USC-SIPI Report #446

Improved Real-Time Magnetic Resonance Imaging of Speech Production

By
Yongwan Lim

December 2020

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
USC Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

Improved Real-Time Magnetic Resonance Imaging of Speech Production

by

Yongwan Lim

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(Electrical and Computer Engineering)

December 2020

Dedication

To my mom, dad, and brother

Acknowledgements

My long journey is about to end. This started on one day of September in 2015, and ever since then, it has never been easy and there were many times I felt frustrated and weak. Without the help and support from many people around me, this dissertation would not have been possible. I would like to express my gratitude to some of the names whose support made me go through hard times.

I have been incredibly fortunate to meet two amazing advisors here at the University of Southern California, Prof. Krishna S. Nayak and Prof. Shrikanth S. Narayanan. They have always been there for me. With their patience, encouragement, and professional guidance, I was able to grow personally and professionally. Prof. Krishna S. Nayak did not just intellectually direct me as my advisor, but also cared about me in so many ways. He always encouraged me to explore my own ideas and guided me to think positively and creatively. I would never know how to thank him. Prof. Shrikanth S. Narayanan is just an incredible person. He always inspired me to see and think big and creative. His sincere care for my study and my well-being is much appreciated. I thank him for his endless support and guidance, and all the new horizons he opened up for me and in speech projects.

Prof. Dani Byrd should also be mentioned here. I was fortunate to work closely with her. I enjoyed all the valuable discussions I have with her about linguistic study design and analysis. I am especially grateful for her guidance and insight for the 3D work.

I would like to express my gratitude to all my qual exam committee members, Prof. Justin P Haldar, Prof. Mahdi Soltanolkotabi, and Prof. Yan Liu for their invaluable advice and feedback on my proposal and during the presentation.

I am very privileged to be a part of the Magnetic Resonance Engineering Laboratory. I thank Sajan Goud Lingala for encouraging and guiding me through research in the early years of my study. I thank Yannick Bliesener for being there as my best friend at USC. I cannot forget the time we shared in and out of our office and during the travels after ISMRMs. I thank Weiyi Chen for being my mate in the speech project and sharing thoughts and ideas. I also thank Ahsan Javed, Vanessa Landes, Xin Miao, Yi Guo, Terrence Jao, Hung P Do for helping me get started when I first joined the lab and creating such a friendly and warm lab environment. I also thank Sreedevi Gutta, Erum Mustaq, Zhibo Zhu, Namgyun Lee, Ziwei Zhao, Ecem Bozkurt, Kübra Keskin, Bochao Li for their company and fun in the lab.

I should mention the incredible collaborators in the Speech Production and Articulation kNowledge group. I would like to thank Colin Vaz, Tanner Sorensen, Weiyi Chen, Miran Oh, Yoonjeong Lee, and Asterios Toutios – we have worked together for collecting data at the hospital for the last five years, almost every Sunday from 4 to 9 or 10 pm, most of the time even working without having dinner. It could have been a very exhausting experience but because of these people being with me, it was a fun and exciting experience. I would like to say to everyone: we've done a great job!

I would like to thank my girlfriend, Heejin Cho. Her love, care, support, and belief in me not only enriched my life but also let me go through all of my years here at USC.

Last but not least, I would like to thank my family, my parents and my younger brother. My family means everything to me and my life. Thank you for always providing me with unconditional support and love and praying for my safety and careers. I know mom and my brother, they must have gone through the toughest time for the last couple of years there in S. Korea, as I do here in LA. Thank you for being there. My thesis is dedicated to them, and especially to my dad who had always been supportive of me and my decision. He would have been very proud of me. I love you.

YONGWAN LIM

University of Southern California

November 2020

Table of Contents

List Of Tables

List Of Figures

Abstract

Human speech is a unique capability that involves complex and rapid movement of vocal tract articulators. To understand the sounds of speech, it is important to see and understand how the different parts of the vocal articulators move to produce sounds. In this sense, real-time magnetic resonance imaging (RT-MRI) has provided powerful insight into speech production because of its ability to non-invasively and safely capture the essential dynamic features of the vocal tract during the speech. RT-MRI has initiated a dramatic scientific change in the nature of speech production research, including an understanding of language, improved speech synthesis and recognition, and several clinical applications.

Despite the great success of RT-MRI in the study of speech production, there would be still unmet needs in improving the quality and quantity of imaging information about the dynamics of vocal tract articulators. This dissertation introduces new tools for RT-MRI of speech production that offer steps toward a better understanding of speech production.

First, I develop a model-based deblurring method for spiral RT-MRI of speech production. This technique estimates and corrects for dynamic off-resonance that appears as spatially and temporally varying blurring in the image. This method is possible to estimate a dynamic field map directly from the phase of single echo-time dynamic images after a coil phase compensation, and I demonstrate this method can be directly applied to

an existing multi-speaker dataset of running speech. I demonstrate improvements in the depiction and tracking of air-tissue articulator boundaries quantitatively using an image sharpness metric, and using visual inspection, and the practical utility of this method on a use case.

Second, I develop a data-driven deblurring method for spiral RT-MRI of speech production. A 3-layer residual convolutional neural network is present to correct image domain off-resonance artifacts without the knowledge of field maps. The mathematical connection between conventional deblurring methods and the proposed network architecture is derived. I propose a model-based framework that leverages the previous model-based method to generate training data with some augmentation strategy. I validate the proposed method using synthetic and real in vivo data with longer readouts, quantitatively using image quality metrics and qualitatively via visual inspection, and with a comparison to conventional methods.

Finally, I develop a new 3D RT-MRI technique for imaging the full 3D vocal tract at high temporal resolution during a natural speech. This technique utilizes an efficient golden-angle stack-of-spirals sampling, undersampling scheme, and constrained reconstruction. I evaluate this technique through in vivo imaging of natural speech production from two subjects and via comparison with interleaved multislice 2D RT-MRI. This promising tool for speech science for the first time enables a quantitative identification of spatial and temporal coordination of important tongue gestures coproduced on and off the midline in the articulation of English consonants /l/ and /s/ via volume-of-interest analysis and allows a direct assessment of vocal tract area function dynamics during natural speaking of utterances.

Chapter 1

Introduction

Magnetic resonance imaging (MRI) is a vital medical imaging modality that has experienced tremendous growth over the past few decades. It is non-invasive, involves no ionizing radiation, provides different types of tissue contrast, and allows for arbitrary imaging planes. The clinical role of MRI is already established in brain, spine, cardiac, abdominal, and joint imaging and now is expanding to include many other applications such as upper airway and interventional imaging.

This dissertation is about real-time MRI (RT-MRI), one emerging research area for which the primary purpose is to image the dynamic process in the human body as they occur. MRI was previously considered as a "slow" imaging modality. Over the last few decades, however, tremendous technical efforts have been made to push the limit of spatiotemporal resolution forward. Some of the important advances have been found in MRI hardware such as strong and fast-switching gradients that enable time-efficient k-space trajectories and phased-array receive coils that become the crucial part of parallel imaging; in sequence developments such as steady-state free precession-based sequences

that allow for efficiently achieving different types of tissue contrast; in theoretical breakthroughs such as advanced imaging models and reconstruction algorithms that allow for recovering image from a limited amount of data, and more recently machine and deep learning technique. Such enormous advances have made real-time imaging feasible and practical in many different invaluable imaging applications. Among the different areas of the body that experience rapid and often irregular motion, its unique value of RT-MRI becomes prominent in imaging the shape and dynamics of human vocal articulators during speech production. This dissertation focuses on developing and improving techniques for RT-MRI that are successfully applied to (but not limited to) speech production application.

1.1 Outline of Contributions

The main contributions in this dissertation are as follows:

Chapter 2: Magnetic Resonance Imaging This chapter contains a basic overview of MR imaging concepts, advanced sampling and reconstruction techniques, and the imaging consideration for speech production applications.

Chapter 3: Model-Based Deblurring for 2D Real-Time MRI of Speech Production This chapter describes a new method that estimates and corrects for dynamic off-resonance to improve the depiction and tracking of vocal tract articulators in spiral RT-MRI of speech production. We show that it is possible to estimate a dynamic field map directly from the phase of single echo-time dynamic images after a coil phase

compensation. We evaluate the present method using simulations and on an existing multi-speaker dataset of running speech. We demonstrate improvements in the depiction of air-tissue boundaries quantitatively using an image sharpness metric, and using visual inspection, and the practical utility of this method on a use case. Prior publication of this work includes [1].

Chapter 4: Data-Driven Deblurring for 2D Real-Time MRI of Speech Production This chapter describes a fast and effective method for deblurring spiral RT-MRI using convolutional neural networks. A 3-layer residual convolutional neural network is present to correct image domain off-resonance artifacts in speech production spiral RT-MRI without the knowledge of field maps. The mathematical connection between conventional deblurring methods and the present network architecture is derived. Training data generation and augmentation strategy are present. We investigate the effect of off-resonance range, shift-invariance of blur, and readout durations on deblurring performance. We validate the present method using synthetic and real data with longer readouts, quantitatively using image quality metrics and qualitatively via visual inspection, and with a comparison to conventional methods. Prior publication of this work includes [2].

Chapter 5: 3D Real-Time MRI of Speech Production Detailed and direct 3D information about airway shape and spatiotemporal dynamics are essential to understanding speech production control and to relating articulation to speech acoustics. This chapter describes a new technique for imaging the full 3D vocal tract at high temporal resolution during natural speech. The present technique is implemented based on

an efficient golden-angle stack-of-spirals sampling and a constrained reconstruction. We evaluate this technique through in vivo imaging of natural speech production from two subjects and via comparison with interleaved multislice 2D RT-MRI. We demonstrate that this promising tool for speech science for the first time enables a quantitative identification of spatial and temporal coordination of important tongue gestures coproduced on and off the midline in the articulation of consonants /l/ and /s/ via volume-of-interest analysis and allows a direct assessment of vocal tract area function dynamics during natural speaking of utterances. Prior publication of this work includes [3].

Chapter 6: Conclusion and Future Work This chapter summarizes the contributions presented in this thesis and outlines areas for future work.

Chapter 2

Magnetic Resonance Imaging

This chapter divides into three parts. The first part provides an introduction to the basics of MR principles which will be useful to understand before moving to the following chapters. The second part reviews some recent technical advances in data sampling and reconstruction. The third part discusses speech imaging application which is the topic of this dissertation.

2.1 Basic MR Principles

2.1.1 Nuclear Magnetic Resonance Physics

Atoms with an odd number of protons and/or an odd number of neutrons possess an intrinsic spin angular momentum (or *spin*). The basic motion of a proton *spin* can be visualized as a spinning gyroscope that is also electrically charged. Such a charged spin produces its own field and is capable of interacting with an external magnetic field.

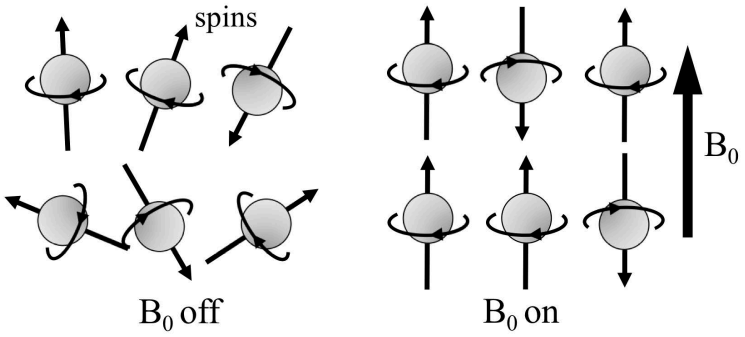

Figure 2.1: Spin orientation in the absence and presence of an external magnetic field. In the absence of any external field, the spins are oriented randomly. In the presence of a $\mathbf{B}_0$ field, spins aline either parallel or anti-parallel, so the net magnization $\mathbf{M}$ becomes non zero.

Sodium (^{23}Na) and phosphorous (^{31}P) are among the elements with such magnetic moment, but hydrogen (^{1}H) is by far the most abundant in the human body, and the only species considered in this dissertation.

Magnetization In the absence of an external field, the spins are oriented randomly as shown in Figure 2.1. Macroscopically, these randomly oriented spins tend to cancel out each other. But when placed in an external magnetic field $\mathbf{B}_0$, the spins tend to align either parallel or anti-parallel to the applied field, with a slightly greater number in the parallel direction, therefore resulting in a non-zero net magnetization $\mathbf{M}$.

Precession At thermal equilibrium, $\mathbf{B}_0$ and $\mathbf{M}$ will align in the same direction (usually denoted as the z-direction or longitudinal direction). When the spins are *perturbed* from

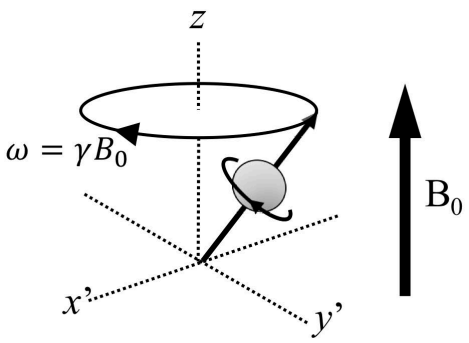

Figure 2.2: Precession of magnetization about the external magnetic field.

their longitudinal orientation, they start precessing about the direction of the $\mathbf{B}_0$ field at a frequency, called the *Larmor* frequency ω_0 (Figure 2.2):

$$\omega_0 = \gamma|\mathbf{B}_0| \tag{2.1}$$

where γ is called the gyromagnetic ratio, an unique constant for each species. For ^{1}H, $\gamma/2\pi = 42.58$ MHz/T.

Excitation The perturbation of the magnetization from the z-axis can be accomplished by applying an oscillating magnetic field $\mathbf{B}_1(t)$. This second field $\mathbf{B}_1(t)$ leads to flip the longitudinal magnetization to produce a transverse component, which becomes detectable by a receive coil. $\mathbf{B}_1(t)$ field is oriented in the x – y plane (usually denoted as transverse plane), perpendicular to the main $\mathbf{B}_0$ field, and tuned to rotate at the resonant frequency ω, creating another resonance condition as illustrated in Figure 2.3. The angle of precession will continue to increase as long as the $\mathbf{B}_1(t)$ is applied at the resonant frequency.

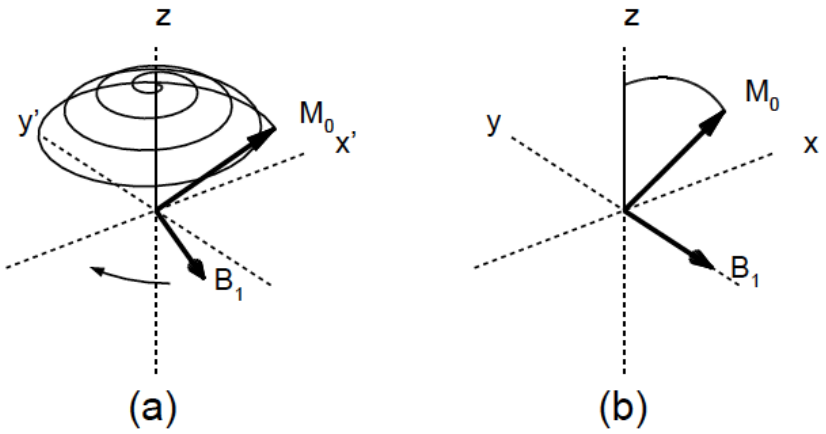

Figure 2.3: Excitation of magnetization by $\mathbf{B}_1(t)$. $\mathbf{B}_1(t)$ induces rotation of magnetization toward transverse plane (figure provided by Brian Hargreaves [4]).

The angle is known as *flip angle*. Since the resonant frequency is in the radio-frequency (RF) range, $\mathbf{B}_1(t)$ field is referred to as *RF field*.

Relaxation Once perturbed by the applied $\mathbf{B}_1(t)$ from the equilibrium orientation, the net magnetization $\mathbf{M}$ returns to its equilibrium position based on two time constants (T_1 and T_2). The longitudinal relaxation rate T_1 specifies the relaxation rate along the z-direction, while the transverse relaxation rate T_2 specifies the relaxation rate in the x – y plane. Fundamentally $T_2 < T_1$ and both relaxation rates vary in different tissues and with field strength.

Bloch Equation The dynamic behavior of the magnetization described above is governed phenomenologically by the Bloch equation:

$$\frac{d\mathbf{M}(t)}{dt} = \mathbf{M}(t) \times \gamma\mathbf{B}(t) - \frac{M_x(t)\mathbf{i} + M_y(t)\mathbf{j}}{T_2} - \frac{(M_z(t) - M_o)\mathbf{k}}{T_1} \tag{2.2}$$

where **i**, **j**, and **k** are the unit vectors in the x, y, and z directions respectively. M_o is the longitudinal magnetization at the equilibrium. $\mathbf{B}(t)$ includes the three magnetic fields applied such as the static field $\mathbf{B}_0$ and RF field $\mathbf{B}_1(t)$ as well as gradient field $\mathbf{G}(t)$, which will be introduced in the subsquent section. From the Bloch equation, the general solution for the transeverse component can be derived as

$$M_{xy}(\mathbf{r},t) = M_{xy}(\mathbf{r},0)e^{-t/T_2(\mathbf{r})}exp\left(-i\gamma\int_0^t B_z(\mathbf{r},\tau)d\tau\right) \tag{2.3}$$

where $M_{xy}(\mathbf{r},t) = M_x(\mathbf{r},t)+iM_y(\mathbf{r},t)$ and $M_{xy}(\mathbf{r},0)$ is the initial transverse magnetization right after excitation. $B_z(\mathbf{r},\tau)$ is the magnetic field only pointed in the z-direction.

Signal Reception The precessing transverse magnetization $M_{xy}(\mathbf{r},t)$ will give rise to an electromotive force in a receiver coil placed near the sample. From Faraday's law of induction, the electromotive force induced in an RF receive coil is proportional to the changes in the magnetic flux caused by the precessing transverse magnetization. The resultant received signal $s_r(t)$ may be written as

$$s_r(t) = \int_{vol} M_{xy}(\mathbf{r},t)dV \tag{2.4}$$

2.1.2 Image Formation

If $\mathbf{B}_1(t)$ is applied only in the presence of $\mathbf{B}_0$, all spins in the volume are excited. Therefore, the received signal shown in Equation 2.4 is contributed by the entire volume. But, in practice, it is desirable to excite only a restricted region of the volume of interest. In addition, the received signal has no spatial information and cannot be distinguished by

spatial location. In the following subsections, we introduce linear gradient field $\mathbf{G}(t)$, briefly discuss the role of it in selective excitation and spatial encoding, and cover subsequent 2D and 3D imaging.

Linear Gradient Field Linear gradient field $\mathbf{G}(t)$ provides spatial selectivity, by generating the longitudinal component of magnetic fields linearly varying proportional to spatial location. Then spatial position can be mapped, proportionally, to the resonance frequency by the Larmor Equation (Equation 2.1) such that $\omega = \gamma(B_0 + \mathbf{G}(t) \cdot \mathbf{r})$.

Selective Excitation *Selective excitation* is achieved by applying the gradient magnetic fields $\mathbf{G}(t)$. Figure 2.4 illustrates the selective excitation. Linear gradient field $\mathbf{G}(t)$ is applied in the direction of desired spatial selectivity, for example, the z-direction. With G_z on, the resonance frequency of spins varies linearly with z such that $\omega = \gamma(B_0 + G_z z)$. $B_1(t)$ is then tuned to the Larmor frequency ω_0, but must have a temporal frequency bandwidth that matches the bandwidth of resonance frequencies of spins in the slice of interest. This can be made possible by varying the amplitude of $B_1(t)$ in time. As a result, only spins in a finite region, say, perpendicular to the z-axis of thickness Δz corresponding to the resonance bandwidth, will be selectively excited.

Spatial Encoding In the presence of both $\mathbf{B}_0$ and $\mathbf{G}(t)$, $B_z(\mathbf{r}, t)$ shown in Equation 2.3 becomes $B_z(\mathbf{r}, t) = B_0 + \mathbf{G}(t) \cdot \mathbf{r}$. Then substituting Equation 2.3 into Equation 2.4 yields

$$s_r(t) = \int_{\Delta z} M_{xy}(\mathbf{r}, 0) e^{-t/T_2(\mathbf{r})} e^{-i\omega_0 t} e^{-i\gamma \int_0^t \mathbf{G}(t)\cdot\mathbf{r}} dr \tag{2.5}$$

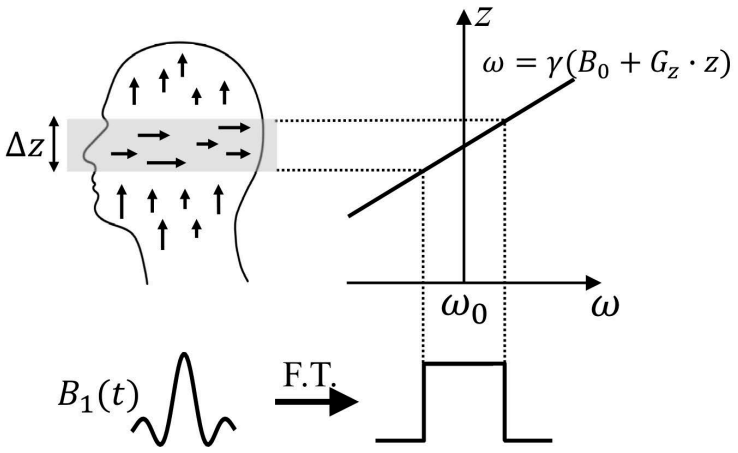

Figure 2.4: Selective excitation. With the linear gradient G_z on, resonance frequency of spins varies linearly with spatial location z. Amplitude modification of $B_1(t)$ field creates limited bandwidth so that only those spins in within the matching frequency band are excited.

After demodulating it by the baseline frequency ω_0 and ignoring T_2 decay, we get the following baseband signal $s(\mathbf{k})$

$$s(\mathbf{k}) = \int_{\Delta z} m(\mathbf{r}) e^{-i2\pi \mathbf{k}(t)\cdot \mathbf{r}} d\mathbf{r} \tag{2.6}$$

where $m(\mathbf{r})$ is a scaled version of the orginal $M_{xy}(\mathbf{r}, 0)$ and $\mathbf{k}$ is defined as

$$\mathbf{k}(t) = \frac{\gamma}{2\pi} \int_0^t \mathbf{G}(t) d\tau \tag{2.7}$$

Equation 2.6 is referred to as the *signal equation* and matches the Fourier transform (FT) relationship between the acquired signal $s(\mathbf{k})$ and the image $m(\mathbf{r})$; $s(\mathbf{k})$ is interpreted as FT of $m(\mathbf{r})$ at a spatial frequency $\mathbf{k}(t)$, which is determined by the time integral of

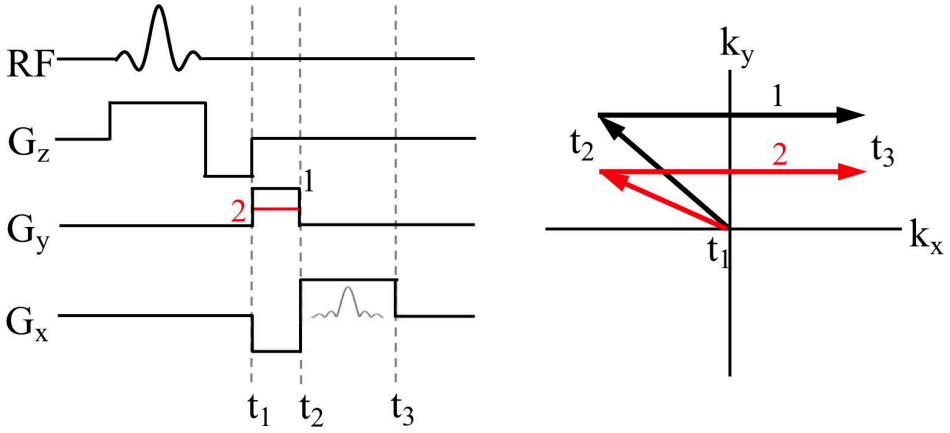

Figure 2.5: 2DFT acquisition. With G_z, a selective excitation RF field excites the slice and the G_y gradient turns on at a time t_1. Once G_y turns off, the signal is read out in the presence of a constant G_x gradient (between t_2 and t_3).

gradient. This Fourier domain of $s(\mathbf{k})$ is known as k-space, the name of which comes from the coordinate $\mathbf{k}(t)$.

2D Imaging Figure 2.5 shows an example of the most commonly used 2D FT acquisition. With the G_z, a selective excitation RF field excites a plane of thickness Δz, perpendicular to the z-direction. Therefore $m(x,y) := \int_{\Delta z} m(\mathbf{r})d\mathbf{r}$ and Equation 2.6 reduces

$$s(k_x,k_y) = \int_x \int_y m(x,y)e^{-i2\pi[k_x(t)x+k_y(t)y]}dydx \tag{2.8}$$

where $k_x(t) = \frac{\gamma}{2\pi}\int_0^t G_x(t)d\tau$ and $k_y(t) = \frac{\gamma}{2\pi}\int_0^t G_y(t)d\tau$. A change in the amplitude of the G_y gradient (also denoted as phase encode gradient) leads to a different line in k-space whereas the signal from each k-space line is read out in the presence of a constant G_x gradient (also denoted as readout gradient).

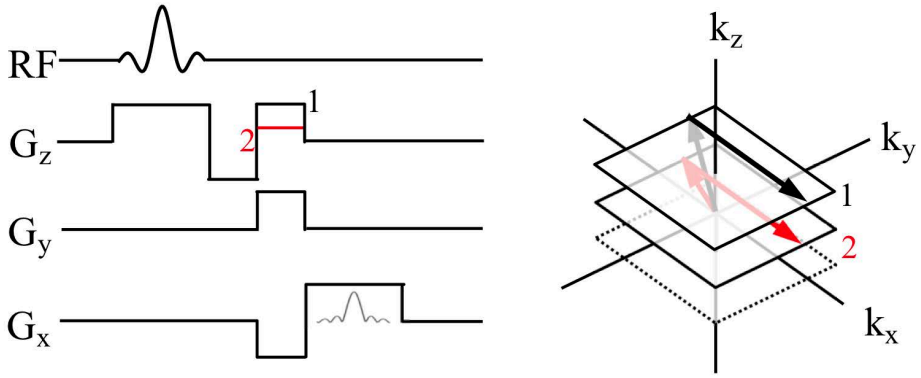

Figure 2.6: 3DFT acquisition.

3D Imaging Figure 2.6 shows an example of 3D FT acquisition. For FT imaging, the extension from 2D to 3D is straightforward by adding one more phase encode gradient along z-axis. With the additional G_z involved, Equation 2.6 can be expressed as

$$s(k_x, k_y, k_z) = \int_x \int_y \int_z m(x, y, z) e^{-i2\pi[k_x(t)x + k_y(t)y + k_z(t)z]} dz dy dx \tag{2.9}$$

Here, similarly, $k_z(t)$ is defined as $k_z(t) = \frac{\gamma}{2\pi} \int_0^t G_z(t) d\tau$.

2.1.3 k-Space Trajectory

2D or 3DFT acquisition is widely used due to the simplicity of reconstruction and robustness to artifacts. But the drawback of such Cartesian sampling is that it requires a large number of encoding lines, resulting in increased acquisition time. There are more time-efficient alternatives to this, having both strengths and weaknesses.

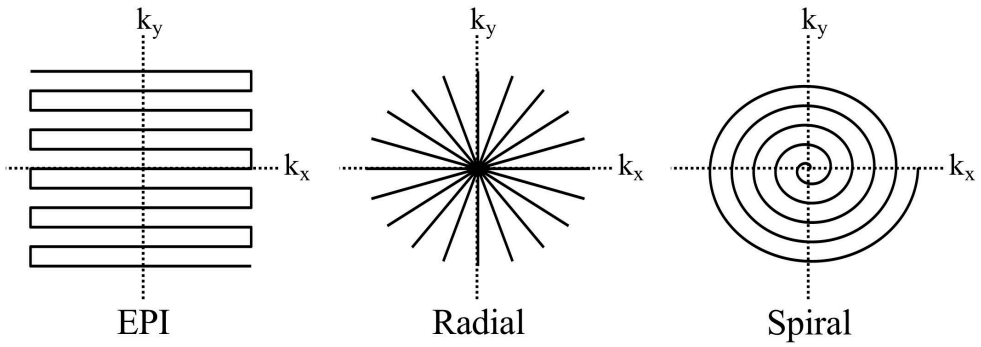

Figure 2.7: Efficient 2D k-space trajectories. EPI, Echo planar imaging; Radial; Spiral.

2D Trajectory Figure 2.7 illustrates efficient 2D sampling trajectories. Echo planar imaging (EPI) is an alternative where multiple Cartesian lines are acquired in a raster-like fashion after each excitation, therefore k-space can be filled only with one or fewer repetitions. One primary limitation is that its prolonged readout time makes it vulnerable to off-resonance, eddy currents, gradient delay, and so on, which give rise to ghosting and distortion artifacts in reconstructed images.

Radial sampling covers k-space by acquiring radial lines at different azimuthal angles. The order of the angle is a design choice but the golden-angle increment is widely adopted. Each readout line at an azimuthal angle passes through the k-space origin, resulting in oversampling the central region of k-space. This offers more robustness to motion than 2DFT and tolerance to spatial aliasing artifact due to undersampling. With moderate angular undersampling along with the golden angle scheme, the spatial aliasing artifacts are usually incoherent in appearance, which can be reduced by advanced reconstruction techniques. While radial sampling is $\pi/2$ less efficient than Cartesian, those favorable properties make it a popular method for rapid imaging.

Spiral sampling usually starts acquiring data from the origin of k-space and ends at the periphery after traversing along a spiral trajectory. Spiral sampling can completely cover k-space with one or a few repetitions to achieve Nyquist sampling of k-space. It provides excellent velocity PSF and reduces motion artifacts due to its natural oversampling at the k-space center. Spiral sampling is well-suited for advanced reconstruction algorithms such as compressed sensing when combined with strategies such as under-sampling and golden angle scheme similar to radial sampling. Because of its favorable properties, it is also widely used for RT-MRI where the capability of capturing rapid motion is crucial such as in cardiac imaging and speech production imaging. However, spiral trajectory is sensitive to off-resonance that appears as signal loss and/or blurring in reconstructed images. These artifacts are most pronounced at high field strength and with long readout duration which is precisely when spiral provides the greatest efficiency. Multi-shot spiral acquisitions with a short readout time ($\leq$ 2.5 ms at 1.5 T) is often used for the practical reason.

3D Trajectory Analogous to the extension from 2D to 3D FT imaging, direct extensions of the aforementioned efficient 2D trajectories are possible by augmenting an additional k_z direction. Figure 2.8 illustrates representative examples of 3D stack-of-stars and stack-of-spiral trajectories. More advanced pure 3D trajectories can be found in the literature, including 3D radial, 3D cone, and so on.

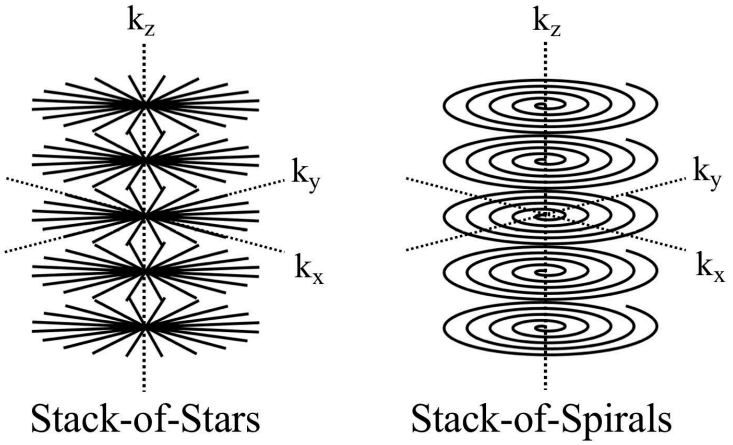

Figure 2.8: 3D k-space trajectories. Stack-of-stars and stack-of-spirals.

2.1.4 Off-Resonance

Thus far we have assumed a perfectly uniform magnetic field B_0. However, in practice, the resonant frequency will not be uniform across samples and shifts in the frequency are often caused by B_0 field inhomogeneities, chemical shift, and susceptibility differences between tissues boundaries. In general, off-resonance can be referenced by deviation frequency $\Delta\omega$ (or Δf in Hz), with respect to the resonant frequency ω_0, which is a function of spatial location. Among several off-resonance sources, susceptibility differences between the soft tissue (water) and air is of most relevant in this dissertation. A typical range of Δf due to the susceptibility difference is proportional to the strength of the external magnetic field B_0; up to $\sim$ 600 Hz at 1.5 T and $\sim$ 1200 Hz at 3 T [5]. Off-resonance reflects the signal equation, say, Equation 2.8 in the case of 2D imaging, as

$$s(k_x, k_y) = \int_x \int_y \left[m(x,y) e^{-i2\pi\Delta f(x,y)t} \right] e^{-i2\pi[k_x(t)x + k_y(t)y]} dy dx \tag{2.10}$$

This equation will be used later in Equations 3.2 and Equations 4.1 in Chapters 3 and 4 in slightly modified forms.

The manifestation of off-resonance in image domain is dependant on the acquisition trajectories and can be analyzed by examining the impulse response. Suppose that a point signal source is located at (x_0, y_0) such that $m(x, y) = {}^2\delta(x - x_0, y - y_0)$. Then the resultant signal can be expressed as $s(k_x, k_y) = e^{-i2\pi\Delta f(x_0,y_0)t}e^{-i2\pi[k_x(t)x_0+k_y(t)y_0]}$. By reconstructing these signals with conventional reconstruction approach, we can determine the point spread function (PSF) of the imaging system as

$$\begin{aligned} &PSF(x, y, x_0, y_0; \Delta f(x_0, y_0)) \\ &\qquad = \int_{k_x}\int_{k_y} s(k_x, k_y)e^{i2\pi[k_x(t)x+k_y(t)y]}dk_y dk_x \\ &\qquad = \int_{k_x}\int_{k_y} \left[e^{-i2\pi\Delta f(x_0,y_0)t}e^{i2\pi[k_x(t)(x-x_0)+k_y(t)(y-y_0)]}\right] dk_y dk_x \end{aligned} \tag{2.11}$$

In 2DFT imaging, $PSF(x, y, x_0, y_0; \Delta f(x_0, y_0)) = {}^2\delta(x - x_0 - \frac{\Delta f(x_0,y_0)}{\gamma G_x/2\pi}, y - y_0)$, resulting in image shift by $\frac{\Delta f(x_0,y_0)}{\gamma G_x/2\pi}$ in the readout direction. In other types of imaging, the artifact becomes complicated. In EPI, the artifact appears as image shift in the phase encoding direction, resulting in a geometic distortion. In radial or spiral imaging, the artifact appears as signal loss and/or blurring. Figure 2.9 illustrates phase error in k-space due to off-resonance and its corresponding PSF in image space for spiral imaging. When $\Delta f \approx 0$ or using very short readouts, the phase accrual due to off-resonance during readout such as $e^{-i2\pi\Delta ft}$ shown in Equation 2.11 can be ignored and the PSF is approximately a sharp

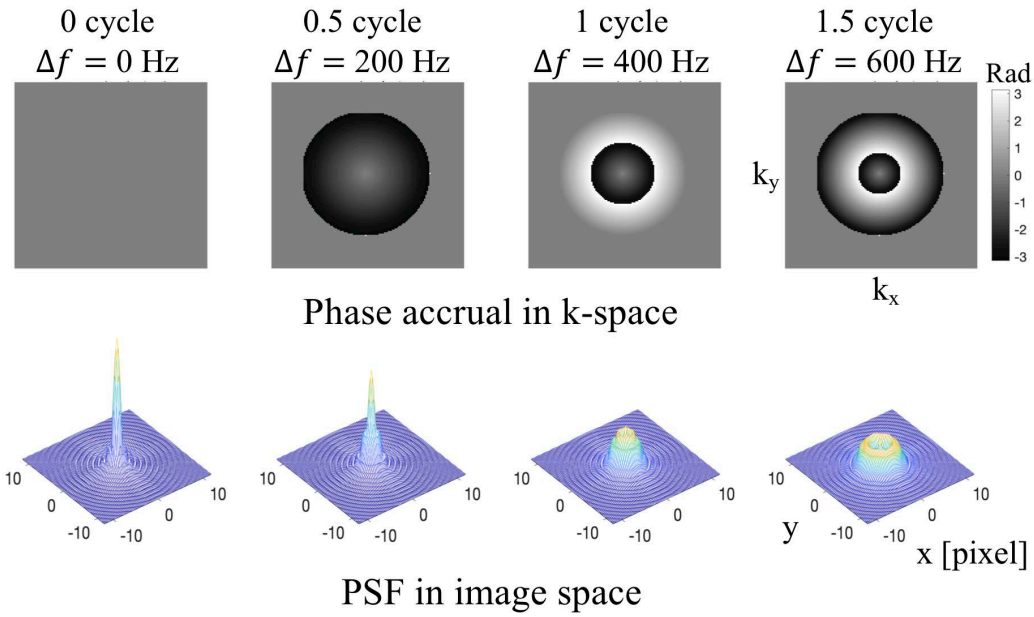

Figure 2.9: PSF in the presence of off-resonance in spiral imaging. Top: phase accrual due to off-resonance ($e^{-i2\pi\Delta ft}$) in k-space. Bottom: corresponding PSFs as a function of off-resonance frequency Δf for spiral imaging. Assuming readout time of 2.5 ms, $\Delta f =$ 400Hz yields one cycle of the phase accrual during the readout.

impulse at $(0, 0)$. However, in the presence of off-resonance, the larger Δf is, the more phase error of $e^{-i2\pi\Delta ft}$ is accured, therefore increasing the spread of PSF.

2.2 Advanced Sampling and Reconstruction

2.2.1 Spatio-Temporal (k-t Space) Sampling

So far, we have assumed a stationary object imaged and the corresponding data acquired onto k-space. For a dynamic object, in contrast to a stationary object, the signal source (i.e., the complex magnetization) at a given spatial location varies with time due to physiological motions, contrast dynamics, or any other type of signal modulation. Thus,

a complete description of a dynamic object requires time as an additional variable. k-t space [6] is an extension of k-space with an additional time axis, t. In the case of 2D dynamic imaging, one can acquire a series of k-space over time and stack all of the 2D k-space into a 3D cube, yielding 3D k-t space (k_x, k_y, and t). Analogously, 3D dynamic imaging yields a corresponding 4D k-t space.

However, because of physical and physiological constraints on the speed of data acquisition, it is usually impractical to acquire the full k-t space data necessary for reconstructing each time frame separately. Therefore, methods have been developed that acquire only a part of the desired k–t space, the rest being recovered or estimated through some model assumption or advanced reconstruction [7, 8, 9, 10, 11, 12]. When and which data points are acquired is referred to as *k-t sampling pattern* or sampling schedule [13]. The k-t sampling pattern affects the fidelity of the reconstructed image. The more one can sample data within k-space, the less one can reconstruct images with aliasing artifact. The more frequently one can sample data along the time axis, the more one can capture quick temporal changes. The k-t sampling pattern is also associated with model assumptions or reconstructions. The k-t sampling patterns for 2D Cartesian, radial, and spiral imaging are illustrated in Figure 2.10. The concept of the k-t sampling pattern shown here is also applicable to 3D imaging.

2.2.2 Image Reconstruction

Gridding and NUFFT For FT imaging where k-space samples are acquired on a rectangular (Cartesian) grid, the image can simply be reconstructed by performing the inverse fast-Fourier transform (FFT). Gridding [14, 15, 16] is one way to reconstruct an

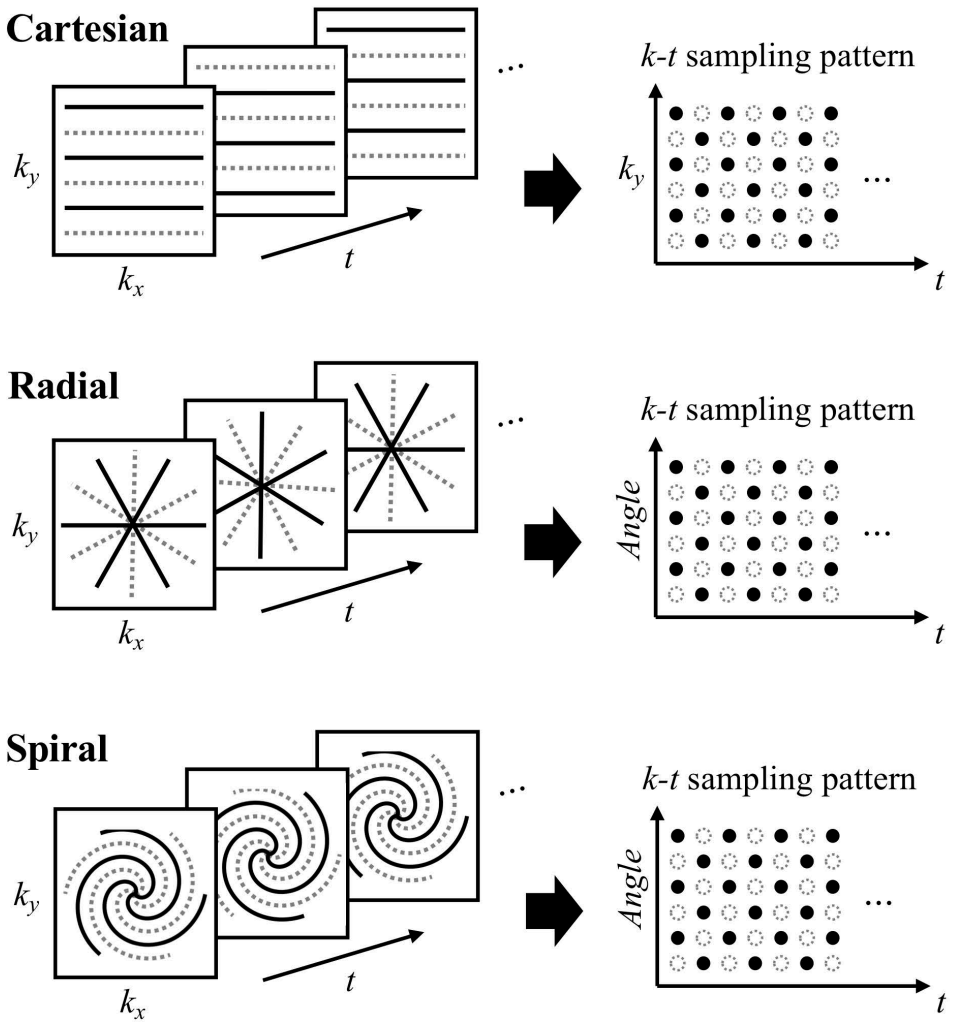

Figure 2.10: k-t sampling pattern for 2D Cartesian, radial, and spiral imaging.

image from k-space samples that do not fall on a regularly spaced Cartesian grid. The general idea known as "gridding" is that the data points lying along some non-Cartesian trajectory through k-space are convolved with a small kernel and resampled onto the Cartesian grid. Then the image is reconstructed by performing the inverse FFT. Non-uniform FFT (NUFFT) [17, 18] is another class of dealing with non-Cartesian samples. Although there are many variations, the NUFFT can be thought of as doing the operation of gridding in the reverse direction; the forward direction of the NUFFT is from the image on the Cartesian grid to k-space samples on a non-Cartesian grid. The forward and its reverse (adjoint) operations of NUFFT are important for iterative reconstruction such as non-Cartesian SENSE [19]. Several design choices exist for both approaches including interpolation kernels, oversampling factors, scaling factors, and density compensation.

View Sharing In many cases, successive image frames in a dynamic series are often more similar than images that are farther apart in time. The idea of *view sharing* is to share data (so-called views) across time frames to take advantage of the similarity among adjacent time frames. View sharing has found its applicability to various spatio-temporal sampling schemes [13]. View sharing [20, 21] is often used interchangeably with *sliding window* reconstruction. Figure 2.11 shows an example of view sharing reconstruction for interleaved spiral acquisitions. Suppose that thirteen spiral interleaves are required to form a single image frame. These 13 interleaves are being continuously and repeatedly acquired while each spiral interleaf or view is being shared across multiple time frames. In doing so, a different image can be generated after each TR, rather than after 13 TRs (for example, without view sharing). Sharing data increases the apparent temporal resolution

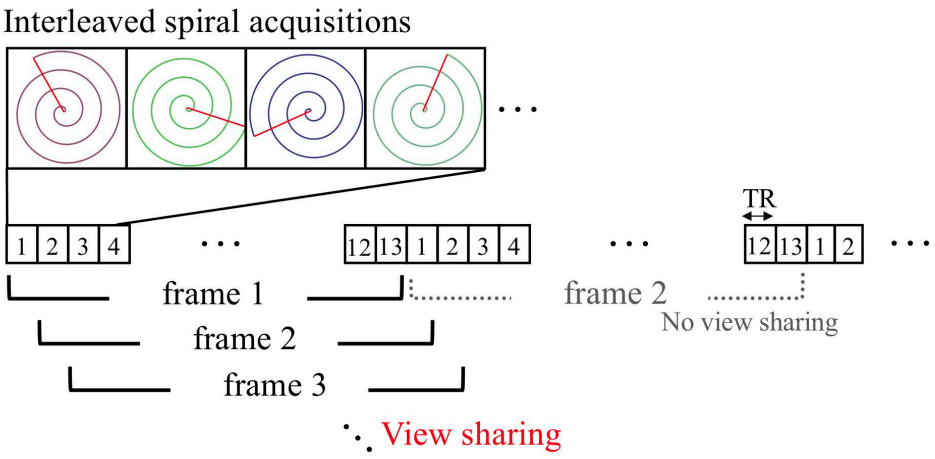

Figure 2.11: View sharing sliding window reconstruction illustrated for interleaved spiral acquisitions. Without view sharing, the complete set of six interleaves is being used to form an image frame without any overlapping. With view sharing, every image frame can be generated by sharing data (views within a sliding window) after each or multiple TRs

since there are 13 times as many time frames but the images may not contain the truly new information [22].

Compressed Sensing Compressed sensing (CS) technique [23] has been established as a promising theoretical framework for accelerating MRI acquisitions. According to CS theory, a signal can perfectly be reconstructed from highly and randomly undersampled data (i.e., sub-Nyquist sampling rate) provided that the signal is sparse in a certain transform domain (sparsifying transform domain). The CS is based on the fact that aliasing artifacts due to randomly undersampled measurements occur as incoherent noise in the transform domain and can effectively be eliminated by using non-linear noise removal algorithms via a minimum l_0 or l_1 norm in the context of constrained reconstruction [24] so that the desired signal can perfectly be recovered.

In dynamic MRI where the images have highly redundant information in both space and time, rapid imaging techniques based on CS have been proposed. The underlying theory can be explained with an example illustrated in Figure 2.12: Obviously, dynamic scenes can perfectly be reconstructed from fully sampled data. If the dynamic scenes have periodic behavior of the motion, the temporal variation of the intensity is also periodic as can be seen in the corresponding spatial–temporal (y–t) space. These scenes can be represented with a few harmonic frequency coefficients in spatial–temporal frequency (y–f) space where f is the FT pair of time t in y–t space. Thus, randomly undersampled patterns in k_y–t space yield incoherent aliasing artifacts in the corresponding y–f space. By using the non-linear reconstruction methods, the incoherent artifacts can be eliminated and the dynamic scenes can be perfectly reconstructed from the randomly undersampled data as same as from the fully sampled data.

Spatial and Temporal Constraints In addition to FT that exploits periodicity in the dynamic time series (e.g., gated cardiac cine), some of the most popular sparsifying transforms in dynamic MRI are finite differences, total variation, and wavelet, which can be applied to exploit spatial and/or temporal sparsity.

In speech imaging applications where the primary features of interest are moving tissue boundaries, temporal finite differences are widely adopted since this encourages pixel intensity to be piecewise constant along time.

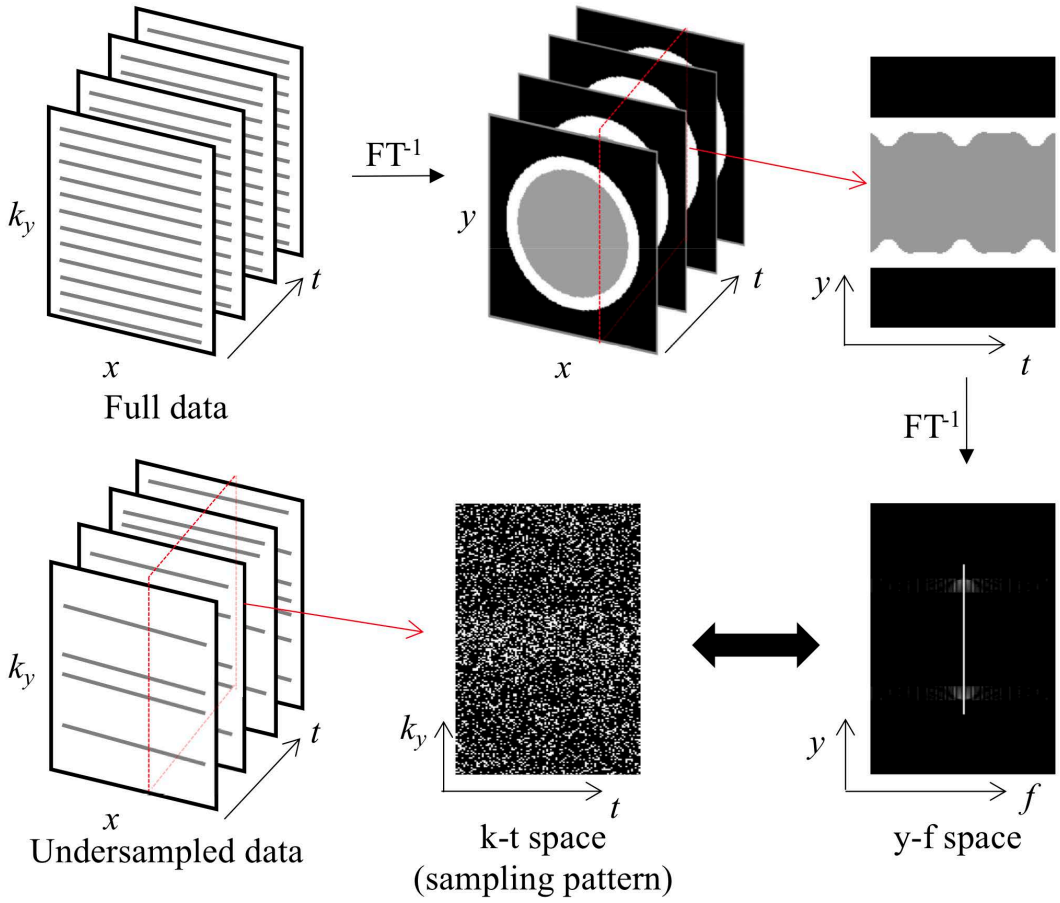

Figure 2.12: Compressed sensing concept in dynamic MRI

2.3 Speech MRI

Speech production is a human's unique capability. It is the result of a well-coordinated interplay of different soft tissue structures such as the lips, tongue, hard palate, soft palate, larynx, and epiglottis as well as several muscles in the upper airway. Speech production mechanism can be summarized as follows [25]: Our brain comes up with a speech plan, which is sent through our bodies as nerve impulses. These nerve signals arrive at muscles, causing them to contract. Muscle movements expand and contract our lungs, allowing them to move air. This air moves through our vocal tract, which we can shape with more muscle movements. By changing the shape of our vocal tract, we can block or release airflow, create vibrations or turbulence, change frequencies and tones, and so on. Through these actions, we produce different speech sounds, which is what we perceive as speech.

To understand the sounds of speech, it is important to see and understand how the different parts of the human body move to produce sounds. Speech scientists, linguists, and clinicians, therefore, have used a wide array of tools to visualize or track the shape and movements of articulators to gain new insights into sound and speech production both in typical speakers and in clinical populations. Figure 2.13 shows the existing speech measurement and/or imaging systems. Table 2.1 lists and evaluates the weaknesses and strengths of each one of those tools.

Electromagnetic articulography (EMA) and electropalatography (EPG) are sensor-based techniques. EMA can track the movement of sensor coils glued to the tongue and other structures in the mouth. EPG can monitor tongue-palate contact from a

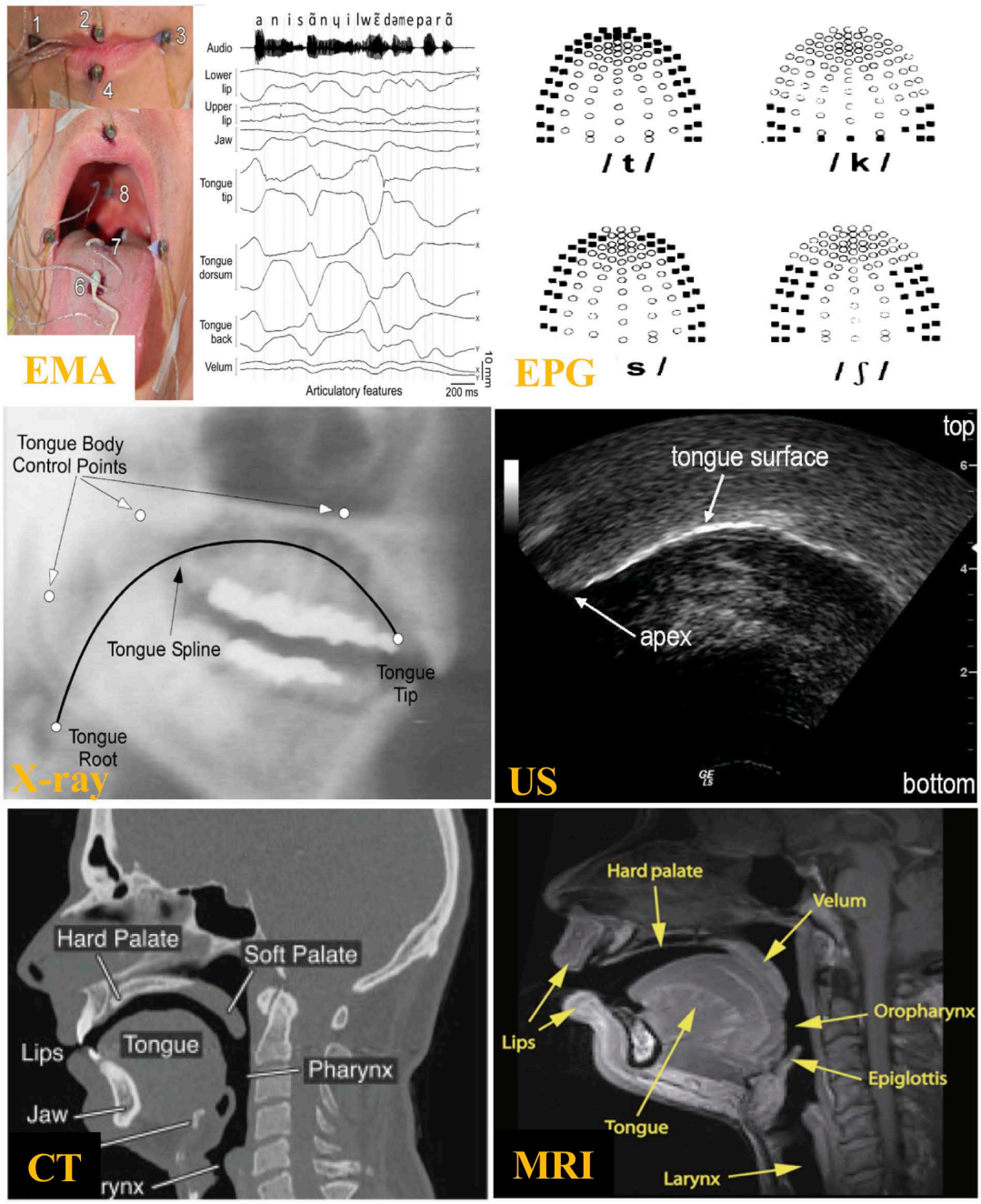

Figure 2.13: Speech measurement and imaging modalities. EMA, Electromagnetic articulography [26]; EPG, Electropalatography [27]; X-ray [28]; US, Ultrasound [29]; CT, Computational tomography [30]; MRI, Magnetic resonance imaging [31];

Table 2.1: Comparison of speech imaging modalities

	EMA	**EPG**	**X-Ray**	**US**	**CT**	**MRI**
Invasive	×	×	✓	–	✓	✓
Ionizing radiation	✓	✓	×	✓	×	✓
Spatial resolution	×	×	✓	✓	✓	✓
Temporal resolution	✓	✓	–	✓	✓	✓
Tissue contrast	×	×	×	×	×	✓
Spatial coverage	×	×	×	×	✓	✓
Cost	✓	×	✓	✓	×	×

strength (✓), weakness (×), and neutral (–).
Abbreviations: EMA, electromagnetic articulography; EPG, electropalatography; US, ultrasound; CT, computer tomography; MRI, magnetic resonance imaging.

customized electro-palate sensor. Both techniques can provide a rapid tracking rate ($\geq$ 100 Hz), but are limited to provide information only from exterior surfaces and in the region covered by the sensors. In addition, the non-invasive set-up for sensors could potentially cause discomfort and abnormal speech. X-ray fluoroscopy is an alternative, non-invasive technique that operates at both high temporal ($\sim$30 Hz or frames per second (fps)) and in-plane spatial resolution ($\leq 0.5 \times 0.5$ mm^2). However, it involves ionizing radiation and is limited to provide projections of the anatomy. Ultrasound is another imaging technique that is relatively low cost and is widely available both in clinical and research settings. One primary limitation is that the tongue is only imaged. X-ray computed tomography (CT) can obtain volumetric images with high spatial resolution and with the delineation of deep soft tissues and bone structure, but the radiation dose may render it not a viable method for frequent use.

Compared with the aforementioned imaging modalities, MRI enables imaging deep soft-tissues in the arbitrary scan plane without radiation or endoscopy and with excellent soft tissue contrast. These favorable properties have made MRI an ideal modality for investigating both the structures and dynamics of the vocal tract. MRI was previously considered as a "slow" imaging modality, but tremendous advances in the field of MRI have made rapid imaging feasible, practical, and readily available.

2.3.1 Speech Real-Time MRI

RT-MRI has gained substantial attention for speech production research because of its unique advantage of monitoring the complete vocal tract during the speech, safely and non-invasively at relatively high spatial and temporal resolution [21]. Figure 2.14 shows an example of 2D RT-MRI of speech production imaged at the mid-sagittal scan plane. RT-MRI continuously captures the dynamic as it is without the need for repetitions or gating as opposed to the gated cine MRI technique. This is particularly valuable as the nature of the movements of the articulators during speech is not necessarily periodic and is unrepeatable. RT-MRI is also compatible with concurrent audio recording, which could aid linguistic analysis. RT-MRI can now be combined with intermittent tagging pulses to visualize internal deformation in the tongue muscles [32, 33].

Applications of RT-MRI in speech production study span several linguistic and speech scientific areas of research including 1) studies of phonetic and phonological phenomena and language acquisition, 2) understanding of the dynamics of vocal tract shaping during speech or non-speech events, 3) modeling of speech production and motor control, 4) speech synthesis and recognition to clinical applications in the treatment planning of

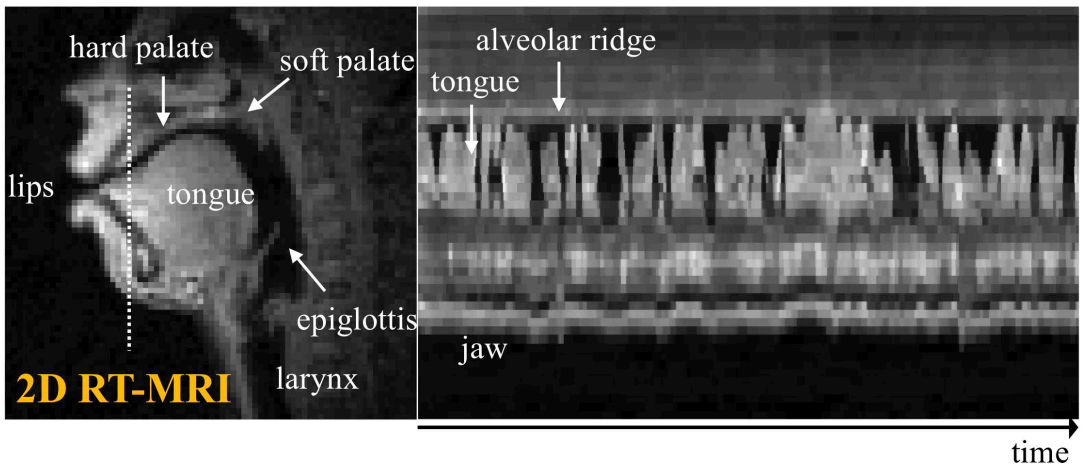

Figure 2.14: Example of 2D RT-MRI of the vocal tract. Left: Mid-sagittal image frame. Right: The intensitivy time profiles corresponding to the cut marked by the dotted line in the image shown on the left. Spiral acquisition; in-plane spatial resolution = 2.4 mm^2; temporal resolution = 12 ms per frame.

speech disorders. Such speech disorder may arise post-surgery in glossectomy or in conditions like velopharyngeal insufficiency and apraxia. The interested readers can refer to review articles for other applications and technical aspects of the speech RT-MRI with a focus on speech [34, 35], speech and sleep [31], and image analysis techniques on RT-MRI of vocal tract motion [36].

2.3.2 Imaging Requirements

Figure 2.15 shows the imaging requirements of spatial and temporal resolution for various speech tasks from the 2014 Speech MRI Summit. Each rectangular zone with an approximate boundary represents a specific speech task in terms of the spatial and temporal resolutions. Although specific imaging parameters would be dictated by the vocal tract regions and speech tasks of interest, the speech MRI generally requires high spatial and temporal resolution. It is recommended in the speech MRI community that an

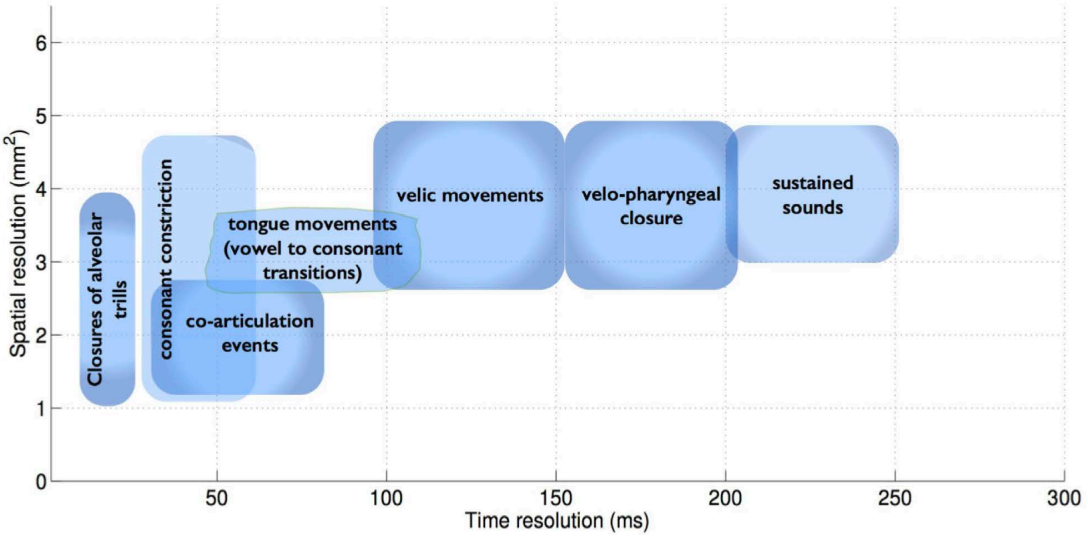

Figure 2.15: Imaging requirements for different speech tasks. A consensus on the spatial and temporal resolution requirements for different speech tasks from 2014 Speech MRI Summit, image courtesy of Lingala [35]. Each rectangular zone with an approximate boundary represents a specific speech task.

in-plane spatial resolution of no more than 3.5 mm^2 and time resolution of below 70 ms are required to study very fast articulatory movements such as those during consonant constrictions and coarticulation events that are major tasks of interest [35].

The current state-of-the-art technique uses non-Cartesian sampling (radial or spiral acquisition) and parallel imaging, combined with constrained reconstruction. This has enabled visualization of 2D dynamic images with spatial resolutions of 1.3 to 2.4 mm^2 at high temporal resolutions of 10 to 60 ms from highly under-sampled MRI data [37, 38, 39, 40, 41, 42, 43]. Imaging 2D mid-sagittal plane is the most preferred (but not enough) as it has provided good utility to speech scientists due to the fact that important information about "place of articulation," which is critical in linguistic contrasts, can be obtained from constriction details in the mid-sagittal plane (such as, for example, in

phonemes /p/, /t/, and /k/ with constrictions are located at the lips, alveolar ridge, and velar region).

2.3.3 Imaging Consideration

Tradeoff In RT-MRI [44], there exists an inherent trade-off between spatial and temporal resolution, spatial (and/or slice) coverage, computation time, as well as noise and artifact reduction [35]. A major challenge in this domain is its requirement of very high temporal resolution since the vocal tract contains several rapidly moving articulators some of which change in less than a few milliseconds. Monitoring such rapid motion has been made possible by restricting the field-of-view to a few 2D planar images and sampling along time-efficient non-Cartesian trajectories at sub-Nyquist rates. However, such a subsampling introduces aliasing artifacts that need to be suppressed during advanced image reconstruction, leading to increased computation time. Furthermore, non-Cartesian trajectory such as spiral sampling is prone to image artifacts such as blurring due to off-resonance. These artifacts impair the delineation of speech articulators near the tissue boundaries, which are of utmost interest for linguistic sciences and clinical diagnosis. Although those non-Cartesian trajectories provide the greatest time efficiency at higher field strength or longer readouts, this is also when the artifacts are most pronounced. This leads the current imaging for speech production to be most often conducted using short duration readouts ($\sim$2.5 ms) and at lower field strength (1.5 T) MRI scanners.

Field Strength Imaging at higher field strength can provide high SNR. Nevertheless, higher field strength ($\geq$ 3T) is often avoided because of its higher susceptibility to off-resonance induced artifacts that significantly degrade image quality around air-tissue boundaries.

Pulse Sequences The mechanism of MR offers the ability to achieve various image contrast. Signal strengths are determined by the MR physical parameters such as spin density, T1, and T2 relaxation rates which differ among tissue types. The imaging sequences can play an important role in producing different contrast by controlling its parameters such as the flip angle, repetition time (TR), and echo time (TE) and by emphasizing differences between the MR physical parameter values.

Among many of the available imaging sequences, two rapid gradient echo sequences are widely used for rapid imaging. Spoiled gradient recalled echo (GRE) sequences are robust to artifacts and provide T1-weighted contrast with a short TR. Balanced steady-state free precession sequences provide higher SNR efficiency than the spoiled sequences. They also provide T2/T1 contrast, which is advantageous in applications such as in cardiac imaging because of the excellent blood–myocardium contrast, although its sensitivity to off-resonance manifests as banding artifacts, limiting its usage at higher field strength.

Although balanced steady-state free precession could provide a higher SNR and better contrast, the banding artifact due to off-resonance is challenging in speech imaging. Besides, speech imaging is mainly interested in looking at a single type of soft-tissue (water) such as the tongue and demands less contrast compared to other applications that necessitate contrast between different types of tissue.

Sampling Trajectory As discussed in "Tradeoff", one needs to consider imaging requirements for speech region and tasks of interest and the intrinsic trade-off for the k-space sampling trajectory of the choice. Non-Cartesian trajectories are generally employed when a higher spatial or temporal resolution is desired.

Receiver Coil Commercially available coils are generally designed for neurovascular or carotid artery imaging and would not be suitable for imaging the upper airway in terms of attainable SNR, comfort level for subjects during speaking, and compatibility with audio recording setup. For those reasons, custom upper airway coil is often used and has also been shown to provide 2-fold to 6-fold higher SNR efficiency than commercially available coils, in upper airway vocal tract regions of interests including the tongue, lips, velum, epiglottis, and glottis [40].

Simultaneous Audio Acquisition Simultaneous acquisition of audio along with RT-MRI data acquisition is essential for subsequent linguistic analysis and investigation of articulation relating to speech acoustics. To avoid MRI acquisition from being interfered with any audio recording setup, a commercially available fiber-optic microphone is widely used in multiple speech production research groups because of its MR-compatibility.

In simultaneous audio acquisition, synchronization is one practical consideration. One possible approach is the hardware-synchronization of an audio sample clock to the MRI scanner's master clock so that the timing between the audio recording and RT-MRI acquisition is internally aligned. Another important consideration is the acoustic noise induced by the rapidly-switching MRI gradient coils during imaging. The microphone is typically placed near the subject's mouth inside an MRI magnet bore at which the

noise level is strongest and often over 100 dB [45, 46]. Along with the audio recording setup, integrated software based on adaptive noise cancellation can offer real-time noise cancellation and/or custom algorithms can be utilized to further improve SNR [47, 48].

2.3.4 USC RT-MRI Protocol

In this dissertation, all experiments described were performed on a commercial 1.5 T scanner (Signa Excite, GE Healthcare, Waukesha, WI) with gradients capable of 40 mT/m magnitude and 150 mT/m/ms slew rate. Figure 2.16 shows our imaging setup at USC. A body coil was used for RF transmission, and a custom eight-channel upper airway coil [40] was used for signal reception. The custom coil is positioned close to the upper airway structures, and has an opening near the mouth for microphone positioning. All developments are based on GRE sequences that provide T1-weighted contrast and on multi-shot spiral-based sampling trajectories. The imaging protocol was approved by our Institutional Review Board.

Any visual or text stimuli were viewed or read in the scanner using a mirror projector setup for presentation [40]. Acoustic audio data were recorded inside the scanner using commercial fiber-optic microphones (Optoacoustics Inc., Yehuda, Israel) simultaneous with the RT-MRI data acquisition using a custom recording setup [47]. The recorded audio was enhanced using a normalized least-mean-square noise cancellation method [47] and was aligned with the reconstructed MRI video sequence to aid linguistic analysis.

The data acquisition schemes presented in Chapters 3, 4, and 5 were implemented using a real-time interactive imaging platform (RT-Hawk, Heart Vista Inc, Los Altos, CA) [50]. Real-time visualization was implemented within this custom platform by using

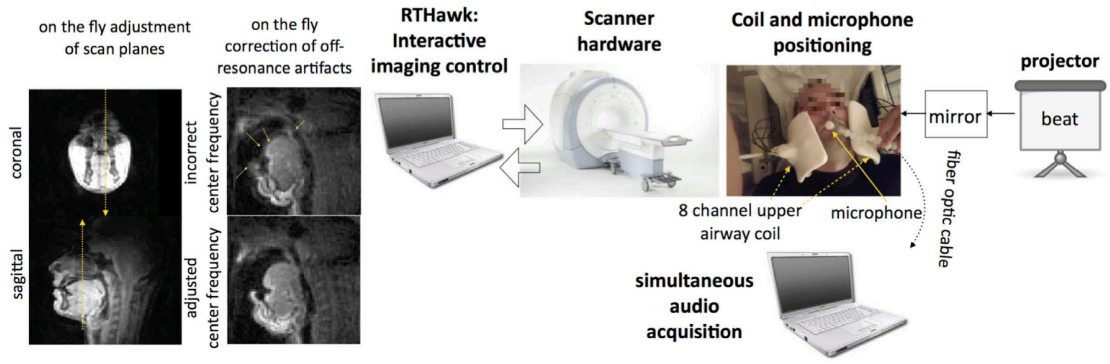

Figure 2.16: USC RT-MRI imaging setup. Image courtesy by Lingala et al [49]

a view sharing sliding window gridding reconstruction [21] to ensure the subject's compliance with stimuli and to detect substantial head movement. In addition to the automatic shimming provided by the prescan calibration from the scanner, a manual adjustment of the center frequency described in Ref. [40] was also performed. Specifically, while the subject being scanned in a neutral open-mouth position, the center frequency was adjusted on-the-fly in a way that the air-tongue boundary is sharp in the mid-sagittal plane.

2.3.5 Unmet Needs

Despite the great success of RT-MRI techniques in providing invaluable imaging tools in the study of speech production research, there would be still unmet needs in improving *quality* and *quantity* of imaging information about the dynamics of articulators during speech production.

2.3.5.1 Artifact Mitigation

While non-Cartesian sampling (for example, spiral acquisition) typically is a preferred choice in the current state-of-the-art techniques to achieve higher spatio-temporal resolutions, it is highly sensitive to errors such as due to off-resonance [5]. Off-resonance gives rise to blurring and/or signal loss in the image domain and degrades image quality substantially at the air-tissue boundaries that surround the vocal tract articulators, which are of utmost interest for linguistic sciences and clinical diagnosis. However, to date, this remains the predominant challenge for speech RT-MRI application as it can cause an error in any linguistic analysis made on RT-MRI images and introduce bias as well as increased variance during data analysis. In addition to that, we often need to compromise scan efficiency to mitigate this artifact. The first two parts of this dissertation (Chapters 3 and 4) deal with deblurring approaches to improve the delineation of articulator boundary in spiral-acquisition-based 2D RT-MRI of speech production.

2.3.5.2 Slice Coverage

Generally, most RT-MRI techniques have been limited to one mid-sagittal imaging plane or a few 2D imaging planes. This is mainly due to the imaging tradeoffs between slice coverage, temporal resolution, and available information from the imaging data. Aiming at a high temporal resolution, the mid-sagittal plane is most informative due to its entire vocal tract coverage from the lips to the glottis. However, speech production involves the movement of articulators occurring in "3D" in nature and linguistically fundamental and interesting speech events cannot be fully understood by only looking at those 2D imaging planes [51, 52]. Examples include lateral shaping of the tongue, such as grooving

or doming, and asymmetries in the tongue shape as well as resonate cavity volume. Also, detailed and direct 3D information about airway shape and spatiotemporal dynamics are essential to understanding speech production control and to relating articulation to speech acoustics. Although several research groups have demonstrated 3D vocal tract imaging techniques [53, 54], those techniques would require a subject to perform speech tasks under constraints against a natural speech production such as sustaining sounds or repeating speech tasks during imaging. There would be still an unmet need for developing a technique that allows for imaging full 3D vocal tract at a high temporal resolution without any constraints during speech production. The second part of this dissertation (Chapter 5) deals with the feasibility of 3D RT-MRI of speech production.

Chapter 3

Model-Based Deblurring for 2D RT-MRI of Speech Production

Spiral RT-MRI is desirable because it allows for a time efficient acquisition, given that spirals can provide higher spatio-temporal resolution than alternative schemes [35, 55]. A key drawback of spiral MRI is signal loss and/or blurring artifacts that result from field inhomogeneity, also called "off-resonance" [56]. This can be significant at air-tissue interfaces due to their magnetic susceptibility difference ($\Delta\chi = 9.41$ parts per million) [5]. Furthermore, these artifacts near the air-tissue boundaries [57] are more pronounced with long spiral readout or at high field strength MRI scanners. To mitigate this artifact, current RT-MRI studies for speech production are most often conducted using short duration readouts ($\sim$2.5 ms) and at lower field strength (1.5 T) MRI scanners [21, 58, 40].

Off-resonance artifacts have a significant potential impact on the analysis of articulatory dynamics, which is of prime interest in speech science. The articulators of interest include the surfaces of the lips, tongue, hard palate, soft palate (velum), and structures along the pharyngeal airway. These are located at air-tissue interfaces and therefore are vulnerable to the artifacts. Previously used speech RT-MRI biomarkers, such as average

pixel intensity [59, 60] in regions of interest (ROI) are prone to error due to artefactual airway area perturbation. Any temporally varying blur of soft tissues can result in changes in the detected patent airway, and will disrupt the estimation of constriction kinematics, such as timing in consonant production [59]. Air-tissue boundary segmentation [61, 62, 63] is required as a pre-processing step in acquiring vocal tract area functions [64] and suffers in the presence of ambiguous boundaries with poor contrast. Velopharyngeal insufficiency [65, 66, 67, 68, 69] is caused by incomplete closure between the soft palate and the posterior and lateral pharyngeal walls, and its assessment can be hampered by signal loss near the soft palate.

Several deblurring methods in spiral scanning have been proposed in the literature [70, 71, 72, 73, 74, 75, 76], most of which require a measurement of a frequency offset image, also called a "field map" [70, 71, 72]. A previous study applied this approach to spiral RT-MRI of vocal tract [77] where spirals with two different echo times (TEs) were obtained in an interleaved fashion and a dynamic field map was estimated using each pair of consecutive images. This field map-based method showed improvement of image quality in the tongue and soft palate. The reconstructed images, however, could suffer from flickering artifact between consecutive images reconstructed with different TEs. This scheme also requires a compromise in temporal and/or spatial resolution [77] and is not applicable to previously-collected single-echo-time data.

An alternative approach is to estimate the field map directly from the dataset itself, known as "auto-focus" [73, 74, 75, 76]. Auto-focus methods employ an image-domain focus metric that provides local information about the presence of residual off-resonance artifacts based on the off-resonance PSF. A widely used metric is the absolute value of the

imaginary component of the image (after correcting for a coil phase) at an image location [74]. It assumes that the imaginary component should be zero when the local effects of off-resonance have been corrected. These methods have shown comparable results to the methods that acquire the field map. However, these are computationally demanding and performance depends on the focus metric used and can be sensitive to experimental factors, such as MRI sequence parameters, SNR, and the accuracy of coil sensitivity maps (especially their phase). Additionally, spurious minima of the focus metric can occur as the range of off-resonance at air-tissue interfaces (600 Hz at 1.5 T) is large enough to produce more than one cycle of phase accrual ($> 2\pi$) even during a short spiral readout ($\sim$2.5 ms) [73, 78, 79].

In this chapter, we demonstrate a simple dynamic off-resonance estimation method for spiral imaging where a dynamic field map is directly estimated from the phase of single-TE dynamic images after a coil phase compensation. We estimate complex coil sensitivity maps from the single-TE scan itself. Our approach does not require a dynamic two-echo measurement of a field map, nor the use of a focus metric. Therefore, it can be performed on conventional real-time spiral data without the need for additional scanning and is not computationally intensive. We evaluate this method using simulations and on an existing multi-speaker dataset of running speech. We demonstrate improvements in the depiction of air-tissue boundaries quantitatively using an image sharpness metric, and using visual inspection, and the practical utility of this method on a use case.

3.1 Theory

3.1.1 Spiral Imaging in the Presence of the Field Inhomogeneity

In spiral MRI, ignoring relaxation and noise, the signal equation of an object with a transverse magnetization $m_0(\mathbf{r})$ is given by

$$s(\tau) = \int_r m(\mathbf{r}) e^{-j2\pi f(\mathbf{r})\tau} e^{-j2\pi[\mathbf{k}(\tau)\cdot\mathbf{r}]} d\mathbf{r} \tag{3.1}$$

where $\tau \in [0, T_{read}]$ is time variable defining $\tau = 0$ as the start of the readout; T_{read} is the readout duration. $\mathbf{r}$ and $\mathbf{k}(\tau)$ are the spatial coordinate and the k-space trajectory, respectively. $m(\mathbf{r}) = m_0(\mathbf{r})C(\mathbf{r})e^{-j2\pi f(\mathbf{r})TE}$; $f(\mathbf{r})$ is the off-resonance frequency presented at $\mathbf{r}$; $C(\mathbf{r})$ is the complex coil sensitivity map.

Consider the image signal ($\tilde{m}(\mathbf{r})$) reconstructed from $s(\tau)$ without off-resonance correction as follows:

$$\tilde{m}(\mathbf{r}) = \int_{r'} m(\mathbf{r}') PSF(\mathbf{r}', \mathbf{r}; f(\mathbf{r}')) d\mathbf{r}' \tag{3.2}$$

where $PSF(\mathbf{r}', \mathbf{r}; f(\mathbf{r}')) = \int_0^{T_{read}} W(\tau) e^{-j2\pi\{f(\mathbf{r}')\tau + \mathbf{k}(\tau)\cdot(\mathbf{r}'-\mathbf{r})\}} d\mathbf{r}$ is a PSF of an imaging system using a particular k-space trajectory in the presence of $f(\mathbf{r})$; $W(\tau)$ denotes the pre-density compensation function for the trajectory. When $f(\mathbf{r})\cdot T_{read} \approx 0$, we can ignore a phase accrual due to off-resonance during the readout. Then, the PSF in Equation 3.2 is a sharp impulse at r so that the image signal in Equation 3.2 can be approximated by $\tilde{m}(\mathbf{r}) \approx m(\mathbf{r}) = m_0(\mathbf{r})C(\mathbf{r})e^{-j2\pi f(\mathbf{r})TE}$.

3.1.2 Field Map Estimation in Spiral Imaging

Consider spiral RT-MRI, where the image time series ($m_i(\mathbf{r}, t)$) for i-th coil is:

$$m_i(\mathbf{r}, t) \approx m_0(\mathbf{r}, t) C_i(\mathbf{r}) e^{-j2\pi f(\mathbf{r},t)TE} \tag{3.3}$$

where t represents time frame; $f(\mathbf{r}, t)$ is dynamic off-resonance; $C_i(\mathbf{r})$ is the complex coil sensitivity map that is spatially smooth and independent of time. Phase accrual during the spiral readout is ignored. Assuming that $m_0(\mathbf{r}, t)$ is real, we can compute an estimate of the dynamic field map, $\hat{f}(\mathbf{r}, t)$, as follows:

$$\hat{f}(\mathbf{r}, t) = \angle \hat{m_0}(\mathbf{r}, t)/(-2\pi TE) \tag{3.4}$$

where $\hat{m_0}(\mathbf{r}, t)$ denotes a coil-composite image using the optimal B1 combination [80], which is given by

$$\hat{m_0}(\mathbf{r}, t) = \Sigma_{i=1}^{N_c} m_i(\mathbf{r}, t) \hat{C}_i^*(\mathbf{r}) \tag{3.5}$$

here $\hat{C}_i(\mathbf{r})$ is an estimate of the sensitivity maps, N_c is the number of coil components; $\hat{C}_i^*(\mathbf{r})$ is the complex conjugate of $\hat{C}_i(\mathbf{r})$.

3.2 Methods

3.2.1 Implementation of Field Map Estimation for Speech RT-MRI

Figure 3.1 illustrates the proposed field map estimation process. The individual coil image frames $m_i(\mathbf{r}, t)$ are first reconstructed from raw k-space $s_i(\mathbf{k}, t)$ using sliding window

view-sharing with NUFFT [81]. For sliding window view-sharing, reconstructions were performed every 4 spirals using a temporal window of 13 spirals (fully sampled k-space). Note that this number matches to a frame rate of dynamic images to be reconstructed in off-resonance correction, which will be described more in Section 3.2.4. The multi-coil images are then merged into composite image frames $\hat{m_0}(\mathbf{r}, t)$ based on Equation 3.5 using complex coil sensitivity maps, whose estimation will be discussed later. $\hat{m_0}(\mathbf{r}, t)$ is then smoothed by convolution with a 3D Hanning window ($\mathbf{r}$-t) with size 3×3×3 to reduce noise, and masked by either of 0 or 1 based on a threshold (2% of maximum of the absolute squared value of the smoothed image) to control uninitialized values in air spaces that result from a lack of image signal. Consequently, a dynamic field map is estimated from the smoothed and masked images of $\hat{m_0}(\mathbf{r}, t)$ based on Equation 3.4.

Complex coil sensitivity maps $\hat{C}_i(\mathbf{r})$ (the 'i' subscript indicates the i-th coil element) are estimated from a temporally averaged and spatially low-pass filtered image. The individual coil image frames $m_i(\mathbf{r}, t)$ (shown in Figure 3.1) are averaged over time and low-pass filtered by a 2D Hanning window with size 15×15 pixels (full-width-half-maximum $\approx$ 8 pixels). Note that this low-pass filter is different from the smoothing applied to $\hat{m_0}(\mathbf{r}, t)$ and is comparable to a low-pass filter that takes 12.5% of the central part of the k-space. These settings were chosen empirically. Then, the resultant image $\overline{m}_i^{low}(\mathbf{r})$ is used to estimate the coil map by $\hat{C}_i(\mathbf{r}) = \overline{m}_i^{low}(\mathbf{r}) / \sqrt{\Sigma_i |\, \overline{m}_i^{low}(\mathbf{r})\, |^2}$. A drawback of this approach is that the spatially smooth portion of the time-averaged field map will be spuriously included in the coil sensitivity map, and will not be corrected, which will be extensively discussed in Section 3.4.

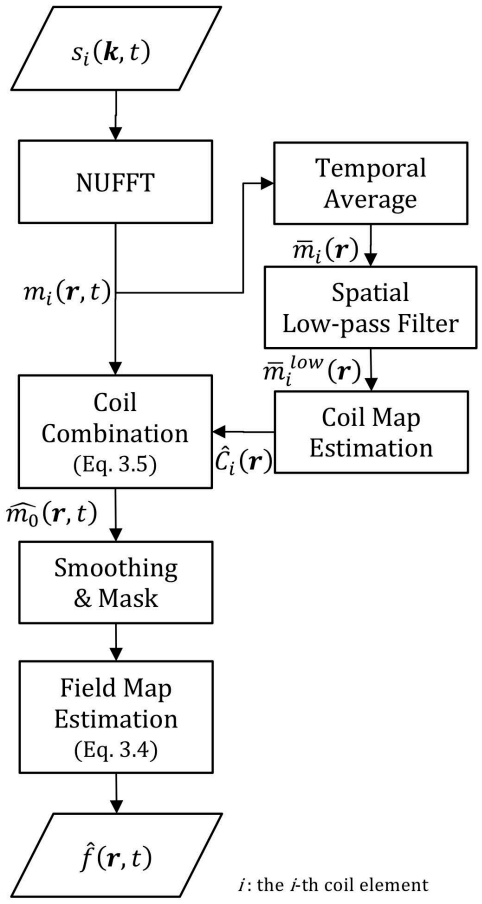

Figure 3.1: Flow-chart illustrating the proposed field map estimation method. The raw image frames from individual coils are first reconstructed from the raw k-space data using view-sharing with NUFFT. The coil sensitivity maps are estimated from the multi-coil image frames after temporal average and spatial low-pass filter. The multi-coil image frames are then merged into composite image frames using the complex coil maps by Equation 3.5. The composite images are smoothed and masked and a dynamic field map is estimated from the phase of the resulting image frames by Equation 3.4.

3.2.2 Simulation

To assess the accuracy of the proposed field map estimation, a simulation was performed with various spiral readout durations as follows: Cartesian images with two TEs (ΔTE = 1 ms) were acquired from a healthy subject at 5 postures including mouth open at

varying angles such as mouth fully open and mouth half open, mouth closed, and tongue tip raised to the front of the palate. For each of the postures, a reference field map was obtained from the phase difference between the images acquired at two TEs divided by ΔTE shown in Figure 3.2(a). Then, for a given spiral trajectory, spiral k-space data were synthesized from the magnitude of the Cartesian image from the first TE based on Equation 3.1. The reference field map was used to simulate off-resonance effects on the synthesized spiral k-space data. Those data simulations were performed with different readout durations varying from 0 ms to 6.3 ms with 0.63 ms increment. Finally, we estimated a field map from the simulated data and attempted to correct for off-resonance based on the estimated field map.

3.2.3 Application to Existing Speech RT-MRI Data

Experiments were performed on a speech RT-MRI dataset collected at our institution using a standardized vocal-tract protocol [49]. It currently contains more than twenty healthy subjects' data on a wide variety of speech tasks to capture salient, static and dynamic, articulatory characteristics of speech production as well as morphological aspects of the vocal tract [49]. Notice that the degree of blurring artifacts in their images varies depending on the subjects and speech tasks. We selected twenty subjects (n = 20, 10F/10M; age 19 – 31 years) with several speech tasks from the dataset.

Imaging was performed using a real-time interactive imaging platform (RT-Hawk, Heart Vista Inc, Los Altos, CA) [50] on a commercial 1.5 T scanner (Signa Excite, GE Healthcare, Waukesha, WI). The body coil was used for RF transmission, and a custom eight-channel upper airway coil [40] was used for signal reception. A 13-interleaf spiral

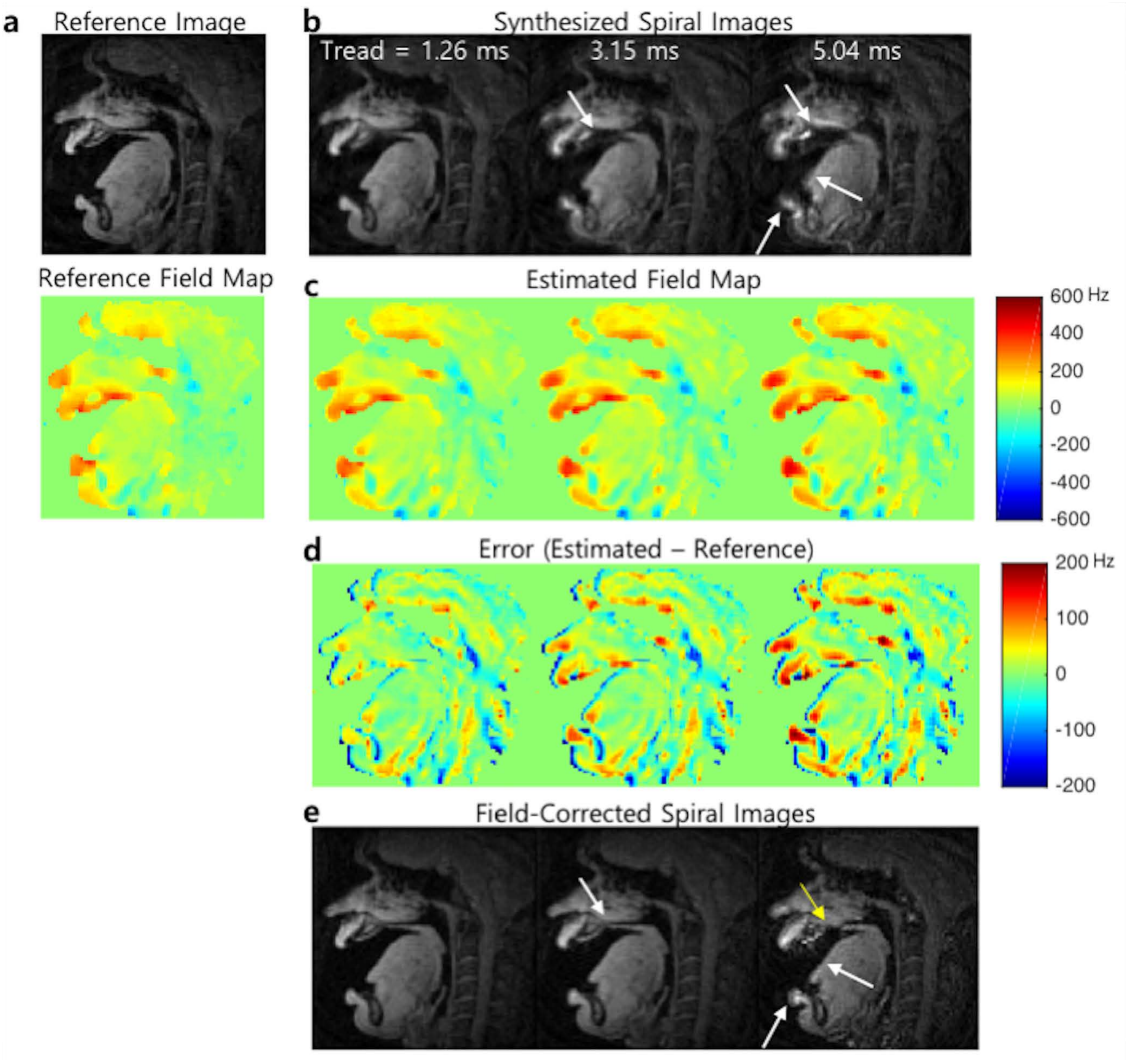

Figure 3.2: Representative simulation results. (a) A magnitude image and reference field map acquired from Cartesian dual-TE acquisition. (b) Synthesized spiral images using the magnitude image and reference field map with different readout durations (1.26, 3.15, and 5.04 ms). Off-resonance blurring is most apparent near the lips, hard palate, and tongue boundary and becomes worse with the longer readouts. (c) Field maps (Unit: Hz) estimated from the phase of the spiral complex images shown in (b). (d) Estimation errors in the field map (error maps amplified by a factor of 3 for better visualization). (e) Spiral images after correction for off-resonance based on the estimated field map represented in (c).

spoiled gradient echo pulse sequence was used. Imaging was performed in the mid-sagittal plane. Imaging parameters used were: $T_{read} = 2.52$ ms, spatial resolution = 2.4×2.4 mm^2,

slice thickness = 6 mm, field of view (FOV) = 200×200 mm^2, TR = 6.004 ms, TE = 0.8 ms, receiver bandwidth = ±125 kHz, and flip angle = 15°. In addition to the automatic shimming provided by the prescan calibration from the scanner, we performed a manual adjustment of the center frequency as described by Lingala et al. [40]. Specifically, we on-the-fly adjusted the center frequency in a way that the air-tongue boundary is sharp in the mid-sagittal plane while the subject being scanned is in a neutral open-mouth position.

3.2.4 Off-resonance Correction

We utilize an iterative approach [82, 83] where the off-resonance exponential term is approximated by a set of bases to improve computational speed and to reconstruct a deblurred image. We integrate this approach into a recent sparse-SENSE reconstruction method [40] that utilizes temporal finite difference constraint to improve time resolution in the time-series of spiral images of speech. Specifically, the off-resonance exponential term shown in Equation 3.1 is approximated by non-exponential bases at each time frame, by using histogram principal components (K=40 bins) and singular value decomposition analysis (L=6) described in Equation 19, 20 from Ref. [83]. Then the approximated bases are incorporated into the imaging model used in the sparse-SENSE reconstruction [40]. Raw k-space data and an estimated coil map are then fed into the reconstruction algorithm as inputs. In turn, it generates a corrected time-series of images. For evaluating the effectiveness of off-resonance correction, the original time-series of images were also reconstructed using the sparse-SENSE reconstruction without the modification. All the images were reconstructed with a temporal resolution of 24 ms/frame (41.66 frames/s,

4 spiral interleaves/frame, and with reduction factor R = 3.25). For implementation, a nonlinear conjugate gradient (CG) algorithm with NUFFT was coded using MATLAB (The MathWorks, Inc., Natick, MA) using 8 cores on a 16-core Intel(R) Xeon(R) CPU E5-2698 v3; 2.30GHz, with 40 MB of L3 cache. The computation time was ≈60 s to estimate the coil sensitivity maps and the field maps for 400 time frames from raw k-space data (≈10 s long dynamic images) and 30 and 180 mins to reconstruct images without and with off-resonance correction, respectively.

3.2.5 Sharpness Score

We introduce an image sharpness measure to investigate the impact of the proposed method on articulator air-tissue boundaries. We quantitatively compare the metric scores between the images with and without correction. We hypothesize that the proposed method would improve the image depiction at air-tissue articulator boundaries in two ways – the blurred-edge width be narrowed and/or the contrast at the edge be enhanced. We define an edge-slope metric for sharpness as follows:

Using a semi-automatic boundary extraction method [62], we extract the superior-posterior (upper) boundary and the inferior-anterior (lower) boundary as shown in Figure 3.3(a). Then, intensity profiles (grid lines) perpendicular to the upper and lower boundary (Figure 3.3(b)) of the patent airway are chosen and extracted from a reconstructed image series with a normalized intensity between 0 and 1, and linearly interpolated to generate ten times greater spatial resolution. Finally, the sharpness score (S) is calculated (Figure 3.3(c)) as follows;

$$S = \alpha CNR/d \tag{3.6}$$

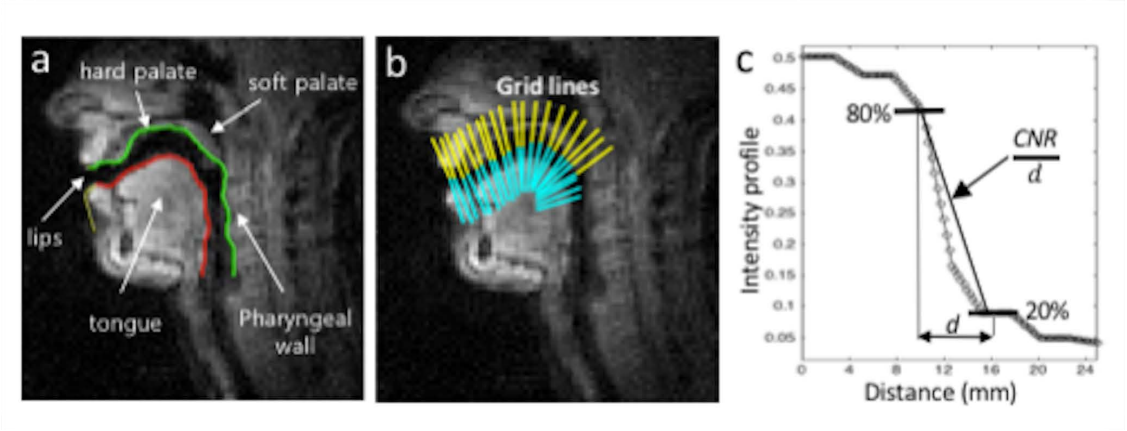

Figure 3.3: Illustration of articulator boundary identification and sharpness score evaluation. (a) Airway boundary segmentation with the upper (superior-posterior) boundary (green, color online) and the lower (inferior-anterior) boundary (red, color online). (b) Gridlines of the upper (yellow) and lower boundaries (cyan) at several locations along the airway are chosen to obtain intensity profiles. (c) Intensity profile of the gridline is plotted where a sharpness metric is measured as a slope between the points of 80% and 20% of the maximum intensity values (CNR/d).

where α is a scaling factor associated with the intensity normalization, $d = |p_{80} - p_{20}|$, and CNR $= (I(p_{80}) - I(p_{20}))/\sigma$; p_{80} and p_{20} are points (nearby the extracted boundary pixel location) at 80% and 20% of the maximum intensity value in grid lines, respectively; $I(p)$ is an intensity value at point p; σ is the standard deviation of an ROI outside the object where there is no signal. The sharpness score was calculated over valid time frames in which a distance between upper and lower boundary pixel locations is greater than 5 pixels. The sharpness score was compared using paired t-tests for statistical analysis, assuming that the samples collected along the grid lines are uncorrelated. A P value of < 0.001 was used to determine statistical significance.

3.2.6 Practical Utility of the Off-resonance Correction

Finally, to determine the practical utility of the off-resonance correction on an end-use case, we measure vocal tract distance, which is the desired metric that is often used in the speech RT-MRI analysis to obtain constriction degree [84, 85, 86] or vocal tract area function [87, 88, 89]. The distance metric is defined as the physical distance between the upper and lower boundaries shown in Figure 3.3(a). The boundaries are extracted using the aforementioned method [62] with the same initialization in both sets of images, without and with off-resonance correction. Distances were measured from both images.

3.3 Results

3.3.1 Simulation

Figure 3.2 shows a representative example (static posture with the mouth fully opened) of simulation results with different spiral readout durations. Off-resonance blurring is seen most clearly at the lips, hard palate, and tongue boundary and becomes more severe with the longer readouts as shown in Figure 3.2(b). As the duration of the readout is longer, the estimated field maps (Figure 3.2(c)) tend to be blurred and amplified in some areas such as the tongue surface and lips surface. Accordingly, high spatial frequency error can be seen in those areas (Figure 3.2(d)). The estimated field map fails to correct for the simulated off-resonance for the longer readout duration (> 5 ms) and the blurred anatomical structures remain unresolved.

3.3.2 Existing Speech RT-MRI Data

Figure 3.4 contains representative mid-sagittal image frames and the corresponding field map estimated for four subjects, which, on visual assessment, presented the most significant blurring artifacts among the twenty subjects. Note that subject numbers of 4, 6, 9, and 13 shown in Figure 3.4 correspond to those shown in Figure 3.5. For every image reconstructed with off-resonance correction, the soft palate, hard palate, and medial surface of the tongue become more intense and sharper compared to the blurred images (see yellow arrows). For all the four subjects, posterior to the alveolar ridge, the hard palate appears sharper up to the soft palate in the deblurred images. Correspondingly, in the estimated field maps, the regions that have shown blurred anatomical structures represent high off-resonance frequency values of > 200 Hz.

Figure 3.6 shows the profiles that are extracted at the solid lines in the sample image frames from the three subjects. For Subject 9, the intensity profile from the deblurred image provides a clear delineation of the soft palate movements. For Subjects 6 and 13, the intensity in the hard palate in the deblurred image sequence is more constant along time than the intensity value in the blurred image sequence. This result agrees with the fact that the hard palate, which is a bony structure covered by a thin layer of tissue, does not change its shape during speech production [63]. Furthermore, the intensity profile from the deblurred image exhibits a sharper boundary between the tongue and air.

Figure 3.7 illustrates one more example of correction result from Subject 4, especially showing the estimated field map over time. As depicted in the off-resonance frequency value vs. time profile, the proposed method enables the capturing of the dynamic change

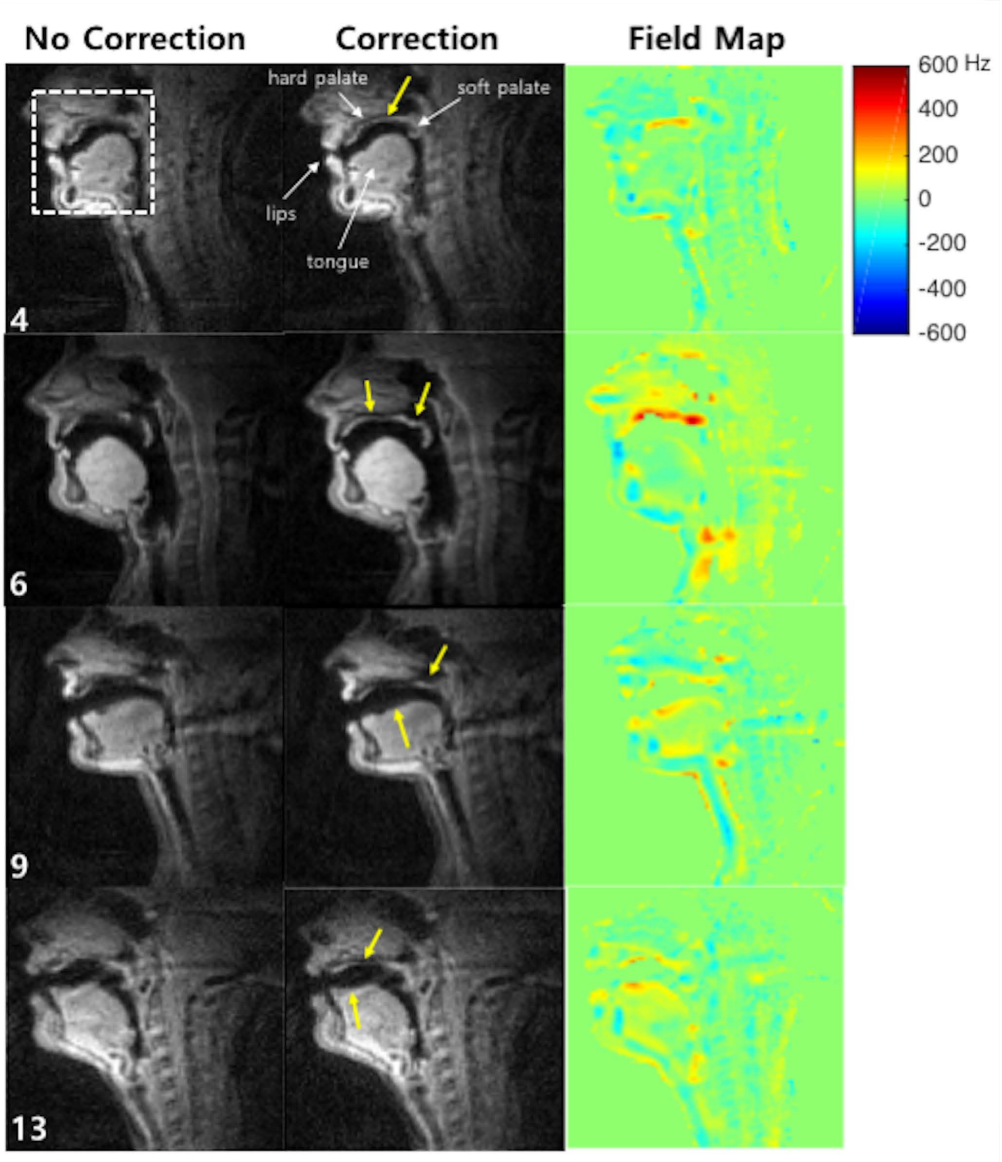

Figure 3.4: Representative mid-sagittal image frames of vocal tracts for four subjects, which, on visual assessment, presented the most significant blurring artifacts and were selected among the twenty subjects. The first and the second columns show images reconstructed with no correction and with correction, respectively. The last column shows the estimated field maps corresponding to those image time frames. Yellow arrows point out the regions that are most affected by off-resonance blurring, and corrected by the proposed method. (The video can be found in Supporting Information Video S1 at Wiley Online.)

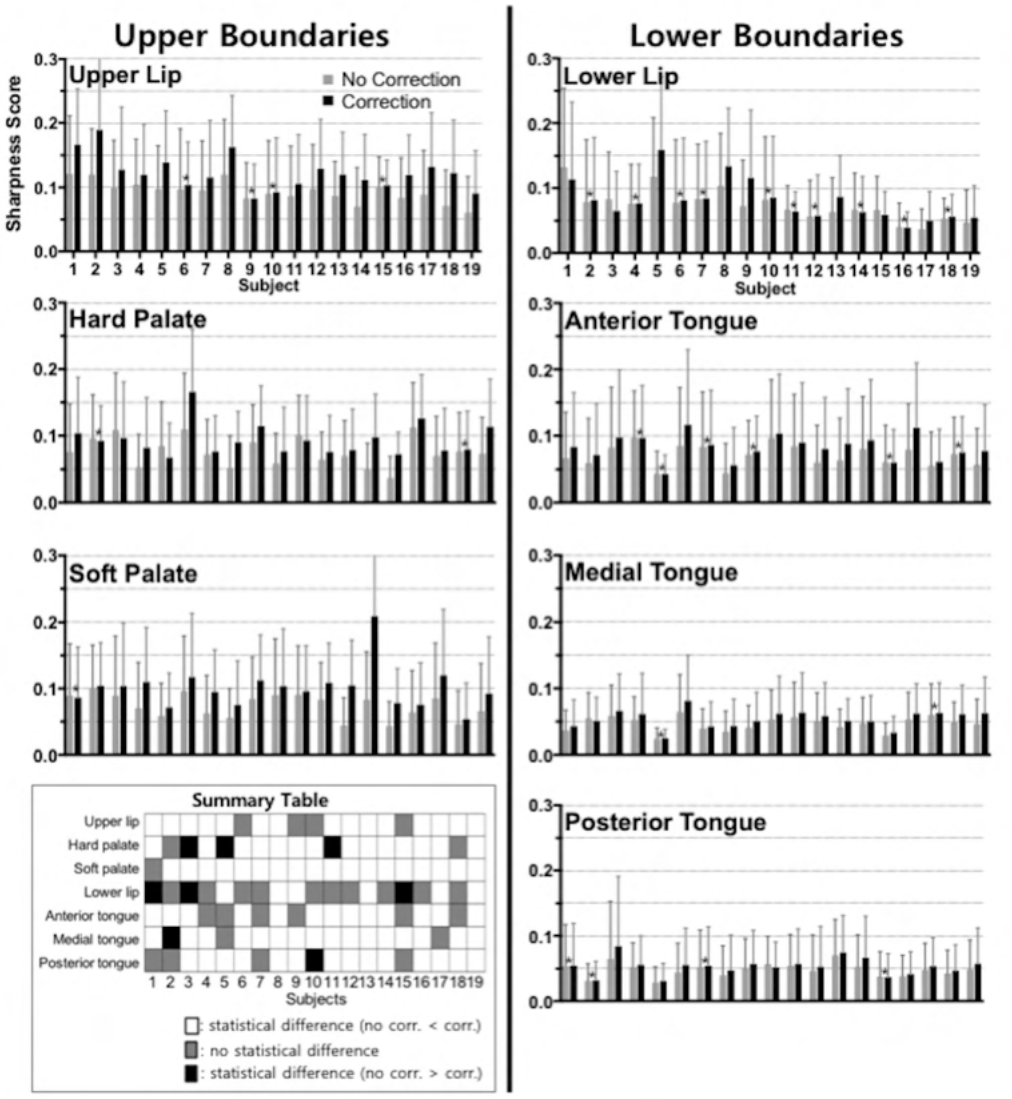

Figure 3.5: Sharpness without and with correction at different articulator boundary locations. Sharpness scores are measured at the upper boundaries (upper lip, hard palate, and soft palate) and lower boundaries (lower lip, anterior-, medial-, and posterior-tongue) along time. The mean and the standard deviation of the sharpness scores over time are shown here where the nineteen subjects are presented in descending order of average uncorrected sharpness score. A paired t-test was performed at each articulator boundary for each individual subject to test for the significance of the sharpness difference. The sharpness scores marked with an asterisk (*) were not found to be statistically different. All remaining scores were found to have significant mean differences ($P < 0.001$). The summary table in the bottom left panel summarizes the significance of the mean sharpness score difference between no correction and correction in three different categories: (white) no correction < correction, (gray) no significant difference between no correction and correction, and (black) no correction > correction.

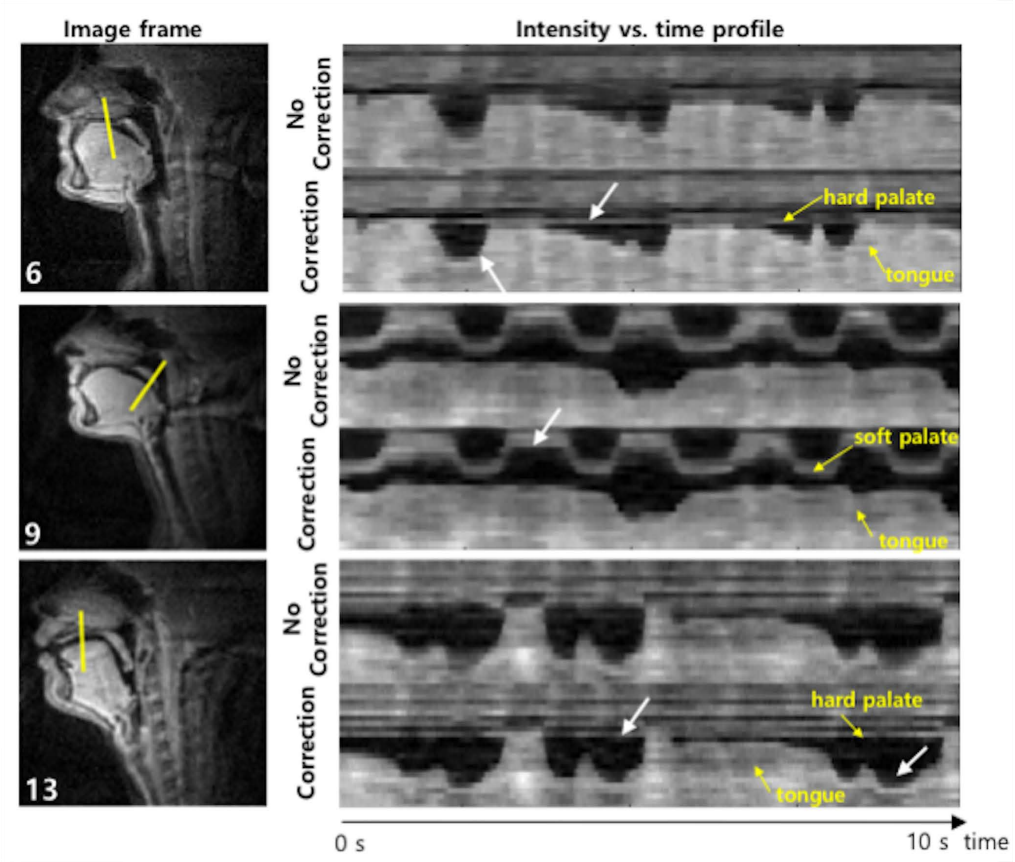

Figure 3.6: Illustration of improved sharpness of articulator boundaries. The first column shows an example frame for three different subjects and the second column shows intensity vs. time profiles marked by the solid lines in the first column images where each of the solid lines corresponds to one of the gridlines shown in Figure 3.3. For all subjects, the intensity time profiles from image sequences reconstructed with correction exhibit sharper boundary between tongue and air than that from image sequences with no correction. For Subject 9, the intensity profile from the correction provides a clear delineation of the soft palate movements. For Subjects 6 and 13, the correction method provides more constant intensity in the hard palate along time than image sequence with no correction.

in off-resonance at the tissue boundaries. Whereas the estimated field map shows high off-resonance frequency values at the hard palate and tongue boundaries over time, it shows a low frequency value at those boundaries during the event of the tongue touching the hard palate because there is no air between the tongue and hard palate (see white arrows).

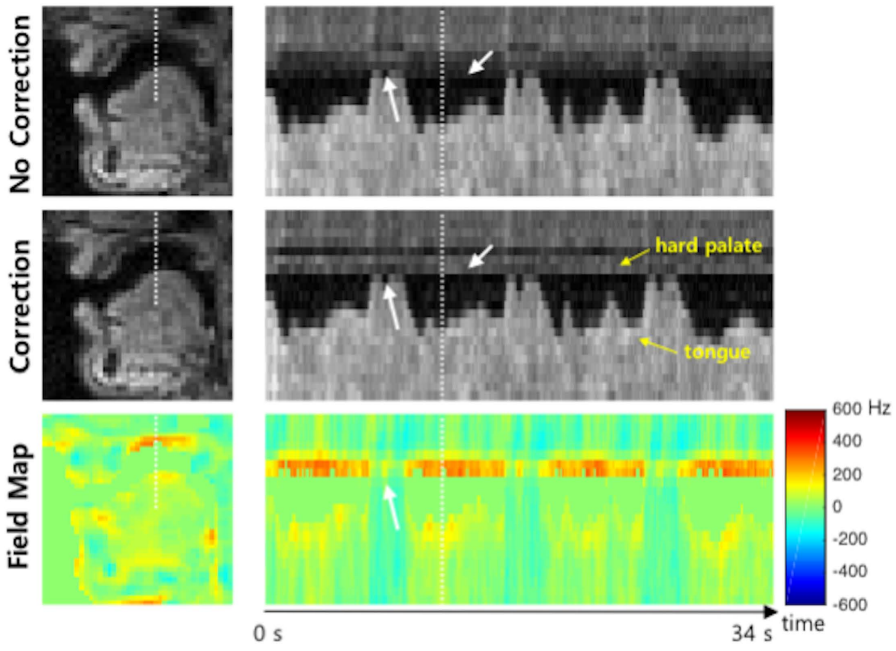

Figure 3.7: Illustration of the estimated field map over time. The first column shows example frames of reconstructed images and field map corresponding to the white dot box shown in Figure 3.4. The second column shows intensity vs. time profiles marked by the dot lines in the first column images. In the estimated field map, high off-resonance frequency values are shown at the hard palate (400 Hz) and tongue (200 Hz) boundaries over time except when the tongue contacts the hard palate. This is because when the tongue touches the hard palate, there is neither air and susceptibility difference between them. (The video can be found in Supporting Information Video S2 at Wiley Online.)

3.3.3 Sharpness Score

Figure 3.5 illustrates the sharpness scores and summary table. Sharpness scores without and with correction were measured at upper airway boundaries (upper lip, hard palate, and soft palate) and lower boundaries (lower lip, anterior-, medial-, and posterior-tongue) described in Figure 3.3 and averaged over time. The boundary extraction method used failed to segment the image from one subject due to low image quality, which was excluded in this sharpness analysis. Overall, the sharpness scores show a statistically significant difference in mean values (correction $>$ no correction, $P < 0.001$) for the subjects tested

at a majority of the boundaries. The lower lip shows negligible sharpness improvement in ten subjects and worse sharpness scores in three subjects when the correction was applied. The hard palate exhibits worse sharpness score in three subjects after correction compared to no correction, whereas fifteen subjects show improvement in sharpness score after correction.

3.3.4 Practical Utility of the Off-resonance Correction

Figure 3.8 illustrates an airway boundary segmentation result based on which the corresponding vocal tract distance is measured from images without and with the correction from Subject 6 shown in Figure 3.4. The uncorrected image exhibits noticeable errors in the segmentation due to off-resonance-induced blurring around the hard palate and soft palate, as indicated with arrows in Figure 3.8(a) and erroneous results on the corresponding vocal tract distance in those areas as shown in Figure 3.8(b).

3.4 Discussion

We have developed a dynamic field map estimation method for spiral RT-MRI where a dynamic field map is directly estimated from the phase of single-TE dynamic images after a coil phase compensation. We estimated complex coil sensitivities from single-echo data itself – temporally averaged and spatially low-pass filtered image. The proposed method could provide partial off-resonance correction for previously collected spiral RT-MRI datasets because it does not require the additional acquisition of the coil sensitivity

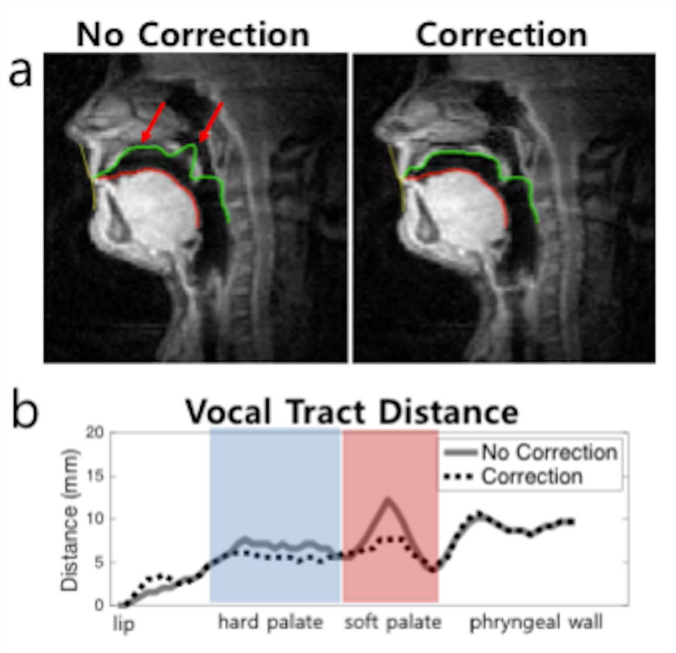

Figure 3.8: Representative illustration of airway boundary segmentation results on images without and with correction from Subject 6. (a) Airway boundary segmentation with the same initialization was performed on images without and with correction, to extract the upper and lower boundaries (green and red contours). As indicated by red arrows, the un-corrected image shows segmentation errors at the hard palate and soft palate due to off-resonance-induced blurring. (b) Vocal tract distance, defined as the distance between the upper and lower boundaries, is plotted. Discernible errors are observed around the hard palate and soft palate in the un-corrected data.

map. The proposed method is simple, computationally less demanding and when combined with the iterative image reconstruction, improves sharpness of the vocal tract articulator boundaries including the upper lip, hard palate, soft palate, and tongue boundaries (except for the lower lip) in a majority of the nineteen subjects tested. This has the potential to improve the downstream analysis of the dynamics of articulators during speech.

The signal equation in Equation 3.3 ignores phase accrual during the spiral readout. This assumption is not strictly true, and becomes less valid for long spiral readout duration and/or large resonant frequency offsets. In most cases, the PSF in Equation 3.2 is

no longer a sharp impulse nor pure real at the origin, which distorts the complex images used for the field map estimation. This PSF distortion is the basis of auto-focus methods. As readout duration is increased, phase and therefore the estimated field map tends to be erroneously blurred and amplified as can be seen in the simulation result (Figure 3.2(c)). These are practical limitations to the proposed method. Our findings suggest that for speech RT-MRI at 1.5 T, the proposed method will fail to work reliably for readout durations > 5 ms. An area of future work is investigating and predicting phase error caused by the non-ideal impulse with longer spiral readout.

An important issue in the field map estimation relates to the accuracy of the coil sensitivity maps. We low-pass-filtered the time-averaged image to estimate the coil map. This stems from the assumption that the coil maps contain only low spatial-frequency information and are stationary. Although the deblurred result demonstrated improvement in the sharpness at the boundaries compared to the original uncorrected images, the correction based on this coil map estimation depends on whether the anatomical structure and its field map are passed by its filtering process and show up in the sensitivity map or not. It corrects field that is not low-pass filtered and the kernel width of the low pass filter needs to be chosen as large as possible not to capture abruptly varying phase due to off-resonance at articulator boundaries while the size needs to be kept at some point to realize the spatially smoothly varying coil phase. However, it would be hard for one to optimize the choice of the size without knowing the object and the coil configuration in detail. In addition, as we described earlier, a precise shimming is required because the zero- and first-order field inhomogeneity is highly likely to be included in the estimated coil map and could be a main source of the error in the estimated field map. An alternative

solution to these limitations of the coil sensitivity map estimation would be to use an additional two-echo, static scan to estimate coil sensitivity maps that are free of phase due to off-resonance and B0 field inhomogeneity [90]. This solution is a work in progress in terms of comprehensive data collection and validation.

Another consideration for the field map estimation is to maintain an acceptable SNR level for the complex image. This is because error in phase is closely related to the SNR of the magnitude image (i.e., $\sigma_\theta = 1/SNR$) [91], as is the field map error (i.e., $\sigma_f = \sigma_\theta/(2\pi \times TE) = 1/(2\pi \times TE \times SNR)$). For example, if SNR = 10 and TE = 0.8 ms, the field map standard deviation is $\sigma_f = 19.9$ Hz. At readout duration of 2.5 and 5.0 ms, this error causes phase accrual error during spiral readout at the edge of the k-space of 18° and 36°, respectively. Therefore, it is important to have sufficient SNR with respect to the given TE and readout duration so that the accuracy of the estimation is less affected by noise. We chose a $3 \times 3 \times 3$ Hanning window (in $\mathbf{r}$-t) to maintain an adequate SNR > 60 in the ROI so that $\sigma_f < 3.3$ Hz theoretically. Note that SNR is approximately increased by $1/\sqrt{\Sigma(w_i)^2}$ where w_i is the weight of the Hanning window. However, the use of a large window could also result in smoothing out high frequency features.

Field map was estimated from images reconstructed using view-sharing with a temporal window of 78 ms (fully sampled k-space, 13 spirals). It is possible that articulator movement within the temporal window ($<<$ 78 ms) could result in temporal blurring of the field map or residual spiral artifact. Temporal blurring could give rise to errors in the artifact-corrected image as there is a discrepancy in the temporal windows between the estimated field maps and the corrected images. For example, if the tongue tip moves so

rapidly that temporal blurring around the tongue tip appears in the field map but not in the image to be reconstructed, there could be unresolved blurring by off-resonance around the tongue tip. Residual spiral artifact that affects the phase of the complex image also could lead to an erroneous field map. This is one of the limitations of the view-sharing scheme used in this work for field map estimation.

We excluded the noise-only area in the estimated field map using a mask. The mask was calculated from the distorted complex images where signal loss often manifests at some boundaries such as the hard palate and soft palate. Therefore, locations containing a high frequency feature could erroneously be masked out as zero. A more sophisticated method for generating field map masks should be investigated to mitigate this type of error.

We measured the sharpness score in several specific air-tissue boundary locations along the vocal tract to quantitatively evaluate the effectiveness of the proposed method. However, no metric is perfect, and the sharpness score was found to be sensitive to several factors. The boundary sharpness score is highly dependent on the location pre-identified as the true boundary. In the presence of signal loss due to off-resonance effect, the semi-automatic boundary segmentation method may fail. Specifically, the boundary location can be incorrectly identified. We often found this case in the original uncorrected image. For example, the boundary at the hard palate and soft palate is ambiguous and segmented erroneously as shown in Figure 3.8(a). In this case, it is hard to fairly compare the scores between the uncorrected and corrected images. To address this problem, in this work, we used a boundary location extracted from the corrected image to measure the score in both the uncorrected and corrected images.

Ultimately, it is important to evaluate the impact of the off-resonance correction on RT-MRI analysis in speech science. For example, in Figure 3.8, we have conducted segmentation of the vocal tract and shown observable improvement in the segmentation and measurement of the vocal tract distance after correction is applied as a use case example in RT-MRI analysis. Nevertheless, since in many cases the improvement would be not so much noticeable by visual inspection as shown in Figure 3.8, a better way to evaluate improvement in the segmentation result would be to compare the segmentation results with manual segmentation results. However, because of the very large number of frames in the RT-MRI datasets, performing a manual segmentation is not practical. Hence, in ongoing work, we are investigating a methodology to evaluate the segmentation results without manual reference.

3.5 Conclusion

We have developed and demonstrated a simple method for estimating a dynamic field map from spiral RT-MRI data of speech and incorporating the correction of the off-resonance into the constrained image reconstruction. We use the base image phase from single-echo data, after some initial processing, to estimate the field map directly by assigning the smoothly varying time-averaged phase to be used as coil phase and the residual high-frequency phase variations to the dynamic field map. We have demonstrated improvements in the depiction of the vocal tract articulators at several air-tissue boundaries both visually and through a sharpness metric, and the practical utility of this method on the boundary segmentation and distance metric as a use case example.

Chapter 4

Data-Driven Deblurring for 2D RT-MRI of Speech Production

Spiral data sampling is used in a variety of MRI applications due to its favorable properties. It requires only a few TRs to achieve Nyquist sampling of k-space, provides excellent velocity PSF [92, 57], and reduces motion artifacts due to its natural oversampling at the k-space center [93, 94]. Spiral sampling is well-suited for advanced reconstruction algorithms such as compressed sensing when combined with strategies such as under-sampling and golden angle scheme [58, 40]. Spiral imaging is also widely used for RT-MRI where the capability of capturing rapid motion is crucial such as in cardiac imaging and speech production imaging [40, 20, 95, 96, 97, 21, 98].

One major limitation of spiral sampling is image blurring due to off-resonance [56]. Off-resonance causes the accumulation of phase error along the readout in the k-space domain, resulting in blurring and/or signal loss in the image domain. To date, this artifact remains the predominant challenge for several RT-MRI applications: In speech RT-MRI it degrades image quality primarily at the air-tissue boundaries which include the vocal tract articulators of interest [77, 1, 35]. In cardiac RT-MRI, it degrades image

quality in the lateral wall and adjacent to draining veins, and around implanted metal (e.g., valve clips, etc) [97, 99]. In interventional RT-MRI, it may degrade image quality around the tools used to perform intervention (depending on the precise composition of the tools) [100]. These artifacts are most pronounced with long readout durations which is precisely when spiral provides the greatest efficiency. In speech production RT-MRI, the convention is to use extremely short readouts ($\leq$ 2.5 ms at 1.5 T) [35].

Many spiral off-resonance correction methods have been proposed in the literature. Most existing methods require prior information about the spatial distribution of the off-resonance, also called the field map, $\Delta f(x, y)$ [101, 102, 70, 71, 72, 78, 103]. For RT-MRI, this field map needs to be updated frequently throughout the acquisition window because of local off-resonance changes as motion occurs. Several research groups have proposed to estimate the dynamic field maps either from interleaved two-TE acquisition using the conventional phase difference method [77, 72, 104] or from single-TE acquisition after coil phase compensation [1]. Common limitations of these approaches are field map estimation errors due to off-resonance induced image distortion and/or reduced scan efficiency which is undesirable for RT-MRI.

Given a field map, the conventional approach to deblur the image is conjugate phase reconstruction (CPR) [101, 102] or one of its several variants [70, 71, 105]. One such variant is frequency-segmentation which reconstructs basis images at demodulation frequencies and applies spatially-varying masks to the basis images to form a desired sharp image. While it is an efficient approximation to assume off-resonance to be spatially varying smoothly, the assumptions are typically violated at air-tissue interfaces. Alternatively, iterative approaches [82, 83] are known to be effective at resolving abruptly varying

off-resonance at the cost of increased computation complexity. Note that neither iterative nor non-iterative approaches are able to overcome the performance dependence on the quality of the estimated field maps.

Recently, convolutional neural networks (CNN) have shown promise in solving this deblurring task. Zeng et al [106] have proposed a 3D residual CNN architecture to correct off-resonance artifacts from long-readout 3D cone scans. Specifically, off-resonance was framed as a spatially varying deconvolution problem. Synthetic data was generated by simulating zeroth-order global off-resonance at a certain range of demodulation frequency. The trained network was applied successfully to long-readout pediatric body MRA scans. Is there an underlying principle that explains why and how CNNs work well in this deblurring task? Perhaps it is the combinatorial nature of nonlinearities such as the rectified linear unit (ReLU) in CNN models. Traditional methods require field maps [101, 102, 70, 71, 72, 78, 103] or focus metrics [74, 73, 107] to estimate the spatially-varying mask. In contrast, CNNs utilize prior information about characteristics of off-resonance in the synthesized training data, while ReLU nonlinearities provide the mask to the convolutional filters, which enables spatially-varying convolution [108]. Once the network is trained, the feedforward operation of CNNs generates a desired sharp image given a blurry image input in an end-to-end manner, without explicit knowledge of field maps.

In this chapter, we attempt to establish a connection between the CNN architecture and traditional deblurring methods. We utilize a compact 3-layer residual CNN architecture to learn the mapping between distorted and distortion-free images for 2D spiral RT-MRI of human speech production. We consider this application [35, 55, 34] because

off-resonance appears as spatially (and temporally) abruptly varying blur, degrades image quality at the vocal tract articulators of interest, and therefore is a fundamental limitation to address. We leverage field maps estimated from a previously proposed dynamic off-resonance correction method that is presented in Chapter 3 [1]. Specifically, we synthesize spatially varying off-resonance by using the estimated field maps with various augmentation strategies. We test the impact of the augmentation strategies on deblurring performance and generalization in terms of several image quality metrics. We evaluate the proposed method using synthesized and real test data sets and compare its performance quantitatively using metrics and qualitatively via visual inspection against conventional deblurring methods.

4.1 Theory

4.1.1 Image Distortion Due to Off-resonance

In spiral MRI, off-resonance results in a spatially varying blur that can be characterized by a PSF [74]. Off-resonance causes the local phase accumulation in the k-space signal. In the spatial domain, this can be viewed as an object being convolved with spatially varying filter kernel (PSF) that is determined by the local off-resonance and trajectory-specific parameters such as a readout time map. Here, we briefly introduce this representation in the discrete domain, which we use throughout this chapter.

In the presence of off-resonance effects, the signal equation after discretization approximation [82, 83] can be expressed as:

$$y_i \approx \sum_{j=1}^{N_p} x_j e^{-j2\pi f_j t_i} e^{-j2\pi(\mathbf{k}_i \cdot \mathbf{r}_j)} \tag{4.1}$$

where $\mathbf{k}_i$ and $\mathbf{r}_j$ represent k-space and spatial coordinates for $i = 1, \ldots, N_d$ and $j = 1, \ldots, N_p$, respectively; y_i is the complex k-space measurement at time $t_i \in [T_E, T_E + T_{read}]$ defining $t_1 = T_E$ as the start of the readout; T_{read} is the readout duration and T_E is echo time. x_j is the transverse magnetization of an object at location $\mathbf{r}_j$. f_j is off-resonance frequency present at location $\mathbf{r}_j$. Here $e^{-j2\pi f_j t_i}$ is the local phase error that is induced by off-resonance present at location $\mathbf{r}_j$ and is multiplied to the k-space signal at $\mathbf{k}_i$. Note that Equation [1] can be expressed in a matrix vector form as $\mathbf{y} = \mathbf{A_f x}$ where $\mathbf{y} = (y_1, \ldots, y_{N_d}) \in \mathbb{C}^{N_d}$, $\mathbf{x} = (x_1, \ldots, x_{N_p}) \in \mathbb{C}^{N_p}$, $\mathbf{f} = (f_1, \ldots, f_{N_p}) \in \mathbb{R}^{N_p}$, and $\mathbf{A_f} \in \mathbb{C}^{N_d \times N_p}$ with $[\mathbf{A_f}]_{ij} = e^{-j2\pi f_j t_i} e^{-j2\pi(\mathbf{k}_i \mathbf{r}_j)}$. In the absence of off-resonance effects (i.e., $f = 0$), A_f is reduced to the conventional (non-uniform) Fourier basis matrix $\mathbf{A_0}$ with $[\mathbf{A_0}]_{i,j} = e^{-j2\pi(\mathbf{k}_i \mathbf{r}_j)}$. Without considering the off-resonance effect, we could reconstruct a blurry image $\tilde{\mathbf{x}} \in \mathbb{C}^{N_p}$ by applying $\mathbf{A_0}^{\mathrm{T}}\mathbf{W}$ to $\mathbf{y}$ as follows:

$$\tilde{\mathbf{x}} = \mathbf{A_0}^{\mathrm{T}}\mathbf{W}\mathbf{y} = \mathbf{A_0}^{\mathrm{T}}\mathbf{W}\mathbf{A_f}\mathbf{x} = \mathbf{H_f} \tag{4.2}$$

where T denotes the conjugate transpose of a matrix. $\mathbf{W} \in \mathbb{R}^{N_d \times N_d}$ is a diagonal matrix defining $[\mathbf{W}]_{i,i} = w_i$ where w_i denotes a density compensation weight at $\mathbf{k}_i$. Here, $\mathbf{H_f} \in \mathbb{C}^{N_p \times N_p}$ is a blurring operator matrix defining $[\mathbf{H_f}]_{j,k} = \Sigma_{i=1}^{N_d} w_i e^{-j2\pi f_k t_i}) e^{j2\pi \mathbf{k}_i(\mathbf{r}_j - \mathbf{r}_k)}$. The k-th column of $\mathbf{H_f}$ corresponds to the discretized PSF for a point-source located at $\mathbf{r}_k$. The effect of off-resonance can be seen as a spatially varying convolution since the PSF is shift-variant due to $e^{-j2\pi f_k t_i}$ with non-zero f_k. Whether the PSF is sharp or blurred is

dependent on the off-resonance frequency (f_k) given the trajectory-specific parameters – trajectory $\mathbf{k}_i$ and time map t_i for $i = 1, \ldots, N_d$. Likewise, the readout time ($T_{read} = t_{N_d}$) determines the shape of the PSF given f_k. For example, the larger f_k and/or the longer T_{read} are, the more phase error of $e^{-j2\pi f_k t_i}$ is accrued, therefore increasing the spread of the PSF.

4.1.2 Approximation of Spatially Varying Blur

The blurring operation is described in Equation 4.2 as a matrix vector multiplication. An approximate analytical solution to the deblurring problem is therefore:

$$\hat{\mathbf{x}} = (\mathbf{H_f}^{\mathrm{T}}\mathbf{H_f})^{+}\mathbf{H_f}^{\mathrm{T}}\tilde{\mathbf{x}} \tag{4.3}$$

where $[\mathbf{H_f}^{\mathrm{T}}\mathbf{H_f}]_{j,k} \approx \Sigma_{i=1}^{N_d} w_i e^{j2\pi(f_j - f_k)t_i} e^{j2\pi\mathbf{k}_i(\mathbf{r}_j - \mathbf{r}_k)}$ and $+$ denotes the pseudo-inverse. Noll et al. have shown that $\mathbf{H_f}^{\mathrm{T}}\mathbf{H_f}$ can be approximated well by an identity matrix under the condition that the phase term due to off-resonance is sufficiently small (i.e., $2\pi|f_j - f_k|t_i \ll \pi/2$) [105]. This is the underlying principle behind CPR and its variants. This condition is met whenever the off-resonance $f(x, y)$ due to B_0 inhomogeneity and susceptibility exhibits smooth spatial variation [5]. Under this assumption, the deblurred image can be obtained by projecting the blurred image onto the column space of $\mathbf{H_f}^{\mathrm{T}}$:

$$\hat{\mathbf{x}} \approx \mathbf{H_f}^{\mathrm{T}}\tilde{\mathbf{x}} \tag{4.4}$$

Note that CPR performs these projections in the frequency domain while other approaches [103, 109] including Equation [4] perform them in the spatial domain.

Next, we approximate $e^{-j2\pi f_k t_i}$ of Equation 4.1 by $e^{-j2\pi f_k t_i} \approx \Sigma_{l=1}^{L} b_{il} c_{lk}$. This approximation is supported by literature [83] for general choices of b_{il} and c_{lk}. For instance, time-segmentation approximates $b_{il} = b_l(t_i)$ and $c_{lk} = e^{-j2\pi f_k t_l}$ for a predetermined set of time points t_l while frequency-segmentation approximates $b_{il} = e^{-j2\pi f_l t_i}$ and $c_{lk} = c_l(f_k)$ for a predetermined set of frequencies f_l. Substituting such an approximation into $\mathbf{H_f}$ yields $[\mathbf{H_f}]_{j,k} \approx \Sigma_{l=1}^{L}[\Sigma_{i=1}^{N_d} b_{il} w_i e^{j2\pi \mathbf{k}_i(\mathbf{r}_j - \mathbf{r}_k)} c_{lk}]$. In matrix form, this can be expressed as:

$$\mathbf{H_f} \approx \Sigma_{l=1}^{L} \mathbf{A_0}^{\mathrm{T}} \mathbf{W} \mathbf{B}_l \mathbf{A_0} \mathbf{C}_l = \Sigma_{l=1}^{L} \mathbf{H}_l \mathbf{C}_l \tag{4.5}$$

where $\mathbf{B}_l \in \mathbb{C}^{N_d \times N_d}$ and $C^l \in \mathbb{C}^{N_p \times N_p}$ are diagonal matrices with $[\mathbf{B}_l]_{i,i} = b_{il}$ and $[\mathbf{C}_l]_{k,k} = c_{lk}$, respectively. Equation 4.5 can be viewed as a decomposition of the shift-variant blurring operator $\mathbf{H_f}$ as a sum of L ($L \ll N_p$; N_p = the number of pixels) convolutions $\mathbf{H}_l$ (i.e., approximately shift-invariant blurring operators) with prior weightings $\mathbf{C}_l$. In frequency-segmentation [105], $\mathbf{H}_l$ can be given by PSFs at a set of L equally spaced off-resonance frequencies and $\mathbf{C}_l$ is a spatially varying mask that has a diagonal element of 1 if a corresponding pixel needs to be assigned to the l-th off-resonance frequency or 0 otherwise. Other types of decomposition are possible for the spatially varying blur operator $\mathbf{H_f}$ in both MR and non-MR literature [83, 110, 111].

Substituting Equation 4.5 into Equation 4.1 yields

$$\hat{\mathbf{x}} \approx \Sigma_{l=1}^{L} \mathbf{C}_l{}^{\mathrm{T}} \mathbf{H}_l{}^{\mathrm{T}} \tilde{\mathbf{x}} = \mathbf{S}\mathbf{C}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\tilde{\mathbf{x}} \tag{4.6}$$

where $\mathbf{S} \in \mathbb{R}^{N_p \times LN_p} = [\mathbf{I}_{N_p} \cdots \mathbf{I}_{N_p}]$, $\mathbf{C}=\begin{bmatrix} \mathbf{C}^1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{C}^L \end{bmatrix}$ and $\mathbf{H} = [\mathbf{H}_1 \cdots \mathbf{H}_L]$. $\mathbf{I}_{N_p} \in \mathbb{R}^{N_p \times N_p}$ is an identity matrix. Equation 4.6 can be interpreted as a blurred image $\tilde{\mathbf{x}}$ being convolved by $\mathbf{H}$ with spatially varying weighting ($\mathbf{C}$), followed by summation ($\mathbf{S}$) along the dimension corresponding to L basis images.

4.1.3 Spatially Varying Deblurring using CNN

Interestingly, the solution described in Equation 4.6 resembles the feedforward operation of a simple two-layer CNN. Let us consider a two-layer CNN:

$$\hat{\mathbf{x}} = \mathbf{D}\mathbf{\Lambda}(\tilde{\mathbf{x}})\mathbf{E}^{\mathrm{T}}\tilde{\mathbf{x}} \tag{4.7}$$

where $\mathbf{D} \in \mathbb{C}^{N_p \times LN_p} = [\mathbf{D}_1 \cdots \mathbf{D}_L]$, $\mathbf{E} \in \mathbb{C}^{N_p \times LN_p} = [\mathbf{E}_1 \cdots \mathbf{E}_L]$, and $\mathbf{\Lambda}(\tilde{\mathbf{x}}) \in \mathbb{R}^{LN_p \times LN_p}$ is a diagonal matrix with 0 and 1 elements that are determined by the nonlinear ReLU activation output (i.e., $[\mathbf{\Lambda}(\tilde{\mathbf{x}})]_{i,i} = 1$ if $[\mathbf{E}^{\mathrm{T}}\tilde{\mathbf{x}}]_i > 0$, otherwise 0). $\mathbf{E}_l$ and $\mathbf{D}_l$ refer to convolution matrices associated with the l-th channel output and input at the first and second layers, respectively. With 1×1 convolutions in the second layer, $\mathbf{D}$ reduces to $\mathbf{D} = [d_1\mathbf{I}_{N_p} ... d_L\mathbf{I}_{N_p}]$ with channel-wise trainable weights d_l for $l = 1, \ldots, L$.

$\mathbf{D}$ in the second layer and $\mathbf{E}$ of the first layer of the CNN perform frequency summation $\mathbf{S}$ and input filtering $\mathbf{H}$ in Equation 4.6, respectively. The convolution matrices $\mathbf{D}$ and $\mathbf{E}$ are learned from the training data, whereas the convolution matrix $\mathbf{H}$ and spatially-varying mask $\mathbf{C}$ are determined by field maps $\mathbf{f}$. More importantly, the zero-one switching behavior of the element-wise ReLU induced nonlinear operator $\mathbf{\Lambda}$ can derive the spatially-varying weight adaptively from the different filtered inputs, and analogously

achieve the spatially varying weighting of the matrix **C** in Equation 4.6. Rather than relying on measuring exam-specific field maps, the CNN would learn from training samples to recognize and undo characteristic effects of off-resonance.

While the implementation of Equation 4.7 would be a direct replicate of current state-of-the-art off-resonance deblurring methods, we take this starting point and build on the following recent advances in machine learning to arrive at the proposed network architecture shown in Figure 4.1. First, we increase the CNN architecture from two to three layers by replacing the single convolutional layer **E** with two convolutional layers with the ReLU in between. For the single convolutional layer, there exist maximum 2^L distinct combinations of different convolution kernels. This is because there are L convolutional filter outputs for each spatial location and summing up the L coefficients at the second layer yields 2^L possible number of combinations due to the one or zero selection of the element-wise ReLU. From one to two cascaded layers, the number of configurations is increased from 2^L to $(2^{L_1} - 1)2^{L_2}$ as also similarly derived for encoder-decoder CNNs by Ye et al [108]. Second, we consider residual learning by adding a skip connection between input and output. Residual learning is widely used for medical image restoration [112, 113, 114, 115] and we experimentally found that it improves deblurring performance (comparison not shown).

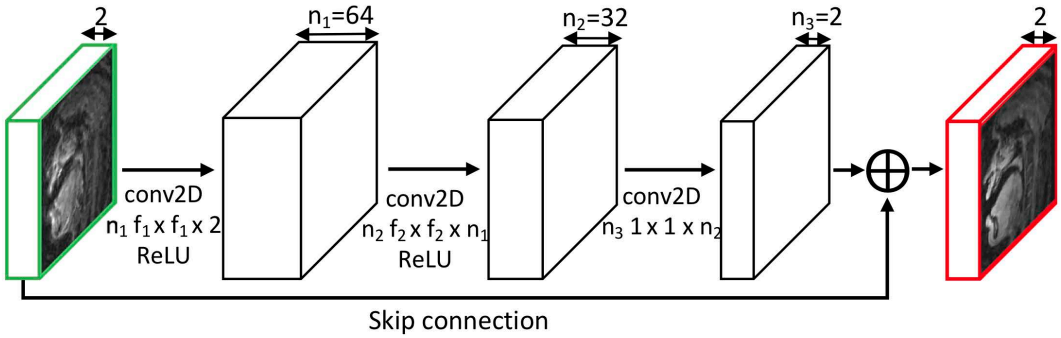

Figure 4.1: Proposed network architecture. The input is distorted complex images, and the output is distortion-free complex images, each consisting of two channels (real and imaginary). The first convolutional layer takes the input distorted image of size 84 $\times$ 84 $\times$ 2 and applies n_1 2D convolutions with filter size f1 $\times$ f1 $\times$ 2 (the last dimension 2 equaling the depth of the input), followed by the ReLU operation. The second layer takes the output of the first layer of size 84 $\times$ 84 $\times$ n_1 and applies n_2 2D convolutions with filter size $f_2 \times f_2 \times n_1$, followed by the ReLU operation. The third layer takes the output of the second layer of size 84 $\times$ 84 $\times$ n_2 and applies 2 2D convolutions with filter size $1 \times 1 \times n_2$. The output of the third layer is added to the input images via the skip connection to generate the final distortion-corrected image of size 84 $\times$ 84 $\times$ 2.

4.2 Methods

4.2.1 Network Implementation Details

The convolutional neural network (Figure 4.1) is comprised of 3 convolutional layers. The network architecture is practically implemented by real-valued operations. Although both input and output of the networks consist of two channels (real and imaginary components), we do not explicitly separate the real and imaginary image processing into separate streams and therefore information between the real and imaginary images is shared between the intermediate layers.

The filter widths are set to $n_1 = 64$, $n_2 = 32$, $n_3 = 2$. We choose $f_1 = 9$, $f_2 = 5$, and $f_3 = 1$. We experimentally determined convolutional filter sizes of f_1 and f_2 that give the

best deblurring performance in terms of image quality metric described in the following section.

We train the model in a combination of L_p loss and L_{gdl} gradient difference loss [116] between the prediction $\hat{\mathbf{x}}$ and ground truth $\mathbf{x}$:

$$L(\hat{\mathbf{x}}, \mathbf{x}) = L_p + \lambda L_{gdl} \tag{4.8}$$

where $L_p(\hat{\mathbf{x}}, \mathbf{x}) = \|\hat{\mathbf{x}} - \mathbf{x}\|_p^p$ and $L_{gdl}(\hat{\mathbf{x}}, \mathbf{x}) = ||\nabla_x \hat{\mathbf{x}}| - |\nabla_x \mathbf{x}|| + ||\nabla_y \hat{\mathbf{x}}| - |\nabla_y \mathbf{x}||$. We choose to use p=1 (i.e., L_1 loss) instead of p = 2 (L_2 loss) because L_1 loss is known to provide a sharp image prediction. We also add L_{gdl} loss because directly penalizing the differences of image gradient also enhances the sharpness of the image prediction [116]. We set $\lambda = 1$. We report the experimental results on the performance of choosing the different values of λ and choosing between L_1 and L_2.

For training the model, we use ADAM optimizer [117] with a learning rate of 0.001, a mini-batch size of 64, and 200 epochs. We implement the network with Keras using Tensorflow backend. The network is trained using a 16-core Intel Xeon E5-2698 v3 CPU and a graphical processing unit of Nvidia Tesla K80.

4.2.2 Generation of Training Data

We acquired data from 33 subjects on a 1.5T scanner (Signa Excite, GE Healthcare, Waukesha, WI) at our institution using a standard vocal-tract protocol [40]. The imaging protocol was approved by our Institutional Review Board. A 13-interleaf spiral-out spoiled gradient echo pulse sequence was used with the body coil for RF transmission and

a custom 8-channel upper airway coil for signal reception. Imaging was performed in the mid-sagittal plane while subjects being scanned followed a wide variety of speech tasks to capture various articulatory postures. Imaging parameters used included: TR = 6.004 ms, TE = 0.8 ms, T_{read} = 2.52 ms, spatial resolution = 2.4 × 2.4 mm^2, slice thickness = 6 mm, FOV = 200 × 200 mm^2, receiver bandwidth = ± 125 kHz, and flip angle = 15°.

Although this protocol uses a short spiral readout (2.52 ms), we found the off-resonance in this data corpus to be diverse and necessary to be corrected to obtain high-quality images. We employed the dynamic off-resonance correction (DORC) method [1] described in Chapter 3 to estimate dynamic field maps and to reconstruct dynamic images in conjunction with off-resonance correction. The resultant dynamic images (and field maps) were of size 84 × 84 × 400 (time) for each subject. We regarded these images ($\mathbf{x}$) and estimated field maps ($\mathbf{f}$) as ground truth. We split 33 subjects into 23, 5, and 5 subjects for the training, validation, and test sets, respectively. The validation set was used for choosing network parameters and performing validation experiments. The test set was used for evaluating the performance of the proposed method.

Blurred images $\tilde{\mathbf{x}}$ were simulated from the ground truth $\mathbf{x}$ with augmented field maps $\mathbf{f}'$ by employing Equation 4.2 ($\tilde{\mathbf{x}} = \mathbf{A_0}^{\mathrm{T}}\mathbf{W}\mathbf{A}_{\mathbf{f}'}\mathbf{x}$) frame by frame as illustrated in Figure 4.2A. Augmentation included a scale α and an offset β to the field map $\mathbf{f}$ such that $\mathbf{f}' = \alpha\mathbf{f} + \beta$ and synthesizing blurred images was based on the augmented field map. Note that $\alpha \neq 0$ would inherit original spatially varying off-resonance blur from the field maps f up to scale, whereas $\alpha = 0$ leads to spatially uniform blur analogous to the work of Zeng et al [106]. We added the offset β to simulate the zeroth-order frequency offsets in image space. Such offsets are a typical result of imperfect shimming. We considered

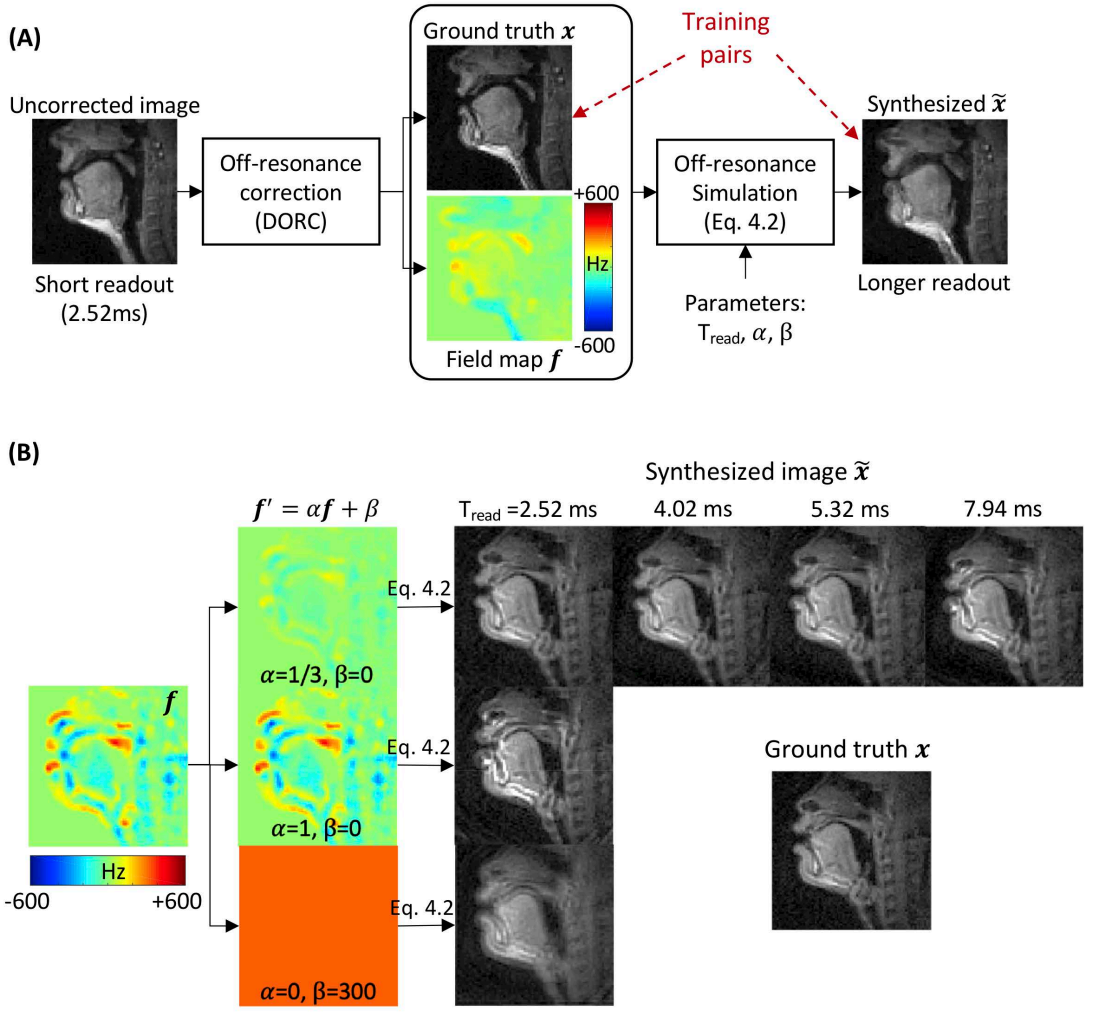

Figure 4.2: Generation of training data. (A) The ground truth image $\mathbf{x}$ and field map $\mathbf{f}$ are obtained from short readout (2.52 ms) data with off-resonance estimation and correction. Blurred images $\tilde{\mathbf{x}}$ are synthesized via simulating Equation 4.2 using the ground truth image $\mathbf{x}$ and field map $\mathbf{f}'$ augmented by α and β and different spiral readout durations (T_{read}). (B) The field maps are augmented by $\mathbf{f}' = \alpha\mathbf{f} + \beta$ with scale α ranging from 0 to 1 and constant offset β from -300 to 300 Hz. We also consider four different spiral readout durations (T_{read} = 2.52, 4.02, 5.32, and 7.94 ms). Those correspond to 13-, 8-, 6-, 4-interleaf spiral-out trajectories, respectively, with the same field of view and in-plane resolution.

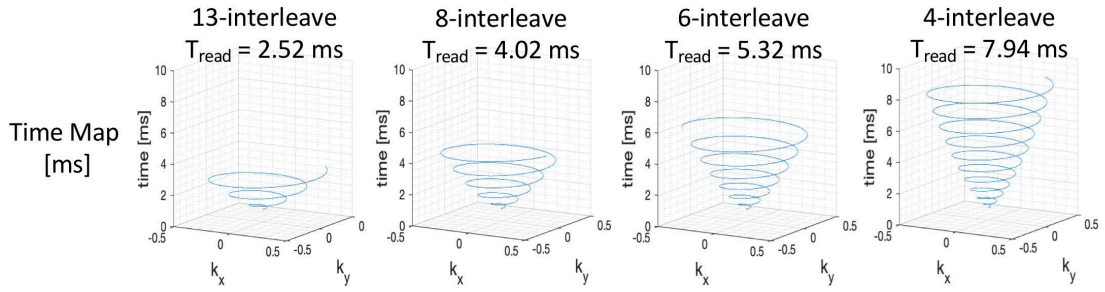

Figure 4.3: Time maps for four different spiral trajectories. Left to right: 13-, 8-, 6-, 4-interleaf spiral-out samplings, with readout times (Tread) of 2.52, 4.02, 5.32, and 7.94 ms, respectively, with the same field of view and in-plane resolution. Only one spiral interleave out of fully sampled interleaves is shown here. Here, TE = 0.8 ms.

four different spiral trajectories as shown in Figure 4.3. Those correspond to 13-, 8-, 6-, 4-interleaf spiral-out samplings, with readout times (T_{read}) of 2.52, 4.02, 5.32, and 7.94 ms, respectively, with the same field of view and in-plane resolution. Figure 4.2B contains examples of synthetic images. During the implementation of $\mathbf{A_f}$ in Equation 4.2, we approximated $e^{-j2\pi f_k t_i}$ as $e^{-j2\pi f_k t_i} = \Sigma_{l=1}^{L} b_{il} c_{lk}$ as described earlier and executed $\mathbf{A_f}$ with $L = 6$ times NUFFT [81] calls.

4.2.3 Validation Experiments

Here, we evaluated the impact of various data augmentation strategies (for f) on deblurring performance and generalization. Specifically, we trained the same network architecture using reference data from 23 subjects synthesized with different combinations of the scale factor α, frequency offset β, readout duration T_{read} (spiral trajectories), and the training data size as summarized in Table 4.1. We then measured the effectiveness of the different configurations by using the validation set (5 subjects) listed in Table 4.1.

EXP I–A. Off-resonance range: The off-resonance frequency range is typically unknown. We examined the impact of off-resonance frequency range (f_{max} denoting the maximum frequency value) in training data on deblurring performance. We trained the network on training data simulated with the 4 different values of fmax using $\alpha \in \{1/6, 1/3, 2/3, 1\}$, resulting in four trained networks, and compared their model performance on validation data with varying fmax as listed in Table 4.1. We considered $\beta = 0$ and $T_{read} = 2.52$ ms for both training and validation sets. Note that the frequency range from -625 to 625 Hz ($f_{max} = 625$ Hz) for the original field maps (i.e., when $\alpha = 1$ and $\beta = 0$).

EXP I–B. Frequency offset and training set size: We added a frequency offset β when synthesizing the training data. This is equivalent to simulating a constant frequency offset over image space. We considered two training configurations; one with $\beta = 0$ and one with $\beta \in \{-300, -200, \ldots, 300\}$. The former had N (= 9200) samples, while the latter had different sample sizes from N to 7N. The range of β from -300 to 300 Hz is deliberately chosen broadly to cover the maximum center frequency error that could be expected.

EXP II. Spatially varying versus spatially uniform blur: Recent work by Zeng et al [106] generated training data by simulating off-resonance at evenly spaced frequencies between ± 500 Hz. This approach, if generalized to spatially varying blur, could benefit situations where the field map is not available for synthesizing spatially varying blur. To test the generalizability of this approach and more importantly the necessity of the field maps for the spatially variant blur, we generated synthetic data of spatially invariant

blur by setting $\alpha = 0$ and $\beta \in \{-600, -550, ..., 600\}$ and of spatially variant blur by using $\alpha = 1$ and $\beta \in \{-300, -200, \ldots, 300\}$, the same setting as in EXP I–B. We then tested each trained network to another configuration setting by considering two validation configurations of spatially variant and invariant blur listed in Table 4.1.

EXP III. Readout duration: We investigated whether a network trained on a particular T_{read} (spiral trajectory) can be generalized to unseen Tread in test time. We used 13-, 8-, 6-, 4-arms trajectories corresponding to T_{read} of 2.52, 4.02, 5.32, 7.94 ms to synthesize blur data, while setting $\alpha = 1$ and $\beta \in \{-300, -200, \ldots, 300\}$.

For all experiments, the accuracy and robustness of deblurring performance were evaluated using multiple image quality metrics. We used the high-frequency normalized error norm (HFEN) [118], due to the expectation that high spatial frequency features would be restored after the blurred boundary is recovered. We also used common metrics, such as peak-signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [119].

Table 4.1: Summary of the parameters used to generate training and validation sets for validation experiments

	Train				**Validation**		
	α	β **(Hz)**	T_{read} **(ms)**	**No. of Samples**	α	β **(Hz)**	T_{read} **(ms)**
EXP I-A	1/6 1/3 2/3 1	{0}	2.52	N[a]	1/6 1/3 2/3 1	{0}	2.52
EXP I-B	1	{0} {-300, -200, ..., 300} {-300, -200, ..., 300} {-300, -200, ..., 300}	2.52	N N 3.5N 7N	1	{0}	2.52
EXP II	1 0	{-300, -200, ..., 300} {-600, -550, ..., 600}	2.52	3.5N 4N	1 0	{0} {-300, -200, ..., 300}	2.52
EXP III	1	{-300, -200, ..., 300}	2.52 4.02 5.32 7.94	7N	1	{-300, -200, ..., 300}	2.52 4.02 5.32 7.94

Each row represents a configuration set of α, β, T_{read}, and number of samples for train set or α, β, T_{read} for validation set.

[a]N (no. of samples) = 9200 84 $\times$ 84 image pairs of distorted and distortion-free complex images.

Abbreviations: EXP, experiment.

4.2.4 Evaluation using Synthetic Test Data

We tested the model on unseen synthetic test data from 5 subjects (independent from training and validation datasets). We used the model trained in EXP III. The test data were simulated from all spiral trajectories without any augmentation (i.e., $\alpha = 1$, $\beta = 0$). For comparison, we applied 1) frequency-segmentation-based multi-frequency interpolation (MFI) [71] and 2) model-based iterative reconstruction (IR) [82] into the synthetically generated test k-space data (y). For MFI, we obtained a deblurred image by $\hat{\mathbf{x}} = {\mathbf{A_f}}^{\mathrm{T}}\mathbf{W}\mathbf{y}$. For IR, we obtained a deblurred image by solving $\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A_f}\mathbf{x}\|_2^2$ iteratively by using conjugate gradient with 16 iterations. In both methods, we used the ground truth field map $\mathbf{f}$ to construct $\mathbf{A_f}$. HFEN, PSNR, and SSIM metrics were used for evaluation.

It is worth noting that the IR method is known to provide more accurate results than the non-IR method for abruptly varying off-resonance in space. This IR approach could provide the best achievable deblurring performance given the ground truth field map $\mathbf{f}$, although it is not available in practice.

4.2.5 Evaluation using Real Experimental Data

We applied the trained network to real data. We acquired spiral RT-MRI data with four readout durations (2.52, 4.02, 5.32, and 7.94 ms) from two subjects and performed image reconstruction as described by Lingala et al [40]. We performed the deblurring on the reconstructed images frame by frame by using the trained network. We compared results with DORC [1]. This auto-calibrated method estimates dynamic field maps from

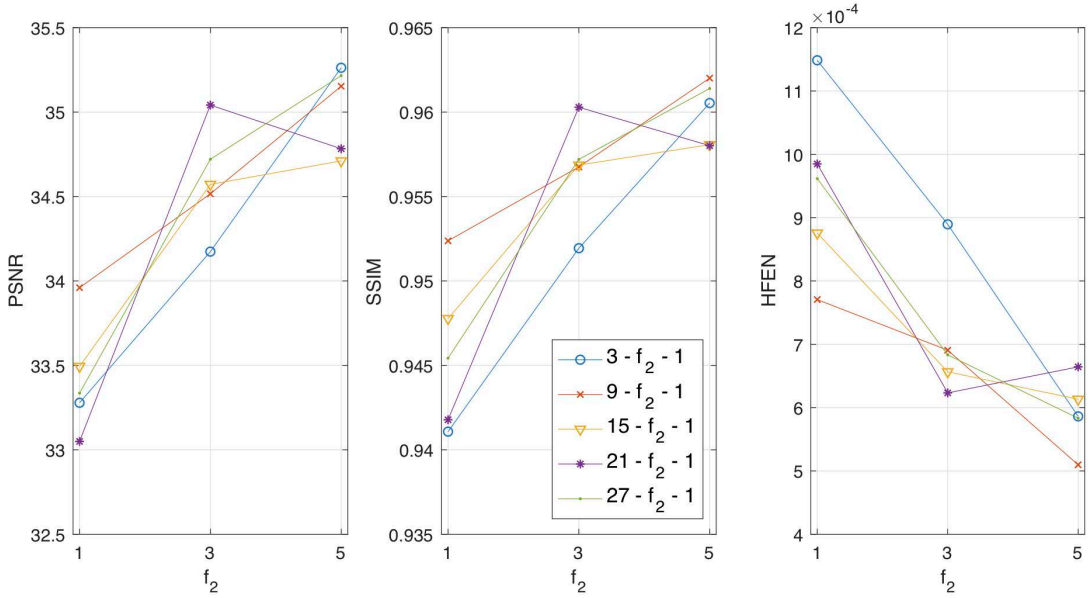

Figure 4.4: Quantitative comparison of deblurring performance for different filter sizes. Image Quality Metrics (PSNR, SSIM, and HFEN) are shown as a function of filter sizes. Each curve with the corresponding color represents a trained network with different filter size of f_1, whereas the horizontal axis represents filter size of f_2. The filter size of the first and second convolutional layers (f_1 and f_2, respectively) were varied from 3 to 27 and from 1 to 5, respectively whereas $f_3 = 1$ was kept constant. We choose $f_1 = 9$, f_2= 5, and f_3= 1 (f_1-f_2-f_3=9-5-1).

single-TE blurred image itself after coil phase compensation, with no scan time penalty and iteratively reconstructs off-resonance-corrected image using the estimated field map.

4.3 Results

4.3.1 Convolution Filter Size

Figure 4.4 shows a comparison of deblurring performance for different filter sizes in terms of PSNR, SSIM, and HFEN. For all comparison metrics, a combination of the filter sizes, $f_1 = 9$, $f_2 = 5$, and $f_3 = 1$, yields the best performance, which is chosen for the rest of the work presented in this chapter.

4.3.2 Loss Function

Figure 4.5 shows the deblurring results as a function of penalty on L_p loss and gradient difference loss in the model training. For L_2 loss, as gradient difference loss penalty λ increases from 0.01 to 1 (3rd and 4th columns), image sharpness is improved visually. This trend can also be observed in the three metrics shown in (A). Compared to L_2 loss, L_1 loss with the same value of λ visually improves the sharpness at the soft palate (4th and 5th columns) as well as it improves the metrics, although the visual difference might not be observed as clearly as that can be observed when increasing λ from 0.01 to 1. We choose to use p = 1 and $\lambda = 1$.

4.3.3 Learned Convolution Kernels

Figure 4.6 visualizes the representative examples of learned kernels from the first and second convolutional layers. Kernels shown at the left and right in panel (A) were applied to real and imaginary input channels, respectively. Some of the kernels are shown to be circular symmetric with varying degrees of sharpness, which is similar to off-resonance-induced PSFs observed in spiral trajectories. It is also observed that some of the learned convolution kernels lack circular symmetry.

4.3.4 Validation Experiments

Figure 4.7A shows deblurring performance (SSIM and PSNR) as a function of the range of off-resonance (f_{max}) in the train and validation sets (EXP I-A). For corrected images, each curve represents a separate network trained with different f_{max}. For no correction (a, black dashed curve), the values of SSIM and PSNR gradually decrease as f_{max} increases in

(A)

L1 or L2	λ (L_{gdl})	PSNR	SSIM	HFEN (x1000)
Baseline (input)		27.17	0.851	4.560
L2	λ=0	33.74	0.949	0.685
	λ=0.01	33.44	0.946	0.759
	λ=0.1	33.38	0.946	0.821
	λ=1	34.68	0.960	0.612
L1	λ=0	34.96	0.960	**0.553**
	λ=0.01	34.69	0.958	0.620
	λ=0.1	34.85	0.959	0.597
	λ=1	**35.20**	**0.962**	0.589

(B)

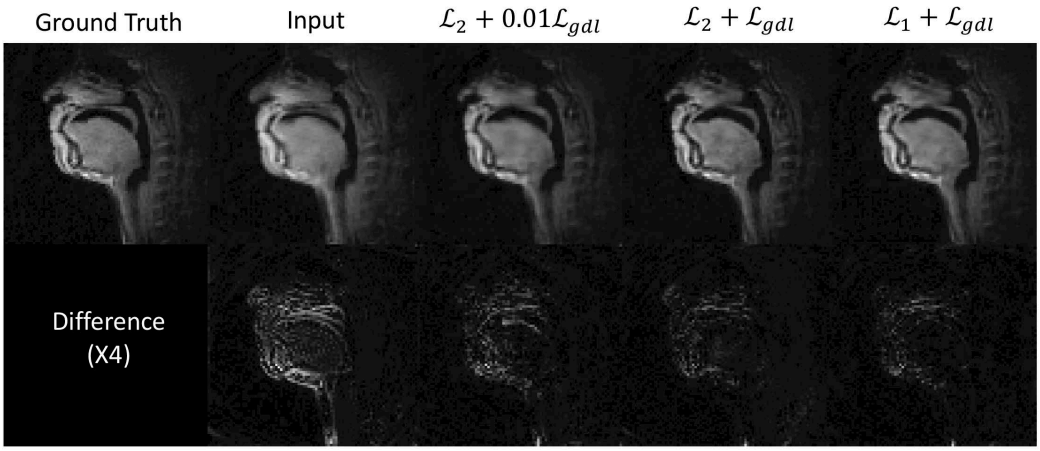

Figure 4.5: Combination of L_p loss and gradient difference loss in the model training. (A) Image Quality Metrics (PSNR, SSIM, and HFEN) are shown. L_1 loss in a combination with gradient difference loss with $\lambda = 1$ exhibits the highest values of PSNR and SSIM over other combinations. (B) Representative image results for combinations of L_p loss and gradient difference loss.

the validation sets, which is likely due to worsened blurring artifact. All but the network trained with f_{max} of 104 Hz (b, blue curve) improve image quality for all f_{max} tested compared with the uncorrected case. Each network is shown to exhibit the highest values of SSIM and PSNR for validation data f_{max}, with which the network is trained. The performance then quickly degrades for f_{max} greater than that of its best performance

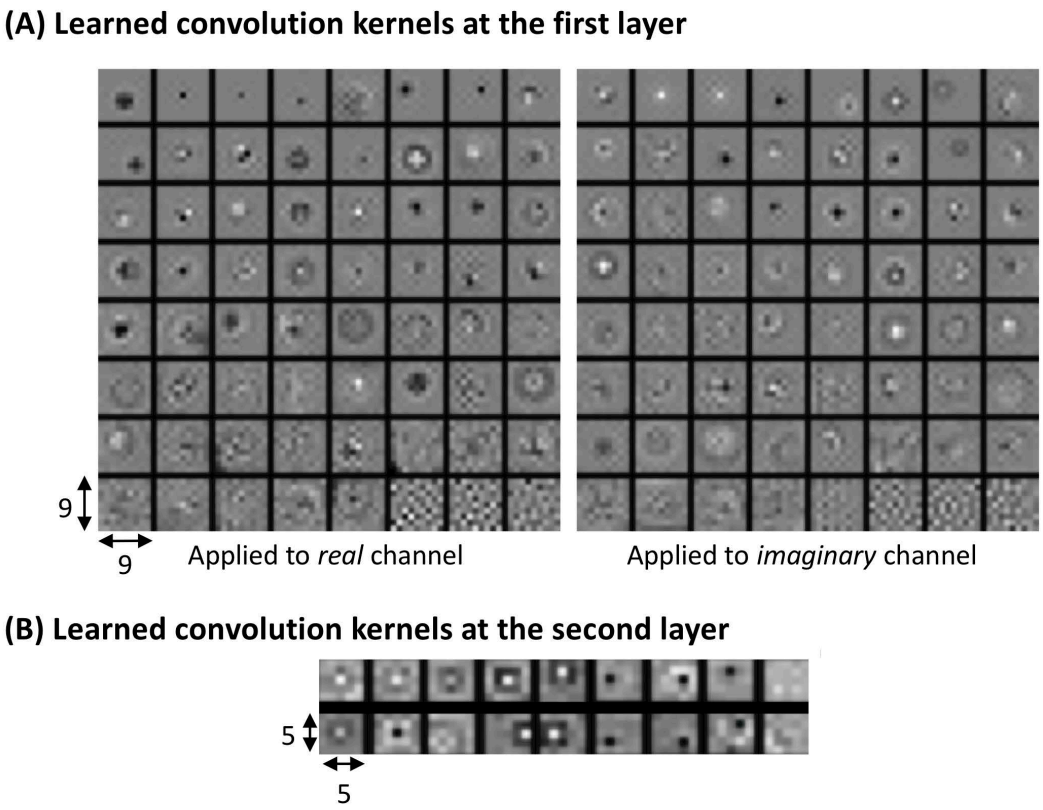

Figure 4.6: Representative examples of learned convolution kernels at the first and second convolutional layers. (A) Kernels shown at the left and right are respectively corresponding to 64 $9 \times 9 \times 1$ kernel weights applied to real and imaginary input channels in the first layer. The majority of kernels exhibit circular symmetry which corresponds to the expected shape of off-resonance PSFs in spiral MRI. (B) 18 $5 \times 5 \times 1$ kernels are visualized out of 32 convolutional kernels of size $5 \times 5 \times 64$ in the second layer. The kernels in the first 8 columns represent structured patterns whereas the kernels shown in the last column represent unstructured patterns.

(see c, orange curve). In Figure 4.7B, representative frames are shown. The uncorrected image presents $f_{max} = 625$ Hz as shown in (a) and the model trained with f_{max} of 625 Hz shown in (e) exhibits the best deblurring performance qualitatively against other models shown in (b-d, f). We observe that it is essential for the frequency range of the training set to be a superset of the validation data.

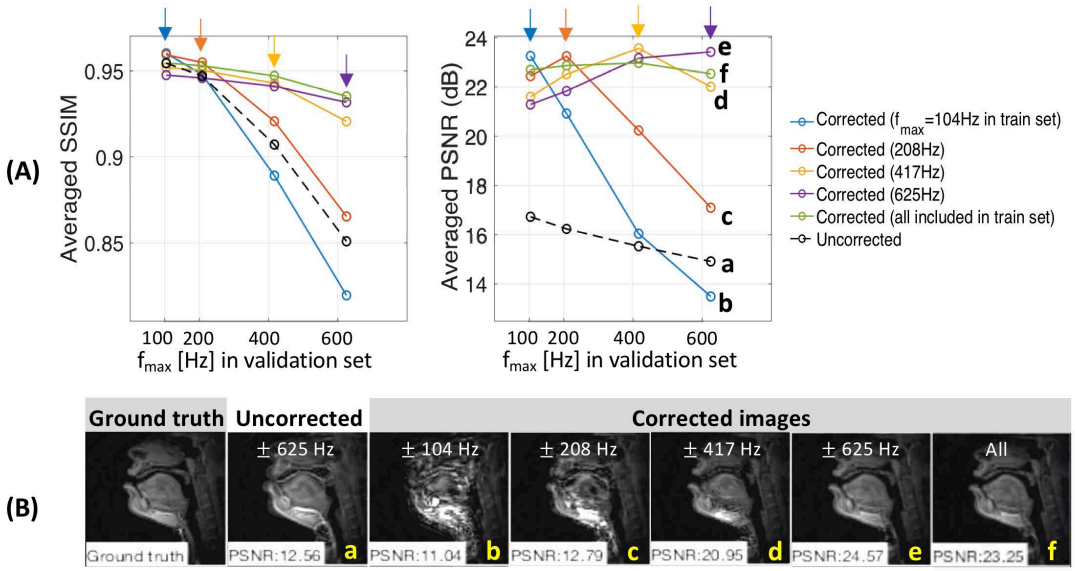

Figure 4.7: Performance depends on the training set (EXP I-A). (A) Averaged SSIM and PSNR as a function of the frequency range (f_{max} denoting the maximum frequency value) for the uncorrected (black dotted) and corrected (non-black solid) images in the validation set. Color (non-black) represents a separate network trained with different f_{max}. Note that higher SSIM and PSNR correspond to better performance. The best performance is achieved when the training and validation datasets share the same range of off-resonance (arrows). When severe off-resonance appears in the validation data as the off-resonance range is increased, performance for the network trained with fmax less than that of the validation data quickly degrades. (B) A representative example of the ground truth, uncorrected, and corrected images. The uncorrected image had f_{max} = 625 Hz and was deblurred using models trained with f_{max} of 104, 208, 417, 625 Hz, and all of them. We observe that it is essential for the frequency range of the training set to be a superset of the validation set.

Table 4.2: Quantitative evaluation of model performance in terms of the PSNR, SSIM, and HFEN, without and with offset β, and as a function of the number of training samples (EXP I-B)

Training data		**PSNR**	**SSIM**	**HFEN**
offset β	**no. of samples**			**($\times$1000)**
Without offset $\beta = 0$	N[a]	32.75	0.979	0.274
With $\beta \in$ {-300,-200,...,300}	N	33.43	0.980	0.125
	3.5N	34.40	0.983	**0.094**
	7N	**34.53**	**0.983**	0.103
No correction		26.82	0.951	0.890

[a]N = 9200 84 $\times$ 84 image pairs of distorted and distortion-free complex images.
Abbreviations: EXP, experiment; HFEN, high-frequency normalized error norm; PSNR, peak SNR; SSIM, structural similarity index.

We also found that adding frequency offsets when synthesizing the training data help the network perform better, which is shown in Table 4.2 (EXP I-B). In addition, as the training samples increase from N to 7N, the performance improvement can be found of 2.1, 0.003, and 0.022 in PSNR, SSIM, and HFEN, respectively.

Table 4.3 presents the average PSNR, SSIM, and HFEN values for the quantitative comparison of model performance on spatially invariant and variant blur (EXP II). The network trained on spatially variant blur has superior PSNR, SSIM, and HFEN values for both validation data of spatially invariant and variant when compared to no correction and the network trained on spatially invariant blur. The network trained on spatially invariant blur improves all the image metrics for the validation data of spatially invariant blur compared to no correction, but when applied to spatially variant blur, it presents even lower values in all the metrics than no correction, indicating it fails to deblur spatially varying off-resonance.

Table 4.3: Quantitative evaluation of model performance on spatially invariant and variant blur in terms of the PSNR, SSIM, and HFEN (EXP II)

Validation data	**Training data**	**PSNR**	**SSIM**	**HFEN (×1000)**
Spatially variant blur	no correction	36.82	0.951	0.890
($\alpha = 1$, $\beta = 0$)	Spatially variant blur ($\alpha = 1$, $\beta \in \{-300, -200, \ldots, 300\}$)	**34.40**	**0.983**	**0.094**
	Spatially invariant blur ($\alpha = 0$, $\beta \in \{-600, -550, \ldots, 600\}$)	26.53	0.934	1.414
Spatially invariant blur	no correction	27.01	0.897	0.346
($\alpha = 0$, $\beta \in \{-300, -200, \ldots, 300\}$)	Spatially variant blur ($\alpha = 1$, $\beta \in \{-300, -200, \ldots, 300\}$)	**35.60**	**0.986**	**0.037**
	Spatially invariant blur ($\alpha = 0$, $\beta \in \{-600, -550, \ldots, 600\}$)	35.54	0.986	0.046

[a] N = 9200 84 × 84 image pairs of distorted and distortion-free complex images.
Abbreviations: EXP, experiment; HFEN, high-frequency normalized error norm; PSNR, peak SNR; SSIM, structural similarity index.

Figure 4.8 presents image metrics as a function of readout duration (T_{read}) in the train and validation sets (EXP III). For each of T_{read} tested, the network trained with the corresponding T_{read} has superior PSNR, SSIM, and HFEN compared to no correction and networks trained with other T_{read} (see arrows). For each trained network, the performance then quickly degrades for T_{read} longer than that of its best performance. In contrast to the individually trained networks, the network trained using all the T_{read} (see green, "All included") exhibits consistent improvement over the T_{read}.

4.3.5 Evaluation using Synthetic Data

Figure 4.9 compares image metric results as a function of T_{read} for images with no correction and after correction using various methods applied to synthetic test data of 5 subjects. The proposed method (purple) is compared against no correction (blue), MFI (red), and IR (orange). For all methods, performance gradually degrades as readout duration increases. Overall, IR has superior PSNR, SSIM, and HFEN values for all readout durations, followed by the proposed method, MFI, and no correction. MFI had lower PSNR than that for no correction for $T_{read} \geq 2.52$ ms (black arrows).

Figure 4.10 contains representative image frames of the ground truth, uncorrected image, and images corrected by various comparison methods. In Figure 4.10A, blurring is clearly seen around the lips, tongue surface, and soft palate in the uncorrected image (yellow arrows). After correction, MFI even deteriorates the delineation of the boundaries in those regions, whereas the IR method almost perfectly resolves the blurring artifact as also clearly observed in the difference images (see Figure 4.10B). The proposed method successfully resolves the blurring artifact in those regions, which is visually comparable to

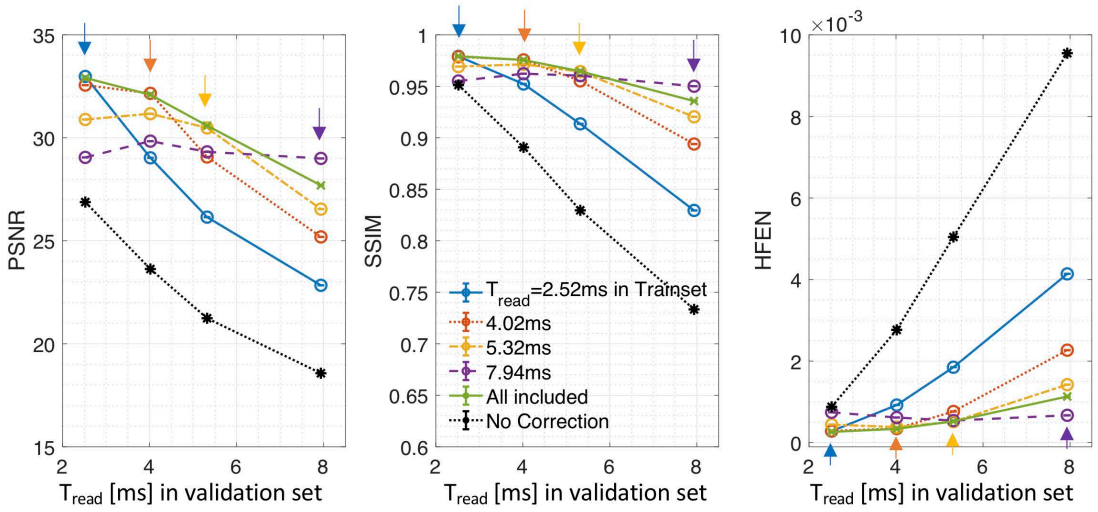

Figure 4.8: Impact of readout duration (EXP III). Image Quality Metrics (PSNR, SSIM, and HFEN) as a function of readout duration (2.52, 4.02, 5.32, and 7.94 ms) for the uncorrected (black) and corrected (non-black) images are shown. For corrected images, color represents a separate network trained with different readout durations; "All included" (green) indicates all of four readout durations are used during training. All metrics are averaged across time and subjects. Note that higher PSNR and SSIM and lower HFEN correspond to better performance. The best performance is almost always achieved when the training and validation datasets share the same readout duration as indicated by arrows in each panel of the image metrics. The performance then quickly degrades for longer readout durations in the validation set than that with which the network was trained.

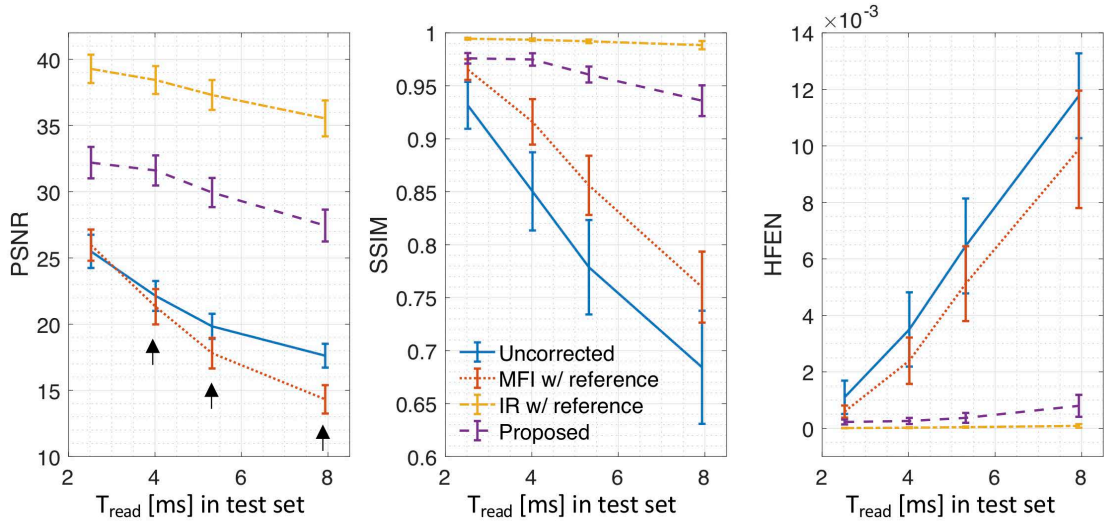

Figure 4.9: Quantitative comparison of deblurring performance for comparison methods using synthetic test data. Image Quality Metrics (PSNR, SSIM, and HFEN) are shown as a function of readout duration (2.52, 4.02, 5.32, and 7.94 ms). All metrics are averaged across time and subjects and error bars were calculated as the standard deviation. Note that higher PSNR and SSIM and lower HFEN correspond to better performance. The proposed method (purple) is compared against no correction (blue), multi-frequency interpolation (MFI, red), and iterative reconstruction (IR, orange). For all methods, performance gradually degrades as readout duration increases. IR performs best for all readout durations, followed by the proposed method, MFI, and no correction. Note that MFI had lower PSNR than that for no correction for readout duration $\geq$ 2.52 ms (black arrows).

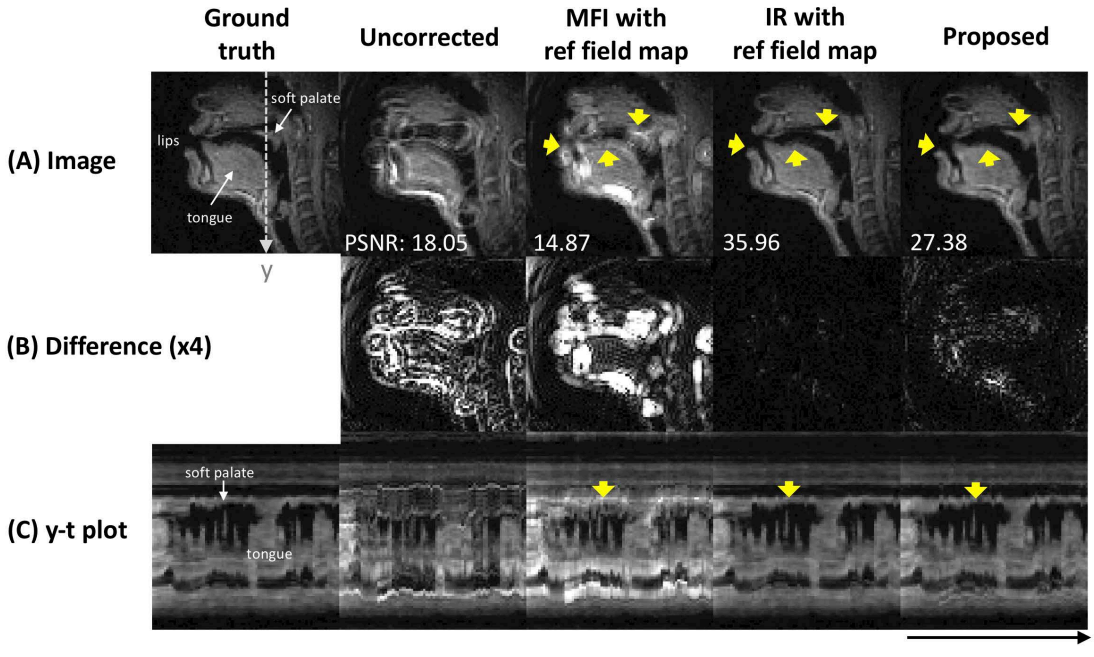

Figure 4.10: Qualitative comparison of deblurred images for comparison methods using synthetic test data. Left to right: the ground truth, uncorrected, multi-frequency interpolation (MFI), model-based iterative reconstruction (IR), and the proposed method. (A) images before and after deblurring with various methods, (B) absolute difference images (amplified by a factor of 4 for better visualization) with respect to the ground truth, and (C) an intensity vs time (y-t) plot marked by a dotted white line in (A). Yellow arrows point out the regions that are most affected by off-resonance blurring and that present contrast in deblurring performance for various methods. The proposed method successfully resolves the blurring artifact, which is superior to uncorrected image and image using MFI, and is visually comparable to IR method that presents the best performance over all others.

the result from the IR. Figure 4.10C shows the intensity time profiles that are extracted at the dotted line marked in the ground truth. Both IR and the proposed methods exhibit sharp boundaries between the tongue and air and around the soft palate, which is consistent over time frames.

4.3.6 Evaluation using Real Experimental Data

Figure 4.11 contains representative experimental data using different spiral readout durations. Image is reconstructed with no off-resonance correction and using the DORC [1]. The proposed method took the uncorrected images (left column) as an input to the trained network and performed deblurring frame-by-frame. In the uncorrected image, off-resonance blurring is most clearly observed at the lower lip (green arrows) and hard palate (red arrows) and becomes severe with the longer readouts. The proposed method can improve the delineation of boundaries in those locations. For example, the lower lip becomes sharper and the structure of the hard plate becomes visible after correction using the proposed method (see red arrows), which is consistent for all readout durations considered. DORC exhibits improved depiction of air-tissue boundaries for short readouts ($\leq$ 5.32 ms) but the signal intensity in several regions becomes spuriously amplified, and the blurred anatomic structures remain still unresolved for longer spiral readouts (7.94 ms) (see yellow arrows).

4.4 Discussion and Conclusion

We have demonstrated a machine learning method for correcting off-resonance artifacts in 2D spiral RT-MRI of human speech production without exam-specific field maps. We trained the CNN model using spatially varying off-resonance blur synthetically generated by using the discrete object approximation and field maps. Once the network is trained, the proposed method is computationally fast (12.3 $\pm$ 2.2 ms per-frame on a single GPU) and effective at resolving spatially varying blur that occurs most significantly at the vocal

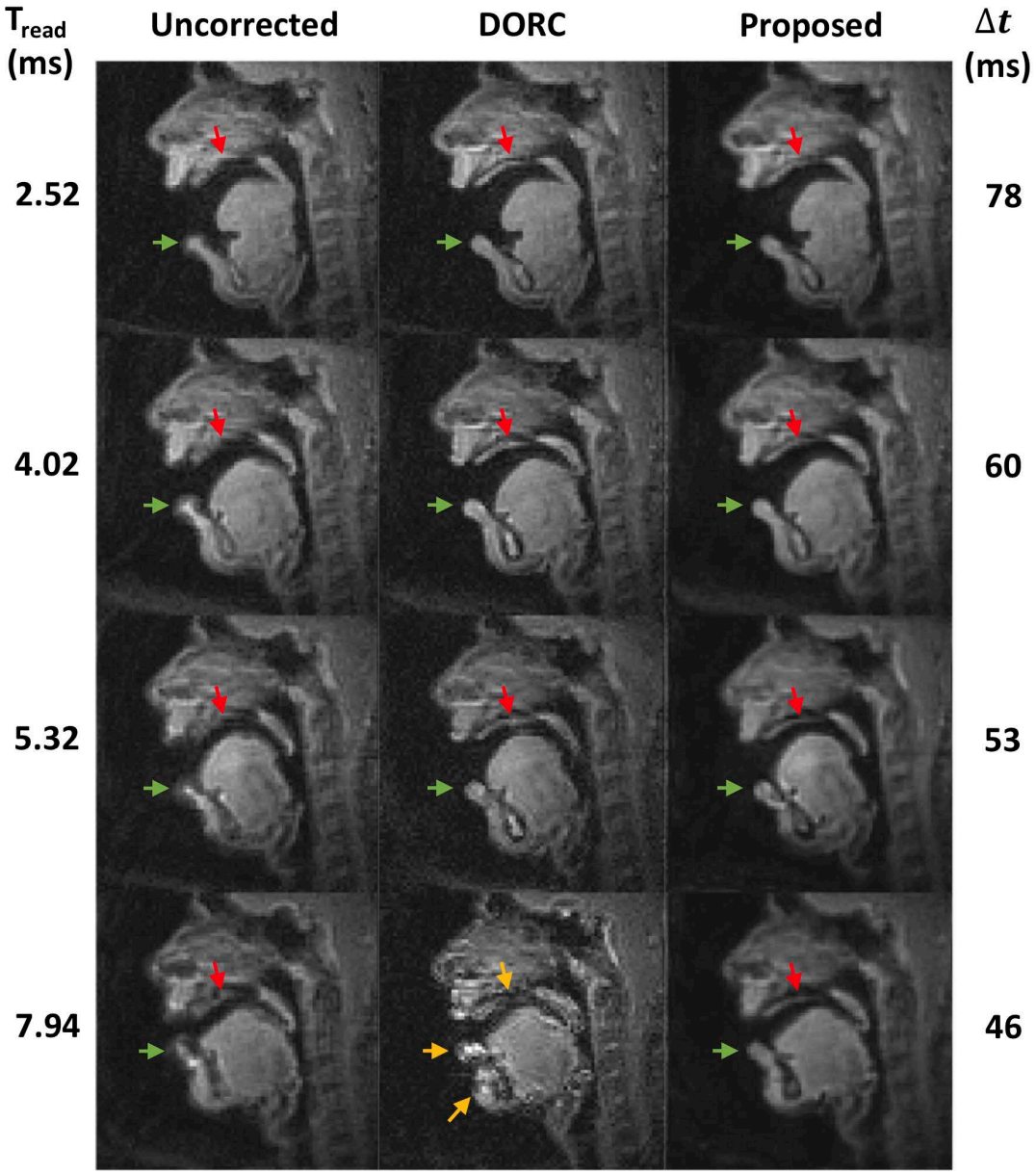

Figure 4.11: Representative experimental results using long readout spirals. Left to right: Image reconstruction with no off-resonance correction, image reconstruction using a previous auto-calibrated dynamic off-resonance correction (DORC), and image deblurred by the proposed method. Top to bottom: readout duration from 2.52 to 7.94 ms and temporal resolution of 78 (13-interleaf) to 46 ms (4-interleaf). Green (lower lip) and red (hard palate) arrows point out the regions that are most affected by off-resonance blurring and corrected by the DORC and proposed methods. The proposed method provides improved delineation of the boundaries, which is consistent for all readout durations considered, whereas DORC fails to resolve the blurred boundaries for a longer readout duration of 7.94 ms (yellow arrows). The video can be found in Supporting Information Video S1 at Wiley Online.

tract air-tissue boundaries of interest. The performance was superior to the current state-of-the-art auto-calibrated method and only slightly inferior to an ideal reconstruction with perfect knowledge of the field map.

We utilized a simple 3-layer residual CNN to learn the deblurring operation based on the training set of paired blurred and ground truth images. The network with the learned parameters is applied to different subjects with different speech patterns and spatio-temporal off-resonance patterns. The results indicate that frame-by-frame blurring is resolved in a matter that is far superior to the correction of the temporal average blur. The CNN performance is invariant to rotation/flipping, although some of the learned convolution kernels lack circular symmetry (Figure 4.6). Our interpretation is that the CNN estimates local deblurring from the features of the input image while allowing for adaptation to the changing off-resonances. Even in blurred images, the necessary information for deblurring remains local (cf., local imaginary components exploited by Noll et al [74]). We speculate that the convolutional filters are able to pick up information from surrounding pixels and use nonlinearities such as the ReLU operation to preserve only the filter outputs that are relevant to deblurring for each spatial location.

To train the proposed network architecture, we synthetically generated spatially varying off-resonance blur using reference data with field maps estimated from the auto-calibrated method [1] with affine linear data augmentation of the field maps ($\mathbf{f}' = \alpha\mathbf{f} + \beta$). This is different from the approach taken by Zeng et al [106] where spatially invariant off-resonance was simulated at a range of off-resonance frequencies. We experimentally found that our network architecture trained for spatially invariant blur would not be able to resolve the spatially varying blur, and the usage of spatially-varying field maps

is valuable (EXP II). This is consistent with one of the critical underlying assumptions pertaining to machine learning that unseen test data comes from the same distribution as training data. We also investigated generalizability for readout duration. We found that the readout duration of the training and test sets should be the same for the network to correct off-resonance without performance degradation (EXP III). This is consistent with expectation, because each readout (trajectory) presents a unique PSF for a given off-resonance frequency, which might not be generalized by the other readouts unless data from the other readouts are also present during the training phase. One exception would be spiral imaging with extremely short echo time.

We compared the proposed method with several conventional methods. For the synthetic test data, MFI exhibited the worst performance (see Fig. 4.9). This is likely due to the air-tissue boundary presenting abrupt spatially-varying off-resonance that would not fulfill the assumption of smooth spatial off-resonance variation. Model-based IR with knowledge of the true field map was superior to all others in terms of PSNR, SSIM, and HFEN. However, these two approaches are impractical as they require knowledge of the field map and are therefore limited by the quality of field map estimates. This is a practical limitation of conventional methods. One such case is shown in Figure 4.11, where auto-calibrated methods do not reliably work at longer readout duration (7.94 ms). This is because the field maps are estimated from the severely distorted images and this error propagates to the estimated field maps and the iterative off-resonance correction procedure. The proposed method avoids this issue and provides superior performance even at a longer spiral readout than the iterative approach.

We used a simple 3-layer CNN architecture, which is motivated by an analogy to traditional deblurring procedures. With the continually growing number of state-of-the-art network architectures, many of which are much deeper, we expect that there is further room for improvement. Nevertheless, there should be a balance between performance and the ability to explain such performance. The purpose of this study was to demonstrate the feasibility of spatially varying off-resonance correction using a simple CNN architecture, and this was achieved. We only examined a single contrast and a single region of the body from the midsagittal imaging plane. A larger study encompassing multiple body regions with different imaging parameters would be valuable in future work.

We considered the speech production application because off-resonance artifacts significantly hamper the detailed speech scientific and linguistic analyses using the dynamic imaging data. The proposed method has shown to provide sharp delineation of articulator boundary with readouts up to $\sim$ 8 ms at 1.5T, which is 3-fold longer than the current standard practice [35] and would provide 1.7-fold improvement in scan efficiency. This would allow for improved accuracy and precision of speech analysis beginning with boundary segmentation [63, 62, 120] that is often impaired by blurring artifact [1]. It would also potentially be feasible to achieve higher temporal resolution using a longer readout with image quality comparable to a short readout (see Fig. 4.11) or to use spiral readouts at higher field strengths such as 3T which is available on more sites and provides higher SNR. The low-latency processing of off-resonance deblurring (12.3 $\pm$ 2.2 ms per frame) without field map would also be valuable for other RT-MRI applications such as

cardiac studies and interventional RT-MRI where off-resonance at the lateral wall, adjacent to draining veins, or around metal implants and tools impedes diagnostic use of RT-MRI.

Chapter 5

3D RT-MRI of Speech Production

In the previous chapters, we have focused on discussing techniques in only 2D imaging. Generally, RT-MRI techniques have been limited to one mid-sagittal imaging plane or a few 2D imaging planes [21, 77, 52, 40, 41, 38, 67, 121, 122]. This has nevertheless provided good utility to speech scientists due to the fact that important information about "place of articulation," which is critical in linguistic contrasts, can be obtained from constriction details in the mid-sagittal plane (such as, for example, in phonemes /p/, /t/, and /k/ with constrictions are located at the lips, alveolar ridge, and velar region). However, vocal tract shaping during speech is enormously complex in geometry and in temporal structuring and cannot be fully understood from mid-sagittal constriction posture along the vocal tract [123]. For example, articulation of English fricative /s/ and lateral approximant /l/ both involve constriction of the tongue tip at the alveolar ridge, but the production of these sounds differ in that [s] has the tongue sides braced and air directed centrally along a groove, while [l] has (one or both) tongue sides lowered, allowing for lateral airflow channels [123]. Detailed and direct three-dimensional (3D) information about airway shape and spatiotemporal dynamics is essential to understanding speech

production control and to relating articulation to speech acoustics. However, in the past shaping imaging for speech has only been available indirectly from mid-sagittal 2D dynamic MRI after transformation to 3D or in static volume from 2D multi-planar imaging or in 3D for non-natural/sustained phonation [87, 124, 88, 89].

Recently, several research groups have demonstrated dynamic 3D MRI of the vocal tract [53, 125, 54]. Burdumy et al proposed an imaging method with 200×200×62 mm^3 spatial coverage using variable density and stack-of-stars radial sampling patterns [53], and measured dynamic modification of articulators during singing and speech tasks. With a temporal resolution of 1.3 s, this approach was restricted to relatively slow speech tasks. Fu et al proposed an imaging method using a combination of 3D cones sampling for a navigator acquisition and Cartesian sampling for image encoding [54]. The method achieved full vocal tract coverage with a high frame rate (166 fps) by employing a partially separable model (low-rank constraints) during reconstruction. This approach inherently requires long acquisition times, potentially resulting in several repetitions of speech tasks, and reconstruction performance may depend on a reliable estimation of temporal basis from the navigator [35]. These constraints may limit its application to natural speech tasks. A review of current state-of-the-art MRI protocols for speech production study can be found in Ref [35].

In this chapter, we address the unmet need for full vocal tract 3D dynamic MRI at high temporal resolution during natural speech, without requiring multiple repetitions of a speech task. We have developed a new technique that achieves 2.4×2.4×5.8 mm^3 spatial resolution and 61 ms temporal resolution over a 200×200×70 mm^3 field-of-view (FOV), using parallel imaging and simple spatiotemporal constraints previously validated in the

context of 2D dynamic MRI [126] (and used in > 50 cases [40, 49]). We extend a 2D spiral gradient sequence [40] to 3D by incorporating a slab excitation and adding phase-encoding along the k_z direction and use spatiotemporal finite difference (FD) constrained reconstruction with an empirically optimized penalty.

5.1 Methods

5.1.1 Data Sampling

The proposed method uses a pseudo-golden angle (GA) stack-of-spirals sampling pattern. Spiral trajectories balance trade-offs among temporal resolution, spatial resolution and signal-to-noise ratio, and have been shown to be robust in speech MRI acquisition [35, 21, 40, 58, 1]. Pseudo-GA increment has previously been used in the context of 2D spiral dynamic MRI [40, 58] and provides a nearly uniform sampling pattern that allows more reduced side-lobe energies of the point spread function and retrospective temporal resolution selection [58]. Most importantly, the pseudo-GA increment (compared to true GA) allows for high quality audio recording because the gradient waveforms and corresponding acoustic noise are periodic. The 2D spiral sequence can be converted to 3D stack-of-spirals sequence by adding phase-encoding lines along the kz direction. We leverage the pseudo-GA spiral sampling in the k_x-k_y plane.

Figure 5.1 illustrates the data sampling scheme. A pseudo-GA spiral sampling is used in the k_x-k_y plane and Cartesian sampling is employed along the kz direction. Each spiral is acquired for all kz phase encodes (linear order) before moving to the next spiral, with a GA increment, $\theta_{GA} = 2\pi \times 2/(\sqrt{}(5) + 1)$. The spiral angle is reset after N interleaves,

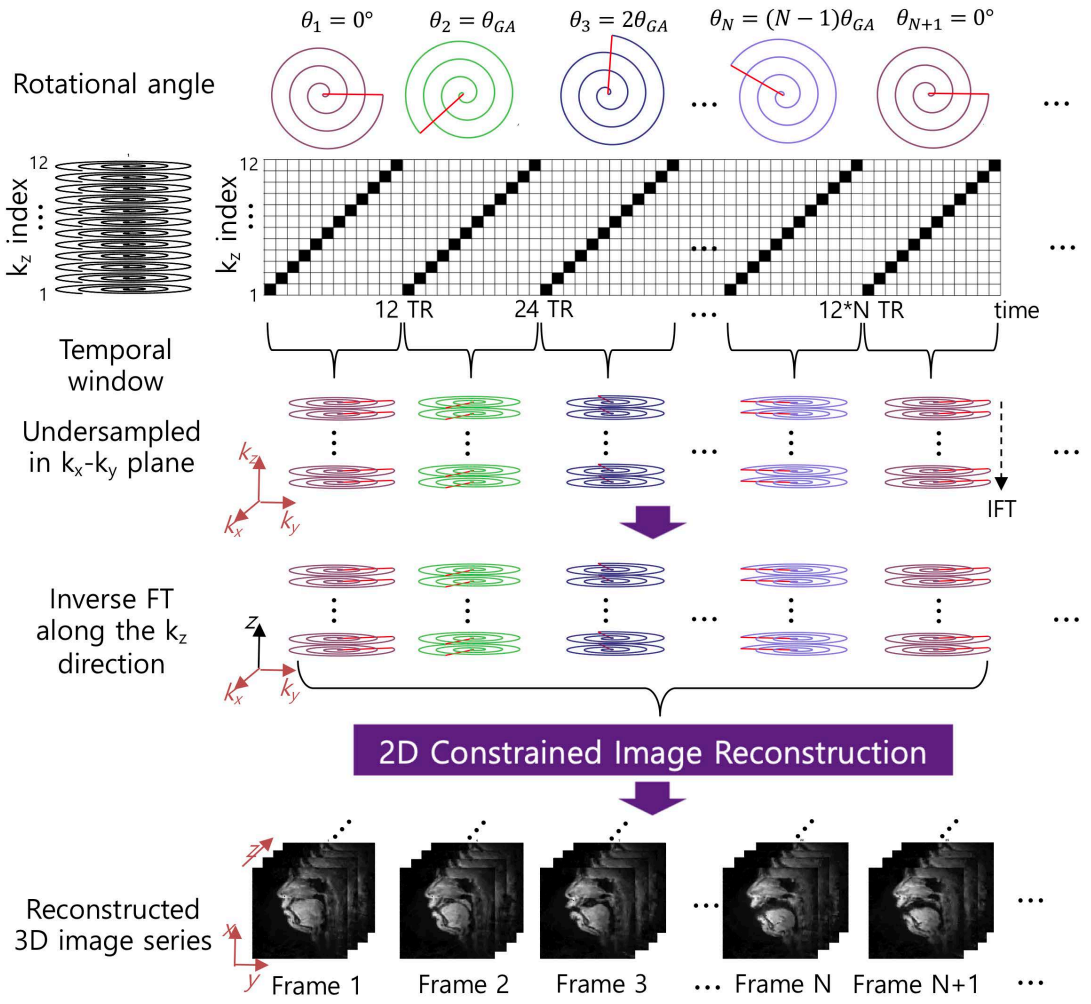

Figure 5.1: An example of a pseudo golden angle stack-of-spirals sampling scheme for 3D RT-MRI. Spiral interleave with a rotation angle is acquired for all k_z phase encodes while the k_z step is sequentially increased. After acquiring all of the kz steps, the rotation angle of spirals is increased by the golden angle, $\theta_{GA} = 2\pi \times 2/(\sqrt{(5)}+1)$. The spiral angle is reset after N interleaves. Inverse Fourier transform is applied to the data collected within a temporal window along the (fully sampled) k_z direction. Then 2D constrained reconstruction is performed slice-by-slice to form a 3D image series.

e.g., after 12·N TRs with 12 phase encoding lines as illustrated in Figure 5.1, where N is a periodicity of the pseudo-GA [40]. We use N=34 in this work.

5.1.2 Image Reconstruction

3D reconstruction is performed slice-by-slice, after inverse Fourier transforming data collected within a temporal window (12 TRs) along the fully sampled k_z direction, as illustrated in Figure 5.1. Note that it is also possible to perform a full 3D reconstruction instead, but given a very large dataset (e.g., 630 samples $\times$ 800 spirals $\times$ 8 channels $\times$ 12 k_z), decoupling the reconstruction into 2D problems is more computationally efficient and practical. We employ a sparse SENSE-based parallel imaging and compressed sensing approach with spatiotemporal first-order FD constraints [126]. Regularization parameters for spatial and temporal sparsity (λ_s and λ_t, respectively) are empirically chosen by visual assessment and once calibrated, are held constant for all studies. Coil sensitivity maps are assumed to be time-invariant and are estimated from time-averaged 3D data from each coil by using ESPIRiT [127]. We perform the reconstruction using the Berkeley Advanced Reconstruction Toolbox [128].

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

5.1.3 3D Dynamic MRI Acquisition

3D slab excitation is achieved by using a minimum phase RF pulse designed with the Shinnar-LeRoux RF design tool software package [129]. The pulse excites a mid-sagittal slab with 5 cm thickness using a flip angle of 5°and a time-bandwidth product of 16, and stop-band and pass-band ripples of 0.5 % and 1 %, respectively. The benefit of using the

Table 5.1: Acquisition parameters for 2D multislice and 3D dynamic MRI protocols

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

minimum phase pulse is that it can provide a sharp slice profile (higher time-bandwidth product) for a given specification and allows for shorter TE because it has an asymmetric pulse shape and requires a short refocusing gradient. 3D data acquisition was performed using a stack-of-spirals spoiled gradient echo readout with imaging parameters presented in Table 5.1.

5.1.4 2D Multislice Dynamic MRI Acquisition

For comparison, we also perform 2D pseudo-GA dynamic MRI with two or three interleaved slices — one mid-sagittal and one or two oblique slices — relevant to the speech task [40], using a previously published approach [52]. The GA increment for the two-

and three-slice sequence occurred every 2 and 3 TRs, respectively. The periodicity of the pseudo-GA (34 interleaves) was the same as in the 3D sequence. Imaging parameters used are listed in Table 5.1. We reconstruct the dynamic image slice-by-slice by using the sparse SENSE-based reconstruction described in "Image Reconstruction" section with 1 spiral interleave per frame with a reduction factor of 13, which corresponds to the temporal resolution of 12 ms per frame and 18 ms per frame for two- and three-slice, respectively.

5.1.5 In-Vivo Speech Experiments

All experiments were performed on a commercial 1.5 T scanner (Signa Excite, GE Healthcare, Waukesha, WI) using a real-time interactive imaging platform (RT-Hawk, Heart Vista Inc, Los Altos, CA) [50] with a gradient strength of 40 $\text{mT}\cdot\text{m}^{-1}$ and a maximum slew rate of 150 $\text{mT}\cdot\text{m}^{-1}\cdot\text{ms}^{-1}$. A body coil was used for RF transmission, and a custom eight-channel upper airway coil [40] was used for signal reception. The imaging protocol was approved by our Institutional Review Board. Two healthy adult volunteers were scanned, after providing written informed consent.

For Speaker 1, audio was recorded inside the scanner simultaneously with data acquisition using a commercial fiber optic microphone (Optoacoustics Ltd., Yehuda, Israel) and a custom recording setup [47]. The recorded speech was then enhanced using a dictionary learning-based acoustic denoising method [48] and was synchronized with the reconstructed dynamic images to aid linguistic analysis.

All the stimuli were read in the scanner using a mirror projector setup for presentation [49]. Speaker 1 (female American English speaker) was scanned with both the 3D and

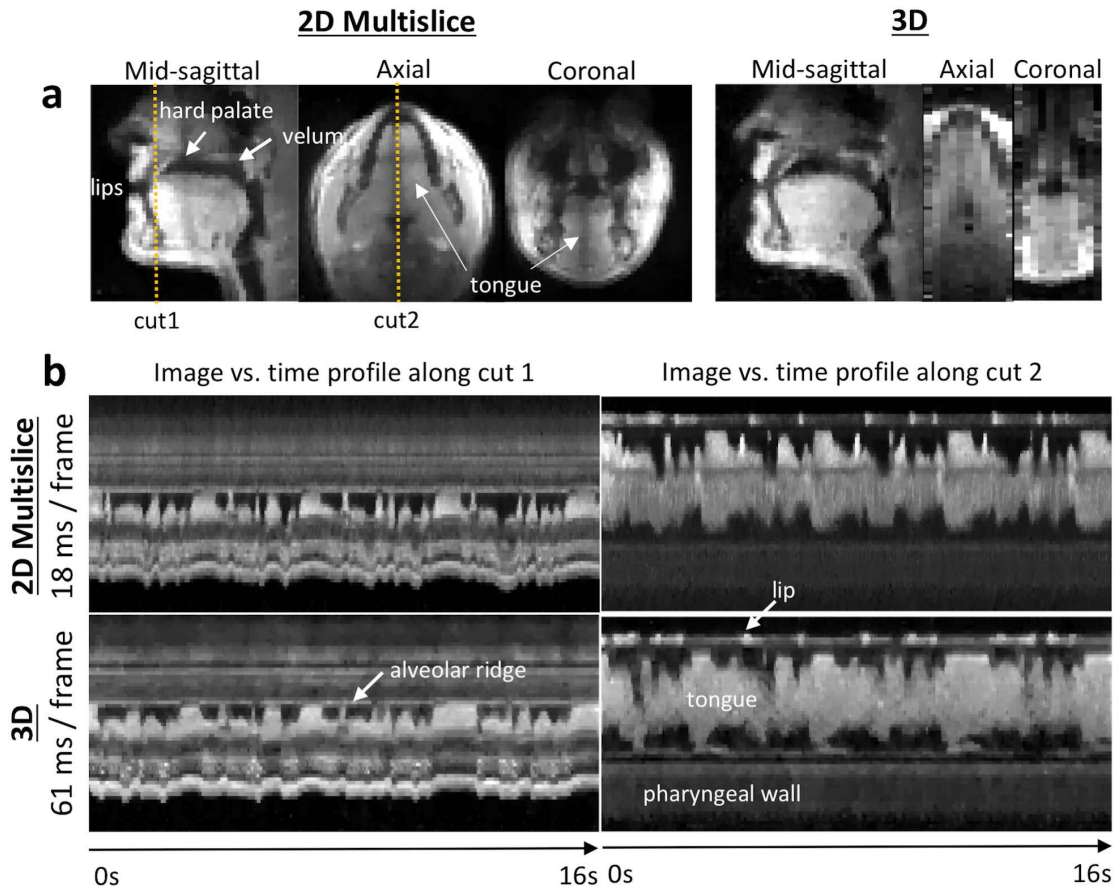

Figure 5.2: Reconstructed images from both 2D multislice and 3D RT-MRI for Speaker 1. (a) Three orthogonal planes (from the left: mid-sagittal, axial, and coronal slices) at consonant /s/ from 2D multislice and 3D RT-MRI. For comparison purpose, three slices are extracted from 3D that would be aligned with those obtained from 2D multislice RT-MRI. See Table 5.1 for acquisition parameters for both the protocols. (b) Illustration of the tongue movements for speech tasks DU 2-5 listed in Table 5.2. Two intensity versus time profiles corresponding to the cuts marked by the dot lines in the images in (a) are shown. Both use the same regularization parameters ($\lambda_t = 0.02$ and $\lambda_s = 0.01$).

Table 5.2: The stimuli for speaker 1.

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

Abbreviations: UD and DU, directions of movements; UD, sides up (groove) to sides down (lateral); DU, the reverse of UD; ., a syllable and/or word boundary; V, a stressed vowel; C, a consonant.
[a]DU1 (the word-initial cluster [ls]) does not exist in English.

2D three-slice sequences with plane locations as shown in Figure 5.2A. The stimuli for Speaker 1 are listed in Table 5.2 and were each spoken twice at a natural speech rate. These stimuli deployed two sounds [s] and [l] with contrasting lingual actions: [s] involves tongue sides up and braced and the tongue surface grooved for central airflow, while [l] involves tongue sides low, allowing lateral airflow. The stimuli placed [s] and [l] temporally "closer together" or "farther apart" in both orders — i.e. [s] preceding [l] and [l] preceding [s] — creating a direction of lingual action of the tongue sides going from up to down or down to up, respectively. Speaker 2 (male native Korean speaker producing English as a second language) was scanned with both the 3D and 2D two-slice (one mid-sagittal and one axial plane at the level of the mid-pharyngeal airway) sequences. This speaker read the English stimuli: "/loo/-/lee/-/la/-/za/-/na/-/za/" repeated twice at a natural rate to produce alternating consonant and vowel sounds. These consonant-vowel syllables utilize consonants ([l], [z], [n]) made with the tongue tip and relatively extreme vowel postures ("ee" [i], "ah" [a], "oo" [u]) made respectively with the tongue body high & front, low & back, and high & back.

5.1.6 Data Analysis

VOI Analysis for Identifying Tongue Actions for [l] and [s] Actions of the tongue tip, sides, and rear (dorsum) are critical in the production of [s] and/or [l], so form the basis of our derived data analysis. In analogy with established region-of-interest analyses [59, 60, 86], volumes-of-interests (VOIs) were designated around three vocal tract locations — the tongue tip (TT), dorsum (TD), and tongue sides (TS) — by manually drawing 2D regions-of-interest in the mid-sagittal and axial image planes and extending

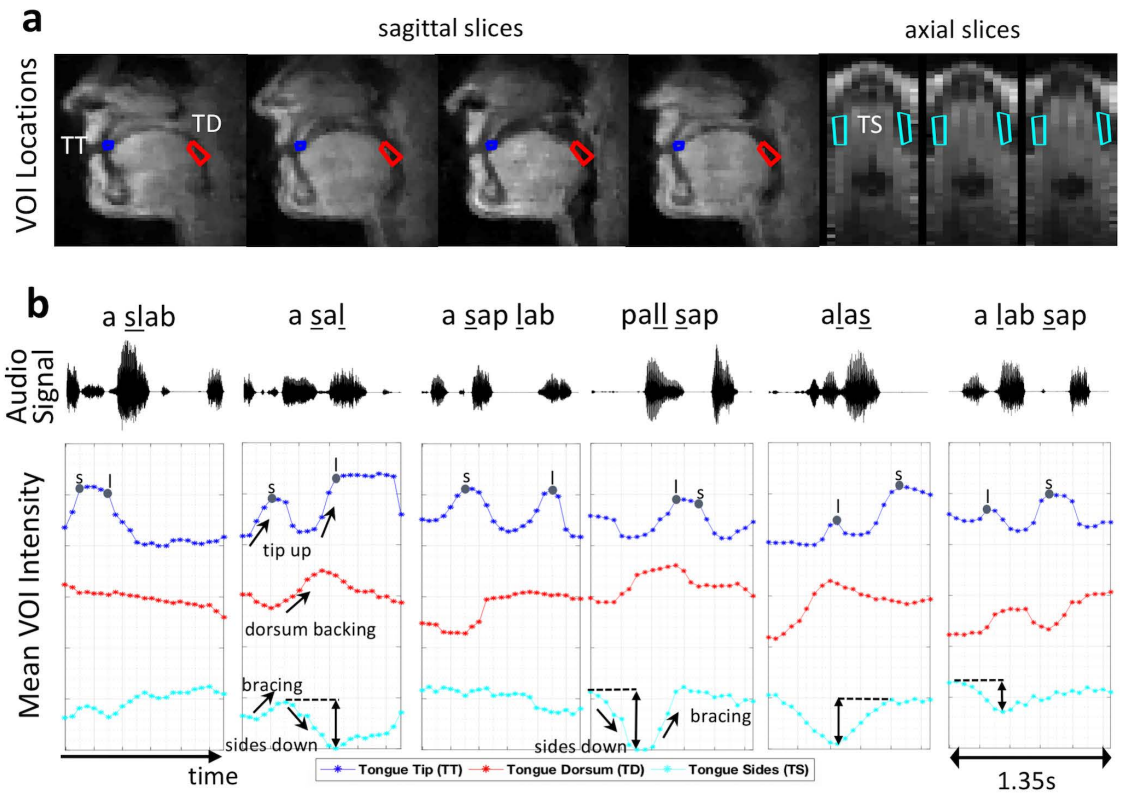

Figure 5.3: VOI analysis for identifying tongue action for [l] and [s]. (a) Placement of VOIs at the tongue tip (blue), back (red), and sides (cyan) overlaid on sagittal and axial images. Illustration of (b) the synchronized denoised audio signals and (c) mean intensity for three VOI locations over time for different stimuli. Mean intensity over time was calculated within each of the VOIs shown in (a). Each time window corresponds to 1.35s with a temporal resolution of 61 ms.

those regions to adjacent parallel image planes as shown in Figure 5.3A. Mean pixel intensity was calculated within each VOI over time. Lingual tissue moving into and out of these VOIs allows the identification of three critical lingual gestures for these sounds: a tongue tip raising gesture, a tongue dorsum backing gesture, and a tongue lateral lowering or dipping gesture. Specifically, the actions of these articulatory gestures are expected to reflect:

1. Between /l/ and /s/, the temporal lag or offset between the two segmental articulations, which should accord with the phonological "temporal distance" between the target consonants, as organized in Table 5.2.

2. Within /l/, the relative coordination of the lingual gestures within the articulation of [l]. In particular, this should accord with prior data from other techniques regarding the internal temporal organization of tongue tip and dorsum gestures for [l] [130, 131].

Measurement of Vocal Tract Area Function The vocal tract area function is defined as the cross-sectional area of the airway as a function of distance from the glottis and is an important measurement in the study of the relation between vocal tract shaping and acoustics. We tested the ability of 3D dynamic MRI to estimate the dynamics of vocal tract area function (using Speaker 2's data). From the mid-sagittal plane, we obtained grid lines that were perpendicular to the airway centerline obtained from an airway boundary segmentation method [62] and extracted angled slices along the grid lines through the 3D volume (61 slices with 2 mm increments). From each of the angled slices, we estimated the airway area [cm^2] encompassed by articulator boundaries from a region growing method [89], applied in this case to the dynamic data. Region growing was performed for each of the angled slices at every time frame independently with seed points automatically chosen as the intersection of the airway centerline from the mid-sagittal plane, and the angled slices. Note that the teeth are not visible in this imaging modality and thus are not reflected in the area function. The resulting error is temporally constant and appears only at the mouth termination region. A subject-specific dental

correction could be performed during post-processing, using additional data that captures the geometry of the teeth [123, 132].

5.2 Results

Figure 5.2 shows representative reconstruction results from 2D multislice and 3D methods for Speaker 1's utterances DU 2-5 (see Table 5.2). The tongue shape at onset of /s/ in the syllable "sap" is shown in three different views in Figure 5.2A; constriction of the tongue tip and grooving of the medial tongue surface are clearly observed in the sagittal and coronal slices, respectively, from both results. Figure 5.2B compares temporal tongue tip dynamics from the 3D result with that from the 2D multislice. The 3D result shares a similar temporal pattern with the tongue tip motion with the 2D multislice result, although it exhibits a slight temporal blurring around the tongue tip compared with its 2D counterpart. Overall, the 3D result provides adequate quality to discern tongue tip actions for the articulation of these consonants in this natural speech task.

Figure 5.3C shows mean pixel intensity curves calculated from three VOIs (TT, TD, and TS) over time for stimuli UD 1, 3, 5, and DU 2, 3, 5. The temporal positions of /s/ and /l/ are measured at their TT mean intensity peaks (i.e., the maximum constriction) as annotated on the time functions in panel C. It is clearly apparent that the articulation /s/ and /l/ are temporally close in "a slab" and "pall sap" and become farther away from each other as other vowel and consonant segments intervene between the two target consonants. This pattern is consistent with the phonological "temporal" organization of the stimuli as listed in Table 5.2.

For /s/ the tongue tip raising motion is the sole critical articulation apparent, whereas for /l/ co-articulation of tongue tip raising, dorsum backing (higher signal in TD), and sides lowering (lower signal in TS) is observed. Interestingly, depending on the position of /l/ in the syllable, distinct spatiotemporal characteristics are observed for the gestures of /l/. In a syllable-final /l/ (e.g., "a sal" and "pall sap"), the tongue dorsum backing is extended for a longer period of time and is more spatially extreme than in a syllable-initial /l/ (e.g., "a sap lab" and "a lab sap"). Similarly, the tongue sides are lowered more in a syllable-final /l/ than in a syllable-initial /l/ as indicated with up-down arrows in Figure 5.3C. The word-internal, intervocalic ambisyllabic /l/ in "alas" shows an intermediate behavior in this regard. In terms of [l]'s internal gestural coordination, its three lingual gestures begin almost simultaneously in syllable-initial position, whereas in syllable-final position the tongue dorsum backing and tongue sides lowering start earlier than the tongue tip raising gesture, leading to a timing lag. This syllable-driven coordination asymmetry has previously been observed for tongue tip-dorsum coordination using point tracking kinematic data on [l] [130, 131]; the proposed protocol not only replicates this finding but also provides new quantitative evidence of a parallel coordination asymmetry involving the tongue sides.

Figure 5.4 shows a direct comparison of tongue shape for /l/ versus /s/ in the phrase "pall sap" for Speaker 1. For both segments, constriction of the tongue tip at the alveolar ridge can be observed in the mid-sagittal images. For /l/, side channels are visible in the axial and coronal slices, as well as tongue body retraction in the coronal slices, all of which funnel air laterally along the tongue sides, whereas for /s/, the tongue is grooved

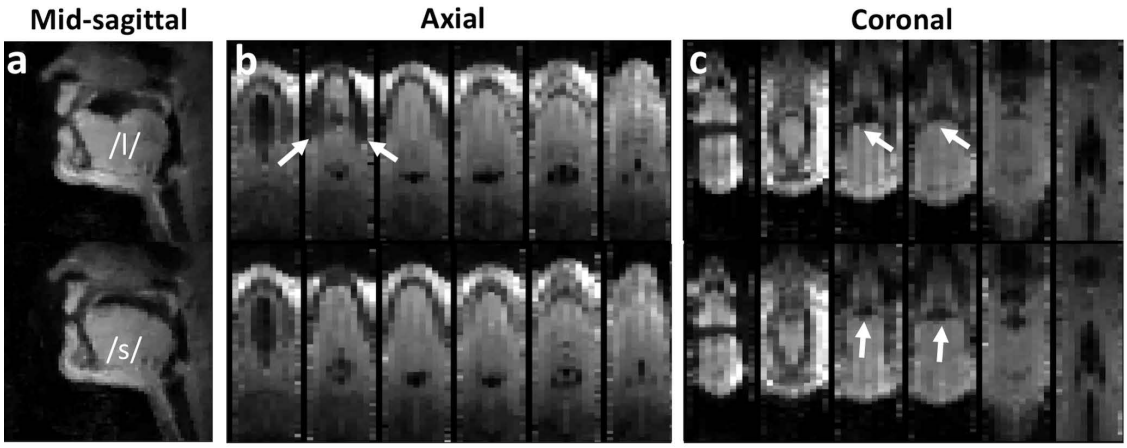

Figure 5.4: Comparison of the vocal tract shape between /l/ and /s/ in the context of "pall sap" for Speaker 1. (a) Mid-sagittal, (b) axial, and (c) coronal views. For both (top) /l/ and (bottom) /s/, mid-sagittal images show constriction of the tongue tip at the alveolar ridge. For /l/, side channels are shown in the axial and coronal slices, as well as the retraction of the tongue rear in the coronal slices; whereas for /s/, grooving of the tongue is shown in the coronal slices. The videos can be found in Supporting Information Video S2 and S3 at Wiley Online.

mid-sagittally as shown in the coronal slices, channeling the airstream anteriorly toward the front teeth.

Figure 5.5 shows vocal tract area function dynamics for the utterances of Speaker 2. Critical constriction events are visible along the length of the vocal tract. Specifically, when consonants /l/, /z/, and /n/ are articulated (e.g., frames 12, 27, 39, 52, 65, 79 shown in Supporting Information Video S4), the relatively rapid tongue tip constrictions used to create these consonants are clearly shown in the area function dynamics (grid line 3). And, when the vowel /ee/ is articulated (frames 31-34 and 117-122), vocalic tongue body constrictions are observable in the palatal region (grid lines 4-7), as is pharyngeal volume expansion (grid line 13-15) associated with /ee/'s tongue body fronting.

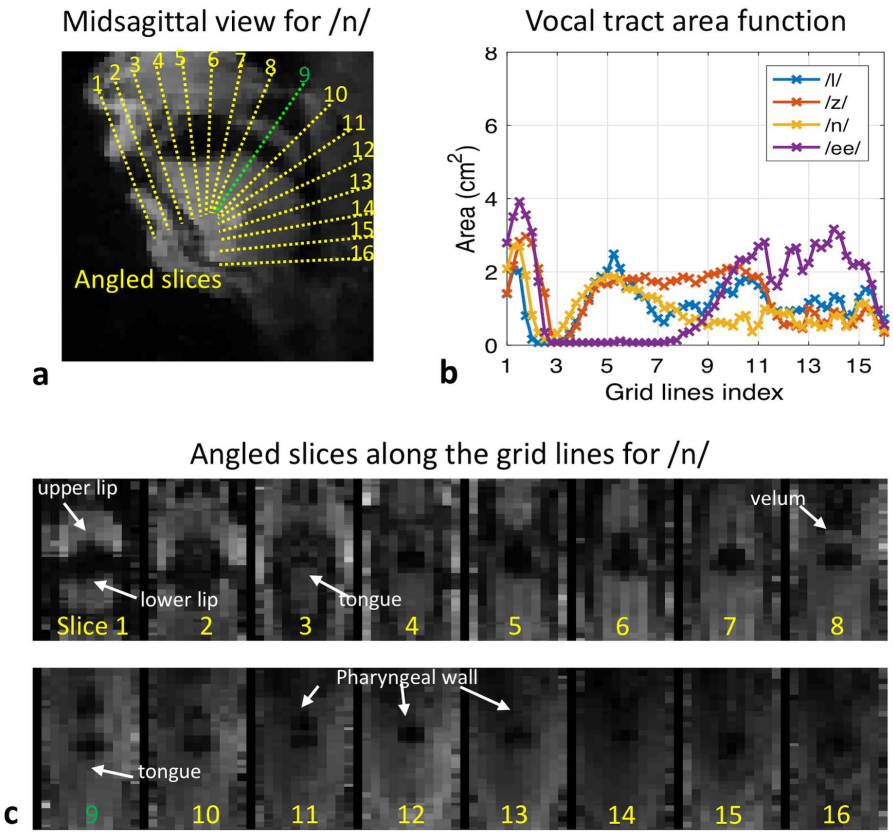

Figure 5.5: Illustration of the capability of estimation of vocal tract area function from 3D RT-MRI for the "na" utterance of Speaker 2. Panel (a) shows an image at the midsagittal plane for /n/ in "na" from dynamic 3D. Grid lines that are perpendicular to the airway centerline are chosen to obtain angled slices shown in Panel (c) (only 16 of the 61 gridlines are shown here). Panel (b) shows the vocal tract area functions for /l/, /z/, /n/, and /ee/ estimated from the 61 angled slices. The videos can be found in Supporting Information Video S4 at Wiley Online.

5.3 Dicussion

We have demonstrated a dynamic 3D imaging technique that provides complete spatial coverage of the human vocal tract, with spatiotemporal resolution adequate to visualize lingual tongue movements occurring during natural speech without the need for task repetition and with results comparable to interleaved multislice 2D dynamic MRI. Based on data obtained using this proposed technique, we developed a VOI analysis to characterize

the coordination of tongue gestures for consonants /l/ and /s/. Earlier point-tracking techniques have established that coordination of the tongue tip and dorsum gestures for American English /l/ varies as a function of syllable position [130, 131, 133]. To our knowledge, the work presented here provides for the first time quantitative imaging data on the magnitude and duration of tongue side movement and on its relative timing variation with respect to the other lingual gestures comprising /l/. Additionally, this technique has allowed us to quantify dynamic vocal tract area functions during natural productions of consonant-vowel syllables having varied consonants articulated with the tongue tip and vowels with varied tongue postures. These area functions show a conservation relation between the changes in area function at different parts of the vocal tract, which is expected to be the case [134].

Validation of our proposed technique is challenging because vocal tract shaping during speech, unlike cardiac or respiratory motion, is not cyclic, and intra-speaker variability makes it difficult to compare the results between methods in a reproducible way (although see Ref [54]). Even after acoustic alignment, the quality of retrospective CINE 3D MRI of the vocal tract is poor [135]. In this work, we use the multislice 2D dynamic MRI as an image quality reference because it provides the current best data quality for natural speech in our experience. However, this 2D method lacks information beyond the acquired (usually mid-sagittal) slices, and this method is applied during a separate production of the speech task. Further validation may be possible with numerical 3D vocal tract phantoms that allow realistic simulation of fluent speech, repeated speech utterances and flexibility of varying speech rate, or with simultaneous acquisition of MRI with another modality such as optical endoscopy.

This work should be considered an initial demonstration of feasibility. The parameters chosen, specifically, the spatial and temporal resolutions may not be optimal for all speech tasks or regions of interest. Higher spatial resolution in the slice (left-right) direction may be needed to precisely measure the vocal tract area function or to precisely identify the borders of smaller articulators such as the laryngeal structures (e.g., the arytenoid cartilages). Higher temporal resolution may be desirable for the study of rapid speech tasks such as alveolar trills, whose rate is about 30 Hz and duration is shorter than 100 ms [35].

In addition, the RF pulse used for 3D slab excitation may be further improved. The slab thickness was designed to be 2 cm thinner than FOV along the slice direction. This margin along with a high TBW allows the avoidance of aliasing in the slice direction due to a transition in the slab profile and/or shifts by resonant offsets that can be up to $\pm$ 625 Hz at 1.5 T at air-tissue interfaces. It is possible that the margin can be reduced. Likewise, there may be room to reduce the TBW and/or employ variable-rate selective excitation pulse [136], which would allow for shorter pulse duration and shorter TR.

Speech production experiments require that the scan operator be able to monitor the articulatory movements to identify when there might be a substantial unexpected change in head positioning, and to identify when speech utterances have been performed correctly per instructions. In the proposed method, these requirements are fulfilled by lower-quality zero-filled linear reconstructions with mediocre temporal resolution (303 ms / frame) and low reconstruction latency ($<$ 10 ms / frame), which were not shown here. Detailed linguistic analysis and computational modeling of speech MRI is almost always performed off-line [36], permitting a high-quality and high-latency reconstruction prior

to processing. The constrained reconstruction temporal window is the only fundamental limit on latency for the proposed method. We found that adequate image quality can be achieved with a temporal window of $\geq$ 16 frames (976 ms) (see Supporting Information Video S1). This indicates that the ultimate minimum latency of the proposed method is approximately 1 second. In the future, this could make it possible to perform real-time analysis with an overall latency of a few seconds.

5.4 Conclusion

We demonstrated a technique for 3D dynamic imaging of the full vocal tract at high temporal resolution during natural speech. The proposed method uses a minimum-phase 3D slab excitation, pseudo GA stack-of-spirals, and spatiotemporal finite difference constrained reconstruction and achieves $2.4\times2.4\times5.8$ mm^3 spatial resolution and 61 ms temporal resolution over a $200\times200\times70$ mm^3 FOV. This technique is evaluated through in-vivo imaging of natural speech production from two subjects with synchronized audio and via comparison with interleaved multislice 2D dynamic MRI. This promising tool for speech science for the first time enables a quantitative identification of spatial and temporal coordination of important tongue gestures coproduced on and off the midline in the articulation of consonants /l/ and /s/ via VOI analysis and allows a direct assessment of vocal tract area function dynamics during natural speaking of utterances.

Chapter 6

Conclusion and Future work

RT-MRI has recently gained substantial attention for speech production research because of its unique advantage of monitoring the complete vocal tract dynamics during the speech, safely and non-invasively at relatively high spatial and temporal resolution. This has been made possible by tremendous technical efforts that push the limit of spatiotemporal resolution forward. However, the current state-of-the-art RT-MRI has several limitations and unmet challenges for speech production application, which often render imaging's operating point below application demands and introduce bias as well as increased variance during data analysis. This dissertation addresses two specific unmet needs for RT-MRI of speech production – 1) off-resonance deblurring and 2) 3D RT-MRI, and presents new tools that improve the quality and quantity of imaging information about the dynamics of articulators, providing steps toward a better understanding of human speech production.

The specific contributions of this dissertation in the field of RT-MRI of speech production can be summarized as follows.

- The development of fully automated off-resonance deblurring for spiral RT-MRI. The dynamic field map is estimated directly from the base image phase from single-echo data with no cost in scan time. Model-based reconstruction is able to correct off-resonance in a previously acquired large corpus of single-echo spiral data. This method improves the depiction of the vocal tract articulators at several air-tissue boundaries both visually and through a sharpness metric, and provides the practical utility on the boundary segmentation and distance metric.

- The conception, design, and implementation of CNN-based off-resonance deblurring for spiral RT-MRI. The network structure is mathematically related to classical conjugate phase reconstruction and a minimal network is designed to achieve the deblurring task. A model-based framework along with a data augmentation scheme is proposed to generate training data. The present method is efficient, effective, and superior to the current state-of-the-art method and only slightly inferior to an ideal reconstruction with perfect knowledge of the field map.

- The development of 3D RT-MRI, which achieves $2.4\times2.4\times5.8$ mm^3 spatial resolution and 61 ms temporal resolution over a $200\times200\times70$ mm^3 FOV. This technique for the first time enables a quantitative identification of spatial and temporal coordination of important tongue gestures coproduced on and off the midline in the articulation of consonants /l/ and /s/ via VOI analysis and allows a direct assessment of vocal tract area function dynamics during natural speaking of utterances.

6.1 Future Work

The following list summarizes specific possibilities for future work in each of the above chapters.

Model-Based Deblurring

- Use an additional static two-echo scan to estimate the coil sensitivity map. The coil sensitivity estimation is a critical step for dynamic field map estimation and the two-echo scan could be conducted at a static posture before dynamic imaging scan and will provide the coil sensitivity maps that are free of phase due to off-resonance and B_0 field inhomogeneity.

- Explore the joint estimation of both dynamic image and field map [137]. Deblurring performance can further be improved by jointly estimating both the image and field map in the model-based reconstruction. In this scheme, the initial guess for the field map estimate would be critical to reaching an optimal solution due to the non-convexity in the joint optimization. The field map estimation proposed in Chapter 3 could be used as an initial input to the joint estimation. This approach would be particularly beneficial when the longer spiral readout is used.

Data-Driven Deblurring

- Explore other non-Cartesian trajectories such as radial or echo-plannar imaging.

- Explore other clinical applications that experience off-resonance artifacts, such as brain imaging near sinuses or imaging near the metal implant.

3D RT-MRI

- Explore more advanced and efficient k-t sampling schemes to improve temporal or spatial resolution and/or enlarge slice coverage. Those include rotated golden-angle scheme, variable density in-plane spiral sampling, variable density sampling along the k_z-t direction, partial Fourier scheme along the k_z direction, and so on.

- Explore other clinical applications that undergo rapid and/or irregular motion but where capturing 3D dynamics is critical, such as cardiac, musculoskeletal, and fetal imaging.

Bibliography

[1] Lim Y, Lingala SG, Narayanan SS, and Nayak KS. Dynamic off-resonance correction for spiral real-time mri of speech. *Magn Reson Med*, 81(1):234–246, 2019.

[2] Lim Y, Bliesener Y, Narayanan SS, and Nayak KS. Deblurring for spiral real-time mri using convolutional neural networks. *Magn Reson Med*, https://doi.org/10.1002/mrm.28393, 2020.

[3] Lim Y, Zhu Y, Lingala SG, Byrd D, Narayanan SS, and Nayak KS. 3d dynamic mri of the vocal tract during natural speech. *Magn Reson Med*, 81(3):1511–1520, 2019.

[4] Hargreaves BA. Spin manipulation methods for efficient magnetic resonance imaging. *PhD thesis, Stanford University*, 2001.

[5] Schenck JF. The role of magnetic susceptibility in magnetic resonance imaging: Mri magnetic compatibility of the first and second kinds. *Med phys*, 23:815–50, 1996.

[6] Xiang QS and Henkelman RM. k-space description for mr imaging of dynamic objects. *Magn Reson Med*, 29:422–428, 1993.

[7] van Vaals JJ, Brummer ME, Dixon WT, and et al. "keyhole" method for accelerating imaging of contrast agent uptake. *J Magn Reson Imaging*, 3:671–675, 1993.

[8] Jones RA, Haraldseth O, Muller TB, Rinck PA, and Oksendal AN. K-space substitution: a novel dynamic imaging technique. *Magn Reson Med*, 29:830–834, 1993.

[9] Madore B, Glover GH, and Pelc NJ. Unaliasing by fourier-encoding the overlaps using the temporal dimension (unfold), applied to cardiac imaging and fmri. *Magn Reson Med*, 42:813–828, 1999.

[10] Tsao J, Boesiger P, and Pruessmann KP. k-t blast and k-t sense: dynamic mri with high frame rate exploiting spatiotemporal correlations. *Magn Reson Med*, 50:1031–1042, 2003.

[11] Jung H, Sung K, Nayak KS, Kim EY, and Ye JC. k-t focuss: a general compressed sensing framework for high resolution dynamic mri. *Magn Reson Med*, 61:103–116, 2009.

[12] Otazo R, Axel L Kim D, and Sodickson DK. Combination of compressed sensing and parallel imaging for highly accelerated first-pass cardiac perfusion mri. *Magn Reson Med*, 64:767–776, 2010.

[13] Tsao J and Kozerke S. Mri temporal acceleration techniques. *J Magn Reson Imaging*, 36:543–560, 2012.

[14] O'Sullivan J. A fast sinc function gridding algorithm for fourier inversion in computer tomography. *IEEE Trans Med Imaging*, 4(4):200–207, 1985.

[15] Jackson JI, Meyer CH, Nishimura DG, and Macovski A. Selection of a convolution function for fourier inversion using gridding: computerised tomography application. *IEEE Trans Med Imaging*, 10(3):473–478, 1991.

[16] Pipe JP and Menon P. Sampling density compensation in mri: rationale and an iterative numerical solution. *Magn Reson Med*, 41(1):179–186, 1999.

[17] Liu QH and Nguyen N. An accurate algorithm for nonuniform fast fourier transforms (nufft). *IEEE Microwave Guided Wave Lett*, 8:18–20, 1998.

[18] Fessler J and Sutton B. Nonuniform fast fourier transforms using minmax interpolation. *IEEE Trans Sig Process*, 51(2):560–574, 2003.

[19] Pruessmann KP, Weiger M, Börnert P, and Boesiger P. Advances in sensitivity encoding with arbitrary k-space trajectories. *Magn Reson Med*, 46(4):638–51, 2001.

[20] Kerr AB, Pauly JM, Hu BS, Li KC, Hardy CJ, Meyer CH, Macovski A, and Nishimura DG. Real-time interactive mri on a conventional scanner. *Magn Reson Med*, 38:355–367, 1997.

[21] Narayanan SS, Nayak KS, Lee S, Sethy A, and Byrd D. An approach to real-time magnetic resonance imaging for speech production. *J Acoust Soc Am*, 115:1771–1776, 2004.

[22] Slavin GS and Bluemke DA. Spatial and temporal resolution in cardiovascular mr imaging: review and recommendations. *Radiology*, 234:330–338, 2005.

[23] Lustig M, Donoho DL, Santos JM, and Pauly JM. Compressed sensing mri. *IEEE Signal Process Mag*, 25(2):72–82, 2008.

[24] Liang Z-P, Boada F, Constable R, Haacke E, Lauterbur P, and Smith M. Constrained reconstruction methods in mr imaging. *Rev Magn Reson Med*, 4:67–185, 1992.

[25] Gick B, Wilson I, and Derrick D. Articulatory phonetics. *(first ed.), Wiley-Blackwell, Malden, MA*, 2013.

[26] Perkell JS, Cohen MH, Svirsky MA, Matthies ML, Garabieta I, and Jackson MT. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *J Acoust Soc Am*, 92(6):3078 – 96, 1992.

[27] Dagenais PA. Electropalatography in the treatment of articulation/phonological disorders. *J Commun Disord*, 28(4):303 – 329, 1995.

[28] Iskarous K. Patterns of tongue movement. *J Phon*, 33(4):363–381, 2006.

[29] Aron M, Kerrien E, Berger MO, and Laprie Y. Coupling electromagnetic sensors and ultrasound images for tongue tracking: acquisition setup and preliminary results. *7th Int Seminar Speech Production (ISSP)*, 2006.

[30] Stavness I, Lloyd JE., Payan Y, and Fels S. Coupled hard–soft tissue simulation with contact and constraints applied to jaw–tongue–hyoid dynamics. *Int J Numer Method Biomed Eng*, 27(3):367–390, 2011.

[31] Kim YC. Fast upper airway magnetic resonance imaging for assessment of speech production and sleep apnea. *Precis Futur Med*, 2(4):131–48, 2018.

[32] Chen W andByrd D, Narayanan SS, and Nayak KS. Intermittently tagged real-time mri reveals internal tongue motion during speech production. *Magn Reson Med*, 82(2):600–613, 2109.

[33] Chen W, Lee N, Byrd D, Narayanan SS, and Nayak KS. Improved real-time tagged mri using realtag. *Magn Reson Med*, 84(2):838–846, 2020.

[34] Scott AD, Wylezinska M, Birch MJ, and Miquel ME. Speech mri: Morphology and function. *Phys Med*, 30:604–618, 2014.

[35] Lingala SG, Sutton BP, Miquel ME, and Nayak KS. Recommendations for real-time speech mri. *J Magn Reson Imag*, 43:28–44, 2016.

[36] Ramanarayanan V, Tilsen S, Proctor M, Töger J, Goldstein L, Nayak KS, and Narayanan S. Analysis of speech production real-time mri. *Comput Speech Lang*, 52:1–22, 2018.

[37] Uecker M, Zhang S, Voit D, Karaus A, Merboldt KD, and Frahm J. Real-time mri at a resolution of 20 ms. *NMR Biomed*, 23(8):986–94, 2010.

[38] Niebergall A, Zhang S, Kunay E, Keydana G, Job M, Uecker M, and Frahm J. Real-time mri of speaking at a resolution of 33 ms: Undersampled radial flash with nonlinear inverse reconstruction. *Magn Reson Med*, 69:477–485, 2013.

[39] Iltis PW, Frahm J, Voit D, Joseph AA, Schoonderwaldt E, and Altenmüller E. High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players. *Quant Imaging Med Surg*, 5(3):374–81, 2015.

[40] Lingala SG, Zhu Y, Kim Y, Toutios A, Narayanan SS, and Nayak KS. A fast and flexible mri system for the study of dynamic vocal tract shaping. *Magn Reson Med*, 77:112–125, 2017.

[41] Lingala SG, Zhu Y, Lim Y, Toutios A, Ji Y, Lo WC, Seiberlich N, Narayanan SS, and Nayak KS. Feasibility of through-time spiral generalized autocalibrating partial parallel acquisition for low latency accelerated real-time mri of speech. *Magn Reson Med*, 78:2275–2282, 2017.

[42] Freitas AC, Ruthven M, Boubertakh R, and Miquel ME. Real-time speech mri: Commercial cartesian and non-cartesian sequences at 3t and feasibility of offline tgv reconstruction to visualise velopharyngeal motion. *Phys Medica*, 46:96–103, 2018.

[43] Ruthven M, Freitas AC, Boubertakh R, and Miquel ME. Application of radial grappa techniques to single- and multislice dynamic speech mri using a 16-channel neurovascular coil. *Magn Reson Med*, 82(3):948–58, 2019.

[44] Nayak KS, Lim Y, Campbell-Washburn A, and Steeden J. Real-time magnetic resonance imaging. *J Magn Reson Imaging*, 10.1002/jmri.27411, 2020.

[45] Counter SA, Olofsson A, Grahn H, and Borg E. Mri acoustic noise: sound pressure and frequency analysis. *J Magn Reson Imaging*, 7:606–611, 1997.

[46] Inouye JM, Blemker SS, and Inouye DI. Towards undistorted and noisefree speech in an mri scanner: correlation subtraction followed by spectral noise gating. *J Acoust Soc Am*, 135:1019–1022, 2014.

[47] Bresch E, Nielsen J, Nayak KS, and Narayanan SS. Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *J Acoust Soc Am*, 120:1791–1794, 2006.

[48] Vaz C, Ramanarayanan V, and Narayanan SS. Acoustic denoising using dictionary learning with spectral and temporal regularization. *IEEE/ACM Trans Audio, Speech, Language Process*, 26:967–980, 2018.

[49] Lingala SG, Toutios A, Toger J, Lim Y, Zhu Y, Kim YC, Vaz C, Narayanan SS, and Nayak KS. State-of-the-art mri protocol for comprehensive assessment of vocal tract structure and function. *In Proceedings of the Annual Conference of INTERSPEECH, San Francisco, CA, USA*, page 475–479, 2016.

[50] Santos JM, Wright GA, and Pauly JM. Flexible real-time magnetic resonance imaging framework. *Conf Proc IEEE Eng Med Biol Soc*, 2:1048–1051, 2004.

[51] Narayanan SS, Byrd D, and Kaun A. Geometry , kinematics , and acoustics of tamil liquid consonants. *J Acoust Soc Am*, 106:1993–2007, 1999.

[52] Kim YC, Proctor MI, Narayanan SS, and Nayak KS. Improved imaging of lingual articulation using real-time multislice mri. *J Magn Reson Imag*, 35:943–948, 2012.

[53] Burdumy M, Traser L, Burk F, Richter B, Echternach M, Korvink JG, Hennig J, and Zaitsev M. One-second mri of a three-dimensional vocal tract to measure dynamic articulator modifications. *J Magn Reson Imag*, 46:94–101, 2017.

[54] Fu M, Barlaz MS, Holtrop JL, Perry JL, Kuehn DP, Shosted RK, Liang ZP, and Sutton BP. High-frame-rate full-vocal-tract 3d dynamic speech imaging. *Magn Reson Med*, 77:1619–1629, 2017.

[55] Bresch E, Kim YC, Nayak KS, Byrd D, and Narayanan SS. Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging. *IEEE Signal Process Mag*, 25:123–129, 2008.

[56] Block KT and Frahm J. Spiral imaging: A critical appraisal. *J Magn Reson Imag*, 21:657–668, 2005.

[57] Meyer CH, Hu BS, Nishimura DG, and Macovski A. Fast spiral coronary artery imaging. *Magn Reson Med*, 28:202–213, 1992.

[58] Kim YC, Narayanan SS, and Nayak KS. Flexible retrospective selection of temporal resolution in real-time speech mri using a golden-ratio spiral view order. *Magn Reson Med*, 65:1365–1371, 2011.

[59] Proctor MI, Lammert A, Katsamanis A, Goldstein L, Hagedorn C, and Narayanan SS. Direct estimation of articulatory kinematics from real-time magnetic resonance image sequences. *In Proceedings of the Annual Conference of INTERSPEECH, Florence, Italy*, page 281–284, 2011.

[60] Lammert A, Ramanarayanan V, Proctor MI, and Narayanan SS. Vocal tract cross-distance estimation from real-time mri using region-of-interest analysis. *In Proceedings of the Annual Conference of INTERSPEECH, Lyon, France*, page 959–962, 2013.

[61] Proctor MI, Bone D, Katsamanis N, and Narayanan SS. Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. *In Proceedings of the Annual Conference of INTERSPEECH, Makuhari, Japan*, page 1576–1579, 2010.

[62] Kim J, Kumar N, Lee S, and Narayanan SS. Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. *In Proceedings of the 10th International Seminar on Speech Production (ISSP), Cologne, Germany*, page 222–225, 2014.

[63] Bresch E and Narayanan SS. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Trans Med Imaging*, 28:323–338, 2009.

[64] Browman C and Goldstein LM. Towards an articulatory phonology. *Phonology Yearbook*, 3:219–252, 1986.

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

[66] Drissi C, Mitrofanoff M, Talandier C, Falip C, Le Couls V, and Adamsbaum C. Feasibility of dynamic mri for evaluating velopharyngeal insufficiency in children. *Eur Radiol*, 21:1462–1469, 2011.

[67] Scott AD, Boubertakh R, Birch MJ, and Miquel ME. Towards clinical assessment of velopharyngeal closure using mri: Evaluation of real-time mri sequences at 1.5 and 3t. *Br J Radiol*, 85:1083–1092, 2012.

[68] Freitas AC, Wylezinska M, Birch MJ, Petersen SE, and Miquel ME. Comparison of cartesian and non-cartesian real-time mri sequences at 1.5t to assess velar motion and velopharyngeal closure during speech. *PLoS ONE*, 11:1–16, 2016.

[69] Bae Y, Kuehn DP, Conway CA, and Sutton BP. Real-time magnetic resonance imaging of velopharyngeal activities with simultaneous speech recordings. *Cleft Palate Craniofac J*, 48:695–707, 2011.

[70] Noll DC, Meyer CH, Pauly JM, Nishimura DG, and Macovski A. A homogeneity correction method for magnetic resonance imaging with time-varying gradients. *IEEE Trans Med Imaging*, 10:629–637, 1991.

[71] Man LC, Pauly JM, and Macovski A. Multifrequency interpolation for fast off-resonance correction. *Magn Reson Med*, 37:785–792, 1997.

[72] Nayak KS, Tsai CM, Meyer CH, and Nishimura DG. Efficient off-resonance correction for spiral imaging. *Magn Reson Med*, 45:521–524, 2001.

[73] Man LC, Pauly JM, and Macovski A. Improved automatic off-resonance correction without a field map in spiral imaging. *Magn Reson Med*, 37:906–913, 1997.

[74] Noll DC, Pauly JM, Meyer CH, Nishimura DG, and Macovski A. Deblurring for non-2d fourier transform magnetic resonance imaging. *Magn Reson Med*, 25:319–333, 1992.

[75] Chen W and Meyer CH. Fast automatic linear off-resonance correction method for spiral imaging. *Magn Reson Med*, 56:457–462, 2006.

[76] Smith TB and Nayak KS. Automatic off-resonance correction in spiral imaging with piecewise linear autofocus. *Magn Reson Med*, 69:82–90, 2013.

[77] Sutton BP, Conway CA, Bae Y, Seethamraju R, and Kuehn DP. Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (flash) at 3 t. *J Magn Reson Imag*, 32:1228–1237, 2010.

[78] Chen W and Meyer CH. Semiautomatic off-resonance correction in spiral imaging. *Magn Reson Med*, 59:1212–1219, 2008.

[79] Lee D, Nayak KS, and Pauly J. Reducing spurious minima in automatic off-resonance correction for spiral imaging. *In Proceedings of the International Society of Magnetic Resonance in Medicine, Kyoto, Japan*, page 2678, 2004.

[80] Roemer PB, Edelstein WA, Hayes CE, Souza SP, and Mueller OM. The nmr phased array. *Magn Reson Med*, 16:192–225, 1990.

[81] Fessler JA and Sutton BP. Nonuniform fast fourier transforms using min-max interpolation. *IEEE Trans Signal Proc*, 51:560–574, 2003.

[82] Sutton BP, Noll DC, and Fessler JA. Fast, iterative, field-corrected image reconstruction for mri. *IEEE Trans Med Imaging*, 22:178–188, 2003.

[83] Fessler JA, Lee S, Olafsson VT, Shi HR, and Noll DC. Toeplitz-based iterative image reconstruction for mri with correction for magnetic field inhomogeneity. *IEEE Trans Signal Proc*, 53:3393–3402, 2005.

[84] Vaz C, Toutios A, and Narayanan SS. Convex hull convolutive non-negative matrix factorization for uncovering temporal patterns in multivariate time-series data. *In Proceedings of the Annual Conference of INTERSPEECH, San Francisco, CA, USA*, page 963–967, 2016.

[85] Kim J, Lammert AC, Kumar Ghosh P, and Narayanan SS. Co-registration of speech production datasets from electromagnetic articulography and real-time magnetic resonance imaging. *J Acoust Soc Am*, 135:EL115–EL121, 2014.

[86] Töger J, Sorensen T, Somandepalli K, Toutios A, Lingala SG, Narayanan SS, and Nayak K. Test–retest repeatability of human speech biomarkers from static and real-time dynamic magnetic resonance imaging. *J Acoust Soc Am*, 141:3323–3336, 2017.

[87] Story BH, Titze IR, and Hoffman EA. Vocal tract area functions from magnetic resonance imaging. *J Acoust Soc Am*, 100:537–554, 1996.

[88] Kim YC, Kim J, Proctor MI, Toutios A, Nayak K, Lee S, and Narayanan SS. Toward automatic vocal tract area function estimation from accelerated three-dimensional magnetic resonance imaging. *ISCA Workshop on Speech Production in Automatic Speech Recognition, Lyon, France*, page 2–5, 2013.

[89] Skordilis ZI, Toutios A, Toger J, and Narayanan SS. Estimation of vocal tract area function from volumetric magnetic resonance imaging. *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA*, page 924–928, 2017.

[90] Robinson S, Grabner G, Witoszynskyj S, and Trattnig S. Combining phase images from multi-channel rf coils using 3d phase offset maps derived from a dual-echo scan. *Magn Reson Med*, 65:1638–1648, 2011.

[91] Brown RW, Cheng YCN, Haacke EM, Thompson MR, and Venkatesan R. Magnetic resonance imaging: Physical principles and sequence design: Second edition. *Hoboken, NJ: John Wiley and Sons*, 2014.

[92] Ahn CB, Kim JH, and Cho ZH. High-speed spiral-scan echo planar nmr imaging-i. *IEEE Trans Med Imaging*, 5:2–7, 1986.

[93] Nishimura DG, Irarrazabal P, and Meyer CH. A velocity k-space analysis of flow effects in echo-planar and spiral imaging. *Magn Reson Med*, 33:549–56, 1995.

[94] Gatehouse PD and Firmin DN. Flow distortion and signal loss in spiral imaging. *Magn Reson Med*, 41:1023–31, 1999.

[95] Yang PC, Kerr AB, Liu AC, Liang DH, Hardy C, Meyer CH, Macovski A, Pauly JM, and Hu BS. New real-time interactive cardiac magnetic resonance imaging system complements echocardiography. *J Am Coll Cardiol*, 32:2049–56, 1998.

[96] Nayak KS, Pauly JM, Kerr AB, Hu BS, and Nishimura DG. Real-time color flow mri. *Magn Reson Med*, 43:251–8, 2000.

[97] Nayak KS, Cunningham CH, Santos JM, and Pauly JM. Real-time cardiac mri at 3 tesla. *Magn Reson Med*, 51:655–660, 2004.

[98] Steeden JA, Kowalik GT, Tann O, Hughes M, Mortensen KH, and Muthurangu V. Real-time assessment of right and left ventricular volumes and function in children using high spatiotemporal resolution spiral bssfp with compressed sensing. *J Cardiovasc Magn Reson*, 20:1–11, 2018.

[99] Reeder SB, Hu HH, Sirlin CB, Group LI, and Diego S. Non-cartesian balanced ssfp pulse sequences for real-time cardiac mri. *Magn Reson Med*, 75:1546–1555, 2016.

[100] Liu H, Martin AJ, and Truwit CL. Interventional mri at high-field (1.5 t): needle artifacts. *J Magn Reson Imag*, 8:214–9, 1998.

[101] Shenberg I and Macovski A. Inhomogeneity and multiple dimension considerations in magnetic resonance imaging with time-varying gradients. *IEEE Trans Med Imaging*, 4:165–74, 1985.

[102] Maeda A, Sano K, and Yokoyama T. Reconstruction by weighted correlation for mri with time-varying gradients. *IEEE Trans Med Imaging*, 7:26–31, 1988.

[103] Makhijani MK and Nayak KS. Exact correction of sharply varying off-resonance effects in spiral mri. *Proc IEEE Int Symp Biomed Imaging, Arlington, VA*, page 730–733, 2006.

[104] Nayak KS and Nishimura DG. Automatic field map generation and off-resonance correction for projection reconstruction imaging. *Magn Reson Med*, 43:151–154, 2000.

[105] Noll DC. Reconstruction techniques for magnetic resonance imaging. page 1991, Ph.D.

[106] Zeng DY, Shaikh J, Holmes S, Brunsing RL, Pauly JM, Nishimura DG, Vasanawala SS, and Cheng JY. Deep residual network for off-resonance artifact correction with application to pediatric body mra with 3d cones. *Magn Reson Med*, 82:1398–1411, 2019.

[107] Lim Y, Lingala SG, Narayanan S, Nayak KS, and Angeles L. Improved depiction of tissue boundaries in vocal tract real-time mri using automatic off-resonance correction. *Proc INTERSPEECH, San Francisco, CA, USA*, page 1765–1769, 2016.

[108] Ye JC and Sung WK. Understanding geometry of encoder-decoder cnns. *arXiv:1901.07647*, 2019.

[109] Ahunbay E and Pipe JG. Rapid method for deblurring spiral mr images. *Magn Reson Med*, 44:491–4, 2000.

[110] Denis L, Thiébaut E, Soulez F, Becker JM, and Mourya R. Fast approximations of shift-variant blur. *Int J Comput Vis*, 115:253–278, 2015.

[111] Miraut D and Portilla J. Efficient shift-variant image restoration using deformable filtering (part i). *EURASIP J Adv Signal Process*, pages 1–20, 2012.

[112] Jin KH, McCann MT, Froustey E, and Unser M. Deep convolutional neural network for inverse problems in imaging. *IEEE Trans Image Process*, 26:1–20, 2016.

[113] Lee D, Yoo J, Tak S, and Ye J. Deep residual learning for accelerated mri using magnitude and phase networks. *IEEE Trans Biomed Eng*, 65:1985–1995, 2018.

[114] Han YS, Yoo J, and Ye JC. Deep residual learning for compressed sensing ct reconstruction via persistent homology analysis. *arXiv:161106391*, 2016.

[115] Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao P, and et al. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE Trans Med Imaging*, 36:2524–35, 2017.

[116] Mathieu M, Couprie C, and LeCun Y. Deep multi-scale video prediction beyond mean square error. *arXiv:1511.05440*, 2016.

[117] Kingma DP and Ba J. Adam: a method for stochastic optimization. *arXiv:1412.6980*, 2014.

[118] Ravishankar S and Bresler Y. Mr image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE Trans Med Imaging*, 30:1028–1041, 2011.

[119] Wang Z, Bovik AC, Sheikh HR, and Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*, 13:600–612, 2004.

[120] Somandepalli K, Toutios A, and Narayanan SS. Semantic edge detection for tracking vocal tract air-tissue boundaries in real-time magnetic resonance images. *Proc INTERSPEECH, Stockholm, Sweden*, page 631–635, 2017.

[121] Kim YC, Hayes CE, Narayanan SS, and Nayak KS. Novel 16-channel receive coil array for accelerated upper airway mri at 3 tesla. *Magn Reson Med*, 65:1711–1717, 2011.

[122] Fu M, Barlaz MS, Shosted RK, Liang ZP, and Sutton BP. High-resolution dynamic speech imaging with deformation estimation. *Conf Proc IEEE Eng Med Biol Soc, Milan, Italy*, pages 1568–71, 2015.

[123] Stone M. Toward a model of three-dimensional tongue movement. *J Phon*, 19:309–320, 1991.

[124] Martins P, Carbone I, Pinto A, Silva A, and Teixeira A. European portuguese mri based speech production studies. *Speech Commun*, 50:925–952, 2008.

[125] Zhu Y, Toutios A, Narayanan SS, and Nayak KS. Faster 3d vocal tract real-time mri using constrained reconstruction. *In Proceedings of the Annual Conference of INTERSPEECH, Lyon, France*, page 1292–1296, 2013.

[126] Chen J, Lingala SG, Lim Y, Toutios A, Narayanan SS, and Nayak KS. Task-based optimization of regularization in highly accelerated speech rt-mri. *In Proceedings of the International Society of Magnetic Resonance in Medicine, Honolulu, HI*, page 1409, 2017.

[127] Uecker M, Lai P, Murphy MJ, Virtue P, Elad M, Pauly JM, Vasanawala SS, and Lustig M. Espirit - an eigenvalue approach to autocalibrating parallel mri: Where sense meets grappa. *Magn Reson Med*, 71:990–1001, 2014.

[128] Tamir JI, Ong F, Cheng JY, Uecker M, and Lustig M. Generalized magnetic resonance image reconstruction using the berkeley advanced reconstruction toolbox. *ISMRM Workshop on Data Sampling and Image Reconstruction, Sedona, AZ*, 2016.

[129] Pauly J, Nishimura D, Macovski A, and Roux P Le. Parameter relations for the shinnar-le roux selective excitation pulse design algorithm. *IEEE Trans Med Imaging*, 10:53–65, 1991.

[130] Delattre P. Consonant gemination in four languages: an acoustic, perceptual, and radiographic study part i. *IRAL Int Rev Appl Linguist Lang Teach*, 9:31–52, 1971.

[131] Sproat R and Fujimura O. Allophonic variation in english /l/ and its implications for phonetic implementation. *J Phon*, 21:291–311, 1993.

[132] Traser L, Birkholz P, Flügge TV, Kamberger R, Burdumy M, Richter B, Korvink JG, and Echternach M. Relevance of the implementation of teeth in three-dimensional vocal tract models. *J Speech Lang and Hear Res*, 60(9):2379, 2017.

[133] Narayanan SS, Alwan AA, and Haker K. Toward articulatory-acoustic models for liquid approximants based on mri and epg data part i the laterals. *J Acoust Soc Am*, 101:1064–77, 1997.

[134] Iskarous K. Patterns of tongue movement. *J Phon*, 33:363–381, 2005.

[135] Zhu Y, Kim YC, Proctor MI, Narayanan SS, and Nayak KS. Dynamic 3-d visualization of vocal tract shaping during speech. *IEEE Trans Med Imaging*, 32:838–848, 2013.

[136] Hargreaves BA, Cunningham CH, Nishimura DG, and Conolly SM. Variable-rate selective excitation for rapid mri sequences. *Magn Reson Med*, 52:590–597, 2004.

[137] Matakos A. Dynamic image and fieldmap joint estimation methods for mri using single-shot trajectories. *PhD thesis, the University of Michigan*, 2013.