Review paper

Speech MRI: Morphology and function

Andrew D. Scotta,b, Marzena Wylezinskaa,c, Malcolm J. Bircha, Marc E. Miquela,c,*

a Clinical Physics, Barts Health NHS Trust, London EC1A 7BE, United Kingdom
b NIHR Cardiovascular Biomedical Research Unit, The Royal Brompton Hospital, Sydney Street, London SW3 6NP, United Kingdom
c Barts and The London NIHR CVBRU, London Chest Hospital, London E2 9JX, United Kingdom

A R T I C L E   I N F O

Article history:
Received 3 March 2014
Received in revised form 24 April 2014
Accepted 1 May 2014
Available online 28 May 2014

Keywords:
Speech imaging
MRI dynamic imaging
Real-time imaging
Vocal tract imaging and modelling
Cleft palate
Vocal tract morphology
Speech production

A B S T R A C T

Magnetic Resonance Imaging (MRI) plays an increasing role in the study of speech. This article reviews the MRI literature of anatomical imaging, imaging for acoustic modelling and dynamic imaging. It describes existing imaging techniques attempting to meet the challenges of imaging the upper airway during speech and examines the remaining hurdles and future research directions.

© 2014 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. All rights reserved.

Introduction

Basics of speech production

The production of human speech is a complex process, see Refs. [1–4] for example, involving numerous organs, namely: the lungs, diaphragm and chest wall; the larynx, pharynx and vocal folds (or cords); the tongue, lips and soft palate (or velum); and the teeth, jaw and nasal cavity. The lungs, driven by the diaphragm and/or chest wall provide the airflow which travels through the lower respiratory tract (the bronchioles, the two bronchial tubes and the trachea). Air enters the upper respiratory tract (which will be the focus of this review) through the larynx which contains the vocal cords, a schematic of the upper airway in mid-sagittal view is represented on Fig. 1. Air is forced through a narrow gap between the vocal folds which vibrate, producing a fundamental frequency plus harmonics. By manipulation of the tension, length and separation of the vocal cords and control of the airflow between them, the fundamental frequency, volume, and therefore the intonation, of speech can be controlled. The remainder of the upper respiratory tract forms a series of connected resonant cavities which can be modified in size and shape using the pharynx, velum, uvula, jaw, tongue and lips. These manipulators modify the formant frequencies of speech. For consonant sounds, speech is further complicated by articulation, which is the partial or full obstruction of the vocal tract by a pair of articulators — the tongue tip and upper teeth, tongue body and hard palate, the lips or the velum and tongue dorsum, for example (Fig. 2).

Clinical assessment of speech

A number of acquired and inherited diseases can affect the function of the speech organs, including cancer [5,6], clefts of the lips and/or palate [4], laryngitis, vocal cord polyps [7], nodules or cysts [8], neurological conditions [9] and some endocrine disorders [10]. Standard diagnostic tools for speech therapists assessing speech disorders include aural assessment, acoustic analysis, oral and nasal airflow measurements [11] and external measurement of the impedance across the vocal folds to measure their movement (electroglottography) [12]. Imaging techniques may also be used, most commonly endoscopy. Stroboscopy is a variant of endoscopy which is used for analysis of vocal fold movement in slow motion [13]. X-Ray fluoroscopy is also used to assess the movement of the vocal folds [14], soft palate and pharyngeal wall movements [15,16]. However, endoscopy is invasive (although minimally) potentially causing discomfort and abnormal speech. It is also limited to
produce en-face images of exterior surfaces within the vocal tract. In contrast, X-ray fluoroscopy is non-invasive and operates at both high temporal (~30 frames per second (fps)) and high in-plane spatial resolution (<0.5 × 0.5 mm² [17]), but is limited to produce projections of the anatomy. The soft-tissue contrast of X-ray fluoroscopy is relatively poor, but the bony structures are clearly seen and the outline of both the soft palate and posterior pharyngeal wall can be identified. Typically, the lateral view is used, but other orientations such as the townes and basal views [15,18,19] have been used to add additional information. Increased contrast [20] is available by coating the vocal tract with barium contrast at the expense of increased patient discomfort [21]. X-Ray imaging also results in an ionising radiation dose to both the patient and operator [22]. Estimates of the average patient effective dose in dysphagia studies, which target similar anatomy, vary between 0.2 mSv [23], 0.4 mSv [24] and 0.85 mSv [22], but are dependent on the exposure time. Speech assessments may occur regularly, resulting in repeated radiation exposures and studies are typically longer than dysphasia protocols in order to obtain a varied speech sample.

Alternative imaging techniques: CT and ultrasound

Other imaging techniques available for imaging the speech articulators include ultrasound and X-ray computed tomography (CT). Ultrasound is relatively low cost, widely available, rapid and free from ionising radiation [25]. Relatively narrow (~2 mm) sections of the mid part of the tongue can be imaged in coronal or sagittal planes, from underneath the chin, at temporal resolutions of around 30 fps and sub-millimetre spatial resolution. However, while the palate and velum may be observed in some lingual positions, air interfaces, and bone are highly reflective and structures beyond these interfaces are not visualised, somewhat limiting the technique. Furthermore, the need for the transducer to be in direct contact with the skin may interfere with normal speech. CT is able to obtain high spatial resolution (<1 × 1 × 1 mm³) volumetric images with delineation of soft tissue and bone, which may be reformatted in any plane. However, radiation doses are typically greater than those from planar X-ray studies. Furthermore, temporal resolution is currently limited to around 165 ms at 3 gantry rotations per second or around 85 ms using state of the art dual source CT systems [26,27] which is insufficient for detailed analysis of tongue, lip and velar motion.

MRI: imaging anatomy and function?

MR imaging provides tomographic images with excellent soft tissue contrast in any plane without the use of ionising radiation and scanners are now commonplace. While MR was previously considered as a “slow” imaging modality, modern techniques, largely developed as a result of the desire to capture or freeze the motion of the heart, can result in temporal resolutions far exceeding those available with CT and even ultrasound [28]. As it is also able to provide images of both the anatomy and function of the vocal tract, MRI as the potential to become the modality of choice in speech imaging. This has resulted in a large body of literature describing the acquisition of MR images in the vocal tract and the potential diagnostic and modelling applications.

Despite the growing body of literature, outside technical MR journals many studies fail to adequately or accurately describe the MR acquisition parameters. Studies often neglect to fully describe important parameters such as resolution [29–34], flip angle [33.35–41], echo train length [33.42–45], or sequence type [30.46.47]. Errors in the description of the methods include echo times (TE) longer than repetition times (TR) [48], descriptions of multi-slice 2D sequences as 3D acquisitions [46] and a steady state turbo spin echo sequence (instead of single shot) [49].

While existing work has considered the application of MRI in a variety of situations, challenges and limitations still exist and further work is required before the modality is used routinely in the vocal tract. For example, large changes in magnetic susceptibility are present at tissue–air interfaces which result in sequence dependent artefacts; signal to noise ratio (SNR) must be traded for temporal and/or spatial resolution; many of the reconstruction techniques for rapid acquisitions are computationally challenging; and imaging is almost universally performed with the subject in the supine position.

No previous review of this large and expanding field has previously been published. In this review we will attempt to provide a rational approach to the body of literature divided into imaging anatomy, imaging for acoustic modelling and dynamic imaging. We will describe existing techniques attempting to meet the challenges of imaging the upper airway during speech and examine the remaining hurdles.

Static imaging of the vocal tract

The excellent soft tissue contrast and lack of ionising radiation have made MR imaging a popular research tool in analysing the...
morphology of the upper respiratory tract in healthy and patient cohorts. Despite the intrinsically dynamic nature of speech production, a large amount of research has been performed while the subject is either silent or during sustained phonation of a relevant sound. For many of these acquisitions, where spatial resolution and contrast are more important than a short acquisition duration, multi-slice 2D spin echo based techniques are used for their resilience to artefacts caused by differences in magnetic susceptibility, resulting in imaging times $>$10 s. The very first MR studies of the upper respiratory tract were performed in this way at low field strengths for assessment of the vocal tract shape [50–52]. However, such long phonations increase the potential for accidental movement or breathing artefact during the sustained phonation and more rapid gradient echo based sequences have been used as an alternative to reduce acquisition duration to 1 s or below (see Section 4).

Imaging the larynx

MR imaging of the larynx (and the pharynx) is common place for the diagnosis and assessment of benign growths, malignant tumours, relapsing polychondritis and necrotising fasciitis [53]. However, tomographic imaging of the vocal folds is less commonly used in assessing speech; as the fundamental frequency of normal speech is between 100 and 250 Hz [54], the opening and closing of the vocal folds are too rapid for the temporal resolution of MR techniques. Despite this, MR imaging at 1.5 T and below has been used to image the vocal folds during sustained phonation in order to aid vocal fold modelling [55], assessment of laryngeal surgery (thyroplasty) [56] and assessment of the angulation and superior–inferior movement of the larynx [44]. The limited duration of phonation, limits such studies to 2D acquisitions, but the multiple thin contiguous slices available from 3D imaging would be preferable. In recent work, high resolution 3D imaging at 3 T was afforded by dividing the phase encode lines between multiple sustained phonations [57]. Static studies (without phonation) reduce the limitations on imaging time but are limited by both voluntary motion and involuntary motion due to respiration, coughing, swallowing and pulsatile carotid flow. In order to address this, Barral et al. [58] used navigator echoes interleaved within a high resolution 3D acquisition to detect and either reacquire or correct phase encode lines corrupted by motion. Studies are typically performed with spin echo based techniques to minimise artefacts caused by sharp changes in magnetic susceptibility present at the tissue–air interfaces. Balanced steady-state free precession (bSSFP) imaging, which is a rapid technique with high SNR has also been demonstrated for real-time imaging of the larynx during phonation (4 fps) [59], but the resulting artefacts due to magnetic susceptibility were severe.

Analysing vocal tract shape

The tomographic nature of MRI with its suitability for use in healthy volunteers means that it is the modality of choice for imaging the vocal tract at rest and during sustained phonation. Data from such studies is typically used in modelling speech production and the analysis of the function of the articulators.

The formant frequencies of speech can be predicted from an area function, which describes the variation in cross-sectional area with length along the vocal tract (see Fig. 3). A number of studies have used MRI to obtain area functions [31,48,52,60–63], typically spin echo based sequences, which are less sensitive to signal loss at the tissue–air-interfaces. Most simplistically, area functions can be derived from measurements of the vocal tract on a single mid-sagittal MRI image or X-ray using empirical relationships relating mid-sagittal width to cross-sectional area [60,64]. However, the casts used to generate the empirical relationships are typically obtained from cadavers and may not fully represent the vocal tract during speech. Alternatively, area functions may be estimated from sparsely spaced 2D MR images positioned perpendicular to the vocal tract along its length [32]. The cross-sectional area at points between the slices can be estimated by fitting a model to the...
measured cross-sectional areas and the mid-sagittal distance obtained from a mid-sagittal image. Demolin et al. [65] used such a technique in 2 healthy subjects with 14 slices positioned along the vocal tract during sustained phonation of French oral vowels. The measured areas were compared with those calculated from published empirical relationships and subject specific models using the mid-sagittal width measured from the same images. Performance of the model varied between vocal tract sections (oral, nasal, pharyngeal, larynx, etc.) and overall, the subject specific models performed better. In order to avoid such empirical relationships altogether, 2D acquisitions with multiple contiguous slices [62,63,66,67] or even 3D acquisitions [68] have been used to capture the vocal tract configuration during sustained phonation. This provides information on vocal tract configuration unavailable with any other imaging modality except X-ray CT, which has limited applicability in healthy subjects. As a result of such studies, measured area functions are available for English fricative consonants [66], vowels, nasals and plosives [62,69], laterals [70], rhotics [71], Tamil liquid consonants [72], French vowels [48] and an extensive collection of European Portuguese consonants and vowels [73] (see Fig. 3). In addition, one study considered the differences in area functions between different voice qualities: normal, “yawny” and “twangy” [74] and another, by the same author [75] considered the changes in vocal tract configuration in the same speaker over 8 years. While the data was not directly acquired for analysis of the vocal tract, work by Fitch et al. [76] demonstrates relationships in 129 young subjects between vocal tract length and subject size, sex and pubescent status. Recently, differences in the morphology of the whole vocal tract among three voice professionals were also demonstrated using 3D volumetric MRI [68].

The accuracy of the predicted formant frequencies generated by the measured area functions varies by subject and phonation. In work which compared the predicted frequencies to those measured in speech samples recorded outside the scanner, the mean absolute error in the first three calculated formants was 7.5 ± 8.0% (range 0.46%–43.6%) and 9.4 ± 6.6% (range 0.2%–22.1%) in a male subject (vowels, nasals and plosives) [62] and female subject (vowels and laterals) [69] respectively. Vocal tract geometry is often simplified to assist in the computation. However, analysis of MRI data has also demonstrated the increased error when the piriform fossa [69,77] and the asymmetry of nasal cavities [78,79] are not included when calculating formants.

Recently, the super-resolution reconstruction technique [80] has been used to improve vocal tract MR images. Multiple orthogonal low-resolution MRI stacks were integrated into one isotropic super-resolution volume. This implementation also incorporated both edge-preserving regularisation to improve SNR and resolution; and motion correction between acquisitions of different stacks. When applied to low-resolution MR stacks derived from library MR datasets (ATR Human Information Science Laboratories MRI database of Japanese vowel production [81]), the resulting resolution was improved by a factor of three. The overall area function derived from super-resolution volumes also produced better prediction of formants, particularly if those formants were sensitive to area perturbation at constrictions (i.e. in the laryngeal cavity).

Beyond area functions, MRI has been used to analyse the positioning of the articulators during various sustained phonations [79], particularly the tongue [70–72,82,83]. Work by Narayanan and Alwan [70] demonstrated the grooving of the tongue along the mid-line with lateral flow channels and a convex shape in the posterior section during the phonation of lateral sounds (e.g. “l”). In contrast, the rhotic sounds [71] were characterised by a more retracted tongue with a convex anterior shape and concave posterior shape. Such an analysis is only possible using MRI as other modalities are either unable to obtain this information or unethical for use in healthy subjects.

Accurate modelling also requires information on the dental geometry. However, the lack of MRI signal from the teeth means they must be superimposed onto the MRI images. Various methods have been used to achieve this, including using data from supplementary CT imaging [62,69], but the associated radiation dose means that it might be difficult to obtain ethical approval. Other work has used dental casts, which may be either measured with callipers [70,71,82,84] or imaged in a bath of water [67,83]. Alternatively, Takemoto et al. [85] acquired MRI data of a subject using blueberry juice as a contrast agent in the oral cavity, leaving the teeth as signal voids. In this case, images of the teeth were obtained in an acquisition with the subject in the prone position and the images obtained for speech modelling were acquired separately in the supine position. However, there is no obvious reason why dental imaging and imaging of sustained phonation could not be performed in the same session in the same patient position, thus reducing the complications of co-registering the two datasets.

MRI data has also been used in combination with other speech analysis tools. Palatography coats the tongue with black powder and electropalatography (EPG) uses an oral insert covered with electrodes attached to the hard palate, to detect the locations of tongue contact [70–72,86]. While MRI and EPG have been used for static vocal tract analysis, electromagnetic articulography [87] (EMA), which tracks miniature coils attached to the speech articulators, has been used to add dynamic information to the static MRI studies [72,88]. Dynamic information on the movement of the tongue and lips may also be obtained from video imaging and has been used to supplement static MRI data in speech modelling [83]. Recently a method was proposed for use in speech simulation [89] based on volumetric MRI and CT data consisting of sustained

Figure 3. Example of area function calculation for Portuguese vowels. Reproduced from Martins et al. [73] with permission from Elsevier.
articulations of vowels and consonants. The model was applied to the synthesis of consonant–vowel syllables which were recognised in 82% of cases in a perception test.

Many such studies have used spin echo based sequences [62,83] which are resistant to signal drop out artefacts at the tissue–air interfaces. A number of studies have also used spoiled gradient echo acquisitions [48,82,90] which are generally faster and potentially allow for a 3D acquisition [73,90], rather than the more common multislice 2D techniques. 3D acquisitions allow the acquisition of multiple thin contiguous non-overlapping slices with better SNR per unit acquisition time than the equivalent non-interleaved multislice 2D acquisition.

Obtaining images at sufficient resolution for multiple sustained sounds is time consuming. For example, Story et al. [62] imaged 22 sustained sounds using axial 5 mm slices at 0.9 × 0.9 mm² resolution to cover the whole vocal tract in one subject, who spent between 7 and 8 h in the scanner over several sessions. Other work [86] has imaged the vocal tract in multiple orientations to produce slices perpendicular to the vocal tract over the maximum length, at the expense of even longer acquisition durations. However, compressed sensing [91] was recently used to reconstruct 3D gradient echo acquisitions of the vocal tract that were obtained with 5 fold undersampling [92]. Images at 1.5 × 1.5 × 2.0 mm³ over a 240 × 240 × 100 mm³ field of view (FOV) were acquired in a time of 7 s. Such studies provide information used in speech synthesis and increase the understanding of speech production. However, there are few if any clinical applications and, unless techniques like compressed sensing are employed, any potential future clinical application is limited by the long acquisition durations required.

Imaging the palate and associated musculature

The soft palate is one of the most frequently studied speech articulators using MRI. The soft-tissue contrast enables non-invasive imaging of the muscular anatomy, providing information which is otherwise only available via surgical or cadaveric techniques. As a result, MRI derived information has been used to describe the normal soft palate and could be used on a per subject basis to guide cleft palate surgery. The levator veli palatini (LVP) muscle (Fig. 2) produces the primary elevating force for the soft-palate and much of the static imaging work in MRI of the soft-palate has been focussed here. The LVP forms a sling, from the lateral attachments to the temporal bone and cartilage surrounding the auditory tube, through the soft palate. In normal subjects [93] MRI data has demonstrated that the LVP is, on average, orientated at 122.4° (male 122.0°, female 122.8°) and the muscle is best visualised from images orientated in this plane (see Fig. 4). Most studies have used multi-slice 2D imaging in the LVP plane, but there is no data on how the angulation varies between subjects and therefore, whether subject specific orientation of LVP imaging planes is required or whether imaging at 122.4° is sufficient in every subject. Several studies in patients [42,45,94] have used multi-slice 2D axial acquisitions or subject specific orientation of imaging which enable the clinical team to determine the morphology of the abnormal LVP in cleft palate patients. Alternatively, recent work [95] used a 3D acquisition with isotropic resolution and a 3D Turbo Spin Echo (TSE) with variable flip angle (SPACE) imaging sequence [96]. The isotropic 3D imaging volume was positioned axially and oblique imaging planes were retrospectively calculated, thus avoiding the requirement to identify the plane of the LVP at acquisition time. Such acquisitions would also be more amenable to reformatting in alternative planes than 2D multi-slice acquisitions, to study the other muscular anatomy of the soft palate. However, swallowing or other head motion often degrades the quality of LVP studies and the nature and severity of the resultant image artefacts are partly dependent on the k-space trajectory. 2D studies are often performed with interleaved slice ordering for temporal efficiency, so motion can result in blurring or mis-registration of a subset of slices. For 3D acquisitions, every reconstructed slice is partly dependent on the data from every k-space location and, therefore, the whole dataset could be ruined by a relatively short period of motion. If, on the other hand, motion only occurs at the outer k-space positions, the severity of the artefact might be very minor. 3D studies are also generally SNR efficient, but studies of the LVP are usually proton density for optimal contrast [93,94] or T2 weighted [43,97–99] (although the use of other contrasts have been described [49,93]) requiring long repetition (TR) times. For 2D studies, the long TR is used to acquire other slices, but for 3D imaging long echo trains must be used to maintain efficiency.

Images of the LVP can also be obtained during sustained phonation, although spatial resolution is compromised to account for the limited acquisition duration. While static imaging has been performed at up to 0.65 × 0.65 mm² in-plane resolution and 2.2 mm thick slices [43] or 0.8 mm isotropic resolution with 3D imaging [95], images of the LVP during sustained phonation have typical spatial resolutions of around 0.65 × 1.0 mm² for 4 mm slices [49].

A number of measurements of the static LVP are often performed, including; length, distance between points of origin, thickness and angle at the points of origin. However, none of these measurements has, as yet, been shown to be a useful measure/
predictor of velopharyngeal function. Furthermore, while a number of studies have measured such parameters, the number of participants is generally small (4–12) and the majority of studies fail to provide estimates of the between subject variance on most variables [49,93,97]. Tian et al. [43] acquired high resolution images in 17 normal subjects in the plane of the LVP during quiet breathing and images during sustained phonation in the same oblique plane and in the axial plane. Five variables including: maximum effective velopharyngeal ratio, maximum pharyngeal constriction ratio, maximum levator shortening ratio, maximum levator stretch ratio, maximum velar height, derived from measured parameters, were used to describe velopharyngeal function. These were presented with mean and standard deviation for the 17 healthy subjects imaged. In follow up work, the same group imaged 30 children (10 normal, 10 with repaired cleft palate and normal velopharyngeal function and 10 with repaired cleft palate and velopharyngeal insufficiency), during nasal breathing and sustained phonation [99]. Of the static measurements, the differences between normal and cleft subjects were larger than the differences between the repaired cleft patients with and without velopharyngeal insufficiencies. The distance between the points where the LVP inserts into the velum is narrower in non-cleft patients and the pharynx is less deep. The patients with residual velopharyngeal insufficiency had larger gaps between the uvula and posterior pharyngeal wall and shorter posterior vela (the distance from the centre of the LVP to the tip of the uvula). Of the dynamic variables, the maximum effective velopharyngeal ratio (maximum of distance from posterior of hard palate to centre of LVP/distance from posterior of hard palate to posterior pharyngeal wall) was significantly greater in the repaired cleft group with correct velopharyngeal function than in the group with velopharyngeal insufficiency. The maximum pharyngeal constriction ratio (1 – [minimal pharyngeal width at phonation/pharyngeal width at rest]) is significantly greater in the velopharyngeal insufficiency patients than in the non-cleft subjects.

Normative data to improve understanding of levator morphology in the presence of a cleft or other condition, was acquired in a sample of 30 subjects (15 males and 15 females) [100] using a high resolution 3D technique. Significant differences between men and women were observed: LVP muscle length, distance between levator insertion, points and angle of origin were all larger in men than that measured in women. However, variations in the relative size of the cranium or height of the individual were not proportionate to the variations observed in the levator muscle.

Other musculature in the velopharyngeal region has been identified in LVP MRI studies [97], but the potential of MRI to study these structures including palatoglossus and palatopharyngeus is yet to be fulfilled.

Dynamic imaging

Imaging at rest or during sustained phonation can only provide limited information on the vocal tract configuration during normal speech. The lack of ionising radiation, excellent soft tissue visualisation and high frame rates now available mean that MR is a powerful technique for the study of both normal and pathological speech. However, achieving the optimum balance in the trade off of SNR, temporal resolution, spatial resolution and subject tolerance is challenging. Studies have suggested that in order to capture the motion of the tongue [101] or soft-palate [102], a minimum temporal resolution of 20 fps is required, but very little work has considered the minimum spatial resolution required [103]. For evaluation of soft palate motion temporal resolution is prioritised. Using a spiral imaging sequence 22 fps at 1.9 × 1.9 mm² has been achieved [102,104], and using techniques widely available on clinical MR systems, 20 fps has been achieved at a reduced special resolution of 2.7 × 2.7 mm² [103].

Triggered acquisitions

High spatial and temporal resolution can be achieved by repeatedly acquiring a subset of the phase encode lines (a segment) required to create an image while the subject performs a speech task (see Fig. 5). The speech task is performed again and the next segment of k-space is repeatedly acquired. This is repeated until k-space is filled ([number of phase encode lines in image/number of lines per segment] times), resulting in an image series with a temporal resolution equal to the duration of the segment. In order to synchronise the production of the speech task with the

![Figure 5. Dynamic imaging methods. Dynamic imaging of speech can be performed using triggered methods (a), retrospectively reordering techniques (b) or real-time imaging (c). Triggered techniques repeatedly acquire a segment of k-space (a subset of the phase encode lines required for an image) while the subject produces a test phrase. The subject repeats the test phrase until enough segments have been acquired to fill k-space. The retrospective reordering technique of Shadle et al. [36] repeatedly acquires full images while the subject freely repeats the test phrase. The k-space data are retrospectively assigned to the imaging frames based on the simultaneously acquired audio data. Real-time imaging uses imaging acceleration techniques to acquire images rapidly while the subject produces the test phrase.](image)
acquisition a trigger signal is required. It is possible to generate a trigger signal from the electrocardiogram (ECG) monitor usually used for cardiac gated studies [40]. However, this means that the subject must synchronise the speech task with his or her heart rate and changes in the speed of the task are difficult to control. Minimal additional effort is required to produce a device which produces a trigger signal at a fixed, user defined repetition rate [37,38]. Such a device can be used to both trigger the scanner (usually via a simulated ECG R-wave) via the ECG interface and provide an audible or visual indicator to the subject.

Problems arise with these techniques when the speech task is not perfectly synchronised with the acquisition, which is more likely after a large number of repetitions of the speech task, or when the speech task is not perfectly reproducible. NessAiver et al. [105] analysed the timing of a repeated speech task and found differences in the onset of speech features of up to 95 ms. Many studies make little effort to minimise the number of repetitions and often acquire only 1 line of k-space per segment [38,39,90] resulting in acquisitions that require up to 256 repetitions of the speech task [39]. Typically cardiac cine imaging techniques acquire a number of lines per segment and sacrifice some signal to noise ratio for reduced imaging times using parallel imaging techniques [106,107]. Alternatively, one group [36,108] has developed a method of retrospectively reconstructing high temporal resolution gated images from a lower temporal resolution acquisition performed during free repetitions of the test phase. The raw k-space data is saved and the relative temporal position of each phase encode line is identified from the spectral characteristics of a simultaneously acquired audio recording. According to their relative temporal positions, the lines are rebinned to form images.

Several studies have also managed to acquire multi-slice data by acquiring additional repetitions of the test phrase accordingly. For example, Takemoto et al. [46] acquired 20 sagittal slices at 33 ms temporal resolution (1 × 2 × 6 mm³ spatial resolution) over 640 repetitions of the test phrase. The retrospective gating technique of Shadle et al. [36] is capable of similarly high spatio-temporal resolution with multi-slice imaging. They acquired three parallel sagittal slices at 1.9 × 1.9 × 5 mm³ spatial resolution and reconstructed to 16 ms temporal resolution using 288 repetitions of the test phrase.

Such techniques have been primarily applied to small groups (1–10) of healthy subjects for analysis of articulation during speech [36,47], including the calculation of area functions for acoustic modelling [40,46] and in multi-slice multi-planar acquisitions for monitoring glottal and tongue movement [39,90]. The same technique has been used for assessment of velopharyngeal function [37,38] and, in conjunction with functional MRI, to assess the brain 

...[flawed] gated techniques are currently the most straightforward way of acquiring dynamic vocal tract images at high temporal and spatial resolution. The highest published acquired temporal resolutions are around 17 ms for a single slice acquisition with a spatial resolution of approximately 1 × 2 × 6 mm³ spatial resolution [46,90]. The predominant limiting factor on temporal and spatial resolution is the subject’s ability to reproducibly repeat the test phrase. However, the minimum TR of the acquisition and the motion of the articulators during this time could also limit the effective temporal and spatial resolution.

Tagged acquisitions

Very few techniques can provide information on the internal deformation of soft tissue structures. Tagged MRI was developed for tracking deformation of the myocardium during the cardiac cycle [110–113]. Such techniques tag the tissue with a spatial variation in the magnetisation available for imaging. As the tissue moves, so do the tags and the deformation of the tags indicates the motion of the tissue. At a similar level the authors of [114,115] used tagged MRI to study the internal deformation of the oral structures, the tongue in particular [114,115]. These early studies used a non-selective inversion recovery technique with selective inversion of a number of planes perpendicular to the imaging plane. After this preparation phase, which is applied while the subject is in the rest position, the subject produces a sound or moves his/her tongue and holds the position. The imaging data is acquired while the subject maintains the position and there is increased image intensity in the planes containing the reinverted tissue. A more commonly used technique is spatial modulation of magnetisation or SPAMM [111], which introduces a sinusoidal variation in the magnetisation and, therefore, the image intensity, see Fig. 6. SPAMM is typically used with cine imaging to dynamically study the internal deformation of tissue. However, the intensity of the tags fades rapidly during the cine images due to longitudinal recovery of magnetisation (time constant T1). A variation of this technique, complementary spatial modulation of magnetisation (CSPAMM) increases tag intensity and reduces tag fade by acquiring two sets of SPAMM tagged images with alternating polarity in the tag pattern and subtracting the two images [105]. Clearly, CSPAMM requires twice as much data and therefore the acquisition duration is twice as long.

Tagged acquisitions are typically acquired as triggered acquisitions, where the tagging preparation is applied immediately after the trigger and the segmented image acquisition is performed afterwards during the speech task. The segmented nature of the acquisition means that temporal and spatial resolution depend on the number of repetitions of the test phrase and the quality of the resultant data depends on the reproducibility of the test phrase. In order to provide two dimensional motion information, either two tagging preparation modules are applied to form a grid like tag pattern, or two images are acquired with perpendicular tagging directions and the grid pattern is created by combining the images.

The imaging data can be acquired using a variety of sequences [112] including non-Cartesian and hybrid echo planar imaging techniques which have advantages for acquiring high temporal resolution data [116,117]. However, all tagged speech studies have used standard gradient echo techniques, such as spoiled fast low angle shot (FLASH) [105] or bSSFP.

Tag tracking can be performed by manual tracing [115], which is time consuming for cine acquisitions, or by automated techniques. A variety of automatic and semi-automatic tracking techniques exist [118–120] but the harmonic phase (HARP) technique is the most widespread. HARP uses SPAMM or CSPAMM data as an input. The sinusoidal variation in the image intensity is represented as an off-centre peak in k-space. Data outside a small region around the off-centre peak is nulled and a Fourier transform is then applied to the data. The magnitude of the resultant data is a low resolution anatomical image while the phase data varies linearly from 0 to 2π in the tagging direction and then wraps with the same period as the tags in the original data. A composite image is formed by multiplying the magnitude and phase of the HARP data and motion analysis is performed by automatically following small, easy to track regions of interest.

In the upper vocal tract, tagged MRI techniques have been used to study the internal deformation of the tongue, predominantly in the mid-sagittal plane. Initial studies performed tagging preparation, then included a delay for movement of the tongue before imaging while the subject maintained the desired tongue position [114,115,121,122]. The results were analysed with reference to the muscular anatomy of the tongue to provide indications of the possible muscles involved in producing Japanese vowels [114,115,121,123], later with electromyographic comparisons [123].
Figure 6. Images derived from tagged MRI data. Subtraction of the two CSPAMM datasets produces the high intensity sinusoidal variations in image intensity (a). HARP processing of the data results in the phase wrapped (b). Horizontally and vertically tagged images can be combined and processed to form a checkerboard like modulation in signal intensity (c). Adapted from Parthasarathy et al. [126], Copyright Acoustic Society of America.

Napadow et al. [122] used a SPAMM technique with a FLASH imaging sequence. Strain maps in the tongue were derived for generic static tongue positions in the mid-sagittal plane (anterior protrusion and sagittal tongue curl) and an axial slice through the tongue (lateral tongue bending). Tagged cine MRI using a segmented imaging technique was first used to analyse tongue deformation during production of the sound /ka/in one example subject [124]. This work used a triggered acquisition in the mid-sagittal plane and manual tracking of tags. The muscles involved in the phonation /k/ and then /a/ were inferred from the regional strains calculated from the images, although the technique cannot rule out passive contraction of a particular muscle caused by a neighbouring fibre bundle. Liu et al. [125] used HARP analysis of multiple sagittal and axial images to derive 3D tongue motion. Parthasarathy et al. [126] applied HARP processing to CSPAMM cine tagged MRI of the tongue acquired in three perpendicular planes. Two images (one horizontally tagged and one vertically) for each of 26 slices were acquired, while the single subject repeated /disuk/ 416 times. The automated nature of HARP analysis meant that the trajectories of individual points, strains, principal strains and velocity fields could be calculated for every frame and every slice. Stone et al. [127] imaged 8 subjects using a CSPAMM tagged technique while the subjects said /i/ then /u/ and processed the data using HARP. The 8 datasets included 3 acquired in the same subject, three different native languages and one patient who had a previous glossectomy. Principal component analysis was used to represent tongue motion as a mean velocity field with dataset–dataset variations. The results demonstrated large intra-subject differences in the same test phrase and similarities in vowel production between a patient with a glossectomy and a native Japanese speaker, and between native Tamil and American English speakers.

Real-time acquisitions

Real-time imaging techniques

In contrast to triggered acquisitions, real-time imaging does not require reproducible and synchronised repeated speech tasks, but achieving sufficient temporal and/or spatial resolution with adequate SNR is even more challenging. A variety of imaging sequences have been used which have various strengths and weaknesses. Sequences widely available on clinical MRI scanners typically achieve low temporal resolutions (2–10 fps), but are easily implemented and well understood. TSE sequences with the zoom [128] and partial Fourier techniques have been used to achieve around 6 fps [65,88,129,130]. The zoom technique uses perpendicular excitation and refocusing pulses for the spin echo to excite a reduced FOV in the phase encoding direction, therefore saving time when imaging. The partial Fourier technique [131] acquires only slightly more than half of the phase encode lines and uses the conjugate symmetry of k-space to recover the rest of the data. Gradient echo sequences have also been used to achieve typically low temporal resolutions (2–10 fps [33,132–134]), although higher temporal resolutions are common in cardiac MRI studies [135,136]. Rapid gradient echo acquisitions can be subdivided depending on how the remaining transverse magnetisation is dealt with at the end of each repeat time (TR). Radiofrequency (RF) spoiled gradient echo techniques or FLASH (fast low angle shot) type sequences [137], discard the remaining transverse magnetisation by modifying the phase of every excitation pulse. Alternatively, without RF spoiling the remaining transverse magnetisation may be used in the next TR to improve SNR. This variant, which we will refer to as steady state free precession (SSFP) [138], requires a constant integral of the gradient waveform on each axis in each TR. While the SNR of the resultant images is increased with respect to the FLASH images, motion during imaging results in signal loss. The balanced SSFP technique (bSSFP) [139,140] uses additional gradients to fully refocus the magnetisation in each TR, further improving SNR. Such sequences are motion resilient and very SNR efficient, but are sensitive to inhomogeneities in the magnetic field caused by imperfections in the magnet or the presence of objects in the field, e.g. a patient. Both FLASH [133,141,142], and bSSFP [33,59,143] acquisitions have been used for real-time imaging of the vocal tract in the mid-sagittal plane. While FLASH acquisitions may be slightly faster, their lower relative SNR may mean that more data must be acquired to compensate.

The highest temporal resolutions have been achieved by using non-Cartesian techniques which sample k-space in a non-rectilinear fashion. The most commonly used non-Cartesian trajectories are radial and spiral, which are most often employed as spoiled (FLASH like) sequences. Spiral trajectories can be time efficient as a relatively large proportion of k-space can be sampled in each TR. However, the images are prone to artefacts [144] including due to the presence of off-resonant spins, caused by magnetic field inhomogeneity or chemical shift (i.e. fat). The artefacts appear as a blurring of the off-resonant tissue in all directions and become more severe with longer spiral trajectories. Methods do exist for reducing [145] the image degradation, but such techniques make the already non-standard reconstruction more complicated and are, therefore, typically only used in speech studies at 3 T, where the off-resonant effects are more common. Radial acquisitions [146] are less sensitive to off-resonance artefacts but are inherently inefficient. However, radial acquisitions are usually undersampled, meaning that an insufficient number of k-space profiles are acquired to fulfill the Nyquist criteria at the k-space periphery. In Cartesian imaging, undersampling results in replicates of the imaged object appearing within the imaged field.
view. In contrast, undersampled radial imaging results in streaking artefacts, which are usually considered less detrimental to diagnostic image quality [147,148]. Radial acquisitions are also relatively motion insensitive as the centre of k-space is heavily oversampled. As for spiral imaging, radial trajectories are primarily confined to research institutions and are not typically available for use on standard clinical MRI scanners found in general radiology departments.

**Application of clinically available real-time techniques**

Despite the relatively low temporal resolutions available, a number of dynamic speech studies have been performed using clinically available real-time techniques. As described above, TSE zoom imaging with partial Fourier acquisition has been used for imaging the vocal tract at low spatiotemporal resolution (2 × 4 × 6 mm³ at 4–6 fps) [65]. Alternatively, the same technique has been used to study velopharyngeal closure [129] at a resolution of 1.5 × 3 × 6 mm³ at 5–6 fps, with imaging performed in the mid-sagittal plane, the axial plane at the height of maximal velopharyngeal closure and in the coronal plane. In a comparison with X-ray videofluoroscopy in 8 subjects there was complete agreement in the pattern of velopharyngeal closure.

The earliest real-time gradient echo studies in the vocal tract [149] used FLASH sequences, had poor temporal resolution (3 fps), low spatial resolution (3 × 6 × 8 mm³) and were used to study vocal tract configurations during sustained vowel production in the mid-sagittal plane and axial slices. While dynamic imaging during sustained phonation avoids blurring and artefacts caused by a failure to completely maintain the phonation during a static acquisition, it does not fully represent normal speech. Higher spatio-temporal resolutions were made possible with the introduction of parallel imaging techniques (SENSE or GRAPPA) and improved scanner hardware [132,141]. Parallel imaging techniques acquire undersampled data with multiple receive coils, each with a spatially varying sensitivity profile [106,107]. The data from each of the coils with information on the coils sensitivity profile is used to resolve the aliasing. Echtenach et al. [132,150–152] have achieved around 8 fps at 1.4 × 2.2 × 11 mm³ using GRAPPA [153] and FLASH gradient echo imaging for evaluating vocal tract configurations in opera singers during register changes. Alternatively, the additional performance obtained using parallel imaging has been used to improve spatial resolution (1 × 1 × 6 mm³ at 2 fps) in studies assessing velopharyngeal function in children with velopharyngeal insufficiency [34,142]. Drissi et al. [33] also used a higher spatial resolution at the same temporal resolution (2 fps) with a bSSFP sequence for assessment of velopharyngeal function in children. However, it is debatable whether such high in-plane spatial resolution is required to determine whether velopharyngeal closure occurs. For assessment of velopharyngeal closure during normal speech, higher temporal resolution is certainly required and the motion blurring at 2 fps is likely to degrade the effective spatial resolution. One potential method for improving spatial or temporal resolution in real-time soft palate imaging is to use adaptive averaging [154]. Increased temporal or spatial resolution can be traded for reduced SNR which is retrospectively recovered. The adaptive averaging algorithm identifies frames where the soft palate is in a similar configuration and selectively averages these frames to improve SNR. This technique is demonstrated in the Supplementary data file 3.

A few studies have used bSSFP techniques to acquire real-time vocal tract images for swallowing [143], assessment of velopharyngeal function [33,103] or assessment of vocal cord configuration [59]. For assessment of velopharyngeal function, Scott et al. [103] compared sequences with spatial resolutions between 1.6 × 1.6 × 10 mm³ and 2.7 × 2.7 × 10 mm³ for temporal resolutions between 9 and 20 fps using SENSE parallel imaging and partial Fourier at both 1.5 T (bSSFP) and 3 T (SSFp). bSSFP sequences were too sensitive to off resonance effects to use at 3 T, but SSFP sequences produced reliably good images. Whereas, bSSFP failed in some cases at 1.5 T, but often produced very good images. For assessment of palate motion the highest temporal resolution was preferred while when assessment of morphology is required some temporal resolution was traded for increased spatial resolution. Examples of acquisitions at 1.5 T (10 fps) and 3 T (20 fps) are shown in the Supplementary data files (files 1 & 2). The only other published comparison of techniques for real-time imaging of the vocal tract compared hybrid echo-planar imaging sequences (hEPI) to FLASH acquisitions for assessment of swallowing [134]. While the hEPI sequence was rejected for use in imaging swallowing due to artefacts, by using a longer echo train length (~11 c.f. 3 in Angnosta et al. [134]) with SENSE and partial Fourier techniques on a modern 1.5 T scanner, other studies [154] were able to obtain high temporal resolution with few artefacts when imaging the soft palate. Images acquired with these methods are shown in the Supplementary data file 3.

**Applications of non-Cartesian imaging techniques**

Initial studies [101,155–157] using spiral imaging in the upper vocal tract obtained similar acquired spatial and temporal resolutions to that possible with Cartesian techniques (2.7 × 2.7 × 5 mm³ at 9 fps). However, as multiple spiral trajectories (each rotated with respect to the last) are required to form a complete image, higher frame rate images can be reconstructed using a sliding window technique see Fig. 7. In the initial spiral work, mid-sagittal images of the vocal tract during speech were acquired at 9 fps and reconstructed using a sliding window at 24 fps [101], example videos acquired using this technique are available at [http://sail.usc.edu/production/rtmr/jasa2004]. This protocol was also successfully applied used in a real time study of sound production by male beatbox artist [158], movie files of different sounds are available at [http://sail.usc.edu/span/beatboxing/]. While each sliding window image has the temporal resolution of the original acquisition, the time difference between the start of each frame can be much less. A spiral-sliding window technique [41] was also used to study the timing differences between motion of the tongue tip and soft palate from mid-sagittal images. Timing differences were demonstrated between the movement of the tongue and the palate, the duration of which was dependent on the position of a nasal consonant sound, [n] [159], in the word and the location of the stress in the word. In another study [160], a similar MRI protocol was applied to investigate articulatory setting in speech production at 22.4 fps. The study was performed on five healthy volunteers and showed significant differences between vocal tract postures adopted during inter-speech pauses and those at absolute rest before speech.

In order to acquire spiral data at higher temporal resolution in the mid-sagittal plane, studies have used a reduced FOV [104]. Signal in the brain and posterior section of the neck is eliminated using saturation bands, preventing it from aliasing back into the reduced FOV. The data was acquired at 3 T, where signal to noise ratio is higher and an alternating TE was used to provide data for field inhomogeneity corrections. The average acquired frame rate was 21 fps (1.9 × 1.9 × 6.5 mm³ spatial resolution), including the time for the saturation pulses, and reconstructed to 30 fps. This technique was used to investigate the influence of the velopharyngeal mechanism on the acoustic characteristics of a nasalised consonant in a dynamic setting. An alternative way of accelerating spiral acquisitions is to use parallel imaging, which is more complex than for Cartesian data. Kim et al. [161] used a golden angle spiral trajectory [162], which results in interleaves with an almost equal radial spacing for any
number of interleaves. Images at multiple temporal resolutions were reconstructed, albeit with lower SNR for higher frame rates. The final combined dataset had higher temporal resolution in more dynamic regions and lower temporal resolution, but better SNR in stationary areas. In another study using accelerated radial imaging Niebergall et al. [163] have achieved 30 fps at $1.5 \times 1.5 \times 10 \, \text{mm}^3$ resolution for imaging the vocal tract in the mid-sagittal plane and the larynx in the coronal plane. An example from this study is shown in the Supplementary data (file 4). In contrast to regular parallel imaging techniques, the coil sensitivity information was calculated iteratively during the parallel imaging reconstruction [164]. Temporal and spatial filters were applied retrospectively [28]. They were able to demonstrate differences in tongue contours between production of vowel sounds individually, within a single word and within in a sentence. The results also demonstrated that production of vowel sounds individually, within a single word and within in a sentence. The results also demonstrated that the actual spoken sounds and audio—video synchronisation. However, the loud volume of the MR scanner during imaging and the presence of strong and time varying magnetic fields, with the requirement to isolate the MRI receive equipment from electromagnetic interference, means that this is challenging. Standard electrical microphones are generally unsuitable for these reasons, although electrical microphones were used in early work at low field [84].

Other recent work has focussed on obtaining real-time speech data in multiple orientations [163,165,166]. This has been achieved using three perpendicular slices, acquired as interleaved radial trajectories, resulting in a frame rate of 156 ms at $2.4 \times 2.4 \times 6 \, \text{mm}^3$. A real-time acquisition of three planes from Ref. [165] is shown in the Supplementary data (file 5). A reduced FOV ($20 \times 20 \, \text{cm}^2$) was enabled by using a custom receive coil (Fig. 8) and SNR is boosted because the interleaved slice ordering results in a longer effective TR and, therefore, increased recovery of longitudinal magnetisation. This was used to show temporal changes in tongue grooving during natural speech. Similar temporal and spatial resolutions with perpendicular or even parallel slices could be achieved using segmented Cartesian techniques, hybrid EPI for example, with parallel imaging. Alternatively, promising initial results demonstrating high frame rates (20 fps, $2.2 \times 2.2 \times 8 \, \text{mm}^3$) with 5 parallel sagittal slices, obtained using an interleaved spiral—Cartesian trajectory and an iterative reconstruction were recently presented [167]. Another group [168] used audio information to align 2D dynamic data into 3D dynamic movies of vocal tract shaping. This technique allowed reconstruction of 3D vocal tract movies with $2.4 \times 2.4 \times 3 \, \text{mm}^3$ spatial resolution and 78 ms temporal resolution. The same group [166] is working towards a true real-time 3D technique using a stack of spiral trajectory and have achieved 17 sagittal images (or kx phase encode steps) at $2.4 \times 2.4 \times 3.0 \, \text{mm}^3$ resolution for a temporal resolution of 1 s. Substantial acceleration is expected with the inclusion of parallel imaging, and highly undersampled trajectories. Model based methods enable reconstruction from heavily undersampled data based on some necessary assumptions. Such an approach was applied to dynamic 3D vocal tract imaging [169] using an undersampled stack of spirals imaging sequence and reconstruction based on a partially separable model [170]. The resultant images had a temporal resolution of 8.6 fps (roughly 8× acceleration), and an effective spatial resolution of $3 \times 3 \times 3.1 \, \text{mm}^3$.

**Other considerations**

**Acoustic recording**

Simultaneous audio recorded during dynamic acquisitions allows verification of the correct pronunciation, acoustic analysis on the actual spoken sounds and audio—video synchronisation. However, the loud volume of the MR scanner during imaging and the presence of strong and time varying magnetic fields, with the requirement to isolate the MRI receive equipment from electromagnetic interference, means that this is challenging. Standard electrical microphones are generally unsuitable for these reasons, although electrical microphones were used in early work at low field [84].

As a result of the difficulties in recording sound during imaging, many studies, particularly for sustained phonation [32,46,48,60,69,78,124] have recorded sound after imaging outside the scanner. This can provide high quality recordings for acoustic analysis, but the recordings are difficult to synchronise with the imaging data and the phonation may vary.

It is standard for MRI scanners to be provided with an intercom system for communication with the patient between acquisitions. However, such systems are not designed for patient communication during imaging and it is difficult to hear the patient speak during many imaging sequences. Despite this, some of the earliest work recorded speech via the intercom [108,171] but results are often unsatisfactory [40]. Another workaround solution is to transfer the sound to a microphone in the scanner control room using plastic tubing and a cup held in front of the subject [34].

Many recent real-time studies use optical microphone based systems, which relay the audio signal from the scanner bore to the control room using fibre optics. The microphone itself consists of two optical fibres and a membrane. The first fibre carries a source light signal into the microphone where a membrane modulates the intensity of the light, which is carried back out of the scanner in the second fibre. The intensity of the light returned is dependent on the displacement of the membrane. A light source and detector/converter are located in the control room. The microphone can be constructed to have a figure of eight type spatial response, so sound entering from both the front and back is significantly attenuated.
whereas sound from only one direction is not. Such devices have been used alone [142], but more commonly with adaptive noise cancellation algorithms [105] to reduce the residual scanner noise. More advanced systems [104,155] use a second optical fibre microphone to record a reference audio signal which is used to reduce the scanner noise.

The high audible noise levels during imaging also mean that subjects may find it difficult to hear their own speech. This can lead to louder and, therefore, unnatural speech. To reduce the effects of this, studies of sustained phonation have played target sounds during imaging [84] and other work has provided audio feedback of the subject’s own voice [132]. Although, in the latter case, the latency of any noise cancellation algorithms used must be considered.

Synchronisation

As there is a finite reconstruction time for each MRI image and a latency associated with many noise cancellation algorithms, simply capturing the video stream from the scanner graphics card with the sound from a fibre optic microphone system is usually inadequate for analysis. Synchronising the recorded audio with video created from the MRI images requires knowledge of the time each of the images was acquired and the timing of the recorded audio. Synchronisation of real-time images with the audio can be performed using the audible start and end time of the scanner noise with accurate information on the sequence timings, from a scanner simulator [104]. Otherwise, it may be possible to record a timing trigger signal from the scanner corresponding to known positions in the image acquisition [155], which can be converted into an audio trigger and recorded simultaneously with the speech audio in a separate channel [154] or used to trigger the start of the audio recording [172]. Automatic synchronisation on the scanner host computer has been performed using custom software written to run on the host and reconstruction computers [142,173].

Receive coil selection

The receive coil used for imaging has an important role in determining the imaging field of view and has a large influence on the image SNR. Phased array coils are standard and are a requirement for parallel imaging techniques. Unless there is a specific reason to use a birdcage coil (to enable transmit/receive for example [104]) there is little reason not to use a phased array coil. While the coverage of a standard head coil may not extend inferiorly enough to cover the vocal tract, except in children [33], head and neck or neurovascular arrays usually cover this region and are often the receive coil of choice for the vocal tract [103,132,142]. However, using a coil with a more localised sensitivity profile can increase SNR or allow for a reduced FOV and, therefore, faster imaging. Commercially available coils for carotid artery imaging have been used [59] to image the larynx, but have a low penetration depth so are not suitable for imaging the other deeper structures of the vocal tract. In order to improve sensitivity, combination of a small flexible phase array coil was used with a bilateral array coil placed on both sides of the neck. In fact the fastest real-time acquisitions [163] were obtained with similar coils.

In order to target specific regions of the vocal tract, purpose built receive coils have been used. Dedicated phased array coils have been developed for larynx imaging as an alternative to using carotid artery coils [39,58]; they cover only the relevant area and can be optimised for imaging at the depth of the larynx. Dedicated coils have also been developed for imaging the oral and velopharyngeal regions of the upper vocal tract [60,173,174]. While standard head or head and neck coils encompass large volumes and have coil elements distributed around this volume, the custom coils tend to only cover the front of the head from the nose to the upper-neck. An example of such a coil designed by Kim et al. [174] and comprising 16-coil elements is illustrated in Fig. 8. It allows for an increased SNR when imaging the upper airways (see Fig. 8c) of between 1.5 (posterior pharyngeal wall) and 8.8 (lower lip) for unaccelerated imaging when compared to a neurovascular array coil. However, such specialised coils are not commercially available and must be designed and built in-house, requiring collaboration from the scanner manufacturer.

The effects of supine patient positioning

The large majority of MRI scanners are the horizontal bore type and the patient lies positioned supine for vocal tract imaging. While this scanner arrangement allows for high field strengths (typically 1.5–7 T), the supine patient position means that gravity acts in the anterior–posterior direction rather than superior–inferior direction when the patient is upright. The effect of supine positioning on

Figure 8. A 16-channel receive coil array for accelerated upper airway MRI at 3 T. (a) Photograph of the 16-channel coil array. The coil array can be held close to the face (b) or folded up and back to permit patient entry and exit. (c) Illustration of coil sensitivity and the upper airway regions of interest (ROIs) used in the evaluation of SNR (1 — upper lip, 2 — lower lip, 3 — front tongue, 4 — mid tongue, 5 — back tongue, 6 — palate, 7 — velum, and 8 — pharyngeal wall). Adapted from Kim et al. [174], Copyright Wiley-Liss, Inc.
the positioning of the speech articulators is subject specific and, therefore, difficult to generalise [175,176]. Studies using ultrasound [175], electromagnetometry [177], X-ray microbeam [178] horizontal bore [130] and open MRI systems [176] have suggested that the predominant tongue motion is a displacement of the tongue body towards the posterior pharyngeal wall in the supine position, although this is not universally true. This gravity induced displacement can be described by a subject-specific rigid body translation [175]. However, there appear to be compensatory mechanisms for protecting the airway patency and allowing normal speech. Acoustic analysis of supine vs. upright vowel sounds [175], found little difference between the formants produced and the tongue motion for specific sounds was similar. A comparison of velar positioning and LVP geometry during phona-

Conclusions: current research, future directions and applications

Despite an increasing interest in the field it is undeniable that MRI has yet to fulfill its full potential in the study of speech. One of the current limitations is the limited coverage in the through-plane direction (typically 1–3 slices) coupled with a temporal resolution that is still too low for many clinical applications. Much of the current research focuses on addressing those issues. For example, a recent novel method for rapid tracking of soft palate motion during speech using navigators was demonstrated for the first time [179]. Pencil beam navigators achieved 37 lines per second and a novel turbo navigator technique 62 lines per second; both exceeding current maximum 2D real-time imaging frame rates. Other research focuses on exploiting spatial–temporal correlations within the data to accelerate dynamic MRI by use of undersampled acquisition schemes and model-based reconstructions. Techniques such as k–t SENSE [180] and k–t FOCUSS [181] have been demonstrated in the vocal tract, the latter appearing to be very promising for speech imaging in clinical practice [182].

Recently, attempts have been made to provide 3D visualisation of the whole vocal tract in real time by applying model based reconstructions [170] to highly undersampled stack of spirals data [169] or via a 3D volumetric navigation strategy which can achieve 100 fps [183].

The acquisition of a larger amount of data, necessary at high temporal resolutions, will inevitably require the use of faster computing as the images will need to be reconstructed on-the-fly to become an integral part of a clinical examination. Graphics processing unit (GPU) computing could provide an answer to this issue and GPU computing has already shown promising results for computationally intensive reconstructions in cardiac MRI [180,184–187].

Despite current hardware and software limitations, speech MRI is advancing in the clinical field. In a recent study using clinical available techniques [188], both children and adults were imaged and a full speech examination was carried out, lead by a speech therapist. However, audio–video synchronisation was performed off-line and a better integration with the current scanner software would require an intensive collaboration with the manufacturer.

One key advantage of using MRI in clinical practice would be to combine anatomical and dynamic information. However, a large number of patients requiring speech assessment have orthodontic devices that could potentially render real-time speech examination difficult or impossible [189]. This highlights the need for an integrated approach if MRI becomes a routine part of the management of those patients. Furthermore, the majority of patients with speech disorders are children of a young age. In our experience, although children tolerate the MRI examination nearly as well as adults, efforts have to be made to complete a full examination in a relatively short period of time, ideally less than half an hour. Motion artefacts are more prevalent in younger patients and it would be valuable to integrate motion-correction schemes such as that described by Anderson et al. [190] in the acquisition. Moving towards true 3D acquisitions would also limit the time required for planning. This is important as the potential range of clinical applications for real-time vocal tract MR imaging is increasing. Apart from the cleft palate, dynamic vocal tract imaging in sleep apnoea [191] and Parkinsonism [192] have been reported.

Finally, beyond clinical applications there is large area of speech and language research [193,194] where the addition of real-time speech MRI to fMRI studies would provide valuable information. Such studies may provide, for example, a better understanding of the mechanisms of phonation; the processes involved in learning a second language; or the mechanisms of singing [159].

Acknowledgements

ADS is currently funded by the Royal Brompton Hospital National Institute of Health Research Cardiovascular Biomedical Research Unit. MEM is partly (20%) funded by the Barts and The London National Institute for Health Research Cardiovascular Biomedical Research Unit. We are grateful to Andreia Freitas for proof-reading the final version of the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ejmp.2014.05.001.

References

[9] Ludlow C, Connor N, Bassich C. Speech timing in Parkinson’s and Hunting-


