Enhancing linguistic research through 2-mm isotropic 3D dynamic speech MRI optimized by sparse temporal sampling and low-rank reconstruction

Riwei Jin1,2 | Ryan K. Shosted3 | Fangxu Xing4 | Imani R. Gilbert5 | Jamie L. Perry5 | Jonghye Woo4 | Zhi-Pei Liang2,6 | Bradley P. Sutton1,2

1Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, Illinois USA
2Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, Urbana, Illinois USA
3Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana, Illinois USA
4Gordon Center for Medical Imaging, Department of Radiology, Massachusetts General Hospital/Harvard Medical School, Boston, Massachusetts USA
5Department of Communication Sciences and Disorders, East Carolina University, Greenville, North Carolina USA
6Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois USA

Correspondence
Bradley P. Sutton, Department of Bioengineering, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.
Email: bsutton@illinois.edu

Funding information
NIH Clinical Center, Grant/Award Number: R01DE027989

Purpose: To enable a more comprehensive view of articulations during speech through near-isotropic 3D dynamic MRI with high spatiotemporal resolution and large vocal-tract coverage.

Methods: Using partial separability model-based low-rank reconstruction coupled with a sparse acquisition of both spatial and temporal models, we are able to achieve near-isotropic resolution 3D imaging with a high frame rate. The total acquisition time of the speech acquisition is shortened by introducing a sparse temporal sampling that interleaves one temporal navigator with four randomized phase and slice-encoded imaging samples. Memory and computation time are improved through compressing coils based on the region of interest for low-rank constrained reconstruction with an edge-preserving spatial penalty.

Results: The proposed method has been evaluated through experiments on several speech samples, including a standard reading passage. A near-isotropic 1.875 × 1.875 × 2 mm³ spatial resolution, 64-mm through-plane coverage, and a 35.6-fps temporal resolution are achieved. Investigations and analysis on specific speech samples support novel insights into nonsymmetric tongue movement, velum raising, and coarticulation events with adequate visualization of rapid articulatory movements.

Conclusion: Three-dimensional dynamic images of the vocal tract structures during speech with high spatiotemporal resolution and axial coverage is capable of enhancing linguistic research, enabling visualization of soft tissue motions that are not possible with other modalities.

KEYWORDS
dynamic speech imaging, low-rank approximation, linguistics, 3D dynamic imaging

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. Magnetic Resonance in Medicine published by Wiley Periodicals LLC on behalf of International Society for Magnetic Resonance in Medicine.

Magn Reson Med. 2023;89:652–664.
1 | INTRODUCTION

Dynamic MRI has found increasing use in scientific research and clinical applications. Its high spatial and temporal resolution offers great potential for capturing both structural and functional articulatory dynamics during speech,\textsuperscript{1,2} examining speech pathologies, following cleft palate surgery,\textsuperscript{3,4} and other areas relevant to the study of synchronic and diachronic phonology.\textsuperscript{5,6} Despite the clear advantages of the nonionizing imaging of dynamic MRI to visualize the tongue and oropharyngeal structures during speech, the technique still poses significant challenges for adequate visualization of the complex vocal tract structures. In particular, human tongue and oropharyngeal structures are one of the most structurally and functionally complicated muscular structures. Due to the innervation by over 13 000 hypoglossal motoneurons, the tongue and oropharyngeal structures are able to generate rapid yet precise speech movements.

A review paper from Lingala et al outlines technical challenges that must be met for effective visualization of speech using dynamic MRI.\textsuperscript{7} Specifically, a minimum 40–150-ms temporal resolution and a 1–4-mm spatial resolution were recommended to avoid missing closure events (i.e., critical moments in speech production when two anatomical structures come in contact with one another). Many important articulatory configurations cannot be observed directly from a midsagittal perspective, such as grooving on the surface of the tongue or asymmetrical velar elevation during velopharyngeal closure. Volume measures for vocal tract channels and cavities, important for acoustic modeling, necessarily require measurements in three dimensions. Because the vocal tract bends approximately 90° at the junction of the posterior oral cavity and the upper pharynx, both coronal and axial planes are needed to accurately measure cross-sectional tract area. In addition, sequences must be robust to magnetic susceptibility effects that arise at the air/tissue interfaces throughout the vocal tract.

High spatial and temporal resolution with spatial coverage over the entire vocal tract along with good resolution in all three dimensions can significantly improve our ability to visualize and analyze complex tongue, velum, and other vocal tract movements. Recent technical processes have greatly expanded the possibilities on high spatiotemporal resolution 3D scanning. For example, Burdumy et al achieved a 1.6-mm isotropic resolution, 62-mm axial coverage, and 1.3 s per image.\textsuperscript{8} Lim et al achieved a 70-mm left–right (through plane for sagittal scan) coverage with 5.8-mm slice thickness, 2.4-mm resolution, and 61-ms time resolution.\textsuperscript{9} Fu et al achieved eight slices with 40-mm total thickness and a temporal resolution of 166 frames per second.\textsuperscript{10}

In our previous works, we sampled temporal navigators with every line of imaging data to achieve a high temporal resolution up to 166 frames per second.\textsuperscript{10–12} However, to achieve the same level of 3D resolution and left–right coverage, these methods may require overly long acquisition time which, due to subject fatigue, can greatly affect the quality of a data set. To address this issue, in this work, we expand the previous methods\textsuperscript{10–12} to achieve high (near) isotropic resolution ($1.875 \times 1.875 \times 2 \, \text{mm}^3$) with significant coverage of thirty-two 3D slice locations across 64 mm of left–right coverage for the sagittal scan and an overall 3D volume acquisition rate of 35.6 Hz (28.1 ms per 3D volume) with a sparse temporal sampling pattern and specialized image-reconstruction approach.

Specifically, we modified our earlier temporal navigation approach to keep acquisition time at a reasonable level.\textsuperscript{10} In addition, by applying a sparser temporal navigation strategy, scanning multiple k-space lines per temporal navigator reduced the overall scan time for data acquisition at the expense of a reduced, but still sufficient, frame rate. Furthermore, to match the sparse temporal sampling pattern, a low-rank reconstruction was performed using the partial separability (PS) model similar to our previous works.\textsuperscript{10–12} Taken together, by applying this combined strategy and analyzing concurrently recorded and synchronized audio waveforms, we can observe complex movements during speech, such as the tongue twisting, tongue grooving, and coarticulation between speech sounds, demonstrating that we can achieve a better visualization of complex 3D motions during speech, compared with the previous works.

2 | METHODS

Exploiting the partial separability of dynamic imaging signals, we sparsely sampled two sets of (k, t)–spaced data in an interlaced manner: an imaging data set (from which spatial basis functions of the PS model were estimated) and a set of temporal navigators in high temporal resolution (which were used for estimation of the temporal basis functions), as shown in Figure 1.\textsuperscript{1,12} The imaging data set is acquired using a Cartesian trajectory with random phase encoding in both the phase and slice directions to provide structural quality images. The navigator data set was acquired using a 3D spiral (cone) trajectory and was only acquired once every four imaging data acquisitions.

The proposed method is different from our previous approach of acquiring temporal navigators with the acquisition of every line of k-space imaging data. Given that our previous frame rates were up to 166 fps with eight 5.0-mm slices,\textsuperscript{10} we instead chose to shorten the total acquisition time, at the expense of a lower frame rate, and acquired...
four k-space lines for every temporal navigator. This can be thought of as reducing the undersampling factor at each imaging time point and reducing the temporal resolution of the model. We group the imaging lines to the closest-in-time sampled navigator signal. This enabled us to expand the 3D coverage while keeping the scan time reasonable. To scan 32 slices across all speech samples, our previous method (2017) needed approximately 30 min, while the new method needs 20 min, a 33% reduction in time.

Figure 1 shows the pulse sequence diagram of our method. We used an RF spoiled gradient-echo acquisition as our base sequence with a TR of 5.5 ms and TE of 1.85 ms. Our imaging acquisition used randomized phase encodings applied in the ky and kz directions with an overall matrix size of 128 × 128 × 32 for a FOV of 24 × 24 × 6.4 cm³. For acquisition of the temporal navigators, we used a 3D spiral-in/spiral-out cone navigator with 6.1-ms TR. The navigator trajectory was designed as a single shot of a 28-shot encoding for a 128-matrix size spiral, with a maximum gradient amplitude of 34 mT/m and slew rate of 120 mT/m/ms, following the analytical k-space design in Glover. The 2D navigators are extended to 3D navigators by adding a linear gradient in the z-direction to have a similar k-space extent in all directions. This design enabled spiral-in/spiral-out acquisitions (approximately matching the TE) to fit within nearly the same TR as the imaging data set.

In our experiments, we scanned 32 slices with 128 phase encodings for each; as a result, we had 4096 lines of 3D k-space to sample. We sampled four lines with each temporal navigator, which resulted in 512 acquisition sets (an acquisition set is a set of one temporal navigator and four associated imaging lines) for each fully sampled frame for model fitting, or about 30 s of scan time. For each fully sampled frame of data, we randomly distributed all phase and slice encodes to avoid repeated acquisition of a line during one fully sampled frame. Note that the PS model enables temporal reconstruction of a full 3D volume at each acquisition set of the acquisition, resulting in 40,960 3D volumes in 80 fully sampled frames, with a temporal spacing of 28.1 ms for our 19.5-min scan.

### 2.1 Partial separability model

We used the PS model to effectively capture spatiotemporal correlations that exist in the dynamic speech images. This model represents dynamic speech signals as a combination of temporal and spatial basis functions. Our previous work had shown that the spatiotemporal correlation assumptions hold well for speech imaging across natural and queued speech samples. Using the PS model, we expressed our desired image time series \( f(\mathbf{r}, t) \) as

\[
f(\mathbf{r}, t) = \sum_{l=1}^{L} c_l(\mathbf{r}) \phi_l(t),
\]

where \( \{c_l(\mathbf{r})\}_{l=1}^{L} \) denotes the spatial basis, and \( \{\phi_l(t)\}_{l=1}^{L} \) represents temporal basis, with \( L \) being the rank of the model, which can be very low (from 30 to 50 in our method).

### 2.2 Coil compression

To reduce the size of the acquired data set to enable efficient computation in image reconstruction, we implemented a coil-compression process that combined data from different receiver coils in a way to capture the
variance across the coil sensitivities in a prespecified region of interest.\textsuperscript{15} We performed singular value decomposition of complex coil-sensitivity maps in the region of interest covering the vocal tract. The descent order of singular values of coils indicates energy-level contribution in the region of interest. For our experiments, the first 10 combined virtual coils out of 20 original coils were used, which contained 95% energy (magnitude) of the oral area. The coil compression scheme enabled a 50% reduction in both computation time and memory requirement.

2.3 \hspace{1em} Reconstruction

To perform spatially regularized PS model–based reconstruction of dynamic speech image, we estimated the temporal basis functions from the navigator data using a singular value decomposition–based method.\textsuperscript{11,12} We then determined the spatial basis by solving a constrained least square problem with a Huber penalty using a conjugate gradient method, as follows:\textsuperscript{16}:

\begin{equation}
\{ \hat{c}(r) \}_{l=1}^{L} = \arg \min_{\{ c(r) \}_{l=1}^{L}} \left\{ \left\| s(k, t) - E \left[ \sum_{l=1}^{L} c(r) \phi_l(t) \right] \right\|_2^2 + \lambda \Phi_{\text{huber}} [D(\hat{c}(r))] \right\}.
\end{equation}

Where \( \{ \hat{c}(r) \}_{l=1}^{L} \) is the estimated spatial basis; \( s(k, t) \) are the measured data; \( E \) is the encoding operator including the effect of Fourier transform \( W \), spatial sensitivity weighting \( S \), and data sampling \( \Omega \); and \( \lambda \) is the regularization parameter. \( D(\cdot) \) is the operator that takes first-order spatial derivative in each dimension. The Huber penalty is an edge-preserving regularization method that penalizes small differences between adjacent pixels quadratically, while large edge differences are penalized by \( L^1 \)-norm\textsuperscript{16} as follows:

\[ \Phi_{\text{huber}}(u) = \begin{cases} 
2 \frac{|u|^2}{M} & |u| \leq M \\
M(2|u| - M) & |u| > M
\end{cases} \]

After the temporal and spatial bases were determined, we reconstructed the time series of images through Equation (1).

To measure the spatial sensitivity, we calculated the coil images by performing the inverse Fourier transform on the time-averaged imaging data across the whole time series. The spatial sensitivity weighting, \( S \), is measured by dividing the individual coil images by the square root of the sum of square images across the coils.

2.4 \hspace{1em} Audio waveform denoising

Audio waveforms were recorded through an MRI-compatible fiber optic microphone (OptoAcoustics, Or Yehuda, Israel) with a temporal sampling rate of 16 000 Hz. To reduce the scanner noise from the recording, we performed waveform denoising using a dictionary learning method with spectral and temporal regularization, as developed by Colin et al.\textsuperscript{17} Our MRI sequence outputted a TTL signal at each TR that was recorded with the audio waveform by the microphone system. This enabled us to use the TTL sequence waveform to accurately align the recorded audio with the MRI images.

2.5 \hspace{1em} Imaging experiments

Imaging experiments were performed on a 3T Siemens Prisma MRI system using a Siemens 20-channel head coil. The acquisition matrix size was 128 \( \times \) 128 \( \times \) 32 with a 240 \( \times \) 240 mm sagittal FOV, 64-mm full thickness left–right across the vocal tract from 32 slices of 2-mm thickness. The spatial resolution was 1.875 \( \times \) 1.875 \( \times \) 2 mm\(^3\), near 2-mm isotropic. We achieved a temporal resolution of 35.6 frames per second in the 3D dynamic imaging.

Four native speakers of American English ranging in age from 22 to 34 were asked to read the “Rainbow Passage”\textsuperscript{18} repeatedly for the 19.5-min scan duration (1 male, 3 female). Data were collected in accordance with a human subjects’ protocol approved through the University of Illinois Urbana-Champaign Internal Review Board (protocol #19106). The rainbow passage was displayed on an LCD screen in the MRI room and was auto-advanced at a preset rate, but subjects were free to speak at a comfortable rate including pauses when necessary.

We acquired 20 full-frame measurements for model fitting, for which the total duration of the scan was 19.5 min. This resulted in 40 960 reconstructed 3D images from the PS model at 35.6 frames per second. The total acquisition time was acceptable to the subjects, but future work will enable a shorter overall acquisition time. We used an engineering station with Intel(R) Xeon(R) CPU E5-2620 at 2.00 GHz and a total memory of 214 GB for data processing. With coil compression, the memory requirement for reconstruction is about 60 GB using MATLAB, and the reconstruction took about 50 h.

3 \hspace{1em} RESULTS

With our new navigation scheme, we have reduced the total acquisition time, but we have also reduced the frame rate to 35.6 fps. To determine whether this level of
temporal resolution was still sufficient to capture articulatory differences relevant to linguistic questions, we examined the impact through a 2D scan using the timing similar to our original acquisition with one imaging line per navigator at 112 fps and our new timing with four imaging lines per navigator at 35.6 fps. The subject was asked to say “Adam climbed a ladder.” This sample contains three instances of central alveolar consonants (the orthographic <d> in “Adam,” “climbed,” and “ladder”). We anticipated that the relevant consonants in “Adam” and “ladder” would be produced as a “flap” /fl/: a consonant that manifests a rapid, primarily up-down, movement of the tongue tip or blade. One authority has emphasized the “momentary” characteristic of this motion. Indeed, the alveolar flap is widely regarded as one of the quickest lingual articulatory gestures associated with the sounds of American English. The duration of the sound lies between 10 and 35 ms. In a study of tongue contact across the roof of the mouth, it was found that the flap consonant had a more retracted tongue position than a typical /d/ (as in “climbed”) (pp. 414–415). We examined the resulting images to determine whether our new imaging paradigm is fast enough to capture the spatiotemporal characteristics of this consonant.

From Figure 2 we observe that although the time resolution is decreased to about 30% of our original frame rate, the flapping tongue movement can still be visualized clearly in the time-strip map both in superior–inferior and anterior–posterior direction, and there is no loss of visualization of articulatory movements.

Having substantiated that our method was capable of capturing some of the fastest movements in speech, we also set out to determine whether our method can differentiate articulatory movements in 3 dimensions. To do so, we compared the tongue position in the consonants /l/ and /t/ in the words “look” and “token” from the Rainbow Passage. The /l/ in “look” is an example of an alveolar lateral consonant: The tongue tip or blade touches the alveolar ridge with the side or sides of the tongue lowered from the roof of the mouth (we will call this unilateral or bilateral lowering). During the production of /t/, which is an alveolar central consonant, the tongue tip or blade briefly raises to touch the alveolar ridge and then lowers; the sides of the tongue are raised during the sound. Thus, the critical distinction between /l/ and /t/ has to do with the lateral edges of the tongue. The articulatory dynamics of these consonants require examination in multiple dimensions. The coronal view, in particular, is necessary for detecting unilateral or bilateral tongue lowering during /l/ and its absence during /t/.

Figure 3 shows a single frame of the 3D data in all three views, where we can see the lateral lowering of the tongue in /l/ in sagittal plains compared with the absence of such lowering during /t/.

To demonstrate that we are able to reliably measure the significant difference between the tongue posture in /l/ and /t/, we present another sample, /l/, in the word “light.” Figure 4 shows the magnitude subtraction map and negative log p-value map for the differences related to sample /l/ook, /t/oken, and /l/ight overlayed on the anatomical...
FIGURE 3 Illustration of three-directional oriented movie frames of two speech samples for each subject. Left images of subjects 1–4 show sagittal, coronal, and axial images during /l/ in the word “look.” Right images show sagittal, coronal, and axial images during /t/ in the word “token.”

coronal images. The t-test was performed across the nine repetitions of the speech sample during the full acquisition. The value shown for the negative log p-value map is $a_i = -\log(p_i)$, where $p_i$ are the p-values from the t-test for each voxel. The larger difference in Figure A, B and higher signal in Figure 4C, D illustrate that the difference in signal intensity between /l/ and /t/ is much greater than between the two different /l/ for all 4 subjects. Also, we can observe that the main difference between /l/ and /t/ is that the lateral edges are lower for /l/, while the middle rises at the indicated position, compared with /t/. In subjects 2 and 3, we observe unilateral lowering of the tongue, made clearly visible in the coronal image, with the right side of the tongue dropping more than the left.

In Figure 4 we observe the dramatic differences in tongue shape for lateral and central alveolar consonants (/l/ and /t/), expressed in the coronal plane. The differences between the lateral consonant /l/ co-articulated with the following /a/ (in “look”) or /a/ (in “light”) are more subtle, but some interesting nuances are still detectable using our method. Based on Figure 4A, C, we can observe the bilateral lowering of the tongue for subjects 1 and 4 in the production of /t/ versus /l/. The primary signal of this difference comes from the dark blue shading over both sides of the tongue (most notably for subject 1). Subjects 2 and 3 exhibit unilateral lowering with a bias toward the speaker’s right side; this gesture is most clearly exhibited by subject 2. Differences between the /l/ consonants in “light” and “look” are mostly related to the vertical position of the tongue, as expected; the tongue is lower for the /l/ in “light” because it precedes a vowel sound that initiates with a low tongue position. This lowering is not particularly dramatic because the tongue is still raised to produce the /l/. The slight difference is indicative of co-articulation with the following vowel sound, as the tongue prepares to produce it.

Another demonstration of our method is the magnitude of deformation maps of three speech samples (Figure 5). We calculate deformation maps from a resting position before the speech starts compared with the position for particular speech samples. We pick a frame of reference from the resting state during the interval between two repetitions. Specifically, the deformations are found using a diffeomorphic image registration scheme between corresponding image volumes. The advanced normalizing tools provide a symmetric estimation of both the forward and backward deformation fields between each two image volumes. We used the cross-correlation as the similarity metric, which ensures good spatial alignments between our image volumes. The value in the maps indicates the deformation magnitude as the number of voxels of movement from resting state to the sample position, the square root of the sum of the squares of displacement in all three directions.

To gain a better sense of the direction of the “movements” between the rest position and the speech sample, we made quiver plots of the 3D deformation fields at the region of interest, indicating the direction of deformation
FIGURE 4 Images of subjects 1–4 producing /l/ in “look.” Underlying anatomical grayscale images of (A)–(D) are the coronal images corresponding to the red dashed line in the midsagittal image (E). Colormap overlays of (A) indicate the difference in magnitude between the subject speaking /l/ in “look” minus /t/ in “token.” B, The magnitude difference between /l/ in “look” and /l/ in “light.” Colormap overlays of (C) indicate testing parameter \( \alpha_i = -\log(p_i) \), where \( p_i \) are t-test p-values of each voxel of nine speech samples between the subject speaking /l/ in “look” and /l/ in “token.” D, The \( \alpha_i \) between the subject speaking /l/ in “look” and /l/ in “light” or movement. Figure 6 provides the quiver plot in three directions for subject 1 to subject 4 saying the consonant /l/ in “look,” /l/ in “light,” and /t/ in “token,” compared with their “resting state.” The directions of arrows indicate movement directions of resting volume from the “resting state” to “sample state.” The lengths of arrows indicate
the number of voxels “moved” from the “resting state” to “sample state.” Subject 3 does not exhibit movements with so great a magnitude as the others, but lingual differences between /l/ and /t/ can still be observed. Subject 4 manifests similar movement patterns as subject 2 during /l/, along with clear raising of velum structure. Some differences may be due to the resting position assumed by subjects between utterances.

An example of a 3D movie of a subject during the reading of the Rainbow Passage can be viewed in Video S1.

4 | DISCUSSION

Our main goal is to demonstrate the capability of our method for showing subtle changes in both spatial and temporal domains during speech. Although we sacrifice some temporal resolution to keep the total acquisition time shorter, we are still able to capture linguistically relevant features. From the results shown in Figure 2, the tongue tip rises abruptly during the /d/ in “Adam,” “climbed” and “ladder,” as expected. The gesture is quickly “reversed”
FIGURE 6 Three-directional deformation arrow maps from “resting state” to “sample state” for the sample /l/ in “look” (A), sample /l/ in “light” (B), and sample /t/ in “token” (C) for subjects 1–4. The underlying images are the “resting state” images in “climbed” and “ladder” but not in “Adam.” This is presumably because the tongue is not required to lower significantly for the unstressed, second vowel in “Adam,” and its position is unspecified during the subsequent bilabial consonant /m/. One noticeable difference between the flapped variant of /d/, predicted to occur in “Adam” and “ladder” but not in “climbed,” is the shadow observed directly underneath the tongue tip. This is likely indicative of the slight retraction of the tongue during the flap, which was also observed by Stone and Hamlet. It is also possible that the upward movement of the tongue in “climbed” is primarily related to the motion of the jaw, whereas during the flap sound, the tongue tip itself flicks upward, leaving a sublingual space below. This dynamic is most visible in “ladder,” where the shadow beneath the tongue tip appears, and subsequently the tongue tip disappears.
from the selected plane of view during the final rhotacized vowel /ɹ/. One might question why the alveolar consonant /l/ in “climbed” and “ladder” does not manifest the same tongue-tip raising movement as the other alveolar sounds. This is most likely because /l/ is produced at a more retracted position. Although the temporal resolution is decreased, we are still able to image these relatively fast articulatory movements. This is in agreement with the temporal resolution requirement specified by Lingala et al.\textsuperscript{7}

Developed from our previous method,\textsuperscript{10} we reduced the total acquisition time from 30 min to 20 min. This reduction will significantly increase the compliance and comfort of subjects. In the current work, several approaches have been made to further ease subjects’ fatigue. First, we changed our speech samples from repeating simple carefully constructed fragments into reading a full passage displayed on a screen. Second, the subjects are made aware that they will read the passage through about 10 times during the scan, giving them a sense of time and progress. However, further reductions are needed, as 20 min is still long, particularly important when imaging pediatric patients. Although not done in the data in this paper, we can break up the 20-min scan into shorter runs and change the passage being read for each run, as our approach does not require explicit repeats of the same passage. We have seen some success with this approach in children with our previous method, although the current proposed shorter overall acquisition time is still likely necessary to achieve high success with children subjects. Further shortening of acquisition time may result from better a priori models and advanced regularization.\textsuperscript{23}

To determine the reconstruction parameters, we made several simulations to explore the effects of rank $L$, regularization parameter $\lambda$, penalization term $L^1$-norm, $L^2$-norm, Huber penalty, and fraction of data being used. To choose a proper rank, we need to assure an adequate number of temporal bases to cover the speech articulatory dynamics. From Figure 7A, insufficient ranks, such as 10 and 30, are not capable of capturing good dynamics. High rank numbers, like 70 and 90, capture more random noise, which will reduce the SNR significantly. The proper rank was found to be approximately $L = 50$, to adequately capture dynamics and reduce the noise. Figure 7B indicates that higher regularization parameter will help in penalizing noise and artifacts. High $\lambda$ value ($\lambda = 7000$) does not show improvements in visualization of articulatory dynamics between $\lambda = 5000$ and 3000, which means $\lambda =$
5000 is an adequate value for the penalty in our data. Evaluation of the discrepancy principle for a particular scanner setup is an important step for adjusting the regularization parameter but should be consistent across scans of a similar imaging protocol.\textsuperscript{24,25} Figure 7C shows the expected property of the three penalization approaches. Based on the theory of Ramani and Fessler,\textsuperscript{26} \(L^1\)-norm penalization can be performed by a splitting-based algorithm with an alternating direction method of multipliers (ADMM). To perform the \(L^2\)-norm penalization, we set a quadratic function to substitute the Huber function (Equation [3]). As shown in Figure 7C, the \(L^2\)-norm resulted in image blurring, whereas \(L^1\)-norm and Huber penalty performed better on denoising while preserving edges. Generally, the \(L^1\)-norm with alternating direction method of multipliers and Huber penalty had a similar reconstruction quality. Because our implementation of the Huber penalty framework is more computationally efficient, we have chosen this for the reconstructions in this paper. Further use of alternating direction method of multipliers should be considered in future work. Figure 7D demonstrates that the quality of the spatiotemporal image is improved as we increase the amount of fully sampled frames included in the reconstruction. With fewer fully sampled frames than the model order (25% case), the image quality is severely degraded due to ill-posedness of the inverse problem. In the 50% sampling case, artifacts persist, especially at the tongue tip. Using more full-frame data than the model order results in high-quality spatiotemporal images.

The benefits of the additional information that high spatial resolution with sufficient axial coverage can achieve is demonstrated in Figure 3. For subjects 1 and 2, the axial images present shadows on the lateral edges of the front of the tongue, whereas this deep shadowing is not present in axial images of /t/. Indeed, the denser shadows on the right side of the tongue are indicative of the unilateral lowering or lingual asymmetry reported by Hamlet.\textsuperscript{27} In this case, the right side of the tongue is lower than the left side. The pattern is not as clear for subjects 3 and 4 but our technique still allows us to observe a different lingual configuration much different from that in /t/. For example, in the axial image of subject 4’s /l/, we observe (in shadow, at the midline of the tongue) the cavity or hollow in the tongue that produces the characteristic antiresonances associated with the acoustics of /l/.

The results of Figure 4 demonstrate our method’s ability to differentiate articulatory motions on a scale of milliseconds. Phonetics textbooks have long observed that the tongue may be lowered bilaterally or unilaterally during the production of /l/.\textsuperscript{28,29} Among 357 subjects, Hamlet\textsuperscript{27} found a nearly equal split between subjects who lower the tongue bilaterally; among those who lower the tongue unilaterally, the split between right side and left side was also nearly equal (with only a slight advantage for dextral lowering). There was no effect of handedness. The acoustic effects of bilateral or unilateral lowering on /l/ are poorly understood; it has been claimed that they make no difference in the acoustic signature of /l/.\textsuperscript{28} However, Hamlet\textsuperscript{27} concludes: “In defective speech, the likelihood is quite strong of encountering both marked anatomical asymmetries … and neurologically based asymmetries of motion.” We note how successfully our imaging method directly captures differences in unilateral and bilateral tongue lowering typical of /l/ (see, for example, Hamlet,\textsuperscript{27} in which speakers self-reported the sensation of coolness on the tongue while maintaining the articulatory posture of /l/ and breathing inward). The difference between (alveolar) lateral and central consonants can be illustrated through comparison of our imaging results for /l/ and /t/.

These results are further enhanced when looking at displacements in Figure 5, which shows differences in magnitude of motion between /l/ and /t/. For example, the magnitude of tongue movements is lower for subjects 1, 2, and 4. Subject 3 shows more lateral tongue movements in /l/ productions; these differences are likely related to the shadows observed on the tongue in Figure 3. There are also obvious velum movements shown for /t/ productions in subjects 1, 2, and 4. We can also observe an asymmetric pattern in /l/ productions, especially from the axial images of subjects 1 and 2, in whom the right side of the tongue moves more than the left side of the tongue. This provides direct imaging evidence of bilateral and unilateral lowering between subjects. This behavior is further evidence of asymmetries in tongue posture\textsuperscript{27} and may help us better understand the acoustic consequences of lingual geometry during lateral production. Thanks to the high temporal resolution our approach achieves, there is potential to observe more subtle articulatory changes and muscle movements of very short duration.

The deformation images of Figures 5 and 6 were formed by extracting multiple instances of the speech samples from the whole scan, usually from 9 to 10, and making the input of deformation estimation to be the average-magnitude images. This approach cannot be used to examine the differences in the speech execution across utterances of the speech sample, but instead the average deformations for the speech sample for the participant. Deformation maps of individual executions of speech samples could be examined for variability in future work.

Figure 6 provides even more information about subject-specific articulatory differences by showing us the directionality of the motion. In subject 1, the tongue moves downward during /l/, but coronal and axial slices also show evidence of an asymmetric pattern of tongue movement. This creates a volume at the right side of the tongue for airflow to pass into the oral cavity (it may also be

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
Subject & /l/ Production & /t/ Production \\
\hline
1 & Lowering & Lowering \\
\hline
2 & Lowering & Lowering \\
\hline
3 & Lowering & Lowering \\
\hline
4 & Lowering & Lowering \\
\hline
\end{tabular}
\caption{Subject-specific articulatory differences.}
\end{table}
indicative of the concavity in the tongue that generates antiresonances specific to the acoustics of /l/). This matches our observations in Figure 5 and provides confirmation of the lingual asymmetry in laterals noted elsewhere. In the sample /t/, the tongue is returning primarily to a low position. In the midsagittal images, we also observe the raising of the velum, a necessary precondition for an oral sound. Coarticulation with the following vowel is also evident, particularly in the midsagittal images. For example, in subject 2, during the lateral consonant in “look,” we find the dorsum of the tongue rising, whereas in “light” we see the root of the tongue retracting toward the posterior pharyngeal wall. These movements are consistent with the articulatory posture of /a/ on the one hand, and (the initial phase) of /a/ on the other. Across subjects, there is less lingual movement during the /t/ in “token,” suggesting that it differs from /l/ in being more resistant to coarticulation with the following vowel.

We note with some concern the high degree of variability between subjects in Figures 5 and 6. This is probably indicative of different postures during the state that we are using as a reference—the “resting” position. Although we might expect the resting position between vocalizations to be consistent across speakers, our data may suggest otherwise. Based on these results, it may be wise to compare only speech samples without reference to “resting” position, unless a protocol can be determined to produce consistent results when a subject is asked to “rest” or “pause” between vocalizations—a topic that we will examine in future work.

The current approach leverages a low-rank model of the dynamic anatomy during speech in the oropharyngeal cavity. This method relies on significant energy being present in the time series for the dynamic motions involved in speech samples. Although explicit repetitions of speech samples are not needed, the articulators involved in the motions need to visit similar postures multiple times across the dynamic time series to create that energy. This is easily accommodated by providing speech samples for the subjects to read and having them read through them multiple times. However, rare events, like occasional swallows, will not be well visualized in the current approach, as the dynamics of the speech articulators will be the major components of the time series. Although the oral phase of swallowing will have overlap with speech movements, the pharyngeal phase is different and will not be represented well in the reconstructed images.

In this paper, we fully sampled the k-space data for each full data frame, providing full sampling of the model. We randomly shuffled the order of the ky and kz encodes with no specific distribution function. Random sampling helps avoid spurious correlations between dynamic motions and the k-space sampling of imaging data and improves the robustness of the reconstruction as has been shown in Lustig et al. \(^{30}\) In future work, we will examine whether the acquisition can be further shortened through variable density random sampling.

5 | CONCLUSIONS

The novel speech MRI approach we developed with 2-mm near-isotropic spatial resolution, 64-mm total axial coverage, and a sufficient temporal resolution of up to 35 fps across 20 min total scan time provides a high spatiotemporal-resolution environment for studying speech dynamics. By analyzing the results of multiple analytic techniques, we have been able to confirm bilateral and unilateral tongue lowering during /l/ in multiple subjects. This unique tongue geometry must be imaged in 3 dimensions (or at least multiple 2D planes) to observe its occurrence. The technique also allows us to capture subtle articulatory changes and muscle movements associated with coarticulatory effects. Because of these developments, our methods will provide rich opportunities for exploring articulatory questions of linguistic relevance in future studies.

FUNDING INFORMATION

The National Institute of Dental & Craniofacial Research of the National Institutes of Health (R01DE027989). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was conducted in part at the Biomedical Imaging Center of the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign.

ORCID

Riwei Jin https://orcid.org/0000-0002-2982-1123
Bradley P. Sutton https://orcid.org/0000-0002-8443-0408

REFERENCES

1. Narayanan S, Nayak K, 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.
2. 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.
3. Kotlarek KJ, Perry JL, Fang X. Morphology of the levator veli palatini muscle in adults with repaired cleft palate. J Craniofac Surg. 2017;28:833-837.
4. Perry JL, Kuehn DP, Sutton BP, Fang X. Velopharyngeal structural and functional assessment of speech in young children
using dynamic magnetic resonance imaging. *Cleft Palate Craniofac J.* 2017;54:408-422.

5. Carignan C, Shosted R, Fu MJ, Liang ZP, Sutton B. The role of the pharynx and tongue in enhancement of vowel nasalization: a real-time MRI investigation of French nasal vowels. *Interspeech.* 2013;3042-3046.

6. Barlaz M, Shosted R, Fu M, Sutton B. Oropharyngeal articulation of phonemic and phonetic nasalization in Brazilian Portuguese. *J Phonetics.* 2018;71:81-97.

7. Lingala SG, Sutton BP, Miquel ME, Nayak KS. Recommendations for real-time speech MRI. *J Magn Reson Imaging.* 2016;43:28-44.

8. 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.

9. 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.

10. 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.

11. Fu M, Zhao B, Carignan C, et al. High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. *Magn Reson Med.* 2015;73:1820-1832.

12. Jin R, Liang ZP, Sutton B. Increasing three-dimensional coverage of dynamic speech magnetic resonance imaging. In: Proceedings of the Annual Meeting of ISMRM [Virtual], 2021. p. 4175.

13. Glover GH. Simple analytic spiral K-space algorithm. *Magn Reson Med.* 1999;42:412-415.

14. Liang Z-P. Spatiotemporal imaging with partially separable functions. *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro,* 2007:988-991.

15. Buehrer M, Pruessmann KP, Boesiger P, Kozerke S. Array compression for MRI with large coil arrays. *Magn Reson Med.* 2007;57:1131-1139.

16. Yu DF, Fessler JA. Edge-preserving tomographic reconstruction with nonlocal regularization. *IEEE Trans Med Imaging.* 2002;21:159-173.

17. Vaz C, Ramanarayanan V, Narayanan S. Acoustic denoising using dictionary learning with spectral and temporal regularization. *IEEE/ACM Trans Audio Speech Lang Process.* 2018;26:967-980.

18. Fairbanks G. The rainbow passage. *Voice and Articulation Drillbook,* Vol. 2. Harper & Row; 1960.

19. Malecot A, Lloyd P. The /t/:/d/ distinction in American alveolar flaps. *Lingua.* 1969;19:264-272.

20. Zue V, Laferriere M. Acoustical study of medial /t/,/d/ in American English. *J Acoust Soc Am.* 1979;66:1039-1050.

21. Stone M, Hamlet S. Variations in jaw and tongue gestures observed during the production of unstressed /d/’s and flaps. *J Phonetics.* 1982;10:401-415.

22. Avants BB, Tustison N, Song G. Advanced normalization tools (ANTS). *Insight J.* 2009;2:1-35.

23. Fu M, Woo J, Liang Z-P, Sutton B. Spatiotemporal atlas based dynamic speech imaging. In: Proceedings of SPIE Medical Imaging, San Diego, California, USA, 2016. pp. 20–28.

24. Wen Y, Chan R. Parameter selection for total-variation-based image restoration using discrepancy principle. *IEEE Trans Image Process.* 2012;21:1770-1781.

25. Blomgren P, Chan T. Modular solvers for image restoration problems using the discrepancy principle. *Numer Linear Algebra Appl.* 2002;9:347-358.

26. Ramani S, Fessler JA. A splitting-based iterative algorithm for accelerated statistical X-ray CT reconstruction. *IEEE Trans Med Imaging.* 2012;31:677-688.

27. Hamlet SL. Handedness and articulatory asymmetries on /s/ and /l/. *J Phonetics.* 1987;15:191-195.

28. Malmberg B. *Phonetics.* Dover Publications; 1963.

29. Ladefoged P. *Preliminariesto Linguistics Phonetics.* The University of Chicago Press; 1971.

30. Lustig M, Santos JM, Donoho DL, Pauly JM. K-t Sparse: high frame rate dynamic MRI exploiting spatio-temporal sparsity. In: Proceedings of the 14th Annual Meeting of ISMRM, Seattle, Washington, USA, 2006. p. 2420.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**Video S1.** 3D movie of a subject during the reading of the Rainbow Passage.

**How to cite this article:** Jin R, Shosted RK, Xing F, et al. Enhancing linguistic research through 2-mm isotropic 3D dynamic speech MRI optimized by sparse temporal sampling and low-rank reconstruction. *Magn Reson Med.* 2023;89:652-664. doi: 10.1002/mrm.29486