We divided each acquisition into multiple time segments (100 time frames per segment, 4–time frame overlap) to manage the memory requirement of the reconstruction.

Reconstruction was performed using *MATLAB* 2017a (MathWorks, Natick, MA) and executed on a 16-core Intel Xeon CPU E5-2698 v3, 2.30 GHz, with 40 MB of L3 cache. Reconstruction time was 73–76 minutes per time segment (100 frames, 6.1 seconds). Note that prior work<sup>19</sup> applied inverse Fourier transform along  $k_z$  and exploited nonlinear reconstruction for each  $x$ – $y$  slice separately, which was possible because  $k_z$  was always fully sampled within each time frame.

All sampling patterns were reconstructed with a temporal resolution of 61 ms per frame (16 fps; 12 TRs). Additionally, we explored a retrospective selection of temporal resolution for case V by reconstructing images with temporal windows of 30.5 ms per frame (33 fps; 6 TRs) and 15.25 ms per frame (66 fps; 3 TRs).

## 2.4 | Regularization parameter selection

We analyzed the sensitivity of regularization parameters  $\lambda_s$  and  $\lambda_t$  in the constrained reconstruction using data acquired from speaker 1 (31-year-old male, native Chinese speaker, English as second language). This speaker was scanned while reading the English stimuli “/loo/-/lee/-/la/-/za/-/na/-/za/,” repeated twice at a natural rate.

Regularization parameters  $\lambda_s$  and  $\lambda_t$  were chosen visually based on image quality in midsagittal views and time-intensity plots by 3 academic linguists with 5 decades of experience using speech RT-MRI. We performed parameter sweeps in two stages. We first performed a coarse sweep spanning 6 orders of magnitude in log scale ( $\lambda_s = 0, 0.001, 0.1$ ;  $\lambda_t = 0.0001, 0.01, 1.0$ ) and then tested a finer region ( $\lambda_s = 0.001, 0.002, 0.004, 0.006, 0.008, 0.010$ ;  $\lambda_t = 0.005, 0.01, 0.02, 0.03, 0.04$ ).

## 2.5 | Experimental optimization

We compared all of the candidate sampling approaches in a prospective experiment with highly repeatable speech tasks produced by 1 volunteer (speaker 2; 18-year-old male, native speaker of American English). The subject was scanned with all five aforementioned 3D RT-MRI data-sampling schemes, as well as interleaved three-slice 2D RT-MRI.<sup>14</sup> Imaging parameters for interleaved 2D RT-MRI are as follows: three

slices,  $2.4 \times 2.4 \text{ mm}^2$  spatial resolution, 6-mm slice thickness,  $200 \times 200 \text{ mm}^2$  FOV, 18-ms temporal resolution (55 fps), and TR/TE = 6.004/0.8 ms. Table 1 lists the speech tasks. Each sentence was spoken twice, once at a “natural” rate and once at a speaking rate of approximately 1.5 times the initial rate, denoted as “speeded.” This stimuli set was designed to test the image quality when we expect large signal fluctuations.

Results from the five candidate sampling patterns (cases I–V; Figure 1) were compared using intensity versus time plots of sagittal, axial, and coronal views. The best sampling scheme was chosen based on boundary sharpness and visual clarity of the dynamic articulators. This (case V) is denoted as the proposed method for the subsequent experiments.

## 2.6 | Speech production experiments

We applied the best-performing sampling and reconstruction approach to a broad set of speech stimuli that were chosen for assessing image improvements. The stimuli were selected to elicit sweeping movements of the tongue body, used to produce relatively wide vocalic airway and rapid movement of the tongue tip coming into contact with the alveolar palate, to produce a consonant constriction.

Two subjects participated in the study of the evaluation for speech application. We evaluated the results from stimuli set 1 (speaker 2; Table 1) using the original (case I) and proposed (case V) methods at 61 ms per frame (16 fps) as well as the reference 2D interleaved method<sup>14</sup> at 18 ms per frame (55 fps). These stimuli were recorded at a normal speaking rate and at a speeded speaking rate. The stimuli set in Table 2 was recorded for speaker 3 (20-year-old male, native speaker of American English), who was scanned with the original, proposed, and reference 2D methods at normal speech rates.

## 2.7 | Speech stimuli

The stimuli set in Table 1 contains four sentences that were designed to elicit an alternating sequence of phonetically low (/æ/, /a/) and high front vowels (/i/, /ɪ/), with no intervening or bordering lingual consonants. These vowel sounds are

**TABLE 1** Stimuli design: Alternating low and high vowels with no intervening lingual consonants

Stimuli	Sequence of vowels
I bop a hip hop beep.	[ai a ə i a i]
I hop a pea poppy.	[ai a ə i a i]
We pop him a ham beef.	[i a i ə æ i]
Peep a happy map meme.	[i ə æ i æ i]

Note: These stimuli were recorded at both normal and speeded (~1.5 times) speaking rates.



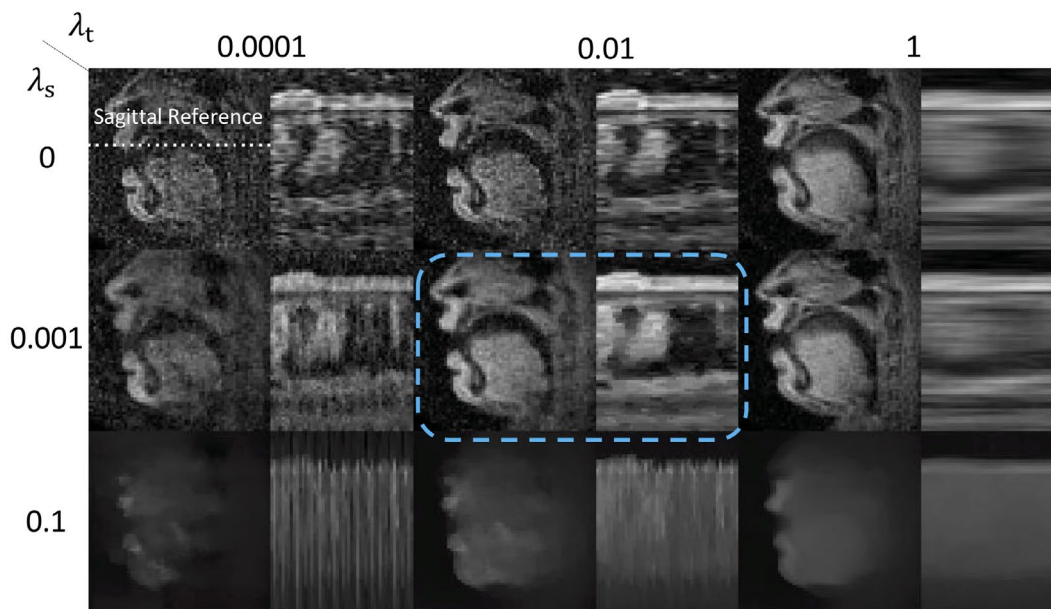
**TABLE 2** Stimuli design: Intrinsically fast tongue-tip consonants

Alveolar consonant	Stimuli		
[r] (falling stress pattern)	word-internal	“I gave ___ a poppy today.”	pita, Otto, Edda
	word-final		Pete, Ott, Ed
	rhotic flap		Peter, otter, Edder
	lateral flap		beatle, bottle, petal
[t] (level stress pattern)		“I gave ___ happily today.”	peat, ott, aid

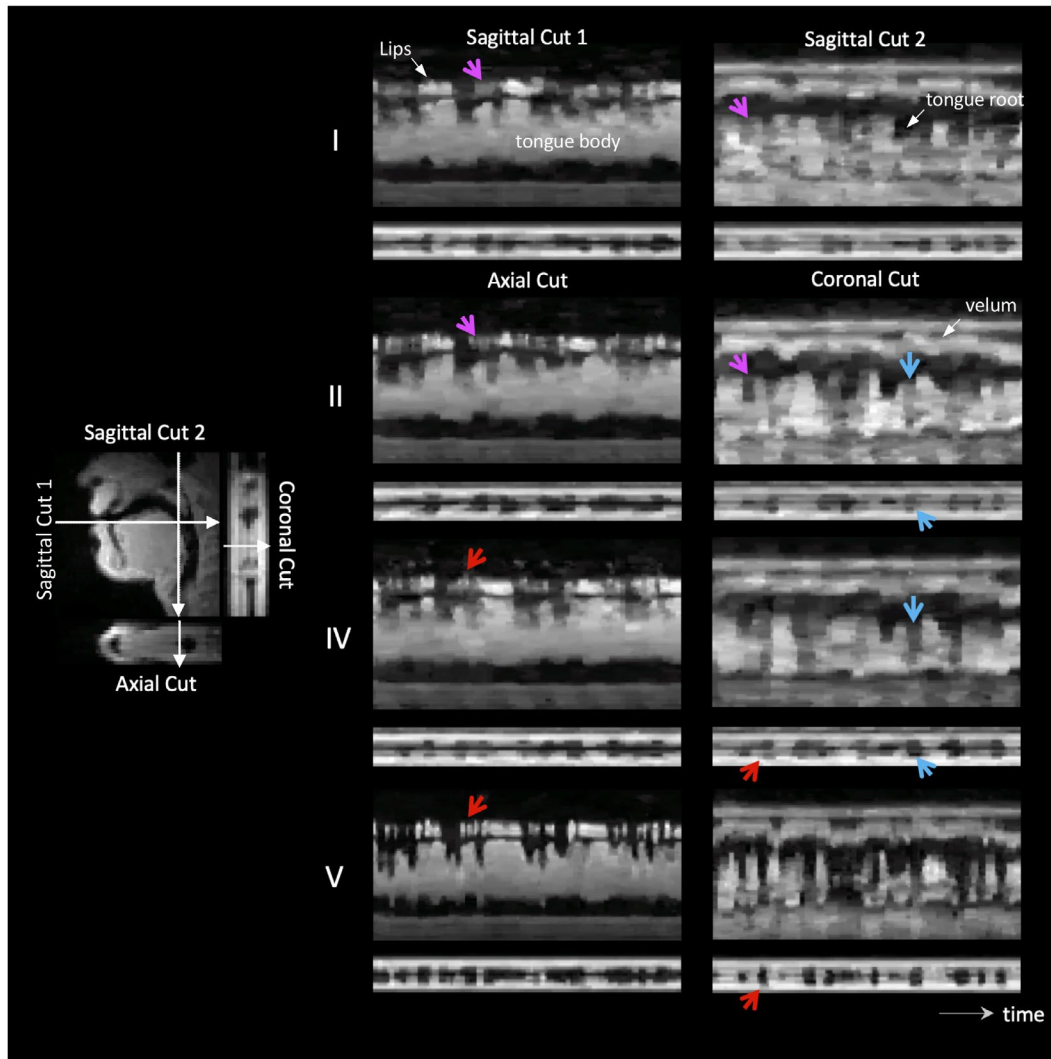
made with tongue postures that are for low vowels retracted and depressed in the mouth, and for high vowels, bunched and domed in the oral palatal cavity. Thus, moving from one type to the other (high to low or low to high) requires the speaker to produce sweeping lingual movements exhibiting large displacements. This allows for the examination of large lingual movements in real, natural speech over a significant spatiotemporal span within the functional vocal tract space. Specifically, in this data set’s sagittal views, the /a/ sound is characterized by tongue retraction producing a narrow constriction of the tongue root in the pharynx, and /æ/ is articulated similarly but with a somewhat flatter tongue surface profile and slightly lesser pharyngeal retraction. For high vowels, /i/ is characterized by a relatively narrow constriction in the oral cavity in close proximity to the hard palate, and /i/ behaves similarly with an only

slightly less constricted posture. These sentence stimuli also contain bilabial stop (closure) consonants (/b/, /p/) intervening between the vowels, during which the lips are approximated, then compressed, and then released, and /h/s, which are articulated without any supralaryngeal constriction. These labial and glottal flanking consonants were chosen so as to minimize any coarticulation with the target vowels by ensuring that the neighboring consonants are not articulated with the tongue. Results were compared in midsagittal and axial views at the time when specific vowels were articulated.

The stimuli set in Table 2 was designed to examine alveolar consonant segments that are created with rapid upward action and closure of the tongue tip articulator; these are among the fastest speech sounds,<sup>34</sup> taking place more rapidly than the vowel articulations elicited in the first stimuli set (Table 1). The recommended minimum temporal resolution for consonant constriction is approximately 60 ms per frame (16 fps) according to Figure 1 in Lingala et al.<sup>3</sup> Fifteen sentences having two different stress patterns were used to elicit the American English alveolar tap/flap consonant [r] (in the falling stress pattern) and eliciting the alveolar consonant [t] (in the level stress pattern [no flapping]). These alveolar consonants are articulated by a single constriction action of the tongue tip, and the tap [r] (eg, the medial consonant in the word “pita”) can be understood as a very rapid stop consonant gesture. The canonical stop consonant [t] (eg the word-final consonant in “peat happily”), whereas still a rapidly articulated tongue tip consonant produces a somewhat longer closure period with greater



**FIGURE 2** Results of parameter sweep during production of “/loo/ - /lee/ - /na/ - /za/ - /la/ - /za/.” From top to bottom, spatial total variation (TV) penalty  $\lambda_s$  ranges from 0, 0.001, to 0.1. From left to right, temporal finite difference (FD) penalty  $\lambda_t$  ranges from 0.0001, 0.01, to 1. A midsagittal view (left) and intensity-time plot (right) with sagittal reference (white dashed line) are shown for each parameter setting. Spatial and temporal blurring is observed as  $\lambda_s$  and  $\lambda_t$  are increased, respectively. Based on the spatiotemporal visualization of articulator movements, we believe that  $\lambda_s$  and  $\lambda_t$  should be chosen in the magnitude of 0.001 and 0.01, respectively (blue dashed boxes). Supporting Information Video S1 contains a denser sampling of the parameter space



**FIGURE 3** Pairwise comparisons. Four intensity-versus-time plots (right) are shown for each sampling strategy. The line locations (left) are shown over sagittal, axial, and coronal view images. Results from four proposed sampling schemes (I, II, IV, and V) are compared. Case V substantially outperforms other methods, providing a high temporal fidelity

lingual contact and compression at the alveolar ridge than for the tap. In addition to word-internal and word-final taps (eg, “pita” vs “Pete a”), two other tap variants are included. A rhotic “flap” is elicited (eg, in the word “Peter”), in which the tongue tip constriction has a retracted posture and trajectory, yielding a significant sublingual cavity. Finally, a laterally released tap is elicited (eg, in the word “petal”), in which the tongue profile is narrowed/stretched front to back (assisted by tongue rear retraction) such that the sides of the tongue lower, potentially pulling away from the palate, so as to allow the lateral airflow required for the following [l] consonant.

## 2.8 | Evaluation

Results were compared using intensity versus time plots during the temporal interval from the end of the vowel /ei/ in the frame sentence’s “gave” to the /a/ or /æ/ vowel in the frame

sentence’s “poppy” or “happily.” The mean value of regions of interest versus time was used to visualize the clarity of detected movements.<sup>35</sup> A  $4 \times 4$  pixel region of interest was placed in the midsagittal views, focusing on tongue-gesture excursion during its maximum consonantal constriction. Intensity normalization among 3D and 2D techniques was done by taking the reference of the middle of the tongue region. Time alignments of different methods were performed manually based on the time frames with the narrowest constriction, such as the tongue tip contacting the alveolar ridge. Note that perfect time alignment across natural tokens cannot be expected or achieved due to normal variations in speech rates, resulting in a drift exhibiting slight temporal misalignments later in time from the constrictions’ anchor point.

We calculated a tongue boundary sharpness score<sup>36</sup> for the original and the proposed methods for stimuli set 1 (Table 1). The steps are illustrated in Figure 8A (see Figure 8A in the Results Section 3.4). The tongue boundary (yellow line in the

midsagittal image frame on the left panel) is extracted over time using a semi-automatic boundary detection method.<sup>37</sup> Gridlines (cyan lines) are constructed perpendicular to the tongue boundary from a linearly interpolated reconstructed image with 8-fold spatial resolution. Intensity profiles (right panel) are obtained along each line. Sharpness score ( $S$ ) for each gridline is calculated as follows:

$$S = \frac{CNR}{d}, \quad (3)$$

where  $CNR$  is the absolute intensity difference between 80% and 20% of the maximum intensity along a gridline divided by the per-pixel noise SD, and  $d$  is the distance, in pixels, between the locations of the 80% and 20% maximum intensity along the gridline.

We selected the blade of the anterior tongue, the body, and the root of the tongue body as locations of interest. Each location includes eight gridlines, and the sharpness score for the eight gridlines are averaged. The evaluated time frames were selected during high-low-high cycles (“hip hop beep” /i/-/a/-/i/; “pea poppy” /i/-/a/-/i/), which start and end at a constricted posture of the tongue body with the hard palate, as is required for a high vowel, and low-high-low cycles (“pop him a ham” /a/-/i/-/æ/; “happy map” /æ/-/i/-/æ/), which start and end at a constricted posture of the tongue root with the rear pharyngeal wall, as is

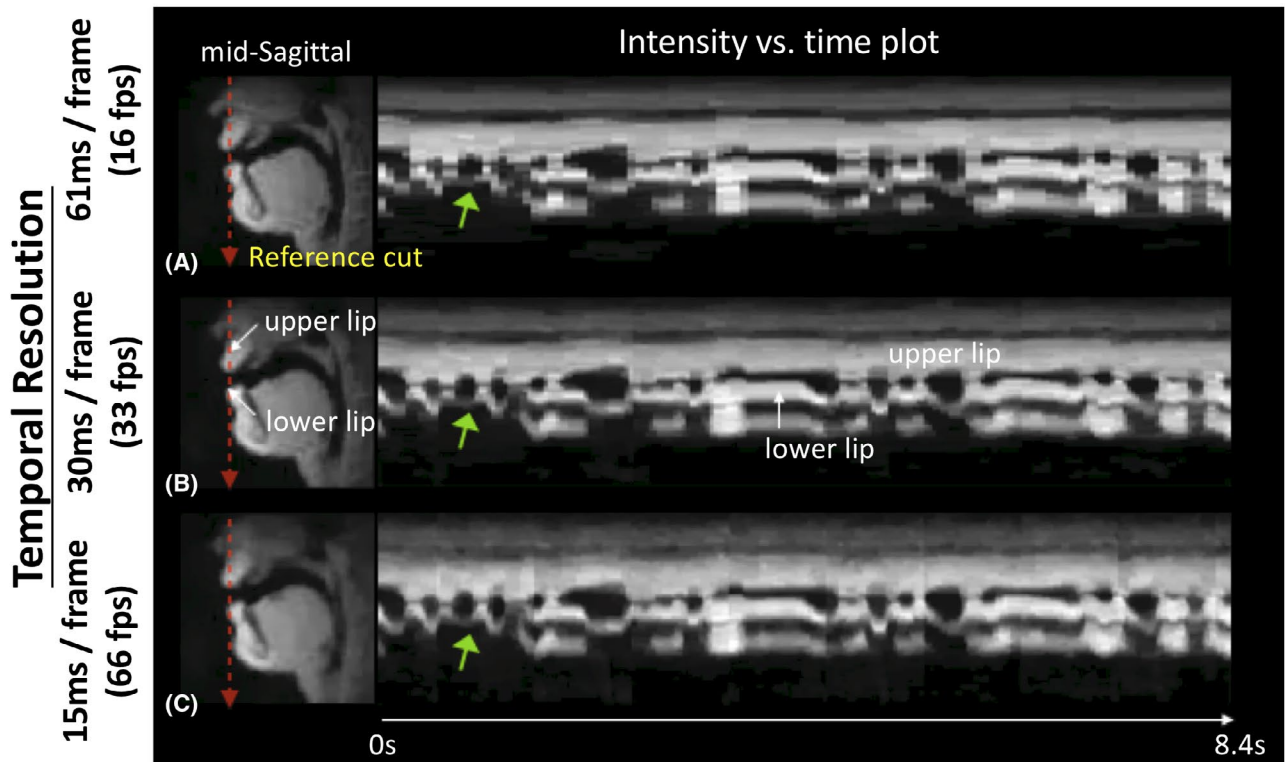
required for a low vowel. We collected these samples in two repetitive experiments at natural and 1.5-times speeded speech rates.

For each location and experiment, we examined the histogram of all sharpness scores over time. The results from the original and proposed methods were assumed to be independent. We tested the null hypothesis that the mean values of the two distributions are equal, using the percentile bootstrap method.<sup>38</sup> The significance level was set to .001.

### 3 | RESULTS

#### 3.1 | Regularization parameter selection

Figure 2 illustrates the selection of regularization parameters  $\lambda_s$  and  $\lambda_t$  in Equation (2) for case I. The spatial regularization term ( $TV_s$ ) controlled by  $\lambda_s$  provides denoising, as visualized in the sagittal plane, but results in excessive spatial smoothing when large (eg,  $\lambda_s = 0.1$ ). The temporal regularization term ( $FD_t$ ) controlled by  $\lambda_t$  suppresses undersampling artifact and recovers intensity changes but results in excessive temporal blurring when large ( $\lambda_t = 1$ ), as best visualized in the intensity-time plots. Supporting Information Video S1 contains results of parameter sweeps performed on a much finer scale. Based on a consensus of 3 expert readers, we chose



**FIGURE 4** Retrospective selection of temporal resolution. (Left) Sagittal view. (Right) Intensity-versus-time plot with reference cut (red dashed lines). We reconstructed stimuli videos using stimuli set 1 at a speeded speech rate with 61 ms per frame (16 frames per second [fps]) (A), 30 ms per frame (33 fps) (B), and 15 ms per frame (66 fps) (C) temporal resolution using  $\lambda_s = 0.008$  and  $\lambda_t = 0.03$

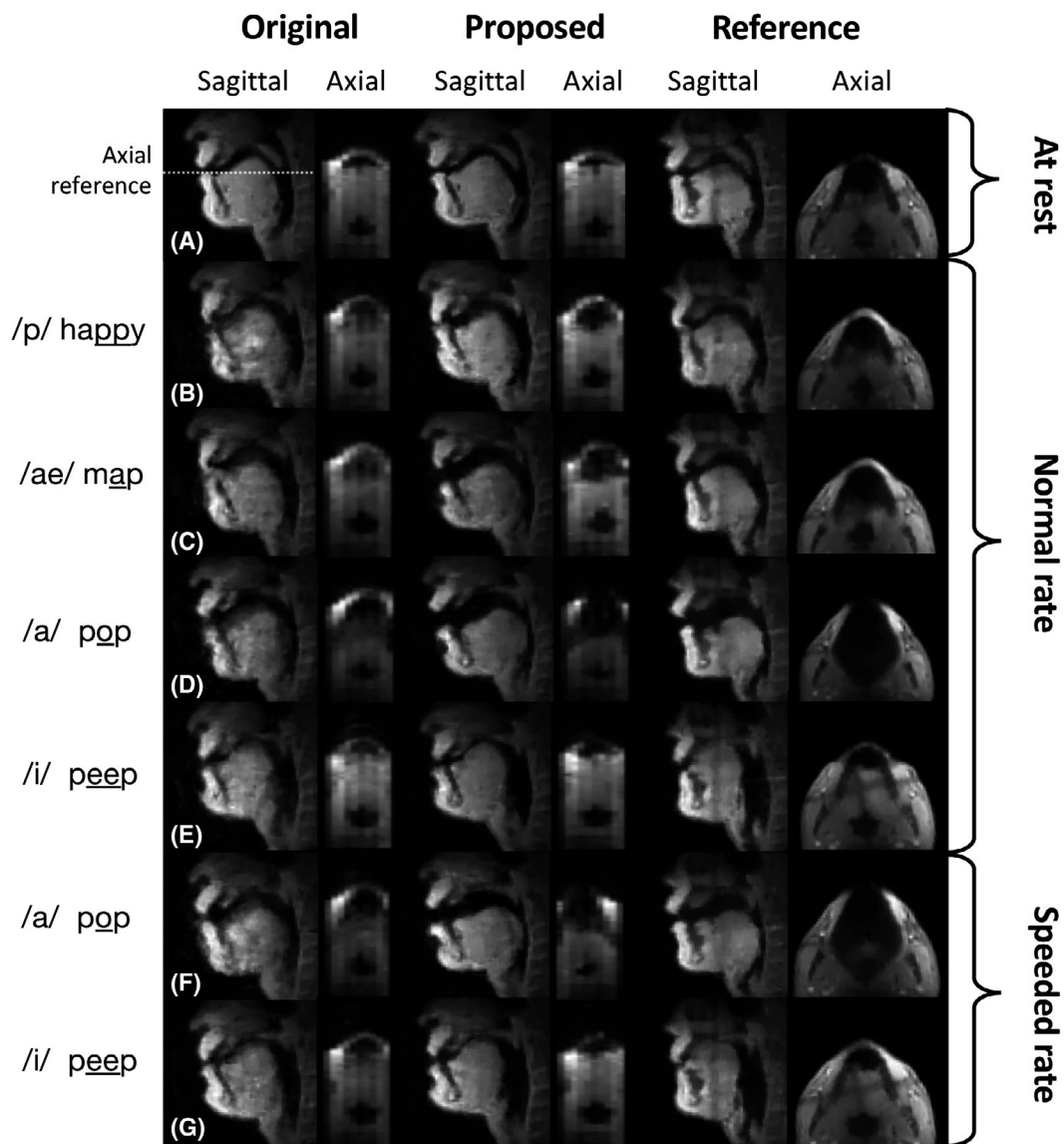
$\lambda_s = 0.008$  and  $\lambda_t = 0.03$ , and applied this to all sampling schemes for the simplicity of pairwise comparisons in the remainder of this work.

### 3.2 | Experimental optimization

Figure 3 presents pairwise comparisons of the candidate sampling methods during a representative utterance. We extracted four different intensity versus time plots from different views that exhibit the lips and tongue movements. In cases I and II we observed sharper articulator boundaries in the midsagittal plane, and clearer lip fluctuations and tongue root movements (pink arrows) for case II than for case I. Similar improvements were observed

when comparing cases III and IV (not shown for brevity). In cases II and IV, we observed improved delineation of tongue body motion in case IV in both midsagittal and coronal views (blue arrows). Similar improvements were observed when comparing cases I and III (not shown for brevity). In cases IV and V, we observed substantial improvements in boundary sharpness and background noise suppression for case V. The region marked with red arrows shows that case V captures the fast movements of lips and tongue that are also seen on interleaved 2D RT-MRI (not shown). We concluded that case V provided the best image quality among the candidate sampling strategies.

Figure 4 illustrates a retrospective selection of temporal resolution for case V, using the results of stimuli set 1 (Table 1) with a speeded speaking rate. The use of finer



**FIGURE 5** Demonstration of speech application of the proposed method in stimuli set 1. Static voiceless images (A) are captured at rest. Six sounds (B-G) are displayed for both normal and speeded speech rates. The results are compared between the original method (case I, first column), the proposed method (case V, second column), and the reference multislice 2D method (third column)



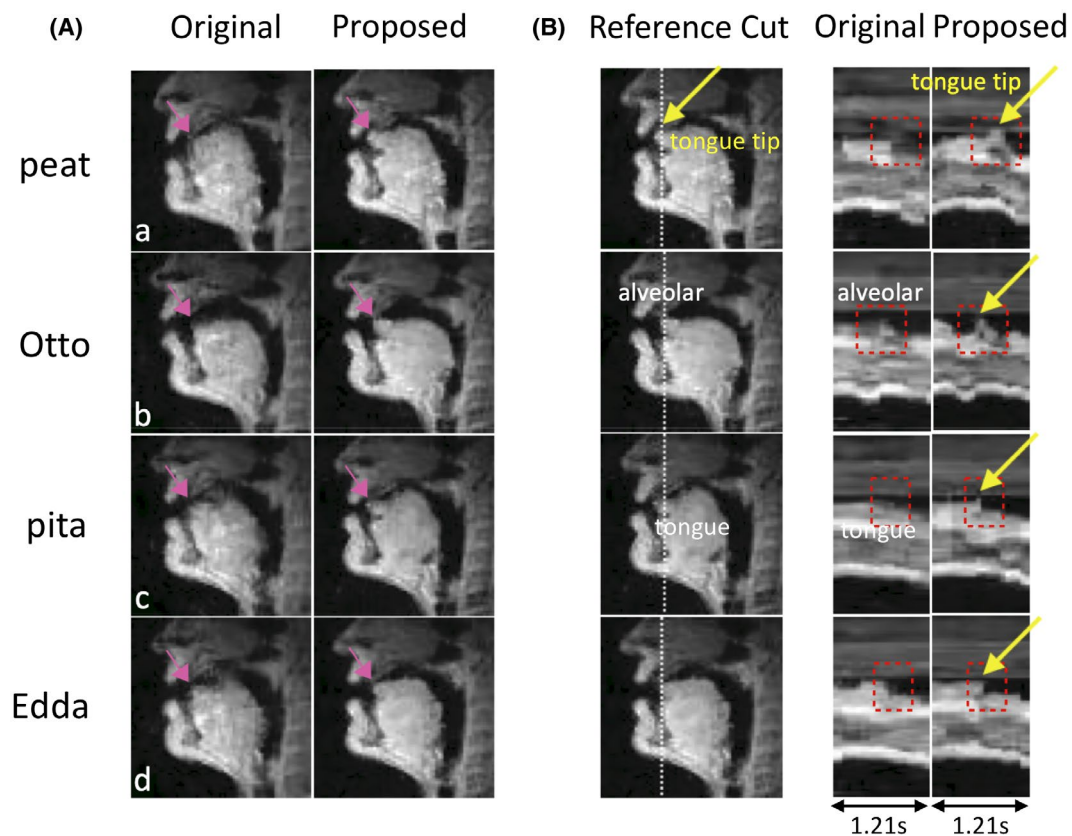
temporal resolution enables capturing rapid opening and closing of upper and lower lips (green arrows in Figure 4C), whereas a relatively better image quality can be preserved with adequate temporal resolution shown in Figure 4A. The proposed method provides a flexible choice of a wide range of temporal resolutions, while a trade-off exists between image quality and temporal resolution.

### 3.3 | Application to speech production

Figure 5 compares the results of the original and proposed data-sampling methods, with the reference interleaved multislice method. Six snapshots are selected from stimuli 1 (Table 1) at normal and speeded speech rates, along with static images in resting state. Both midsagittal and axial views are shown. The high (/i/) and low (/a/, /æ/) vowels shown in Figure 5 are monosyllabic and bounded by labial consonants. First, static images at resting state share the same information among original, proposed, and reference methods. Second, for the /p/ sound (Figure 5B), selected at the earliest closure of the lower lip, we observed blurring of the lower lip and jaw with the original method. The proposed method does not have this issue and is consistent with the interleaved 2D

reference. Third, for low vowels (Figure 5C,D), we observed blurring at the anterior tongue and tongue root with the original method. The proposed method shows clearer tongue surface boundaries, which correspond with the 2D reference. Fourth, for high vowels (Figure 5E), we observed blurring at the superior surface of the tongue with the original method. The blurring near the same location is improved in the proposed method. The reference method shares the same information as the proposed method. Fifth, for the speeded speech rate (Figure 5F,G), the proposed method demonstrates improved sharpness of articulator boundaries compared with the original approach. See also Supporting Information Video S2, showing dynamic tongue movement cycles.

Figure 6 qualitatively characterizes the improvements of the proposed method compared with the original using the stimuli set 2 (Table 2). Four examples are shown: alveolar consonant [t] (Figure 6A) and word-internal consonant [r] (Figure 6B-D). Midsagittal views (Figure 6A) at the time of constriction and the intensity-versus-time plots (Figure 6B) during a temporal interval of 1.21 seconds are shown. Reference cut in Figure 6B demonstrates different directions of tongue movements. First, the proposed method depicts a single contact of the tongue tip to the alveolar ridge in midsagittal views (marked with pink arrows in Figure 6A). This

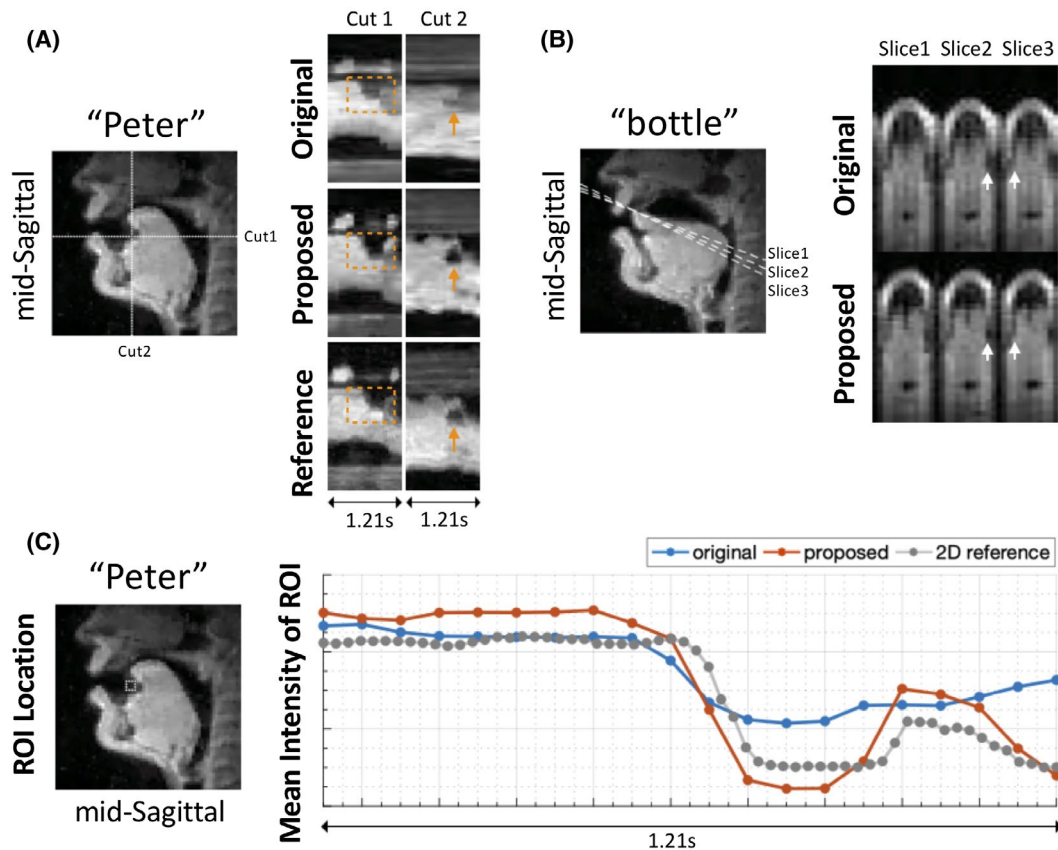


**FIGURE 6** Demonstration of speech application of the proposed method (case V) in stimuli set 2. (A) Midsagittal views of four selected alveolar consonants at the time of constriction, for the original (case I, first column) and the proposed (case V, second column) methods. (B) Intensity-versus-time plots with the reference cut marked as dashed lines in the midsagittal views

closure moment is also captured by the intensity-versus-time plots (marked with yellow arrows in Figure 6B). This critical event could not be detected by the original method in space nor in time. Second, the midsagittal results in Figure 6A are blurry around the tongue tip and the blade during constriction when using the original method. Third, the creation of the constriction is clearly captured by the proposed method as the tongue tip quickly rises to form and then lowers to release its closure contact (period within red dashed boxes in Figure 6B) during “peat” (a) and “Otto” (b). The sharp boundaries reflecting the retraction motion (period within red dashed boxes in Figure 6B) during the medial consonants of “pita” (c) and “Edda” (d) were observed in the proposed method. These were not captured by the original method. See also Supporting Information Video S3.

Figure 7 demonstrates the tongue-shaping geometries for rhotic and lateral flaps in stimuli set 2. Three displays are selected to assess imaging improvements. Intensity-versus-time plots are shown to demonstrate the temporal fidelities in capturing the sublingual airway cavity in the rhotic flap and release in “Peter” (Figure 7A). The tilted axial views are shown to demonstrate the creation of the tongue’s side channels while articulating the laterally released tap in “bottle”

(Figure 7B). The mean value-versus-time plots are displayed for the original, proposed, and reference methods (Figure 7C). In Figure 7A, we observed in both cut 1 and cut 2 the creation of a sublingual cavity, as the tongue tip quickly rises and retracts (period within the orange boxes). However, similar regions are blurred in the original method. From Figure 7B, clear dark channels at the sides of the tongue are detectable along the axis from tongue tip to the retracting back of the tongue. Channel formation for lateral lingual airflow (along one or both sides of the tongue) is a critical linguistic feature of this consonant production. Figure 7C presents a region-of-interest analysis of the formation for the sublingual cavity also shown in Figure 7A. A pixel intensity drop in this vocal tract region is observed for the proposed and the reference methods during the time of the sublingual cavity formation, but the original method shows a shallow curve with poorly defined onset and end, as compared with the sharper and greater magnitude reflection of the cavity formation in the proposed method. Overall, the results in the proposed method match closely with the reference. In summary, the results illustrate the capability of the proposed method for capturing more complex geometries associated with rapid raising and retraction of the tongue tip and narrowing of the



**FIGURE 7** Illustration of sublingual cavity during a rhotic flap and the tongue side channeling during a laterally release tap. (A) Intensity-versus-time displays. Horizontal (cut 1) and vertical (cut 2) lines are selected from midsagittal plane (dashed lines). (B) Tilted axial views, indicating the stretched tongue sides. (C) Mean of region of interest (ROI) versus time plots among original (case I), proposed (case V), and 2D reference methods

tongue profile created by anterior–posterior stretching for the tongue.

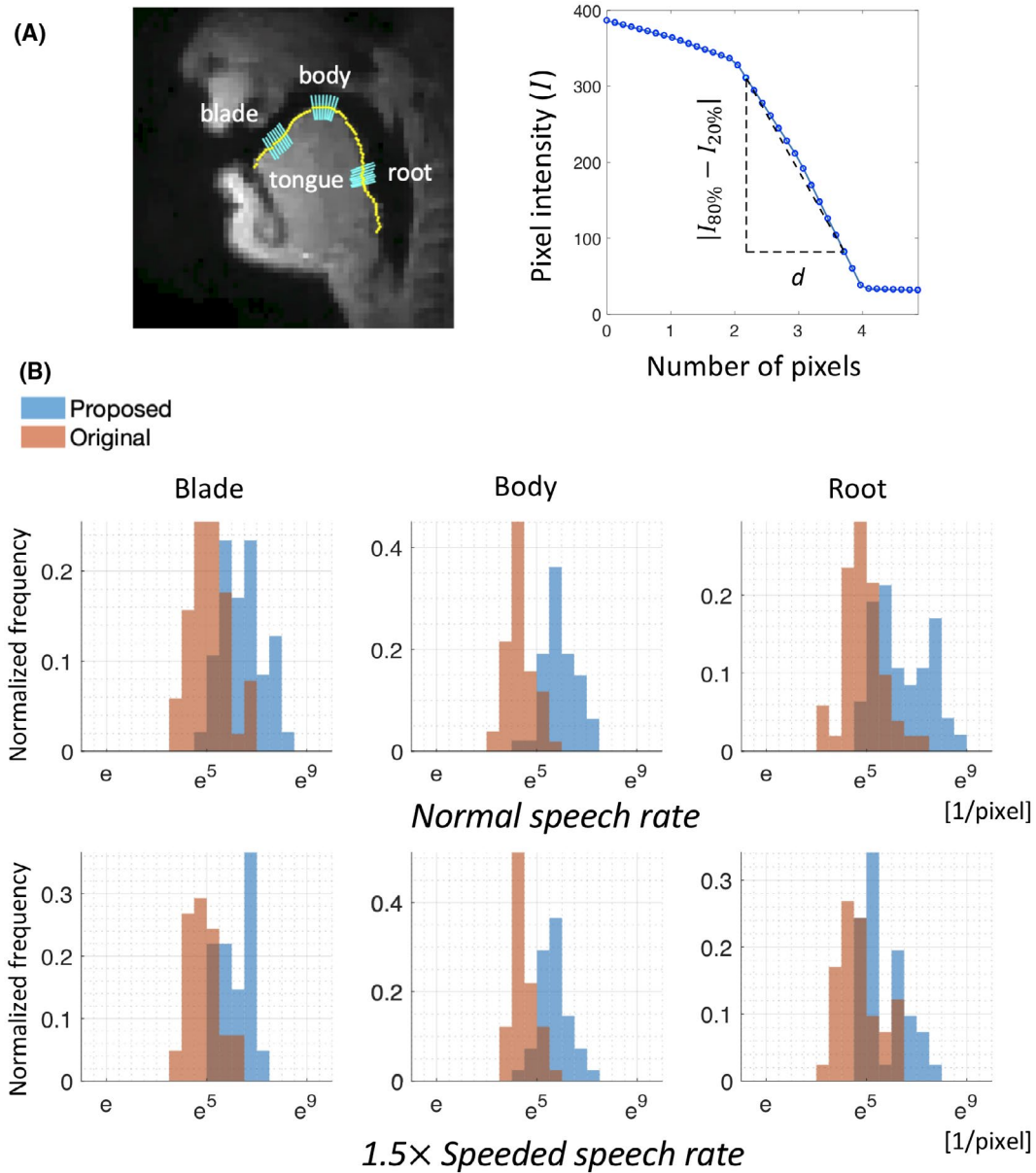
### 3.4 | Quantitative evaluation

Figure 8B illustrates the quantitative evaluation of tongue boundary sharpness in stimuli set 1 (Table 1). The sharpness scores are measured at the blade, body, and root of the tongue. Histograms of sharpness score for the original (red) and the proposed (blue) methods are plotted on a log scale for the three locations at normal and 1.5-times speeded speech rates.

For all three locations and both speech rates, we found a statistically significant difference in the mean tongue boundary sharpness, with  $P < .001$ . The  $P$ -values for the tongue blade, body, and root are .0001, <.0001, and <.0001 for the normal speech rate, and <.0001, <.0001, and .0003 for the 1.5-times speeded speech rate, respectively.

## 4 | DISCUSSION

We demonstrate improved 3D RT-MRI of speech production using stack-of-spiral sampling with advanced data-sampling



**FIGURE 8** Boundary sharpness score and the evaluation of the original and the proposed methods. (A) Extracted tongue boundary (yellow) and constructed gridlines (cyan) in the blade, body, and root of the tongue in a midsagittal plane displayed on the left. The boundary sharpness score is calculated based on the intensity profile (right) along each gridline. (B) Normalized histograms of tongue boundary sharpness score for the original and the proposed methods, for three locations of interest and two speech rates (normal and 1.5 times speeded).  $e$ : Euler's number (approximately equal to 2.718)

strategies and spatio-temporally constrained reconstruction. We perform a step-by-step comparison to figure out the best sampling strategy. We demonstrate the application of the improved 3D RT-MRI to capture alternating high-low vowels in both normal and speeded speech rates in healthy adults. This indicates that the proposed method provides adequate temporal resolution to capture natural movement of articulators, such as lips and tongue, in speech production studies, needing no speech slowing, prolonged or sustained speech, or repetitions. Furthermore, we show the capability of the proposed method to visualize fast alveolar consonant segments during natural speech, along with capturing complex 3D tongue shape geometries. This approach provides a better depiction of rapidly moving articulators and roughly 2-fold finer temporal resolution as compared with the prior state-of-the-art.<sup>19</sup> Furthermore, 3D speech RT-MRI now provides approximately 10-fold greater spatial coverage compared with mature 2D speech RT-MRI technology, with desired capabilities of capturing rapid temporal articulatory events. We believe this will enable significant new insights in speech science.

The substantial improvements in temporal resolution and spatial coverage stand to allow more complete characterization of complex vocal tract geometries, such as those found in lingual grooving, channeling, and concavity found in phonetic sibilant, lateral, retroflex, and other consonants (eg, /s, l, ʃ, t/) as well as the tongue side bracing found for most constricted speech sounds including mid and high vowels and the labial protrusion critical to certain vowels (eg, /u, y/) and consonants (eg, /ɹ, w, f/). Finally, a thorough and accurate spatiotemporal characterization of 3D vocal tract shaping is critical for models connecting articulation to the acoustic resonance and noise structure that it produces and that characterizes the transmitted speech signal. Beyond typical speech production, a complete characterization of 3D vocal tract dynamics may prove important for describing and remediating speech following surgery, such as in cancer patients having glossectomies.

Beyond its utility in dynamic speech imaging, the proposed data-sampling and reconstruction method may benefit other applications, particularly those in which dynamic edge information is critical, such as (1) dynamic MRI of cardiac function,<sup>39-41</sup> in which the endocardial and epicardial contours are of greatest importance; (2) dynamic MRI of the upper and lower airway,<sup>42,43</sup> where the pharyngeal airway and trachea are most critical; and (3) dynamic MRI of joint motion,<sup>44,45</sup> where the movement of bones and cartilage are most critical. The proposed sampling strategy may also benefit a 3D stack-of-radial (a.k.a. stack-of-stars) approach due to its similar distribution in 3D  $k$ -space.<sup>27,46</sup> This data-sampling scheme has the benefit of mitigating motion-induced temporal blurring, especially with respect to radial trajectory that has been proved to be robust to motion artifacts.<sup>47</sup>

Data sampling and constrained reconstruction are intertwined. It is likely that there are tailored constrained reconstruction approaches that would work better for each sampling strategy. Possibilities include regularized nonlinear inversion<sup>48</sup> and partial separability model-based reconstruction.<sup>17</sup> Reconstruction time may be reduced by using parallelized GPU computation, as individual time segments can be processed independently. The GPU-based reconstruction has been investigated by multiple groups and has provided 3-200-fold reduction in computation time for MRI-constrained reconstruction.<sup>49,50</sup> This would be a valuable direction for future work, especially for interactive RT-MRI applications that require low-latency reconstruction.

## 5 | CONCLUSIONS

We demonstrate improved 3D RT-MRI of human speech production using innovative ( $k$ ,  $t$ ) data sampling. We propose a stack-of-spiral sampling trajectory with variable density randomized temporal order along  $k_z$  and golden-angle increment in the  $k_x$ - $k_y$  plane sampling trajectory, combined with spatio-temporally constrained reconstruction. This scheme shows superior resistance to motion artifacts and outperforms other proposed sampling schemes. The improvements in spatiotemporal resolution and in ameliorating blurring allow better visualization of fast lip and tongue movements of both large and small articulatory magnitude, and are successful at imaging even very rapid consonant segments and complex 3D tongue shaping geometries such as sublingual cavities and lateral channels. Both dynamic speech production imaging as well as other imaging applications in which dynamic edge information and/or mitigating off-resonance are critical stand to benefit from this advance.

## ACKNOWLEDGMENT

We thank our scan volunteers and acknowledge the support and collaboration of the Speech Production and Articulation kNoWledge (SPAN) group at the University of Southern California, Los Angeles, California.

## DATA AVAILABILITY STATEMENT


The code and sample datasets that support the findings of this study are openly available in GitHub at [https://github.com/usc-mrel/Improved\\_3DRT\\_Speech](https://github.com/usc-mrel/Improved_3DRT_Speech).

## ORCID

Ziwei Zhao  <https://orcid.org/0000-0003-0281-1141>

Yongwan Lim  <https://orcid.org/0000-0003-0070-0034>

Dani Byrd  <https://orcid.org/0000-0003-3319-5871>

Shrikanth Narayanan  <https://orcid.org/0000-0002-1052-6204>

[org/0000-0002-1052-6204](https://orcid.org/0000-0002-1052-6204)

Krishna S. Nayak  <https://orcid.org/0000-0001-5735-3550>

[org/0000-0001-5735-3550](https://orcid.org/0000-0001-5735-3550)



## REFERENCES

1. Scott AD, Wylezinska M, Birch MJ, Miquel ME. Speech MRI: morphology and function. *Phys Medica*. 2014;30:604-618.
2. Bresch E, Kim Y, Nayak K, Byrd D. Seeing speech: capturing vocal tract shaping and analysis. *IEEE Signal Process Mag*. 2008;25:123-132.
3. Lingala SG, Sutton BP, Miquel ME, Nayak KS. Recommendations for real-time speech MRI. *J Magn Reson Imaging*. 2016;43:28-44.
4. Lingala SG, Zhu Y, Kim YC, Toutios A, Narayanan S, Nayak KS. A fast and flexible MRI system for the study of dynamic vocal tract shaping. *Magn Reson Med*. 2017;77:112-125.
5. 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.
6. 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.
7. Kim YC, Narayanan SS, Nayak KS. Accelerated 3D MRI of vocal tract shaping using compressed sensing and parallel imaging. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2009, Taipei, Taiwan. pp 389-392.
8. 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.
9. Narayanan S, Nayak K, Byrd D, Lee S. An approach to real-time magnetic resonance imaging for speech production. *J Acoust Soc Am*. 2003;113:2258.
10. Kim YC, Narayanan SS, Nayak KS. Accelerated three-dimensional upper airway MRI using compressed sensing. *Magn Reson Med*. 2009;61:1434-1440.
11. Proctor MI, Bone D, Katsamanis N, Narayanan S. Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. In: Proceedings of the 11th Annual Conference on the International Speech Communication Association (INTERSPEECH), Makuhari, Chiba, Japan, 2010. pp 1576-1579.
12. Silva S, Teixeira A. Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Comput Speech Lang*. 2015;33:25-46.
13. Ramanarayanan V, Tilsen S, Proctor M, et al. Analysis of speech production real-time MRI. *Comput Speech Lang*. 2018;52:1-22.
14. Kim YC, Proctor MI, Narayanan SS, Nayak KS. Improved imaging of lingual articulation using real-time multislice MRI. *J Magn Reson Imaging*. 2012;35:943-948.
15. Fu M, Barlaz MS, Shosted RK, Liang ZP, Sutton BP. High-resolution dynamic speech imaging with deformation estimation. In: Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 2015. pp 1568-1571.
16. Narayanan S, Byrd D, Kaun A. Geometry, kinematics, and acoustics of Tamil liquid consonants. *J Acoust Soc Am*. 1999;106:1993-2007.
17. 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.
18. 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 Imaging*. 2017;46:94-101
19. 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.
20. Pipe JG, Ahunbay E, Menon P. Effects of interleaf order for spiral MRI of dynamic processes. *Magn Reson Med*. 1999;41:417-422.
21. Deng W, Zahneisen B, Stenger VA. Rotated stack-of-spirals partial acquisition for rapid volumetric parallel MRI. *Magn Reson Med*. 2016;76:127-135.
22. Scheffler K, Hennig J. Frequency resolved single-shot MR imaging using stochastic k-space trajectories. *Magn Reson Med*. 1996;35:569-576.
23. Nayak KS, Nishimura DG. Randomized trajectories for reduced aliasing artifact. In: Proceedings of the ISMRM 6th Scientific Meeting & Exhibition, Sydney, Australia, 1998. p 670.
24. Tsai CM, Nishimura DG. Reduced aliasing artifacts using variable-density k-space sampling trajectories. *Magn Reson Med*. 2000;43:452-458.
25. Liao JR, Pauly JM, Brosnan TJ, Pelc NJ. Reduction of motion artifacts in cine MRI using variable-density spiral trajectories. *Magn Reson Med*. 1997;37:569-575.
26. Lee JH, Hargreaves BA, Hu BS, Nishimura DG. Fast 3D imaging using variable-density spiral trajectories with applications to limb perfusion. *Magn Reson Med*. 2003;50:1276-1285.
27. Zhou Z, Han F, Yan L, Wang DJJ, Hu P. Golden-ratio rotated stack-of-stars acquisition for improved volumetric MRI. *Magn Reson Med*. 2017;78:2290-2298.
28. Santos JM, Wright GA, Pauly JM. Flexible real-time magnetic resonance imaging framework. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Francisco, California, 2004. pp 1048-1051.
29. 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.
30. Kerr AB, Pauly JM, Hu BS, et al. Real-time interactive MRI on a conventional scanner. *Magn Reson Med*. 1997;38:355-367.
31. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Mach Learn*. 2010;3:1-122.
32. Uecker M, Ong F, Tamir JJ, et al. The BART toolbox for computational magnetic resonance imaging. In Proceedings of the ISMRM 23th Scientific Meeting & Exhibition, Toronto, Canada, 2015. p. 2486.
33. 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.
34. Miller AL, Finch KB. Corrected high-frame rate anchored ultrasound with software alignment. *J Speech, Lang Hear Res*. 2011;54:471-486.
35. Lammert A, Ramanarayanan V, Proctor M, Narayanan S. Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis. In: Proceedings of the 14th Annual Conference on the International Speech Communication Association (INTERSPEECH), Lyon, France, 2013. pp 959-962.
36. Lim Y, Lingala SG, Narayanan SS, Nayak KS. Dynamic off-resonance correction for spiral real-time MRI of speech. *Magn Reson Med*. 2019;81:234-246.
37. Kim J, Kumar N, Lee S, Narayanan S. 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.
38. Wilcox R. *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*. Boca Raton, FL: CRC Press; 2011.
  39. Mascarenhas NB, Muthupillai R, Cheong B, Pereyra M, Flamm SD. Fast 3D cine steady-state free precession imaging with sensitivity encoding for assessment of left ventricular function in a single breath-hold. *Am J Roentgenol*. 2006;187:1235-1239.
  40. Barkauskas KJ, Rajiah P, Ashwath R, et al. Quantification of left ventricular functional parameter values using 3D spiral bSSFP and through-time non-Cartesian GRAPPA. *J Cardiovasc Magn Reson*. 2014;16:1-13.
  41. Montalt-Tordera J, Kowalik G, Gotschy A, Steeden J, Muthurangu V. Rapid 3D whole-heart cine imaging using golden ratio stack of spirals. *Magn Reson Imaging*. 2020;72:1-7.
  42. Persak SC, Sin S, McDonough JM, Arens R, Wootton DM. Noninvasive estimation of pharyngeal airway resistance and compliance in children based on volume-gated dynamic MRI and computational fluid dynamics. *J Appl Physiol*. 2011;111:1819-1827.
  43. Kim YC, Marc Lebel R, Wu Z, Davidson Ward SL, Khoo MCK, Nayak KS. Real-time 3D magnetic resonance imaging of the pharyngeal airway in sleep apnea. *Magn Reson Med*. 2014;71:1501-1510.
  44. Henrichon SS, Foster BH, Shaw C, et al. Dynamic MRI of the wrist in less than 20 seconds: normal midcarpal motion and reader reliability. *Skeletal Radiol*. 2020;49:241-248.
  45. Shaw CB, Foster BH, Borgese M, et al. Real-time three-dimensional MRI for the assessment of dynamic carpal instability. *PLoS One*. 2019;14:1-18.
  46. Mendes JK, Adluru G, Likhite D, et al. Quantitative 3D myocardial perfusion with an efficient arterial input function. *Magn Reson Med*. 2020;83:1949-1963.
  47. Block KT, Chandarana H, Milla S, et al. Towards routine clinical use of radial stack-of-stars 3D gradient-echo sequences for reducing motion sensitivity. *J Korean Soc Magn Reson Med*. 2014;18:87.
  48. Uecker M, Zhang S, Voit D, Karaus A, Merboldt KD, Frahm J. Real-time MRI at a resolution of 20 ms. *NMR Biomed*. 2010;23:986-994.
  49. Freiberger M, Knoll F, Bredies K, Scharfetter H, Stollberger R. The agile library for biomedical image reconstruction using GPU acceleration. *Comput Sci Eng*. 2013;15:34-44.
  50. Wu XL, Gai J, Lam F, et al. Impatient MRI: Illinois Massively Parallel Acceleration Toolkit for image reconstruction with enhanced throughput in MRI. In: *Proceedings of the International Symposium on Biomedical Imaging*, Chicago, IL, 2011. pp 69-72.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**VIDEO S1** Movie display of parameter sweeps in finer scale with the production of “/loo/ - /lee/ - /na/ - /za/ - /la/ - /za/.” From top to bottom, spatial total variation (TV) penalty  $\lambda_s$  is chosen to be 0.001, 0.002, 0.004, 0.006, 0.008, and 0.010. From left to right, temporal regularization penalty  $\lambda_t$  is 0.005, 0.01, 0.02, 0.03, and 0.04. A midsagittal view (left) and intensity-time plot (right) are shown for each parameter setting. The results illustrate the same effects as Figure 2. A reasonable parameter set can be chosen in the valley, where  $\lambda_s$  ranges from 0.002 to 0.008 and  $\lambda_t$  ranges from 0.01 to 0.03

**VIDEO S2** Tongue movement cycles during high-low-high or low-high-low tongue postures of stimuli set 1, at a natural speech rate. Axial and coronal views are shown with reference to the image on the left. A,B, High-low-high cycles start and end at a constricted posture of the tongue body with the hard palate, as is required for a high vowel; “hip hop beep” (/i/-/a/-/i/) (A) and “pea poppy” (/i/-/a/-/i/) (B). C,D, Low-high-low cycles start and end at a constricted posture of the tongue root with the rear pharyngeal wall, as is required for a low vowel: “pop him a ham” (/a/-/i/-/æ/) (C) and “happy map” (/æ/-/i/-/æ/) (D). Pink arrows highlight major imaging improvements at the top domed portion of the tongue body, at the retracted constriction of the tongue root, and at the lips in a relatively open posture, for the proposed method compared with the original method. The proposed method shares the same information as the reference method

**VIDEO S3** Tongue tip (“alveolar”) tap consonant segments from stimuli set 2 (total clip duration = 1.21 seconds), showing the 3D original, 3D proposed, and 2D reference methods. The videos are approximately aligned temporally from the end of the vowel in the frame sentence prior word (“gave” /ei/) to the vowel in frame sentence following word (“poppy” /a/ or “happily” /æ/). The viewer is directed to the upward movement and contact of the tongue tip at the alveolar ridge by the pink arrows. The red arcs shown in “otter” illustrate the tongue retraction pathway

**How to cite this article:** Zhao Z, Lim Y, Byrd D, Narayanan S, Nayak KS. Improved 3D real-time MRI of speech production. *Magn Reson Med*. 2021;85:3182–3195. <https://doi.org/10.1002/mrm.28651>