Introduction

Several techniques are available for the acquisition of data on the kinematics of speech production. Electromagnetic articulography (EMA) (Schönle et al., 1987) uses electromagnetic fields to track the positions of small coil sensors adhering to the articulators in two or three dimensions with sampling rates up to 500 Hz. Electropalatography (EPG) (Hardcastle et al., 1989) uses an artificial palate with embedded electrodes to record linguapalatal contact, typically at 100–200 Hz. Ultrasound can be used to image the tongue (Stone, 2006; Whalen et al., 2005) or larynx (Celata et al., 2017; Moisik, Lin and Esling, 2014) at 30–100 Hz. Despite their availability, these techniques are limited in their coverage of the vocal tract. EMA provides rich data about the movement of sensors on lingual and labial fleshpoints, but such sensors/markers cannot be easily placed at posterior locations on the tongue, on the velum, in the pharynx, or at the larynx; hence these technologies are limited in the spatial coverage of the complex vocal tract geometry. Additionally, EMA does not provide information as to the passive vocal tract structures such as the palate (both hard and soft) and pharyngeal wall that are important landmarks for constriction location and airway shaping. EPG is restricted to contact measurements of the tongue at the hard palate and typically does not exceed a 100 Hz sampling rate. Further, EPG does not record kinematic information of the tongue but rather the consequence of this movement for (obstruent) constriction formation. Ultrasound cannot consistently or reliably image the tongue tip, or opposing vocal tract surfaces such as the hard and soft palate and hence the airway shaping.

X-ray radiation has been used to image the sagittal projection of the entire vocal tract at rates typically between 10 and 50 frames per second (Badin et al., 1995; Delattre, 1971; Munhall, Vatikiotis-Bateson and Tohkura, 1995; Wood, 1982), providing rich dynamic data with superior coverage of the entire vocal-tract configuration. However, its use for speech research has today been abandoned for health and ethical reasons, since X-ray energy exposes subjects to unacceptable levels of radiation. Magnetic resonance imaging (MRI) has been used to capture images of static configurations of the vocal tract, also with very good coverage of its global configuration, but it does this while subjects sustain (and sometimes phonate) continuant speech sounds over unnaturally long periods of time, thus
producing static airway shaping information rather than dynamic speech production information (Clément et al., 2007; Narayanan, Alwan and Haker, 1995; Story, Titze, and Hoffman, 1996). While MRI has for some time been considered to be a slow imaging modality, modern techniques that were largely developed to capture the motion of the heart can now yield temporal resolutions exceeding those available with X-ray or ultrasound. In tagged or triggered MRI acquisition methods, notably repetitive cine-MRI (Stone et al., 2001; Takemoto et al., 2006), articulatory dynamics of running speech can be reconstructed from large numbers of repetitions (which should ideally be identical) of short utterances.

Real-time magnetic resonance imaging

Currently, significant advances in MR acquisition software, reconstruction strategies, and customized receiver coil hardware development have allowed real-time MRI (rtMRI) to emerge as an important and powerful modality for speech production research (Bresch et al., 2008; Narayanan et al., 2004; Niebergall et al., 2013; Scott et al., 2014; Sutton et al., 2009). RtMRI provides dynamic information from the entire midsagittal (or other) plane of a speaker’s upper airway during arbitrary, continuous spoken utterances with no need for repetitions. Sampling rates can now be achieved that are acceptable for running speech, and noise cancellation can yield an acceptable synchronized speech audio signal. RtMRI can capture not only lingual, labial and jaw motion, but also articulatory motion of the velopharyngeal mechanism and of the laryngeal articulatory mechanism, including shaping of the epilaryngeal tube and the glottis. Additionally, and in contrast to many other imaging or movement tracking modalities, rtMRI can acquire upper and rear airway structures such as the hard and soft palate, pharyngeal wall, and details of the structures of the laryngeal mechanism. Though EMA still has superior temporal resolution and audio quality (as well as price and accessibility), it cannot parallel rtMRI as a source of dynamic information about overall vocal tract movement and airway shaping.

An MRI scanner consists of electromagnets that surround the human body and create magnetic fields. MR images are formed through the interaction between externally applied magnetic fields and nuclear magnetic spins in hydrogen atoms present in the water molecules of the human body. A static magnetic field (typically 1.5, 3, or 7 Tesla) serves to polarize the hydrogen atoms. Atoms are then excited using a radiofrequency magnetic field and a set of linear magnetic field gradients that are dynamically changing over very short periods of time according to pre-designed pulse sequences. After excitation, as the hydrogen atoms return to equilibrium, they emit signals that represent samples in the spatial Fourier transform domain (which is referred to as k-space) of the excited area. Given enough samples, an inverse Fourier transform can reconstruct a map of the density of hydrogen atoms in the excited area. Typically, atoms in a single thin plane are excited, and the k-space and Fourier transforms are two-dimensional. The orientation of the slice is determined by the gradient fields.

In typical MRI applications, the k-space may be sampled for several seconds, or even minutes, in order to generate a single high-quality image. However, in order to capture the dynamics of speech production in fine temporal resolution, especially with no need of repetitions, the sampling needs to happen in amounts of time much shorter than a second (Lingala, Sutton, et al., 2016). For such real-time MRI, the k-space is sampled only partially (but in a highly principled manner) in order to generate images of sufficient quality for further analysis. Thus, real-time MRI is subject to a compromise between temporal resolution and image quality. Image quality mainly comprises two factors – spatial resolution and signal-to-noise ratio – that can to some extent be independently controlled.
An efficient way to sample the k-space is along spirals, as shown in Figure 2.1, which is achieved by appropriate pulse-sequence design (Narayanan et al., 2004). Samples covering such a spiral can be acquired in little more than 6 ms. Successive spirals are rotated by certain angles with respect to each other, and by combining samples from a few spirals, an image can be formed. In the visualization of Figure 2.1, an image is formed by combining k-space information from five successive spirals.

In a typical implementation of vocal-tract rtMRI that has been extensively used (Narayanan et al., 2014; Toutios and Narayanan, 2016), 13 spirals were combined to form an image, thus leading to a temporal resolution of 78 ms per image. Each image is comprised of 68 by 68 pixels, with a spatial resolution of 3 by 3 mm per pixel. Videos of vocal production were generated by combining successive images, or frames. New frames were generated by overlapping information from six spirals, which yielded videos with a frame rate of 23 frames per second.

The more recent and currently used rtMRI acquisition protocol applies a constrained reconstruction scheme (Lingala et al., 2017). This method exploits redundancy across

![Figure 2.1](image.png)
frames based on the fact that the critical dynamic information for speech production is concentrated at the edges between air and tissue, enabling a reconstruction with a temporal resolution of 12 ms, that is, a frame rate of 83 frames per second, with 84 by 84 pixels per image and 2.4 by 2.4 mm per pixel. A short example is shown in Figure 2.2. Crucial to the success of this method is the use of a custom array of coils surrounding the vocal tract that receive the signals emitted by the excited hydrogen atoms. These custom coils have superior sensitivity in regions corresponding to the pertinent vocal-tract structures (lips, tongue, velum, epiglottis, glottis) compared to receiver coils typically used in clinical MRI.

In most cases, rtMRI images a thin slice (~5 mm thick) at the mid-sagittal plane. However, it is possible to image any other plane, such as parasagittal, coronal, axial, or an arbitrary oblique plane. These views can offer particular insight to speech production research questions involving the shaping of constrictions, laterality, and stretching, compression, concavities/grooving and posturing of the tongue. It is also possible to image concurrently two or three slices, with the frame rate divided by a factor of two or three respectively – essentially switching slices rapidly between spiral acquisitions, giving rise to multi-slice dynamic videos. Figure 2.3 shows an example of such imaging. This technique can offer particular insight to the creation of more complex articulatory shaping and its concurrent larynx movement.

![Figure 2.2](image-url) Successive midsagittal real-time MRI frames of a male speaker uttering “don’t ask me.” The 40 frames shown span about 480 ms (video reconstructed at 83 frames per second). The International Phonetic Alphabet (IPA) annotation serves as a rough guide for the segments being produced in the image sequence.
Figure 2.3  Concurrent multi-slice real-time MRI for the sequence /baθa/. The two lines on the top mid-sagittal image show the orientations of the axial and coronal slices. The arrow on the coronal image during /θ/ shows a tongue groove channel (which would be very difficult to observe with other imaging modalities).

Speech audio is typically recorded concurrently with rtMRI in articulatory research. Acquiring and synchronizing the audio with the imaging data presents several technical challenges. Audio can be recorded using a fiber-optic microphone (Garthe, 1991), but the overall recording setup typically needs to be customized (Bresch et al., 2006). Synchronization between audio and images can be achieved using a trigger signal from the MRI scanner. That said, a significant challenge in audio acquisition is the high level of noise generated by the operation of the MRI scanner. It is important that this noise be canceled satisfactorily in order to enable further analysis of the speech acoustic signal. Proposed audio de-noising algorithms specifically targeting the task can exploit the periodic structure of the MR noise generated by pulse sequences (Bresch et al., 2006), or not (Vaz, Ramanarayanan and Narayanan, 2013); the latter algorithms can be used even when the applied pulse sequence leads to non-periodic noise.

RtMRI image processing and analysis

Reconstructed rtMRI data have the form of high-frame rate videos depicting the hydrogen density of tissue in a thin vocal-tract slice, most usually the midsagittal. Some speech
production phenomena may be studied simply by manually inspecting these videos and measuring the timing of articulatory events identified in the image sequences. To assist in this, our team at USC has developed and made freely available a graphical user interface (GUI) that allows users to browse the videos frame-by-frame, inspect synchronized audio and video segments in real time or at slower frame rates, and label speech segments of interest for further analysis with the supporting tool set.

Some speech production studies require more elaborate image processing and analysis of the rtMRI videos. Unsupervised segmentation of regions corresponding to the mandibular, maxillary and posterior areas of the upper airway has been achieved by exploiting spatial representations of these regions in the frequency domain, the native domain of MRI data (Bresch and Narayanan, 2009; Toutios and Narayanan, 2015). The segmentation algorithm uses an anatomically informed object model and returns a set of tissue boundaries for each frame of interest, allowing for quantification of articulator movement and vocal tract aperture in the midsagittal plane. The method makes use of alternate gradient vector flows, non-linear least squares optimization, and hierarchically optimized gradient descent procedures to refine estimates of tissue locations in the vocal tract. Thus, the method is automatic and well-suited for processing long sequences of MR images. Obtaining such vocal-tract air-tissue boundaries enables the calculation of vocal-tract midsagittal cross-distances, which in turn can be used to estimate area functions via reference sagittal-to-area transformations (Maeda, 1990; McGowan, Jackson and Berger, 2012; Soquet et al., 2002). See Figure 2.4 for sample results deploying this processing method.

![Figure 2.4](image.png)

*Figure 2.4  Example of region segmentation (white outlines) of articulators in rtMRI data. The word spoken by the female subject is “critical” with rough IPA annotation shown. The first frame corresponds to a pre-utterance pause posture.*
A limitation of this unsupervised regional segmentation method is that it is slow, requiring significant computational resources. To address the issue of computational resources, our team has also started developing a convolutional deep neural network, trained on examples of video frames with corresponding air-tissue boundaries derived via the original segmentation method (Somandepalli, Toutios, and Narayanan, 2017). Once this neural network is in place and fully tested, deriving high-quality air-tissue boundaries from new rtMRI frames will be instantaneous.

While air-tissue boundary detection is important for capturing the posture of individual articulators over time, it is often enough to observe the dynamics of the formation and release of constrictions in a specific region of the vocal tract. As a faster (yet less accurate) alternative, a method of rapid semi-automatic segmentation of rtMRI data for parametric analysis has been developed that seeks pixel intensity thresholds distributed along tract-normal grid lines and defines airway contours constrained with respect to a tract-centerline constructed between the glottis and lips (Kim, Kumar et al., 2014; Proctor et al., 2010). A version of this rapid method has been integrated in the aforementioned GUI.

Pixel intensity in an MR image is indicative of the presence or absence of soft tissue; consequently, articulator movement into and out of a region of interest in the airway can be estimated by calculating the change in mean pixel intensity as the articulator of interest moves (increasingly) into and out of the region of interest. Using this concept, a direct image analysis method has been developed that bypasses the need to identify tissue boundaries in the upper airway (Lammert, Proctor, and Narayanan, 2010). In this approach to rtMRI speech dynamic image analysis, constriction location targets can be automatically estimated by identifying regions of maximally dynamic correlated pixel activity along the palate and at the lips, and constriction and release gestures (goal-directed vocal tract actions) can be identified in the velocity profiles derived from the smoothed pixel intensity functions in vocal tract regions of interest (Proctor et al., 2011). Such methods of pixel intensity-based direct image analysis have been used in numerous studies examining the compositionality of speech production, discussed in what follows.

**Examining the compositionality of speech production**

Speech is dynamic in nature: It is realized through time-varying changes in vocal tract shape that emerge systematically from the combined effects of multiple constriction events distributed over space (i.e., subparts of the vocal tract) and over time. Understanding the spatiotemporal dynamics of speech production is fundamental to linguistic studies.

Real-time MRI allows pursuing this goal through investigating the compositionality of speech into cognitively controlled goal-directed vocal tract action events, called gestures (Browman and Goldstein, 1992). Of specific interest are: (a) the compositionality in space, i.e., the deployment of gestures distributed spatially over distinct constriction effector systems of the vocal tract; (b) the compositionality in time, i.e., the deployment of gestures co-produced temporally; and (c) the characterization of articulatory setting, i.e., the postural configuration(s) that vocal tract articulators tend to be deployed from and return to in the process of producing fluent and natural speech. Each of these areas of study will exhibit differences among languages that reflect the range of biologically viable linguistic systems.

An example study on the compositionality of speech production in space examined retroflex stops and rhotics in Tamil (Smith et al., 2013). The study revealed that in some contexts these consonants may be achieved with little or no retroflexion of the tongue tip. Rather, maneuvering and shaping of the tongue so as to achieve post-alveolar contact varies across...
vowel contexts. Between back vowels /a/ and /u/, post-alveolar constriction involves the curling back of the tongue tip, but in the context of the high front vowel /i/, the same constriction is achieved by tongue bunching. Results supported the view that so-called retroflex consonants have a specified target constriction in the post-alveolar region but indicate that the specific articulatory maneuvers employed to achieve this constriction are not fixed. The superposition of the consonantal constriction task with the tasks controlling the shaping and position of the tongue body for the surrounding vowels (in keeping, for example, with Öhman, 1966) leads to retroflexion of the tongue in some cases and tongue bunching in others.

An example line of research on gestural compositionality in time examined the coordination of velic and oral gestures for nasal consonants. For English /n/ (Byrd et al., 2009), it was found that near-synchrony of velum lowering and tongue tip raising characterizes the timing for [n] in syllable onsets, while temporal lag between the gestures is characteristic for codas, supporting and extending previous findings for /m/ obtained with a mechanical velotrace (Krakow, 1993). In French – which, unlike English, contrasts nasal and oral vowels – the coordination of velic and oral gestures was found to be more tightly controlled, to allow for the distinction between nasal vowels and nasal consonants (Proctor, Goldstein, et al., 2013). But while the nature of the coordinative relation was different between French and English, the timing of the corresponding gestures as a function of prosodic context varied in the same way.

Regarding the characterization of articulatory setting, research using rtMRI of speech has supported the hypothesis that pauses at major syntactic boundaries (i.e., grammatical pauses) but not at ungrammatical pauses (e.g., word search) are planned by a high-level cognitive mechanism that also controls and modulates the rate of articulation around these prosodic junctures (Ramanarayanan et al., 2014). This work further hypothesizes that postures adopted during grammatical pauses in speech are more mechanically advantageous than postures assumed at absolute rest, i.e., that small changes in articulator positions during grammatical pauses would produce larger changes in speech task variables than small changes during absolute rest. This hypothesis was verified using locally weighted linear regression to estimate the forward map from low-level articulator variables to high-level task variables (Lammert et al., 2013). The analysis showed that articulatory postures assumed during grammatical pauses in speech, as well as speech-ready postures, are significantly more mechanically advantageous than postures assumed during absolute rest.

RtMRI data, complemented by speech audio, have afforded new insights into the nature and execution of speech production goals, the relationship between speech articulation and acoustics, and the nature of variability in speech motor control. Acquisition methodologies have now matured enough to enable speech production research across languages with a variety of articulatory shaping maneuvers and temporal patterning of articulation. Examples of studies enabled by rtMRI, from our team and other research groups, include, in addition to the aforementioned ones, work that has examined the production of nasals in English, French (Carignan et al., 2015; Proctor, Goldstein, et al., 2013), and Brazilian Portuguese (Barlaz et al., 2015; Meireles et al., 2015); liquids in English (Harper, Goldstein, and Narayan, 2016; Proctor and Walker, 2012) and Korean (Lee, Goldstein, and Narayan, 2015); English diphthongs (Hsieh and Goldstein, 2015), (Proctor et al., 2016); English and Spanish coronal stops (Parrell and Narayan, 2014); Lebanese Arabic coronal “emphatic” (uvular-ized) consonants (Israel et al., 2012); Tamil retroflexes (Smith et al., 2013); Puerto Rican Spanish rhotics (Monteserín, Narayan, and Goldstein, 2016); and Khoisan clicks (Proctor et al., 2016).
Finally, rtMRI can provide a unique view of speech compositionality in breakdown. Previous work has demonstrated the potential of rtMRI to study the characteristics of speech production of people suffering from verbal apraxia, i.e., the inability to execute a voluntary movement despite being able to demonstrate normal muscle function (Hagedorn et al., 2017), and people who have undergone surgical removal of part of their tongue (glossectomy) because of cancer (Hagedorn et al., 2014). Dynamical articulatory imaging can provide detailed and quantifiable characterizations of speech deficits in spatiotemporal coordination and/or execution of linguistic gestures and of compensatory maneuvers that are adopted by speakers in the face of such hurdles. This rtMRI data can have important potential benefits including for therapeutic and surgical interventions; additionally, an examination of the linguistic system in breakdown can contribute to a better understanding of the cognitive structure of healthy speech production.

**A data-driven computational speech production framework**

The collection of extensive amounts of rtMRI data enables computational modeling work that can advance the refinement of existing speech production models and the development of new ones. Of particular interest is modeling speech production across different individuals, in order to explore how individual vocal-tract morphological differences are reflected in the acoustic speech signal and what articulatory strategies are adopted in the presence of such morphological differences to achieve speech invariance, either perceptual or acoustic. One of the long-term objectives of this ongoing work is to improve scientific understanding of how vocal-tract morphology and speech articulation interact and to explain the variant and invariant aspects of speech properties within and across talkers. Initial work with rtMRI has focused on individual differences in the size, shape and relative proportions of the hard palate and posterior pharyngeal wall. Specific aims have been to characterize such differences (Lammert, Proctor, and Narayanan, 2013b), to examine how they relate to speaker-specific articulatory and acoustic patterns (Lammert, Proctor, and Narayanan, 2013a), and to explore the possibility of predicting them automatically from the acoustic signal (Li et al., 2013). Moreover, rtMRI may help characterize individual differences of other vocal-tract structures such as the epilaryngeal tube and the glottis (Moisik, Lin, and Esling, 2014).

In more recent work, a factor analysis was applied to air-tissue boundaries derived by the previously mentioned automatic segmentation algorithm (Bresch and Narayanan, 2009). The method, which was inspired by older articulatory models based on limited amounts of X-ray data (Harshman, Ladefoged, and Goldstein, 1977; Maeda, 1990), decomposes the vocal-tract dynamics into a set of articulatory parameter trajectories corresponding to relative contributions (degrees of freedom) of the jaw, tongue, lips, velum, and larynx, and operating on speaker-specific vocal-tract deformations (Toutios and Narayanan, 2015). Constrictions along the vocal tract were also measured from the segmentation results and a locally linear mapping from model parameters to constriction degrees was found using a hierarchical clustering process with a linearity test (Sorensen et al., 2016).

Having a locally linear mapping between linguistically critical constrictions along the vocal tract and the overall vocal-tract shaping, as represented compactly by the parameters of the articulatory model, enables the application of dynamical system modeling to animate the vocal tract towards the achievement of such constrictions. The assumption that deformations of the vocal-tract are governed by dynamical systems (more specifically, critically damped oscillators) that operate on the constrictions is a central concept in the theory of Articulatory Phonology (Browman and Goldstein, 1992) and Task Dynamics (Saltzman and...
Vocal tract imaging and analysis

Munhall, 1989), its computational counterpart. The framework (factor analysis model; mapping between articulatory parameters and constrictions; animation with dynamical systems) was applied to a large number of English speakers to identify speaker-specific strategies that govern the tongue–jaw and lip–jaw coordinative synergies by which different individuals achieve vocal-tract constrictions (Sorensen et al., 2016).

Initial work has also been done toward using this framework for the synthesis of realistic vocal-tract dynamics from dynamical systems specifications. Critically damped oscillators involved in speech production (as proposed by Articulatory Phonology) are characterized by vectors of targets and stiffnesses, operating on task variables (constrictions along the vocal tract). The proposed framework enables casting the same dynamical systems to operate on the parameters of the articulatory model, which, in turn, can readily construct the midsagittal vocal tract shaping dynamics. The approach has been put forward for synthesizing vocal-tract dynamics for VCV sequences (where C was a voiced plosive), and these synthesized dynamics were used as inputs to an articulatory-to-acoustic simulator (Maeda, 1982), which generated satisfactory acoustic results (Alexander et al., 2017). Stiffness and target vectors in that work were set manually, but this is considered only as an essential first step toward fitting these parameters to the actual rtMRI data in an analysis-by-synthesis setup, which will be equivalent to uncovering the spatiotemporal structure of speech motor control commands, under the assumption of Task Dynamics (Saltzman and Munhall, 1989).

Extending the scope of Articulatory Phonology, one may forgo the assumption of critically damped oscillators and look instead to statistically decompose vocal-tract kinematics into a set of (free-form) spatiotemporal bases, or primitives, a small number of which are activated at any time in the course of a speech utterance. This has been achieved using a novel convolutive non-negative matrix factorization (NMF) algorithm (Vaz, Toutios, and Narayanan, 2016), in conjunction with the articulatory model and the parameter-to-constriction mapping. Vocal-tract dynamics were resynthesized efficiently from derived bases and activation patterns, however the association of these to phonological units remains challenging.

Strategies for modeling articulatory behavior can be beneficial towards goals beyond understanding linguistic control regimes and individual differences. These tools can help shed light to on paralinguistic aspects of articulatory behavior. One such case is the expression of emotion. The state-of-the-art in speech emotion research has predominantly focused on surface speech acoustic properties; there remain open questions as to how speech properties covary across emotional types, talkers, and linguistic conditions. Given the complex interplay between the linguistic and paralinguistic aspects of speech production, there are limitations to uncovering the underlying details just from the resultant acoustics. As will be discussed later in this chapter, a large rtMRI database of emotional speech has been collected, and analysis of this data using some of the tools described earlier is under way. Finally, rtMRI is being used to study different types of vocal performance, including Western Classical Soprano singing (Bresch and Narayanan, 2010) and Human Beatboxing performance (Proctor, Bresch, et al., 2013). This work investigates how human vocal organs are utilized in different performance styles, how performers adopt articulatory strategies to achieve specific acoustic goals, how their articulation in performance resembles or differs from that of spoken speech, and how percussive and linguistic gestures are coordinated.

Databases

In order to facilitate speech production research using rtMRI across the speech community, a body of rtMRI data, with synchronized and de-noised audio, has been made publicly available.
available. The USC-TIMIT database (Narayanan et al., 2014) includes rtMRI data from ten speakers (five male, five female), each producing a set of 460 phonetically balanced sentences from the MOCHA-TIMIT set.

The USC-EMO-MRI database includes rtMRI data from ten actors (five male, five female), each enacting four different emotions in the MRI scanner (neutral, happy, angry, sad) while repeating a small set of sentences multiple times (Kim, Toutios et al., 2014).

The USC Speech and Vocal Tract Morphology MRI Database (Sorensen et al., 2017) includes data from 17 speakers (eight male, nine female). In addition to rtMRI, the dataset also includes three-dimensional volumetric MRI data of vocal tract shapes during sustained speech sounds. The rtMRI data include consonant-vowel-consonant sequences, vowel-consonant-vowel sequences, read passages, and spontaneous speech. One of the passages was produced in five different speaking styles: normal, fast, slow, whispered and shouting. Volumetric MRI was acquired using an accelerated protocol that used 7 seconds to scan the full volume of the vocal tract. This (relatively) short scan time enabled the acquisition of volumetric MRI of the full set of vowels and continuant consonants of American English.

While rtMRI data in the aforementioned databases were collected using earlier acquisition protocols at 23 frames per second, the latest technology with improved temporal resolution is now showcased in a collection, available online, in which four expert phoneticians produce sounds of the world’s languages as denoted in the IPA, with some supplementary English words and phonetically balanced texts (Toutios et al., 2016).

Open challenges

Real-time MRI for speech production research still presents challenges. First, rtMRI is currently done in a supine position, not the common upright posture for speech. Much literature has been devoted to the assessment of differences in speech articulation between the two positions (Kitamura et al., 2005; Stone et al., 2007; Tiede, Masaki, and Vatikiotis-Bateson, 2000; Traser et al., 2014). It has been suggested that positional differences are quite limited and that compensatory mechanisms, at least in healthy subjects, are sufficiently effective to allow the acquisition of meaningful speech data in a supine position (Scott et al., 2014). The potential use of upright or open-type scanners would fully remove this consideration, and there have been a few studies that demonstrate the utility of such scanners for upper-airway structures (Honda and Hata, 2007; Perry, 2010).

The MRI scanner is a very noisy environment, and subjects need to wear earplugs during acquisition, thus infringing on natural auditory feedback. Though it may be speculated the subjects would speak much louder than normal or that their articulation would be significantly affected as a result, it has been observed that these statements held true only rarely (Toutios and Narayanan, 2016). It is possible that somatosensory feedback compensates for the shortfall in auditory feedback (Katseff, Houde, and Johnson, 2012; Lametti, Nasir, and Ostry, 2012) and/or that the impairment in feedback is not exceedingly severe or perturbing.

Because of the large magnetic fields involved, people need to be excluded from being subjects in MRI research if they have prosthetics such as pacemakers, defibrillators, or (most) cochlear implants; these subjects can be identified and excluded in a screening process (Murphy and Brunberg, 1997). Otherwise, subject comfort is usually not an issue for adult healthy subjects and for scan durations (overall time spent in the scanner) of up to 90 minutes (Lingala, Toutios, et al., 2016).

Dental work is not a safety concern but may pose imaging issues. However, the disruptions associated with most dental work do not consistently degrade image quality. In
general, image quality is subject-dependent (and in some cases it can be difficult to even maintain constant quality throughout the speech sample) (Lingala, Sutton, et al., 2016). It has been observed that the impact of dental work appears to be more prominent when such work resides on the plane that is imaged and that the impact is often quite localized around the dental work (Toutios and Narayanan, 2016). For example, orthodontic permanent retainers at the upper incisors result in loss of midsagittal visual information from a small circle (typically with diameter up to 3 cm) around the upper incisors.

The teeth themselves are not visible in MRI due to their chemical composition. Various methods have been used to superimpose teeth onto MRI images, including using data from supplementary Computed Tomography (CT) imaging (Story, Titze, and Hoffman, 1996), dental casts (Alwan, Narayanan, and Haker, 1997; Narayanan, Alwan, and Haker, 1997), or MRI data acquired using a contrast agent in the oral cavity such as blueberry juice (Takimoto et al., 2004) or ferric ammonium citrate (Ng et al., 2011) that leaves the teeth as signal voids. Recently, a method was proposed that reconstructs the teeth from a single three-dimensional MRI scan, where the speaker sustains a specially designed posture: Lips closed and tongue tightly in contact with the teeth (Zhang, Honda, and Wei, 2018).

Finally, even though MRI scanners are today commonplace (albeit expensive), they are not portable. Thus, rtMRI is not amenable to fieldwork studies. However, by providing unparalleled dynamic images of the entire vocal tract, rtMRI can help develop models that predict global tongue shaping from partial information, such as that provided by a portable ultrasound or EPG.

Conclusion

Research in the area of speech production has long sought to obtain accurate information about the movement and shaping of the vocal tract from larynx to lips. Dynamic, real-time articulatory data are crucial for the study of phonemic inventories, cross-linguistic phonetic processes, articulatory variability, and phonological theory. Such data afford insights into the nature and execution of speech production, the cognitive mechanisms of motor control, the relationship between speech articulation and acoustics, and the coordination of goals postulated in models of speech production. This chapter has presented an overview of recent advances in vocal tract imaging focusing on real-time MRI, and has reviewed examples of applications.

Real-time MRI presents an unprecedented opportunity for advancing speech production research. Current state-of-the-art in MR acquisition software, reconstruction strategies, and receiver coil development have allowed real-time MRI (rtMRI) to provide clear benefits in terms of spatial vocal tract coverage with very good temporal resolution as compared to other techniques for articulatory data acquisition (e.g., Bresch et al., 2008; Narayanan et al., 2004). RtMRI provides dynamic information from the entire midsagittal (or other) plane of a speaker’s upper airway during arbitrary, continuous spoken utterances with no need for repetitions, and noise-cancellation techniques yield an acceptable synchronized speech audio signal. Many important early findings in speech production research were based on the study of X-ray videos of the vocal tract that collected limited amounts of data per speaker from a limited number of speakers, before being abandoned because of serious health concerns. Real-time MRI does not have such safety constraints. Real-time MRI further enables dynamic imaging of any arbitrary slice of interest (sagittal, coronal, axial, or oblique) in the vocal tract from larynx to lips, thus offering a most comprehensive means of observing the dynamics of vocal tract shaping. Finally, image
processing and data analysis techniques are rapidly advancing the quantification and interpretation of this valuable real-time articulatory data.

Acknowledgments

Work supported by NIH grant R01DC007124 and NSF grant 1514544. The authors wish to thank Professor John Esling for valuable feedback on an earlier version of this chapter.

Notes

2 Narayanan S. et al., SPAN: Resources, http://sail.usc.edu/span/resources.html

References


