High-Frame-Rate Full-Vocal-Tract 3D Dynamic Speech Imaging

Maojing Fu[1,2], Marissa S. Barlaz[3], Joseph L. Holtrop[2,4], Jamie L. Perry[5], David P. Kuehn[6], Ryan K. Shosted[2,3], Zhi-Pei Liang[1,2], Bradley P. Sutton[2,4]

[1] Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois

[2] Beckman Institute of Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois

[3] Linguistics, University of Illinois at Urbana-Champaign, Urbana, Illinois

[4] Bioengineering, University of Illinois at Urbana-Champaign, Urbana, Illinois

[5] Communication Sciences and Disorders, East Carolina University, Greenville, North Carolina

[6] Speech and Hearing Science, University of Illinois at Urbana-Champaign, Urbana, Illinois

Running title: High-Frame-Rate Full-Vocal-Tract 3D Dynamic Speech Imaging

Submission category: Note

Correspondence to:

Maojing Fu

4259 Beckman Institute for Advanced Science and Technology

405 N Mathews Ave, Urbana, IL 61801, USA

E-mail: mfu2@illinois.edu

Approximate word count for the manuscript body: 2800 words.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version record. Please cite this article as doi:10.1002/mrm.26248.

ABSTRACT

Purpose: To achieve high temporal frame rate, high spatial resolution and full-vocal-tract coverage for three-dimensional (3D) dynamic speech MRI by using low-rank modeling and sparse sampling.

Methods: 3D dynamic speech MRI is enabled by integrating a novel data acquisition strategy and an image reconstruction method with the Partial Separability (PS) model: a) a self-navigated sparse sampling strategy that accelerates data acquisition by collecting high-nominal-frame-rate cone navigators and imaging data within a single TR, and b) a reconstruction method that recovers high-quality speech dynamics from sparse $(\mathbf{k}, t)$-space data by enforcing joint low-rank and spatiotemporal total variation (TV) constraints.

Results: The proposed method has been evaluated through in vivo experiments. A nominal temporal frame rate of 166 frames per second (defined based on a repetition time of 5.99 ms) was achieved for an imaging volume covering the entire vocal tract with a spatial resolution of $2.2 \times 2.2 \times 5.0$ mm^3. Practical utility of the proposed method was demonstrated via both validation experiments and a phonetics investigation.

Conclusion: 3D dynamic speech imaging is possible with full-vocal-tract coverage, high spatial resolution and high nominal frame rate to provide dynamic speech data useful for phonetic studies.

Keywords: partial separability, sparsity, low-rank approximation, cone navigation, dynamic speech imaging

INTRODUCTION

Dynamic MRI is useful for research and clinical studies in speech, especially in capturing structural and functional changes of the vocal tract. Recent applications of dynamic MRI to speech-related studies include investigating articulatory dynamics (1–3), studying phonetic variability (4–7), learning language variation (8–14), examining physiological defects (15–17), monitoring swallow functions (18, 19) and observing professional singing or voice training (20, 21). A review of the clinical needs and impact of dynamic speech MRI can be found in (22). Despite its wide applications, the clinical impact of dynamic speech MRI is still limited by the intrinsic trade-offs between speed, coverage and resolution.

Imaging speed is critical for many speech-related studies because structural changes within short intervals, even as short as 10 ms (9, 23), may contain important information about speech function. Significant efforts have been made to improve the speed of speech MRI, resulting in a number of promising methods, which include a) fast-scanning methods with specialized pulse sequences (24, 25) or parallel imaging (26, 27), and b) model-based methods utilizing support constraints (28, 29), low-dimensional signal models (30–34) and sparsity constraints (35–39). Effective integration of these methods has also led to greater imaging speed (23, 40–45).

Full-vocal-tract three-dimensional (3D) coverage and high spatial resolution are also important for clinical applications of speech MRI. Broad spatial coverage is especially useful for phonetics studies as sound production may require coordinated movements from multiple structures with complex geometries at different locations of the vocal tract. In particular, 3D coverage over the entire vocal tract has shown great benefits for a number of speech-related studies (22, 46). High spatial resolution, on the other hand, is needed to delineate the gestures of small-scaled articulators. For example, a tagged-MRI study has imaged the tip of the tongue with a high spatial resolution of 1.9 mm in order to properly model its motion (47). A review of imaging protocols for dynamic speech MRI and the associated acquisition and reconstruction tools can be found in (48).

Although the past decade has seen significant improvement separately in the speed, coverage or resolution of speech MRI, it remains challenging to achieve all three properties at the same time. Recently, the partial separability (PS) model-based methods have shown great potential to balance the trade-offs between these properties. For instance, a two-dimensional (2D) multislice-based approach has enabled visualization of 8-slice speech dynamics at 12.8 fps (23). A 3D acquisition-based approach also allowed articulatory gestures to be captured at 8.1 fps (46). Expanding upon our earlier approaches (23, 49, 50), this work aims at achieving

full 3D dynamic speech MRI with simultaneous high temporal frame rate, broad spatial coverage and high spatial resolution. This goal is achieved by integrating novel data acquisition strategies with PS model-based 3D acquisition and reconstruction methods. A nominal frame rate of 166 fps and a spatial resolution of 2.2 $\times$ 2.2 $\times$ 5.0 mm^3 are achieved with full vocal-tract coverage across 8 slices. The practical utility of our approach has been systematically evaluated by numerical simulations, validation experiments and a phonetics study on American English flaps.

THEORY

Subspace Modeling

In a dynamic speech MRI experiment, the measured data from $(\mathbf{k}, t)$-space can be expressed as:

$$d(\mathbf{k}, t) = \int I(\mathbf{r}, t) e^{-i2\pi \mathbf{k}\cdot\mathbf{r}} d\mathbf{r} + \eta(\mathbf{k}, t), \tag{1}$$

where $I(\mathbf{r}, t)$ is the desired image series and $\eta(\mathbf{k}, t)$ is the measurement noise. For speech MRI experiments, $I(\mathbf{r}, t)$ often manifests strong spatiotemporal correlation because a) speech motions are composed of movements of a limited number of articulators; and b) the temporal profiles representing the bulk motions of the articulators bear certain level of resemblance across different speech sounds. The PS model quantitatively represents this spatiotemporal correlation with Lth-order partially separable functions (30):

$$I(\mathbf{r}, t) = \sum_{l=1}^{L} \psi_l(\mathbf{r}) \phi_l(t), \tag{2}$$

where $\{\psi_l(\mathbf{r})\}_{l=1}^{L}$ denotes a set of spatial basis functions and $\{\phi_l(t)\}_{l=1}^{L}$ denotes a set of temporal basis functions. Also, it has been shown that the Casorati matrix $\hat{\mathbf{I}}$ defined over the point set $\{I(\mathbf{r}_n, t_m)\}_{n,m=1}^{N,M}$,

$$\hat{\mathbf{I}} = \begin{bmatrix} I(\mathbf{r}_1, t_1) & \cdots & I(\mathbf{r}_1, t_M) \\ \vdots & \ddots & \vdots \\ I(\mathbf{r}_N, t_1) & \cdots & I(\mathbf{r}_N, t_M) \end{bmatrix}, \tag{3}$$

is a low-rank matrix with its rank upper bounded by L (30, 32), where N and M are the number of spatial encodings and temporal frames. This implies that $\hat{\mathbf{I}}$ lives in an L-dimensional subspace ($L \ll \min\{N, M\}$) and allows the factorization $\hat{\mathbf{I}} = \mathbf{UV}$, where columns of $\mathbf{U} \in \mathbb{C}^{N\times L}$ span the spatial subspace of $\hat{\mathbf{I}}$ and rows

of $\mathbf{V} \in \mathbb{C}^{L\times M}$ span the temporal subspace of $\hat{\mathbf{I}}$ (23, 30). By leveraging the correlations between the spatial and temporal information, the PS model enables an acquisition and reconstruction method that simultaneously achieves high spatial resolution and high temporal frame rate, as outlined below.

Data Acquisition

The $(\mathbf{k}, t)$-space is sparsely sampled to obtain two data sets, the navigator and imaging data, in an interleaved fashion (23, 49, 50). Specifically, the navigator data are acquired using a cone trajectory (23) that traverses extended 3D $\mathbf{k}$-space within short temporal intervals. This cone trajectory is chosen because it allows a good trade-off between navigation speed, signal-to-noise ratio (SNR), and $\mathbf{k}$-space coverage in both the low- and high-spatial-frequency regions (23). The imaging data are acquired from distributed $\mathbf{k}$-space using Cartesian trajectories with random phase encoding orders. The use of Cartesian trajectories greatly simplifies the reconstruction problem and results in low image distortions from magnetic susceptibility (23).

We accelerate the acquisition with a "self-navigation" strategy. Unlike previous approaches that collect navigator data with a separate radio frequency (RF) excitation (23, 49, 50), this self-navigation strategy combines the acquisition of both navigator and imaging data into one single repetition time (TR) using a multi-echo readout (53). This combined acquisition of both data sets is particularly desirable for speech imaging applications not only because it effectively increases the imaging speed by shortening TR, but also because it avoids missing temporal components that associate with important articulatory dynamics. Figure 1 illustrates the proposed acquisition strategy with a simplified pulse sequence diagram.

Additional considerations are given to reduce the sensitivity of the acquisition to eddy currents effects. For the above "self-navigation" strategy, it should be noted that a rephasing gradient is added prior to navigator acquisition in the second gradient echo. This rephasing gradient ensures the navigator trajectory starts from the center of $\mathbf{k}$-space at each TR. Considering this, it is desirable to minimize the length of this gradient in pursuance of shorter TR (so that the imaging speed is increased). However, eddy currents generated from the prior imaging acquisition and rephasing gradients vary from TR to TR due to the random phase encoding and directly impact the temporal signatures of the navigator data. This results in biased estimation of the temporal subspace and unwanted temporal dynamics in the reconstructions. To address this issue, we chose the shortest length for the rephasing gradient that result in no noticeable eddy current effects, but also maintains a high imaging speed, determined by trial-and-error tuning of its gradient duration and ramps while examining the

resulting reconstructions for contamination from eddy currents.

Image Reconstruction

Image reconstruction based on the PS model requires determining both the temporal and spatial subspaces, i.e., matrix $\mathbf{U}$ and $\mathbf{V}$, from the acquired data. In this work, we choose to determine $\mathbf{U}$ and $\mathbf{V}$ in two separate steps. Specifically, $\mathbf{V}$ is determined by performing singular value decomposition (SVD) on the navigator data (23, 30). With $\mathbf{V}$ determined, estimation of $\mathbf{U}$ can be formulated as a least-squares fitting problem (30) and is greatly simplified as compared to alternative methods that jointly estimate $\mathbf{U}$ and $\mathbf{V}$.

PS model-based image reconstruction using joint low-rank and sparsity constraints has been developed in (33). The use of joint constraints has been shown to improve the conditioning of the PS model fitting problem, and the improvement is especially apparent when high model order is required but limited measurements are available (23, 33). Also, the spatiotemporal total variation (TV) constraint has been applied to dynamic imaging in (33, 54, 55), such as dynamic myocardial perfusion MRI (54). Compared with a spatial-spectral sparsity constraint (23), the spatiotemporal TV constraint simultaneously penalizes finite differences in the spatial and temporal domains, so that the articulatory dynamics are preserved as spatiotemporal edges in the reconstructions. Extending upon these previous approaches (23, 33, 54, 55), we develop a method that jointly imposes the low-rank and spatiotemporal TV constraints.

The image reconstruction problem can be formulated as follows:

$$\hat{\mathbf{U}} = \underset{\mathbf{U}\in\mathbb{C}^{N\times L}}{\arg\min} \sum_{q=1}^{Q} ||\mathbf{\Omega}\{\mathbf{F}\mathbf{S}_q\mathbf{U}\mathbf{V}\} - \mathbf{d}_q||_2^2 + \lambda\, \mathrm{TV}\{\mathbf{U}\mathbf{V}\}, \tag{4}$$

where Q denotes the number of receiver coils, $\mathbf{\Omega}\{\cdot\}$ denotes a sparse sampling operator corresponding to the acquisition of the imaging data (and vectorizing the acquired data in a columnwise fashion), $\mathbf{F}$ denotes a spatial Fourier transform matrix, $\mathbf{S}_q$ denotes the sensitivity map of the qth coil, $\mathbf{d}_q$ denotes the sparsely acquired imaging data samples from the qth receiver coil, and λ denotes a regularization parameter. The TV operator is defined as $\mathrm{TV}\{\mathbf{U}\mathbf{V}\} = \sum_{j=1}^{M}\sum_{i=1}^{N} ||\mathbf{D}_i\mathbf{U}\mathbf{V}_j||_1$, where $\mathbf{V}_j$ is the jth column of V, $\mathbf{D}_i \in \mathbb{C}^{3\times N}$ is a gradient operator taking finite differences at the ith pixel of the image along spatially horizontal, spatially vertical and temporal directions (finite difference along the slice direction was not incorporated due to computational considerations). A numerical algorithm based on half-quadratic regularization with continuation is applied to solve Eq. 4 (55).

METHODS

Experiments were performed on a Siemens Trio scanner (Siemens Medical Solutions, Erlangen, Germany) with the following features: a field strength of 3 T, a gradient strength of 40 mTm^{-1}, a maximum slew rate of 176 $\mathrm{Tm}^{-1}\mathrm{s}^{-1}$ and a 12-channel head receiver coil. Based on the proposed self-navigation strategy, a FLASH sequence has been developed to acquire data with the following parameters: a TR of 5.99 ms, an echo time (TE) of 1.85 ms for the imaging data, a TE of 3.25 ms for navigator data, an acquisition matrix size of $128 \times 128 \times 8$, a FOV of $280 \times 280 \times 40$ mm^3 and a spatial resolution of $2.2 \times 2.2 \times 5.0$ mm^3. When acquiring the necessary data that targets at a model order of around 70, as was done in this work, the acquisition time was 7 min 12 s (increasing the model order by 1 requires an increase of approximately 6.13 s in acquisition time). With the proposed image reconstruction method, the recovered image sequence allows visualizing the entire vocal tract at a nominal frame rate of 166 fps (defined based on the reconstruction of a full 3D volume at each TR of 5.99 ms).

Prior to the acquisition of the dynamic imaging data, a pilot scan was performed to determine the sensitivity profiles of the receiver coils. The estimated sensitivity profiles were assumed to be time-invariant for the subsequent image reconstruction. During the acquisitions, the voice of subjects was simultaneously recorded at a sampling rate of 8 kHz through a fiber-optic microphone with active noise cancellation (Dual Channel FOMRI, Optoacoustics, Or Yehuda, Israel). The head motion of each subject was minimized by fixing the positions of the patient's head in the receiver coil with foam pads. Informed consents were obtained for all subjects and the experiment was carried out in accordance with regulations of the Institutional Review Board at the University of Illinois at Urbana-Champaign.

Numerical Simulations

A generic numerical phantom for 3D dynamic speech MRI has been created and simulation studies have been performed to characterize the performance of the proposed method. The phantom was designed to simulate multi-channel, complex-valued dynamic speech imaging data. Specifically, this numerical phantom was constructed from an initial reconstruction from an in vivo dynamic MRI experiment, where the subject was requested to produce repetitions of /loo/-/la/-/lee/-/la/ sounds at his own speaking rate. The created numerical

phantom had a matrix size of 128 × 128 × 8, a FOV of 280 × 280 × 40 mm^3, a spatial resolution of 2.2 × 2.2 × 5.0 mm^3, a TR of 5.99 ms and a total number of 71680 time frames.

Simulated data acquisition followed the $(\mathbf{k}, t)$-space sampling strategy as described in the THEORY section. At each TR, the imaging data were created by taking samples along one Cartesian line in 3D $\mathbf{k}$-space according to a randomized phase encoding order; the navigator data were created by sampling from a 3D cone trajectory in $\mathbf{k}$-space using an NUFFT-based routine (56). Sensitivity profiles were taken directly from the initial scan. White Gaussian noise was added to data from each receiver coil, such that the simulated data had a noise level that was comparable to the in vivo acquisitions. With this simulated sampling strategy, a full set of data was acquired after sampling 71680 time frames from the numerical phantom (equivalent to an acquisition length of 7 min 12 s). Reconstruction from simulated data was performed using the proposed method: $\mathbf{V}$ was first determined by performing SVD on the navigator data; $\mathbf{U}$ was then estimated according to the strategy as described in the THEORY section. A model order of 70 and a regularization parameter of 1.31 × 10^{-6} were chosen as described in the DISCUSSION section. The following reconstruction error was used to quantitatively assess reconstruction quality (33),

$$\text{error} = \frac{||\mathbf{I}_p - \mathbf{UV}||_{\text{F}}}{||\mathbf{I}_p||_{\text{F}}}, \tag{5}$$

where $\mathbf{I}_p$ represents the numerical phantom and $||\cdot||_{\text{F}}$ represents the Frobenius norm. Simulation results on the generic numerical phantom were summarized in Supporting Figure S1 of the supplementary document.

A modified numerical phantom was also created based on the generic phantom to characterize the frame rate achievable with the proposed method. Specifically, the generic phantom was augmented with a high-temporal-frequency flashing pattern - a bright cube was positioned above the subject's forehead and appeared on every other time frame. This flashing pattern was appropriate for characterizing the imaging speed because it requires an effective frame rate of at least 166 fps to be properly captured. Simulation results on the modified numerical phantom were summarized in Supporting Figure S2 of the supplementary document.

Validation Experiments

Four volunteers participated in the validation experiments. Two volunteers were male and the other two were female. All volunteers were native speakers of American English and they had an age range of 23 to 38 years. During data acquisition, volunteers were requested to recurrently produce /loo/-/lee/-/la/-/za/-/na/-/za/ at their natural speaking paces.

To investigate whether the proposed method indeed allows for improved spatiotemporal dynamics, we conducted two additional acquisitions to compare the performance of our 3D imaging method versus that of a previous 2D multi-slice method (23). A male speaker of American English volunteered for both experiments - the 3D acquisition followed the imaging protocol as described above, whereas the 2D acquisition followed a dynamic 8-slice imaging protocol described in (23), where the full frame rate is split across the 8 slices due to interleaved acquisition of each slice, resulting in a nominal frame rate of 12.8 fps. In order to enable roughly synchronized comparison of the associated temporal dynamics, the subject was requested to produce sound following visual cues ("karaoke" scripts of /loo/-/lee/-/la/-/za/-/na/-/za/ sounds) synchronized for both acquisitions. Notice that the mismatch in temporal articulatory motion is minimized with the visual cue, but a certain level of temporal misalignment may still exist with this experiment design.

Application to Phonetics Studies

The proposed method was also applied to a phonetics study on the articulation of flaps - a subset of consonant sounds that are challenging to study because they occur for a brief duration of $\sim$20 ms (57). Many existing methods in articulatory phonetics lack sufficient frame rate to capture the tongue postures associated with these brief events. In this paper, particular interest was placed on applying our method to analyze tongue postures in American English flaps - sounds that are characterized by a single, short closure made with the apex of the tongue contacting the alveolar ridge (57). These flaps are usually realized when an alveolar stop (/t/ or /d/) occurs intervocalically after a stressed syllable. Traditional phonological theories claim that these two flaps (/t/ and /d/) lose all distinction (58) in their acoustic characteristics and, therefore, in their underlying articulatory gestures, as well. Recent experimental studies, however, have implied that a slight acoustic distinction may exist between these flaps (59). It is worth noting that this claim has only been demonstrated using acoustic evidence (60), without imaging evidence of the underlying articulatory differences.

In order to determine whether the articulation of these two flaps manifests any gestural difference, we acquired these flaps with carrier phrases in a dynamic MRI experiment. Specifically, a single female speaker of Mid-Atlantic American English (a dialect known to demonstrate acoustic differences (61)) volunteered as the subject. The speaker was requested to repeat the minimal pair "writing" and "riding" in a carrier phrase ("I said **X** to you") at a normal speaking rate for the length of acquisition. After acquisition, the boundaries of the /t/ and /d/ flaps from the carrier phrase were annotated using the synchronously acquired acoustic signal

and a representative frame associated with each flap was manually selected after the annotation. In addition, a rectangular region of interest (ROI) that included the tongue and oral cavity was defined afterwards on the reconstructed imaging volume for the convenience of ensuing phonetics analysis.

With the annotated flaps and the predefined ROI, two previously developed phonetic analysis methods were applied to reveal the distinction between the flaps. Specifically, a deformation-based analysis method (62, 63) was employed to measure the vertical distances between the tongue tip and alveolar ridge for /t/ and /d/ flaps, respectively, among every frame within the duration of each flap across 137 occurrences. In addition, another analysis method based on the principal component analysis (64) was also applied. Visualization of the principal components (PC) projected back onto the original pixels in the image was displayed using a heat map, which allowed for identification of the relationship between PCs and pixel intensity in different parts of the image. This in turn allowed us to infer the association between PCs and differential movements of the tongue.

RESULTS

Validation Experiments

Figure 2 shows tongue gestures of the upper vocal tract from a 3D imaging experiment where a subject was asked to produce /loo/-/lee/-/la/-/za/-/na/-/za/ sounds. Specifically, Figure 2a and 2b show tongue gestures at the onset of the /l/ sound and /a/ sound in the /la/ syllable, respectively. Although the /l/ and /a/ sounds transition within a brief duration ($\sim$20 ms), apparent differences in tongue gestures are still captured with great spatial detail: the tip of the tongue is elevated towards the alveolar ridge to prepare for the production of /l/ sound, while the tongue retracts to a resting position to produce /a/ sound. In addition, it is noticed that the velum at slice 4 is not in full contact with the velopharyngeal wall. This is not unexpected because /l/ is often classified as a "liquid" consonant, whose production does not require full buildup of intraoral air pressure and tight velopharyngeal closure.

We then investigated if an increased nominal frame rate of the 3D acquisition improves spatiotemporal dynamics over a previous lower-nominal-frame-rate 2D acquisition. Figure 3 shows direct comparison of temporal dynamics from the proposed method (with a nominal frame rate of 166 fps) versus that from a previous 2D multi-slice method (23), where the full frame rate is split across the 8 slices and results in a nominal frame rate of 12.8 fps. Specifically, representative temporal profiles are taken along strips across the tongue tip from

both reconstructions (slight temporal mismatch in tongue motion can be observed as the reconstructions are performed upon two separate, but temporally guided experiments as described in the Methods section). As can be seen, the temporal profile from the 3D method displays sharper temporal transitions of the tongue motion compared with its 2D couterpart. By comparing the associated temporal dynamics along the dashed line segments, it is apparent that a similar temporal motion pattern is shared by the two reconstructions, but the 3D method offers richer spatiotemporal information. Even for regions that involves less motion, such as the lower chin, the 3D method still provides enhanced spatiotemporal dynamics.

To further investigate the shaping of the tongue, we examined the images during the time point of contact between the tongue and the alveolar ridge. As our method captures an imaging volume covering the vocal tract, great flexibility is allowed to visualize tongue gestures in arbitrary planes. For instance, Figure 4a shows mid-sagittal, coronal and axial views of the tongue during the production of /l/ sound. At this time point, the tongue comes into full contact with the alveolar ridge and its gesture is well captured across all view planes. When the tongue retracts to its resting position and continues to the /a/ sound, as seen in Figure 4b, its gesture is reflected in coronal and axial planes as darkened intensity. In addition, incomplete velopharyngeal closure is observed as in Figure 2. This is reasonable because the /l/ sound is immediately followed by the "low" vowel /a/ which itself is often produced with a lowered soft palate position as shown in Figure 4b.

Application to Phonetics Studies

The proposed method provides high-quality spatiotemporal dynamics to systematically study the production of flaps, which is a challenging task to perform if only acoustic recordings are available. Figure 5 shows representative imaging and statistical results in the analysis of /t/ and /d/ flaps during the phrases "I said writing / riding to you". Figure 5a compares the respective tongue gestures between the two flaps - the /t/ flap has a slightly more superior tongue position compared with the /d/ flap. We then quantitatively measured the averaged tongue tip - alveolar ridge distances with a deformation-based method (62) for the /d/ and /t/ flaps over a normalized duration (the original durations are $\sim$66 ms for the /d/ flaps and $\sim$54 ms for the /t/ flaps). As shown in Figure 5b, a larger averaged distance for the /d/ flaps is observed at the beginning of the normalized duration, while the distances from the two flaps converge at the end. In addition, the above results are validated with a phonetic analysis based on both heat maps and principal component (PC) scores (64). As seen with Figure 5c, the /t/ flaps have a higher correlation with PC1 than the /d/ flaps, as evidenced by "greener" pixels in the

heat map around the tongue tip. The higher intensity of green pixels suggests that the /t/ flaps exhibit greater anterior movements (64) in the tongue apex region in its articulation. This result was statistically verified in a one-way ANOVA ($F = 104.7$, $p < 0.001$). The above analyses demonstrate the effectiveness of the proposed method in analyzing sounds that are otherwise difficult to study from acoustic recordings.

DISCUSSION

The proposed method improves imaging speed and spatial coverage by integrating a self-navigation scheme into 3D acquisitions. The self-navigation scheme allows faster sampling as it removes the overhead associated with a separate RF excitation and, instead, collects the navigator data at a later echo time in each TR. In particular, a TR of 5.99 ms is achieved with the proposed self-navigation scheme, while an overall TR of 9.81 ms per slice was used with a "separate RF excitation" approach for previous 2D multislice acquisitions (23). Moreover, the proposed self-navigation scheme optimized the length of the refocusing pulse to ensure there exists no significant impact from the eddy current effects on the reconstruction while maintaining a relatively short TR.

The proposed method reveals subtle temporal and spatial changes in the production of American English flaps. It is known in phonetics that a voiced consonant (such as the /d/ flap) lengthens the sounds preceding it (65), while a voiceless consonant (such as the /t/ flap) does not. The present study provides sufficient frame rate to corroborate this with the finding that the sounds preceding the /d/ flaps are on average $\sim$12 ms longer than those preceding the /t/ flaps. This finding may suggest that a neutralization in vowel space is involved for the first element of the sounds preceding the /t/ flap, whereas a fuller representation of a low back vowel is realized for the /d/ flap. To verify this hypothesis, we compared the associated spatial details (i.e., averaged tongue-tip - alveolar ridge distances) over a normalized duration. The results agree with the hypothesis and show that a lower position of the tongue indeed placed the tongue tip in a more posterior position for the /d/ flaps compared with the /t/ flaps. Our method reveals the subtle difference between two perceptually similar flaps, which are difficult to distinguish via acoustic recordings. The results allow us to conclude that the /t/ and /d/ flaps have articulatory differences, while sharing a general gesture.

Although our method has demonstrated a number of merits, several aspects need to be considered to obtain high-quality reconstructions. First, it is important to rigorously define the true temporal resolution achievable

from our method. In contrast with a linear imaging method, whose resolution property can be characterized by a shift-invariant point spread function (PSF), our method is a nonlinear imaging method and its resolution is difficult to characterize because the associated PSF may be shift-varying or object-dependent. Although improved spatiotemporal dynamics are demonstrated in Figure 3, a rigorous measure of the true temporal resolution is challenging to define as it may vary from voxel-to-voxel, frame-to-frame and non-linearly interact with the spatiotemporal features of the object being imaged. To avoid potentially misleading statements on the truly-achieved resolution, a uniform nominal frame rate was used as an empirical measure of the temporal resolution. It should also be noted that other empirical measures exist to describe the resolution for nonlinear imaging methods. For instance, the local PSF (66) for our method has a full width half maximum of 1.1 voxel and corresponds to a nominal frame rate of 151 fps, but systematic evaluation of the effectiveness of these measures is needed for future research.

Practical limits exist in the available acceleration from the proposed acquisition strategy. Although a high nominal frame rate is achieved to allow potentially the study of short speech patterns, the actual scan time increases on the order of 6 s for each model order increase and the total acquisition length is on the order of 7 min with the current imaging protocol. Also, practical limit in acceleration exists because the self-navigated scheme requires playing out a gradient rephaser prior to the navigator (as illustrated in Figure 1). Higher frame rate requires increased slew rates in the rephaser, which consequently induces eddy currents that we have seen compromise the quality of the navigator. Careful control on gradient switching and slew rate limits, therefore, is essential to prevent the impact of eddy currents on the accuracy of the estimated temporal subspace. In this paper, we carefully limited the length of the rephaser to be 890 μs and we have observed in preliminary studies that shorter durations result in eddy current artifacts as higher-temporal-frequency information spread throughout the reconstructed image. Our method may benefit from theoretical characterization and systematic validations of the eddy current effects in the future.

The proposed method requires the experimenter to determine an appropriate model order L. In this work, L is chosen empirically based on visual inspection of reconstruction quality including the discerniblity of small-sized articulators, the level of temporal blurring during frame-to-frame transitions, and the overall readability of the reconstructed articulatory motion. Based on these features, an L of 70 was chosen in this work and it has consistently yielded good empirical result for the reported experiments. It should be noted, however, that there exists other quantitative-metric-driven approaches (67, 68) to guide the selection of L, although integration of these approaches and systematic evaluation of their performance will be performed in future research.

Like many imaging methods that involve regularized reconstructions, the performance of our method is influenced by the selection of regularization parameters. In this work, the proposed formulation assigns a uniform regularization parameter λ to the spatiotemporal TV constraint along all spatial and temporal dimensions. It should be noted that a refined selection of λ, such as assigning a separate regularization parameter along each of the three spatial dimension and the temporal dimensions, may lead to improved reconstruction quality. Also, the value of λ in this work is chosen based on the discrepancy principle as discussed in (33, 69), while other alternative methods exists (70) and systematic evaluation of these methods will be an interesting future research topic.

The proposed method may pose a computational burden in the context of clinical applications due to the underlying high-dimensional optimization problem involved. For instance, reconstruction time of 71680 frames at 166 fps (defined based on a TR of 5.99 ms) from a data set obtained from a 7 min 12 s scan with 12 receiver channels was around 12 hour 36 min on a 32-core 512GB-memory workstation without code optimization. Acceleration of computation can be realized by leveraging computational methods, such as those exploiting graphical processing units (71), but adaptation of these methods to such a large-scale optimization problem may not be trivial and is beyond the scope of this work.

CONCLUSIONS

High-frame-rate 3D full-vocal-tract dynamic speech MRI has been achieved by exploiting a) an acquisition strategy with 3D spatial encoding and a volumetric self-navigated scheme, and b) an image reconstruction method based on joint low-rank and spatiotemporal-TV constraints. The proposed method has been validated in speech imaging experiments, achieving a nominal imaging speed of 166 fps (defined based on a TR of 5.99 ms) with a spatial resolution of $2.2 \times 2.2 \times 5.0$ mm^3 for an imaging volume covering the entire vocal tract. Its effectiveness has also been demonstrated through a phonetic study on American English flaps.

ACKNOWLEDGMENTS

This work was partially supported by NIH-1R03DC009676-01A1 and NSF-1121780. The authors would also like to acknowledge Dr. Eric Shaffer for providing computational facilities.

References

1. S. M. R. Ventura, D. R. S. Freitas, J. M. R. Tavares. Toward dynamic magnetic resonance imaging of the vocal tract during speech production. J Voice 2011;25:511–518.

2. M. Echternach, M. Markl, B. Richter. Dynamic real-time magnetic resonance imaging for the analysis of voice physiology. Curr Opin Otolaryngol Head Neck Surg 2012;20:450–457.

3. A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, J. Frahm. 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.

4. M. Proctor, L. Goldstein, A. Lammert, D. Byrd, A. Toutios, S. Narayanan. Velic coordination in French nasals: A real-time magnetic resonance imaging study. In Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013, p. 577.

5. C. Smith, M. Proctor, K. Iskarous, L. Goldstein, S. Narayanan. Stable articulatory tasks and their variable formation: Tamil retroflex consonants. In Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013, p. 2006.

6. L. Lu, A. Lammert, V. Ramanarayanan, S. Narayanan. A comparative cross-linguistic study of vocal tract shaping in sibilant fricatives in English, Serbian and Mandarin using real-time magnetic resonance imaging. In Proceedings of the 21st International Congress on Acoustics, Montreal, Canada, 2013, p. 291.

7. A. Zourmand, S. Mirhassani, H. Ting, S. Bux, K. Ng, M. Bilgen, M. Jalaludin. A magnetic resonance imaging study on the articulatory and acoustic speech parameters of Malay vowels. Biomed Eng Online 2014;13:103–123.

8. C. Carignan, R. Shosted, M. Fu, Z.-P. Liang, B. Sutton. The role of the tongue and pharynx in enhancement of vowel nasalization: A real-time MRI investigation of French nasal vowels. In Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013, pp. 340.

9. N. Wong, M. Fu, Z.-P. Liang, R. Shosted, B. Sutton. Observations of perseverative coarticulation in lateral approximants using MRI. In Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013, p. 612.

10. R. Shosted, Z. Hermes, M. Fu, L.-H. Ning, Z.-P. Liang, B. Sutton. Articulating emphasis in Gulf and Levantine Arabic: An rt-MRI approach. In Proceedings of the 1st Experimental Arabic Linguistics Conference, Al Ain, United Arab Emirates, 2013.

11. M. Barlaz, M. Fu, M. Dubin, Z.-P. Liang, R. Shosted, B. Sutton. Lingual differences in Brazilian Portuguese oral and nasal vowels: An MRI study. In Proceedings of the 18th International Congress of Phonetic Sciences, Scotland, UK, 2015.

12. M. Barlaz, M. Fu, C. Carignan, Z.-P. Liang, R. Shosted, B. Sutton. Deformation-based articulatory representations of speech sounds. In Proceedings of the 15th Conference on Laboratory Phonology, Ithaca, USA, 2016, P. 169.

13. Z. Hermes, M. Fu, S. Rose, R. Shosted, B. Sutton. Representations of Place and Airstream Mechanism: A real-time MRI study of Tigrinya ejectives. In Proceedings of the 15th Conference on Laboratory Phonology, Ithaca, USA, 2016, P. 133.

14. Z. Hermes, M. Barlaz, R. Shosted, M. Fu, B. Sutton. The Articulatory Configuration of the Pharynx during the Voiced and Voiceless Pharyngeal Fricatives / ʕ / and / ħ/ in Gulf and Levantine Arabic: A Real-Time Magnetic Resonance Imaging Study. In Proceedings of the 30th Annual Symposium on Arabic Linguistics, Stony Brook, USA, 2016.

15. B. Atik, M. Bekerecioglu, O. Tan, O. Etlik, R. Davran, H. Arslan. Evaluation of dynamic magnetic resonance imaging in assessing velopharyngeal insufficiency during phonation. J Craniofac Surg 2008;19:566–572.

16. B. Wein, M. Drobnitzky, S. Klajman, W. Angerstein. Evaluation of functional positions of tongue and soft palate with MR imaging: initial clinical results. J Magn Reson Imaging 1991;1:381–383.

17. Y. Suto, T. Matsuo, T. Kato, I. Hori, Y. Inoue, S. Ogawa, T. Suzuki, M. Yamada, Y. Ohta. Evaluation of the pharyngeal airway in patients with sleep apnea: value of ultrafast MR imaging. Am J Roentgenol 1993;160:311–314.

18. K. Kumar, V. Shankar, R. Santosham. Assessment of swallowing and its disorders: A dynamic MRI study. Eur J Radiol 2012;82:215–219.

19. S. Zhang, A. Olthoff, J. Frahm. Real-time magnetic resonance imaging of normal swallowing. J Magn Reson Imaging 2012;35:1372–1379.

20. J. Sundberg. Articulatory configuration and pitch in a classically trained soprano singer. J Voice 2009;23:546–551.

21. M. Proctor, E. Bresch, D. Byrd, K. Nayak, S. Narayanan. Paralinguistic mechanisms of production in human beatboxing: A real-time magnetic resonance imaging study. J Acoust Soc Am 2013;133:1043–1054.

22. A. D. Scott, M. Wylezinska, M. J. Birch, M. E. Miquel. Speech MRI: Morphology and function. Eur J Phys 2014;30:604–618.

23. M. Fu, B. Zhao, C. Carignan, R. Shosted, J. Perry, D. Kuehn, Z.-P. Liang, B. Sutton. High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. Magn Reson Med 2015;73:1820–1832.

24. M. Uecker, S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt, J. Frahm. Real-time MRI at a resolution of 20 ms. NMR Biomed 2010;23:986–994.

25. B. P. Sutton, C. A. Conway, Y. Bae, R. Seethamraju, D. P. Kuehn. Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3 T. J Magn Reson Imaging 2010;32:1228–1237.

26. K. Pruessmann, M. Weiger, M. Scheidegger, P. Boesiger. SENSE: Sensitivity encoding for fast MRI. Magn Reson Med 1999;42:952–962.

27. M. Griswold, P. Jakob, R. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, A. Haase. Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn Reson Med 2002;47:1202–1210.

28. J. Tsao, P. Boesiger, K. Pruessmann. k-t BLAST and k-t SENSE: Dynamic MRI with high frame rate exploiting spatiotemporal correlations. Magn Reson Med 2003;50:1031–1042.

29. C. Mistretta, O. Wieben, J. Velikina, W. Block, J. Perry, Y. Wu, and K. Johnson. Highly constrained backprojection for time-resolved MRI. Magn Reson Med 2006;55:30–40.

30. Z.-P. Liang. Spatiotemporal imaging with partially separable functions. in In Proceedings of IEEE International Symposium on Biomedical Imaging, Washington D.C., USA, 2007, pp. 988–991.

31. H. Pedersen, S. Kozerke, S. Ringgaard, K. Nehrke, W. Kim. k-t PCA: temporally constrained k-t BLAST reconstruction using principal component analysis. Magn Reson Med 2009;62:706–716.

32. J. Haldar and Z.-P. Liang. Spatiotemporal imaging with partially separable functions: a matrix recovery approach. In Proceedings of IEEE International Symposium on Biomedical Imaging, Rotterdam, The Netherlands, 2010, pp. 716–719.

33. B. Zhao, J. Haldar, A. Christodoulou, Z.-P. Liang. Image reconstruction from highly undersampled (k, t)-space data with joint partial separability and sparsity constraints. IEEE Trans Med Imaging 2012;31:1809–1820.

34. A. Christodoulou, H. Zhang, B. Zhao, T. Hitchens, C. Ho, Z.-P. Liang. High-resolution cardiovascular MRI by integrating parallel imaging with low-rank and sparse modeling. IEEE Trans Biomed Eng 2013;60:3083–3092.

35. M. Lustig, J. Santos, D. Donoho, J. Pauly. k-t SPARSE: High framerate dynamic MRI exploiting spatio-temporal sparsity. In Proceedings of the 14th Annual Meeting of ISMRM, Seattle, USA, 2006, p. 2420.

36. U. Gamper, P. Boesiger, S. Kozerke. Compressed sensing in dynamic MRI. Magn Reson Med 2008;59:365–373.

37. E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, S. Narayanan. Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging. IEEE Signal Proc Mag 2008;25:123–132.

38. H. Jung, K. Sung, K. Nayak, E. Kim, J. Ye. k-t FOCUSS: A general compressed sensing framework for high resolution dynamic MRI. Magn Reson Med 2009;61:103–116.

39. D. Liang, E. DiBella, R.-R Chen, L. Ying. k-t ISD: Dynamic cardiac MR imaging using compressed sensing with iterative support detection. Magn Reson Med 2012;68:41–53.

40. R. Otazo, D. Kim, L. Axel, D. Sodickson. Combination of compressed sensing and parallel imaging for highly accelerated first-pass cardiac perfusion MRI. Magn Reson Med 2010;64:767–776.

41. S. Lingala, Y. Hu, E. DiBella, M. Jacob. Accelerated dynamic MRI exploiting sparsity and low-rank structure: k-t SLR. IEEE Trans Med Imaging 2011;30, no. 5, pp. 1042–1054.

42. M. Usman, C. Prieto, T. Schaeffter, P. Batchelor. k-t group sparse: A method for accelerating dynamic MRI. Magn Reson Med 2011;66:1163–1176.

43. H. Li, M. Haltmeier, S. Zhang, J. Frahm, A. Munk. Aggregated motion estimation for image reconstruction in real-time MRI. arXiv preprint 2013; arXiv:1304.5054.

44. A. D. Scott, R. Boubertakh, M. J. Birch, M. E. Miquel. Adaptive averaging applied to dynamic imaging of the soft palate. Magn Reson Med 2013;70:865–874.

45. Y.-C. Kim, M. I. Proctor, S. S. Narayanan, K. S. Nayak. Improved imaging of lingual articulation using real-time multislice MRI. J Magn Reson Imag 2012;35:943–948.

46. Y. Zhu, Y. Kim, M. Proctor, S. Narayanan, K. Nanak. Dynamic 3D visualization of vocal tract shaping during speech. In Proceedings of the 19th Annual Meeting of ISMRM, Montreal, Canada, 2011, p. 4355.

47. M. Stone, E. Davis, A. Douglas, M. NessAiver, R. Gullapalli, W. Levine, A. Lundberg. Modeling the motion of the internal tongue from tagged cine-MRI images. J Acoust Soc Am 2001;109:2974–2982.

48. S. Lingala, B. Sutton, M. Miquel, K. Nayak. Recommendations for real-time speech MRI. J Magn Reson Imag 2015; doi:10.1002/jmri.24997;

49. M. Fu, A. Christodoulou, A. Naber, D. Kuehn, Z.-P. Liang, B. Sutton. High-frame-rate multislice speech imaging with sparse sampling of (k, t)-space. In Proceedings of the 20th Annual Meeting of ISMRM, Melbourne, Australia, 2012, p. 12.

50. M. Fu, B. Zhao, J. Holtrop, D. Kuehn, Z.-P. Liang, B. Sutton. High-frame-rate full-vocal-tract imaging based on the partial separability model and volumetric navigation. In Proceedings of the 21st Annual Meeting of ISMRM, Salt Lake City, USA, 2013, p. 4269.

51. C. Brinegar, S. S. Schmitter, N. N. Mistry, G. A. Johnson, Z.-P. Liang. Improving temporal resolution of pulmonary perfusion imaging in rats using the partially separable functions model. Magn Reson Med 2010;64:1162–1170.

52. A. G. Christodoulou, B. Zhao, H. Zhang, C. Ho, Z.-P. Liang. Four-dimensional MR cardiovascular imaging: Method and applications. In Proceedings of IEEE International Symposium on Biomedical Imaging, Chicago, USA, 2011, pp. 3732–3735.

53. A. G. Christodoulou, T. K. Hitchens, Y. L. Wu, C. Ho, Z.-P. Liang. Improved subspace estimation for low-rank model-based accelerated cardiac imaging. IEEE Trans Biomed Eng 2014;61:2451–2457.

54. G. Adluru, C. McGann, P. Speier, E. Kholmovski, A. Shaaban, E. DiBella. Acquisition and reconstruction of undersampled radial data for myocardial perfusion magnetic resonance imaging. J Magn Reson Im 2009;29:466–473.

55. B. Zhao, J. P. Haldar, A. G. Christodoulou, Z.-P. Liang. Further development of image reconstruction from highly undersampled (k, t)-space data with joint partial separbility and sparsity constraints. In Proceedings of IEEE International Symposium on Biomedical Imaging, Chicago, USA, 2011, pp. 1593–1596.

56. B. Sutton, D. Noll, J. Fessler. Fast, iterative image reconstruction for MRI in the presence of field inhomogeneities. IEEE Trans Med Imaging 2003;22, no. 2, pp. 178–188.

57. P. Ladefoged, I. Maddieson. The sounds of the world's languages. Blackwell 1996.

58. N. Warner, B. Tucker. Phonetic variability of stops and flaps in spontaneous and careful speech. J Acoust Soc Am 2011;130:1606–1617.

59. R. Port, F. Mitleb, M. O'Dell. Neutralization of obstruent voicing in German is incomplete. J Acoust Soc Am 1981;70(Suppl.1):13.

60. W. Herd, A. Jongman, J. Sereno. An acoustic and perceptual analysis of /t/ and /d/ flaps in American English. J Phon 2010;38:504–516.

61. A. Braver. Imperceptible incomplete neutralization: Production, non-identifiability, and non-discriminability in American English flapping. Lingua 2014;152:24–44.

62. M. Barlaz, M. Fu, Z.-P. Liang, R. Shosted, B. Sutton. The emergence of nasal velar codas in Brazilian Portuguese: An rt-MRI Study. In Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 2015, pp. 2660–2664.

63. M. Fu, M. Barlaz, R. Shosted, Z.-P. Liang, B. Sutton. High-resolution dynamic speech imaging with deformation estimation. In Proceedings of IEEE International Symposium on Biomedical Imaging, Milan, Italy, 2015, PP. 2281–2285.

64. C. Carignan, R. Shosted, M. Fu, Z.-P. Liang, B. Sutton. A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French. J Phon 2015;50:34–51.

65. N. Umeda. Vowel duration in American English. J Acoust Soc Am 1975;58:434–445.

66. T. Wech, D. Stab, J. C. Budich, A. Fischer, J. Tran.-Gia, D. Hahn, H. Kostler. Resolution evaluation of MR images reconstructed by iterative thresholding algorithms for compressed sensing. Med Phys 2012;39:4328–4338.

67. P. Stoica, S. Yngve. Model-order selection: a review of information criterion rules. IEEE Signal Process Mag 2004;21:36–47.

68. K. Bumham, D. Anderson. Model selection and multimodel inference: a practical information-theoretic approach. Springer 2002.

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

70. S. Ramani, Z. Liu, J. Rosen, J. Nielsen, J. A. Fessler. Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods. IEEE Trans Image Process 2012;21:3659–3672.

71. J. Gai, N. Obeid, J. Holtrop, X.-L. Wu, F. Lam, M. Fu, J. Haldar, W. Hwu, Z.-P. Liang, B. P. Sutton. More IMPATIENT: A gridding-accelerated Toeplitz-based strategy for non-Cartesian high-resolution 3D MRI on GPUs. J Parallel Distrib Comput 2013;73:686–697.

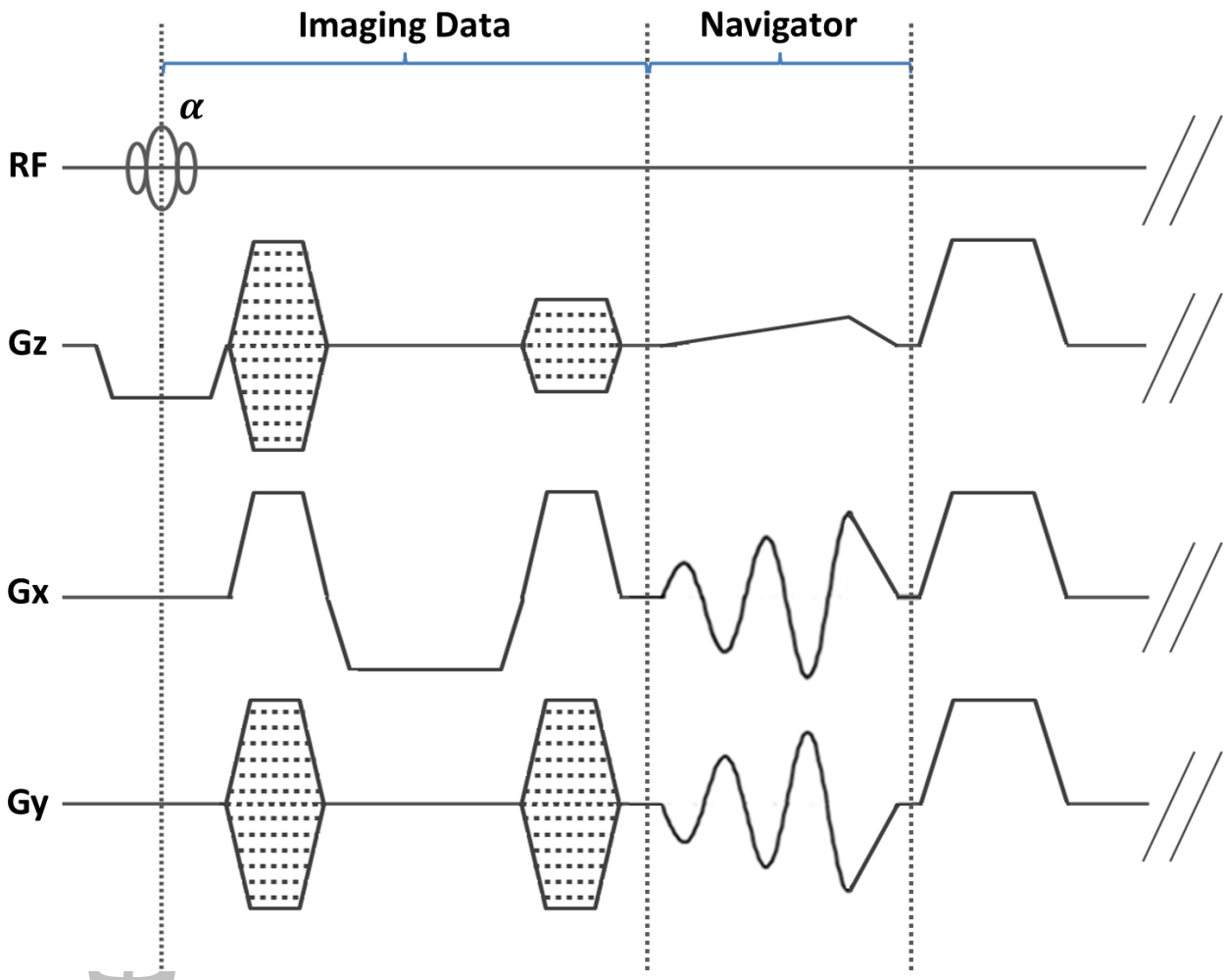


Figure 1: A simplified pulse sequence diagram for the proposed "self-navigated" data acquisition strategy. Within a single TR, the imaging data set is acquired at early echo time using a Cartesian trajectory with random phase encoding, while the navigator data set is acquired at later echo time using a cone trajectory.

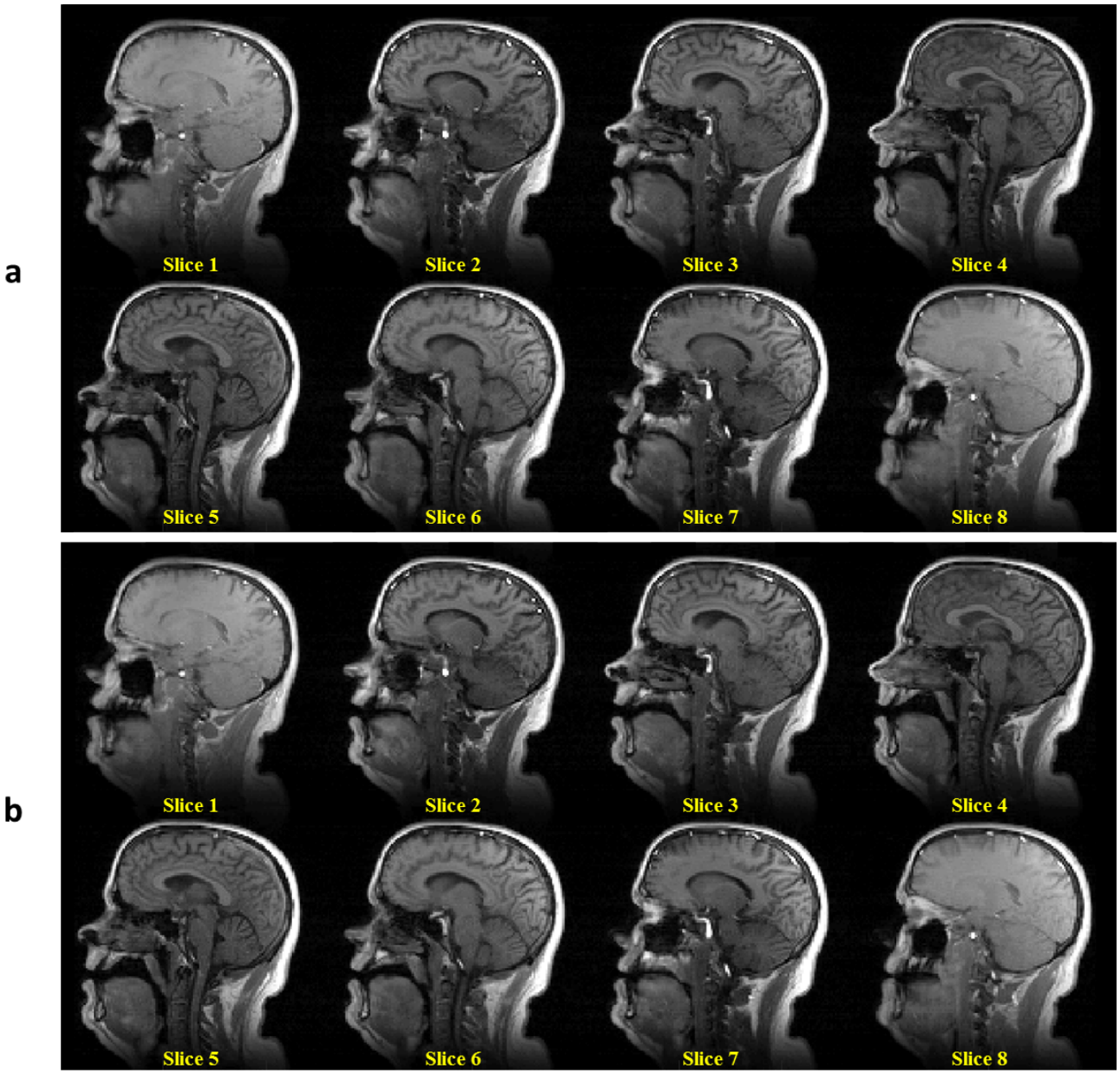


Figure 2: Mid-sagittal articulator gestures of the upper vocal tract during the production of /loo/-/lee/-/la/-/za/-/na/-/za/ sounds. Articulator gestures of /l/ sound in the /la/ syllable is shown in (a). Articulator gestures of /a/ sound in the /la/ syllable is shown in (b).

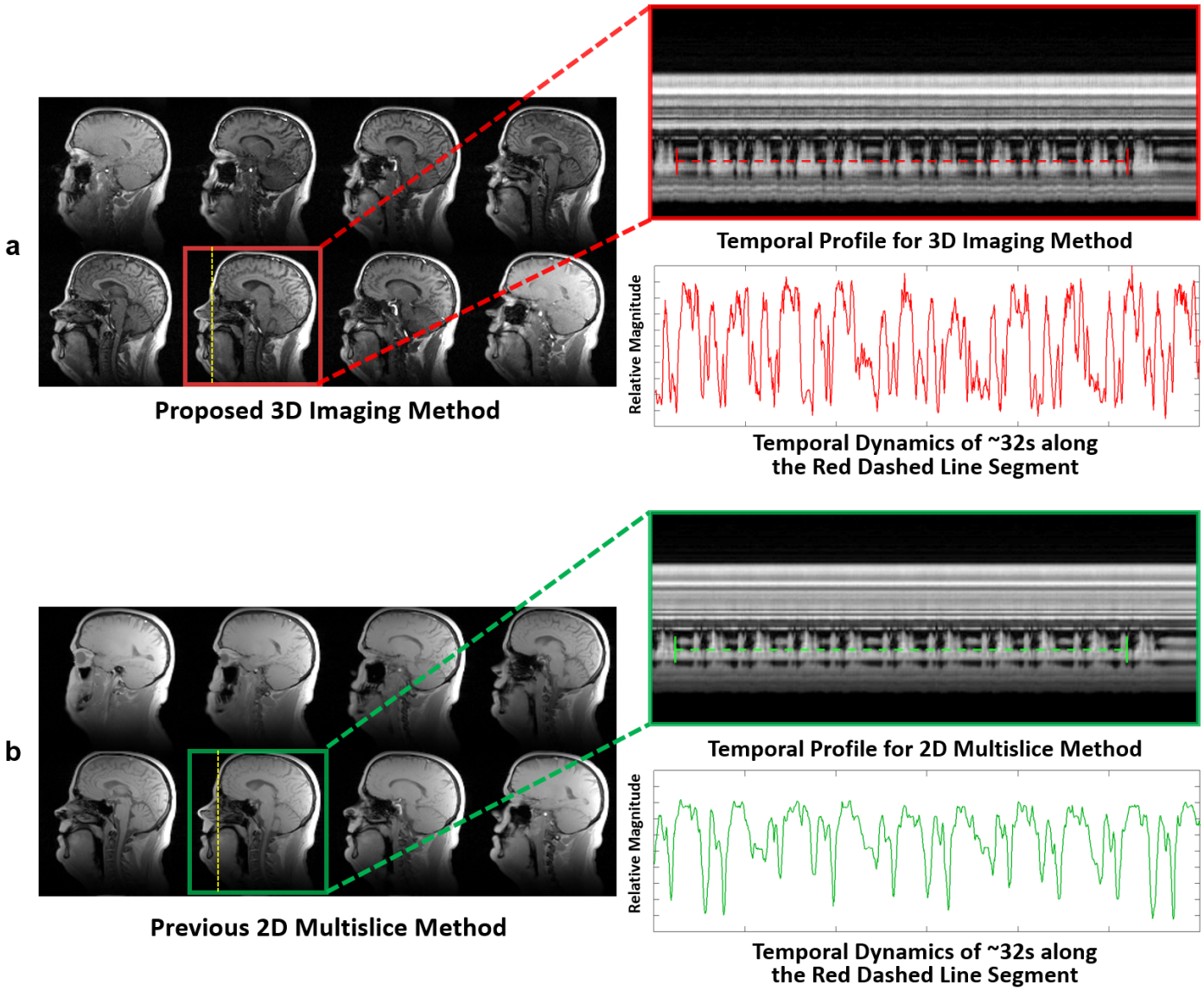


Figure 3: Comparison of temporal dynamics from the proposed method and from a previous 2D multi-slice method: (a) the temporal profile along a strip across the tongue tip from a 3D reconstruction; the associated temporal dynamics along a red dashed line segment; (b) the temporal profile along a strip across the tongue tip from a 2D multi-slice reconstruction; the associated temporal dynamics along a green dashed line segment. Improved temporal dynamics and shaper temporal transitions are seen with the 3D reconstruction.

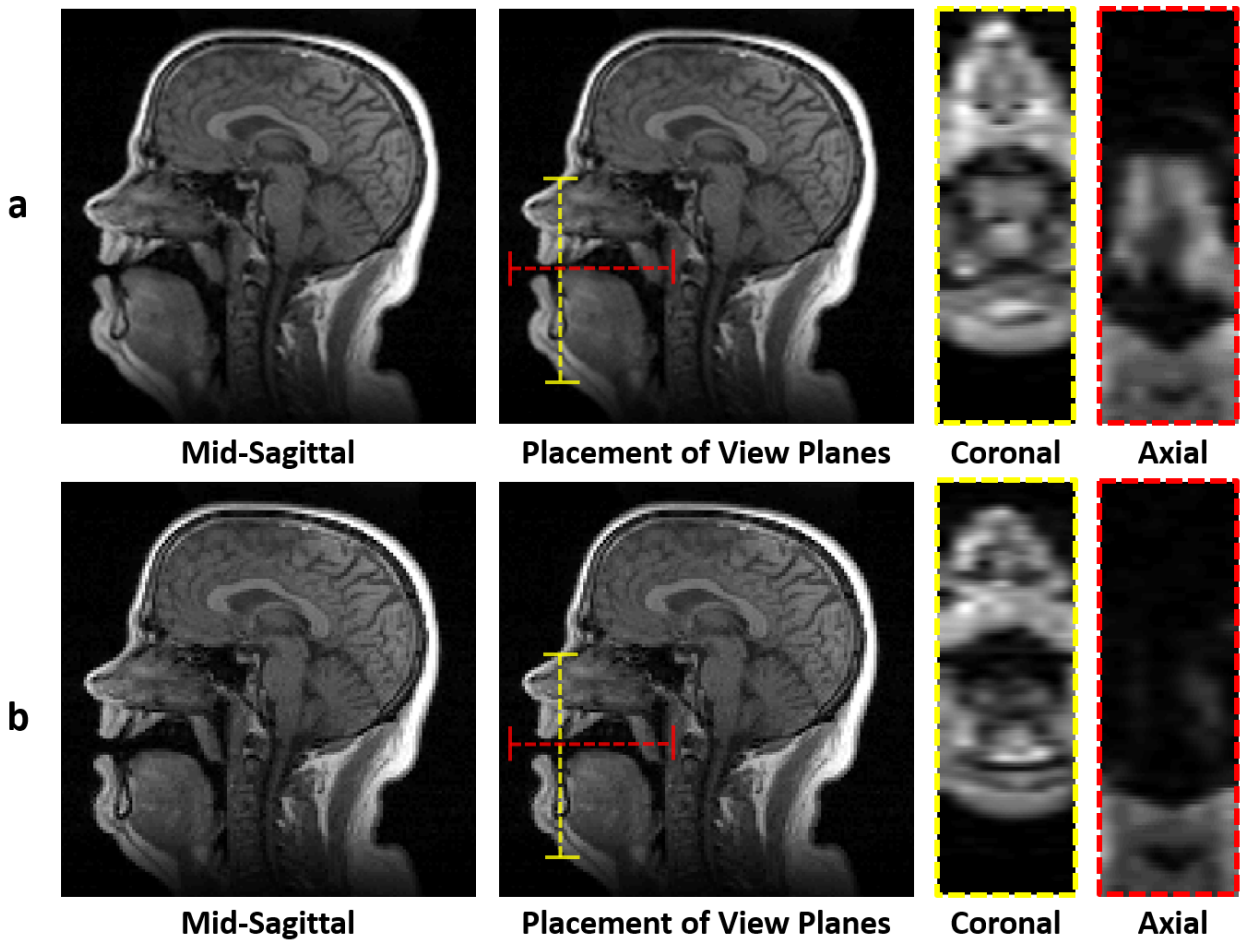


Figure 4: Visualization of sound production at mid-sagittal, coronal and axial planes: (a) the production of /l/ in the /la/ sound; (b) the production of /a/ in the /la/ sound. The placement of coronal and axial planes are indicated with yellow and red colors, respectively.

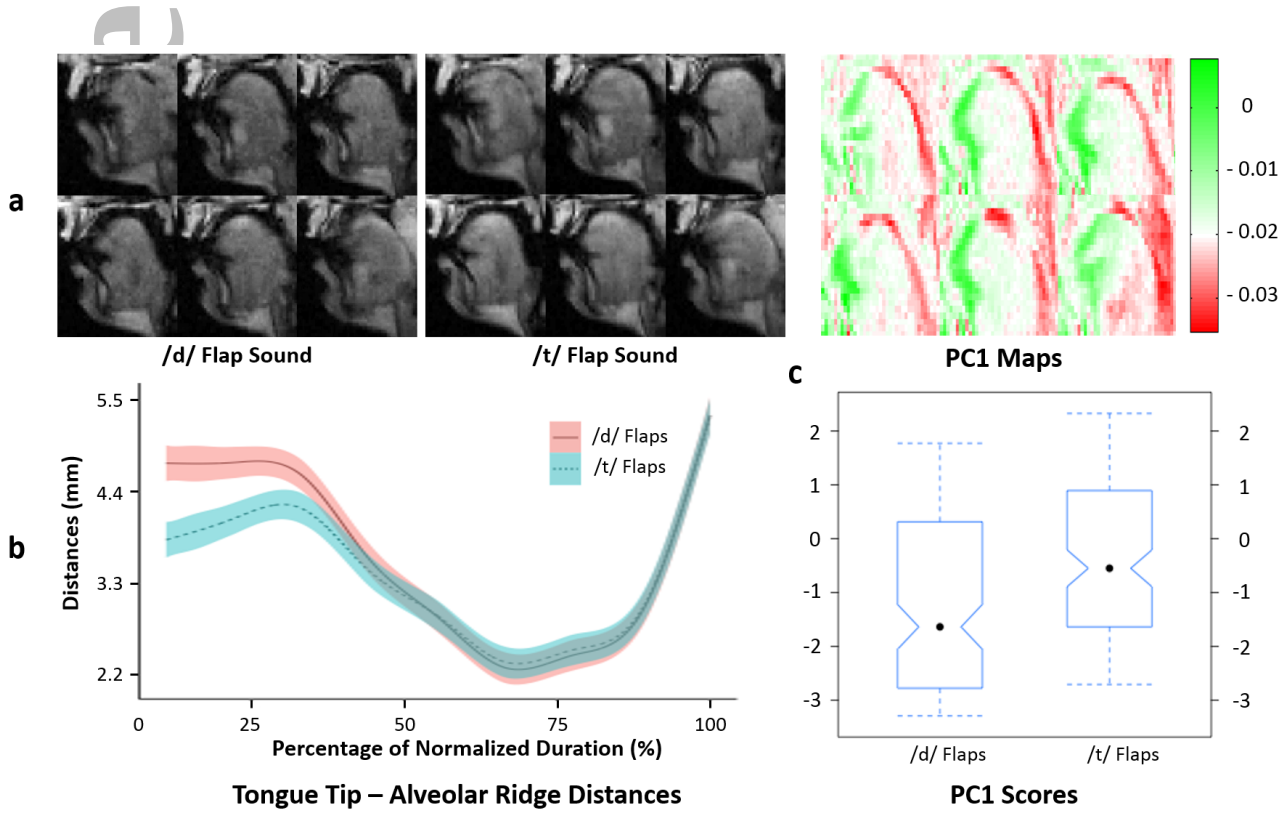


Figure 5: Mid-sagittal reconstructions and phonetics analyses of the production of American English flaps: (a) representative tongue gestures at the production of the /d/ and /t/ flaps; (b) averaged distances between the tongue tip and the alveolar ridge for the /d/ and /t/ flaps over a normalized duration; (c) top row shows spatial maps associated with the first principal component (PC1) for the flaps; bottom row shows statistical results of higher PC1 score for /t/ flap than /d/ flap sound (suggesting higher level of tongue apex protrusion and tongue blade elevation).

Figure 1:

A simplified pulse sequence diagram for the proposed "self-navigated" data acquisition strategy. Within a single TR, the imaging data set is acquired at early echo time using a Cartesian trajectory with random phase encoding, while the navigator data set is acquired at later echo time using a cone trajectory.

Figure 2:

Mid-sagittal articulator gestures of the upper vocal tract during the production of /loo/-/lee/-/la/-/za/-/na/-/za/ sounds. Articulator gestures of /l/ sound in the /la/ syllable is shown in (a). Articulator gestures of /a/ sound in the /la/ syllable is shown in (b).

Figure 3:

Comparison of temporal dynamics from the proposed method and from a previous 2D multi-slice method: (a) the temporal profile along a strip across the tongue tip from a 3D reconstruction; the associated temporal dynamics along a red dashed line segment; (b) the temporal profile along a strip across the tongue tip from a 2D multi-slice reconstruction; the associated temporal dynamics along a green dashed line segment. Improved temporal dynamics and shaper temporal transitions are seen with the 3D reconstruction.

Figure 4:

Visualization of sound production at mid-sagittal, coronal and axial planes: (a) the production of /l/ in the /la/ sound; (b) the production of /a/ in the /la/ sound. The placement of coronal and axial planes are indicated with yellow and red colors, respectively.

Figure 5:

Mid-sagittal reconstructions and phonetics analyses of the production of American English flaps: (a) representative tongue gestures at the production of the /d/ and /t/ flaps; (b) averaged distances between the tongue tip and the alveolar ridge for the /d/ and /t/ flaps over a normalized duration; (c) top row shows spatial maps associated with the first principal component (PC1) for the flaps; bottom row shows statistical results of higher PC1 score for /t/ flap than /d/ flap sound (suggesting higher level of tongue apex protrusion and tongue blade elevation).

Supporting Figure S1:

Comparison of spatial details and temporal dynamics of: a) the numerical phantom; and b) recon-struction of simulated data. Temporal profiles ($\sim$40s) along a strip at the yellow line across the tongue tip on the 4th mid-sagittal plane are compared for the phantom (red) and the reconstruction (green). Absolute difference of the temporal profiles (scaled by a factor of 2) is also shown.

Supporting Figure S2:

Mid-sagittal reconstructions and phonetics analyses of the production of American English flaps: (a) representative tongue gestures at the production of the /d/ and /t/ flaps; (b) averaged distances between the tongue tip and the alveolar ridge for the /d/ and /t/ flaps over a normalized duration; (c) top row shows spatial maps associated with the first principal component (PC1) for the flaps; bottom row shows statistical results of higher PC1 score for /t/ flap than /d/ flap sound (suggesting higher level of tongue apex protrusion and tongue blade elevation).

Supplementary Document

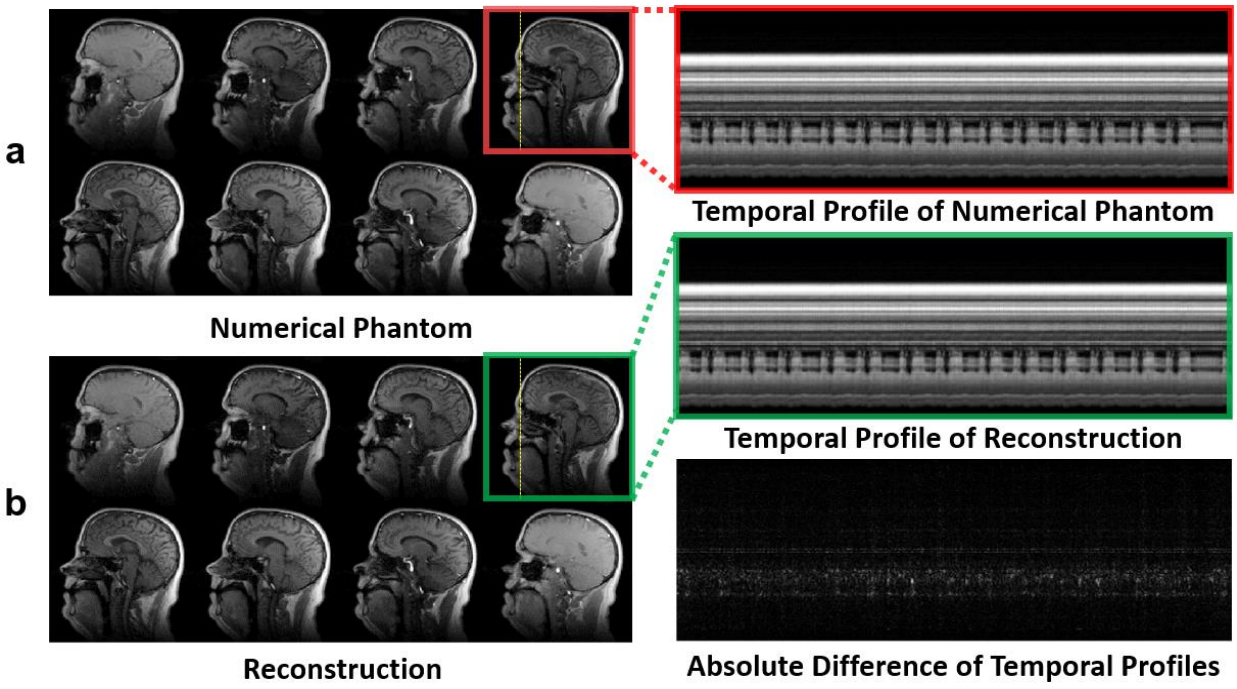
In this supplementary document, we provide results from the numerical simulations as validation to the performance of the proposed method. The paper has presented reconstructions from in vivo acquisitions of real speech tasks, demonstrating that our method is capable of resolving motion on the order of the nominal frame rate and with good spatial resolution. However, given that no other dynamic speech imaging method is capable of spatially and temporally resolving the motions as demonstrated in the paper, a validation experiment is in general challenging since no suitable existing data or model is available.

Further, we provide simulation results in this document to demonstrate that the nominal frame rate can be achieved with the proposed method. Although the simulations yield good empirical results, the reader should keep in mind that our method is a nonlinear imaging method and its exact performance results on a particular data set may depend on the quality and characteristics of that specific data set.

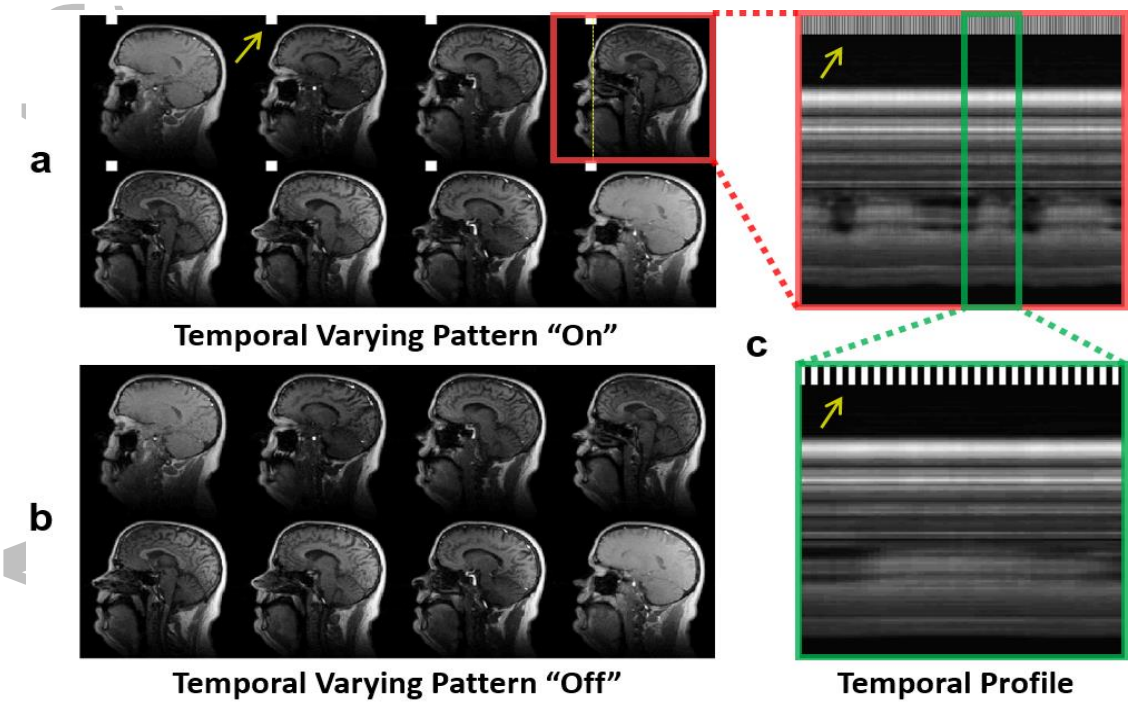
SIMULATION RESULTS

Supporting Figure S1 shows representative simulation results from the generic phantom. The reconstructed spatiotemporal dynamics are consistent with those in the phantom and has a reconstruction error of 0.0472. Specifically, Supporting Figure S1a and S2b show tongue gestures from the same time frame for the phantom and the reconstruction, respectively. Articulatory gestures of the phantom are well captured with great spatial details in the reconstruction. To further evaluate the quality of the reconstructed dynamics, temporal profiles for both the numerical phantom (red box) and the reconstruction (green box) are compared. As seen, the reconstruction faithfully represents the temporal dynamics of the phantom without significant temporal blurring, which is also evident by examining the absolute difference in the temporal profiles. Supporting Figure S1 demonstrates that our method is capable of capturing high-quality spatiotemporal details for dynamic speech MRI.

Supporting Figure S2 shows representative results from the modified phantom. In particular, Supporting Figure S2a shows a reconstructed time frame with the added “on” pattern – the added bright cube positions above the subject’s forehead and is indicated by a yellow arrow. As contrast, Supporting Figure S2b shows an ensuing time frame that has the “off” pattern. Supporting Figure S2c shows a reconstructed temporal profile along a strip across the subject’s tongue tip. As seen, the temporal varying pattern and the dynamics of articulatory motion are both well captured by the reconstruction. Even in the zoom-in view of the reconstructed temporal profile, the level of temporal blurring is small. Supporting Figure S2 demonstrates that the proposed method is capable of capturing temporal events that require a frame rate of 166 fps. It may also be feasible to validate our method’s capability to capture temporal events at higher frame rates if proper numerical phantoms were designed. The construction and evaluation on such phantoms will be performed on future research.



Supporting Figure S1: Comparison of spatial details and temporal dynamics of: a) the numerical phantom; and b) reconstruction of simulated data. Temporal profiles (~40s) along a strip at the yellow line across the tongue tip on the 4th mid-sagittal plane are compared for the phantom (red) and the reconstruction (green). Absolute difference of the temporal profiles (scaled by a factor of 2) is also shown.



Supporting Figure S2: Characterization of the nominal frame rate using a modified numerical phantom with a flashing temporal pattern every other frame: a) a time frame that has a bright cube that is positioned above the subject's forehead; b) a time frame that has no added bright cube; c) the temporal profile along the yellow line across the tongue tip on the 4th mid-sagittal plane (top) and its zoom-in view (bottom). It is obvious from the zoom-in view that the temporal blurring in the reconstructed temporal profile is small.