HIGH-RESOLUTION FULL-VOCAL-TRACT DYNAMIC SPEECH MAGNETIC RESONANCE IMAGING

BY

MAOJING FU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Zhi-Pei Liang, Chair
Associate Professor Brad P. Sutton
Professor Minh N. Do
Associate Professor Ryan K. Shosted
Assistant Professor Jonghye Woo, Massachusetts General Hospital and Harvard University

ABSTRACT

Dynamic magnetic resonance imaging (MRI) holds great promise for speech-related studies because of its potential to investigate velopharyngeal motion and physiological properties jointly in real time. However, many applications of dynamic speech MRI are limited by the technical tradeoffs in imaging speed, spatial coverage, spatial resolution and clinical interpretation. In particular, high-resolution dynamic speech MRI with full-vocal-tract coverage and phonetically meaningful interpretation remains a challenging goal for many speech researchers. This dissertation develops novel model-based dynamic speech MRI approaches to enable high-resolution, full-vocal-tract 3D dynamic speech MRI with quantitative characterization of the articulatory motion.

Our approaches include technical developments in imaging models, data acquisition strategies and image reconstruction methods: (a) high-spatiotemporal-resolution speech MRI from sparsely sampled data is achieved by employing a low-rank imaging model that leverages the spatiotemporal correlations in dynamic speech motion; (b) a self-navigated sampling strategy is developed and employed to acquire spatiotemporal data at high imaging speed, which collects high-nominal-frame-rate cone navigators and randomized Cartesian imaging data within a single TR; (c) quantitative interpretation of speech motion is enabled by introducing a deformation-based sparsity constraint that not only improves image reconstruction quality but also analyzes articulatory motion by a high-resolution deformation field; and (d) accurate assessment of subject-specific motion as opposed to generic motion pattern is realized by using a low-rank plus sparse imaging model jointly with a technique to construct high-quality spatiotemporal atlas. Regional sparse modeling is further introduced to assist effective motion analysis in the regions of interest.

Our approaches are evaluated through both simulations on numerical phantoms and *in vivo* validation experiments across multiple subject groups. Both simulation and experimental results allow visualization of articulatory dynamics with a frame rate of 166 frames per second, a spatial resolution of $2.2 \times 2.2 \times 5.0$ mm^3, and a spatial coverage of $280 \times 280 \times 40$ mm^3 covering the entire upper vocal tract across 8 mid-sagittal slices. Deformation fields yielded from our approaches share an identical spatiotemporal resolution that characterizes accurate soft-tissue motion. With a high-quality atlas, the low-rank and the sparse components are reconstructed to reveal both subject-specific motion and generic speech motion across a specific subject group.

The effectiveness of our approaches is demonstrated through practical phonetics investigations that include (a) integrative imaging and acoustics analysis of velopharyngeal closure; (b) understanding the formation and variation in a variety of languages, American English, North Metropolitan French, Brazilian Portuguese and Levantine Arabic; and (c) analyzing motion variability of a particular subject with respect to a specific subject group. The capabilities of our method have the potential for precise assessment of the oropharyngeal dynamics and comprehensive evaluation of speech motion.

To my family, for their unfailing love and support

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to my advisor, Professor Zhi-Pei Liang, for all the support, vision and protection that he has given me over the past six years. The advice and suggestions from him are always insightful and have reshaped my way of thinking, studying and living.

I owe a great debt of gratitude to Professor Brad Sutton, who offered me the opportunities to explore dynamic speech MRI. The weekly one-on-one meeting with him is always invigorating, and I am very grateful for his guidance and patience throughout my graduate study.

I also owe thanks to the other members of my preliminary exam committee: Professors Minh Do, Ryan Shosted and Jonghye Woo, whose vision, comments and suggestions have defined the path of my dissertation research and have shaped the scope of this dissertation.

I am lucky to have many friends, lab mates and collaborators across multiple research groups at the University of Illinois: Bo Zhao, Fuquan Ren, Zhenghua Wu, Guangpu Shao, Xi Peng, Huiqian Du, Xiaobo Qu, Chao Ma, Fan Lam, Qiang Ning, Yudu Li, Rong Guo, Dan Li and Shengwen Guo from Professor Liang's group; Cheng Ouyang, David Ho, Joe Holtrop, Jiading Gai, Curtis Johnson, Giang-Chau Ngo, Aaron Anderson, Nate Wetter and Genevieve Labelle from Professor Sutton's group; Panying Rong, Li-Hsin Ning, Marissa Barlaz, Zainab Hermes, Christopher Carignan, Nicole Wong and Di Wu from Professor Shosted's group; Xiaochun Lai, Tan Nguyen; and Professors Ling-Jian Meng, Jamie Perry, Aaron Johnson and David Kuehn. Their conversation and collaboration made my dissertation research a joy to experience.

I am truly grateful towards Jennifer Carlson, Laurie Fisher and Youakim Pascal for offering me the TA positions to help me survive the extreme difficulties and financial struggles in the past three years. I am also blessed to have worked with the nicest colleagues and the smartest students for the courses that I TA'ed for. It has always been a pleasure to work with my fellow TAs and course instructors: Ali Yekkehkhany, Weihao Gao, Cheng Chen, Ge Yu, Fardad Raisali, Dimitrios

Katselis, Professors Juan Alvarez, Bruce Hajek, Mark Hasegawa-Johnson, Yi Lu and Yihong Wu.

I am extremely fortunate to have had internship experience from both Intuitive Surgical Incorporated at Sunnyvale and Massachusetts General Hospital at Boston. I worked on a variety of exciting projects and broadened my horizon on machine learning, image segmentation and registration and object-oriented programming. It has always been enjoyable to work with them: Jun Ma, Bai Wang, Tao Zhao, Shiyang Chen, Zhe (Sophie) Li, Wenlong Lu, Amir Refaee, Samuel Huang, Jesse Mao, Nuo (Jerry) Xu, Pechin Lo, Chi Zhang, Yixin Gao and Ray Chen from the Walle team at ISI; Jonghye Woo, Fangxu Xing, Yongjin Chang and Georges El Fahkri from the Gordon imaging center at MGH.

I would like to acknowledge the Computational Science and Engineering program and the Graduate College for their support through the CSE fellowship and the Dissertation Travel Grant. Also, I am very grateful to Professors Eric Shaffer and Ryan Shosted for providing me unlimited access to and countless computational hours on their servers, from which the majority of results in this dissertation were generated.

I would like to thank my family for their unfailing love, support and understanding. They are always my oasis of smile, strength and confidence. Especially I would like to thank my wife, Meng, for her support and encouragement during all times of difficulty. This dissertation is specifically dedicated to my family.

This dissertation would not have been possible without the help and support from many other friends. Although I cannot list all of them, I would like to express my thanks to everyone who helped me get to this point.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

1D	One dimensional
2D	Two dimensional
3D	Three dimensional
4D	Four dimensional
ANOVA	Analysis of variance
ANTs	Advanced normalization tools
BP	Basis pursuit
CS	Compressed (Compressive) sensing
CT	Computerized Tomography
DFT	Discrete Fourier transform
DL	Dictionary learning
EMA	Electromagnetic articulographs
F1	The first formant
F2	The second formant
F3	The third formant
FFT	Fast Fourier transform
FID	Free induction decay
FLASH	Fast low-angle shot
FOCUSS	Focal underdetermined system solver
FOV	Field of view

FPS	Frames per second
fMRI	Functional MRI
GPU	Graphics processing unit
GRAPPA	Generalized autocalibrating partially parallel acquisition
LDDMM	Large deformation diffeomorphic metric mapping algorithm
LPSF	Localized point spread function
MP	Matching pursuit
MR	Magnetic resonance
MRI	Magnetic resonance imaging
NMR	Nuclear magnetic resonance
NUFFT	Non-uniform fast Fourier transform
OMP	Orthogonal matching pursuit
PC	Principal component
PC1	The first principal component
PC2	The second principal component
PS	Partial separability
PSF	Point spread function
RF	Radio frequency
ROI	Region of interest
SENSE	Sensitivity encoding
SNR	Signal-to-noise ratio
SPSF	Simulated point spread function
SSFP	Steady state free precession
SVD	Singular value decomposition
SyN	Symmetric image Normalization
TE	Echo time
TR	Repetition time

TV	Total variation
T_1	Spin-lattice relaxation time
T_2	Spin-spin relaxation time
VFS	Video fluoroscopy

LIST OF SYMBOLS

$\mathbf{C}$	The Casorati matrix formed from spatiotemporal data.
$\mathbf{d}$	The sparsely sampled spatiotemporal data.
$\mathbf{E}$	The imaging operator consisting of sensitivity encoding and sparse sampling.
$\mathbf{F}$	The Fourier encoding matrix.
$d(\mathbf{k}, t)$	The sparsely sampled spatiotemporal data from $(\mathbf{k}, t)$-space.
l	The index of model order.
L	The model order.
M	The number of temporal frames.
N	The number of spatial encodings.
$\mathbf{I}$	The desired spatiotemporal image.
$I(\mathbf{r}, t)$	The desired spatiotemporal image of speech motion.
$\mathbf{k}$	The spatial locations in the Fourier space.
t	The time instances during which the spatiotemporal samples are acquired.
T	Tesla.
$\mathbf{U}$	The spatial subspace of the spatiotemporal image.
$\mathbf{V}$	The temporal subspace of the spatiotemporal image.
$\mathbf{r}$	The spatial locations in the image space.
Q	The number of receiver coils.
$\mathbf{X}$	The articulated nasal vowels.
$\eta(\mathbf{k}, t)$	The measurement noise during spatiotemporal data acquisition.

λ	The regularization parameter.
λ_1	The first regularization parameter.
λ_2	The second regularization parameter.
$\mathbf{\Omega}$	The sparse sampling operator corresponding to imaging data acquisition.
ϕ_q	The qth interpolation kernel.
$\psi_l(\mathbf{r})$	The lth spatial basis function.
$\phi_l(t)$	The lth temporal basis function.

CHAPTER 1

INTRODUCTION

1.1 Problem statement

In a typical dynamic speech MRI experiment, the acquired Fourier encoded magnetic resonance (MR) data, $d(\mathbf{k}, t)$, from the $(\mathbf{k}, t)$-space can be expressed as

$$d(\mathbf{k}, t) = \int I(\mathbf{r}, t) e^{-i2\pi\mathbf{k}\cdot\mathbf{r}} d\mathbf{r} + \eta(\mathbf{k}, t), \tag{1.1}$$

where $I(\mathbf{r}, t)$ is the desired spatiotemporal images of speech motion and $\eta(\mathbf{k}, t)$ is the measurement noise. Unlike in conventional MRI applications, dense sampling in the $(\mathbf{k}, t)$-space is challenging during dynamic speech MRI data acquisition: on the one hand, the articulators in the vocal tract move rapidly during speech production and require high sampling speed to capture; on the other hand, the sampling speed in dynamic speech MRI is constrained due to physical limits of the data acquisition hardware and the physiological concerns on peripheral stimulation on human. The situation is more difficult when it entangles with the "curse of dimensionality" – the required amount of measured data grows exponentially with the increased resolution requirements of the imaging problem according to the Nyquist-Shannon sampling theorem. This dilemma renders high-resolution full-vocal-tract dynamic speech MRI a challenging task and has led to various technical trade-offs between the imaging speed, spatial resolution, temporal resolution, spatial converge, signal-to-noise ratio and clinical interpretation of speech motion.

Aiming at addressing the above challenges, this dissertation centers on developing model-based approaches to achieve high-resolution full-vocal-tract dynamic speech MRI with phonetically meaningful interpretation of speech motion. The hypothesis is that high-quality dynamic speech MRI is achievable by leveraging the latest developments in spatiotemporal imaging modeling, fast data acquisition strategies and advanced reconstruction methods. In particular, this dissertation proposes to include (a) a low-rank imaging model that leverages spatiotemporal correlation in speech motion to recover high-quality dynamic images; (b) an accelerated model-based data acquisition strategy to capture fast articulatory dynamics; (c) an image reconstruction method that combines a low-rank image model with deformation estimation to allow high-quality reconstruction and accurate analysis of articulatory motion; (d) an advanced imaging model that enables effective decomposition of subject-specific motion patterns as opposed to the generic speech motion pattern; and (e) evaluation of the practical utility of the above technical elements in terms of systematic simulation, *in vivo* experiments and phonetics investigation. These developments allow dynamic speech MRI to serve as a powerful imaging and diagnostic tool for a variety of speech-related research and applications.

1.2 Motivation

Dynamic magnetic resonance imaging holds great promise for real-time visualization of the vocal articulators and the associated vocal muscles. Compared with existing dynamic imaging techniques, dynamic speech MRI has multiple technical advantages: dynamic speech MRI allows visualizing the anatomical structures along arbitrary imaging planes without invasive procedures, as compared with endoscopy and ultrasound; and dynamic speech MRI provides excellent soft-tissue contrast without using ionizing radiation, as compared with video fluoroscopy and positron emission tomography. Recent development in MRI hardware, sequence developments and advanced imaging models have effectively enhanced the technical capabilities of MRI and enabled significant progress in various dynamic imaging applications.

Dynamic speech MRI has recently started to demonstrate its usefulness both for scientific research and clinical studies in speech. The utilization of dynamic speech MRI to capture structural and functional changes of the vocal tract has led to a broad spectrum of applications. These applications include studying complex soft-tissue geometries in the upper vocal tract [1–3]; detecting structural defects, functional disorder and motor dysfunctions [4–7]; and studying the evolution and variations of important languages [8–14]. Although dynamic speech MRI has found widespread use in a variety of scenarios, the state-of-the-art dynamic imaging methods still suffer from low spatiotemporal resolution, limited spatial coverage and poor clinical interpretation. Developing advanced dynamic imaging techniques that can simultaneously allow high-resolution visualization of speech motion, broad spatial coverage covering the entire upper vocal tract and effective phonetics interpretation of the articulatory motion is critical towards deeper understanding of basic phonetics studies and better clinical diagnostics.

An effective dynamic speech imaging should at least provide six desirable properties: (1) non-invasive procedures to visualize articulatory dynamics without interrupting natural speech production; (2) good soft-tissue contrast to clearly visualize the vocal articulators and the associated driving muscles in the vocal tract; (3) sufficient spatial resolution differentiate the gestures of small-sized articulators; (4) high imaging speed to capture rapid temporal movements of the vocal tract articulators: high imaging speed is a critical property for dynamic speech MRI applications because brief temporal events may carry important information that reflects the structural and functional changes of the vocal tract [15]; (5) quantitative characterization of the reconstructed articulatory motion with respect to a ground truth image; and (6) effective separation and analysis of subject-specific motion patterns as opposed to the subject group. Although the past decade has seen significant improvement separately in each of the above properties, in general it remains challenging to maximize all properties at the same time. Also, the associated clinical tools to effectively interpret the reconstructed images have been generally lacking. These difficulties have prevented dynamic speech MRI from greater scientific impact and clinical influence.

1.3 Overview of contributions

This dissertation centers on the developments of dynamic speech MRI methods that simultaneously provide high-speed, high-resolution, full-vocal-tract-coverage imaging methods of speech dynamics with quantitative characterization of speech motion both for a single subject and across the entire subject group. The main contributions in this dissertation are summarized as follows:

- An imaging model has been employed to represent the spatiotemporal structure of dynamic speech images and recover high-quality spatiotemporal speech dynamics from highly undersampled measured data. An image reconstruction formulation has been proposed to leverage the signal properties of speech and integrate both the imaging model and sparse modeling. Efficient algorithms-based on half-quadratic regularization have been implemented and optimized. Performance has been systematically evaluated through simulation, *in vivo* experiments and phonetic investigations.

- Model-based data acquisition strategies have been developed to capture fast transitions of dynamic articulatory motion. Developments in data acquisition strategies include the design and implementation of fast, low-angle shot (FLASH) pulse sequences; design and optimization of novel trajectories both for 2D and 3D dynamic speech MRI applications; design and implementation of sparse sampling schemes appropriate for the imaging model; and the development of imaging and operation protocols for practical phonetics studies.

- A novel dynamic speech MRI method based on deformation analysis has been proposed to not only enhance the quality of image reconstruction, but also to allow quantitative analysis of the articulatory motion by high-resolution deformation fields. These two goals are simultaneously achieved by integrating a deformation-based sparsity constraint into the image reconstruction formulation. This method has been further developed to allow automatic tracking or segmentation of articulatory gestures for various phonetics applications. Performance of this method has been evaluated by simulation, *in vivo* experiments and phonetics investigations.

- To allow characterization of the articulatory gestures across multiple subjects, a novel dynamic speech MRI method has been proposed to incorporate a spatiotemporal atlas into a low-rank plus sparse imaging model. A spatiotemporal atlas provides strong prior information for the image reconstruction problem and the low-rank plus sparse model allows effective decomposition of the generic and subject-specific speech motion. Regional sparse modeling has also been integrated to improve the performance of our method in targeted regions. The resulting imaging method not only facilitates the image reconstruction through a spatiotemporal atlas, but also enables effective interpretation of speech motion separately for the generic pattern and the subject-specific motion. The performance of this method has been evaluated by simulation, *in vivo* experiments and statistical phonetic analysis.

1.4 Organization of dissertation

This dissertation is organized as follows:

- Chapter 2 reviews the necessary background information for the ensuing chapters, covering historical development of speech imaging techniques, a detailed review of important dynamic speech MRI methods from signal processing approaches and data acquisition approaches, and a detailed review of the clinical and phonetics applications.

- Chapter 3 presents our method that enables high-resolution, full-vocal-tract 3D dynamic speech MRI. It presents technical developments in the imaging model, accelerated data acquisition strategy, algorithm analysis, simulation studies, results from validation experiments and various examples from phonetics investigations.

- Chapter 4 introduces our method that integrates deformation estimation into dynamic speech MRI to simultaneously improve reconstruction quality and allow accurate motion analysis. It consists of technical developments in problem formulation, solution algorithm, simulation studies and validation experiments. The practical utility of our method has been validated through multiple phonetics investigations.

- Chapter 5 presents our work on a spatiotemporal atlas-based dynamic speech MRI method. It presents technical developments on a low-rank plus sparse imaging model, a method of constructing spatiotemporal atlas images of both high image quality and high throughput and an image reconstruction formulation that integrates the imaging model with a spatiotemporal atlas. Experimental results and phonetics investigations are presented to demonstrate the practical value of our method for clinical and scientific applications.

- Chapter 7 provides the conclusion to this dissertation.

CHAPTER 2

BACKGROUND

Chapter 2 aims to provide a detailed review of dynamic speech imaging techniques, covering its technical development over the past three decades, signal-processing-based approaches to dynamic speech MRI, data-acquisition-based approaches to dynamic speech MRI and the representative clinical applications of dynamic speech MRI. The detailed review provided in this chapter serves as important background information for the ensuing chapters. Specifically, Section 2.1 provides an overview of the development of dynamic speech imaging methods over the past three decades, covering conventional dynamic speech imaging methods and the evolution of MR-based dynamic speech imaging methods. Section 2.2 presents the signal-processing-based approaches to dynamic speech MRI, covering partial separability model-based imaging, parallel imaging and compressed sensing-based imaging. Section 2.3 presents the data-acquisition-based approaches to dynamic speech MRI, covering spatiotemporal sampling patterns, fast pulse sequences, advanced data acquisition hardwares and a range of scanning protocols. The last section in this chapter presents representative dynamic speech MRI applications, covering both clinical and non-clinical applications.

2.1 Development of speech imaging

2.1.1 Conventional dynamic speech imaging

Speech is a vital bodily function directly related to the quality of life. Given its importance, effective techniques have been developed to monitor and measure the dynamics in the vocal tract. These techniques can be roughly divided into two categories, non-imaging-based and imaging-based techniques. This dissertation provides an overview of the imaging-based techniques, while the interested readers can refer to [16] for a review of the non-imaging-based techniques.

The imaging-based techniques are far superior to their non-imaging-based counterparts in providing directly visualization of both localized articulatory behaviors and global articulatory movements. This visualization capability is of great importance in speech-related studies because articulatory motion consists of bulk motion from hard-tissue structures (the hard palate, the jaw and the pharyngeal wall), as well as localized motion from soft-tissue structures (the tongue, the velum and the epiglottis). Having the capability to simultaneously visualize both patterns of motion (which is not available from non-imaging-based techniques) helps understand how these structures behave and change individually and interact with others collectively [17]. Four imaging methods have traditionally dominated dynamic speech imaging methods: video endoscopy, ultrasound, video fluoroscopy (VFS) and computerized tomography (CT). This dissertation gives an overview of each of these imaging methods.

Video endoscopy

Video endoscopy is by its nature an optical imaging modality that allows visualization of the interior of the human body by using a small-sized camera. A flexible endoscope, usually equipped with its own light source at the front end, is inserted through the oral and nasal cavities to reach the region of interest. Structures of the surface of the inner cavities are imaged, and the acquired images are transferred back through an optical fiber to radiologists for visual evaluation [18]. Figure 2.1 provides an example of a typical endoscopic image of the surface of the pharynx obtained by using video endoscopy.

Video endoscopy provides straightforward detection of the surface lesions in the vocal tract. In particular, the abnormal tissues can be easily identified by comparing their shapes and those of the normal tissues [19]. Comparison results can be more targeted by employing certain types of dyes to enhance image contrast [20, 21]. However, one of the main potential limitations of video endoscopy in clinical applications is the light source used by the endoscopic imaging system – imaging information obtained with video endoscopy is usually constrained by the FOV and

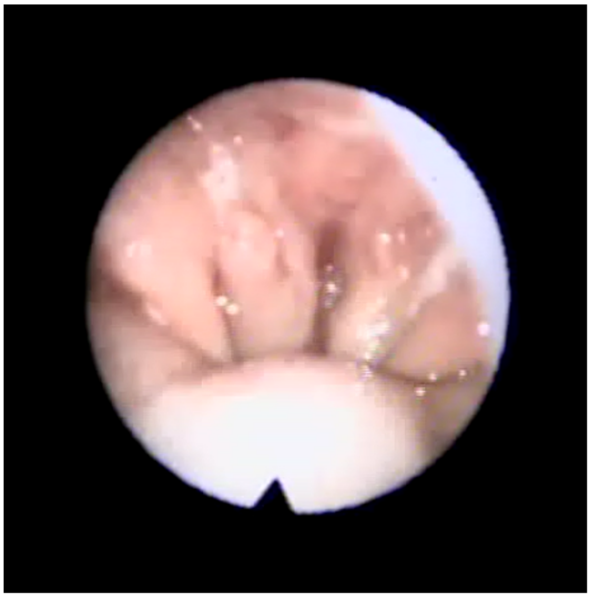


Figure 2.1: A typical image of the surface of the pharynx obtained by using video endoscopy (image courtesy of Dr. Brad Sutton and Dr. Jamie Perry).

the resolution of its front-end camera. The dependence on camera performance makes it difficult to detect abnormal structures beneath the surface of the vocal tract cavity with video endoscopy. In addition, video endoscopy can be regarded as an invasive imaging method through which the subjects suffer great discomfort. These potential limitations of video endoscopy makes it a less commonly used imaging method for dynamic speech imaging applications.

Ultrasound

Ultrasound is a typical non-invasive imaging modality. It employs a range of high-frequency sound waves for the purpose of imaging. Specifically, an ultrasound transducer focuses high-frequency sound waves into a sound beam. The resultant sound beam is further introduced into the human body to visualize an inner structure [22, 23]. The introduced sound beam bounces back when it reaches air-tissue boundaries in the vocal tract. When the sound beam is bounced back, it can be captured by the transducer in the form of voltage signals [22]. The locations of the soft-tissue structures can be determined by calculating the timing, amplitude and even phase [23] of the returned sound beam. This non-invasive feature of ultrasound has made it a popular imaging modality for studying the oropharyngeal structure.

Ultrasound has its own drawback: the sound beam produced by the transducer has limited penetration depth due to signal loss across the air-tissue boundaries. This prevents the clinicians from getting a visualization of the inner structures of the vocal tract, even when the transducer is placed close to the lower chin [22]. In addition, ultrasound has limited soft-tissue contrast compared with MRI, CT or video fluoroscopy. These drawbacks have limited the use of ultrasound for dynamic speech imaging applications [24–26].

Video fluoroscopy

Video fluoroscopy integrates the imaging power of conventional X-ray fluoroscopy with accelerated visualization techniques to provide a sequence of X-ray images of the vocal tract. Video fluoroscopy generates high-frequency electromagnetic waves that pass through soft tissues of the human body and get picked up by the X-ray receiver. As different tissues and structures have different absorptions along its beam path, the received X-ray carries information about tissue structures as two-dimensional (2D) projections.

Video fluoroscopy has long been treated as the "gold standard" imaging modality for clinical dynamic speech imaging experiments [27–30]. However, video fluoroscopy is an imaging modality that relies on the use of ionizing radiation, rendering it inappropriate for dynamic speech imaging scans that require long scan time across a large number of subjects. As given in Figure 2.2, video fluoroscopy obtains images from a single projection angle, which prevents visualization of complex 3D soft-tissue structures.

Computerized tomography

Computerized tomography (CT) exceeds video fluoroscopy in its capability to differentiate spatial features. Unlike video fluoroscopy, CT obtains imaging data from more than one projection

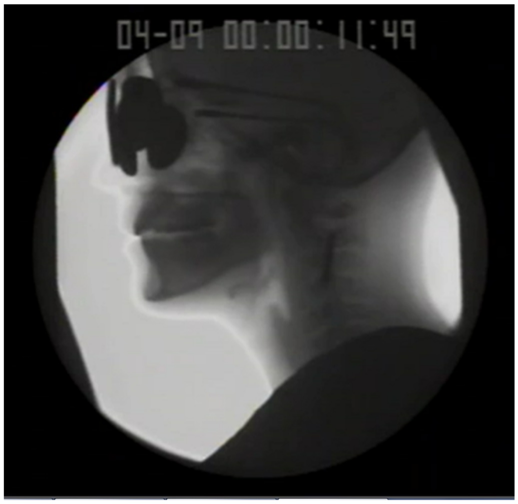


Figure 2.2: A typical image of the upper vocal tract obtained by using video fluoroscopy (image courtesy of Dr. Brad Sutton and Dr. Adrienne Perlman). This image was obtained while the subject was swallowing liquid.

angle. This allows the ensemble of data to be combined and reconstructed according to the projection slice theorem. As a result, CT is capable of differentiating anatomical features in space instead of providing overlapping projected spatial images. Given this characteristic, CT has long been regarded as a competitive alternative to video fluoroscopy for clinical applications related to speech [31, 32]. Recent developments in CT have also resulted in higher spatial resolution, faster imaging speed and the presence of multimodal CT imaging techniques. These developments in CT have led to a range of exciting applications in speech [31, 32]. Like video fluoroscopy, CT is by its nature an imaging modality relying on the use of ionizing radiation. This characteristic of CT renders it inappropriate for large-scale phonetics studies despite many of its favorable properties.

2.1.2 MR-based dynamic speech imaging

Compared with conventional modalities for dynamic speech imaging, MRI provides excellent soft-tissue contrast along arbitrary view planes without the risk of ionizing radiation. This favorable property has made MRI an ideal imaging modality for investigating both the structures and dynamics in the vocal tract.

Early applications of MRI to speech-related studies mainly centered on imaging static or prolonged utterance [33]. Specifically, the subjects were requested to hold articulatory gestures during the length of the experiment. As a result, the sounds of interest are largely limited to simple vowels by which the subjects can easily maintain articulatory gestures over a relatively long time, such as 2 s or more in early applications [33]. Typical vowels for these experiments include vowel sounds such as /a/, /i/ and /u/. As a result, typical regions of interest mainly include the tongue body and the soft palate [34]. Despite these successful reports, prolonged utterance significantly limits the variety of sounds that can be studied with MRI. This is because the important transitions between different sounds and the dynamics within each sounds were completely beyond observation with these early methods.

An important improvement over the early speech MRI methods was achieved by the utilization of gating techniques. Gating techniques are based on the assumption that there exist repetitions of the carrier phrase during speech. Therefore, this assumption can be used to relax the sampling requirements if the speech contains certain repetitions or quasi-repetitive patterns. Specifically, a portion of $\mathbf{k}$-space data is collected during each repetition, and the ensemble of required $\mathbf{k}$-space samples is collected at the end of all the repetitions. As a result, acquiring $\mathbf{k}$-space data across multiple repetitions can be treated as identical to instantaneous sampling of the required samples in a single repetition. Compared with the earlier techniques, the gating techniques have allowed complex speech samples, rather than static ones, to be studied using MRI [35–38]. However, the performance of gating techniques greatly relies on the speaker's ability to reproduce consistent motion across all the repetitions. Motion variability may heavily corrupt speech dynamics for even simple utterances, such as "golly" [39]. This drawback has switched people's attention to dynamic speech imaging methods that allow natural sound production in real time.

High-resolution, three-dimensional (3D) dynamic speech MRI, as the latest development in speech imaging, holds great promise for scientific research and clinical applications in speech. Broad spatial coverage is especially useful both for phonetics studies and clinical application. This is because the sound production may require coordinated movements from multiple structures with complex geometries at different locations of the vocal tract. Having a broad spatial coverage allows these coordinated movements to be captured and compared at the same time. In particular, 3D spatial coverage over the entire vocal tract has shown great benefits for a number of speech-related studies [35, 40]. High spatiotemporal resolution, on the other hand, is needed to delineate the gestures of small-scaled articulators and fast transitions of articulatory dynamics. 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 [41]. Other studies have also indicated that structural changes within short intervals, even as short as 10 ms [10–14], may contain important information about speech function. In this way, increasing the imaging speed of dynamic speech MRI is critical towards capturing the subtle temporal transitions that may help reveal the structural and functional changes of the articulators. A review of imaging protocols for dynamic speech MRI and the associated acquisition and reconstruction tools can be found in [42]. Some important imaging protocols, data acquisition methods and image reconstruction tools will also be discussed in detail in the following sessions.

2.2 Dynamic speech MRI: Signal processing approaches

Dynamic speech MRI brings great challenges to MR data acquisition and image reconstruction. Compared with the rapidly changing articulatory motion, the relatively slow sampling speed of conventional MR data acquisition prevents collection of the full Nyquist samples. Therefore, advanced signal processing approaches leveraging the physics and signal properties of dynamic speech MRI are needed to properly capture the fast transitions of speech motion.

2.2.1 Partial separability model-based imaging

Acceleration in the imaging speed of dynamic speech MRI can benefit from embracing the generic properties of the images representing speech motion. As a generic signal property [43], partial separability (PS) exploits the fact that many speech images have a high degree of spatiotemporal correlation and, as a result, exist in a subspace of lower dimensionality. It is worth noting that partial separability is a natural property that is often observed in images or image sequences from a variety of applications, including spectroscopic imaging [44, 45], cardiac imaging [46–48] and speech imaging [42,49–51]. Based on the PS model, the dynamic MR images representing speech motion can be represented as [43]:

$$I(\mathbf{r},t)=\sum_{l=1}^{L}\psi_l(\mathbf{r})\phi_l(t), \tag{2.1}$$

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 up to model order L. It should be noted that the Lth-order PS model induces a low-rank structure: the Casorati matrix $\hat{\mathbf{I}}$ defined over the point set $\{I(\mathbf{r}_n,t_m)\}_{n,m=1}^{N,M}$,

$$\hat{\mathbf{I}}=\begin{bmatrix} I(\mathbf{r}_1,t_1) & \cdots & I(\mathbf{r}_1,t_M)\\ \vdots & \ddots & \vdots\\ I(\mathbf{r}_N,t_1) & \cdots & I(\mathbf{r}_N,t_M)\end{bmatrix}, \tag{2.2}$$

is a low-rank matrix with its rank upper bounded by L [43, 52], where N and M are the number of spatial encodings and temporal frames. This implies that $\hat{\mathbf{I}}$ exists in an L-dimensional subspace ($L \ll \min\{N, M\}$) and can be represented by a matrix factorization $\hat{\mathbf{I}}=\mathbf{U}\mathbf{V}$, 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 $\hat{\mathbf{I}}$ [43, 52]. It should be noted that the PS model represents a generic property and, as a result, the Casorati matrix in the $(\mathbf{k},t)$-space can be expressed as:

$$\hat{\mathbf{C}}=\begin{bmatrix} d(\mathbf{k}_1,t_1) & \cdots & d(\mathbf{k}_1,t_M)\\ \vdots & \ddots & \vdots\\ d(\mathbf{k}_N,t_1) & \cdots & d(\mathbf{k}_N,t_M)\end{bmatrix}, \tag{2.3}$$

the rank of which is also upper bounded by L [43, 52]. Similarly, this implies that $\hat{\mathbf{C}}$ exists in an L-dimensional subspace and allows the factorization $\hat{\mathbf{C}}=\mathbf{\Phi}\mathbf{\Psi}$, where columns of $\mathbf{\Phi}\in\mathbb{C}^{N\times L}$ span the spatial subspace of $\hat{\mathbf{C}}$, and rows of $\mathbf{\Psi}\in\mathbb{C}^{L\times M}$ span the temporal subspace of $\hat{\mathbf{C}}$ [43,52]. Compared with full-rank or higher-rank matrices, low-rank matrices rely on fewer degrees of freedom to represent the signal of interest by leveraging the correlation between spatial and temporal information. In this way, the PS model allows high-quality speech dynamics to be captured with high spatial resolution and high temporal frame rate.

Conventional imaging methods towards recovering the Casorati matrices, $\hat{\mathbf{I}}$ or $\hat{\mathbf{C}}$, are performed based on the rank minimization-based approaches. For instance, a typical rank minimization-based method solves the following problem,

$$\hat{\mathbf{C}} = \arg\min_{\mathbf{C}} \mathrm{rank}(\mathbf{C}) \text{ s.t. } ||\mathbf{d} - \mathbf{E}\mathbf{C}||_2 \leq \epsilon, \tag{2.4}$$

where $\mathbf{d}$ is the sparsely sampled data from the dynamic speech MRI experiment; $\mathbf{E}$ is the encoding operator that includes a sparse sampling operation, which retains the elements of $\mathbf{C}$ at the sampled locations; and ϵ is the tolerance level for data discrepancy. It has been indicated that $\mathbf{C}$ can be successfully recovered with high probability when certain numerical conditions are satisfied [53–55].

Compared with the matrix completion-based approaches, the reconstruction problem can be simplified when either the temporal or spatial subspace of $\mathbf{C}$ is predetermined. For instance, the PS model [43] applies an explicit rank constraint to reconstruct $\mathbf{C}$ in an elegant way: determination of the temporal subspace and the spatial subspaces of $\mathbf{C}$ is performed in two separate, but synergistically connected, steps. As a result, the image reconstruction problem requires fewer spatiotemporal samples and becomes more straightforward compared with the matrix completion-based approaches. In particular, the image reconstruction problem is equivalent to recovering the spatial subspace $\mathbf{\Phi}$ of $\mathbf{C}$ if the temporal subspace $\mathbf{\Psi}$ is predetermined. Mathematically, the image reconstruction problem can be expressed as:

$$\hat{\mathbf{\Psi}} = \arg\min_{\mathbf{\Psi}} ||\mathbf{d} - \mathbf{E}\mathbf{\Psi}\mathbf{\Phi}||_2^2. \tag{2.5}$$

As can be seen, enforcing an explicit rank constraint (predetermination of the temporal subspace) significantly simplifies the reconstruction problem and greatly reduces the number of spatiotemporal samples required to properly recover the speech dynamics [52]. Given these advantages of the PS model, the following chapters of this dissertation will focus on detailed strategies of applying PS model-based data acquisition and image reconstruction methods to capture high-quality spatiotemporal dynamics of speech.

2.2.2 Compressed sensing-based imaging

The compressed (or compressive) sensing theory exploits the compressibility or sparsity in natural images to recover image from undersampled data. Compared with methods that rely on spatiotemporal support constraint to facilitate image recovery [56–59], sparsity in a transformed domain may be a more generic signal property: a sparsifying transform can be applied to transform the image of interest into a certain transformed domain where only a few non-zero coefficients exist. The compressed sensing scheme allows effective sampling of compressed image data, instead of sampling at the amount specified by the Nyquist-Shannon theorem, before recovering the sparsely sampled data through sparsity-promoting algorithms [60–64]. Compared with conventional sampling strategies, the sampling strategies guided by the compressed sensing theories allow the inherent correlation in the signal of interest to be exploited, and consequently relax the sampling requirements. Therefore, the compressed sensing-based approaches, often referred to as sparse modeling-based approaches, have found extensive use in a variety of MR applications [65–67]. The interested reader can also refer to [16] for a detailed review of the classical compressed sensing-based imaging methods in the context of general dynamic MRI.

Mathematically, the compressed sensing theory aims to obtain a solution for the following problem

$$\mathbf{d} = \mathbf{E}\mathbf{I} + \epsilon, \tag{2.6}$$

where $\mathbf{d}$ denotes the measurement data according to a certain sampling strategy, $\mathbf{E}$ denotes an encoding operator that models the imaging process and ϵ denotes the measurement noise. Without loss of generality, if we assume the existence of a transform $\mathbf{T}$ such that $\mathbf{I}$ can be mapped to a vector that is sparse (containing a limited number of non-zero elements and a large number of zero elements), the image recovery problem can be rewritten as follows,

$$\hat{\mathbf{I}} = \arg\min_{\mathbf{I}} ||\mathbf{T}\mathbf{I}||_0 \text{ s.t. } ||\mathbf{d} - \mathbf{E}\mathbf{I}||_2^2 < \epsilon, \tag{2.7}$$

where $||\cdot||_0$ denotes the l_0 norm that calculates the cardinality of non-zero elements in a certain vector. It should be noted that the use of $||\cdot||_0$ results in a non-convex optimization problem that is NP-hard, i.e., the problem cannot be settled within polynomial time, and consequently a common alternative to the above optimization problem relies on the use of the l_1 norm as surrogate. In practice, the l_1 norm is commonly used because it is regarded as the tightest convex relaxation of the l_0 norm, while other surrogate functions have been used to promote desirable image property. Without loss of generality, the image recovery problem with l_1 norm as a surrogate metric can be rewritten in the following form,

$$\hat{\mathbf{I}} = \arg\min_{\mathbf{I}} ||\mathbf{T}\mathbf{I}||_1 \text{ s.t. } ||\mathbf{d} - \mathbf{E}\mathbf{I}||_2^2 < \epsilon. \tag{2.8}$$

If we introduce the Lagrange multiplier, the above problem can be further rewritten as:

$$\hat{\mathbf{I}} = \arg\min_{\mathbf{I}} ||\mathbf{T}\mathbf{I}||_1 \, + \, \lambda||\mathbf{d} - \mathbf{E}\mathbf{I}||_2^2 < \epsilon, \tag{2.9}$$

where λ denotes the Lagrange multiplier. The compressed sensing theory asserts that the image $\mathbf{I}$ in the above formulation can be exactly recovered with high probability when the restricted isometry property is satisfied [68–71]. In practice, however, the restricted isometric property is not straightforward to validate. Also, the available acceleration in imaging speed by using compressed sensing alone is often limited. In order to demonstrate this, numerical simulations have been performed based on a numerical phantom developed in [51] and the comparison of reconstruction quality with multiple methods is shown in Figure 2.3. As can be seen, Figure 2.3a shows a representative mid-sagittal image and the associated temporal profile (taken with a vertical strip across the tip of the tongue) from the gold standard. Figure 2.3b illustrates a representative mid-sagittal image and the associated temporal profile (taken with a vertical strip across the tip of the tongue) from a low-rank model-based method. Figure 2.3c shows a representative mid-sagittal image and the associated temporal profile (taken with a vertical strip across the tip of the tongue) from a compressed sensing-based method. Figure 2.3d shows a representative mid-sagittal image and the associated temporal profile (taken with a vertical strip across the tip of the tongue) from a sliding window-based method. As can be seen, the compressed sensing-based method suffer from compromised image quality and degraded temporal dynamics when the required acceleration in imaging speed is high (simulated with less sampled data). The sliding window-based method, on the other hand, suffers from significant spatial and temporal blurring. As a result, compressed sensing methods are often used in complementary to other imaging strategies to yield improved reconstruction quality.

2.3 Dynamic speech MRI: Data acquisition approaches

The previous section focuses on introducing signal processing-based approaches to dynamic speech MRI. Successful application of these approaches for practical phonetics experiments requires corresponding data acquisition schemes. This section focuses on providing a detailed review of the data acquisition approaches towards dynamic speech MRI, covering a variety of spatiotemporal sampling patterns, fast pulse sequences, data acquisition hardware and imaging protocols.

2.3.1 Spatiotemporal sampling

The design and optimization of the spatiotemporal sampling pattern is critical towards the image quality in dynamic speech MRI reconstructions. A variety of sampling strategies have been proposed to efficiently and effectively traverse the $(\mathbf{k}, t)$-space during dynamic speech MRI data acquisition. This subsection focuses on the representative $(\mathbf{k}, t)$-space sampling trajectories and analyzes the influence of each type of sampling trajectories on the quality of the resulting reconstruction.

Cartesian sampling is perhaps the most commonly used sampling trajectory in MRI due to its straightforwardness and simplicity in implementation [72]. Specifically, Cartesian sampling on a 2D image plane or a 3D image volume is often the default option for a series of imaging sequences on existing commercial platforms [73]. For MRI applications that involve high frequency and great amplitude variation motion, Cartesian sampling has also been shown to reduce magnetic susceptibility artifacts [16, 72, 74]. When combined with half-Fourier sampling techniques, Cartesian

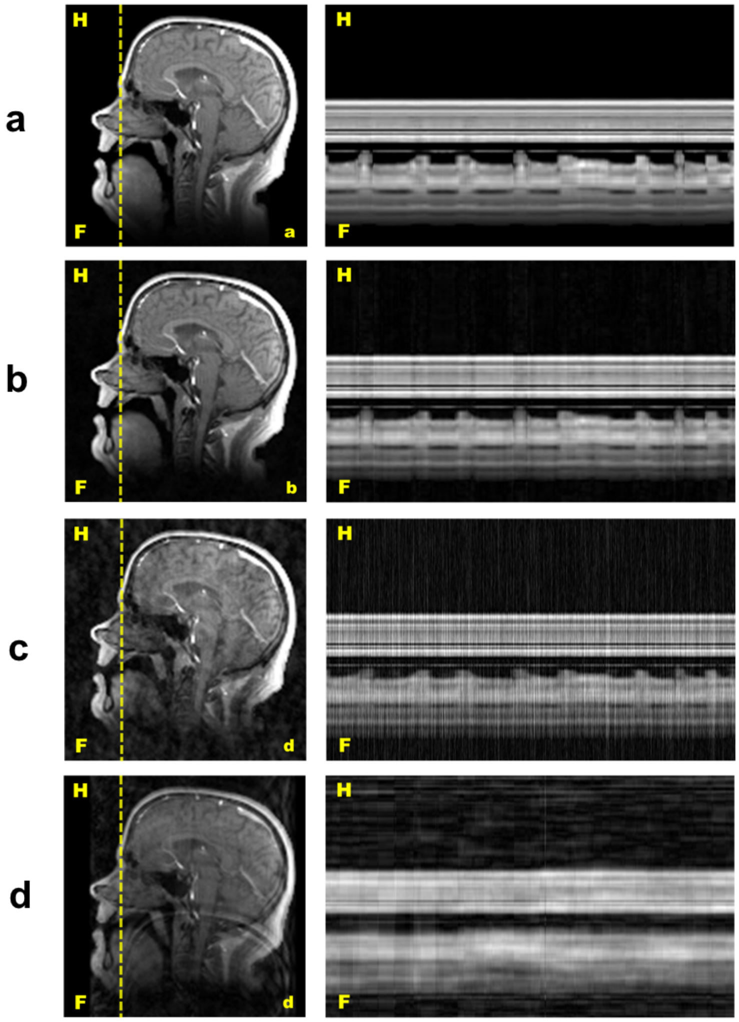


Figure 2.3: Comparison of reconstruction quality with multiple methods on a numerical phantom: (a) a representative mid-sagittal image and the associated temporal profile (taken with a vertical strip across the tip of the tongue) from the gold standard; (b) a representative mid-sagittal image and the associated temporal profile from a low-rank model-based method; (c) a representative mid-sagittal image and the associated temporal profile from a compressed sensing-based method; and (d) a representative mid-sagittal image and the associated temporal profile from a sliding window-based method. As can be seen, the compressed sensing-based method suffer from compromised image quality and degraded temporal dynamics when the required acceleration in imaging speed is high (simulated with less sampled data). The sliding window-based method, on the other hand, suffers from significant spatial and temporal blurring.

sampling also allows accelerated sampling by a factor of 2 to 3 without significantly changing the underlying imaging sequence or protocol [42].

Compared with Cartesian sampling, non-Cartesian sampling trajectories are often employed to achieve a good trade-off between spatial resolution, temporal resolution and spatial coverage [50, 51, 74–76]. A typical non-Cartesian sampling trajectory is the spiral trajectory, which has been shown as a good trade-off between SNR, $\mathbf{k}$-space coverage and sampling speed [77–79]. These desirable properties of the spiral trajectory come from the fact that it is intrinsically densely-sampled in the center of $\mathbf{k}$-space, while quickly spiraling out towards the edge of the $\mathbf{k}$-space. This unique combination of the spiral trajectory gives it unprecedented advantage over the Cartesian trajectory [50, 51, 74–76]. Other options for non-Cartesian sampling patterns include radial sampling patterns and zig-zag sampling patterns. A detailed analysis of the superior performance of non-Cartesian trajectories to their Cartesian counterparts can be found in [16].

Advanced model-based data acquisition schemes often rely on a combination of multiple spatiotemporal sampling patterns. For instance, PS model-based dynamic speech MRI applies a composite sampling strategy where two types of sampling trajectories were employed [50, 51, 74–76] - a Cartesian trajectory is used to acquire the imaging data with random phase encoding orders, while a spiral or cone trajectory was used to acquire the navigation data with extended $\mathbf{k}$-space coverage. It should be noted that the cone trajectory [80–82] can be viewed as a modified spiral trajectory with consideration of an additional spatial dimension in the $\mathbf{k}$-space. The combination of multiple spatiotemporal sampling patterns allows the advantage of each sampling pattern to be fully exploited, and consequently the acquired data can be better utilized to assist the ensuing image reconstruction problem. Recent developments in other fast MR applications also indicate the potential of a variety of composite sampling trajectories [83–85].

2.3.2 Pulse sequences

A variety of fast pulse sequences have been developed to accommodate the physical and physiological requirements of dynamic speech MRI experiments. In general, both gradient-echo-based and spin-echo-based sequences have been developed and used for imaging the vocal tract. The spin-echo-based sequences are often used to visualize static structures of the upper vocal tract or used as pre-scans prior to the dynamic data acquisition [86–90], while the gradient-echo-based sequences are often used for dynamic speech MRI experiments [87, 91–93]. It is worth noting that recent advances in steady state free precession (SSFP)-based sequences have also been used to improve the signal-to-noise ratio (SNR) of dynamic speech MRI acquisition [35, 94], while a potential disadvantage of SSFP-based data acquisition remains in its sensitivity to field inhomogeneities [95, 96]. To obtain optimized data acquisition quality, the imaging protocol often has to take into consideration the interactions between the pulse sequences, the quality of shimming and the field strength of the scanner.

Simultaneous high imaging speed with improved image quality can be achieved when the gradient-echo-based sequences are properly integrated with model-based data acquisition schemes. An example of successful integration is the implementation of a spiral FLASH sequence to sample the $\mathbf{k}$-space at high imaging speed [97,98]. This sequence has been further modified and optimized to properly capture the fast transition of articulatory motion at an interval as short as 5.99 ms [51]. Specific parameters of the optimized sequence are as follows: a TR of 5.99 ms, a TE of 1.85 ms for the imaging data, a TE of 3.25 ms for navigator data, an acquisition matrix size of 128 × 128 × 8, an FOV of 280 × 280 × 40 mm^3 and a spatial resolution of 2.2 × 2.2 × 5.0 mm^3. Prior to the acquisition of the dynamic imaging data using the above-mentioned sequence, 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.

2.3.3 Parallel imaging

Parallel imaging-based methods improve the quality of dynamic speech MRI by exploiting the potential of multichannel receiver coils. For a typical dynamic speech MRI experiment based on parallel imaging techniques, the required data are independently collected from $\mathbf{Q}$ receiver channels. A typical usage of parallel imaging in dynamic speech MRI is using 12 head channels together with 4 neck channels. During the reconstruction step, the data collected from these acquisition channels are combined into the reconstructed image sequence $I(\mathbf{r}, t)$. The use of multiple phased array coils in a dynamic speech MRI experiment allows the spatiotemporal signal to be better recovered from the collected data even when they are sampled under the Nyquist rate. This property is especially favorable for dynamic speech MRI experiments where fast transitions of articulatory motion are anticipated.

The application of parallel imaging techniques to accelerate data acquisition in dynamic speech MRI is rooted in Papoulis's multichannel sampling theorem of band-limited signals. For a spatiotemporal band-limited signal, $I(\mathbf{r}, t)$, representing the articulatory motion, if we assume that $I(\mathbf{r}, t)$ has a spatial support of width W and the spatiotemporal samples are acquired using Q acquisition channels, Papoulis's multichannel sampling theorem asserts that $I(\mathbf{r}, t)$ can be recovered from the underlying data without significant aliasing when the sampling speed is above the sampling speed of Q / W with a proper choice of interpolation kernel ϕ_q [99–102]. This statement also holds true for its equivalent in the Fourier space $d(\mathbf{k}, t)$ [99–102].

Examples of representative parallel imaging techniques include SENSE [103], GRAPPA [104] and their variants [105–108]. Although these techniques may differ in their specific strategies to combine multichannel data, invariably they distribute the phase encoding portions of the field of view among an array of receiver coils and carry out these tasks in parallel. Although the multichannel sampling allows a certain acceleration factor, it should be noted that the acceleration enabled by using multiple sampling channels are still limited. Specifically, the acceleration is still limited to approximately 3 to 4 folds for 2D dynamic speech MRI applications and roughly 8 folds for 3D dynamic speech MRI applications. It is also worth noting that the multiband acquisitions or simultaneous multi-slice acquisitions can acquire up to 12 slices simultaneously using 32- or 64-channel coils for brain imaging, but the acceleration achievable is dependent on the slice orientation and the coil geometry. Appropriate coils for speech imaging are not readily available, although some developments have been made [109–112]. Another potential problem involves the reduced SNR penalty, while the other potential problem involves the measurement noise from each receiver coil [103–108]. Given these practical considerations of using parallel imaging techniques, they are often combined with model-based imaging methods to enhance the quality or speed of dynamic speech MRI experiments.

2.3.4 Acquisition hardware

Scanner specification

Performance of the scanner plays an important role in determining the acquisition speed and the quality of the acquired data. Subjects of dynamic speech MRI experiments are often placed in a supine position in the MR scanner. The head motion of each subject was minimized by fixing the positions of the subject's head in the receiver coil with foam pads. Guidance of speech samples (carrier phrases) is often provided through a projection screen. For a typical dynamic speech MRI experiment on a Siemens Trio scanner (Siemens Medical Solutions, Erlangen, Germany), the system specifications are as follows: a field strength of 3 T, a gradient strength of 40 mTm^{-1}, a maximum slew rate of 176 $\mathrm{Tm}^{-1}\mathrm{s}^{-1}$, a 12-channel head receiver coil and a 4-channel neck receiver coil. Similar specifications can also be found in a variety of other scanners [113, 114]. Informed consent is obtained for all subjects, and the experiment needs to be carried out in accordance with the protocols from the Institutional Review Board at the University of Illinois at Urbana-Champaign.

Audio recording

Audio recording during MR data acquisition allows additional information about speech dynamics to be obtained in the form of an acoustic signal. In general, audio recording starts at the same time as MR data acquisition and will be later synchronized with the reconstructed spatiotemporal images of articulatory motion in the post-processing stage. As dynamic speech MRI acquires spatiotemporal samples at a relatively fast sampling speed, the temporal alignment of the acquired acoustic signal and the reconstructed image is critical towards accurate phonetic analysis of the articulatory motion as opposed to the features in the acoustic signal. In addition, the recorded acoustic signal is often contaminated by acoustic noise due to fast transition of gradients in dynamic MR data acquisition. As a result, noise canceling techniques are also important towards the ensuing phonetic analysis. These two challenges have made signal processing of the acoustic signal an important task in properly integrating the reconstruction and the recorded audio.

Recent developments in signal processing techniques have witnessed multiple successful commodity products in providing actively-noise-canceled acoustic recording along MR data acquisition. It has been reported in our own works [51, 76] that the voice of the 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) during data acquisition. Similar setups have also been used in a range of experiments for other purposes, including signal processing applications, photonics applications and biosensing applications [115–125]. Advanced voice enhancement algorithms have also been proposed to improve the quality of the voice recordings [39, 126–128]. These efforts have already facilitated the understanding of the relationship between acoustics and articulation of nasal and oral vowels [129].

2.3.5 Imaging protocol

The choice of imaging protocols involves making decisions on various aspects of dynamic speech MRI. These decisions include understanding of the underlying clinical and scientific purposes of dynamic speech MRI experiments, the specific imaging requirements for a dynamic speech MRI experiment, the requirements for data acquisition (choices of scanner, field strength, gradient strength, slew rates, pulse sequences, spatiotemporal sampling patterns, receiver coils and audio recording facilities), the requirements for image reconstruction (imaging model, image reconstruction formulation and optimization algorithms), as well as the requirements for post-processing of the reconstructed images (audio-video synchronization and phonetics and acoustic analysis). Although these considerations may overlap largely with other subsections above, it is important to make a high-level integrated decision covering all potential aspects of the imaging process in order to achieve optimized image quality.

Recently there has been a growing number of works that provide detailed comparison of the differences between different aspects of the imaging protocol. Among these works, a recent review based on a recent dynamic speech MRI summit held at University of Southern California provides an objective and detailed summary of the widely used dynamic speech MRI techniques [42]. Compared with other existing dynamic speech MRI techniques, it should be noted that the methods presented in this dissertation outperforms their counterparts in terms of imaging speed and spatiotemporal resolution. Detailed comparison of the image quality in clinical and scientific applications is beyond the scope of this dissertation, but will be an interesting research topic in the future.

2.4 Dynamic speech MRI: Clinical and phonetics applications

The previous section focuses on providing a detailed review of the data acquisition aspects in dynamic speech MRI. This section focuses on giving a detailed review of a wide range of applications of dynamic speech MRI both for clinical applications and scientific research. These applications are roughly divided into two categories – clinical and other applications of dynamic speech MRI.

2.4.1 Clinical applications of dynamic speech MRI

Clinical applications of speech MR imaging involve monitoring pathological changes, whether inherited or acquired, in soft-tissue structures of physiological functions of the organs in the vocal tract. Typical organs of interest include the tongue, the lips, the velum, the epiglottis, the pharyngeal wall and the nasal cavity. The geometric shapes and physiological functions of these organs may change significantly in speech diseases and disorders. Speech MRI, as an imaging modality that provides good soft-tissue contrast without using ionizing radiation or causing significant subject discomfort, holds great promise for monitoring the structures, dynamics and pathological states of these organs, especially when compared with endoscopy, ultrasound or CT. Representative applications include using speech MRI to monitor the structure and motion of the cleft palate [27, 130–132], the cleft lips [133, 134], various tongue cancers [135–137], the laryngitis [138–140] and multiple types of damages to the vocal cord [141–143].

The uses of speech MRI in clinical applications are not limited to qualitatively visualizing geometric structures. It has also been used to provide quantitative measurements: physiological parameters have been obtained from various clinical experiments to facilitate clinical assessment of the changes in speech functions. Useful parameters have been obtained from various speech motion, including swallowing movements [144–147], lateral pharyngeal wall motion [6, 148–150] and the rate of airflow [151, 152]. These clinical applications and explorations of MRI have provided valuable intuition towards better health care for subjects suffering from speech diseases and disorders.

While speech MRI has been employed in various applications to monitor pathological, structural and functional changes in the vocal tract, there are certain challenges and limitations in these applications. Although the majority of clinical researchers have acknowledge the potential and effectiveness of applying MRI in speech-related topics, the potential of MRI has not been fully utilized. For instance, there has been limited effort in optimizing the temporal resolution, the spatial resolution and the SNR of the associated imaging protocols. Other technical aspects, such as the choices of pulse sequences, image reconstruction formulations and optimization algorithms are less discussed. While optimization of these aspects may not be the focus for clinical applications of dynamic speech MRI, it is reasonable to believe that better performance or more accurate results can be obtained if these aspects are taken into consideration.

2.4.2 Phonetics applications of dynamic speech MRI

Speech MR imaging is also promising for facilitating the studies of a broad spectrum of speech-related studies, including speech production, speech motor functions and language evolution. During speech production, there exists a range of variations in the articulator postures depending on the subjects' vocal tract structures, language features and the context of the carrier phrases [153–156]. It has been demonstrated in previous studies that speech MRI (including dynamic and static speech MRI) is effective in capturing the subtle differences in the postures of vocal tract articulators from utterance to utterance [157–160]. The unique advantages of dynamic speech MRI also find great use in quantitatively analyzing speech production. These analyses include extracting parameters to evaluate speech motor functions [161–163] and to calculate the area functions at various locations in the vocal tract [1, 164]. Recent progress in multi-modal speech imaging has also witnessed a growing interest in employing dynamic speech MRI jointly with other imaging modalities, such as ultrasound, electromagnetic articulographs (EMA), computerized tomography (CT) or advanced sensing techniques [165–173], to obtain multi-model measurements and analyses of the underlying speech motions. Detailed review of these approaches is beyond the scope of this dissertation.

The applications of dynamic speech MRI are not just limited to analyzing or understanding speech movements themselves. Images and analytic results from dynamic speech MRI also are important towards understanding of subtle language features and language variations. Dynamic speech MRI has been extensively applied to study a variety of widely used languages. In particular, these languages include English [12, 129, 174], Mandarin [175–179], French [10–12, 51, 180], Arabic [8, 13, 14, 50, 76, 181, 182] and Portuguese [9, 183]. Beyond the references provided in this dissertation, there exists a large and growing number of novel applications of dynamic speech MRI that may eventually lead to a deeper understanding of the mechanism, function and dynamics of speech.

CHAPTER 3

HIGH–RESOLUTION 3D DYNAMIC SPEECH MRI

3.1 Introduction

The past decade has seen significant improvement separately in the speed, coverage or resolution of speech MRI. However, it remains challenging to maximize 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 [43]. For instance, a two-dimensional (2D) multi-slice-based approach has enabled visualization of 8-slice speech dynamics at 12.8 fps [74, 75, 184]. A 3D acquisition-based approach also allowed articulatory gestures to be captured at 8.1 fps [40]. Expanding upon earlier approaches [50, 51, 74–76, 184], this chapter 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 × 2.2 × 5.0 mm^3 are achieved with full vocal-tract coverage across 8 slices. The practical utility of this approach has been systematically evaluated by numerical simulations, validation experiments and phonetics investigations.

3.2 Imaging model

As discussed in Section 1.1, the measured data from the $(\mathbf{k}, t)$-space can be expressed as

$$d(\mathbf{k}, t) = \int_{\text{object}} I(\mathbf{r}, t) e^{-i2\pi\mathbf{k}\cdot\mathbf{r}} d\mathbf{r} + \eta(\mathbf{k}, t). \tag{3.1}$$

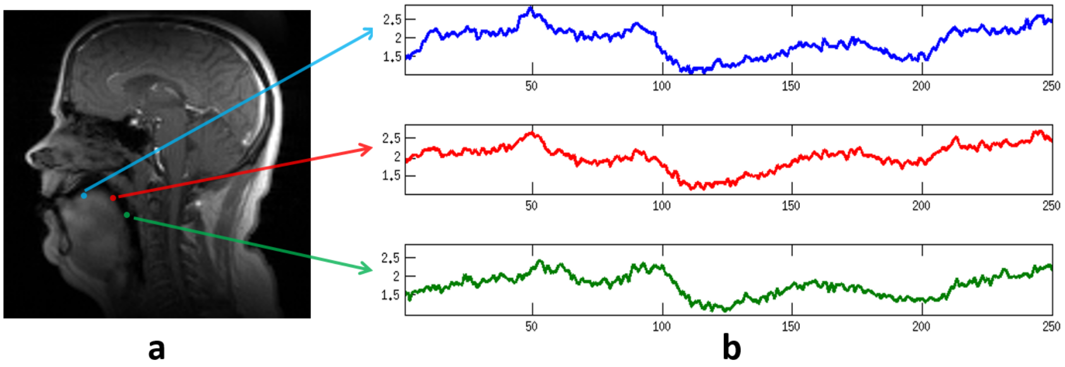


Figure 3.1: Strong spatiotemporal correlation exists in the dynamic image sequence of speech. Three spatial points are taken on the surface of the tongue body as indicated by the colored arrows in (a). As can be seen with (b), although the temporal dynamics at these three spatial points are different at each time point, the overall temporal evolution remain highly correlated. This figure shows that strong spatiotemporal correlation exists in the spatiotemporal speech motion.

For speech MRI experiments, $I(\mathbf{r},t)$ often manifests strong spatiotemporal correlation because (a) spatial images corresponding to articulations of similar sounds are highly correlated, (b) image voxels within certain bulk articulators (e.g. the tongue or velum) share similar temporal dynamics, and (c) only a few driving muscles are involved in the production of a specific syllable. Strong spatiotemporal correlation can be illustrated with Figure 3.1, where three spatial points are taken along the surface of the tongue body from a dynamic speech imaging sequence. As shown in Figure 3.1b, the associated temporal dynamics along these three points are different at each time point but demonstrate strong spatiotemporal correlation.

The PS model quantitatively represents the strong spatiotemporal correlation with a set of partially separable functions [43]:

$$I(\mathbf{r},t) = \sum_{l=1}^{L} \psi_l(\mathbf{r})\phi_l(t), \tag{3.2}$$

where L is the model order, $\{\psi_l(\mathbf{r})\}_{l=1}^{L}$ is a set of spatial basis functions and $\{\phi_l(t)\}_{l=1}^{L}$ is a set of temporal basis functions. The PS model implies that a data matrix $\hat{\mathbf{I}}$ (referred to as a Casorati

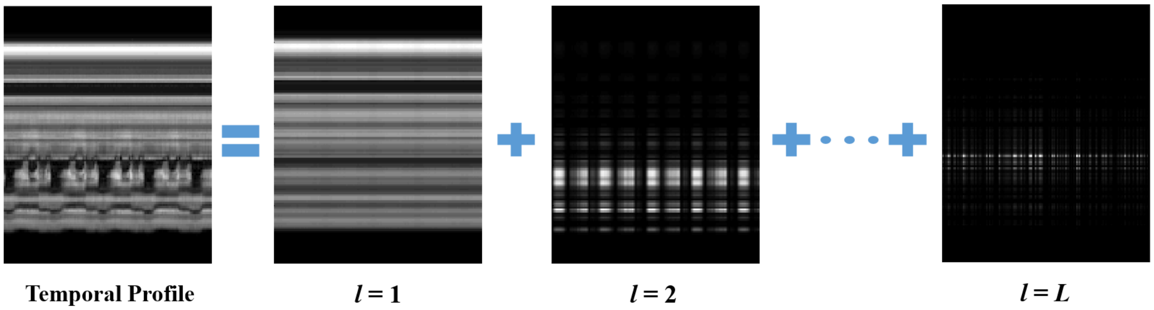


Figure 3.2: Spatiotemporal dynamics at the tip of the tongue can be well-represented by the partial separability model. As can be seen, the temporal profile at the tongue tip can be well-represented as the combination of a finite number of components. These components are represented in the partial separability model as the spatiotemporal dynamics at different model orders.

matrix in [43]) defined over any point set $\{I(\mathbf{r}_n, t_m)\}_{n,m=1}^{N,M}$,

$$\hat{\mathbf{I}} = \begin{bmatrix} I(\mathbf{r}_1, t_1) & \cdots & I(\mathbf{r}_1, t_M) \\ \vdots & \ddots & \vdots \\ I(\mathbf{r}_N, t_1) & \cdots & I(\mathbf{r}_N, t_M) \end{bmatrix}, \tag{3.3}$$

has a rank upper bounded by L [43, 52], where N denotes the number of spatial encoding steps and M denotes the number of image frames in the dynamic image sequence. As given in Figure 3.2, Equation 3.2 induces a low-rank structure and Equation 3.3 implies that $\hat{\mathbf{I}}$ exists in an L-dimensional subspace ($L \ll \min\{N, M\}$). Equation 3.2 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 $\hat{\mathbf{I}}$ [43]. 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.

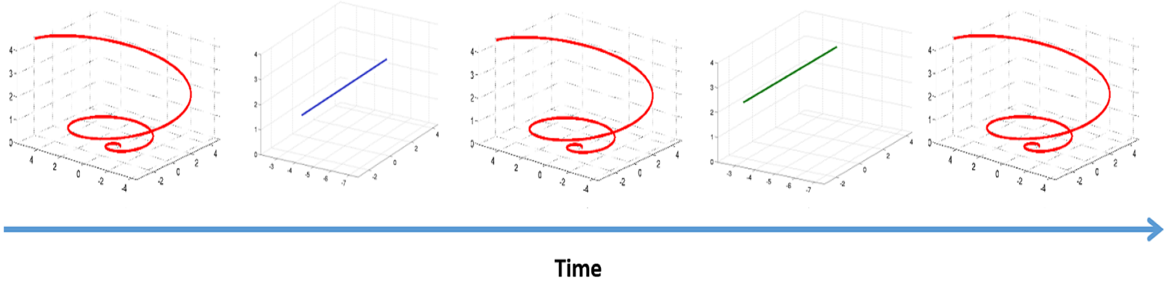


Figure 3.3: Demonstration of the proposed data acquisition strategy with a simplified $(\mathbf{k}, t)$-space sampling pattern. The $(\mathbf{k}, t)$-space is sparsely sampled to obtain two data sets, the navigator data and the imaging data, in an interleaved fashion. Due to the different function of these two data sets, separate considerations were given to the spatiotemporal sampling patterns. The navigator data are sampled with a cone trajectory that traverses the distributed 3D $\mathbf{k}$-space within short temporal intervals. The imaging data are acquired from distributed $\mathbf{k}$-space using Cartesian trajectories with random phase encoding orders.

3.3 Data acquisition

The $(\mathbf{k}, t)$-space is sparsely sampled to obtain two data sets, the navigator and imaging data, in an interleaved fashion [43, 50, 51, 74–76, 184, 185]. Figure 3.3 illustrates the data acquisition strategy with a simplified $(\mathbf{k}, t)$-space sampling pattern. For these two data sets, the navigator data is used to estimate $\mathbf{V}$, while the imaging data is used to estimate $\mathbf{U}$. Although each data set serves a different purpose, both data sets contribute to reproducing $\hat{\mathbf{I}}$.

Due to the different function of these two data sets [43], separate considerations were imposed in terms of acquisition trajectory and sampling requirements. Specifically, the navigator data are acquired using a cone trajectory [50, 74] 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, SNR, and $\mathbf{k}$-space coverage in both the low- and high-spatial-frequency regions: (a) the cone trajectory provides high gradient efficiency to traverse 3D $\mathbf{k}$-space compared with other trajectories, (b) the cone trajectory receives high SNR due to its natural oversampling at the center of $\mathbf{k}$-space and (c) the cone trajectory has the potential to cover broader $\mathbf{k}$-space within certain time constraints or slew rate limits [50, 74, 186, 187]. 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 [51, 184].

A self-navigation strategy is exploited to accelerate the PS model-based acquisition. Unlike previous approaches that collect navigator data with a separate radio frequency (RF) excitation [74, 75], this self-navigation strategy combines the acquisition of both navigator and imaging data into a single repetition time (TR) using a multi-echo readout [47]. 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 3.4 illustrates the proposed acquisition strategy with a simplified pulse sequence diagram.

Additional considerations are given to reduce the sensitivity of the acquisition to eddy current 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 current 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, the shortest length is chosen for the rephasing gradient, which results 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 current.

3.4 Image reconstruction

Given the acquired navigation data, the principal component analysis (PCA) or singular value decomposition (SVD) is performed to estimate the temporal subspace, i.e., matrix $\mathbf{V}$ [43, 185]. Specifically, $\mathbf{V}$ is constructed from the L- most significant right singular vectors of $\hat{\mathbf{C}}$ [43]. With

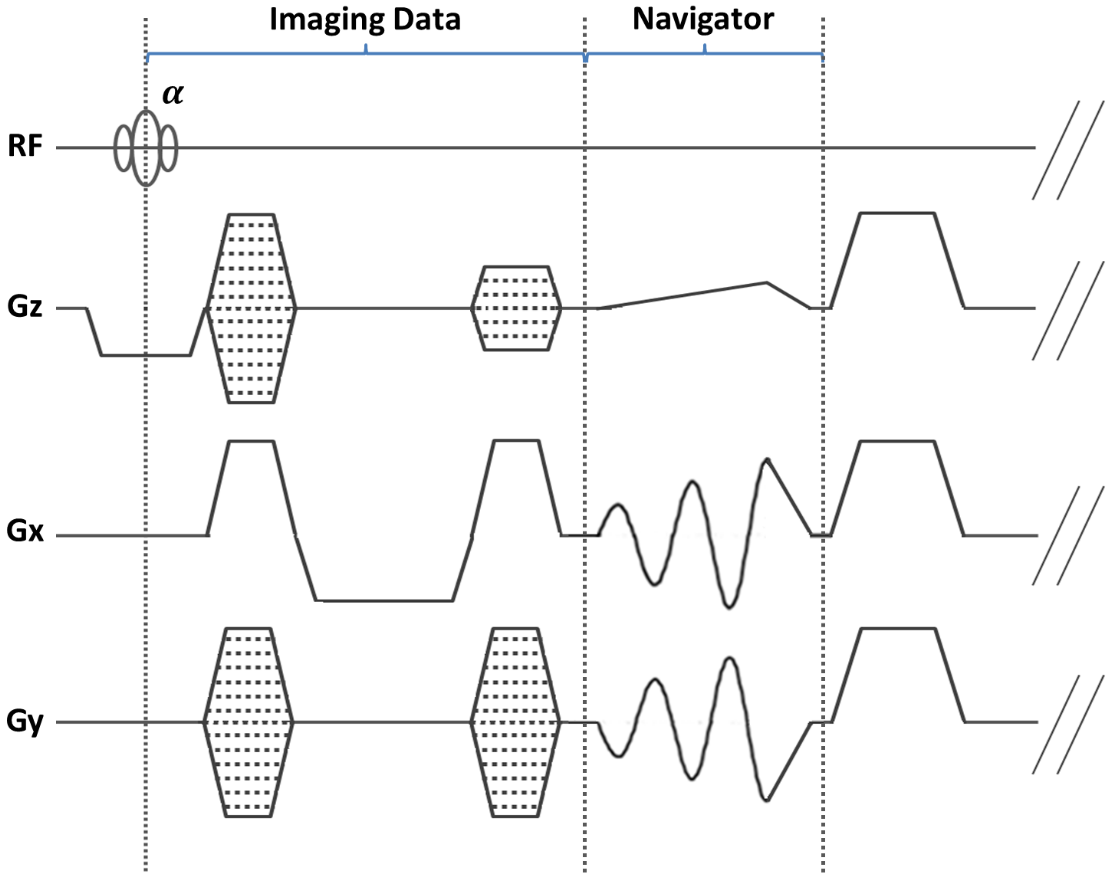


Figure 3.4: A simplified pulse sequence diagram for the proposed “self-navigated” data acquisition strategy. This strategy is proposed to accelerate data acquisition. Within a single TR, the imaging data set is acquired at early echo time using a Cartesian trajectory with random phase encoding, while the navigator data set is acquired at later echo time using a cone trajectory.

$\mathbf{V}$ estimated, one can determine $\mathbf{U}$ from the imaging data by solving a least-squares problem. By separating the estimation of $\mathbf{V}$ and $\mathbf{U}$ in two steps, our method yields a convex and significantly simplified reconstruction problem (compared with methods that simultaneously determine $\mathbf{U}$ and $\mathbf{V}$).

Also, it should be noted that direct determination of $\hat{\mathbf{U}}$ from least-squares fitting is usually ill-conditioned, especially when a higher L is used but limited samples are available [48]. Additional constraint needs to be employed to regularize the ill-conditioned image reconstruction problem. The sparsity constraint has been found effective in existing low-rank constrained reconstruction problems [47, 48]. Combining the sparsity constraint with the low-rank constraint has also been shown to improve the conditioning of the PS model fitting problem. Further, the spatiotemporal total variation (TV) constraint has been applied to other dynamic imaging applications [48, 188]. Compared with a spatial-spectral sparsity constraint [48, 188], 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 [47,48,188], we further developed a method to jointly impose the low-rank and spatiotemporal TV constraints.

The image reconstruction problem can be formulated as follows:

$$\hat{\mathbf{U}} = \underset{\mathbf{U}\in\mathbb{C}^{N\times L}}{\arg\min} \sum_{q=1}^{Q} ||\boldsymbol{\Omega}\{\mathbf{F}\mathbf{S}_q\mathbf{U}\mathbf{V}\} - \mathbf{d}_q||_2^2 + \lambda\,\mathrm{TV}\{\mathbf{U}\mathbf{V}\}, \tag{3.4}$$

where Q denotes the number of receiver coils, $\boldsymbol{\Omega}\{\cdot\}$ 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. In particular, the TV operator is defined as

$$\mathrm{TV}\{\mathbf{U}\mathbf{V}\} = \sum_{j=1}^{M}\sum_{i=1}^{N} ||\mathbf{D}_i\mathbf{U}\mathbf{V}_j||_1, \tag{3.5}$$

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 Equation 3.4 [48].

3.5 Simulation studies

3.5.1 Simulation setup

Numerical simulations were conducted to evaluate the performance of our method. Particular emphasis of the simulation study is placed on demonstrating that high-resolution 3D speech MRI can be achieved with our method. Given that no other dynamic speech imaging method is capable of spatially and temporally resolving speech motion as proposed in this dissertation, a numerical phantom needs to be developed to properly examine the performance of our method in Section 3.4. 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 the specific data sets acquired from *in vivo* experiments.

A generic numerical phantom for 3D dynamic speech MRI was created and simulation studies have been performed to characterize the performance of our 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 $280 \times 280 \times 40$ mm^3, a spatial resolution of $2.2 \times 2.2 \times 5.0$ mm^3, a TR of 5.99 ms and a total number of 71,680 time frames.

Simulated data acquisition followed the $(\mathbf{k}, t)$-space sampling strategy as described in Section 3.3. 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 [189]. 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 data set was acquired after sampling 71,680 time frames from the numerical phantom (equivalent to an acquisition length of 7 min 12 s). Reconstruction from simulated data was performed using our 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 Section 3.4. A model order of 70 and a regularization parameter of 1.31×10^{-6} were chosen based on empirical evaluation of image quality. The following reconstruction error was used to quantitatively assess reconstruction quality [48],

$$\text{error} = \frac{||\mathbf{I}_p - \mathbf{U}\mathbf{V}||_{\mathrm{F}}}{||\mathbf{I}_p||_{\mathrm{F}}}, \tag{3.6}$$

where $\mathbf{I}_p$ represents the numerical phantom and $|| \cdot ||_{\mathrm{F}}$ represents the Frobenius norm.

A modified numerical phantom was also created based on the generic phantom to characterize the frame rate achievable with our method. Specifically, the generic phantom was augmented with a high-temporal-frequency flashing pattern: a bright square 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.

3.5.2 Simulation results

Figure 3.5 shows representative simulation results from the generic phantom. The reconstructed spatiotemporal dynamics are consistent with those in the phantom and have a reconstruction error of 0.0472. Specifically, Figure 3.5a and Figure 3.5b 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 both for the numerical phantom (red box) and the reconstruction (green box) are compared. As seen, the reconstruction faithfully represents the

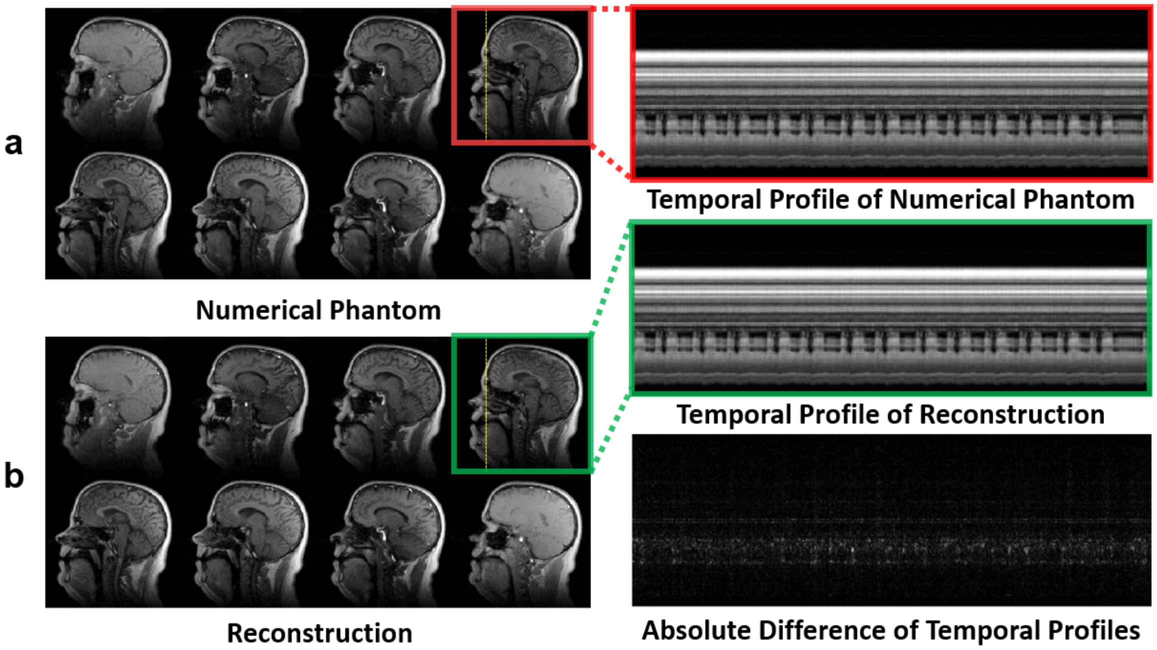

a
b
Numerical Phantom
Reconstruction
Temporal Profile of Numerical Phantom
Temporal Profile of Reconstruction
Absolute Difference of Temporal Profiles

Figure 3.5: Comparison of spatial details and temporal dynamics. The comparison is performed between: (a) the numerical phantom; and (b) reconstruction of the simulated data. The temporal profiles (∼40 s) 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.

temporal dynamics of the phantom without significant temporal blurring, which is also evident by examining the absolute difference in the temporal profiles. Figure 3.5 demonstrates that our method is capable of capturing high-quality spatiotemporal details for dynamic speech MRI.

Figure 3.6 shows representative results from the modified phantom. In particular, Figure 3.6a shows a reconstructed time frame with the added "on" pattern: the added bright square positions above the subject's forehead. As contrast, Figure 3.6b shows an ensuing time frame that has the "off" pattern. Figure 3.6c 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. Figure 3.6 demonstrates that our method is capable of capturing temporal events that require a frame rate of 166 fps.

The performance of our method has been further evaluated on the above numerical phantom using a simulated point spread function (SPSF). In particular, the SPSF is obtained by measuring our method's frequency response, i.e., by comparing the output spectrum of the reconstruction as opposed to the input spectrum of the phantom. As seen in Figure 3.7, the SPSF has a full width half maximum of 1.1 voxel and corresponds to a nominal frame rate of 151 fps. It is noticed that the SPSF provides an empirical description of the frequency response of our method and, despite its own limitation, the results from SPSF depends on this specific numerical phantom and may not be indicative to the true resolution property.

3.6 Experimental studies

3.6.1 Validation experiments

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 chapter, 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 our 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

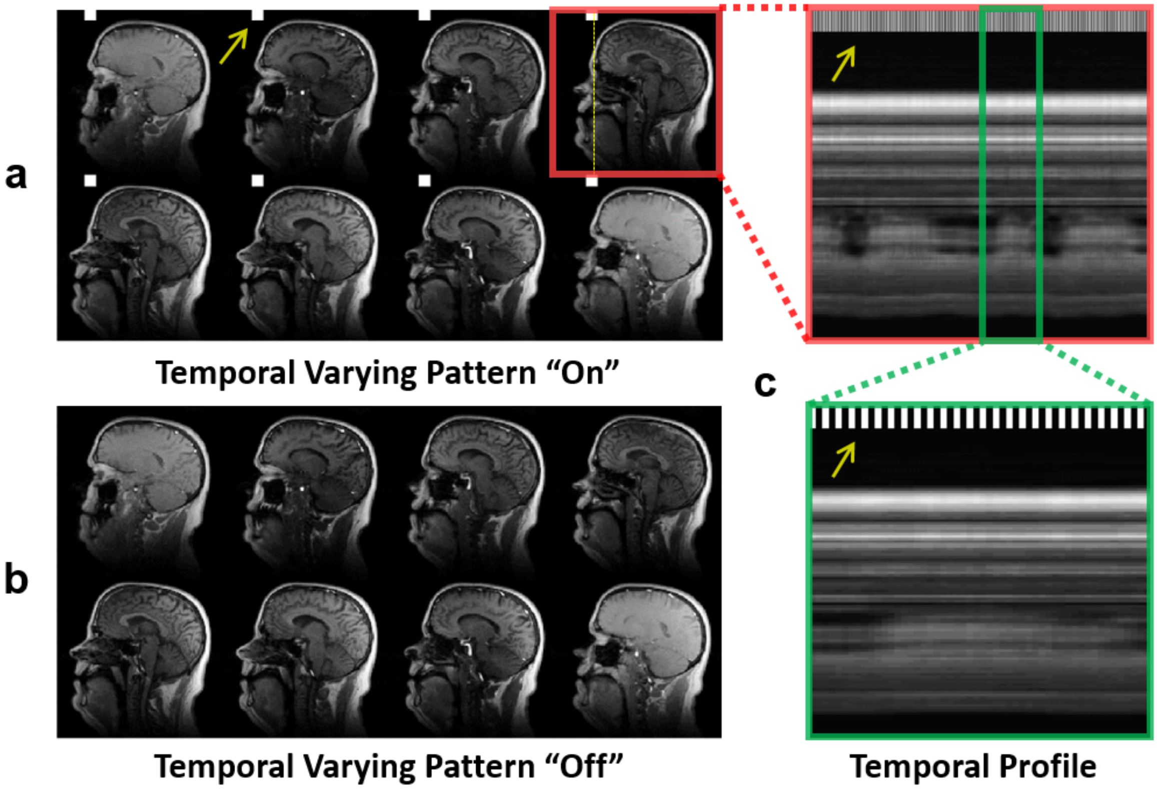


Figure 3.6: Characterization of the nominal frame rate using a modified numerical phantom with a flashing temporal pattern every other frame: (a) a time frame that has a bright square that is positioned above the subject's forehead; (b) a time frame that has no added bright square; (c) the temporal profile along the yellow line across the tongue tip on the fourth 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 relatively small.

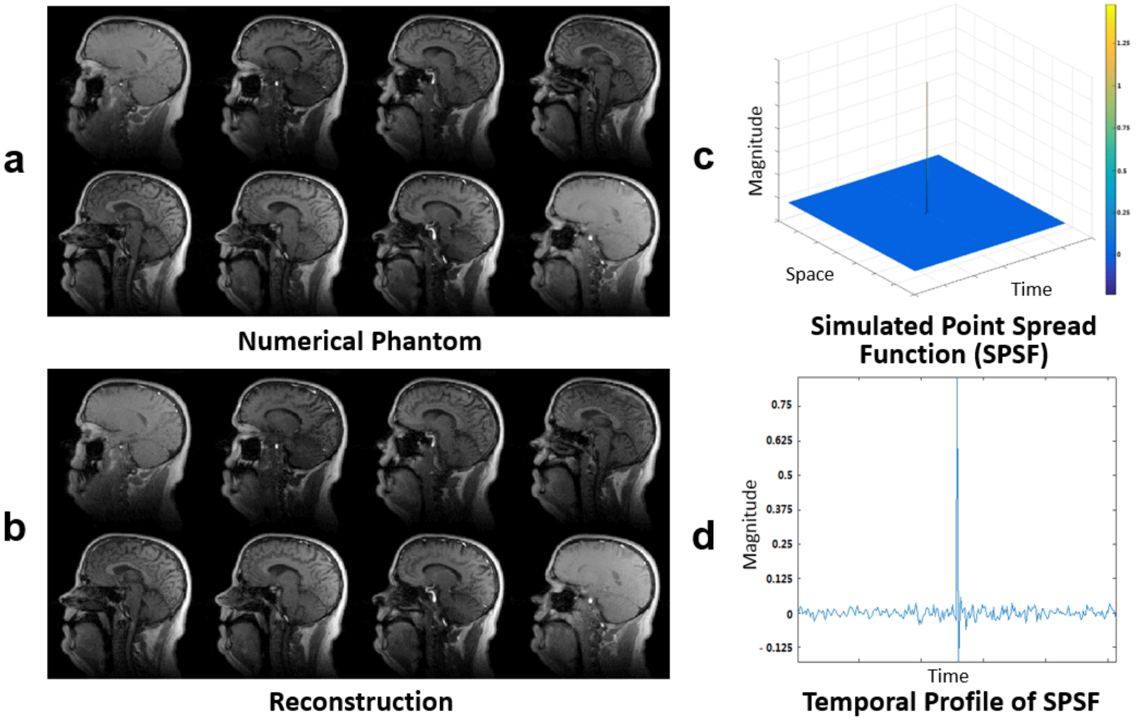


Figure 3.7: Evaluation of the point spread function of our method based on a numerical phantom: (a) articulatory gestures from the numerical phantom; (b) articulatory gestures from the reconstruction; (c) simulated point spread function (SPSF) calculated from the frequency response of our method; (d) the simulated point spread function evaluated across time has a full width half maximum of 1.1 voxel.

the 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 subject's head in the receiver coil with foam pads. Informed consent was obtained for all subjects and the experiment was carried out in accordance with protocols from the Institutional Review Board at the University of Illinois at Urbana-Champaign.

Eight volunteers participated in the validation experiments. Six volunteers were male and two were female. All volunteers were native speakers of American English, and they had an age range of 22 to 38 years. During data acquisition, volunteers were requested to recurrently produce /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ at their natural speaking pace. To investigate whether our method indeed allows for high-quality spatiotemporal dynamics, two additional acquisitions were conducted to compare the performance of our 3D imaging method as opposed to that of a previous 2D multi-slice method [50]. 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 [50], 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 sounds following visual cues ("karaoke" scripts of the /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 experimental design.

3.6.2 Experimental results

Figure 3.8 shows tongue gestures of the upper vocal tract from a 3D imaging experiment where a subject was asked to produce the /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ sounds. Specifically, Figure 3.8a and Figure 3.8b show tongue gestures at the onset of the /l/ sound and the /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 the /l/ sound, while the tongue retracts to a resting position to produce the /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.

The shaping of the tongue is then investigated. The images during the time point of contact are plotted 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 3.9a shows mid-sagittal, coronal and axial views of the tongue during the production of the /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 3.9b, its gesture is reflected in coronal and axial planes as darkened intensity. In addition, incomplete velopharyngeal closure is observed as in Figure 3.8. 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 3.9b.

The temporal dynamics of our method is also evaluated. Figure 3.10 shows spatial images and temporal profiles from a 3D imaging experiment, in which a different subject was asked to produce an identical carrier phrase as in Figure 3.9. Spatial images are shown in the first row of Figure 3.10. Representative temporal dynamics are taken from a line segment on a central mid-sagittal slice and are plotted as opposed to time. Sharp temporal transitions of tongue motion are well-captured and presented in the second row of Figure 3.10. In addition, a transverse line segment is placed right beneath the alveolar ridge to capture the contact from the tongue tip. As illustrated in the second row of Figure 3.10, the temporal profile of tongue-tip contact from the transverse line segment matches well with its mid-sagittal counterpart.

To further investigate if an increased nominal frame rate of the 3D acquisition improves spatiotemporal dynamics over a previous lower-nominal-frame-rate 2D acquisition. Figure 3.11 shows direct comparison of temporal dynamics from our method (with a nominal frame rate of 166 fps) as opposed to that from a previous 2D multi-slice method [50], where the full frame rate is split across the 8 slices and results in a nominal frame rate of 12.8 fps. Specifically, representative

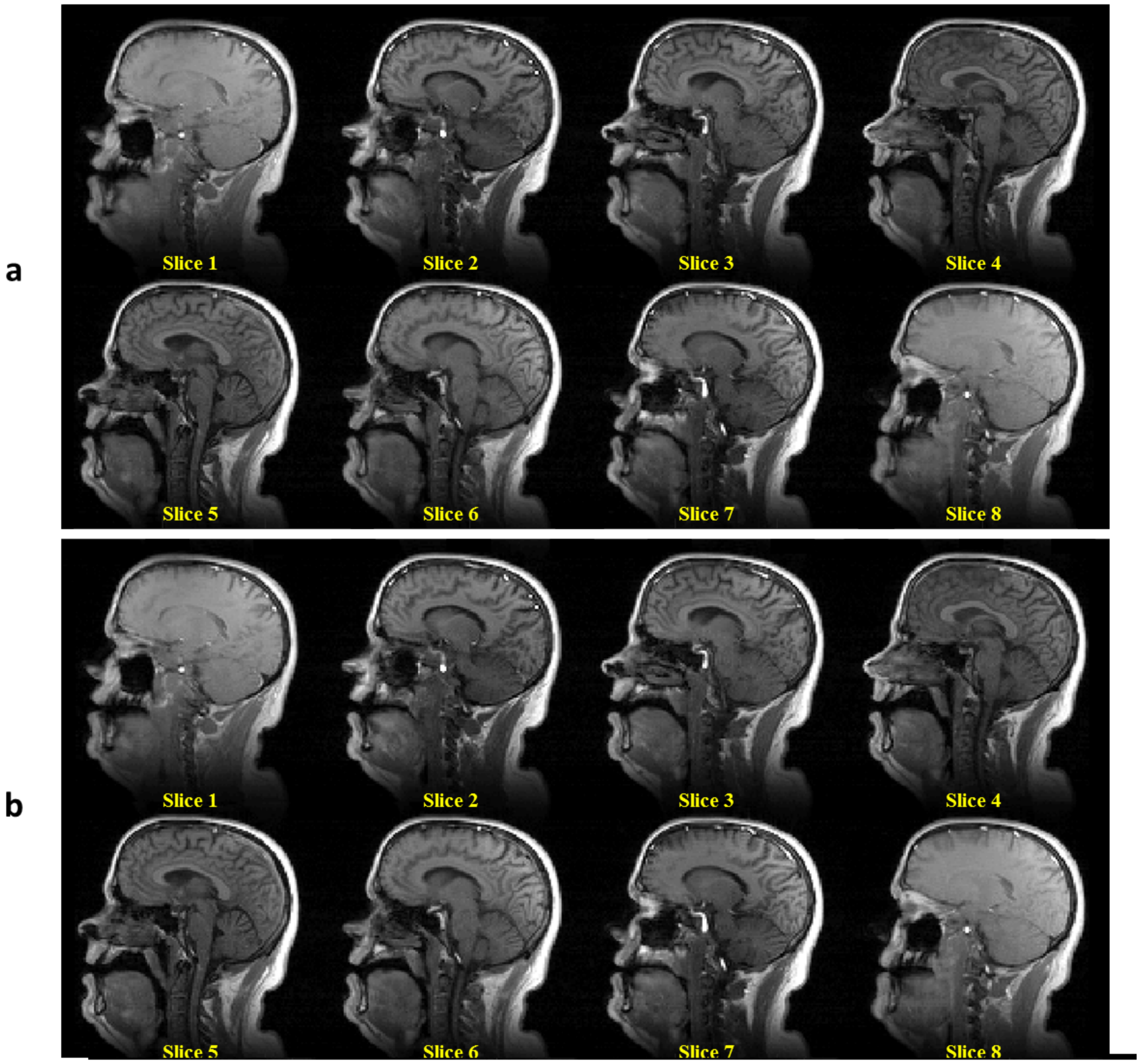


Figure 3.8: Mid-sagittal articulatory gestures in the upper vocal tract during the production of the /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ sounds. The articulatory gestures of the /l/ sound in the /la/ syllable is shown in (a). The articulatory gestures of the /a/ sound in the /la/ syllable are shown in (b). Although the /l/ and /a/ sounds transition within a brief duration (∼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 the /l/ sound, while the tongue retracts to a resting position to produce the /a/ sound.

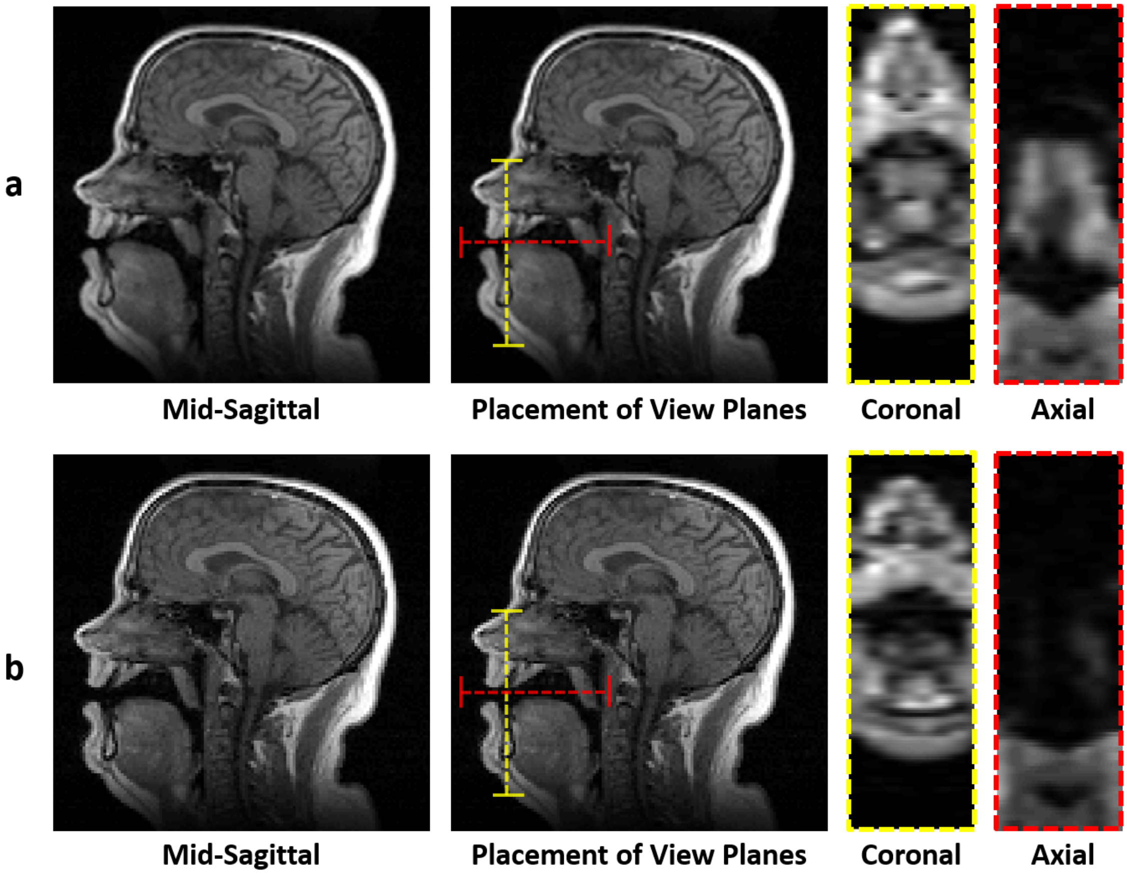


Figure 3.9: Visualization of sound production at the mid-sagittal, coronal and axial planes: (a) the production of /l/ in the /la/ sound: at this time point, the tongue comes into full contact with the alveolar ridge and its gesture is well-captured across all view planes; (b) the production of /a/ in the /la/ sound. When the tongue retracts to its resting position and continues to the /a/ sound, its gesture is reflected in coronal and axial planes as darkened intensity. The placements of coronal and axial planes are indicated with yellow and red colors, respectively.

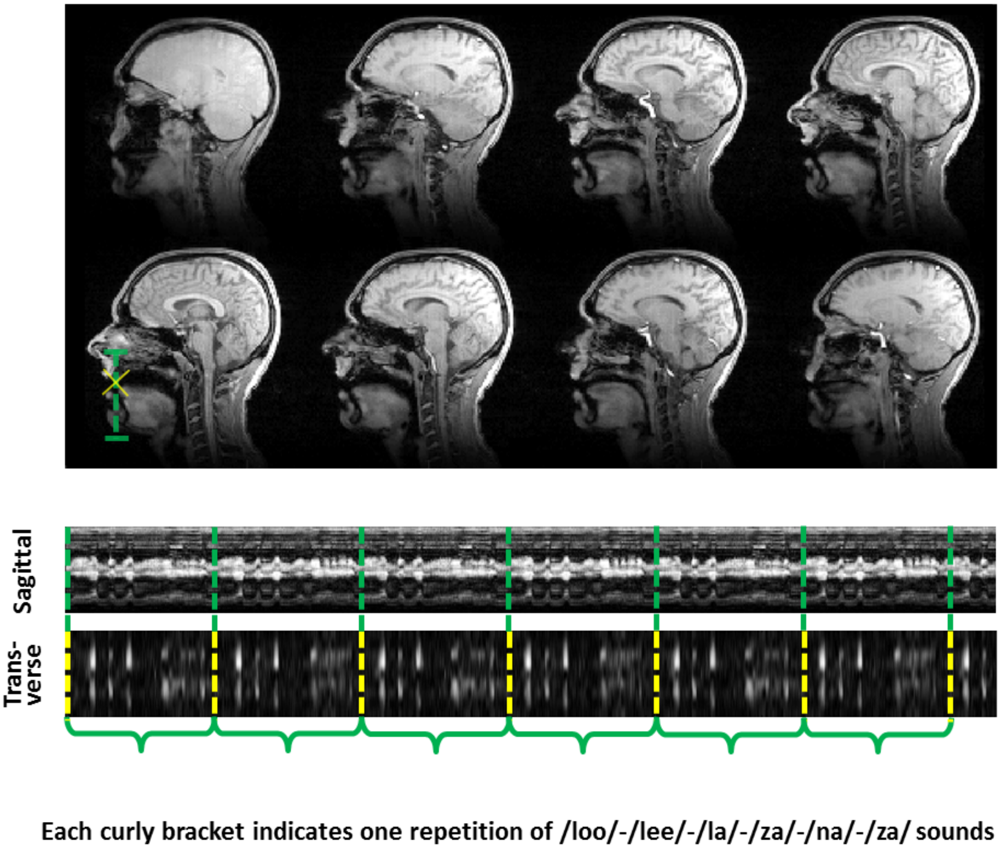

Sagittal
Trans-
verse
Each curly bracket indicates one repetition of /loo/-/lee/-/la/-/za/-/na/-/za/ sounds

Figure 3.10: Temporal dynamics of tongue motion from an imaging volume covering the entire vocal tract. The temporal dynamics are taken from a line segment on a central mid-sagittal slice and are plotted as opposed to time. Sharp temporal transitions of tongue motion are well-captured and presented. The bottom rows demonstrate the dynamics along a line segment from the mid-sagittal plane, as well as on a transverse plane placed beneath the alveolar ridge.

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 Section 3.6). As can be seen, the temporal profile from the 3D method displays sharper temporal transitions of the tongue motion compared with its 2D counterpart. 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.

3.7 Phonetics investigations

3.7.1 Phonetics investigations

Our method was first employed to systematically study the production of flaps, which is a challenging task to perform if only acoustic recordings are available. 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 – is known to be of particular phonetics interest [190]. Many existing methods in articulatory phonetics lack sufficient frame rate to capture the tongue postures associated with these brief events. In this dissertation, 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 [190]. These flaps are usually realized when an alveolar stop (the /t/ or the /d/) occurs intervocalically after a stressed syllable. Traditional phonological theories claim that these two flaps (the /t/ and the /d/) lose all distinction [191] 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 [192]. It is worth noting that this claim has only been demonstrated using acoustic evidence [193], without imaging evidence of the underlying articulatory differences.

In order to determine whether the articulation of these two flaps manifests any gestural differ-

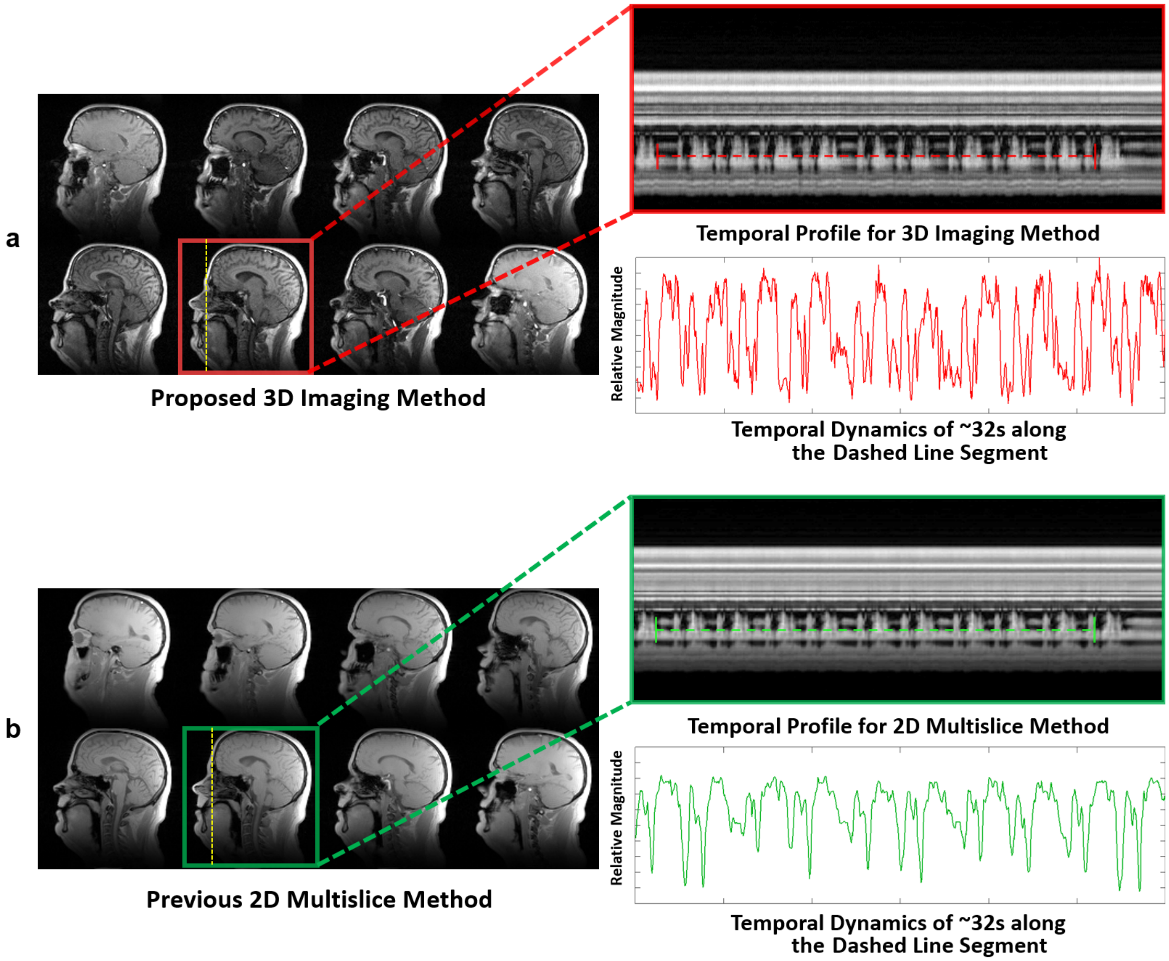

a
Proposed 3D Imaging Method
Temporal Profile for 3D Imaging Method
Relative Magnitude
Temporal Dynamics of ~32s along the Dashed Line Segment
b
Previous 2D Multislice Method
Temporal Profile for 2D Multislice Method
Relative Magnitude
Temporal Dynamics of ~32s along the Dashed Line Segment

Figure 3.11: Comparison of temporal dynamics from our 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 sharper temporal transitions are seen with the 3D reconstruction.

ence, these flaps were acquired 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 [194]) 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 the /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 phonetic 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 [12, 13, 184] was employed to measure the vertical distances between the tongue tip and alveolar ridge for the /t/ and the /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 [180] 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.

3.7.2 Investigation results

Figure 3.12 shows representative imaging and statistical results in the analysis of the /t/ and the /d/ flaps during the phrases "I said writing / riding to you." Figure 3.12a compares the respective tongue gestures between the two flaps. The /t/ flap has a slightly more superior tongue position compared with the /d/ flap. The averaged tongue tip - alveolar ridge distances is then quantitatively measured with a deformation-based method [12, 13, 184] for the /d/ and the /t/ flaps over a normalized duration (the original duration are ∼66 ms for the /d/ flaps and ∼54 ms for the /t/ flaps). As shown in Figure 3.12b, a larger averaged distance for the /d/ flaps is observed at the beginning of

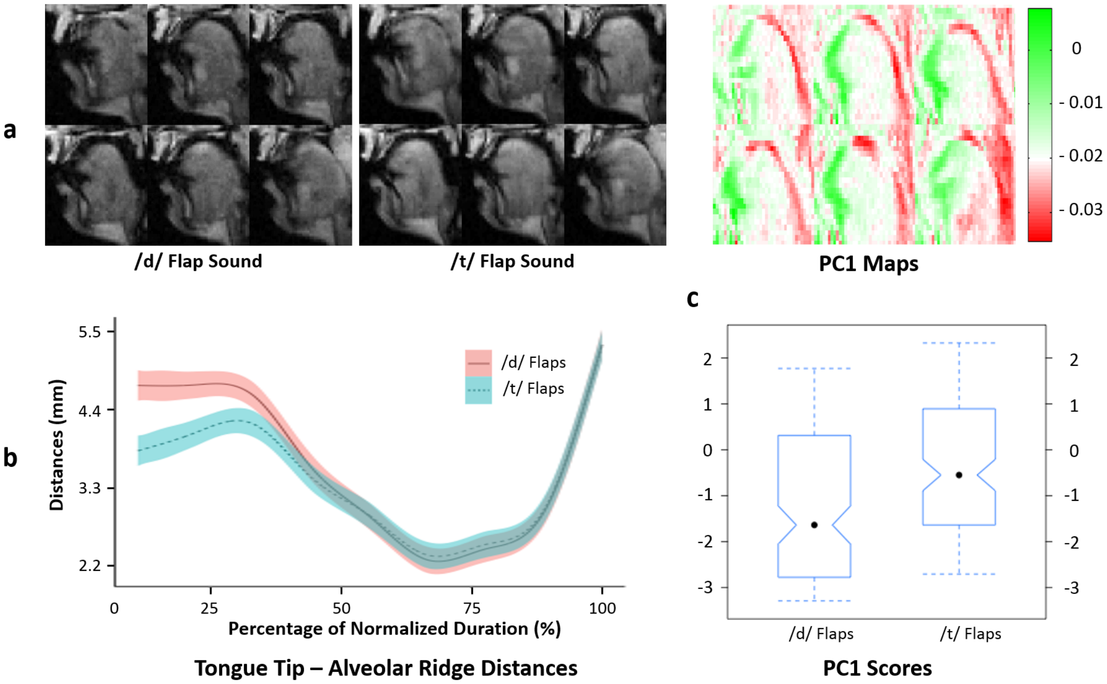


Figure 3.12: Mid-sagittal reconstructions and phonetics analyses of the production of American English flaps: (a) representative tongue gestures at the production of the /d/ and the /t/ flaps; (b) averaged distances between the tongue tip and the alveolar ridge for the /d/ and the /t/ flaps over a normalized 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 the /t/ flap than the /d/ flap sound (suggesting higher level of tongue apex protrusion and tongue blade elevation).

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 [180]. As seen with Figure 3.12c, 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 [180] in the tongue apex region during articulation. This result was statistically verified in a one-way ANOVA (F = 104.7, $p < 0.001$). The above analyses demonstrate the effectiveness of our method in analyzing sounds that are otherwise difficult to study from acoustic recordings.

Our method was additionally applied to analyze the level of nasalization during the production of a speech sample. Particular emphasis in the second phonetics investigation was placed on velopharyngeal activity. With regard to velopharyngeal activity, nasalization refers to the coupling between the nasal and oral cavities due to velar movement [195]. The level of velopharyngeal coupling is influenced by the fast interaction between the velum and the velopharyngeal aperture. Previous work has indicated the potential of dynamic MRI to examine the spatiotemporal variation in the aperture and provide imagery basis for phonetic assessments of nasalization level [196, 197]. However, analysis of velopharyngeal activity at this frame rate has not been possible.

French was chosen as the carrier language for the studies of nasalization. Unlike English, French contains nasal vowel sounds, oral vowel sounds and nasalization effects during speech. Therefore, it is an especially interesting language to study in relation to the coupling between the nasal and oral cavities to produce nasalized speech sounds, such as the alveolar nasal consonant as introduced in [10, 198–200]. This experiment focused on a dialect of French, northern metropolitan French, due to its prevalence. The carrier phrase was chosen as " Il retape $\mathbf{X}$ parfois," where $\mathbf{X}$ denotes articulation of three nasal vowels: /ɑ̃/, /ɛ̃/ or /ɔ̃/. A healthy female subject was recruited for the experiment, during which she was requested to produce the phrase recurrently. No other requirement was imposed on the subject's speaking pace. This experiment was conducted in accordance with an approved protocol through the Institutional Review Board at the University of Illinois at Urbana-Champaign.

Figure 3.13 shows images from multiple imaging planes from an *in vivo* experiment. Our method allowed opening sizes of the vocal tract to be simultaneously investigated. Specifically, these images focus on comparing vocal tract opening sizes for three nasal vowels, the /ɑ̃/, /ɛ̃/ and /ɔ̃/ sounds. In the anterior portion of the vocal tract, as illustrated with Figure 3.13a and Figure 3.13b, the /ɑ̃/ sound has the largest opening size while the /ɔ̃/ sound has the smallest. An opposite observation, as given in Figure 3.13c, is seen in the mid-posterior portion of the vocal tract, where /ɑ̃/ has the smallest opening size and /ɔ̃/ has the largest. Nearly identical opening sizes are observed in the inferior portion of the vocal tract, as given in Figure 3.13d. The results in Figure 3.13 demonstrate that our method provides sufficient spatial coverage to capture variation in vocal tract opening sizes at different spatial locations.

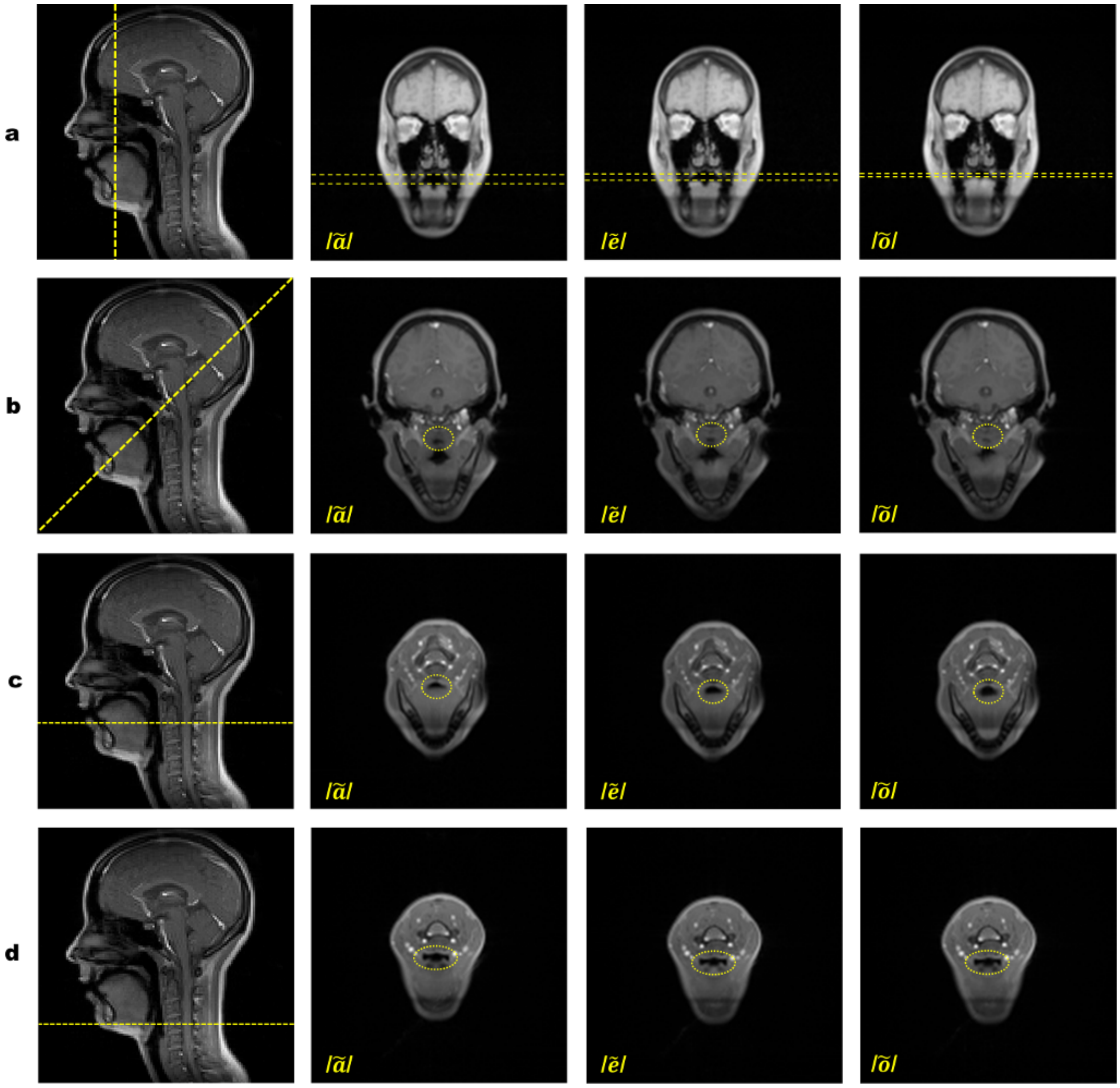


Figure 3.13: Comparison of multiple nasal vowels /ɑ̃/, /ɛ̃/ and /ɔ̃/ sounds: (a) /ɑ̃/ has the largest distance between the median portion of the tongue and the palate; (b) /ɑ̃/ has largest velopharyngeal opening size; (c) /ɑ̃/ has the smallest opening between the root of the tongue and the pharynx; and (d) three vowels have nearly identical opening between the epiglottis and the pharyngeal wall.

3.8 Discussion

Our method provides broad spatial coverage to capture speech movements for phonetic analysis. Sufficient spatial coverage is critical for phonetic studies that investigate the relations between multiple regions in the vocal tract. For the three nasal vowels sounds, traditional phonetic analysis of French language has taken for granted that the superior region of the vocal tract undergoes a greater level of motion compared with the inferior region. However, imaging evidence has been lacking. In this chapter, this hypothesis is verified by applying our method to examine movements at different levels of the vocal tract and the results are summarized in the results section. As can be seen, the opening sizes of these cavities vary significantly in the superior vocal tract but remain nearly identical in the inferior portion. The results confirm the above hypothesis and may suggest a pivot-like structure of muscle motion: the superior vocal tract may be more deformable while the inferior vocal tract may be more stable.

Our method improves imaging speed and spatial coverage by integrating a self-navigation scheme into 3D acquisition. 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 multi-slice acquisitions [51]. 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.

Our 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 [201], while a voiceless consonant (such as the /t/ flap) does not. This 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 lower back positioned vowel sound is realized for the /d/ flap. To verify this hypothesis, the associated spatial details (i.e., averaged tongue tip to alveolar ridge distances) are compared 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 by acoustic recordings. The results allow us to conclude that the /t/ and the /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 3.11, a rigorous measure of the true temporal resolution is challenging to define as it may vary from voxel to voxel and from 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 [202] for our method has a full width half maximum of 1.1 voxel and corresponds to a nominal frame rate of 151 fps.

Practical limits exist in the available acceleration from our acquisition strategy. Although a high nominal frame rate is achieved to potentially allow 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 this imaging protocol. Also, a practical limit in acceleration exists because the self-navigated scheme requires playing out a gradient rephaser prior to the navigator. Higher frame rate requires increased slew rates in the rephaser, which consequently induces eddy current that compromise the quality of the navigator. Careful control of gradient switching and slew rate limits, therefore, is essential to prevent the impact of eddy current on the accuracy of the estimated temporal subspace. In this chapter, the length of the rephaser was carefully chosen and limited to be 890 μs, and it has been observed in our preliminary studies that shorter duration results in eddy current artifacts as higher-temporal-frequency information spreads throughout the reconstructed image. Our method may benefit from further theoretical characterization and systematic validations of the eddy current effects.

Like many imaging methods that involve regularized reconstructions, the performance of our method is influenced by the selection of regularization parameters. In this chapter, our 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 our method is chosen based on the discrepancy principle as discussed in [203], while other alternative methods exists [204], and systematic evaluation of these methods will be an interesting further research topic.

Our 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 71,680 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 [205], but adaptation of these methods to such a large-scale optimization problem may not be trivial and is beyond the scope of this chapter.

3.9 Summary

High-frame-rate 3D full-vocal-tract dynamic speech MRI has been achieved by exploiting: (a) a novel 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. Our 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 phonetic studies on American English flaps and French nasalization.

CHAPTER 4

DYNAMIC SPEECH MRI WITH DEFORMATION ESTIMATION

Dynamic speech magnetic resonance imaging (MRI) has been considered as a promising technique for speech-related studies. In Chapter 3, the potential of dynamic speech MRI in capturing the structural and functional changes of the velopharyngeal region of the vocal tract has been demonstrated with simulation results, *in vivo* experiments and phonetics studies on American English flaps and French nasalization. This chapter presents an image reconstruction method that not only allows high-quality visualization of speech motion, but also quantitative description of the motion at the same time. In particular, Section 4.2 provides a brief review of the imaging model. Section 4.3 gives detailed description of our method. Section 4.4 describes the setup and the results from the numerical simulation. Section 4.5 provides details on the setup of the validation experiments and the phonetics investigations. Section 4.6 provides results from the validation experiments and the phonetics investigations. Discussion and summary on our method will be given at the end of the chapter.

4.1 Introduction

Dynamic speech magnetic resonance imaging has been considered as a promising technique for speech-related studies. Despite the great potential and wide application of dynamic speech MRI, its practical utility has still been limited by two obstacles: the intrinsic trade-offs between speed and resolution, and the lack of quantitative analysis on the reconstructed articulatory motions.

As discussed in Chapter 2 and Chapter 3, a desirable dynamic speech MRI technique should offer at least three properties to overcome the above two obstacles: First, it should provide sufficient imaging speed to capture temporal transitions of articulatory gestures. Second, the technique should have adequate spatial resolution to differentiate fine-scaled spatial features. Finally, the technique should be able to quantitatively analyze speech motion. In practice, there exist multiple ways to analyze the articulator gestures, with respect to a reference gesture, but among these methods it has been shown that the deformation field holds great potential in providing quantitative and objective descriptions of speech motion [206, 207]. In particular, if the deformation field is estimated with high spatial resolution, it can serve as a useful tool for representing and quantifying articulatory motion in both horizontal and vertical directions. Conventionally, deformation fields are extracted from reconstructed images as separate image registration steps [208, 209]. However, the accuracy of estimated motion relies heavily on that of the reconstructed images. Recent research has shown great potential in improving the accuracy by simultaneously estimating the image sequence and deformation field [210].

Expanding upon our earlier methods [50,51,74–76] and inspired by our earlier conference publication [184], this chapter presents a novel imaging method that realizes all the above-mentioned requirements. In particular, our method in this chapter simultaneously estimates a high-resolution dynamic image sequence and an associated dense deformation field. Both 2D and 3D dynamic speech MRI experiments were performed to demonstrate the effectiveness of our methods: the 2D dynamic speech MRI experiments were performed to achieve a spatial resolution of 2.2 mm $\times$ 2.2 mm with a nominal imaging speed of 100 fps, while the 3D dynamic speech MRI experiments were performed to achieve a spatial resolution of 2.2 mm $\times$ 2.2 mm $\times$ 5.0 mm, a nominal imaging speed of 166 fps and an imaging volume covering the entire upper vocal tract with 8 mid-sagittal slices without slice gap. This capability not only captures fine-scaled articulatory motion, but also enables quantitative characterization of motion through a precise deformation field, which shares an identical spatiotemporal resolution with that of the reconstruction. The practical utility of our method is systematically examined through *in vivo* experiments and a phonetics study on American English. While some results are demonstrated with 2D dynamic speech MRI experiment data, it should be noted that our method is a general imaging methodology and extends naturally to all 2D and 3D dynamic speech MRI applications.

4.2 Imaging model

The imaging model follows in general with that presented in Section 3.2. This section only provides a brief overview of the imaging model for the sake of brevity, while particular focus is placed on the incorporation of deformation-based methods into the imaging formulation in Section 4.3. The acquired data $d(\mathbf{k}, t)$ from the $(\mathbf{k}, t)$-space can be expressed as

$$d(\mathbf{k}, t) = \int_{\text{object}} I(\mathbf{r}, t) e^{-i2\pi\mathbf{k}\cdot\mathbf{r}} d\mathbf{r} + \eta(\mathbf{k}, t). \tag{4.1}$$

In practice, $d(\mathbf{k}, t)$ is often sampled below the Nyquist rate due to physical and physiological concerns, and the recovery of $I(\mathbf{r}, t)$ is often ill-conditioned. However, the partial separability model allows recovery of high-quality dynamic image sequence from sparsely sampled data by assuming that the spatiotemporal dynamics to be partially separable to the L^{th} order [43]. In addition, the PS model implies the rank of the Casorati matrix $\hat{\mathbf{I}}$ is upper bounded by L and allows the following factorization:

$$\mathbf{I} = \mathbf{U}\mathbf{V}, \tag{4.2}$$

where $\mathbf{U}$ is the spatial subspace and $\mathbf{V}$ is the temporal subspace. $I(\mathbf{r}, t)$ can be effectively recovered from highly undersampled data when both $\mathbf{U}$ and $\mathbf{V}$ are determined [43].

4.3 Image reconstruction

As presented in Section 3.2, the image reconstruction problem requires estimating both $\mathbf{U}$ and $\mathbf{V}$ from the acquired data. In general, $\mathbf{U}$ and $\mathbf{V}$ can be determined using either a combined approach or a separate approach [43, 48, 50, 51, 76, 188]. This section presents an approach that determines $\mathbf{U}$ and $\mathbf{V}$ separately from two individual steps: (a) singular value decomposition is first performed on the navigator data set to determine the temporal subspace, or matrix $\mathbf{V}$ [43] and (b) with a predetermined $\mathbf{V}$, $\mathbf{U}$ can be estimated by directly solving a least-squares problem [43].

Determination of $\mathbf{U}$ from direct least-squares fitting may lead to an ill-conditioned problem [48, 188], and sparsity constraints can be incorporated to regularize the reconstruction problem.

This section chooses to enforce a sparsity constraint on the deformed spatiospectral image $\mathbf{D}_{\text{def}}\{\mathbf{I}\}$, where $\mathbf{D}_{\text{def}}\{\cdot\}$ is a composite imaging operator consisting of two transforms: (1) a non-rigid spatiotemporal warping transform from $\mathbf{I}$ to a reference image $\mathbf{I}_{\text{ref}}$, and (2) a temporal Fourier transform applied on the deformed image $\mathbf{D}_{\text{ref}}\{\mathbf{I}\}$. The warping transform lies at the core of $\mathbf{D}_{\text{def}}\{\cdot\}$. This transform allows nonlinear image registration based on a deformation field, $\mathbf{\Theta}$, that represents the pixel-wise coordinate displacements between $\mathbf{I}$ and $\mathbf{I}_{\text{ref}}$ along both horizontal and vertical directions. Specifically, the warping transform is performed with a routine based on the demons algorithm [211–214]. The proposed deformation-based sparsity constraint is appropriate for dynamic speech MRI applications because the temporal profiles of the deformed image are more uniform than those on the original image. As a result, signal representations of the deformed images in the $(\mathbf{x}, f)$-domain are sparser than those of the un-deformed images [184, 210]. A conceptual illustration of the motivation behind this approach is given in Figure 4.1.

With the proposed deformation-based sparsity constraint, the image reconstruction problem can be written as

$$\hat{\mathbf{U}} = \arg\min_{\mathbf{U}} ||\mathbf{d} - \mathbf{E}\{\mathbf{F_S}\mathbf{U}\mathbf{V}\}||_2^2 + \lambda\, ||\mathbf{D}_{\text{ref}}\{\mathbf{U}\mathbf{V}\}||_1, \tag{4.3}$$

where $\mathbf{E}$ is a composite imaging operator that consists of sensitivity encoding across multiple receiver coils, as well as sparse sampling; $\mathbf{F_S}$ is a spatial Fourier transform matrix and λ is a regularization parameter. The formulation in Equation 4.3 is similar to the previous formulation in Equation 3.4 with a novel constraint that yields two-fold benefits - on the one hand, the proposed constraint provides stronger regularization, compared with that used in Equation 3.4, for the estimation of the spatial subspace $\mathbf{U}$ due to the sparser $(\mathbf{x}, f)$ spectrum; on the other hand, our formulation yields a high-resolution deformation field $\mathbf{\Theta}$ as an important by-product that allows quantitative description of the articulatory motion. An algorithm based on half-quadratic minimization with continuation procedures is applied to solve the above optimization problem. At each continuation step, it is noticed that the operator $\mathbf{D}_{\text{def}}\{\cdot\}$ provides an update of the deformation field $\mathbf{\Theta}$.

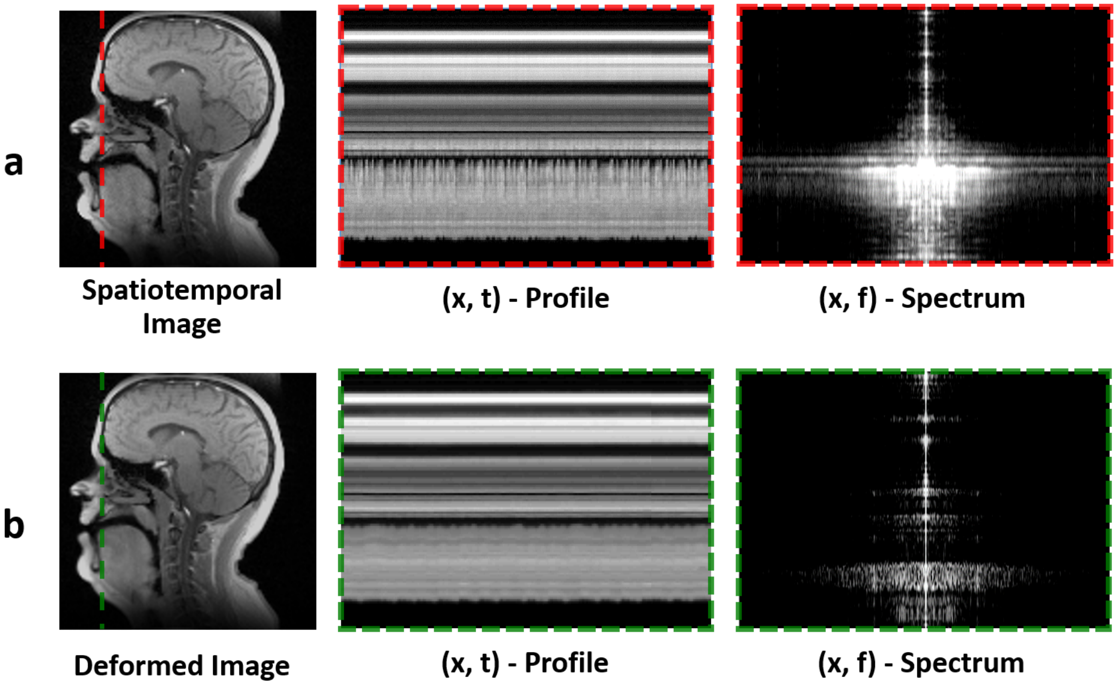


Figure 4.1: Comparison of the $(\mathbf{x}, t)$ profiles and $(\mathbf{x}, f)$ spectrum for the dynamic and deformed image sequences along vertical strips across the tip of the tongues: (a) the spatial position, $(\mathbf{x}, t)$ profile and $(\mathbf{x}, f)$ spectrum for the dynamic image sequence, (b) the spatial position, $(\mathbf{x}, t)$ profile and $(\mathbf{x}, f)$ spectrum for the deformed image sequence. Compared with the dynamic image sequence, the deformed one demonstrates more uniformed $(\mathbf{x}, t)$ profile and sparser $(\mathbf{x}, f)$ spectrum.

4.4 Simulation studies

4.4.1 Simulation setup

Numerical simulations were conducted to evaluate the performance of our method. Particular emphasis of the simulation study was placed on demonstrating the estimated deformation field from our method is accurate. Given that no other dynamic speech MRI method is capable of presenting the spatiotemporal motion and deformation at the resolution as provided in this dissertation, a numerical phantom needs to be developed to properly examine the performance of our method. It should be noted that the simulations performed in this dissertation yielded good empirical results, but 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 the specific data sets acquired from *in vivo* experiments.

A generic numerical phantom for 2D dynamic speech MRI has been created. Simulation studies have been performed to characterize the effectiveness of estimating the spatiotemporal motion and the spatiotemporal deformation field. 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 /za/ - /na/ - /za/ sounds at his own speaking rate. The created numerical phantom had a matrix size of 128 $\times$ 128 $\times$ 1, a FOV of 280 $\times$ 280 $\times$ 6.5 mm^3, a spatial resolution of 2.2 $\times$ 2.2 $\times$ 6.5 mm^3, a TR of 9.8 ms and a total number of 8960 time frames.

Simulated data acquisition followed the $(\mathbf{k}, t)$-space sampling strategy as described in this chapter. 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 [189]. 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. Image reconstruction and deformation field estimation from the simulated data were performed using our method and are not discussed here for the sake of brevity.

4.4.2 Simulation results

Figure 4.2 compares the estimated deformation field with regards to the ground truth deformation field. Specifically, each row in Figure 4.2 contains four images – the spatiotemporal reconstruction, the ground truth deformation field, the estimated deformation field and the residual deformation field (from left to right and top to bottom). The residual deformation field has been scaled up by a factor of 5 to assist visualization. Figure 4.2a shows the comparison when the tongue is elevated towards the alveolar ridge. Figure 4.2b shows the comparison when the tongue is positioned at a neutral position. As can be seen, the reconstruction faithfully represents the spatial details of the numerical phantom without significant spatial distortion. In addition, the yielded deformation field represents accurate estimation of the articulatory motion with respect to the reference image. This comparison demonstrates that our method is capable of both high-quality reconstruction and accurate estimation of the deformation field.

Figure 4.3 examines further the estimated deformation field. Specifically, each row in Figure 4.3 contains three images – the residual deformation field, the reference image and the resulting image with the residual deformation field applied to the reference image (from left to right). Figure 4.3a shows the comparison when the tongue is elevated towards the alveolar ridge. Figure 4.3b shows the comparison when the tongue is positioned at a neutral position. As can be seen, the estimated deformation field faithfully captures the articulatory motion such that the residual deformation field only causes minor spatiotemporal deformation on the reference image. This comparison demonstrates that our method is capable of estimating accurate deformation field.

4.5 Experimental studies

4.5.1 Validation experiments

To demonstrate the effectiveness of our method, *in vivo* experiments were performed on a Siemens Trio scanner (Siemens Medical Solutions, Erlangen, Germany) with a field strength of 3 T, a gradient strength of 40 $\mathrm{mTm^{-1}}$ and a maximum slew rate of 176 $\mathrm{Tm^{-1}s^{-1}}$. A 12-channel head receiver

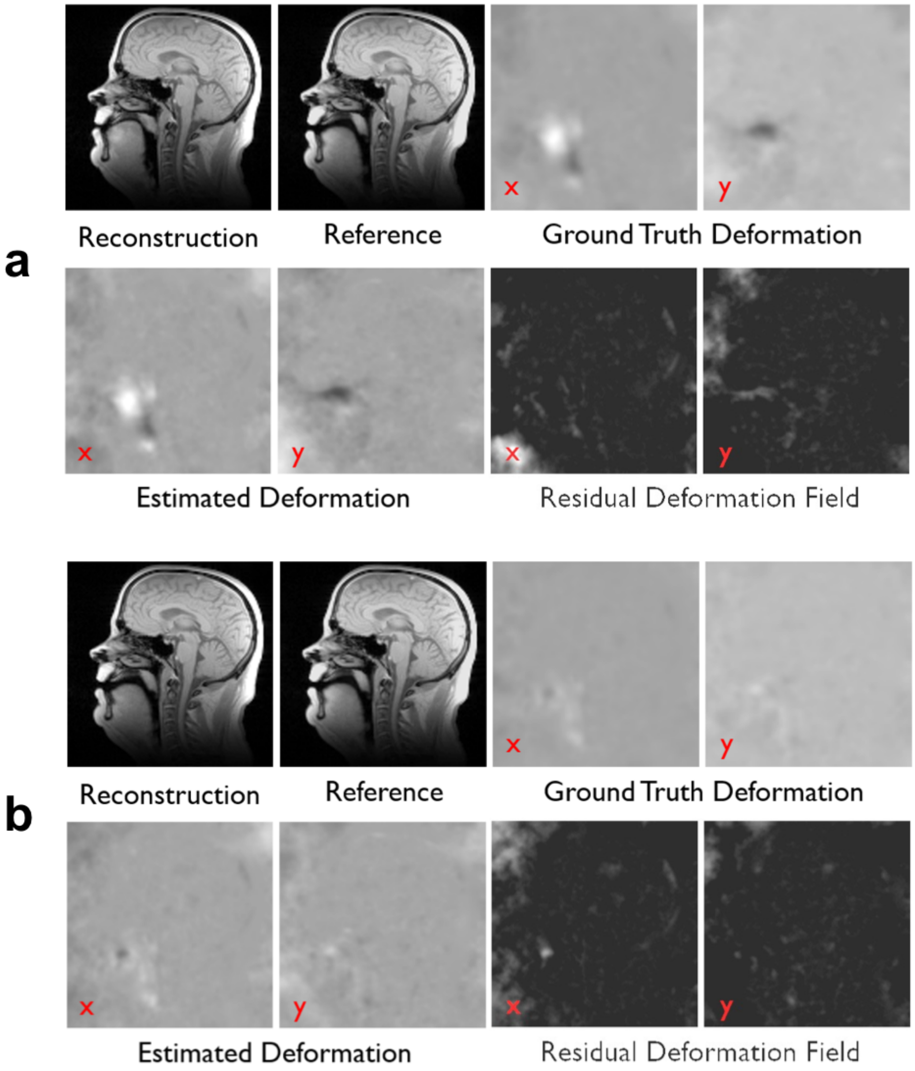


Figure 4.2: Comparison of the estimated deformation field with regards to the ground truth deformation field: (a) the spatiotemporal reconstruction, the ground truth deformation field, the estimated deformation field and the residual deformation field (from left to right and top to bottom) when the tongue is elevated towards the alveolar ridge. The residual deformation field has been scaled up by a factor of 5 to assist visualization; and (b) the spatiotemporal reconstruction, the ground truth deformation field, the estimated deformation field and the residual deformation field (from left to right and top to bottom) when the tongue is positioned at a neutral position. The residual deformation field has been scaled up by a factor of 5 to assist visualization. This comparison demonstrates that our method is capable of both high-quality reconstruction and accurate estimation of the deformation field.

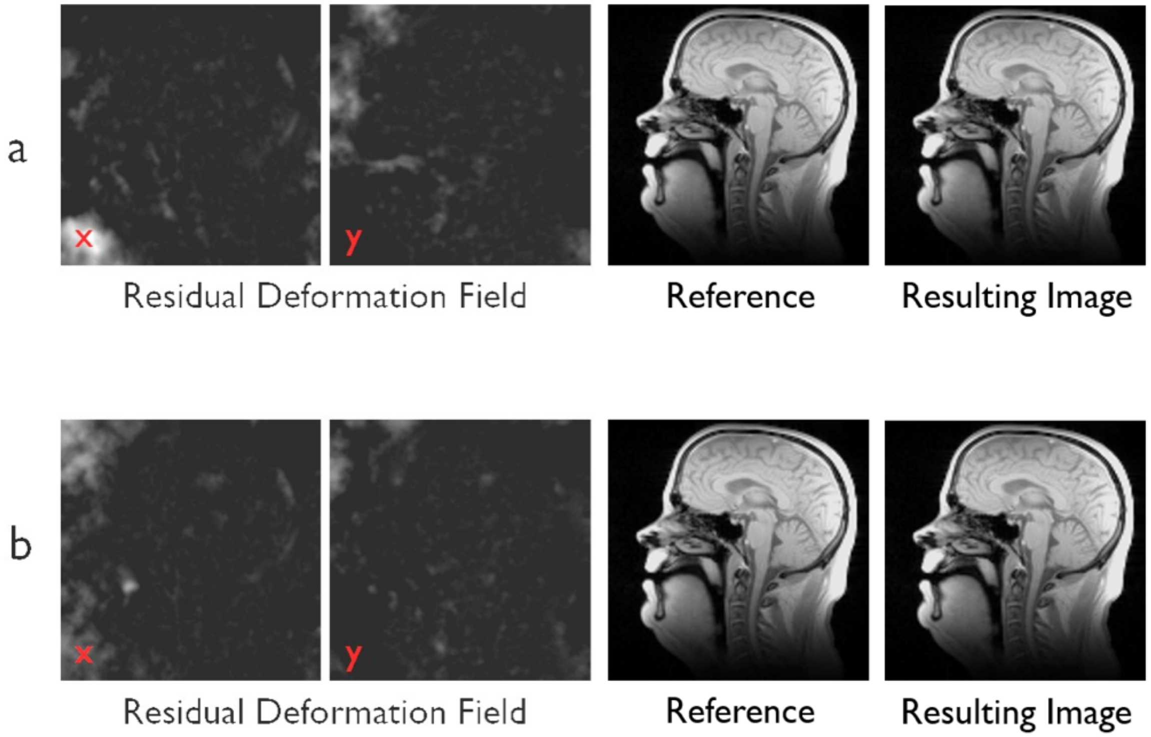


Figure 4.3: Examination of the estimated deformation field - the residual deformation field is applied to the reference image to demonstrate the level of residual deformation between the estimated deformation field and the ground truth: (a) the residual deformation field, the reference image and the resulting image with the residual deformation field applied to the reference image (from left to right) when the tongue is elevated towards the alveolar ridge. The residual deformation field has been scaled up by a factor of 5 to assist visualization; and (b) the residual deformation field, the reference image and the resulting image with the residual deformation field applied to the reference image (from left to right) when the tongue is positioned at a neutral position. This comparison demonstrates that our method is capable of estimating accurate deformation field.

coil was used to image the subject. The acquisition details are different for the 2D and the 3D dynamic speech MRI experiments, and are described separately as follows.

For 2D dynamic speech MRI experiments, a FLASH sequence was developed to interleave a spiral navigation acquisition and a Cartesian imaging acquisition with a TR of 9.78 ms. The navigation and imaging data acquisition had a TE of 0.85 ms and 2.3 ms, respectively. Due to the nonlinear and non-stationary nature of our formulation, a nominal frame rate of 100 fps (defined by 1/TR) is used to describe the imaging speed capturing speech motion and deformation. Other parameters were: acquisition matrix size = 128 $\times$ 128, FOV = 280 mm $\times$ 280 mm, spatial resolution = 2.2 mm $\times$ 2.2 mm and a slice thickness of 6.5 mm. When acquiring data that targets a model order of around 70, the acquisition time was 1 min 42 s.

For 2D dynamic speech MRI experiments, five volunteers participated in the validation experiments. These subjects had an age range of 23 to 38 and were requested to produce /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ recurrently at their natural speaking pace. The voices of these subjects were simultaneously recorded at a sampling rate of 22 kHz through a fiber-optic microphone with active noise cancellation (Dual Channel FOMRI, Optoacoustics, Or Yehuda, Israel). Informed consent was obtained for all subjects and the experiment was carried out in accordance with protocols from the Institutional Review Board at the University of Illinois at Urbana-Champaign.

For 3D dynamic speech MRI experiments, the self-navigation strategy was applied to acquire data. A FLASH sequence has been developed to acquire data with the following parameters: a TR of 5.99 ms, a 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, the acquisition time was 7 min 12 s. With our 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). Two volunteers participated in the 3D dynamic speech MRI validation experiments, the procedure of which followed largely with the 3D data acquisition strategies discussed in Chapter 3. For brevity of this dissertation, the details of the experiments are omitted but the interested readers are encouraged to refer to Chapter 3 for more details.

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 the 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 subject's head in the receiver coil with foam pads. Informed consent was obtained for all subjects and the experiment was carried out in accordance with protocols from the Institutional Review Board at the University of Illinois at Urbana-Champaign.

4.5.2 Phonetics investigations

To examine the practical utility of our method, the method was applied to a phonetics study to examine the differences in articulatory gesture of oral vowels and their nasal counterparts. Particular interest was placed on the level of pharyngeal constriction in front and back vowel congeners in American English. Traditionally, the measurement of constriction was obtained by manual analysis on low-temporal-resolution images, which is challenging and tedious especially for the analysis of image sequences that have thousands of image frames. As a result, automatic analysis on high-temporal-resolution image sequence has long been lacking.

Given our method, a natural phonetics question to ask is: is it possible to enable automatic tracking and measuring the level of constriction by leveraging the deformation field? Starting from this question, imaging data were collected from a volunteer speaker of American English using both 2D and 3D dynamic speech MRI protocols. The targeted sounds for phonetics investigation were /a/, /u/, /e/ vowels as opposed to the longer /a$\sim$/, /u$\sim$/ and /e$\sim$/ vowels in American English. Other details of data acquisition were carried out following the identical procedures as described in the previous section and are omitted in this subsection for brevity.

The reconstructed spatiotemporal image and deformation fields were analyzed using an approach proposed in [184]. In particular, the contours pertaining to the tongue root and the posterior mediopharyngeal wall were first extracted from the reference image. Using this contour information, together with the associated deformation field, the positions of the tongue root and those of

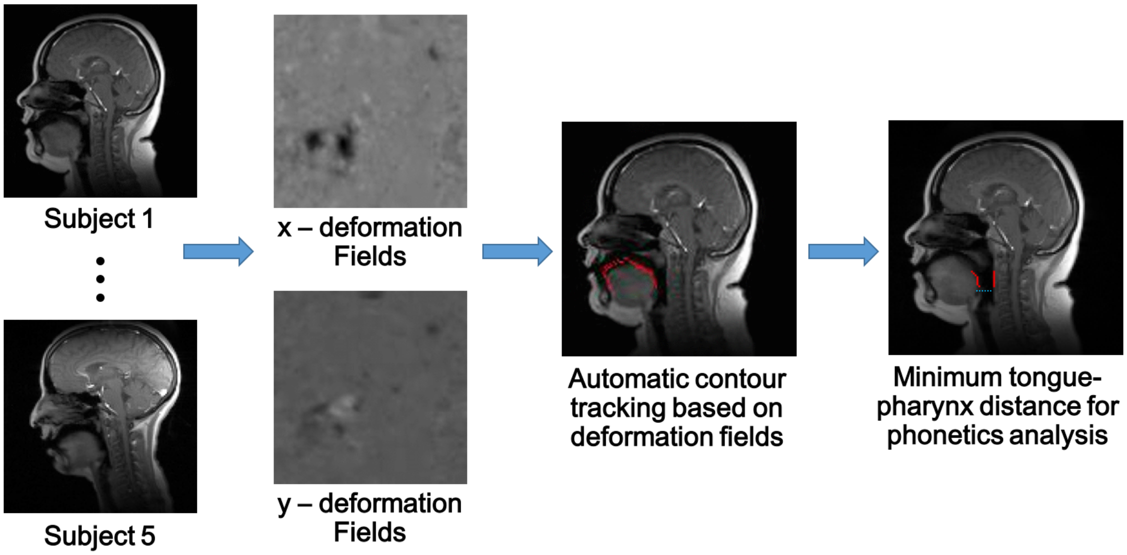

Subject 1
Subject 5
x – deformation Fields
y – deformation Fields
Automatic contour tracking based on deformation fields
Minimum tongue-pharynx distance for phonetics analysis

Figure 4.4: Demonstration of automatic constriction measurement: The contours pertaining to the tongue root and the posterior mediopharyngeal wall were first extracted from the reference image. Using this contour information, together with the associated deformation field, the positions of the tongue root and those of the posterior pharyngeal wall at each temporal frame were calculated. The calculated positions were further converted into distances between the tongue root and the pharyngeal wall for linguistic analysis.

the posterior pharyngeal wall at each temporal frame were calculated. The calculated positions were further converted into distances between the tongue root and the pharyngeal wall for linguistic analysis. Figure 4.4 illustrates the essential steps of automatic constriction measurement.

4.6 Experimental results

4.6.1 Validation experiments

Figure 4.5 shows the reconstruction of a mid-sagittal slice and its associated deformation field from an *in vivo* experiment. Articulatory gestures in producing /ee/ during the /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ sounds are shown with fine spatial details in Figure 4.5a. The reference image, with

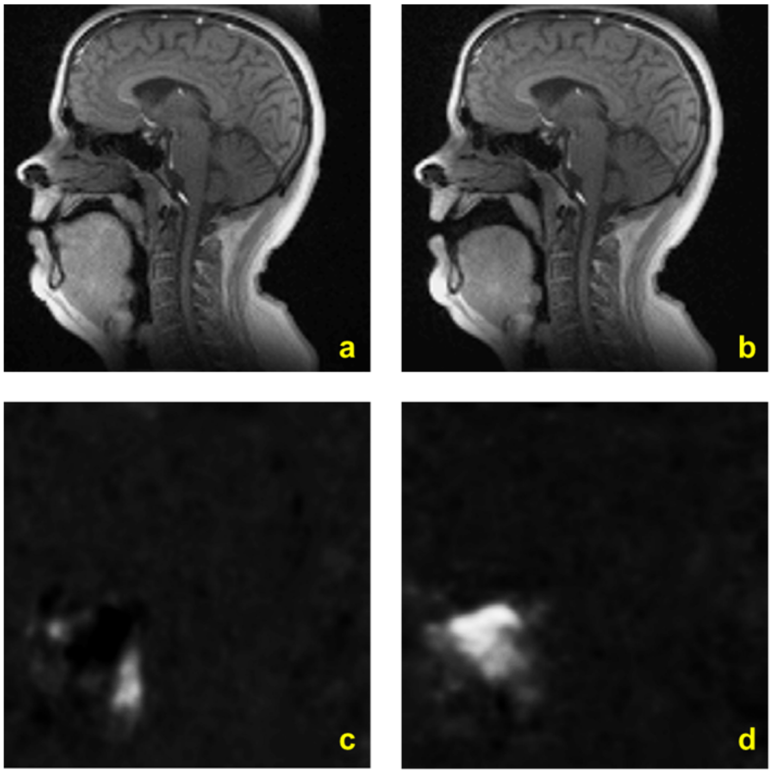


Figure 4.5: Demonstration of the reconstructed articulatory gesture and the associated deformation field: (a) the articulatory gesture during the production of the /ee/ sound in the /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ sounds; (b) the reference image used for deformation estimation; (c) the estimated deformation field in the horizontal direction; (d) the estimated deformation in the vertical direction.

respect to Figure 4.5a, is obtained from the identical subject at a representative neutral position and is shown in Figure 4.5b. Compared with the reference image in Figure 4.5b, it is obvious that the bulk tongue body in Figure 4.5a elevates towards the roof of the mouth in the vertical direction, with slight horizontal translation at the tongue tip and tongue root. This trend of tongue motion is faithfully capture by the deformation fields: the deformation field in the horizontal direction is shown in Figure 4.5c and the deformation field in the vertical direction is shown in Figure 4.5d.

Figure 4.6 shows the reconstruction and the temporal profiles from an 3D *in vivo* experiment where the subject was requested to produce the /ai/ - /au/ - /wi/ - /a/ - /u/ - /i/ sounds. The reconstructed articulatory gesture of the /a/ sound from the entire upper vocal tract is free from significant artifacts and is shown in Figure 4.6a. The temporal profile along a vertical strip across the tip of the tongue taken from the fourth mid-sagittal slice (indicated as a green dashed line) is shown in Figure 4.6b. As is seen, the temporal profile in Figure 4.6b captures the temporal dynamics of the contact between the tip of the tongue and the alveolar ridge with fine temporal details. The temporal profile along a horizontal strip across the soft palate (velum) from the fourth mid-sagittal slice (indicated as a red dashed line) is shown in Figure 4.6c. The speech motion along the horizontal strip, including the motion of the lips and that of the soft palate, are free from significant temporal blurring.

Figure 4.7 shows the reconstruction from mid-sagittal imaging planes and the re-sliced coronal imaging planes from a 3D *in vivo* experiment where the subject was requested to produce the /ai/ - /au/ - /wi/ - /a/ - /u/ - /i/ sounds. The reconstructed articulatory gestures of the /a/ sound are shown from eight consecutive mid-sagittal slices in Figure 4.7a without any slice gap. Re-sliced reconstructions are shown on Figure 4.7b covering five coronal imaging planes. These coronal images progress from the front end of the vocal tract towards the root of the tongue (indicated with five different colors) and are re-sliced based on the same reconstruction in Figure 4.7a.

Figure 4.8 examines the accuracy of the estimated deformation field by measuring the difference between the deformed image and the reference. Specifically, the comparison was performed between two methods: (a) the PS-sparse and (b) our method. The deformed image for the PS-sparse method is estimated as a separate step after the image reconstruction. The deformed image for our method is jointly estimated with the reconstruction itself according to our formulation. The differences between the deformed images and the reference image are shown in Figure 4.8c and Figure 4.8f, respectively for PS-sparse and our method. As can be seen, our method yields smaller difference at the back of the velum (as indicated with yellow arrows in the zoomed-in view inside the dashed green box). This suggests that the simultaneously estimated deformation field allows more accurate description of articulatory motion (because its deformed image better matches with the reference), especially compared with a separately estimated deformation field that is prevalent in many phonetics studies.

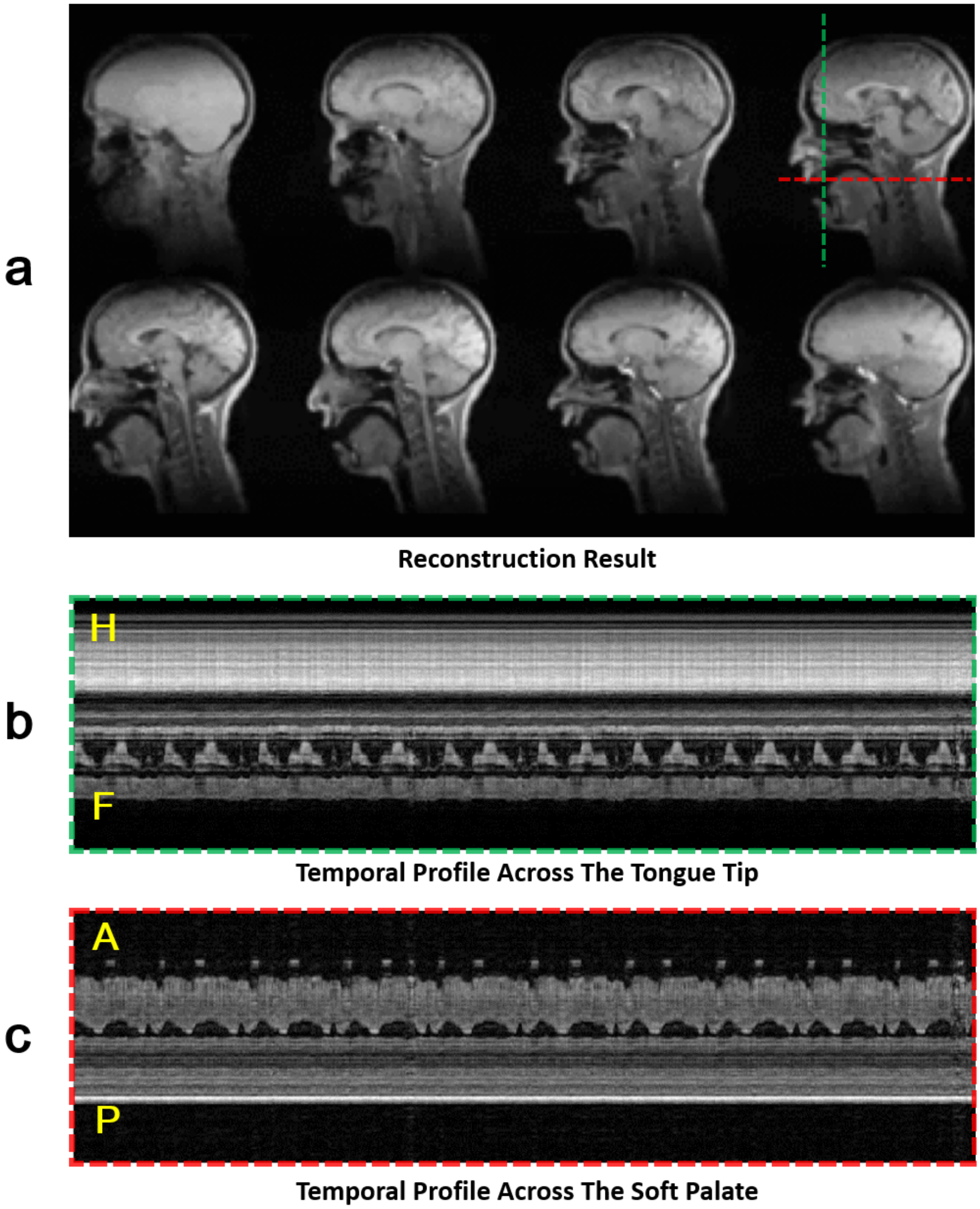

a
Reconstruction Result
b
H
F
Temporal Profile Across The Tongue Tip
c
A
P
Temporal Profile Across The Soft Palate

Figure 4.6: Reconstruction results and temporal profiles from a 3D *in vivo* experiment where the subject was requested to produce the /ai/ - /au/ - /wi/ - /a/ - /u/ - /i/ sounds: (a) the reconstructed articulatory gesture of the /a/ sound from the entire upper vocal tract; (b) temporal profile along a vertical strip across the tip of the tongue taken from the fourth mid-sagittal slice (indicated as a green dashed line); (c) temporal profile along a horizontal strip across the soft palate (velum) from the fourth mid-sagittal slice (indicated as a red dashed line).

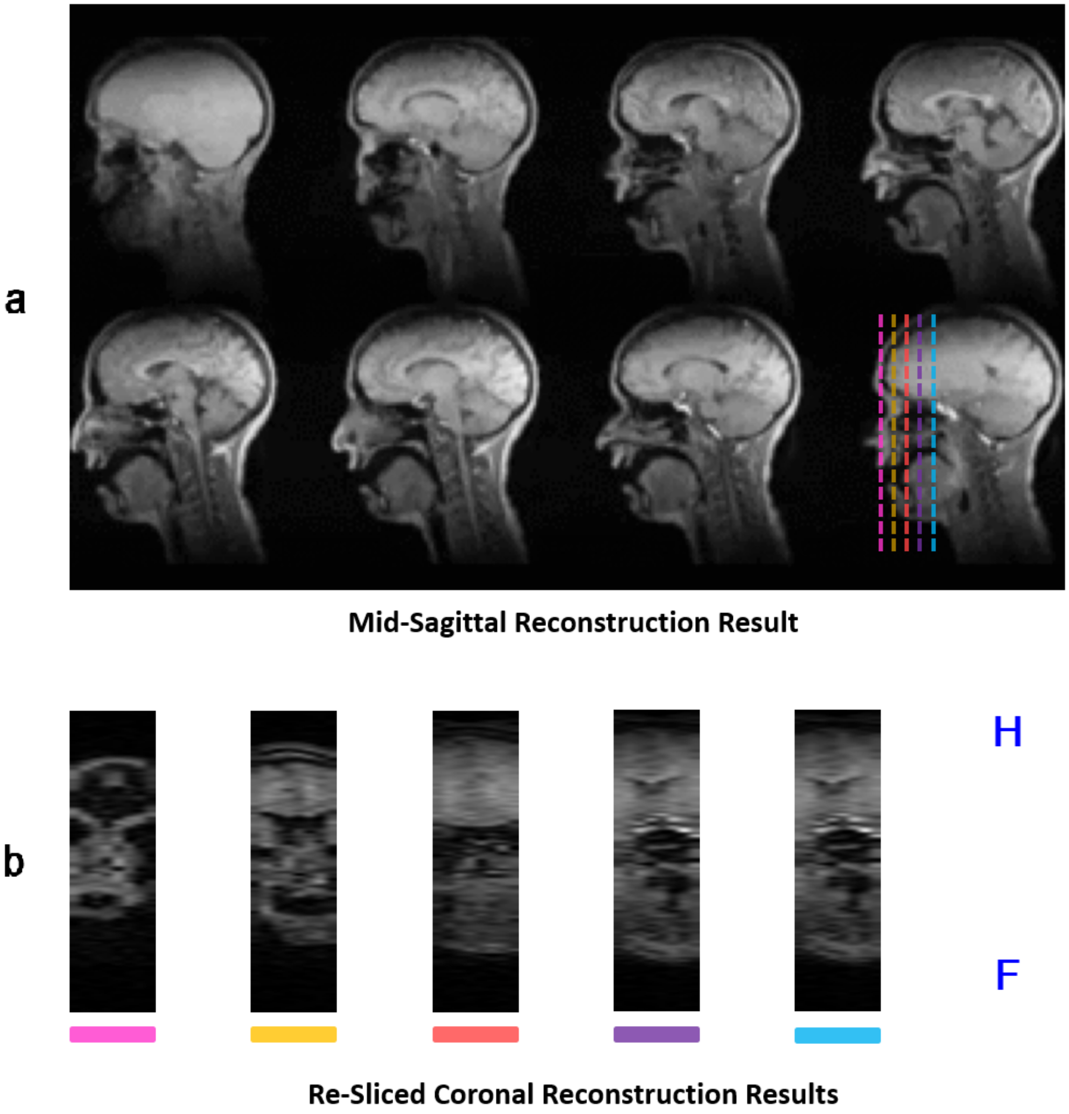

a
Mid-Sagittal Reconstruction Result
b
H
F
Re-Sliced Coronal Reconstruction Results

Figure 4.7: Reconstruction results from mid-sagittal imaging planes and the re-sliced coronal imaging planes from a 3D *in vivo* experiment where the subject was requested to produce the /ai/ - /au/ - /wi/ - /a/ - /u/ - /i/ sounds: (a) the reconstructed articulatory gesture of the /a/ sound from eight consecutive mid-sagittal slices without slice gap; (b) Re-sliced reconstructions on five coronal imaging planes that progress from the front end of the vocal tract towards the root of the tongue (indicated with five different colors).

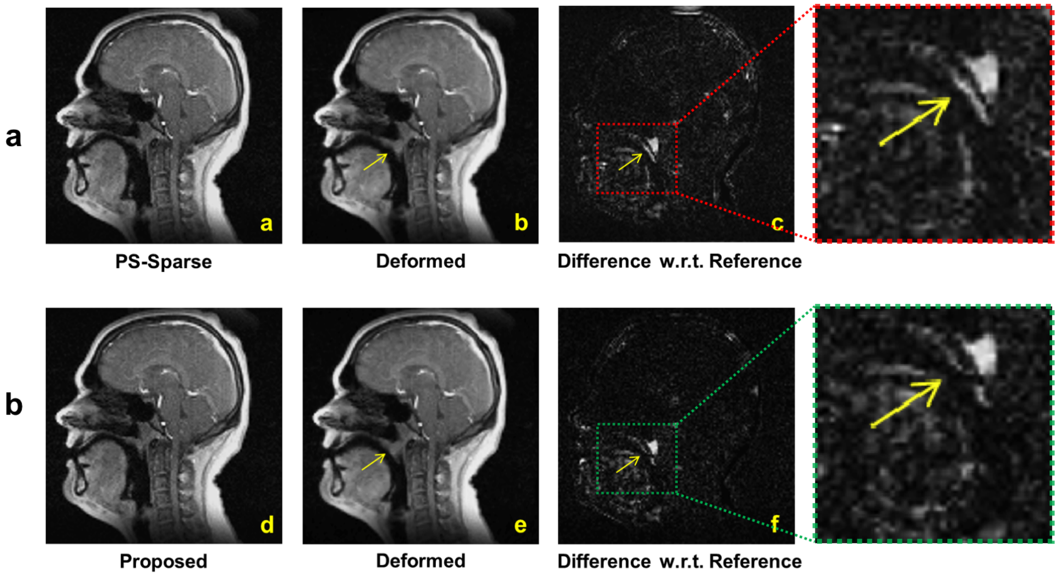


Figure 4.8: Comparison on the accuracy of the estimated deformation field by measuring the difference between the deformed image and the reference. Specifically, the comparison was performed between two methods: (a) PS-sparse and (b) proposed method. The deformed image for the PS-sparse method is estimated as a separate step after the image reconstruction. The deformed image for our method is jointly estimated with the reconstruction itself according to our formulation. The differences between the deformed images and the reference image are shown in (c) and (f), respectively. Our method yields better deformed images with reduced difference (as shown in the zoomed-in view inside the dashed green box) and therefore suggests better estimation of the articulatory motion.

4.6.2 Phonetics investigations

With the capability to jointly estimate dynamic image sequences and deformation fields, our method opens up a range of opportunities in quantitatively analyzing articulatory motion. Figure 4.9 shows obvious difference in the minimum distances between the tongue root the and posterior pharyngeal walls between the /a/ and the /a$\sim$/ sounds (t = 15.426, p $\lesssim$ 0.0001), with the oral vowel having a wider pharyngeal opening than the nasal vowel. Also, there is obvious difference between /u/ and /u$\sim$/ sounds (t = 17.6502, p $\lesssim$ 0.0001), but with the nasal vowel having a wider pharyngeal aperture. There is less difference between the /e/ and the /e$\sim$/ sounds in pharyngeal distance (t = - 0.66, p = 0.5). This is an indication of enhanced F1 differences in the non-front vowels F1 is predicted to rise for the /a$\sim$/ sound in comparison with the /a/ sound, and to lower for the /u/ sound in comparison with the /u$\sim$/ sound. Furthermore, these results are in line with previous findings on oral articulation of oral and nasal vowel pairs in Brazilian Portuguese [9, 183, 184], indicating that the distinctions between oral and nasal vowel articulation are achieved by more complex speech motion than the canonical view of nasalization through velum lowering, which involve multiple sections of the vocal tract.

4.7 Discussion

Our method is capable of simultaneously offering high-spatiotemporal-resolution dynamic image sequences and the associated dense deformation fields for visualizing, analyzing and quantifying articulatory motion. Compared with its counterpart that estimates deformation fields as separate image registration, our method allows more accurate motion analysis. Although the practical value of our method has been demonstrated in the *in vivo* experiments, several aspects may affect the quality of reconstruction and computation efficiency. These aspects include the selection of model order, regularization parameters and the computational requirement.

Our method requires the experimenter to determine an appropriate model order L. In this chapter, L was chosen empirically based on visual inspection of reconstruction quality including the discernibility of small-sized articulators, the level of temporal blurring during frame-to-frame

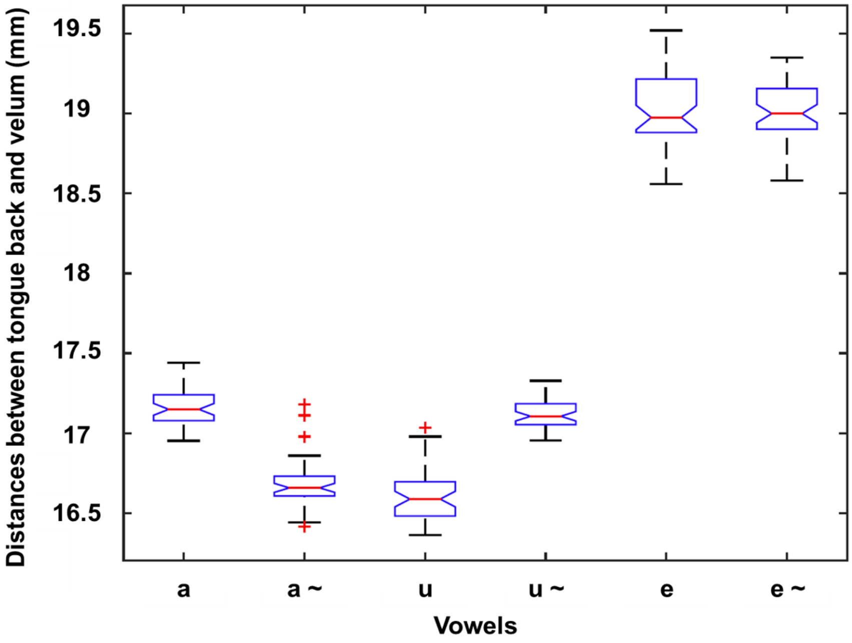


Figure 4.9: Minimum tongue-velum distances for oral (the /a/, /u/, /e/ sounds) and nasal (the /a∼/, /u∼/, /e∼/ sounds) vowel pairs. The statistics in the comparison are summarized in the box plot and the outliers are indicated with red crosses. As is seen, obvious difference is observed in the minimum distances between the tongue root the and posterior pharyngeal walls between the /a/ and the /a∼/ sounds. There is also obvious difference between the /u/ and the /u∼/ sounds, while there is less difference between the /e/ and the /e∼/ sounds in pharyngeal distance.

transitions, and the overall readability of the reconstructed articulatory motion. Based on these features, an L of 65 was chosen in this chapter and it has consistently yielded good empirical results for the reported experiments. It should be noted, however, that there exist other quantitative-metric-driven approaches [215–217] to guide the selection of L, although integration of these approaches and systematic evaluation of their performance will be out of the scope of this dissertation.

Our method requires selection of the regularization parameter λ. In the reconstructions presented in this chapter, λ is also chosen based on visual inspection of image quality based on motion fidelity and temporal blurring. Specifically, a $\lambda = 4.67 \times 10^{-3}$ was chosen for the above reconstructions. However, it is worth noting that our method provides relatively robust performance over a range of λ values. This robustness mainly results from the complementary role of signal model and the proposed constraint. In addition, there exist some theoretical results to guide the selection of regularization parameters [203]. The integration of these methods into our method will be systematically investigated in further studies.

Our method may pose a computational burden in its clinical applications due to the high dimensional optimization problem and the nonlinear registration routine involved. For instance, the reconstruction time for a total of 8960 2D time frames at a nominal speed of 100 fps (corresponding to a 1 min 29 s 2D dynamic speech MRI scan) from 12 receiver channels was around 1.5 hour on a 24-core 128GB-memory SUN workstation without code optimization. Fortunately, acceleration in computation time can be realized by taking advantage of a decoupled computation structure that decomposes a large-scale reconstruction problem into separate smaller ones along one spatial dimension [51]. In addition, fast solvers for key arithmetic components have been implemented for the proposed problem based on graphical processing units [205, 218, 219], although code optimization for computational efficiency is beyond the scope of this chapter.

4.8 Summary

In summary, this chapter presents a novel method for high resolution dynamic speech MRI with the capability to quantitatively describe articulatory motion. Our method enables simultaneous estimation of a high-resolution dynamic image sequence and a high-resolution deformation field.

We performed both 2D and 3D dynamic speech MRI experiments to demonstrate the effectiveness of our methods: the 2D dynamic speech MRI experiments were performed to achieve a spatial resolution of 2.2 mm × 2.2 mm with a nominal imaging speed of 100 fps, while the 3D dynamic speech MRI experiments were performed to achieve a spatial resolution of 2.2 mm × 2.2 mm × 5.0 mm, a nominal imaging speed of 166 fps and an imaging volume covering the entire upper vocal tract with 8 mid-sagittal slices without slice gap. The practical utility of the method has been validated through *in vivo* experiments and a phonetics study on American English. Our method serves as a promising tool for visualizing and characterizing articulatory motion in speech-related studies.

CHAPTER 5

SPATIOTEMPORAL ATLAS–BASED DYNAMIC SPEECH MRI

As introduced in Chapter 3 and Chapter 4, dynamic speech MRI has been recognized as a promising method for visualizing the articulatory motion of speech in both scientific research and clinical applications. However, precise characterization of the gestural and acoustical properties of the vocal tract remains a challenging task for dynamic speech MRI because it simultaneously requires: (1) reconstructing high-quality spatiotemporal images from highly undersampled data; (2) improving the quality of the reconstruction by incorporating strong prior knowledge that genuinely reflects the underlying articulatory dynamics; and (3) quantitatively interpreting the reconstructed images that contain great motion variability. This chapter aims to propose a novel imaging method that simultaneously meets these three requirements by integrating a spatiotemporal atlas into a partial separability model-based imaging framework. Through the novel use of an atlas-driven imaging model, our method in this chapter is capable of capturing high-quality articulatory dynamics simultaneously at high imaging speed and at high spatiotemporal resolution. Moreover, our method enables quantitative characterization of variability of speech motion (comparison of the reconstructed articulatory dynamics with respect to the generic motion pattern across all subjects) through the spatial residual components and the reconstructed sparse components.

5.1 Introduction

Dynamic speech magnetic resonance imaging (MRI) has demonstrated its usefulness for both scientific research and clinical studies in speech. The interested readers are encouraged to refer to Chapter 3 for high-resolution full-vocal-tract 3D dynamic speech MRI and Chapter 4 for high-resolution dynamic speech MRI with deformation estimation. The utilization of dynamic speech

MRI to capture structural and functional changes of the vocal tract has led to a range of exciting applications, such as studying complex soft-tissue geometries [1], detecting structural defects and motor dysfunctions [4], as well as understanding articulatory motion related to language formation and variation [8–14]. The interested readers are also encouraged to refer to Chapter 2 for a detailed review of a variety of application of dynamic speech MRI to speech-related studies.

Driven by clinical needs and scientific curiosity about the fast moving structures in the oral and pharyngeal regions, many recent studies of speech are in pursuit of dynamic speech MRI techniques that are capable of high imaging speed and high spatial resolution. This pursuit, however, in turn results in dynamic imaging data sets with high spatiotemporal dimensions and great motion variability, which consequently creates challenges for effective clinical interpretation of the spatiotemporal reconstructions. As a result, a critical need for the speech-imaging community remains a practical dynamic speech MRI method that simultaneously fulfils three requirements: high imaging speed, high spatial resolution and accurate characterization of the underlying articulatory motion.

Recent developments in dynamic speech MRI techniques have led to dynamic images that contain great variation in articulatory motion. For instance, our approach in Chapter 3 has enabled capturing articulatory dynamics over a temporal interval of as short as 6 ms across an imaging volume covering the entire upper vocal tract. Along with the fast growing capability in dynamic speech MRI, however, imaging methods that provide simultaneous high-quality spatiotemporal reconstruction and accurate analysis of the articulatory dynamics have been developing slowly. The approach to address this need would allow the extraction of the subject-specific articulatory dynamics that sits on top of the "standard" speech motion associated with a carrier phrase, and consequently holds great promise for simultaneously providing high-quality reconstruction. It is worth noting that recent developments in the low-rank plus sparse model have demonstrated certain potential in answering this question [220–228] and have shown improvements in several MR imaging applications [220–222]. However, this low-rank plus sparse model has never been employed to reconstruct highly undersampled dynamic speech MRI data, neither has it been used to show any potential in dynamic speech imaging. One hypothesis of this chapter is that the idea of the low-rank plus sparse model can be improved and incorporated into the partial separability model-based imaging framework to both improve reconstruction quality and facilitate analysis of the articulatory motion.

Developments in dynamic speech MRI techniques have also led to significantly increased spatiotemporal dimensions that prevent effective data analysis. As introduced in Chapter 3 and Chapter 4, a better spatial resolution and a larger number of time frames from advanced dynamic speech MRI methods not only bring unprecedented capability to visualize speech motion, but also lead to greater difficulty to properly analyze the reconstruction. Although dynamic speech MRI experiments have be performed over a relatively large subject population [9, 183], few clinical speech analysis tools exist in general to provide objective and effective interpretation over a large number of data sets. In order to allow objective and effective analysis of generic speech motion, a "standard" is required to represent the generic temporal motion pattern from all the groups of interest. This "standard" serves as a metric upon which the reconstructed spatiotemporal dynamics are compared and analyzed in terms of articulatory pattern and motion variability. However, it is not an easy task to construct a "standard": the spatial features and the contrast properties in this image sequence should be in register with those in the reconstructions in order to minimize the analysis error. As a result, the construction of such an image sequence has long been considered difficult.

The possibility to meet the above stringent requirements has only been opened up recently - a technique has been proposed to construct a high-quality spatiotemporal atlas from a group of subjects [229–231]. In particular, this technique allows high-temporal-resolution articulatory dynamics and high-spatial-resolution anatomical features to be preserved during the atlas construction steps [229–231]. This high-quality atlas-creation technique, and its integration with model-based spatiotemporal imaging methods, holds great promise for a better dynamic speech MRI method.

Inspired by the developments in Chapter 3 and Chapter 4 and extending upon the previous conference publications [232, 233], this chapter proposes a novel dynamic speech MRI method that leverages a spatiotemporal atlas to simultaneously allow high imaging speed, high spatial resolution and accurate characterization of the underlying articulatory motion. Of particular novelty are the low-rank plus sparse model and the integration of the atlas into dynamic speech MRI through a spatiotemporal sparsity constraint. The low-rank plus sparse model allows separate modeling of the generic speech motion and the subject-specific articulatory movement, which not only solves the image reconstruction problem, but also improves image reconstruction quality. The integration of a spatiotemporal atlas into the imaging formulation allows the dynamic articulatory motion to be compared against the atlas in the form of spatial residual components. Further, the proposed model enables accurate comparison of motion at an identical spatiotemporal resolution (of the reconstruction) either on a single or across multiple subjects. The practical utility of our method has been demonstrated through an *in vivo* experiment.

5.2 Imaging model

The proposed dynamic speech MRI method aims to integrate the idea of a low-rank plus sparse model and a high-quality spatiotemporal atlas into a PS model-based imaging framework. This section first briefly describe the individual components - a low-rank plus sparse model-based imaging method, an improved method based on the low-rank plus sparse model with regional sparse modeling and a method that constructs high-quality spatiotemporal atlas-based on group-wise diffeomorphic registration. Upon completion of describing each individual components, this section then describes the integration of these components.

5.2.1 Low-rank plus sparse model

As discussed in Section 1.1, the measured spatiotemporal data from the $(\mathbf{k}, t)$-space can be expressed with the following equation:

$$d(\mathbf{k}, t) = \int_{\text{object}} I(\mathbf{r}, t) e^{-i2\pi\mathbf{k}\cdot\mathbf{r}} d\mathbf{r} + \eta(\mathbf{k}, t). \tag{5.1}$$

In typical dynamic speech MRI experiments, the measured $(\mathbf{k}, t)$-space data $d(\mathbf{k}, t)$ is often sparsely sampled due to physical and physiological limitations. Reconstruction of $I(\mathbf{r}, t)$ from sparsely sampled data $d(\mathbf{k}, t)$ is an underdetermined problem. As a result, sophisticated modeling of the spatiotemporal dynamics based on the anatomical and physiological features of speech motion is critical towards proper recovery of $I(\mathbf{r}, t)$ from $d(\mathbf{k}, t)$. As introduced in Chapter 3 and Chapter

4, the PS model allows high-quality reconstructions of $I(\mathbf{r}, t)$ from $d(\mathbf{k}, t)$ by assuming $I(\mathbf{r}, t)$ to be Lth order partially separable [43]. Further, the PS model assumes that $I(\mathbf{r}, t)$ exists in a low-dimensional subspace and induces a low-rank structure on the Casorati matrix $\mathbf{I}$ that is defined over any point set $I\{\mathbf{r}_n, t_m\}_{n,m=1}^{N,M}$, where N and M are the number of spatial encodings and time frames. This low-rank structure allows $\mathbf{I}$ to be factorized as $\mathbf{I} = \mathbf{UV}$ [52], where $\mathbf{U}$ and $\mathbf{V}$ represents the associated spatial and temporal subspaces, respectively. $I(\mathbf{r}, t)$ can be effectively reconstructed from the sparsely sampled data when both $\mathbf{U}$ and $\mathbf{V}$ have been determined.

Recent developments in the low-rank plus sparse model [220, 221] (or often referred to as a robust principal component analysis model [222–226]) have revealed the potential of incorporating an additional sparse component to represent residual spatiotemporal dynamics that are not captured by the low-rank model. In particular, the low-rank plus sparse model expresses the spatiotemporal dynamics to be a combination of the generic spatiotemporal motion (often known as the background motion) and the additional spatiotemporal motion that is unique to the speaker or the carrier phrase. Mathematically, the generic spatiotemporal motion (or the background motion) is represented with a low-rank matrix and the additional spatiotemporal motion is represented with a sparse matrix. Mathematically, this idea can be expressed as

$$I(\mathbf{r}, t) = L(\mathbf{r}, t) + S(\mathbf{r}, t), \tag{5.2}$$

where $L(\mathbf{r}, t)$ represents the generic spatiotemporal motion and $S(\mathbf{r}, t)$ represents the additional spatiotemporal motion. As is shown in Equation 5.2, the low-rank component is designed to capture the spatially and temporally correlated dynamics of speech, while the sparse component is designed to model the additional spatiotemporal dynamics that are built on top of the low-rank matrix. Specifically in dynamic speech MRI, the low-rank plus sparse model allows greater flexibility in modeling the spatiotemporal motion due to the separation of a low-rank component and a sparse component. In addition, the separation of the spatiotemporal motion into two parts greatly facilitates the ensuing phonetic analysis on the associated motion. Therefore, the hypothesis of this chapter is that the low-rank plus sparse model can properly represent the spatiotemporal motion; and the resulting separation of spatiotemporal motion is useful towards meaningful phonetic analysis on the articulatory dynamics.

The low-rank plus sparse model can be seamlessly integrated with the PS model-based spatiotemporal imaging framework. It is noted that the low-rank component that represents the generic spatiotemporal dynamics can be effectively expressed in the PS model [43]. Mathematically, this integration can be expressed as in matrix-vector form as

$$\mathbf{I} = \mathbf{U}\mathbf{V} + \mathbf{S}, \tag{5.3}$$

where $\mathbf{U}$ represents the spatial subspace of $\mathbf{I}$, $\mathbf{V}$ represents the temporal subspace of $\mathbf{I}$ and $\mathbf{S}$ represents the additional spatiotemporal speech motion in $\mathbf{I}$. It should be noted that the PS model is a powerful signal model that represents the universal signal property for a variety of dynamic speech MRI applications [50, 51, 74–76]. Therefore, the integration of the low-rank plus sparse model into the PS model-based spatiotemporal imaging framework allows the spatiotemporal dynamics of speech to be tailored to a specific speaker or a specific carrier phrase [232].

5.2.2 Regional sparse modeling

The representation power of the low-rank plus sparse model can be further enhanced by incorporating regional sparse modeling. In particular, Equation 5.2 assumes a global low-rank plus sparse structure, where the background component and additional component of the underlying articulatory motion are combined across all the spatiotemporal dimension. In practice, however, the additional component of any speech motion may be more obvious in the vocal tract region where most of the moving articulators appear. In the non-articulatory regions, such as the brain, the neck and the nasal cavity, the additional component does not carry significant soft-tissue movements and can therefore be captured properly with the background component alone. In the interest of focusing on modeling the spatiotemporal dynamics for major articulators, this subsection aims at enhancing the original assumption of the low-rank plus sparse model and proposes to incorporate a regional sparse model to capture the additional articulatory motion in the regions of phonetics interest. In particular, the articulatory region of regional sparse modeling is specified by a spatiotemporal mask predetermined by the user, and the resulting model can be mathematically represented as

$$I(\mathbf{r},t) = L(\mathbf{r},t) + \Omega_{\text{vocal tract}}\{S(\mathbf{r},t)\}, \tag{5.4}$$

where $\Omega_{\text{motion}}(\mathbf{r},t)\{\cdot\}$ represents an imaging operator that enforces a predefined spatiotemporal mask onto the associated image sequence of speech. The spatiotemporal region that this mask covers should be indicative of the vocal tract area in which major articulatory motion occurs. An illustration of this model can be found in Figure 5.1. It should be noticed that the spatiotemporal mask $\Omega_{\text{motion}}(\mathbf{r},t)$ shown in Figure 5.1 aims at illustrating the general concept, while in practice the spatiotemporal mask can be spatiotemporal varying and can change based on the regions of interest and articulator of interest.

Like its primitive counterpart, the low-rank plus sparse model with regional sparse modeling can also be integrated with the PS model-based spatiotemporal imaging framework. Mathematically, this integration can be expressed as in matrix-vector form as

$$\mathbf{I} = \mathbf{U}\mathbf{V} + \boldsymbol{\Omega}\mathbf{S}, \tag{5.5}$$

where $\boldsymbol{\Omega}$ represents an imaging operator that enforces a spatiotemporal support constraint to realize regional sparse modeling of speech motion in the designated region of interest. Compared with Equation 5.3, the spatiotemporal model represented in Equation 5.5 is tailored to focus on a particular region of phonetics interest. This formulation allows prior knowledge of the articulatory motion to be injected into the imaging process and consequently represents a more accurate imaging model for a particular carrier phrase of interest.

5.3 Image reconstruction

A novel dynamic speech MRI method has been proposed to integrate the proposed imaging model with a spatiotemporal atlas-based constraint to achieve high-resolution spatiotemporal imaging of speech. In particular, the formulation proposes to enforce two constraints to regularize the image reconstruction problem: (1) A sparsity constraint was chosen to be enforced on the regional sparse component in the spatiotemporal image. This sparsity constraint is suitable for facilitating dy-

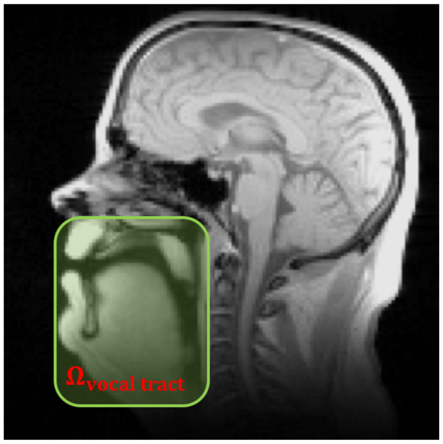


Figure 5.1: Our method models the spatiotemporal dynamics of speech motion with a low-rank plus regional sparse model: (a) the quasi-static regions, such as the brain, the nasal cavity or the background, are modeled with the background component; and (b) regions that contain major articulatory motion are modeled with both the background and the sparse component to capture both the generic motion and the additional motion of speech. These regions are schematically indicated with a spatiotemporal mask $\Omega_{\text{vocal tract}}$ in the image.

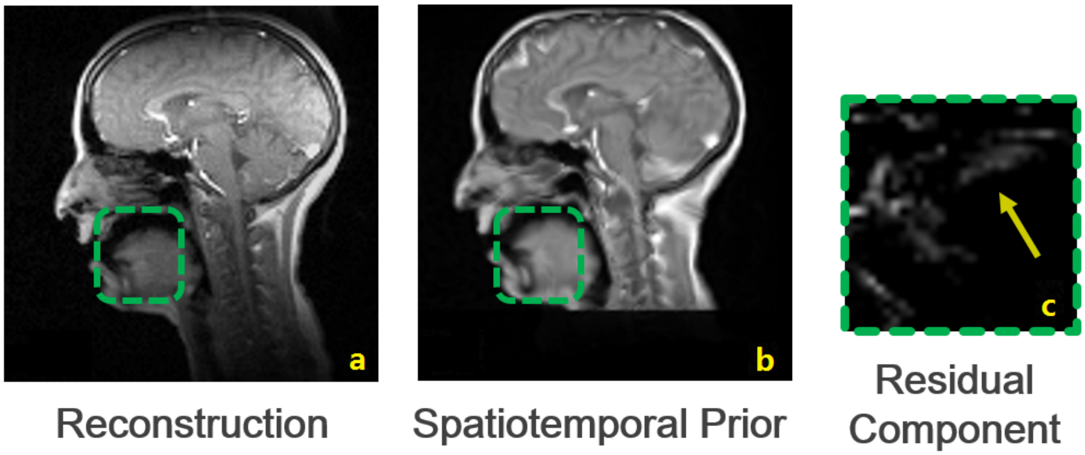


Figure 5.2: The proposed consistency constraint enforces the spatiotemporal difference between the candidate reconstruction and the spatiotemporal prior constructed from a set of high-quality spatiotemporal atlas images. This figure is a typical frame of the reconstructed image (left), a typical frame of the constructed spatiotemporal prior image (middle) and the associated residual component (right). The spatiotemporal prior is capable of approximating the true spatiotemporal motion and contrast changes of the reconstruction. The residual component is capable of picking up the difference between the reconstruction and the prior image.

namic speech MRI reconstruction due to the reasons discussed in Section 5.2.2. (2) A consistency constraint was chosen to be enforced on the spatiotemporal difference between the low-rank component, $\mathbf{UV}$, and the spatiotemporal prior, $\mathbf{I}_{\mathrm{p}}$, constructed from a set of high-quality spatiotemporal atlas images [232]. It should be noted that there exist multiple ways to construct $\mathbf{I}_{\mathrm{p}}$ [229–231], and this chapter chooses to construct $\mathbf{I}_{\mathrm{p}}$ based on a nearest neighbor classifier [234–238] with a distance metric based on the Lipschitz norm [229–231]. An illustration of a typical frame of the reconstruction, the spatiotemporal prior, $\mathbf{I}_{\mathrm{p}}$, and the associated residual component, $\mathbf{UV} - \mathbf{I}_{\mathrm{p}}$, is provided with Figure 5.2. As seen with Figure 5.2, conceptually the proposed consistency constraint is appropriate because $\mathbf{I}_{\mathrm{p}}$ is designed to approximate the true spatiotemporal dynamics, and this constraint picks up the spatial and temporal differences between $\mathbf{UV}$ and $\mathbf{I}_{\mathrm{p}}$.

The two proposed constraints work in complementary to regularize the ill-conditioned image reconstruction problem. The resulting image reconstruction formulation, can be mathematically written as

$$\{\hat{\mathbf{U}}, \hat{\mathbf{S}}\} = \arg\min_{\mathbf{U}} ||\mathbf{d} - \mathbf{E}\{\mathbf{U}\mathbf{V} + \mathbf{\Omega}\mathbf{S}\}||_2^2 + \lambda_1 ||\mathbf{U}\mathbf{V} - \mathbf{I}_{\mathrm{p}}||_2^2 + \lambda_2 ||\mathbf{\Omega}\mathbf{S}||_1, \qquad (5.6)$$

where $\mathbf{U}$ represents the desired spatial subspace, $\mathbf{S}$ represents the desired sparse component, $\mathbf{E}\{\cdot\}$ represents a composite imaging operator that incorporates sensitivity encoding, undersampling in the $(\mathbf{k}, t)$-space and the associated Fourier encoding processes, λ_1 represents a regularization parameter for the consistency constraint (the first constraint in Equation 5.6), λ_2 represents a regularization parameter for the sparsity constraint (the second constraint in Equation 5.6) and $\mathbf{\Omega}$ represents an imaging operator that enforces a spatiotemporal mask onto the associated image. The above constrained optimization problem can be solved efficiently using an algorithm based on additive half-quadratic minimization with continuation procedures [232]. A schematic summary of our method is provided in Figure 5.3.

5.4 Spatiotemporal atlas construction

Construction of an objective spatiotemporal atlas of the targeted articulatory motion is the key to improving the quality of the reconstructions and the accuracy of the phonetic analysis. This chapter chooses to construct a high-quality spatiotemporal atlas in two sequential steps: (1) construction of a generic spatiotemporal atlas and (2) the creation of a subject-specific atlas.

5.4.1 Generic spatiotemporal atlas

The atlas construction strategies proposed in [229] to construct a generic atlas in multiple steps: (1) an initial reconstruction is performed based on the methods proposed in Chapter 3. In particular, the choice of initial reconstruction is completely subject to the user's discretion, and this chapter chooses to initialize with the methods in Chapter 3, while other methods may also be employed for initialization; (2) a common spatial space [229] is defined from the initial reconstructions by selecting a time frame that has a representative articulatory posture; (3) all the time frames are

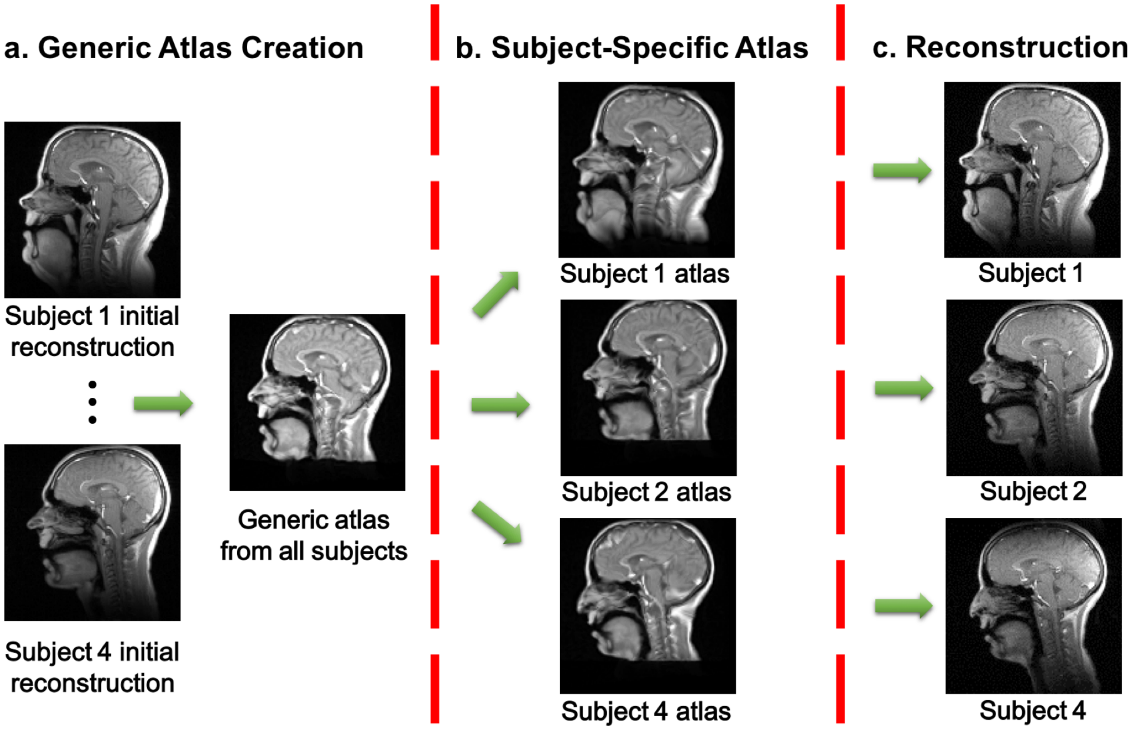

a. Generic Atlas Creation
b. Subject-Specific Atlas
c. Reconstruction
Subject 1 initial reconstruction
Subject 4 initial reconstruction
Generic atlas from all subjects
Subject 1 atlas
Subject 2 atlas
Subject 4 atlas
Subject 1
Subject 2
Subject 4

Figure 5.3: Our method integrates a high-quality spatiotemporal atlas into a low-rank plus sparse model-based imaging framework. This figure summarizes our method: (a) generation of initial spatiotemporal atlas from four initial reconstructions; (b) generation of subject-specific atlas employing spatial and temporal warping procedures; and (c) reconstruction of spatiotemporal image by jointly enforcing the proposed constraints.

spatially warped to a predefined common space. The associated spatial warping procedure is performed with the symmetric image normalization (SyN) method [214] using the large deformation diffeomorphic metric mapping algorithm (LDDMM) [239]. Specifically, it should be noted that one can employ routines in the advanced normalization tools (ANTs) open source software library [240] to perform this step, as was done in this chapter; and (4) a generic atlas is constructed by temporally realigning the preliminary atlas using a temporal alignment strategy based on the Lipschitz norm, which is defined as [229]

$$\mathrm{Lip}(\xi_{\mathrm{t_i}}^{\mathrm{t_s}}, \mathbf{r}) := \inf\{\, ||{\xi_{\mathrm{t_i}}^{\mathrm{t_s}}}^{-1}(\mathbf{r}_1) - {\xi_{\mathrm{t_i}}^{\mathrm{t_s}}}^{-1}(\mathbf{r}_2)|| \leqslant \mathrm{Lip}(\xi_{\mathrm{t_i}}^{\mathrm{t_s}}, \mathbf{r})\, ||\mathbf{r}_1 - \mathbf{r}_2||, \mathbf{r}_1, \mathbf{r}_2 \in \mathbf{r}, \mathbf{r}_1 \neq \mathbf{r}_2\}, \quad (5.7)$$

where $\xi_{\mathrm{t_i}}^{\mathrm{t_s}}$ represents a diffeomorphism between a time frame $\mathrm{t_i}$ in the initial guess and another time frame $\mathrm{t_s}$ of the subject. Particularly in this chapter, the initial guess has been set to the reconstruction of a representative subject, while other choices exist to choose the initial guess [229–231]. For each subject, the temporal alignment procedure finds the candidate time frame $\mathrm{t_i}$ from the neighboring time frames (with a window size of 20 time frames) by

$$\widehat{\mathrm{t_i}} = \underset{\mathrm{t_i} \in \{\mathrm{t_i}-10,\cdots,\mathrm{t_i}+9\}}{\arg\min} \mathrm{Lip}(\xi_{\mathrm{t_i}}^{\mathrm{t_s}}, \mathbf{r}). \quad (5.8)$$

A preliminary spatiotemporal atlas is finally constructed by applying a group-wise diffeomorphic registration on each time frame across all subjects [230, 231]. Specifically, a cross correlation criterion is employed to assist the group-wise image registration problem [214, 240]. A schematic overview of the proposed atlas construction method is provided in Figure 5.4.

The construction of a spatiotemporal atlas is not limited to 2D dynamic speech MRI applications. As a generic method, this technique can also find its use in dynamic speech MRI of higher dimensions. Figure 5.5 provides an example where this generic atlas-construction technique is applied to construct a 3D spatiotemporal atlas. As can be seen with Figure 5.5, the upper image represents a representative frame from the 3D reconstruction, while the lower image represents the corresponding 3D atlas image. Although the spatiotemporal dimensions have significantly increased from 2D speech MRI to 3D speech MRI, our method is still capable of constructing high-quality atlas images. The image shown in Figure 5.5, for instance, is devoid of significant

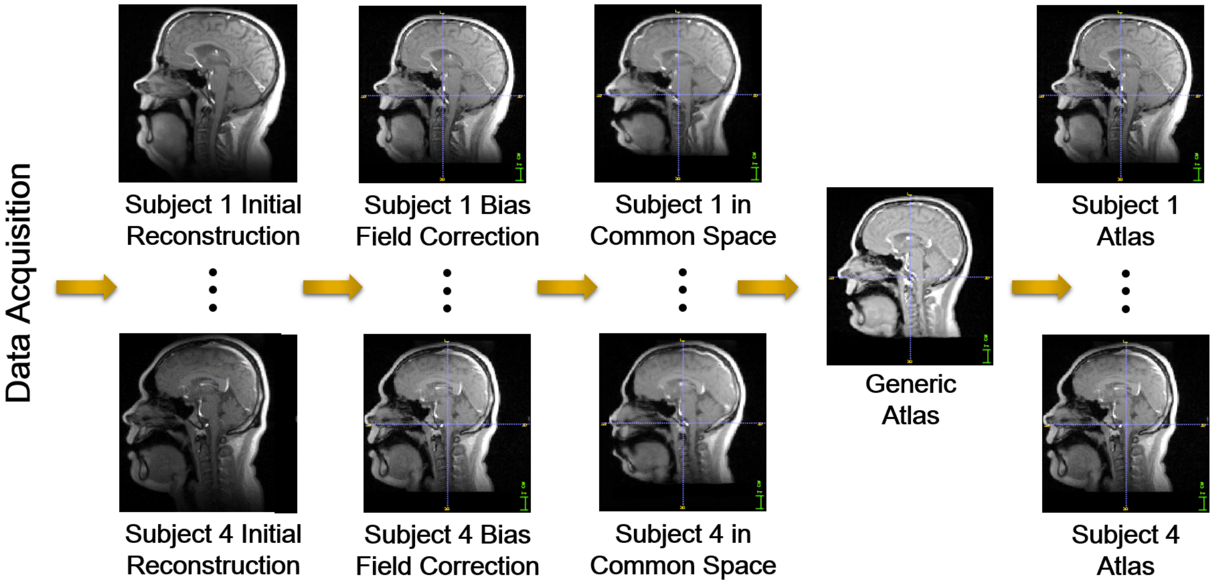

Data Acquisition
Subject 1 Initial Reconstruction
Subject 1 Bias Field Correction
Subject 1 in Common Space
Subject 1 Atlas
Generic Atlas
Subject 4 Initial Reconstruction
Subject 4 Bias Field Correction
Subject 4 in Common Space
Subject 4 Atlas

Figure 5.4: Our method constructs a subject-specific spatiotemporal atlas in multiple steps: (a) a common spatial space is defined from the initial reconstructions by selecting a time frame that has a representative articulatory posture; (b) all time frames are spatially warped to the predefined common space; (c) a preliminary spatiotemporal atlas is constructed by temporally realigning the preliminary atlas using a temporal alignment strategy based on the Lipschitz norm; and (d) a generic atlas is constructed by applying a group-wise diffeomorphic registration on each time frame across all subjects.

contrast difference as compared with the reconstruction.

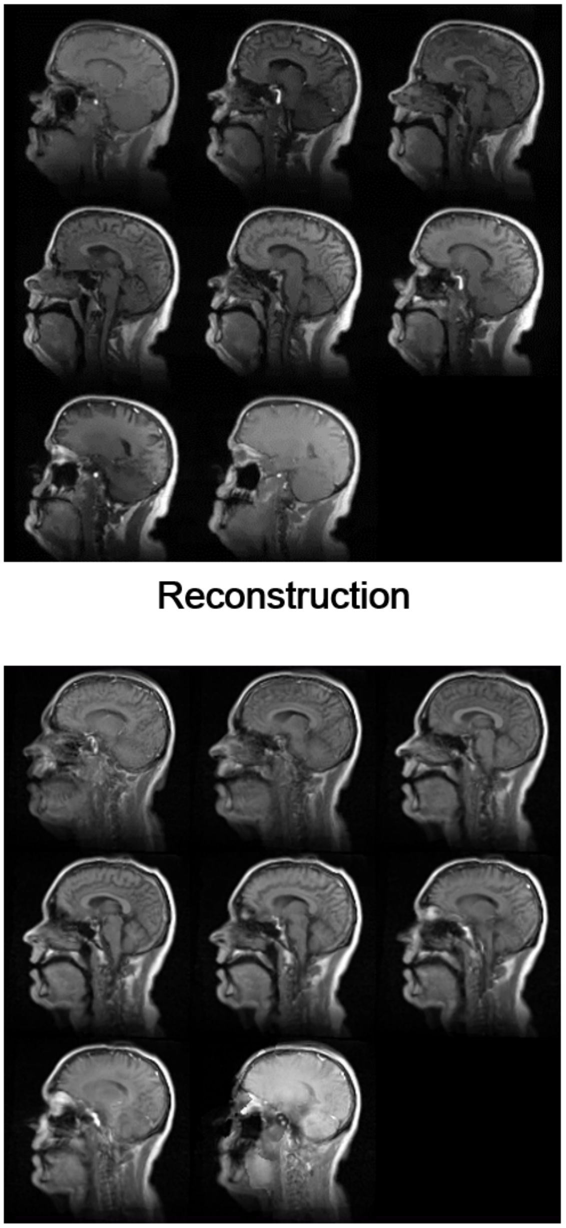
5.4.2 Subject-specific spatiotemporal atlas

There exist multiple ways to construct the subject-specific atlas-based on the steps described in Section 5.4.1 [229–231]. This chapter chooses to apply an inverse spatial transform to warp the generic atlas back to each subject's image sequence by using the SyN method described in Section 5.4.1 [214, 239, 240]. The resultant subject-specific atlas is not only devoid of significant contrast difference or temporal misalignment with the initial reconstruction, but is also an objective representation of the generic motion pattern across all the subjects [230, 231].

5.5 Experimental studies

5.5.1 Experimental design

To demonstrate the effectiveness of our method, a validation experiment was performed on a Siemens Trio scanner (Siemens Medical Solutions, Erlangen, Germany) with the following specifications identical with those described in the previous chapters. Four volunteers (three males and one female) participated in the validation experiment. These subjects had an age range of 23 to 38 and were requested to produce /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ recurrently at their natural speaking pace. Informed consent was obtained and the experiments were carried out in accordance with local protocols from the Institutional Review Board at the University of Illinois at Urbana-Champaign.



Reconstruction

Generic 4D Atlas

Figure 5.5: A set of 3D generic spatiotemporal atlas constructed using the method introduced in Chapter 5. The upper image shows a representative frame from the reconstruction. The lower image shows the corresponding generic atlas image. The subject-specific atlas image is devoid of significant contrast difference as compared with the reconstruction.

5.5.2 Experimental results

Figure 5.6 shows the articulatory movements from *in vivo* experiments where the subjects were asked to produce recurrent the /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ sounds at their own speeds. Although these four subjects have diverse anatomical features and distinct speaking pace, their articulatory movements are invariably well-captured with our method. For instance, Figure 5.6a shows the tongue gestures at neutral positions during the production of the /la/ sound. Figure 5.6b shows protrusion of the tongue (as indicated with yellow arrows) in preparation for the production of the /loo/ sound. Compared with Figure 5.6a, the elevation of the tongue towards the alveolar ridge are well described in Figure 5.6b with fine spatial details across all the subjects, despite differences in their vocal tract anatomies and variance in speech motion patterns.

Figure 5.7 shows the effectiveness of our method in capturing the low-rank and the sparse components. In particular, this chapter focuses on comparing the reconstructed components among two temporally adjacent time instances where the articulatory motions differ slightly in localized regions. As can be seen, Figure 5.7a shows a time instance when the tongue is in touch with the alveolar ridge and the velum is raised towards the pharyngeal wall, and Figure 5.7b shows an ensuing time instance when the tongue remains touching the alveolar ridge and the velum is lowered towards the root of the tongue. Each row in Figure 5.7a and Figure 5.7b contains three horizontally arranged images: the reconstructed image frame, the reconstructed low-rank component and the estimated sparse component. The articulatory movements in (a) and (b) differ slightly due to the temporal adjacency of these two image frames, and most difference are found in localized regions around the velum or at the tip of the tongue. However, the proposed method is still capable of capturing the difference in articulatory motion. Difference in localized articulatory motion is well-captured in both the velum and the tongue tip.

Figure 5.8 shows that the proposed low-rank plus sparse model is capable of capturing dynamic subject-specific motion that exists on top of the generic motion. In particular, Figure 5.8a shows an overlaid image where the low-rank component is color coded with green color and the sparse component is color coded with cyan color. Although the relative locations are less obvious when reading the gray-scale images, the color coded image in Figure 5.8a allows the position of the sparse component to be clearly observed in relation to the underlying low-rank component.

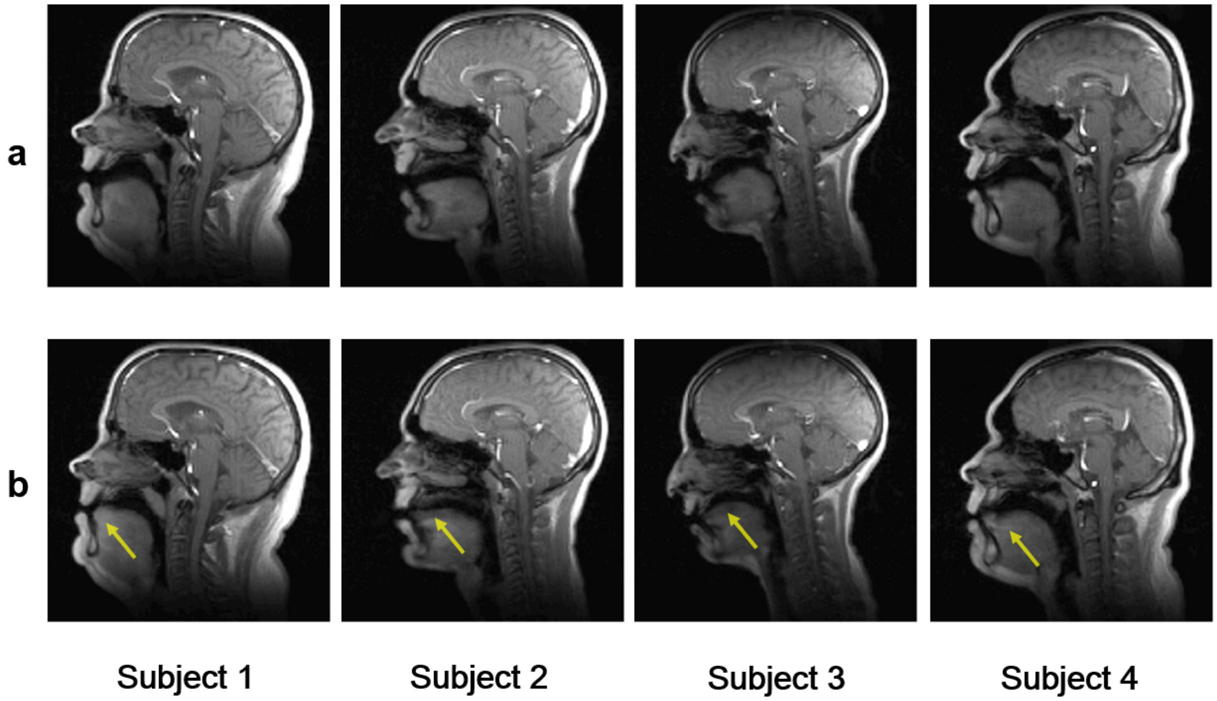


Figure 5.6: Articulator gestures from all subjects when they produce the /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ sounds at: (a) a resting position during the /la/ sound; and (b) an elevated position where the tongue protrudes towards the lips (as indicated with yellow arrows) and the velum elevates towards the velopharyngeal wall. Although these four subjects have diverse anatomical features and distinct patterns of speech motion, their articulatory movements are invariably well-captured with our method.

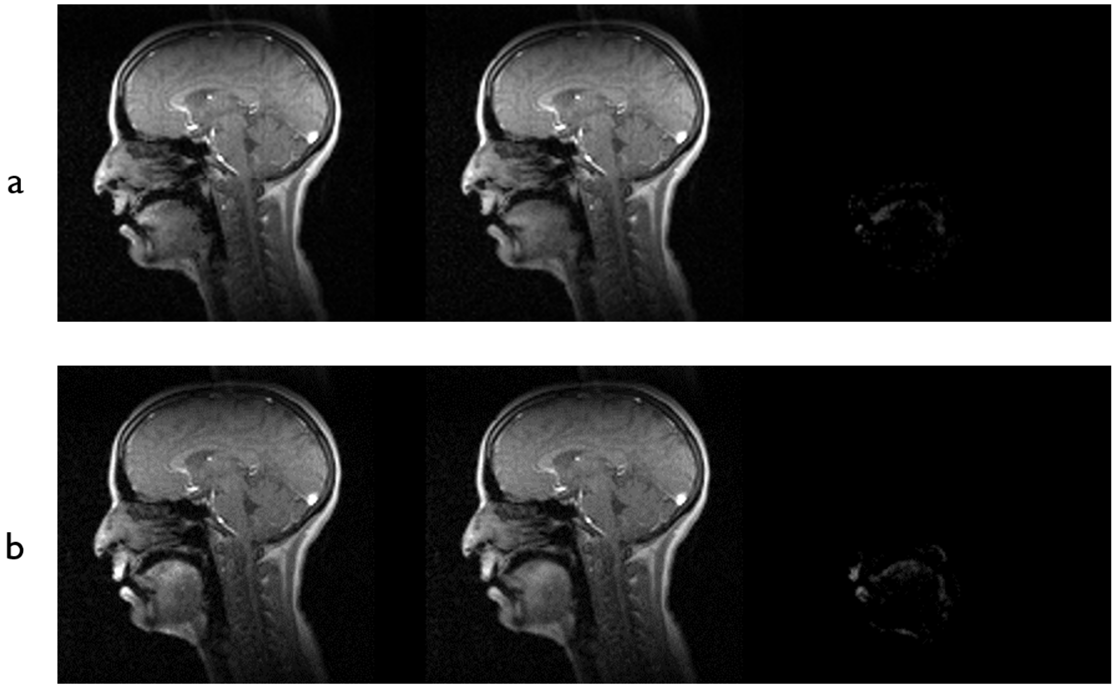


Figure 5.7: The reconstructed articulatory gesture compared with the associated low-rank and sparse component. The reconstruction are obtained from an *in vivo* data set where the subject was requested to produce the /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ sounds. Reconstruction results from two temporally adjacent time instances are shown: (a) a time instance when the tongue is in touch with the alveolar ridge and the velum is raised towards the pharyngeal wall; and (b) an ensuing time instance when the tongue remains touching the alveolar ridge and the velum is lowered towards the root of the tongue. Each row of the results contains three horizontally arranged images: the reconstructed image frame, the reconstructed low-rank component and the estimated sparse component. As can be seen, the articulatory movements in (a) and (b) differ slightly in very localized regions, but our method is still capable of capturing the difference in articulatory motion in both the low-rank and the sparse component.

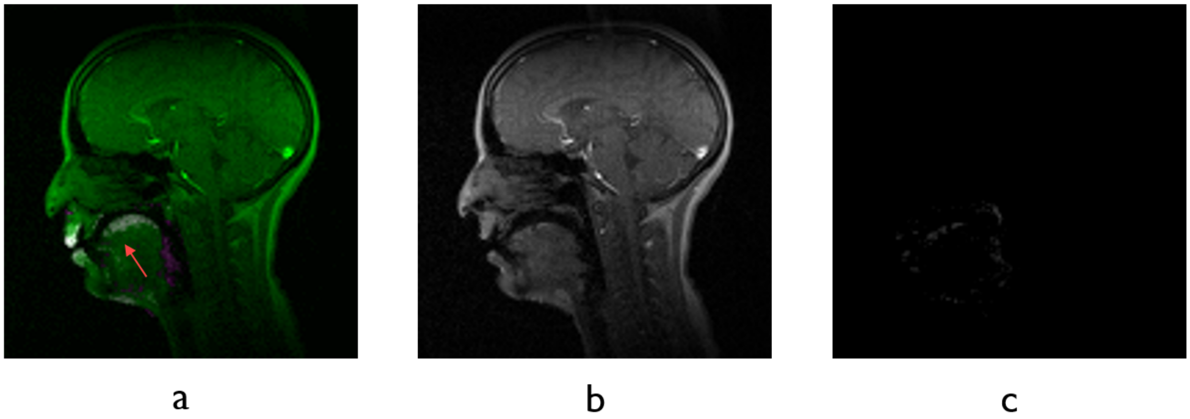


Figure 5.8: The reconstructed sparse component is associated with the subject-specific articulatory motion. This figure shows (a) an overlaid image where the low-rank component is color coded with green color and the sparse component is color coded with cyan color; (b) the low-rank component as a gray-scale image; and (c) the sparse component as a gray-scale image. It is obvious from (a) that the reconstructed sparse component in (c) is spatially in register with the reconstruction. In addition, the sparse component mainly exists in regions where subject-specific movements are likely to happen, such as the lips and the surface of the tongue. These images show that our low-rank plus sparse model is capable of capturing additional dynamic subject-specific motion that exists on top of the generic motion.

Figure 5.8b illustrates the low-rank component as a gray-scale image and Figure 5.8c shows the associated sparse component as a gray-scale image. It is obvious from the comparison between Figure 5.8a, b and c that the reconstructed sparse component is spatially in register with the reconstruction. In addition, the sparse component mainly exists in regions where subject-specific movements are likely to happen, such as the lips and the surface of the tongue. Results in Figure 5.8 reveal the connection of the underlying articulatory motion with the reconstructed low-rank and sparse component.

The separation of the low-rank and the sparse components in the imaging model allows the articulatory gestures to be compared with both the generic motion and the subject-specific motion. Specifically, our method introduces a spatiotemporal atlas into the imaging framework. The spatiotemporal atlas and the associated spatiotemporal prior image provides a standard image frame that is in register with the subject's speech motion and is devoid of major contrast difference. This benefit of a spatiotemporal atlas and the prior provides a metric to evaluate the subject's motion variability targeted to a specific carrier phrase. It should be noticed that, without our method or a spatiotemporal atlas, such evaluation may purely depend on the evaluator's subjective feeling and is therefore hardly objective. With our method, however, objective evaluation of motion and motion variability is enabled, in addition to the high-quality reconstructions available from our method.

Figure 5.9 shows representative phonetic analysis results available from our method. In particular, the tongue contours for all /a/ sounds during the production of /loo/-/lee/-/la/-/za/- /na/-/za/ phrases are extracted from both the reconstruction and the spatiotemporal prior image [12]. For example, a representative tongue contour from the reconstructed image is illustrated on the right of Figure 5.9. The envelopes of all tongue contours from the atlas are plotted with pink color and those from the reconstruction are plotted with cyan color. As seen, the tongue contours from the spatiotemporal prior image (pink) are spatially and temporally aligned with the reconstruction tongue contours (cyan), but demonstrates a thinner envelope. This suggests the spatiotemporal prior image is representative of the generic pattern of speech, which is expected given the "averaging" effect during the atlas construction procedures [229–231].

The ability to reconstruct the low-rank component and the sparse component also provides other perspectives in analyzing the underlying articulatory motion. One way to analyze motion would be to read into the temporally averaged motion pattern. However, one typical problem is that the temporally averaged motion pattern does not contained much additional information because the differences have been reduces during the process of taking averages. However, such differences can be separated and clearly analyzed with the arrival of the sparse component.

In particular, Figure 5.10 demonstrates the possibility of performing statistical analysis of the associated articulatory motion even with the temporally averaged results. As can be seen, a temporally averaged reconstruction image frame in Figure 5.10a, while a temporally averaged low-rank component image frame in Figure 5.10b. As discussed, simple visual inspection of the images in Figure 5.10a and Figure 5.10b does not reveal significant difference of articulatory motion between the reconstruction and the low-rank component. This is because the difference is sometimes subtle and has been further reduced during the process of taking averages. However, the difference can be clearly revealed from Figure 5.10c, which is obtained by temporally averaging the sparse

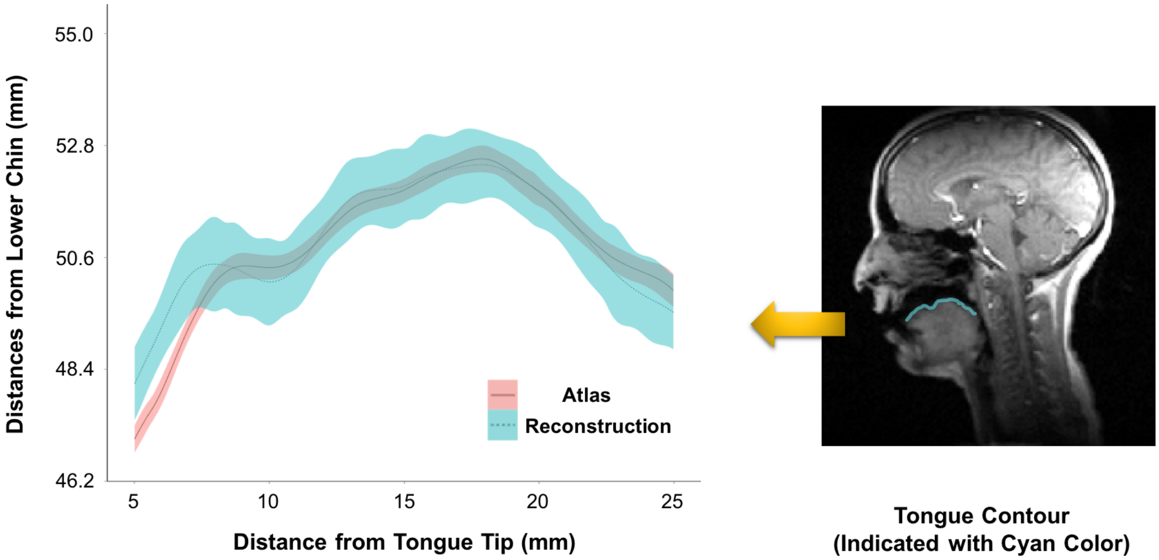


Figure 5.9: Envelopes of tongue blade contours of the spatiotemporal prior image (pink color) and the reconstruction (cyan color) for all /a/ sounds. The envelope from the spatiotemporal prior image is approximately aligned with that from the reconstruction , but demonstrates smaller motion variance. Representative tongue contours from atlas are reconstructions are shown on the right.

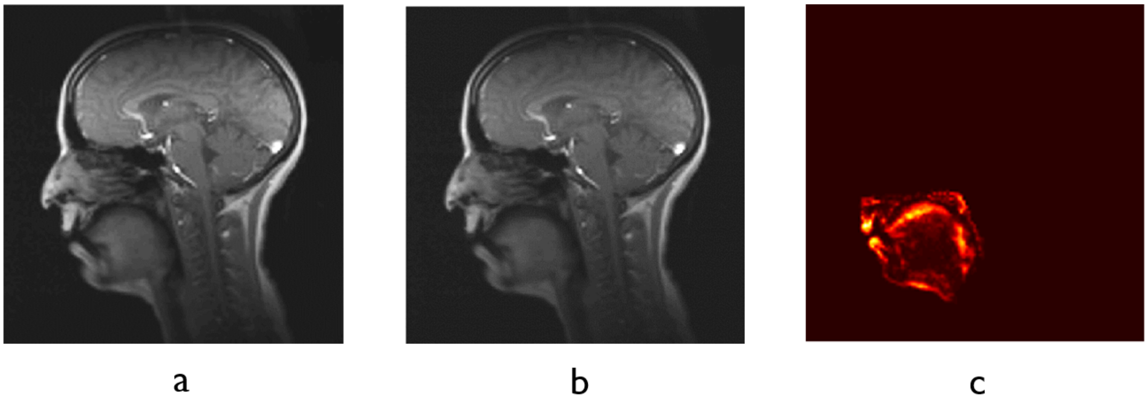


Figure 5.10: Statistical analysis of the reconstructed sparse component allows the subject-specific motion patterns to be effectively analyzed. This figure shows (a) a temporally averaged reconstruction image frame and (b) a temporally averaged low-rank component image frame. Visual inspection of the images in (a) and (b) does not reveal significant difference of articulatory motion between the reconstruction and the low-rank component. However, the difference can be clearly revealed from (c), which is obtained by temporally averaging the sparse component along all spatial dimensions. The result indicates that the major motion variation of this speaker mainly exists in the following regions: the lips, the surface of the tongue, the back of the velum as well as the tongue dorsum. This result indicates that the speaker has a more vibrant articulatory motion as compared to the generic pattern of speech represented by the atlas image.

component along all spatial dimensions. The result indicates that the major motion variation of this speaker mainly exists in the following regions: the lips, the surface of the tongue, the back of the velum as well as the tongue dorsum. This result, even without significant effort to analyzed, intuitively indicates that the speaker has a more vibrant articulatory motion as compared to the generic pattern of speech represented by a spatiotemporal atlas.

Figure 5.11 shows the effectiveness of capturing high-quality articulatory movements and high-quality temporal profile from an *in vivo* experiment. Specifically in this experiment, this subject was requested to produce the /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ sounds. As can be seen with Figure 5.11a, two representative articulatory gestures are shown from time frame 4779 and time frame 4951 of the reconstructed image sequence. The articulatory gestures are particularly chosen as typical time frames that represent the protruded and the retracted motion of the tongue, respectively. Both types of movements are well-captured by our method. In Figure 5.11b, a temporal profile ($\sim$18 s in length) is taken along a vertical strip across the tongue tip. The movements of the tongue tip, especially its contact against the alveolar ridge, is captured with high temporal fidelity in the provided temporal profile. Figure 5.11 demonstrates that both the articulatory movements and the temporal dynamics of the subject are well-captured by our method.

Figure 5.12 demonstrates the effectiveness of our atlas construction method. In particular, this is demonstrated by comparing a representative time frame from the reconstruction in Figure 5.12a, with a corresponding time frame from the spatiotemporal prior in Figure 5.12b. The residual component is also shown in Figure 5.12c to assist the comparison. By comparing Figure 5.12a and Figure 5.12b, it is noted that the gestures of the tongue and the velum resemble, in general, in the reconstruction and the spatiotemporal prior image, while being dissimilar in subtle spatial structures and contrast details.

Figure 5.13 demonstrates the effectiveness of capturing temporal dynamics by comparing the temporal profiles of the reconstruction and the spatiotemporal prior image. Temporal profiles along the tongue tip from the subject are shown in Figure 5.13a and Figure 5.13b. It is noticed that the spatiotemporal prior image is capable of representing the generic articulatory motion. Specifically, the upward and downward movements of the tongue tip are well-represented in Figure 5.13b. However, the generic motion from the spatiotemporal prior image lacks sufficient temporal details to correctly describe detailed temporal dynamics, such as the contact of the tongue tip towards the alveolar ridge.

Figure 5.14 aims to demonstrate the benefits of applying the low-rank plus sparse model to capture the spatiotemporal dynamics of speech. Compared with imaging models that represent the spatiotemporal dynamics as low-rank and sparse, the low-rank plus sparse model allows temporal dynamics to be analyzed for the low-rank component and the sparse component, respectively. In particular, Figure 5.14 shows the temporal dynamics from an *in vivo* experiment where a subject was requested to produce the /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ sounds. In particular, the temporal profiles are taken from a vertical strip across the tip of the tongue. Figure 5.14a shows the position of the strip; Figure 5.14b shows the temporal profile of the reconstruction; Figure 5.14c shows the temporal profile of the low-rank component; and Figure 5.14d shows the temporal profile of the sparse component. As can be seen, the temporal profile of the low-rank component represents the

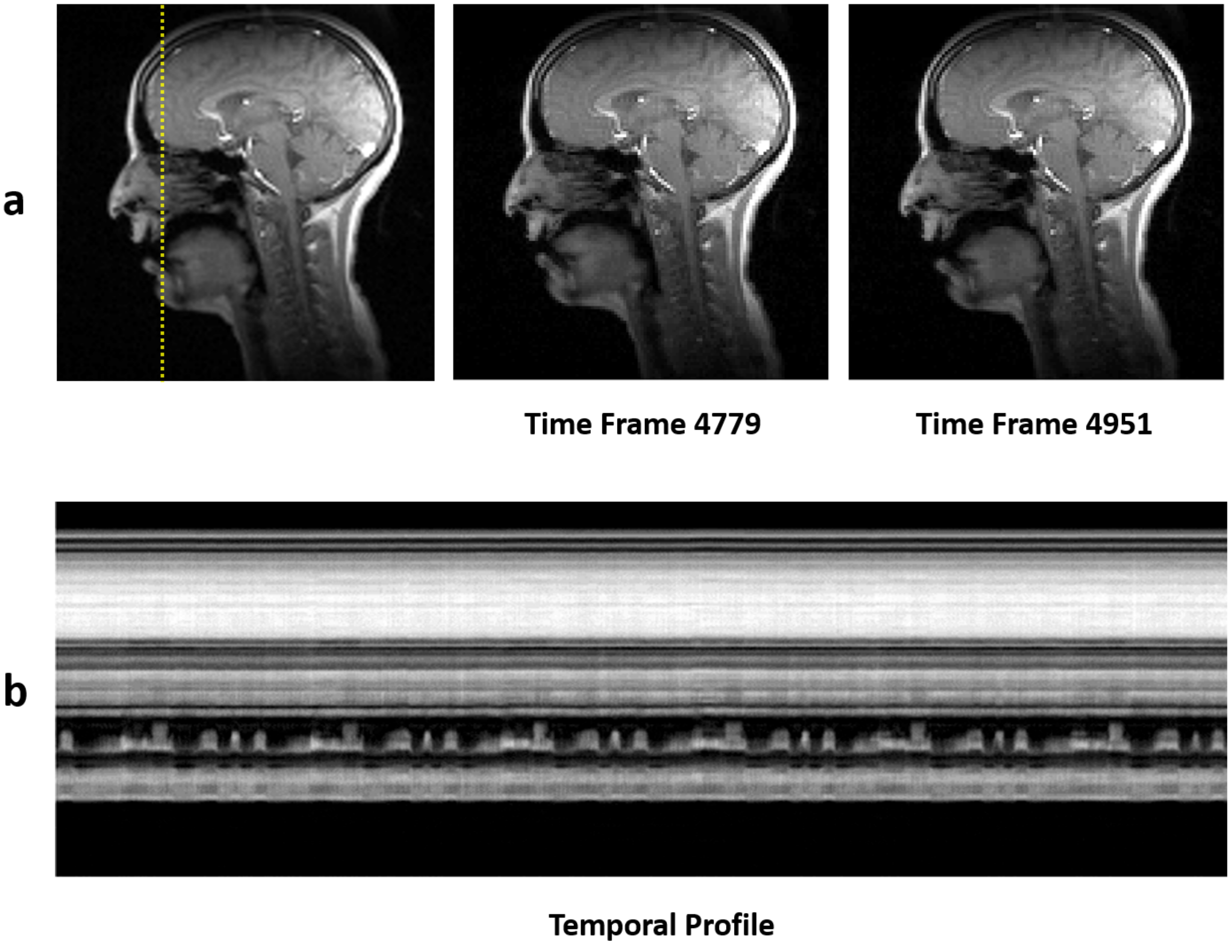


Figure 5.11: The articulator gestures and the temporal profile from an *in vivo* experiment where a subject was requested to produce the /loo/ - /lee/ - /la/ - /za/ - /na/ - /za/ sounds: (a) two representative articulatory gestures are shown from time frame 4779 and time frame 4951, representing protruded and retracted motion of the tongue, respectively; (b) the temporal profile (∼18 s in length) is taken along a vertical strip across the tongue tip. Both the articulatory movements and the temporal dynamics are well-captured by our method.

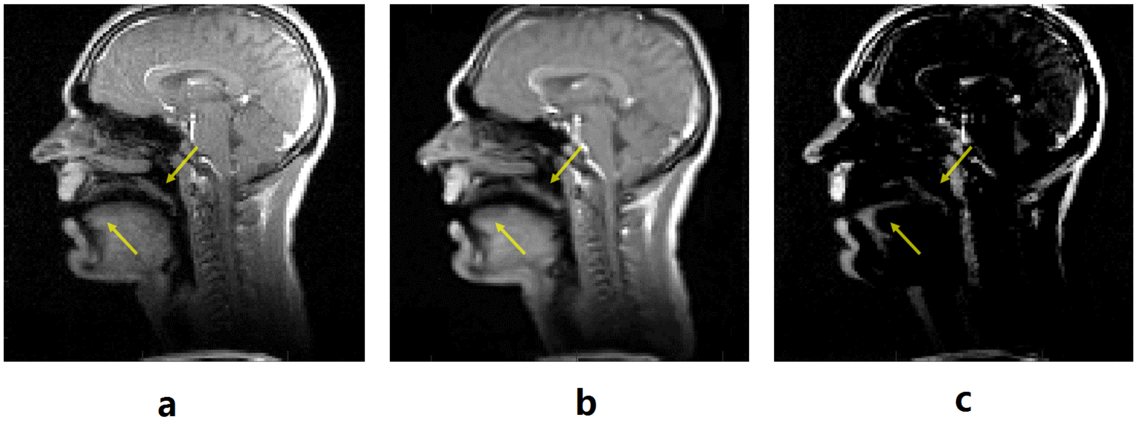


Figure 5.12: Comparison of (a) the reconstruction, (b) the spatiotemporal prior and (c) the residual component. It is noted that the gestures of the tongue and the velum resemble in the reconstruction and the spatiotemporal prior, while being different in subtle spatial structures and contrast details.

generic motion at the tip of the tongue, including the upward and downward movements of the tip of the tongue towards the alveolar ridge, as well as those of the lower jaw. However, additional contrast and subtle temporal transitions are picked up by the sparse component. The application of the low-rank plus sparse model jointly with a spatiotemporal atlas decomposes the temporal dynamics of speech into two perspectives: the generic movements that are common among all speakers that contribute to the atlas, as well as the additional dynamic movements that are specific to the subject, which faithfully represent the subject's pattern of speech over time. The integration of the proposed imaging model with a spatiotemporal atlas not only separates these two patterns of speech, but also integrates them to capture the full temporal dynamics to represent the subject's motion.

Our method allows temporal dynamics of speech to be decomposed and subtle temporal motion to be analyzed in detail. Figure 5.15 illustrates the temporal dynamics from an identical experiment as in Figure 5.14, but attempts to analyze the temporal motion in greater detail. In particular, the temporal transitions are organized as follows: Figure 5.14a shows a representative mid-sagittal frame and the temporal profile of the reconstruction; Figure 5.14b shows a representative mid-sagittal frame and the temporal profile of the low-rank component; and Figure 5.14c

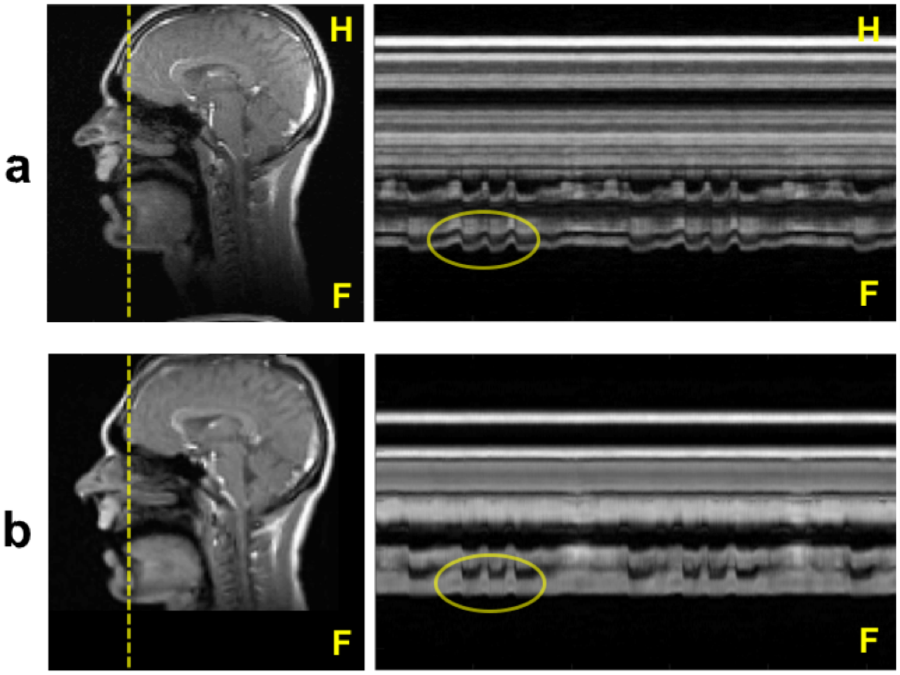


Figure 5.13: Comparison of temporal profiles between the reconstruction and the atlas. Temporal profiles taken from a vertical strip on the subject across the tongue tip and the alveolar ridge are compared: (a) the subject's reconstruction and (b) the associated spatiotemporal prior image. Note that the spatiotemporal prior image is capable of representing the generic articulatory motion, such as the upward and downward motion of the tongue tip. However, the lack of temporal details needs to be compensated by the proposed model.

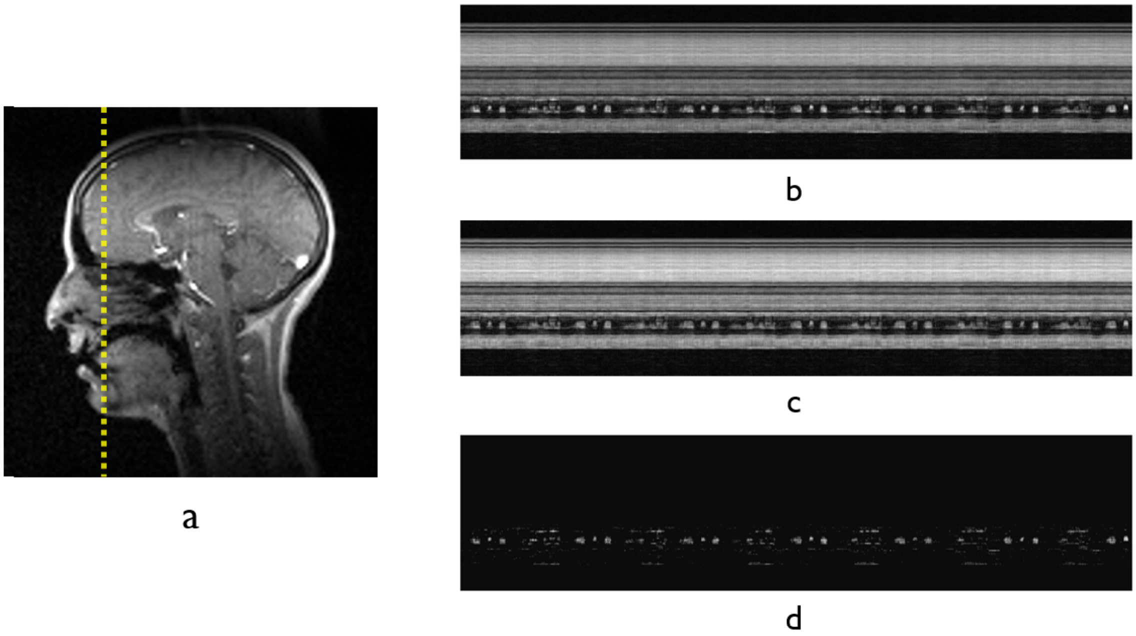


Figure 5.14: The low-rank plus sparse model allows temporal dynamics to be analyzed separately for the low-rank and the sparse components. This figure shows the temporal dynamics taken from a vertical strip across the tip of the tongue: (a) the position of the vertical strip; (b) the temporal profile of the reconstruction; (c) the temporal profile of the low-rank component; and (d) the temporal profile of the sparse component.

shows a representative mid-sagittal frame and the temporal profile of the sparse component. Compared with Figure 5.14, the Figure 5.15 zooms in the temporal events of the speaker's motion and reveals in detail how the sparse component interacts with the low-rank component to recover the full temporal dynamics associated with the carrier phrase. In particular, the sparse component compensates for the subtle contrast differences in small-sized regions such as the tip of the tongue and the tip of the alveolar ridge. As a result, the temporal transitions in the reconstruction are more natural due to the complementary roles of the low-rank and sparse components.

5.6 Discussion

Our method provides high-resolution imaging of speech and accurate assessment of articulatory motion by integrating a spatiotemporal atlas into low-rank plus sparse model-based dynamic speech MRI. The integration of a low-rank plus sparse model into dynamic speech MRI not only improves the spatial details and temporal dynamics in the reconstruction, but also allows quantitative characterization of articulatory motion in the form of low-rank and sparse components. Although the merits of our method have been demonstrated in the *in vivo* experiments above, several aspects of the method need to be considered. These aspects include the proposed sparsity constraint, the selection of model order and regularization parameter, and the computational requirements.

A high-quality atlas requires accurate spatiotemporal registration of the generic atlas towards the spatiotemporal prior image. Unlike many existing works [241, 242] that aim to solve the spatiotemporal registration problem in one single step, this chapter chooses to solve the registration problem in two separate steps: spatial warping-based on the reference frame and temporal alignment based on the Lipschitz norm. In this way, the original high dimensional image registration problem can be reduced into a lower dimensional problem that register images in a frame-wise fashion. Compared with the approaches from the existing works [241, 242], the approach used in this chapter can significantly simplify the spatiotemporal registration problem and increase the throughput of the generation of the spatiotemporal prior image. The effectiveness of this strategy

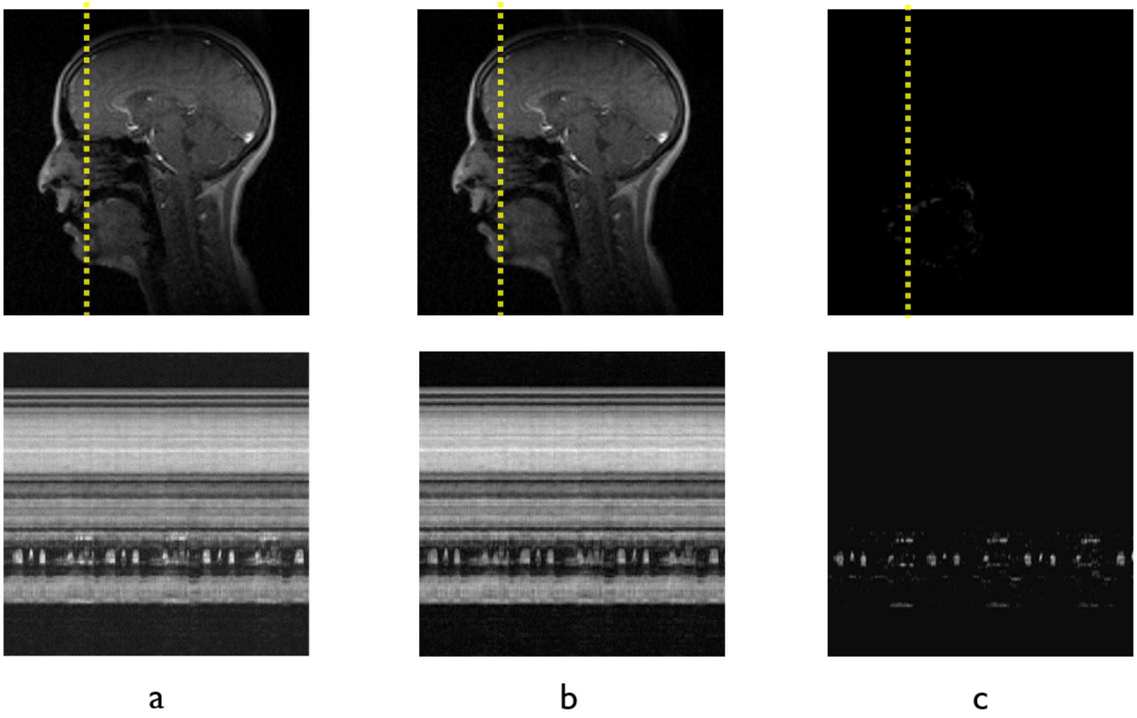


Figure 5.15: Temporal motion analyzed in detail: (a) a representative frame and the temporal profile of the reconstruction; (b) a representative frame and the temporal profile of the low-rank component; and (c) a representative frame and the temporal profile of the sparse component. Compared with Figure 5.14, this figure zooms in the temporal events of the speaker's motion and depicts how the sparse component interacts with the low-rank component to recover the full temporal dynamics associated with the carrier phrase. As can be seen, the sparse component compensates for the subtle contrast differences in small-sized regions such as the tip of the tongue and the tip of the alveolar ridge.

has been demonstrated in both previous [229–231] and current work. However, quantitative characterization of its performance will be an interesting topic for future research.

The performance of the proposed sparsity constraint depends largely on the sparsity level available from the sparse component. Although it is generally assumed that the low-rank component of the imaging model captures the generic articulatory motion and the sparse component should hence be intrinsically sparse, in practice the sparsity level may be suboptimal due to the noise or the suboptimal performance of the image registration algorithm. Aiming at this issue, this chapter further enhances the sparsity level by incorporating a regional sparse model: the sparse component is only estimated for the regions that are likely to contain meaningful articulatory motion. In other regions inside the vocal tract, the non-sparse components are naturally suppressed, and hence the sparse component can be more focused on the "dynamic movements that matter." It should be noticed, however, that regional sparse modeling is not necessarily required and the performance of the proposed model is generally robust because of the complementary role between the consistency constraint and the sparsity constraint. Systematic characterization of the performance of the proposed sparsity constraint under different sparsity levels will be carried out in future studies.

Our method also requires the selection of the model order L, and two regularization parameters λ_1 and λ_2. In this chapter, L, λ_1 and λ_2 were chosen based on visual inspection of image quality based on the quality of spatial details and the level of temporal blurring. Specifically, this chapter chose L = 40, $\lambda_1 = 2.25 \times 10^{-2}$ and $\lambda_1 = 1.75 \times 10^{-2}$ for the validation experiments. It is noted that this set of regularization parameters are chosen empirically and may be sub-optimal. However, there exist some theoretical results that may shed light upon the choices of the model order [243] and the regularization parameters [244]. The integration of these theoretic selection criteria into the selection of optimal parameters for our method should be carried out in the future.

Our method may pose a computational burden in its clinical applications due to the high dimensional optimization problem involved in reconstructions and the computationally-expensive procedures to construct a spatiotemporal atlas. For instance, the construction of an atlas for a total of 8960 time frames at a nominal speed of 102 fps (corresponding to a 1 min 28 s scan) took approximately 47 hours on a standard workstation. With the spatiotemporal atlas predetermined, the reconstruction time was around 4.5 hour on a 24-core SUN workstation without code optimization. Fortunately, acceleration in computation time may be realized by taking advantage of the strong computational power of the graphical processing units [205, 218, 219]. The implementation of the associated algorithm and optimization of computational efficiency will be investigated in future research.

5.7 Summary

This chapter proposes a novel dynamic speech MRI method to capture articulatory dynamics with simultaneous high imaging speed and accurate characterization of the articulatory motion. This is achieved by integrating a spatiotemporal atlas into a low-rank plus sparse imaging model to properly capture the spatiotemporal dynamics of speech. The resulting low-rank and sparse components enable quantitative characterization of articulatory motion. In particular, the generic articulatory motion can be well-represented by the low-rank component and the subject-specific motion is represented by the sparse component. Further, the practical utility of our method has been validated through an *in vivo* experiment. Our method serves as a promising tool for comparing and characterizing articulatory motion with regards to the generic motion pattern.

CHAPTER 6

CONCLUSION

The scientific and clinical value of dynamic speech MRI has long been limited by its intrinsic trade-offs between imaging speed, spatiotemporal resolution, spatial coverage and motion characterization. This dissertation has shown that high-resolution, full-vocal-tract 3D dynamic speech MRI with quantitative characterization of generic and individual speech motion can be achieved with our methods.

A PS model-based imaging method has been shown to capture the detailed spatiotemporal structure of dynamic speech images and recover high-quality spatiotemporal speech dynamics from highly undersampled measured data. An image reconstruction formulation has been shown to successfully leverage signal properties and integrate the imaging model and sparse modeling. A PS model-based data acquisition strategy has been developed to capture fast transitions of dynamic articulatory motion. Efficient algorithms based on half-quadratic regularization have been implemented and optimized. Performance has been systematically evaluated through simulation and *in vivo* experiments, achieving a nominal imaging speed of 166 fps with a spatial resolution of 2.2 × 2.2 × 5.0 mm^3 for an imaging volume covering the entire vocal tract. Its effectiveness has also been demonstrated through phonetic studies on American English flaps and French nasalization.

A dynamic speech MRI method based on deformation analysis has been shown to not only enhance the quality of the reconstruction, but also allow quantitative analysis of the articulatory motion by high-resolution deformation fields. Our method enables simultaneous estimation of a high-resolution dynamic image sequence and a high-resolution deformation field. Both 2D and 3D dynamic speech MRI experiments were performed to demonstrate the effectiveness of our methods - the 2D dynamic speech MRI experiments were performed to achieve a spatial resolution of 2.2 mm × 2.2 mm with a nominal imaging speed of 100 fps, while the 3D dynamic speech MRI experiments were performed to achieve a spatial resolution of 2.2 mm × 2.2 mm × 5.0 mm, a nominal imaging speed of 166 fps and an imaging volume covering the entire upper vocal tract with 8 mid-sagittal slices without slice gap. This method has been further developed to allow automatic tracking of articulatory gestures for a phonetics study on American English. Performance of this method has been evaluated by simulation, *in vivo* experiments and phonetics investigations.

To allow characterization of the subject-specific articulatory motion and the generic motion pattern, this dissertation has also proposed to incorporate a spatiotemporal atlas into a low-rank plus sparse model-based imaging framework. This approach not only improves the image reconstruction quality through the use of a spatiotemporal atlas, but also enables meaningful separation of speech motion into the low-rank and sparse components. In addition, regional sparse modeling has been introduced to assist the analysis of subject-specific motion over a designated area of interest. Successful analyses of motion have been performed on both 2D and 3D dynamic speech MRI experiments from various subjects across multiple subject groups. Our method enables objective analysis of subject-specific speech motion of a particular subject as opposed to the generic pattern of speech across a particular group. Our methods serves as promising tools for comparing and characterizing articulatory motion for large-scale phonetics studies.

REFERENCES

[1] D. Aalto, O. Aaltonen, R.-P. Happonen, P. Jääsaari, A. Kivelä, J. Kuortti, J.-M. Luukinen, J. Malinen, T. Murtola, R. Parkkola et al., “Large scale data acquisition of simultaneous MRI and speech,” *Applied Acoustics*, vol. 83, pp. 64–75, 2014.

[2] M. Echternach, J. Sundberg, S. Arndt, M. Markl, M. Schumacher, and B. Richter, “Vocal tract in female registers a dynamic real-time MRI study,” *Journal of Voice*, vol. 24, no. 2, pp. 133–139, 2010.

[3] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, “Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging,” *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 123–132, 2008.

[4] S. Ha, D. P. Kuehn, M. Cohen, and N. Alperin, “Magnetic resonance imaging of the levator veli palatini muscle in speakers with repaired cleft palate,” *Cleft Palate-Craniofacial Journal*, vol. 44, no. 5, pp. 494–505, 2007.

[5] H. Shinagawa, T. Ono, E.-I. Honda, S. Masaki, Y. Shimada, I. Fujimoto, T. Sasaki, A. Iriki, and K. Ohyama, “Dynamic analysis of articulatory movement using magnetic resonance imaging movies: Methods and implications in cleft lip and palate,” *Cleft Palate-Craniofacial Journal*, vol. 42, no. 3, pp. 225–230, 2005.

[6] C. Drissi, M. Mitrofanoff, C. Talandier, C. Falip, V. Le Couls, and C. Adamsbaum, “Feasibility of dynamic MRI for evaluating velopharyngeal insufficiency in children,” *European Radiology*, vol. 21, no. 7, pp. 1462–1469, 2011.

[7] B. Atik, M. Bekerecioglu, O. Tan, O. Etlik, R. Davran, and H. Arslan, “Evaluation of dynamic magnetic resonance imaging in assessing velopharyngeal insufficiency during phonation,” *Journal of Craniofacial Surgery*, vol. 19, no. 3, pp. 566–572, 2008.

[8] R. Shosted, M. Fu, A. Benmamoun, Z.-P. Liang, and B. P. Sutton, “Using partially separable functions to image spatiotemporal aspects of Arabic pharyngealization,” *Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 2091–2091, 2012.

[9] M. Barlaz, M. Fu, J. Dubin, R. Shosted, Z. Liang, and B. Sutton, “Lingual differences in Brazilian Portuguese oral and nasal vowels: an MRI study,” in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015. [Online]. Available: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0819.pdf

[10] C. Carignan, R. Shosted, M. Fu, Z.-P. Liang, and B. P. Sutton, “The role of the pharynx and tongue in enhancement of vowel nasalization: a real-time MRI investigation of French nasal vowels.” in *Proceedings of the INTERSPEECH Conference*, 2013, pp. 3042–3046.

[11] N. Wong, M. Fu, Z.-P. Liang, R. Shosted, and B. P. Sutton, “Observations of perseverative coarticulation in lateral approximants using MRI.” in *Proceedings of the INTERSPEECH conference*, 2013, pp. 612–616.

[12] M. Barlaz, M. Fu, Z.-P. Liang, R. Shosted, and B. P. Sutton, “Deformation-based articulatory representations of speech sounds.” in *Proceedings of the 15th Conference on Laboratory Phonology*, 2016. [Online]. Available: https://labphon.org/labphon15/long_abstracts/LabPhon15_Revised_abstract_200.pdf

[13] Z. Hermes, M. Fu, S. Rose, R. Shosted, and B. P. Sutton, “Representations of place and airstream mechanism: A real-time MRI study of Tigrinya ejectives.” in *Proceedings of the 15th Conference on Laboratory Phonology*, 2016.

[14] Z. Hermes, M. Barlaz, R. Shosted, M. Fu, and B. P. Sutton, “The articulatory configuration of the pharynx during the voiced and voiceless pharyngeal fricatives in Gulf and Levantine Arabic: A real-time magnetic resonance imaging study.” in *Proceedings of the 30th Annual Symposium on Arabic Linguistics*, 2016.

[15] D. Kuehn, “Cineradiographic investigation of velar movement variables among normals,” in *Cleft Palate Journal*, vol. 12, 1976, pp. 338–339.

[16] M. Fu, “Dynamic speech imaging with low-rank approximation,” Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2012.

[17] M. Stone, “Laboratory techniques for investigating speech articulation,” *The Handbook of Phonetic Sciences*, pp. 11–32, 1997.

[18] K. A. Kendall and R. J. Leonard, *Laryngeal Evaluation: Indirect Laryngoscopy to High-Speed Digital Imaging*. Thieme, 2011.

[19] Z. Li, H. Bakhshaee, L. Helou, L. Mongeau, K. Kost, C. Rosen, and K. Verdolini, “Evaluation of contact pressure in human vocal folds during phonation using high-speed videoendoscopy, electroglottography, and magnetic resonance imaging,” in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, 2013, p. 060306.

[20] S. Rajan, M. Kurien, A. Gupta, S. Mathews, R. Albert, and D. Tychicus, “Velopharyngeal incompetence in patients with cleft palate, flexible video pharyngoscopy and perceptual speech assessment: a correlational pilot study,” *Journal of Laryngology & Otology*, vol. 128, no. 11, pp. 986–990, 2014.

[21] D. D. Deliyski, R. E. Hillman, and D. D. Mehta, “Laryngeal high-speed videoendoscopy: Rationale and recommendation for accurate and consistent terminology,” *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 5, pp. 1488–1492, 2015.

[22] G. J. Capilouto, E. D. Frederick, and H. Challa, "Measurement of infant tongue thickness using ultrasound: a technical note," *Journal of Clinical Ultrasound*, vol. 40, no. 6, pp. 364–367, 2012.

[23] A. Lee, N. Zharkova, F. Gibbon, M. Ball, and F. Gibbon, "Vowel imaging," *Handbook of Vowels and Vowel Disorders*, pp. 138–159, 2013.

[24] A. Fenster and D. B. Downey, "3-D ultrasound imaging: a review," *IEEE Engineering in Medicine and Biology magazine*, vol. 15, no. 6, pp. 41–51, 1996.

[25] J. D'hooge, A. Heimdal, F. Jamal, T. Kukulski, B. Bijnens, F. Rademakers, L. Hatle, P. Suetens, and G. Sutherland, "Regional strain and strain rate measurements by cardiac ultrasound: Principles, implementation and limitations," *European Heart Journal-Cardiovascular Imaging*, vol. 1, no. 3, pp. 154–170, 2000.

[26] T. R. Nelson and D. H. Pretorius, "Three-dimensional ultrasound imaging," *Ultrasound in Medicine & Biology*, vol. 24, no. 9, pp. 1243–1270, 1998.

[27] J. L. Perry, D. P. Kuehn, B. P. Sutton, and X. Fang, "Velopharyngeal structural and functional assessment of speech in young children using dynamic magnetic resonance imaging," *Cleft Palate-Craniofacial Journal*, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27031268

[28] M. E. Spector, E. Callaway, E. L. McKean, and M. E. Prince, "Videofluoroscopic-guided botulinum toxin injections for pharyngoesophageal spasm after total laryngectomy," *The Laryngoscope*, vol. 123, no. 2, pp. 394–397, 2013.

[29] H. Su, A. Khorsandi, J. Silberzweig, A. Kobren, M. Urken, M. Amin, R. Branski, and C. Lazarus, "Temporal and physiologic measurements of deglutition in the upright and supine position with videofluoroscopy (VFS) in healthy subjects," *Dysphagia*, vol. 30, no. 4, pp. 438–444, 2015.

[30] M. Lafer, S. Achlatis, C. Lazarus, Y. Fang, R. C. Branski, and M. R. Amin, "Temporal measurements of deglutition in dynamic magnetic resonance imaging versus videofluoroscopy," *Annals of Otology, Rhinology & Laryngology*, vol. 122, no. 12, pp. 748–753, 2013.

[31] M. T. Heller, C. C. Meltzer, M. B. Fukui, C. A. Rosen, S. Chander, M. A. Martinelli, and D. W. Townsend, "Superphysiologic fdg uptake in the non-paralyzed vocal cord: Resolution of a false-positive PET result with combined PET-CT imaging," *Clinical Positron Imaging*, vol. 3, no. 5, pp. 207–211, 2000.

[32] C. T. Ferrand, "Speech science: an integrated approach to theory and clinical practice," *Ear and Hearing*, vol. 22, no. 6, p. 549, 2001.

[33] M. Yoshikawa, M. Yoshida, K. Tsuga, Y. Akagawa, and M. E. Groher, "Comparison of three types of tongue pressure measurement devices," *Dysphagia*, vol. 26, no. 3, pp. 232–237, 2011.

[34] A. M. Sulter, D. G. Miller, R. F. Wolf, H. K. Schutte, H. P. Wit, and E. L. Mooyaart, "On the relation between the dimensions and resonance characteristics of the vocal tract: a study with MRI," *Magnetic Resonance Imaging*, vol. 10, no. 3, pp. 365–373, 1992.

[35] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech MRI: Morphology and function," *Physica Medica*, vol. 30, no. 6, pp. 604–618, 2014.

[36] K. Bettens, F. L. Wuyts, and K. M. Van Lierde, "Instrumental assessment of velopharyngeal function and resonance: a review," *Journal of Communication Disorders*, vol. 52, pp. 170–183, 2014.

[37] J. Perry and G. Schenck, "Instrumental assessment in cleft palate care," *SIG 5 Perspectives on Speech Science and Orofacial Disorders*, vol. 23, no. 2, pp. 49–61, 2013.

[38] A. A. Kane, J. A. Butman, R. Mullick, M. Skopec, and P. Choyke, "A new method for the study of velopharyngeal function using gated magnetic resonance imaging," *Plastic and Reconstructive Surgery*, vol. 109, no. 2, pp. 472–481, 2002.

[39] M. S. NessAiver, M. Stone, V. Parthasarathy, Y. Kahana, and A. Paritsky, "Recording high quality speech during tagged cine-MRI studies using a fiber optic microphone," *Journal of Magnetic Resonance Imaging*, vol. 23, no. 1, pp. 92–97, 2006.

[40] Y. Zhu, Y.-C. Kim, M. I. Proctor, S. S. Narayanan, and K. S. Nayak, "Dynamic 3-D visualization of vocal tract shaping during speech," *Medical Imaging, IEEE Transactions on*, vol. 32, no. 5, pp. 838–848, 2013.

[41] M. Stone, E. P. Davis, A. S. Douglas, M. NessAiver, R. Gullapalli, W. S. Levine, and A. Lundberg, "Modeling the motion of the internal tongue from tagged cine-MRI images," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 2974–2982, 2001.

[42] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech MRI," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 28–44, 2016.

[43] Z.-P. Liang, "Spatiotemporal imaging with partially separable functions," in *Biomedical Imaging (ISBI), 6th IEEE International Symposium on*, 2007, pp. 181–182.

[44] F. Lam and Z.-P. Liang, "A subspace approach to high-resolution spectroscopic imaging," *Magnetic Resonance in Medicine*, vol. 71, no. 4, pp. 1349–1357, 2014. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/mrm.25168/abstract

[45] C. Ma, F. Lam, C. L. Johnson, and Z.-P. Liang, "Removal of nuisance signals from limited and sparse 1H MRSI data using a union-of-subspaces model," *Magnetic Resonance in Medicine*, vol. 75, no. 2, pp. 488–497, 2016. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/mrm.25635/abstract

[46] A. G. Christodoulou, S. D. Babacan, and Z.-P. Liang, "Accelerating cardiovascular imaging by exploiting regional low-rank structure via group sparsity," in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, 2012, pp. 330–333.

[47] A. G. Christodoulou, H. Zhang, B. Zhao, T. K. Hitchens, C. Ho, and Z.-P. Liang, "High-resolution cardiovascular MRI by integrating parallel imaging with low-rank and sparse modeling," *Biomedical Engineering, IEEE Transactions on*, vol. 60, no. 11, pp. 3083–3092, 2013.

[48] B. Zhao, J. P. Haldar, A. G. Christodoulou, and Z.-P. Liang, "Image reconstruction from highly undersampled-space data with joint partial separability and sparsity constraints," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 9, pp. 1809–1820, 2012.

[49] S. G. Lingala, Y. Hu, E. DiBella, and M. Jacob, "Accelerated dynamic MRI exploiting sparsity and low-rank structure: kt SLR," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 5, pp. 1042–1054, 2011.

[50] M. Fu, B. Zhao, C. Carignan, R. K. Shosted, J. L. Perry, D. P. Kuehn, Z.-P. Liang, and B. P. Sutton, "High-resolution dynamic speech imaging with joint low-rank and sparsity constraints," *Magnetic Resonance in Medicine*, vol. 73, no. 5, pp. 1820–1832, 2015.

[51] M. Fu, M. Barlaz, J. Holtrop, R. K. Shosted, J. L. Perry, D. P. Kuehn, Z.-P. Liang, and B. P. Sutton, "High-resolution dynamic speech imaging with joint low-rank and sparsity constraints," *Magnetic Resonance in Medicine (In Press)*, 2016.

[52] J. P. Haldar and Z.-P. Liang, "Spatiotemporal imaging with partially separable functions: a matrix recovery approach," in *Biomedical Imaging: From Nano to Macro, IEEE International Symposium on*, 2010, pp. 716–719.

[53] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[54] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung, "Guarantees of Riemannian optimization for low rank matrix completion," *arXiv preprint arXiv:1603.06610*, 2016. [Online]. Available: https://arxiv.org/abs/1603.06610

[55] M. Bai, X. Zhang, G. Ni, and C. Cui, "An adaptive correction approach for tensor completion," *SIAM Journal on Imaging Sciences*, vol. 9, no. 3, pp. 1298–1323, 2016.

[56] J. Tsao, P. Boesiger, and K. P. Pruessmann, "k-t BLAST and k-t SENSE: dynamic MRI with high frame rate exploiting spatiotemporal correlations," *Magnetic Resonance in Medicine*, vol. 50, no. 5, pp. 1031–1042, 2003.

[57] D. Xu, K. F. King, and Z.-P. Liang, "Improving k-t SENSE by adaptive regularization," *Magnetic Resonance in Medicine*, vol. 57, no. 5, pp. 918–930, 2007.

[58] N. Aggarwal and Y. Bresler, "Patient-adapted reconstruction and acquisition dynamic imaging method (paradigm) for MRI," *Inverse Problems*, vol. 24, no. 4, p. 045015, 2008.

[59] C. Mistretta, O. Wieben, J. Velikina, W. Block, J. Perry, Y. Wu, and K. Johnson, "Highly constrained backprojection for time-resolved MRI," *Magnetic Resonance in Medicine*, vol. 55, no. 1, pp. 30–40, 2006.

[60] D. Liang, E. V. DiBella, R.-R. Chen, and L. Ying, "k-t ISD: dynamic cardiac MR imaging using compressed sensing with iterative support detection," *Magnetic Resonance in Medicine*, vol. 68, no. 1, pp. 41–53, 2012.

[61] M. Usman, C. Prieto, T. Schaeffter, and P. Batchelor, "k-t group sparse: a method for accelerating dynamic MRI," *Magnetic Resonance in Medicine*, vol. 66, no. 4, pp. 1163–1176, 2011.

[62] M. Lustig, J. M. Santos, D. L. Donoho, and J. M. Pauly, "Kt SPARSE: High frame rate dynamic MRI exploiting spatio-temporal sparsity," in *Proceedings of the 13th Annual Meeting the International Society of Magnetic Resonance and Medicine, Seattle*, vol. 2420, 2006.

[63] U. Gamper, P. Boesiger, and S. Kozerke, "Compressed sensing in dynamic MRI," *Magnetic Resonance in Medicine*, vol. 59, no. 2, pp. 365–373, 2008.

[64] H. Jung, K. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye, "k-t FOCUSS: a general compressed sensing framework for high resolution dynamic MRI," *Magnetic Resonance in Medicine*, vol. 61, no. 1, pp. 103–116, 2009.

[65] J. D. Trzasko, C. R. Haider, E. A. Borisch, N. G. Campeau, J. F. Glockner, S. J. Riederer, and A. Manduca, "Sparse-CAPR: Highly accelerated 4D CE-MRA with parallel imaging and nonconvex compressive sensing," *Magnetic Resonance in Medicine*, vol. 66, no. 4, pp. 1019–1032, 2011.

[66] M. Murphy, M. Alley, J. Demmel, K. Keutzer, S. Vasanawala, and M. Lustig, "Fast-SPIRiT compressed sensing parallel imaging MRI: Scalable parallel implementation and clinically feasible runtime," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 6, pp. 1250–1262, 2012.

[67] F. Huang, W. Lin, G. R. Duensing, and A. Reykowski, "k-t sparse GROWL: Sequential combination of partially parallel imaging and compressed sensing in k-t space using flexible virtual coil," *Magnetic Resonance in Medicine*, vol. 68, no. 3, pp. 772–782, 2012.

[68] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9, pp. 589–592, 2008.

[69] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[70] M. A. Davenport and M. B. Wakin, "Analysis of orthogonal matching pursuit using the restricted isometry property," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4395–4401, 2010.

[71] F. Krahmer and R. Ward, "New and improved johnson-lindenstrauss embeddings via the restricted isometry property," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 3, pp. 1269–1281, 2011.

[72] A. C. Freitas, M. Wylezinska, M. J. Birch, S. E. Petersen, and M. E. Miquel, “Comparison of Cartesian and non-Cartesian real-time MRI sequences at 1.5 T to assess velar motion and velopharyngeal closure during speech,” *PloS One*, vol. 11, no. 4, p. e0153322, 2016.

[73] J. d’Arcy, D. Collins, I. Rowland, A. Padhani, and M. Leach, “Applications of sliding window reconstruction with cartesian sampling for dynamic contrast enhanced MRI,” *NMR in Biomedicine*, vol. 15, no. 2, pp. 174–183, 2002.

[74] M. Fu, B. Zhao, J. Holtrop, J. Perry, D. Kuehn, Z. Liang, and 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 the International Society of Magnetic Resonance and Medicine, Salt Lake City, Utah, USA*, 2013, p. 4269.

[75] M. Fu, A. G. Christodoulou, A. T. Naber, D. P. Kuehn, Z.-P. Liang, and B. P. Sutton, “High-frame-rate multislice speech imaging with sparse sampling of (k, t)-space,” in *Proceedings of the 20th Annual Meeting the International Society of Magnetic Resonance and Medicine, Melbourn, Australia*, vol. 12, 2012.

[76] M. Fu, M. S. Barlaz, J. L. Holtrop, J. L. Perry, D. P. Kuehn, R. K. Shosted, Z.-P. Liang, and B. P. Sutton, “High-frame-rate full-vocal-tract 3D dynamic speech imaging,” *Magnetic Resonance in Medicine*, 2016. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/mrm.26248/abstract

[77] G. H. Glover and C. S. Law, “Spiral-in/out BOLD fMRI for increased SNR and reduced susceptibility artifacts,” *Magnetic Resonance in Medicine*, vol. 46, no. 3, pp. 515–522, 2001.

[78] G. H. Glover and S. Lai, “Self-navigated spiral fMRI: Interleaved versus single-shot,” *Magnetic Resonance in Medicine*, vol. 39, no. 3, pp. 361–368, 1998.

[79] G. H. Glover et al., “Simple analytic spiral k-space algorithm,” *Magnetic Resonance in Medicine*, vol. 42, no. 2, pp. 412–415, 1999.

[80] P. T. Gurney, B. A. Hargreaves, and D. G. Nishimura, “Design and analysis of a practical 3D cones trajectory,” *Magnetic Resonance in Medicine*, vol. 55, no. 3, pp. 575–582, 2006.

[81] P. Irarrazabal and D. G. Nishimura, “Fast three dimensional magnetic resonance imaging,” *Magnetic Resonance in Medicine*, vol. 33, no. 5, pp. 656–662, 1995.

[82] E. Staroswiecki, N. K. Bangerter, P. T. Gurney, T. Grafendorfer, G. E. Gold, and B. A. Hargreaves, “In vivo sodium imaging of human patellar cartilage with a 3D cones sequence at 3 T and 7 T,” *Journal of Magnetic Resonance Imaging*, vol. 32, no. 2, pp. 446–451, 2010.

[83] S. Kecskemeti, K. Johnson, Y. Wu, C. Mistretta, P. Turski, and O. Wieben, “High resolution three-dimensional cine phase contrast mri of small intracranial aneurysms using a stack of stars k-space trajectory,” *Journal of Magnetic Resonance Imaging*, vol. 35, no. 3, pp. 518–527, 2012.

[84] H. Jung, J. Park, J. Yoo, and J. C. Ye, “Radial k-t FOCUSS for high-resolution cardiac cine MRI,” *Magnetic Resonance in Medicine*, vol. 63, no. 1, pp. 68–78, 2010.

[85] E. B. Welch, A. Manduca, R. C. Grimm, H. A. Ward, and C. R. Jack Jr, "Spherical navigator echoes for full 3D rigid body motion measurement in mri," *Magnetic Resonance in Medicine*, vol. 47, no. 1, pp. 32–41, 2002.

[86] A. Van der Linden, M. Verhoye, V. Van Meir, I. Tindemans, M. Eens, P. Absil, and J. Balthazart, "In vivo manganese-enhanced magnetic resonance imaging reveals connections and functional properties of the songbird vocal control system," *Neuroscience*, vol. 112, no. 2, pp. 467–474, 2002.

[87] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[88] H. K. Vorperian, R. D. Kent, L. R. Gentry, and B. S. Yandell, "Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: Preliminary results," *International Journal of Pediatric Otorhinolaryngology*, vol. 49, no. 3, pp. 197–206, 1999.

[89] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: a study using magnetic resonance imaging," *Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[90] T. Baer, J. Gore, L. Gracco, and P. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *Journal of the Acoustical Society of America*, vol. 90, no. 2, pp. 799–828, 1991.

[91] M. Uecker, S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt, and J. Frahm, "Real-time MRI at a resolution of 20 ms," *NMR in Biomedicine*, vol. 23, no. 8, pp. 986–994, 2010.

[92] D. Demolin, S. Hassid, T. Metens, and A. Soquet, "Real-time MRI and articulatory coordination in speech," *Comptes rendus biologies*, vol. 325, no. 4, pp. 547–556, 2002.

[93] K. Mády, R. Sader, A. Zimmermann, P. Hoole, A. Beer, H.-F. Zeilhofer, and C. Hannig, "Assessment of consonant articulation in glossectomee speech by dynamic MRI." in *Proceedings of the INTERSPEECH Conference*. Citeseer, 2002.

[94] A. Scott, R. Boubertakh, M. Birch, and M. Miquel, "Towards clinical assessment of velopharyngeal closure using MRI: evaluation of real-time MRI sequences at 1.5 and 3 t," *British Journal of Radiology*, 2014.

[95] X. Feng, J. Inouye, S. Blemker, K. Lin, K. Borowitz, T. Altes, T. Kovach, W. El-Nahal, C. Pelland, and C. H. Meyer, "Assessment of velopharyngeal function with multi-planar high-resolution real-time spiral dynamic MRI," *Proceedings of the 21st Annual Meeting the International Society of Magnetic Resonance and Medicine (Scientific Sessions)*, vol. 1228, 2013.

[96] B. Hargreaves, "Rapid gradient-echo imaging," *Journal of Magnetic Resonance Imaging*, vol. 36, no. 6, pp. 1300–1313, 2012.

[97] B. P. Sutton, C. A. Conway, Y. Bae, R. Seethamraju, and D. P. Kuehn, “Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3 T,” *Journal of Magnetic Resonance Imaging*, vol. 32, no. 5, pp. 1228–1237, 2010.

[98] G.-C. Ngo, J. L. Holtrop, M. Fu, F. Lam, and B. P. Sutton, “High temporal resolution fmri with partial separability model,” in *Engineering in Medicine and Biology Society (EMBC), the 37th Annual International Conference of the IEEE*, 2015, pp. 7482–7485.

[99] H. Ur and D. Gross, “Improved resolution from subpixel shifted pictures,” *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 2, pp. 181–186, 1992.

[100] J. Shi, Y. Chi, and N. Zhang, “Multichannel sampling and reconstruction of bandlimited signals in fractional Fourier domain,” *IEEE Signal Processing Letters*, vol. 17, no. 11, pp. 909–912, 2010.

[101] J. Brown, “Multi-channel sampling of low-pass signals,” *IEEE Transactions on Circuits and Systems*, vol. 28, no. 2, pp. 101–106, 1981.

[102] D. Wei, Q. Ran, and Y. Li, “Reconstruction of band-limited signals from multichannel and periodic nonuniform samples in the linear canonical transform domain,” *Optics Communications*, vol. 284, no. 19, pp. 4307–4315, 2011.

[103] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, P. Boesiger et al., “SENSE: Sensitivity encoding for fast MRI,” *Magnetic Resonance in Medicine*, vol. 42, no. 5, pp. 952–962, 1999.

[104] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase, “Generalized autocalibrating partially parallel acquisitions (GRAPPA),” *Magnetic Resonance in Medicine*, vol. 47, no. 6, pp. 1202–1210, 2002.

[105] M. Blaimer, F. A. Breuer, N. Seiberlich, M. F. Mueller, R. M. Heidemann, V. Jellus, G. Wiggins, L. L. Wald, M. A. Griswold, and P. M. Jakob, “Accelerated volumetric MRI with a SENSE/GRAPPA combination,” *Journal of Magnetic Resonance Imaging*, vol. 24, no. 2, pp. 444–450, 2006.

[106] M. Blaimer, F. Breuer, M. Mueller, R. M. Heidemann, M. A. Griswold, and P. M. Jakob, “SMASH, SENSE, PILS, GRAPPA: how to choose the optimal method,” *Topics in Magnetic Resonance Imaging*, vol. 15, no. 4, pp. 223–236, 2004.

[107] M. Uecker, P. Lai, M. J. Murphy, P. Virtue, M. Elad, J. M. Pauly, S. S. Vasanawala, and M. Lustig, “ESPIRiT an eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA,” *Magnetic Resonance in Medicine*, vol. 71, no. 3, pp. 990–1001, 2014.

[108] P. J. Koopmans, “Two-dimensional-NGC-SENSE-GRAPPA for fast, ghosting-robust reconstruction of in-plane and slice-accelerated blipped-CAIPI echo planar imaging,” *Magnetic Resonance in Medicine*, 2016.

[109] Y.-C. Kim, M. I. Proctor, S. S. Narayanan, and K. S. Nayak, "Improved imaging of lingual articulation using real-time multislice MRI," *Journal of Magnetic Resonance Imaging*, vol. 35, no. 4, pp. 943–948, 2012.

[110] K. S. Nayak, C. H. Cunningham, J. M. Santos, and J. M. Pauly, "Real-time cardiac MRI at 3 tesla," *Magnetic Resonance in Medicine*, vol. 51, no. 4, pp. 655–660, 2004.

[111] K. S. Nayak, J. M. Pauly, D. G. Nishimura, and B. S. Hu, "Rapid ventricular assessment using real-time interactive multislice MRI," *Magnetic Resonance in Medicine*, vol. 45, no. 3, pp. 371–375, 2001.

[112] Y. Yang, W. Engelien, S. Xu, H. Gu, D. A. Silbersweig, and E. Stern, "Transit time, trailing time, and cerebral blood flow during brain activation: Measurement using multislice, pulsed spin-labeling perfusion imaging," *Magnetic Resonance in Medicine*, vol. 44, no. 5, pp. 680–685, 2000.

[113] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein et al., "USCTIMIT: a database of multimodal speech production data," Tech. Rep., 2013. [Online]. Available: http://sail.usc.edu/span/usc-timit/usctimit_report.pdf

[114] T. Sorensen, A. Toutios, L. Goldstein, and S. Narayanan, "Characterizing vocal tract dynamics across speakers using real-time mri," *Proceedings of the INTERSPEECH conference*, pp. 465–469, 2016.

[115] M. Fu, J. Zhuang, F. Hou, X. Ning, Q. Zhan, and Y. Shao, "Extracting human gait series based on the wavelet transform," *Acta Physica Sinica*, vol. 59, no. 6, pp. 4343–4350, 2010.

[116] M. Fu, J. Zhuang, F. Hou, Q. Zhan, Y. Shao, and X. Ning, "A method for extracting human gait series from accelerometer signals based on the ensemble empirical mode decomposition," *Chinese Physics B*, vol. 19, no. 5, p. 058701, 2010.

[117] F. Hou, X. Ning, J. Zhuang, X. Huang, M. Fu, and C. Bian, "High-dimensional time irreversibility analysis of human interbeat intervals," *Medical Engineering & Physics*, vol. 33, no. 5, pp. 633–637, 2011.

[118] M. Zhang, C. Ge, M. Lu, Z. Zhang, and B. T. Cunningham, "A self-referencing biosensor based upon a dual-mode external cavity laser," *Applied Physics Letters*, vol. 102, no. 21, p. 213701, 2013.

[119] M. Zhang, J. Peh, P. J. Hergenrother, and B. T. Cunningham, "Detection of protein–small molecule binding using a self-referencing external cavity laser biosensor," *Journal of the American Chemical Society*, vol. 136, no. 16, pp. 5840–5843, 2014.

[120] C.-Y. Lu, C.-Y. Ni, M. Zhang, S. L. Chuang, and D. H. Bimberg, "Metal-cavity surface-emitting microlasers with size reduction: Theory and experiment," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 5, pp. 1–9, 2013.

[121] A. Matsudaira, C.-Y. Lu, M. Zhang, S. L. Chuang, E. Stock, and D. Bimberg, "Cavity-volume scaling law of quantum-dot metal-cavity surface-emitting microlasers," *IEEE Photonics Journal*, vol. 4, no. 4, pp. 1103–1114, 2012.

[122] C. Ge, M. Lu, Z. Zhang, B. T. Cunningham et al., "A self-referencing biosensor based upon a dual-mode external cavity laser," in *Proceedings of CLEO: Science and Innovations*. Optical Society of America, 2013, pp. CM2H–6.

[123] C.-Y. Lu, M. Zhang, S. L. Chuang, E. Stock, and D. Bimberg, "Size dependence of quantum-dot metal-cavity surface-emitting microlaser," in *Proceedings of 2012 International Semiconductor Laser Conference (ISLC)*, 2012.

[124] M. Zhang, "External cavity laser biosensor for label-free detection," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2015.

[125] M. Zhang and B. T. Cunningham, "Liquid-tuned plasmonic external cavity laser," in *Proceedings of CLEO: Applications and Technology*. Optical Society of America, 2014, pp. AF1L–3.

[126] D. Aalto, J. Malinen, P. Palo, O. Aaltonen, M. Vainio, R.-P. Happonen, R. Parkkola, and J. Saunavaara, "Recording speech sound and articulation in MRI," in *Biodevices*, 2011, pp. 168–173.

[127] R. A. Mattson, T. J. Krochta, and H. K. Tuy, "Voice actuated volume image controller and display controller," 1994, US Patent 5,303,148.

[128] J. Malinen and P. Palo, "Recording speech during MRI: Part II," in *MAVEBA*, 2009, pp. 211–214.

[129] M. Barlaz, C. Carignan, R. Shosted, S. Johnson, Z.-P. Liang, M. Fu, and B. Sutton, "Understanding the relationship between acoustics and articulation of nasal and oral vowels," in *Proceedings of the 5th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan, Honolulu, Hawaii, USA*, 2016, p. 3pSC.

[130] D. P. Kuehn and K. T. Moller, "Speech and language issues in the cleft palate population: The state of the art," *Cleft Palate-Craniofacial Journal*, vol. 37, no. 4, pp. 348–348, 2000.

[131] A. Kummer, *Cleft Palate & Craniofacial Anomalies: effects On Speech and Resonance*. Nelson Education, 2013.

[132] J. C. McGowan III, H. Hatabu, D. M. Yousem, P. Randall, and H. Y. Kressel, "Evaluation of soft palate function with MRI: application to the cleft palate patient." *Journal of Computer Assisted Tomography*, vol. 16, no. 6, pp. 877–882, 1992.

[133] S. G. Lingala, Y. Zhu, Y.-C. Kim, A. Toutios, S. Narayanan, and K. Nayak, "High-frame-rate real-time imaging of speech production," *SPIE News Room*, 2015. [Online]. Available: http://spie.org/x113691.xml

[134] S. Balaji, "Bilateral cleft lip and palate, hypertelorism with agenesis of corpus callosum," *Indian Journal of Dental Research*, vol. 27, no. 1, p. 100, 2016.

[135] L. Lam and N. Samman, "Speech and swallowing following tongue cancer surgery and free flap reconstruction – a systematic review," *Oral Oncology*, vol. 49, no. 6, pp. 507–524, 2013.

[136] J. Lee, J. Woo, F. Xing, E. Z. Murano, M. Stone, and J. L. Prince, "Semi-automatic segmentation of the tongue for 3D motion analysis with dynamic MRI," in *Biomedical Imaging (ISBI), IEEE 10th International Symposium on*, 2013, pp. 1465–1468.

[137] C. Hagedorn, A. Lammert, M. Bassily, Y. Zu, U. Sinha, L. Goldstein, and S. S. Narayanan, "Characterizing postglossectomy speech using real-time MRI," in *International Seminar on Speech Production, Cologne, Germany*, 2014.

[138] A. Hemraj, J. Back, and T. George, "1263: a case of laryngitis caused by an unusual suspect," *Critical Care Medicine*, vol. 43, no. 12, pp. 317–318, 2015.

[139] W.-G. WANG and J.-P. LU, "Clinial analysis of 106 children with acute laryngitis," *Journal of Medical Recapitulate*, vol. 1, p. 060, 2012.

[140] S. Padilla, A. Sebring, P. Jani, J. Kane, and C. Clardy, "1262: aspirin-induced acute interstitial nephritis in a pediatric liver transplant patient," *Critical Care Medicine*, vol. 43, no. 12, p. 317, 2015.

[141] C.-C. Huang, Y.-S. Leu, C.-F. J. Kuo, W.-L. Chu, Y.-H. Chu, and H.-C. Wu, "Automatic recognizing of vocal fold disorders from glottis images," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 228, no. 9, pp. 952–961, 2014.

[142] M. Zañartu, B. D. Erath, S. D. Peterson, R. E. Hillman, and G. R. Wodicka, "Modeling incomplete glottal closure due to a posterior glottal opening and its effects on the dynamics of the vocal folds," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, 2013, p. 060239.

[143] L. S. Reder and R. A. Franco Jr, "Benign vocal fold lesions," *Practical Laryngology*, p. 27, 2015.

[144] K. V. Kumar, V. Shankar, and R. Santosham, "Assessment of swallowing and its disorders a dynamic MRI study," *European Journal of Radiology*, vol. 82, no. 2, pp. 215–219, 2013.

[145] M. Ohkubo, T. Higaki, K. Nishikawa, M. Otonari-Yamamoto, T. Sugiyama, R. Ishida, and M. Wakoh, "Optimal contrast enhancement liquid for dynamic MRI of swallowing," *Journal of Oral Rehabilitation*, vol. 43, no. 9, pp. 678–682, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27328011

[146] W. G. Pearson Jr and A. C. Zumwalt, "Visualising hyolaryngeal mechanics in swallowing using dynamic MRI," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 2, no. 4, pp. 208–216, 2014.

[147] K. Sinko, C. Czerny, R. Jagsch, A. Baumann, and C. Kulinna-Cosentini, “Dynamic 1.5-T vs 3-T true fast imaging with steady-state precession (true fisp)-MRI sequences for assessment of velopharyngeal function,” *Dentomaxillofacial Radiology*, vol. 44, no. 8, p. 20150028, 2015.

[148] F. Özgür, G. Tuncbilek, and A. Cila, “Evaluation of velopharyngeal insufficiency with magnetic resonance imaging and nasoendoscopy.” *Annals of Plastic Surgery*, vol. 44, no. 1, pp. 8–13, 2000.

[149] H. S. Magen, A. M. Kang, M. K. Tiede, and D. Whalen, “Posterior pharyngeal wall position in the production of speech,” *Journal of Speech, Language, and Hearing Research*, vol. 46, no. 1, pp. 241–251, 2003.

[150] D. F. Johns, R. J. Rohrich, and M. Awada, “Velopharyngeal incompetence:: a guide for clinical evaluation,” *Plastic and Reconstructive Surgery*, vol. 112, no. 7, pp. 1890–1898, 2003.

[151] B. N. Johnson, C. Russell, R. M. Khan, and N. Sobel, “A comparison of methods for sniff measurement concurrent with olfactory tasks in humans,” *Chemical Senses*, vol. 31, no. 9, pp. 795–806, 2006.

[152] S. D. Edwards, “Minimally invasive apparatus for internal ablation of turbinates,” 1998, US Patent 5,827,277.

[153] D. H. Klatt, “Linguistic uses of segmental duration in English: acoustic and perceptual evidence,” *Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1208–1221, 1976.

[154] R. Jakobson and M. Halle, *Fundamentals of Language*. Walter de Gruyter, 2002, vol. 1.

[155] J. B. Pierrehumbert, “Phonological representation: Beyond abstract versus episodic,” *Annual Review of Linguistics*, vol. 2, pp. 33–52, 2016.

[156] C. A. Fowler, “The segment in articulatory phonology,” *The Segment in Phonetics and Phonology*, p. 25, 2015.

[157] M. F. Assaneo, J. Sitt, G. Varoquaux, M. Sigman, L. Cohen, and M. A. Trevisan, “Exploring the anatomical encoding of voice with a mathematical model of the vocal system,” *NeuroImage*, vol. 141, pp. 31–39, 2016.

[158] K. E. Bouchard and E. F. Chang, “Control of spoken vowel acoustics and the influence of phonetic context in human speech sensorimotor cortex,” *Journal of Neuroscience*, vol. 34, no. 38, pp. 12 662–12 677, 2014.

[159] S. Aryal and R. Gutierrez-Osuna, “Accent conversion through cross-speaker articulatory synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7694–7698.

[160] G. Yule, *The Study of Language*. Cambridge University Press, 2014.

[161] W. Ziegler and H. Ackermann, “Neural bases of phonological and articulatory processing,” *The Oxford Handbook of Language Production*, p. 275, 2014.

[162] W. Ziegler and H. Ackermann, “Neuromotor speech impairment: It’s all in the talking,” *Folia Phoniatrica et Logopaedica*, vol. 65, no. 2, pp. 55–67, 2013.

[163] S. Serafini, J. M. Komisarow, W. Gallentine, M. A. Mikati, M. J. Bonner, P. G. Kranz, M. M. Haglund, and G. Grant, “Reorganization and stability for motor and language areas using cortical stimulation: case example and review of the literature,” *Brain Sciences*, vol. 3, no. 4, pp. 1597–1614, 2013.

[164] L. Traser, C. Spahn, B. Richter, T. Baumann, M. Schumacher, and M. Echternach, “Real-time and three-dimensional MRI for diagnosis of pharyngoceles,” *Journal of Magnetic Resonance Imaging*, vol. 40, no. 1, pp. 55–57, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24395345

[165] B. T. Cunningham, M. Zhang et al., “Biosensing lasers using external cavities coupled with photonic and plasmonic crystals,” in *Optical Sensors*. Optical Society of America, 2015, pp. SeT2D–1.

[166] B. T. Cunningham, M. Zhang, Y. Zhuo, L. Kwon, and C. Race, “Recent advances in biosensing with photonic crystal surfaces: a review,” *IEEE Sensors Journal*, vol. 16, no. 10, pp. 3349–3366, 2015.

[167] W. Chen, K. D. Long, M. Lu, V. Chaudhery, H. Yu, J. S. Choi, J. Polans, Y. Zhuo, B. A. Harley, and B. T. Cunningham, “Photonic crystal enhanced microscopy for imaging of live cell adhesion,” *Analyst*, vol. 138, no. 20, pp. 5886–5894, 2013.

[168] M. Zhang, M. Lu, C. Ge, and B. T. Cunningham, “Plasmonic external cavity laser refractometric sensor,” *Optics Express*, vol. 22, no. 17, pp. 20 347–20 357, 2014.

[169] M. Zhang, C.-P. Huang, G.-D. Wang, and Y.-Y. Zhu, “Theory of extraordinary light transmission through sub-wavelength circular hole arrays,” *Journal of Optics*, vol. 12, no. 1, p. 015004, 2009.

[170] R. D. Peterson, W. Chen, B. T. Cunningham, and J. E. Andrade, “Enhanced sandwich immunoassay using antibody-functionalized magnetic iron-oxide nanoparticles for extraction and detection of soluble transferrin receptor on a photonic crystal biosensor,” *Biosensors and Bioelectronics*, vol. 74, pp. 815–822, 2015.

[171] W. Chen, K. D. Long, H. Yu, Y. Tan, J. S. Choi, B. A. Harley, and B. T. Cunningham, “Enhanced live cell imaging via photonic crystal enhanced fluorescence microscopy,” *Analyst*, vol. 139, no. 22, pp. 5954–5963, 2014.

[172] Y. Zhuo, H. Hu, W. Chen, M. Lu, L. Tian, H. Yu, K. D. Long, E. Chow, W. P. King, S. Singamaneni et al., “Single nanoparticle detection using photonic crystal enhanced microscopy,” *Analyst*, vol. 139, no. 5, pp. 1007–1015, 2014.

[173] W. Chen, K. D. Long, J. Kurniawan, M. Hung, H. Yu, B. A. Harley, and B. T. Cunningham, "Planar photonic crystal biosensor for quantitative label-free cell attachment microscopy," *Advanced Optical Materials*, vol. 3, no. 11, pp. 1623–1632, 2015.

[174] M. Fu and B. P. Sutton, "High-resolution, full-vocal-tract 4D real time speech imaging," in *Proceedings of the Proceedings of Midwest Speech and Language Days*, 2014.

[175] H. Ci, A. Graan, G. González, P. Thompson, A. Hill, and J. S. Duncan, "Mandarin functional MRI language paradigms," *Brain and Behavior*, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27781139

[176] Y. Wang, J. Dang, X. Chen, J. Wei, H. Wang, and K. Honda, "An MRI-based acoustic study of Mandarin vowels." in *Proceedings of the INTERSPEECH Conference*, 2013, pp. 568–571.

[177] M. Proctor, L. H. Lu, Y. Zhu, L. Goldstein, S. Narayanan et al., "Articulation of Mandarin sibilants: a multi-plane realtime MRI study," *Proceedings of Speech Science and Technology*, pp. 113–116, 2012.

[178] J. Zhang, J. Wei, C. Zhang, D. Huang, and J. Dang, "Visualization of Mandarin articulation driven by ultrasound data," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*, 2014, pp. 363–366.

[179] L.-H. Ning, D. Wu, M. Fu, Z.-P. Liang, and B. Sutton, "Tongue shape of Mandarin retroflex consonants: Real-time magnetic resonance imaging data," in *Proceedings of the Midwest Speech and Language Days*, 2013.

[180] C. Carignan, R. K. Shosted, M. Fu, Z.-P. Liang, and B. P. Sutton, "A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of French," *Journal of Phonetics*, vol. 50, pp. 34–51, 2015.

[181] R. Shosted, H. Zainab, M. Fu, L.-H. Ning, Z.-P. Liang, and B. P. Sutton, "Articulating emphasis in Gulf and Levantine Arabic: a rt-MRI approach," in *Proceedings of the Experimental Arabic Linguistics Conference (EXAL)*, 2013.

[182] R. Shosted, M. Fu, L.-H. Ning, A. Benmamoun, Z.-P. Liang, and B. P. Sutton, "An MRI study of gutturals and glottalics in Arabic and Tigrinya," in *Proceedings of the Illinois Symposium on Semitic Linguistics (ISSL)*, 2012.

[183] M. Barlaz, M. Fu, Z.-P. Liang, R. Shosted, and B. Sutton, "The emergence of nasal velar codas in brazilian portuguese: an rt-MRI study," in *Proceedings of 16th Annual Conference of the International Speech Communication Association*, 2015.

[184] M. Fu, M. S. Barlaz, R. K. Shosted, Z.-P. Liang, and B. P. Sutton, "High-resolution dynamic speech imaging with deformation estimation," in *Engineering in Medicine and Biology Society (EMBC), the 37th Annual International Conference of the IEEE*, 2015, pp. 1568–1571.

[185] A. S. Gupta and Z.-P. Liang, “Dynamic imaging by temporal modeling with principal component analysis,” in *Proceedings of 9th Annual Meeting the International Society of Magnetic Resonance and Medicine*, 2001, p. 10.

[186] E. Adalsteinsson, P. Irarrazabal, S. Topp, C. Meyer, A. Macovski, and D. M. Spielman, “Volumetric spectroscopic imaging with spiral-based k-space trajectories,” *Magnetic Resonance in Medicine*, vol. 39, no. 6, pp. 889–898, 1998.

[187] M. S. Blasco, S. Krishnan, D. Moratal, S. Ramamurthy, and M. Brummer, “High spatial frequencies are more dynamic than low spatial frequencies in cardiac motion,” in *Proceedings of the 17th Annual Meeting the International Society of Magnetic Resonance and Medicine, Hawaii, USA*, 2009, pp. 1425–1434.

[188] B. Zhao, J. P. Haldar, C. Brinegar, and Z.-P. Liang, “Low rank matrix recovery for real-time cardiac MRI,” in *Biomedical Imaging: From Nano to Macro, IEEE International Symposium on*, 2010, pp. 996–999.

[189] B. P. Sutton, D. C. Noll, and J. A. Fessler, “Fast, iterative image reconstruction for MRI in the presence of field inhomogeneities,” *Medical Imaging, IEEE Transactions on*, vol. 22, no. 2, pp. 178–188, 2003.

[190] P. Ladefoged and I. Maddieson, “The sounds of the world’s languages,” *Language*, vol. 74, no. 2, pp. 374–376, 1998.

[191] N. Warner and B. V. Tucker, “Phonetic variability of stops and flaps in spontaneous and careful speech,” *Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1606–1617, 2011.

[192] R. Port, F. Mitleb, and M. O’Dell, “Neutralization of obstruent voicing in German is incomplete,” *Journal of the Acoustical Society of America*, vol. 70, no. S1, pp. S13–S13, 1981.

[193] W. Herd, A. Jongman, and J. Sereno, “An acoustic and perceptual analysis of /t/ and /d/ flaps in American English,” *Journal of Phonetics*, vol. 38, no. 4, pp. 504–516, 2010.

[194] A. Braver, “Imperceptible incomplete neutralization: Production, non-identifiability, and non-discriminability in American English flapping,” *Lingua*, vol. 152, pp. 24–44, 2014.

[195] D. P. Kuehn, “A cineradiography investigation of velar movement variables in two normals,” *Cleft Palate-Craniofacial Journal*, vol. 13, pp. 88–103, 1976.

[196] D. Byrd, S. Tobin, E. Bresch, and S. Narayanan, “Timing effects of syllable structure and stress on nasals: a real-time MRI examination,” *Journal of Phonetics*, vol. 37, no. 1, pp. 97–110, 2009.

[197] J. L. Perry, B. P. Sutton, D. P. Kuehn, and J. K. Gamage, “Using MRI for assessing velopharyngeal structures and function,” *Cleft Palate-Craniofacial Journal*, vol. 51, no. 4, pp. 476–485, 2014.

[198] I. P. Association, *Handbook of the International Phonetic Association: a Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[199] A. C. Cohn, "Phonetic and phonological rules of nasalization." Ph.D. dissertation, dissertation.[UCLA Working Papers in Phonetics 76], 1990.

[200] A. C. Cohn, *Phonetic and Phonological Rules of Nasalization*. University Microfilms, 1993.

[201] N. Umeda, "Vowel duration in American English," *Journal of the Acoustical Society of America*, vol. 58, no. 2, pp. 434–445, 1975.

[202] T. Wech, D. Stäb, J. C. Budich, A. Fischer, J. Tran-Gia, D. Hahn, and H. Köstler, "Resolution evaluation of MR images reconstructed by iterative thresholding algorithms for compressed sensing," *Medical Physics*, vol. 39, no. 7, pp. 4328–4338, 2012.

[203] Y.-W. Wen and R. H. Chan, "Parameter selection for total-variation-based image restoration using discrepancy principle," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 1770–1781, 2012.

[204] S. Ramani, Z. Liu, J. Rosen, J. Nielsen, and J. A. Fessler, "Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods," *Image Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 3659–3672, 2012.

[205] J. Gai, N. Obeid, J. L. Holtrop, X.-L. Wu, F. Lam, M. Fu, J. P. Haldar, W. H. Wen-mei, Z.-P. Liang, and B. P. Sutton, "More IMPATIENT: A gridding-accelerated toeplitz-based strategy for non-Cartesian high-resolution 3D MRI on GPUs," *Journal of Parallel and Distributed Computing*, vol. 73, no. 5, pp. 686–697, 2013.

[206] V. Sanguineti, R. Laboissière, and D. J. Ostry, "A dynamic biomechanical model for neural control of speech production," *Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1615–1627, 1998.

[207] M. Reyes-Gomez, N. Jojic, and D. P. Ellis, "Towards single-channel unsupervised source separation of speech mixtures: The layered harmonics/formants separation-tracking model," in *ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing*, 2004.

[208] H. Jung and J. C. Ye, "Motion estimated and compensated compressed sensing dynamic magnetic resonance imaging: What we can learn from video compression techniques," *International Journal of Imaging Systems and Technology*, vol. 20, no. 2, pp. 81–98, 2010.

[209] M. Usman, D. Atkinson, F. Odille, C. Kolbitsch, G. Vaillant, T. Schaeffter, P. G. Batchelor, and C. Prieto, "Motion corrected compressed sensing for free-breathing dynamic cardiac MRI," *Magnetic Resonance in Medicine*, vol. 70, no. 2, pp. 504–516, 2013.

[210] S. G. Lingala, E. DiBella, and M. Jacob, “Deformation corrected compressed sensing (DC-CS): a novel framework for accelerated dynamic MRI,” *Medical Imaging, IEEE Transactions on*, vol. 34, no. 1, pp. 72–85, 2015.

[211] J.-P. Thirion, “Image matching as a diffusion process: an analogy with Maxwell’s demons,” *Medical image analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[212] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, “Non-parametric diffeomorphic image registration with the demons algorithm,” *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007*, pp. 319–326, 2007.

[213] H. Wang, L. Dong, J. O’Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung, “Validation of an accelerated demons algorithm for deformable image registration in radiation therapy,” *Physics in Medicine and Biology*, vol. 50, no. 12, p. 2887, 2005.

[214] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, “Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain,” *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[215] P. Stoica and Y. Selen, “Model-order selection: a review of information criterion rules,” *Signal Processing Magazine, IEEE*, vol. 21, no. 4, pp. 36–47, 2004.

[216] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach.* Springer Science & Business Media, 2003.

[217] M. Peruggia, “Model selection and multimodel inference: a practical information-theoretic approach,” *Journal of the American Statistical Association*, vol. 98, no. 463, pp. 778–779, 2003.

[218] X.-L. Wu, J. Gai, F. Lam, M. Fu, J. P. Haldar, Y. Zhuo, Z.-P. Liang, W.-M. Hwu, and B. P. Sutton, “IMPATIENT MRI: Illinois massively parallel acceleration toolkit for image reconstruction with enhanced throughput in MRI,” in *Biomedical Imaging: From Nano to Macro, IEEE International Symposium on*, 2011, pp. 69–72.

[219] X.-L. Wu, Y. Zhuo, J. Gai, F. Lam, M. Fu, J. P. Haldar, W.-M. Hwu, Z.-P. Liang, and B. P. Sutton, “Advanced MRI reconstruction toolbox with accelerating on GPU,” in *Proceedings of SPIE*, vol. 7872, no. 1, 2011, p. 78720Q.

[220] R. Otazo, E. Candes, and D. K. Sodickson, “Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components,” *Magnetic Resonance in Medicine*, vol. 73, no. 3, pp. 1125–1136, 2015. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/mrm.25240/abstract

[221] R. Otazo, C. Emmanuel, and D. K. Sodickson, “Low-rank & sparse matrix decomposition for accelerated DCE-MRI with background & contrast separation,” in *Proceedings of the ISMRM Workshop on Data Sampling and Image Reconstruction.* Citeseer, 2013.

[222] A. Ganesh, Z. Lin, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast algorithms for recovering a corrupted low-rank matrix," in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2009 3rd IEEE International Workshop on*, 2009, pp. 213–216.

[223] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[224] W.-t. Tan, G. Cheung, and Y. Ma, "Face recovery in conference video streaming using robust principal component analysis," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 3225–3228.

[225] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, vol. 61, 2009.

[226] X. Liang, X. Ren, Z. Zhang, and Y. Ma, "Repairing sparse low-rank texture," in *European Conference on Computer Vision*. Springer, 2012, pp. 482–495.

[227] M. Chen, A. Ganesh, Z. Lin, Y. Ma, J. Wright, and L. Wu, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," *Coordinated Science Laboratory Report no. UILU-ENG 09-2214*, 2009.

[228] Y.-H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries." in *ISMIR*, 2013, pp. 427–432.

[229] J. Woo, F. Xing, J. Lee, M. Stone, and J. L. Prince, "Construction of an unbiased spatio-temporal atlas of the tongue during speech," in *Information Processing in Medical Imaging*. Springer, 2015, pp. 723–732.

[230] J. Woo, J. Lee, E. Z. Murano, F. Xing, M. Al-Talib, M. Stone, and J. L. Prince, "A high-resolution atlas and statistical model of the vocal tract from structural MRI," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 3, no. 1, pp. 47–60, 2015.

[231] J. Woo, M. Stone, and J. L. Prince, "Multimodal registration via mutual information incorporating geometric and spatial context," *Image Processing, IEEE Transactions on*, vol. 24, no. 2, pp. 757–769, 2015.

[232] M. Fu, J. Woo, M. S. Barlaz, R. K. Shosted, Z.-P. Liang, and B. P. Sutton, "Spatiotemporal-atlas-based high-resolution dynamic speech MRI," in *Proceedings of the 24th Annual Meeting and Exhibition of ISMRM*, 2016, p. 874.

[233] M. Fu, J. Woo, Z.-P. Liang, and B. P. Sutton, "Spatiotemporal-atlas-based dynamic speech imaging," in *Proceedings of SPIE Medical Imaging Conference*. International Society for Optics and Photonics, 2016, pp. 978 804–978 804.

[234] K. P. Murphy, *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

[235] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2227–2240, 2014.

[236] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in neural Information Processing Systems*, 2005, pp. 1473–1480.

[237] E. Alpaydin, *Introduction to Machine Learning*. MIT Press, 2014.

[238] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 2013.

[239] Y. Cao, M. Miller, R. L. Winslow, L. Younes et al., "Large deformation diffeomorphic metric mapping of vector fields," *Medical Imaging, IEEE Transactions on*, vol. 24, no. 9, pp. 1216–1230, 2005.

[240] B. B. Avants, N. Tustison, and G. Song, "Advanced normalization tools (ANTS)," *Insight Journal*, vol. 2, pp. 1–35, 2009.

[241] M. De Craene, G. Piella, O. Camara, N. Duchateau, E. Silva, A. Doltra, J. Dhooge, J. Brugada, M. Sitges, and A. F. Frangi, "Temporal diffeomorphic free-form deformation: application to motion and strain estimation from 3D echocardiography," *Medical Image Analysis*, vol. 16, no. 2, pp. 427–450, 2012.

[242] A. Gholipour, C. Limperopoulos, S. Clancy, C. Clouchoux, A. Akhondi-Asl, J. A. Estroff, and S. K. Warfield, "Construction of a deformable spatiotemporal MRI atlas of the fetal brain: evaluation of similarity metrics and deformation models," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*. Springer, 2014, pp. 292–299.

[243] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *Signal Processing Magazine, IEEE*, vol. 21, no. 4, pp. 36–47, 2004.

[244] S. Ramani, Z. Liu, J. Rosen, J. Nielsen, and J. A. Fessler, "Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods," *Image Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 3659–3672, 2012.