Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 52 (2018) 1–22

www.elsevier.com/locate/csl

Analysis of speech production real-time MRI☆

Vikram Ramanarayanan[a,b,*], Sam Tilsen[c], Michael Proctor[d], Johannes Töger[e], Louis Goldstein[f], Krishna S. Nayak[f], Shrikanth Narayanan[f]

[a] *Educational Testing Service R&D, San Francisco, CA, United States*
[b] *University of California, San Francisco, CA, United States*
[c] *Cornell University, Ithaca, New York*
[d] *Macquarie University, New South Wales, Australia*
[e] *Lund University, Lund, Sweden*
[f] *University of Southern California, Los Angeles, CA, United States*

Received 11 July 2017; received in revised form 19 February 2018; accepted 11 April 2018
Available online xxx

Abstract

Recent advances in real-time magnetic resonance imaging (RT-MRI) have made it possible to study the anatomy and dynamic motion of the vocal tract during speech production with great detail. The abundance of rich data on speech articulation provided by medical imaging techniques affords new opportunities for speech science, linguistics, clinical and technological research and application development, but also presents new challenges in audio–video data analysis and data modeling. We review techniques used in analysis of articulatory data acquired using RT-MRI, and assess the utility of different approaches for different types of data and research goals.

Keywords: Real-time magnetic resonance imaging; Speech production; Region of interest; Vocal tract; Medical image processing; Speech science

1. Introduction

Detailed articulatory data are a critical source of information about human speech production, its biomechanical properties, and linguistic underpinnings. However, a perennial challenge has been access to realistic and useful articulatory data. Techniques that have been used to measure speech articulation (summarized in Table 1) include X-ray microbeam (XRMB: Westbury et al., 1990), electropalatography (EPG: Hardcastle, 1972), electromagnetic articulography (EMA: Perkell et al., 1992; Wrench, 2000) and ultrasound (Stone and Davis, 1995; Whalen et al., 2005). Although some of these techniques are invasive, they are able to capture articulatory information at high sampling rates to varying degrees. However, none of these modalities provide a complete view of all vocal tract articulators, which is important for studying vocal tract posture.

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.
* Corresponding author at: University of California, San Francisco, CA, United States.
E-mail address: vramanarayanan@ets.org (V. Ramanarayanan).

https://doi.org/10.1016/j.csl.2018.04.002
0885-2308/

Table 1
Articulatory measurement techniques.

Characteristic	XRMB	EMA	Ultrasound	EPG	RT-MRI
Order of typical sampling rate (Hz)	100	500	50–300	100	5 to > 100
Relative spatial resolution	Low	Low	Medium	High	High
View of vocal tract	Fleshpoints	Fleshpoints	Tongue	Tongue–palate contact	Full view
Supine position?	No	No	No	No	Yes
Invasive?	Yes	Yes	No	Yes	No
Example database (with citation)	Wisconsin X-ray microbeam database (Westbury et al., 1990)	Edinburgh MOCHA database (Wrench, 2000)	Haskins HOCUS (Whalen et al., 2005)	Edinburgh MOCHA database (Wrench, 2000)	USC MRI–TIMIT database (Narayanan et al., 2014)

More recently, developments in real-time magnetic resonance imaging (henceforth, RT-MRI) have enabled examination of shaping along the entirety of the vocal tract during speech production, providing a means for observing and quantifying the 'choreography' of the articulators, in space and time, including structural/morphological characteristics of speakers in conjunction with their articulation dynamics and acoustics (Narayanan et al., 2004). While RT-MRI typically has an intrinsically lower frame rate than the other modalities, recent advances in parallel imaging and sparse reconstruction have helped to significantly improve the temporal resolution of acquired data including multiplane imaging (Iltis et al., 2015; Fu et al., 2015; 2017; Lingala et al., 2016). Importantly, RT-MRI offers a clear advantage over other methods with respect to patient safety, relative non-invasiveness of imaging, and the ability to image in 3D or simultaneously in multiple arbitrary 2D planes. On the other hand, MRI is typically more expensive and less accessible – especially for field studies – compared to other sensing modalities. Another consideration in studies using RT-MRI is the effect of gravity due to the supine position subjects assume in order to be scanned using MRI (Subtelny et al., 1972; Engwall, 2003). However, in an X-ray microbeam study of two Japanese subjects, Tiede et al. (2000) concluded that while the effects of the supine posture were significant for sustained vowel production, they were minimal for running speech production.

Data collected using RT-MRI can be used to inform important questions in linguistic theory, speech modeling and clinical research. Many advances in RT-MRI spatial and temporal resolution have been driven by the need to investigate phonetic and phonological phenomena, such as vowel nasalization in Portuguese (Teixeira et al., 2012) and French (Carignan et al., 2015), liquid consonant behavior in English (Proctor and Walker, 2012) and Korean (Lee et al., 2015), coarticulation in VCV sequences (Demolin et al., 2002), and characterization of click consonants in African languages (Proctor et al., 2016). RT-MRI has also been used to investigate vocal tract shaping during non-speech events, such as those observed during beatboxing (Proctor et al., 2013a) and singing (Burdumy, 2016). Such research is also important for automatic speech and speaker recognition technologies, as any speech or speaker modeling procedure must reflect the structure of the underlying physical system in order to be effective (Rose et al., 1996). For example, automatic speech recognition (ASR) can benefit from knowledge of the coordination of the vocal tract articulators and the resulting acoustics; this can help reduce apparent token-to-token variability, removing a confounding factor in general pattern recognition algorithms (Frankel and King, 2001; Mcdermott and Nakamura, 2006; McGowan, 1994). In addition, speakers exhibit substantial differences in many aspects of their individual vocal tract morphology, all of which have the potential to alter acoustic output or force speakers to adjust their articulation in compensation. Incorporating such knowledge into speaker modeling could likewise improve speaker recognition performance. Finally, RT-MRI also allows researchers to examine important clinical questions regarding disordered vocal processes in patients, including speech and swallowing. Clinical uses of RT-MRI include imaging patients who have undergone glossectomy (partial removal of the tongue) to treat oral cancer (Mády et al., 2003), and in patients with apraxia of speech (Hagedorn et al., 2014) or sleep apnea (Drissi et al., 2011).

Relatively rich RT-MRI datasets, such as the USC MRI-TIMIT database (Narayanan et al., 2014), allow researchers to study a wide range of the aforementioned linguistic, speech and clinical phenomena in a systematic manner. However, datasets collected using RT-MRI also pose significant recording, processing and analysis challenges (Silva and Teixeira, 2016). For instance, it is non-trivial to collect the amounts of data that are required for answering different linguistic and modeling questions in a statistically meaningful manner, due to the labor, cost and expertise involved. This affects the reproducibility of the observations and analyses across different subjects and demographic backgrounds, which in turn has implications for the inferences made using this data, both by humans and machine

algorithms. In many applications, since we largely rely on experts (physicians, radiologists) to interpret and analyze such imaging data (as with other human-centric medical datasets), there are generally no gold standard datasets or annotations of the data, which means that there is no accepted standard benchmark for evaluating performance of automated analysis algorithms. The large number of moving parts in the data collection chain from acquisition to reconstruction to analysis to interpretation also pose challenges to automation. Silva and Teixeira (2016) identify key steps in this workflow, and address further considerations for quantifying and comparing data across speakers.

1.1. A Taxonomy for RT-MRI analysis

This review paper describes the wide variety of RT-MRI analysis techniques used for speech research, and presents a unified taxonomy within which these techniques can be understood and extended. The taxonomy, schematically represented in Fig. 1, is organized according to the broad class of image processing used, type of output information obtained, and auxiliary or prior information required for each method. Analysis techniques are also classified according to difficulty of implementing each method of processing, and the level of abstractness of the output representation. Within such a taxonomy, we can broadly define four classes of image analysis techniques: those based on (i) *basis decomposition or matrix factorization techniques* at the level of the raw or processed images, (ii) *pixel-* or *region-of-interest (ROI)-based*, (iii) *grid-based*, and (iv) *contour-based*. The first class of techniques typically operates on the whole image and includes techniques such as principal components analysis (PCA) and convolutive nonnegative matrix factorization (cNMF) that operate on the original articulatory data to obtain relatively abstract, spatio-temporal basis functions of articulatory movement (Ramanarayanan et al., 2013a; 2016) that have been shown, in some cases, to be akin to linguistic gestures (Browman and Goldstein, 1995). The second class of methods – *ROI-based* – require manual definition of specific regions of interest where the relative variation (or covariation) of ensembles of pixels can be used to obtain insights into various linguistic and clinical questions of interest, or provide intermediate features for further modeling (Lammert et al., 2010; Tilsen et al., 2016). The third class of methods are *grid-based* in that they require the superimposition of an appropriate (typically semi-polar) reference coordinate system on the image, allowing extraction of vocal tract area functions by computing points of intersection between gridlines and soft tissue (Maeda, 1979; Proctor et al., 2010b). While this is more complex to implement relative to the previous methods, it also provides more easily interpretable lower-level features. The final

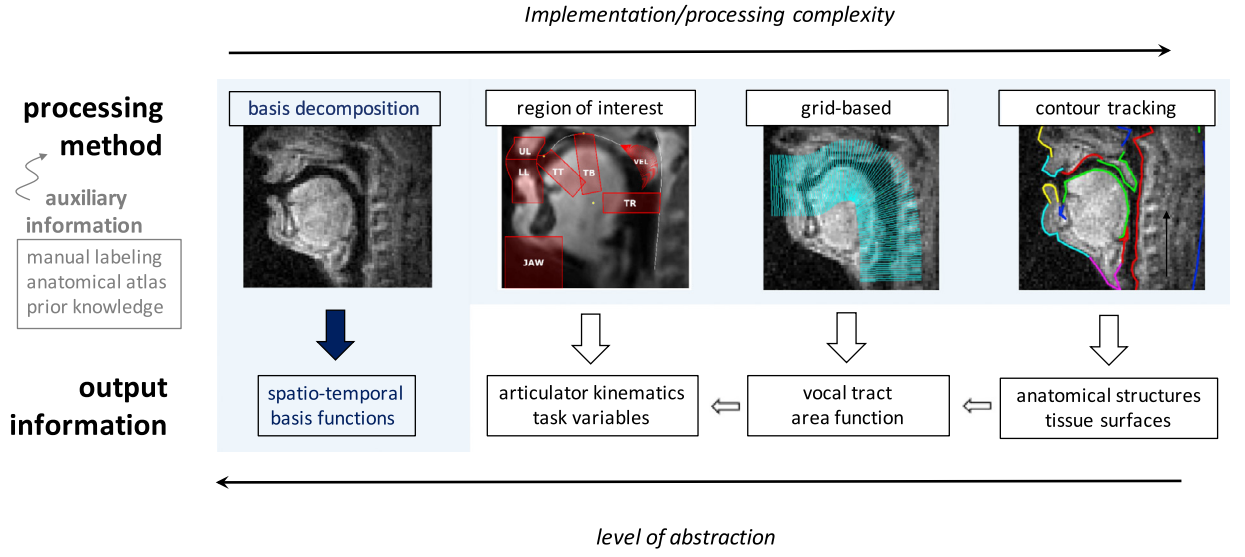

Fig. 1. A schematic overview of different RT-MRI analysis methods. The blue colored box indicates that basis decomposition techniques like principal component analysis (PCA), factor analysis (FA), convolutive nonnegative matrix factorization (cNMF), etc. can be applied to both the raw image as well as features obtained using all processing methods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

class of analysis methods involve extraction of all tissue boundaries, including those that do not directly define the geometry of the upper airway, but which are critically involved in speech production. These methods allow more detailed specifications of anatomical structures to be used directly for analysis, or as intermediate features for modeling. Examples of such methods include image segmentation (which yield vocal tract air-tissue contour outlines) as well as higher-level features that can be derived from these contours, such as vocal tract constriction variables or geometric articulatory coordinates.

The rest of this paper expands on each of these different classes of analysis methods, and is organized as follows: Sections 2–5 briefly describe the four aforementioned classes of processing methods, followed by a discussion of issues of reproducibility and reliability with respect to RT-MRI analyses (Section 6). We then briefly highlight and summarize some of key linguistic and speech science (Section 7) and clinical studies (Section 8) where such analysis methods have been applied.

2. Basis decomposition or matrix factorization techniques

Basis decomposition refers to a family of methods that factors a given data matrix (for instance, a single MR image, or an entire MR image sequence where each column contains a vector representation of the raw image or image features) or tensor (for instance, a video sequence) into a *basis matrix* consisting of a finite number of invariant spatial and/or temporal patterns and an *activation matrix* that describes how weighted combinations of those patters can reconstruct the original data.[1] These techniques allow extraction of the invariant aspects of data from the different sources of variability contained within. Such techniques are versatile in that they can be applied to any vector representation of the RT-MRI data – ranging from the vectorized image itself represented by raw pixel intensities to vectorized contour outline representations. The principal components analysis (PCA) family of techniques (Jolliffe, 2002), in particular, is a widely used technique that represents data as a weighted sum of orthogonal factors that account for maximal variance in the data (i.e., each column of the basis matrix is orthogonal to the other columns; see for example Maeda, 1990). PCA has been used to analyze patterns of vocal tract shaping across both spatial (e.g. Yehia and Tiede, 1997) and temporal domains. Carignan et al. (2015), for example, applied principal component analysis to cohorts of pixels in midsagittal images comparing the velum and tongue body articulation in French nasal vowel production. Targeted image regions are modeled as sets of n pixel intensities (columns/variables) for each of m images (rows/observations) in their experimental corpus, for each MR slice. Grayscale heatmaps can then be used to interpret PCs in articulatory terms, with whiter regions corresponding to presence of soft tissue. Factor analysis, a related technique that does not require orthogonality of components among others, has also been used to analyze tongue-shaping behavior (Harshman et al., 1977).

Other methods such as temporal decomposition and convolutive non-negative matrix factorization (or cNMF) have been successfully applied to time-varying articulatory data (both raw as well as processed features) to capture both spatial and temporal invariances. For instance, Jung et al. (1996) used weighted temporal decompositon techniques, first proposed by Atal (1983), to derive gestural score-based representations (see Browman and Goldstein, 1995) directly from articulator movement records. More recently, Ramanarayanan et al. (2013b) applied a cNMF technique with sparseness constraints to extract meaningful "primitive" representations of articulatory movement, akin to the physical realization of linguistic gestures proposed by Articulatory Phonology theory (Browman and Goldstein, 1995). This family of algorithms factors an $M \times N$ data matrix (the columns of which, for example, can be formed by converting each $I \times J$ MRI image of an input video sequence into a column vector of shape $M \times 1$, where $M = I * J$) into a "basis tensor" that contains a fixed number of spatio-temporal patterns (or primitives) and an "activation matrix" that captures the time instants when each of those spatio-temporal patterns occurred in the data. Subsequent work has since expanded upon these results to obtain more mathematically-valid and linguistically-interpretable primitives (see Ramanarayanan et al., 2016; Vaz et al., 2016, for graphical examples of such primitives).

[1] In linear algebra, a basis is a set of linearly independent vectors that, in a linear combination, can represent every vector in a given vector space. Such a set of vectors can be collected together as columns of a matrix – a matrix so formed is called a basis matrix. More generally, this concept can be extended from vectors to functions, i.e., a basis in a given function space would consist of a set of linearly independent basis functions that can represent any function in that function space. For further details and mathematical treatments of basis decomposition techniques including unified views of different matrix factorization methods, see Strang (2006), Li and Ding (2006), Singh and Gordon (2008), Ding et al. (2010).

3. Direct image analysis: looking at specific vocal tract regions of interest (ROIs)

Direct image analysis methods analyze pixel intensity in a specific region of the image. Most methods of this type involve regions-of-interest (ROIs). ROIs are often defined with the goal of identifying image pixels in which spatiotemporal variation in pixel intensity provides information regarding a relevant articulatory feature (e.g. jaw elevation) or task variable (e.g. lip aperture). ROIs can be as small as a single pixel, although it is often desirable to define ROIs as multi-pixel regions. An advantage of using direct image analysis methods is that they are relatively simple to implement and interpret. For example, Niebergall et al. (2013) show how articulatory dynamics and coarticulation are visible in spatiotemporal profiles of pixel intensity from single-pixel-wide lines along which lingual and labial articulators move. A different example comes from Freitas et al. (2016), who use the ratio of the mean and standard deviation of a velum ROI to assess signal artifacts. A number of studies have examined velum control in nasal vowels and consonants using ROI-based methods (cf. Teixeira et al., 2012; Carignan et al., 2013; Byrd and Saltzman, 2003).

3.1. ROI selection considerations

ROIs usually define a static, contiguous region whose boundaries are determined relative to anatomical structures and/or articulator positions observed across a sequence of image frames. Manual ROI definition is normally guided by knowledge of vocal tract anatomy, in order to ensure that extracted features are useful for analyses. The robustness of ROI-based features depends on the characteristics of the structures which are used to define ROIs, which in turn depends on the specific articulators/task variables for which an ROI is defined. Below we discuss a couple of examples.

An ROI defined to capture information regarding the tongue root (TR, see the second image panel of Fig. 1) should be delimited by the posterior pharyngeal wall. The posterior pharyngeal wall is fairly easy to identify because it is comprised of soft tissue, and it is relatively static across frames during speech. However, speakers can produce pharyngeal constriction gestures that change the location of the pharyngeal wall. Determining the anterior boundary of a TR ROI is somewhat more problematic. This boundary should be anterior enough such that the tongue root surface always remains in the ROI, but a boundary that is too anterior will reduce the dynamic range of features extracted from the ROI and may overlap with other ROIs, inducing a correlation (which is often undesirable). The superior and inferior boundaries of a TR ROI should generally avoid capturing velar and epiglottal movement.

Many of the same issues arise when defining ROIs to capture information regarding the lips. The inferior boundary of a lower lip (LL) ROI (cf. Fig. 1) should be sufficiently inferior so that the superior surface of the lower lip cannot move outside of the boundary. A problematic decision regarding the superior LL-ROI boundary and inferior upper lip (UL)-ROI boundary must be made in order to define separate LL and UL ROIs. This is problematic because the spatial location of labial contact can vary according to many speech-related factors. Hence more complex grid/edge-based methods are preferable. TT and TB ROIs can be challenging to define because the relevant articulators are constrained by hard-tissue structures: the anterior palate and dentition. Because these regions of the vocal tract do not show up very clearly in most imaging protocols, some interpolation and extrapolation might be necessary to approximate their locations. Alternatively, additional image processing can be used to improve SNR in key regions: Scott et al. (2013) use adaptive averaging to better resolve the soft palate.

3.2. ROI pixel intensity

One class of measures derived from ROIs involves aggregate pixel intensity. The sum (or alternatively, mean or median) intensity of pixels in an ROI reflects the amount of tissue in the region. Variations in the pixel intensity in a region over time thus can be used to analyze articulator dynamics (see e.g. Bresch and Narayanan, 2009; Lammert et al., 2010; Shosted et al., 2012; Silva et al., 2013; Tilsen et al., 2016). For example, as the tongue root (TR) is retracted for a low, back vowel, the pixel intensity will increase in the TR ROI. The usefulness of such features depends on the extent to which the ROI isolates the relevant articulator movement.

Beyond reliance on manual image labeling, there are several drawbacks to features derived from ROI pixel intensity. First, the intensity values obtained from these methods are not measures of articulator positions, even if they can indirectly represent those positions. Moreover, the intensity of a single pixel or the average over a group of

pixels is quite difficult to interpret in any physical unit due to complexities in the relation between tissue density, MRI signal levels, and reconstructed image values. Second, pixel intensity features do not reflect the direction of tissue motion. For example, if the lower lip exhibits both vertical and horizontal components of motion through the LL ROI, these cannot be distinguished.

3.3. ROI intensity centroid

Another method for extracting features from an ROI involves calculating the 2-dimensional intensity centroid in an ROI. The intensity centroid (IC) is analogous to a center of mass – simply picture the ROI as a 3-dimensional object whose elevation at each pixel coordinate corresponds to its intensity: the IC is where the object would balance on a fulcrum. Technically, the IC is the intensity-weighted average value of the horizontal and vertical coordinates in an ROI (Tilsen et al., 2016). The IC thus provides a measure of where intensity within an ROI is distributed. For example, when the LL is protruded, this can be distinguished from LL retraction, even when there is no aggregate change in pixel intensity. The ROI-IC has the advantage of providing information regarding movement direction, and if desired a principal components analysis can be conducted to determine the main axis of motion in an ROI.

While the ROI-IC also provides a feature with physically interpretable units – pixel coordinates – these do not index the position of any articulator or tissue boundary. The IC also has the disadvantage of being highly sensitive to ROI location: if the relevant articulator leaves the ROI, the centroid measure can change discontinuously.

3.4. Inter-ROI correlation

Relations between ROIs are also informative. Because pixel intensities from distinct regions of interest can be intrinsically temporally-aligned, correlations between regions provide information regarding articulatory activity. Multiple-ROI analysis may therefore be suitable for investigating coordination relationships between articulators (Proctor et al., 2010a). Teixeira et al. (2012), for example, examine velic coordination in European Portugese nasal vowel production using two ROIs located at the lips and nasopharynx. In the case of repetitive speech tasks, these data can then be treated using standard methods for comparing signal similarity and examining inter-signal timing, such as correlation analysis. This has proven to be a useful approach for identifying speech errors (Lammert et al., 2011), even when covert (not apparent in the acoustic signal), as demonstrated in an analysis of apraxic speech (Hagedorn et al., 2017).

3.5. Disadvantages of direct image analysis

Direct image analysis methods have the disadvantage of relying on manual image labeling or pixel selection. ROIs are typically manually defined from image inspection or semi-automatically generated from manually identified anatomical landmarks. Because of this, the replicability of manual ROI-based methods may be influenced by inter-labeler consistency. ROI-based methods are also potentially problematic for articulatory feature extraction. The extracted features are necessarily indirect representations of articulator kinematics/vocal tract geometry: ROI-based features are expressed in units of pixel intensity and pixel location, rather than units of articulator position or constriction size. We note also that co-registration of image frames is a prerequisite for automating ROI analyses across a sequence of frames. This serves to maintain a constant anatomical location of the ROI. An affine transformation is usually applied to rotate and translate each frame, in order to compensate for head movement in the imaging plane. Off-sagittal plane head movement cannot be compensated for by the co-registration process and therefore introduced noise in direct image analyses.

4. Grid based analysis

Analysis of the vocal tract with reference to a superimposed grid has a long tradition in speech science. The technique was originally developed to analyze X-ray data (Heinz and Stevens, 1964; Ladefoged et al., 1971; Mermelstein, 1973), and later adapted for use in pioneering structural MRI studies of speech production (Baer et al., 1991; Greenwood et al., 1992; Demolin et al., 2000). The use of grid-based analysis has a dual motivation: it provides a straightforward method of capturing vocal tract shape in a complex image, and doing so in a sparse parametric form

that can be reconciled with popular models of speech production and phonological representation. These techniques have primarily been used in analysis of static MR or X-ray images, and are applicable to RT-MRI when applied to each frame of an image sequence.

4.1. Analysis grid geometry

Analysis begins with construction of a coordinate system to be superimposed on images of the vocal tract. Rectangular grids have been used to analyze articulation within a Cartesian coordinate space (Stone et al., 2001), but midsagittal articulatory analysis most frequently uses a polar coordinate system. A semipolar grid constructed around an origin located near the center of the resting tongue body will follow the curve of the vocal tract from the mid-pharynx to the alveolar ridge. Another advantage of this method is that radial gridlines remain roughly normal to the tract midline throughout the midsagittal region of interest (Fig. 2), allowing for consistent estimation of vocal tract aperture throughout the oral cavity (Greenwood et al., 1992; Demolin et al., 2000).

The analysis grid is typically located and defined with respect to key anatomical landmarks, including the posterior pharyngeal wall, hyoid, highest point of the hard palate, upper teeth and lips, to ensure optimal coverage of the speech articulators and resonant cavities (Öhman, 1967; Mermelstein, 1973). A single semi-polar analysis grid can be used to cover the entire vocal tract (e.g. Zhang et al., 2016), or additional geometries can be introduced to follow the morphology of the tract more closely. The lower pharynx can be tracked with a set of horizontal lines extending from the end of the semipolar grid to the glottis, or into the trachea, and the anterior part the tract to the alveolar ridge may be covered with a series of vertical (Mermelstein, 1973; Engwall and Badin, 1999) or tract-normal gridlines (Beautemps et al., 1995), or a second semi-polar grid tracking the change in tract curvature (Proctor et al., 2010b).

The density of the analysis grid will have important implications for speech MRI analysis, as gridline spacing will determine how sparsely tissue boundaries will be sampled. In high resolution images, gridlines may be spaced at a fixed angle or distance (e.g. Öhman, 1967 specifies 5° and 5 mm), but uniform spacing may not always provide adequate coverage of complex anatomy such as the epiglottis or the uvula, especially if the subject's head and/or the image FOV are small. Closely-spaced gridlines may help resolve anatomical detail and improve tracking of rapid

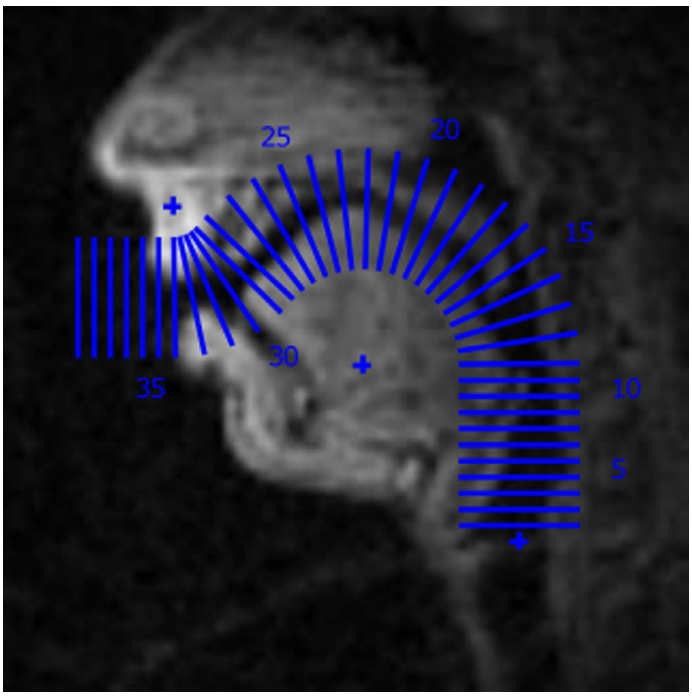

Fig. 2. Composite semi-polar analysis grid, defined with respect to glottal, lingual and palatal landmarks, superimposed on a midsagittal MR image of a male vocal tract (Proctor et al., 2010b).

articulator movement, for example around the alveolar ridge, or where variability in signal quality demands higher spatial sampling, such as the velum. Variable gridline spacing – sometimes dynamically adjusted (Engwall and Badin, 1999) – has been used in different regions of the tract to accommodate these factors.

Additional factors must be considered when designing a grid for analysis of RT-MRI data, as opposed to static structural MRI. If a grid is defined with respect to anatomical structures which move during speech, either the grid must be shifted between image frames, or a translation must be applied when comparing vocal tract configuration across frames, to correct for the movement relative to the grid (Proctor et al., 2010b). If the subject's head is immobilized during image acquisition, the passive structures of the upper oral tract and rear pharyngeal wall will remain fixed across frames, but motion of the larynx and lips will need to be accommodated.

4.2. Locating anatomical structures

Once a grid has been superimposed on an MR image, anatomical structures are located by finding the intersections between each gridline and tissues boundaries of interest – typically the tongue, lips, velum, and passive structures of the midsagittal oral cavity. Tissue boundaries can be manually identified on each gridline (Greenwood et al., 1992), but are more commonly extracted automatically or semi-automatically.

Algorithms can detect air-tissue transitions by locating abrupt changes in pixel intensity along each gridline (e.g. Engwall and Badin, 1999; Kim et al., 2014). The analysis grid guides the algorithm by defining limits on the search space, typically bounded longitudinally by the glottis and the most anterior point of the lips; a search algorithm may assume that each gridline between these extremities will intersect exactly two tissue boundaries defining the inner and outer limits of the midsagittal vocal tract. Intensity profiles calculated along tract-normal gridlines may contain multiple local extrema, and the actual tract boundaries do not always align with the steepest intensity slopes, so accurate determination of tract geometry is not a trivial task. Zhang et al. (2016) use combined multi-directional Sobel operators to more robustly identify candidate tissue boundary points in noisy MR images.

A computationally efficient method of locating tract outlines is to construct a graph through the set of candidate boundary nodes (Fig. 3) in which edges are weighted according to other geometric constraints such as distance and curvature, and/or with reference to tissue boundary locations on temporally-adjacent image frames (Zhang et al., 2016). Two optimal paths from glottis to lips can then be calculated, corresponding to inner (tongue) and outer (palate and rear pharyngeal wall) vocal tract boundaries. Tissue boundary estimates are constrained by multiple configurable search parameters, and can be calculated using Dijkstra's algorithm (Proctor et al., 2010b), the Viterbi

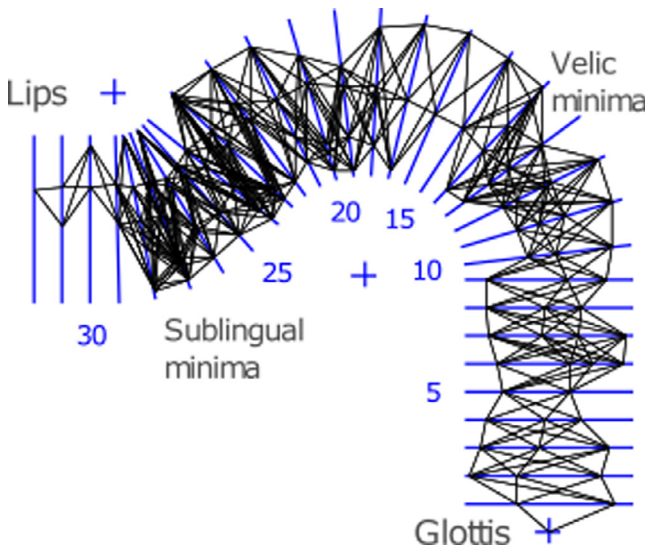

Fig. 3. Use of a graph to compute vocal tract boundary locations. Nodes represent candidate points (locally maximal changes in pixel intensity) on tissue boundaries. Terminal nodes are defined at glottal and mid-labial points. Tract outlines computed from optimal paths between terminal nodes along edges weighted by image features and geometric constraints.

algorithm (Kim et al., 2014; Zhang et al., 2016) or other optimization techniques. These approaches typically require some prior knowledge of the anatomical configuration of the tract, which is used to further constrain the search space.

Although vocal tract analysis is prototypically performed in the midsagittal plane, similar techniques can be used to analyze tract geometry in MRI data acquired from coronal, axial, and oblique imaging planes (Kim et al., 2012). An advantage of grid-based approaches is that the coordinate system imposed on one set of images can guide acquisition from intersecting planes. For example, coronal cross sectional images can be acquired through the anterior part of the tract, and axial images from the pharynx, at regularly spaced planes corresponding to the gridlines on the midsagittal analysis grid (Badin et al., 1998; Engwall and Badin, 1999), allowing for three-dimensional vocal tract models to be constructed.

4.3. Extracting speech models

Grid-based analysis of speech MRI data has proven popular because the information extracted using this method of analysis is highly interpretable in phonetic terms, and can directly inform standard models of speech production. Midsagittal aperture functions can be computed directly from tissue boundary intersection points on each gridline, from which area functions can be derived (Mermelstein, 1973; Beautemps et al., 1995; Engwall and Badin, 1999).

Grid-based analysis has also proven especially useful in real-time MRI for examining articulation in specific regions of the vocal tract. Because the analysis grid defines a consistent coordinate space, information can be reliably extracted from identical regions of the tract in image sequences to examine changes in constriction over time (Demolin et al., 2000; Proctor and Walker, 2012).

5. Image segmentation and higher-level feature extraction

While direct pixel analysis or grid-based analysis provide a robust and convenient analysis, intuitively meaningful contour-based representations may also be required, depending on the application. Examples include outlines or contours that delineate the tongue and vocal tract structures. In the case of XRMB or EMA, contours may be obtained by fitting a smooth spline through all pellet points. However, in the case of MRI and ultrasound, more involved image processing is required to segment air-tissue boundaries. See Fig. 4 for graphical examples of such features.

5.1. Unsupervised image segmentation

Unsupervised region segmentation of the upper airway, jaw and supraglottal articulators is a viable approach given sufficiently large datasets, and may be used for processing long sequences of MR images. Sampaio and Jackowski (2017) deploy level set functions (Li et al., 2010) to segment anatomical structures in midsagittal sequences of 120 frames acquired at 10 f.p.s., using an iterative technique robust to the absence of one or more articulators. A robust tool for segmentation in image Fourier space (also called k-space) has been developed based on an algorithm that uses an anatomically-informed object model, returning a set of tissue boundaries for each frame of interest (Bresch and Narayanan, 2009). This technique allows for quantification of articulator movement and vocal tract aperture in the midsagittal plane (see first panel of Fig. 4). In recent years, there has been increasing research on automated unsupervised and semi-supervised vocal tract image segmentation[2] that leverages a wide range of machine learning and computer vision tools, such as active shape models (ASM: Cootes et al., 1995), active appearance models (AAM: Cootes et al., 2001), and mesh deformation and registration methods (see for example Silva and Teixeira, 2015; Hewer et al., 2014; Harandi et al., 2015; Labrunie et al., 2016; Eryildirim and Berger, 2011). More robust segmentation may be achieved through combination of these methods. Asadiabadi and Erzin (2017), for example, combine active shape and active contour models (ACM, or 'snakes': Kass et al., 1988) to reduce boundary tracking errors when articulator occlusion occurs. Raeesy et al. (2013) use recursive boundary subdivision (RBS) to generate an initial training dataset that automatically finds landmarks in the images, and then deploy oriented active shape models (OASM) to locate tissue boundaries.

[2] Note that while there is also work on expert manual segmentation, we primarily focus this review on automatic analysis methods.

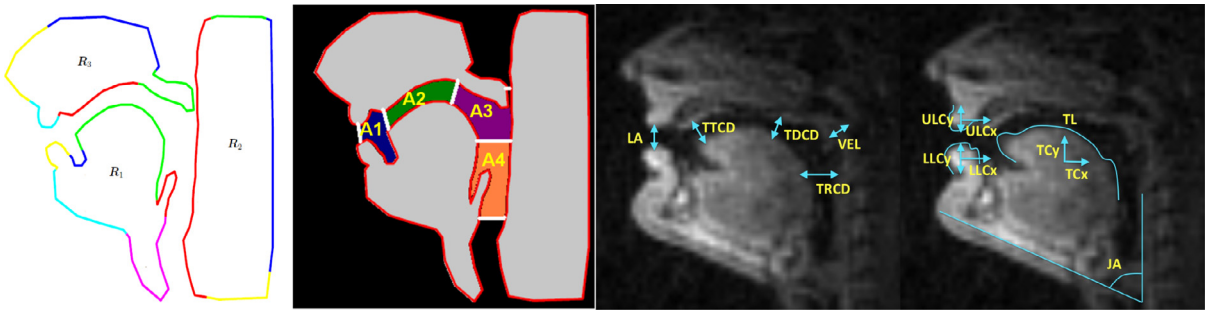

Fig. 4. Examples of meaningful features that can be computed from RT-MRI data. (a) Contour outlines (Bresch and Narayanan, 2009). (b) Meaningful cross-distances can be computed that segment the vocal tract into areas A1–A4 (Ramanarayanan et al., 2010). (c) Cross-distances in more detail (lip aperture (LA), velic aperture (VEL), and constrictions of the tongue tip (TTCD), tongue dorsum (TDCD) and tongue root (TRCD). (d) Articulatory posture variables – jaw angle (JA), tongue centroid (TC) and length (TL), and upper and lower lip centroids (ULC and LLC).

5.2. Higher-level features: constriction cariables and articulator positions

Once vocal tract contours are obtained, we can further compute other meaningful representations – such as area functions (introduced earlier), cross-distances, and other postural variables (Mermelstein, 1973; Maeda, 1990). A pervasive problem with analyzing parametric vocal tract data extracted from MRI of different speakers is that of cross-speaker alignment. Whichever technique has been used to quantify tract configuration, differences in speaker anatomy make it difficult to compare data across speakers. In the case of grid-based analyses, for example, different numbers of gridlines might be required for different speakers, and gridlines with the same index might intersect different vocal tracts at different places of articulation. Ramanarayanan et al. (2013a) proposed a method to automatically derive cross-distances that are computed at points where constrictions are made in the vocal tract during normal speech production, such as the alveolar ridge for coronal stop consonants, or the lips for labial stops. Hence they are conducive to meaningful comparison across subjects. In addition, other meaningful postural features such as the angle of the jaw or the centroid of the tongue can be computed from segmented MRI data.

6. Reproducibility of RT-MRI data

When acquiring and analyzing RT-MRI speech data, several factors contribute variability to the final result, in both wanted and unwanted ways. Wanted variability comes, for example, from intra-speaker variations and emotional state, in cases where these effects are under study. Unwanted variability may be caused by variations in subject positioning, scan plane positioning and manual steps in post-processing and evaluation. Unwanted data variability may be negligible in small single-center studies (i.e. a study performed at a single site by one research group alone) and in cross-sectional studies where all the data is collected and analyzed by a single researcher at one point in time. However, it is critical to understand and mitigate unwanted sources of variability to tackle important questions with high scientific impact such as (a) long-term speech development and rehabilitation, (b) clinical applications where the analysis result may influence diagnosis and treatment, and (c) to conduct larger studies involving several research sites separated geographically.

The sources of variability can be classified into (1) MRI acquisition technology variability, including scanner parameters, noise and image reconstruction, (2) MRI operator variability, such as subject positioning and scan plane alignment, (3) image analysis including subjective differences in manual delineation and initialization of automatic methods, and (4) physiologic variation such as short-term speaker variability (including repetition-to-repetition) and longer-term changes in speech production (e.g., due to emotional state, development, and aging) (Töger et al., 2016). In the above, groups 1–3 generally introduce unwanted variability, which may confound comparisons or even mask important information in the collected data. In contrast, the variability introduced in group 4 may be highly desired for many research aims. Therefore, all steps in the acquisition, processing and analysis of data should be optimized to minimize the influence of groups 1–3, and preserve the physiological variability in group 4 in the resulting data.

In the post-processing context of this review, analysis methods should be designed to not introduce bias or variability of its own. Furthermore, analysis methods should be robust to human error, so that data produced at different points in time or at different research sites can be compared without introducing confounding differences.

Several practical strategies that are useful to quantify and reduce unwanted variability can be borrowed from medical imaging research. In medical imaging, unwanted variability has long been in the spotlight to ensure accurate and reliable diagnosis and treatment decisions for patients. Recently, methods and terminology have been standardized by the *Quantitative Imaging Biomarker Alliance* (QIBA), organized by the Radiological Society of North America (RSNA) (Kessler et al., 2015). An important and quite simple test of the robustness of a post-processing method is to let two researchers evaluate data from a subset of subjects independently, to determine the *interobserver variability*. A high interobserver variability suggests that data evaluation is ambiguous and must be specified more thoroughly, while a low variability reinforces the objectivity of the data. Further, more extensive tests include *repeatability*, where the same MRI scan is performed twice on the same day under the same conditions, also called a *test-retest study*, and *reproducibility*, where the operator, measuring equipment and spatial location of the measurement are changed.

In speech and upper airway research using static MRI, a limited number of studies have determined interobserver variability (Arens et al., 2001; Chi et al., 2011; Fitch and Giedd, 1999; Vorperian et al., 2005; Echternach et al., 2016) and repeatability (Welch et al., 2002) of static upper airway measures. To the best of our knowledge, only a single study has investigated the test-retest repeatability of RT-MRI biomarkers (Töger et al., 2016). This shows an open opportunity to further strengthen the RT-MRI speech analysis community and prepare the field for larger speech studies and scientific breakthroughs, e.g., through large multi-center population studies, long-term longitudinal speech development studies and clinical applications.

7. Linguistic understanding and speech technology applications

The analysis techniques reviewed in Sections 2–6 have been used to investigate a wide range of linguistic and speech phenomena. RT-MRI has transformed laboratory phonology by providing new levels of detail about the global configuration of the vocal tract and the way it is reshaped over time. Because many advances in real-time MRI of the vocal tract have been driven by investigations into phonological phenomena and language-specific studies of speech production, methods for analyzing these data are often intrinsically phonetically-motivated. A representative survey of techniques used to analyze RT-MRI data in linguistic studies is provided in Table 2.

7.1. Study of linguistic phenomena

Grid-based analysis of MRI data has often been used in studies focusing on specific aspects of speech production, where articulation can be intuitively compared across speakers, tokens, and points in time, in target regions of the vocal tract defined by a range of grid lines; for example, to characterize pharyngeal articulation in Arabic emphatic consonant production (Israel et al., 2012). Grid-based analysis is also a popular and straightforward method for tracking articulator movement (e.g. of the larynx, Demolin et al., 2002; or velum, Teixeira et al., 2012), and quantifying changes in vocal tract aperture along a targeted gridline (Demolin et al., 2000; Carignan et al., 2015). In an innovative multi-modal study of speech imitation (combining structural vocal tract imaging with fRMI of the brain), Carey et al. (2017) use grid-based analysis to quantify labial protrusion during a vowel production task.

Other types of phonetic studies require detailed information about global tongue shaping, in which case tissue segmentation analysis may be a more appropriate way to process RT-MRI data. Midsagittal lingual outlines have been used to study details of lingual and velic posture in French vowels (Delvaux et al., 2002), and articulation of multi-gestural liquid consonants in English (Proctor and Walker, 2012) and Tamil (Proctor et al., 2009). Midsagittal analysis of lingual-palatal contact patterns extracted from high temporal resolution MRI sequences have been used to examine details of click consonant production in African languages (Proctor et al., 2016).

Region-of-interest analysis is a useful method for investigating gestural timing relationships and patterns of constriction formation and release in specific regions of the vocal tract; it has been used to characterize Korean liquid consonant production (Lee et al., 2015), and Italian singleton/geminate contrasts (Hagedorn et al., 2011). This approach may also be more suitable when, due to differences in vocal tract morphology, tissue outlines or grid-based analysis do not allow for easy comparison of tokens produced by different speakers. ROI analysis has proven to be a

Table 2
Linguistic understanding and speech technology applications.

Linguistic Phenomenon	Language	Class	Methods	Representations	Analyses	Reference
Pausing behavior in spontaneous speech	English	Basis decomp	Intensity gradient	Midsag boundary difference	Gradient frame energy	Ramanarayanan et al. (2009)
Identifying articulatory primitives	English	Basis decomp	cNMF	Midsag basis vectors	Mean constrictions	Ramanarayanan et al. (2011)
Articulatory settings for speech	English	Basis decomp	ACMs	Midsag tract area descriptors	VTADs	Ramanarayanan et al. (2013a)
Articulatory settings for speech	English	Basis decomp	LW regression	Cross-section, jacobians	Tract variables	Ramanarayanan et al. (2014)
Phonetic class recognition	English	Basis decomp	PSPI	Optimal regions, PSPI matrices	Max. pixel information	Prasad and Ghosh (2016)
Coarticulation, artic. compensation	French	Direct		Midsag images	Qualitative	Demolin et al. (1997)
Velic coordination in nasalisation	French	Direct		Midsag images	Inter-frame timing	Proctor et al. (2013b)
Patterns of vowel reduction	English	Direct		Mean midsag images	Mean image geometry	Proctor et al. (2015)
Geminate consonant production	Italian	Direct	ROI	Constriction kinematics	Tissue velocities	Hagedorn et al. (2011)
Liquid consonant allophony	Korean	Direct	ROI	Constriction kinematics	Intensity ratios	Lee et al. (2015)
Nasal vowel, liquid articulation	Portuguese	Direct	ROI	Constriction kinematics, CSA	Intensity ratios	Teixeira et al. (2012)
Nasal vowel production	French	Direct	ROI, PCA	Midsag PC heatmaps, CSA	PC loadings, apertures	Carignan et al. (2015)
Pharyngealization and emphasis	Arabic	Direct	ROI, PCA	Constriction kinematics, PCs	Aperture timeseries	Shosted et al. (2012)
Vowel and consonant production	French	Direct, grid	ROI	Midsag aperture, location	Tissue kinematics	Demolin et al. (2002)
Articulation of nasal vs. oral vowels	French	Direct, grid		Midsag tissue boundaries	Qualitative	Demolin et al. (2002)
Articulation of nasal vs. oral vowels	Portuguese	Direct, contour	ROI, Region growing	Tissue boundaries	Tissue kinematics	Silva et al. (2013)
Vowel & consonant production	German	Direct, contour	ROI	Midsag & coronal image	Aperture kinematics	Niebergall et al. (2013)
Articulatory coordination in vowels	French	Grid		Midsag area functions	Aperture kinematics	Demolin et al. (2000)
Liquid consonant production	Tamil	Grid		Midsag aperture functions	Grid-based aperture	Proctor et al. (2009)
Emphatic consonant articulation	Arabic	Grid		Midsag tissue boundaries	Grid-based aperture	Israel et al. (2012)
Liquid-vowel coarticulation	English	Grid		Midsag tissue boundaries	Qualitative	Proctor and Walker (2012)
Click consonant production	Nama	Grid		Midsag tissue boundaries	Qualitative	Proctor et al. (2016)
Fricative production	Mandarin	Grid		Midsag & coronal boundaries	Qualitative	Proctor et al. (2012)
Tongue shaping in liquid consonants	English	Grid		Midsag tissue boundaries	Lingual curvature	Smith (2014)
Lip rounding in imitative vowel prod.	English	Grid		Midsag tissue boundaries	Labial protrusion	Carey et al. (2017)
Articulation of nasal vs. oral vowels	Portuguese	Contour	AAMs	Midsag tissue boundaries	Artic. feature comparison	Silva and Teixeira (2015)
Articulation of nasal vs. oral vowels	Portuguese	Contour	AAMs	Midsag tissue boundaries	Artic. feature comparison	Silva and Teixeira (2016)
Syllable structure and nasalization	English	Contour	ACMs	Midsag tissue boundaries	Aperture kinematics	Byrd et al. (2009)
Vowel & consonant production	Swedish	Contour	Threshold, spline	Midsag tissue boundaries	Tract variables	Engwall (2004)

popular method in studies of nasalization, where the state of the velum, whose physiology varies considerably across individuals, can nevertheless be robustly tracked by monitoring the intensity of a cohort of pixels located in the naso-pharynx (Teixeira et al., 2012; Carignan et al., 2015).

7.2. *Automatic speech recognition*

There have been several production-oriented approaches to automatic speech recognition using both direct and estimated data as well as using recognition models developed based on speech production knowledge (Rose et al., 1996; Deng et al., 1997; Frankel and King, 2001; Mcdermott and Nakamura, 2006; Ramanarayanan et al., 2012). For example, Frankel and King (2001) showed improvement in speech recognition accuracy by combining acoustic and articulatory features from a talker. However, it is not practical to assume the availability of direct articulatory measurements from a talker in real-world speech recognition scenarios. To address this challenge, a number of techniques have been proposed (Deng et al., 1997; Lee et al., 2003; Ma and Deng, 2004) where, instead of relying on features from direct articulatory measurements, abstracted articulatory knowledge is incorporated in designing models (e.g., Dynamic Bayesian Networks (DBNs), Hidden Markov Models (HMMs)) which can be gainfully used for automatic speech recognition. A summary of such techniques can be found in Mcdermott and Nakamura (2006). Deng (1998) proposed an integrated Bayesian framework for ASR that consists of a hard-wired lexical compilation/representation component (which attempts to generalize ideas of feature overlap proposed by phonological theories so that the acoustic space can be modeled with fewer atomic speech units) as well as a stochastic acoustic mapping component. Multi-steam architectures (Metze and Waibel, 2002) have been also proposed as an alternative approach where linguistically derived articulatory (or more generally, phonetic) features are estimated from the acoustic speech signal, typically using deep neural networks (DNNs), and then used to either replace or augment acoustic observations in an existing HMM based speech recognition system. More recently, Ghosh and Narayanan (2011a)

have used estimated articulatory features obtained through subject-independent acoustic-to-articulatory inversion (AAI) to address the challenge of unavailability of direct speech production data during speech recognition. While these techniques have been primarily applied to EMA or X-ray data, they can in principle be extended to data obtained using RT-MRI.

7.3. Speaker morphology and recognition

Understanding the interplay of vocal tract structure, articulation and acoustics has technological applications for automatic speaker recognition. Vocal tract length normalization is one example of morphological knowledge that has already provided performance benefits to automatic speech and speaker recognition (Eide and Gish, 1996; Lee and Rose, 1998; Welling et al., 2002). Possibilities exist for providing normalization of the acoustic signal for other structural differences that impact a variety of phonemes (Lammert et al., 2013b). An essential component of this normalization, in terms of making if practically useful, is to accurately predict morphological characteristics of a speaker, such as vocal tract length (Lammert and Narayanan, 2015), among others, directly from the acoustic signal (i.e., morphological inversion). Predictions of this kind may subsequently lead to applications in speaker recognition. These features will be unique to an individual speaker, making them ideal for biometric applications.

Li et al. (2016) recently proposed a practical feature-level and score-level fusion approach by combining acoustic and estimated articulatory information for both text independent and text dependent speaker verification. They demonstrated that the articulatory constraints introduced by inverted articulatory features help to reject wrong password trials and improve the performance after score level fusion, achieving more than 15% relative equal error rate reduction for speaker verification tasks.

7.4. Speech synthesis

There is a large body of work in the literature on leveraging articulatory data and representations for speech synthesis applications. See Kröger and Birkholz (2009) for a nice overview of different articulatory synthesis models – categorized as either vocal-tract models, acoustic models, glottal models or noise-source models. Recent advances in such modeling are informed by phonological theory (Kröger and Birkholz, 2007), real articulatory data such as from MRI (Birkholz and Kröger, 2006), and the understanding of coarticulation phenomena (Birkholz, 2013). Furthermore, many synthesis models that are not directly articulatory in nature have also successfully leveraged machine learning techniques such as Gaussian Mixture Models (GMMs) or Deep Neural Networks (DNNs) to incorporate articulatory information into the synthesis process (see for example Rahim et al., 1993; Toda et al., 2004; Ling et al., 2009; 2013).

7.5. Modeling articulatory dynamics and speech motor control

Speech production data can facilitate understanding of forward (Lammert et al., 2013a; Ramanarayanan et al., 2014) and inverse models (Ghosh and Narayanan, 2011b) of the vocal tract, and build generative models of vocal tract shape dynamics (Bresch et al., 2010). In the latter, the authors investigated the application of statistical graphical models that can capture the spatio-temporal dependencies between various articulators in a data-driven manner. This study indicates that if we combine (a) an explicit multistream transcription with (b) appropriate techniques for extracting articulatory time-functions along with (c) the appropriate statistical models, we are well-positioned to derive phonological information directly from articulatory data. In related work, Katsamanis et al. (2011) proposed a modeling framework to validate different articulatory representations using articulatory recognition, which affords an understanding of the usefulness of a given representation in analyzing speech articulation.

The availability of articulation data such as from RT-MRI offers new scientific inquiry possibilities. Consider the case of speech motor control. One popular theory of motor control is the inverse dynamics model, i.e., in order to generate and control complex behaviors, the brain needs to explicitly solve systems of coupled equations. Mussa-Ivaldi et al. (1999) and Hart and Giszter (2010) instead argue for a less computationally complex viewpoint wherein the central nervous system uses a set of '*primitives*' to "solve" the inverse dynamics problem. Articulatory movement primitives may be defined as a dictionary or template set of articulatory movement patterns in space and time, weighted combinations of the elements of which can be used to represent the complete set of coordinated spatio-

temporal movements of vocal tract articulators required for speech production. Ramanarayanan et al. (2011) recently proposed an algorithm to automatically extract such primitives from speech articulation data. Subsequent work has since expanded upon these results to obtain more mathematically-valid and linguistically-interpretable primitives (Ramanarayanan et al., 2016; Vaz et al., 2016). Consider further the case of coarticulation in speech, where the position of an articulator/element may be affected by the previous and following target (Ostry et al., 1996). Using the idea of motor primitives enables exploration of how the choice, ordering and timing of a given movement element within a well-rehearsed sequence can be modified through interaction with its neighboring elements (coarticulation) (Sosnik et al., 2004).

8. Clinical applications

Speech RT-MRI provides a unique window into upper airway function that is potentially useful in a variety of clinical applications. The vast majority of clinical MRI procedures are of static tissue (e.g. brain, spine, joints) or of periodic motion (e.g. cardiac motion or pulsatile blood flow). With recent developments in RT-MRI, it is also possible to study non-periodic processes such as speech, swallowing and velopharyngeal function, which will severely restrict quality of life when dysfunctional. While swallowing is not strictly a speech-related function, swallowing depends on the same anatomical structures and physiological functions as speech, and can be imaged and analyzed using similar methods, and it is therefore included in this review in the hopes that the two research fields can benefit from each other. Table 3 summarizes a literature review of current clinical applications of RT-MRI in the fields of surgery, swallowing function, velopharyngeal insufficiency (VPI) and speech disorders. Selected studies from each field are discussed in more detail below.

In advanced cases of tongue cancer, parts of the tongue must often be surgically removed in a procedure called glossectomy. Fig. 5 shows the RT-MRI data of a patient after glossectomy. An animated version with simultaneously recorded audio can be found in the Supplemental Files (online). The reduced volume of the tongue is evident, as are the altered articulatory patterns compared to healthy control subjects. Glossectomy will inevitably impact the dynamics of speech production, as demonstrated in a study of five advanced tongue cancer patients post-glossectomy by Hagedorn et al. (2014). Vowel frequency analysis (F1 and F2) and qualitative assessment by observers showed that speech production was significantly altered post-glossectomy. In this study, ROI-based analysis (see Section 3) was used to study constriction events during various speech tasks. Compensatory mechanisms in the post-glossectomy patients included the use of labial stops instead of coronal stops and laterals, and fricatives being produced using a constriction between the tongue dorsum and palate instead of at the alveolar ridge. This study showed that RT-MRI can capture altered speech dynamics in patients post-surgery. Future uses of RT-MRI in glossectomy patients may lead to better understanding of the speech defects incurred by different surgical strategies. Patients may be imaged before and after surgery to specifically tailor speech therapy programs to optimize recovery of speech intelligibility.

Swallowing function is of great clinical importance, especially in stroke patients. The current standard of care is videofluoroscopy (VF), which provides high spatial resolution, but poor tissue contrast and requires ionizing

Table 3
Clinical applications.

Application	Study population	Method	Representations	Analyses	Ref
Speech disorders	Apraxic speech (n=1)	ROI	Midsag. images	Aperture kinematics	Hagedorn et al. (2017)
Surgery, speech disorders	Pre- and post-surgery tongue cancer (n=8)	Grid	Midsag. images	Aperture kinematics	Mády et al. (2003)
Surgery, speech disorders	Post-surgery tongue cancer (n=5)	ROI	Midsag. images	Aperture kinematics	Hagedorn et al. (2014)
Surgery, swallow function	Tongue cancer (n=4), controls (n=3)	Manual annotation	Midsag. + cor. images	Tissue kinematics	Zu et al. (2013)
Swallow function	Controls (n=10)	Manual annotation	Midsag. + cor. + ax. images	Qualitative, timings	Olthoff et al. (2014)
Swallow function	Brain stem infarct (n=3), controls (n=10)	Direct image	Midsag. + cor. + ax. images	Qualitative	Vijay Kumar et al. (2012)
Swallow function	Controls (n=10)	Manual annotation	Midsag. + oblique ax. images	Qualitative, timings	Zhang et al. (2012)
VPI	Children/young adults, suspected VPI (n=7)	Direct image	Midsag. + cor. + ax. images	Qualitative	Beer et al. (2004)
VPI	Controls (n=10)	Manual annotation	Midsag. images	Aperture kinematics	Bae et al. (2011)
VPI	Children/young adults with VPI (n=11)	Manual annotation	Midsag. + ax. images	Tissue kinematics	Drissi et al. (2011)
VPI	Children with VPI (n=5)	Direct image	Midsag. + cor. + ax. images	Qualitative	Sagar and Nimkin (2014)
VPI	Controls (n=6)	Direct image	Midsag. images	Tissue dimensions	Scott et al. (2012)

VPI = velopharyngeal insufficiency. All study subjects are adults unless otherwise noted. Midsag. = midsagittal, cor. = coronal, ax. = axial. Controls = healthy volunteers.

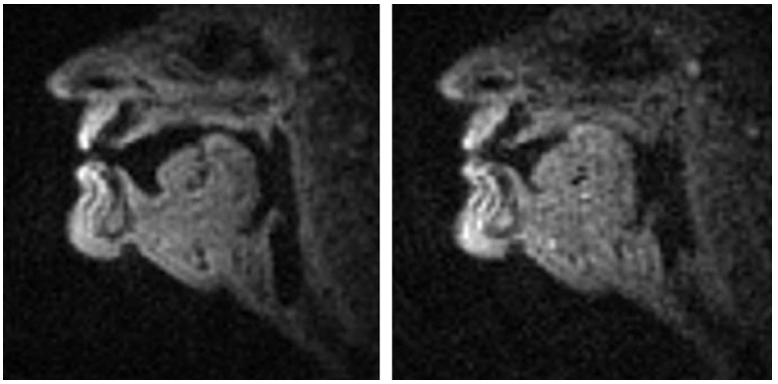

Fig. 5. *RT-MRI analysis of post-glossectomy speech*. Left: resting posture showing missing coronal tissue and reconstructed anterior part of tongue dorsum. Right: articulation of final consonant in 'bit', revealing simultaneous dorsal and labial constrictions [k͡p], substituting for the coronal stop /t/.

radiation (X-rays). For this application, RT-MRI can potentially provide better tissue contrast and reduce the radiation burden for patients. In an early study on swallow function, Zu et al. (2013) compared four controls to three patients with tongue squamous cell carcinoma. RT-MRI images were analyzed by measuring oral transit time (OTT) and pharyngeal transit time (PTT) of a swallowed bolus of yogurt. The change in submental muscle length (SM) and the distance between the hyoid bone and thyroid cartilage (TH) were quantified. OTT and PTT were longer in patients, which was consistent with previous VF studies. Furthermore, the patients showed altered motion patterns in SM and TH, reflecting their dysfunctional swallow mechanics. The measurement of SM is unique to RT-MRI due to the excellent soft tissue contrast. In summary, RT-MRI promises to be a versatile and sensitive method for clinical evaluation of swallow function. One important weakness of RT-MRI is that the overwhelming majority of scanners image the patient in the supine position, while swallowing is typically performed while upright. MRI scanners for upright scanning do exist, although with significantly limited imaging performance.

Velopharyngeal insufficiency (VPI) is a condition where the velum does not completely close against the pharynx, with consequences for speech, eating (swallowing and chewing) and breathing. Causes for VPI include cleft palate, adenoid gland removal, muscular weakness and motor speech disorders. Nasendoscopy and videofluoroscopy (VF) are traditionally used for diagnosis and follow-up, but are limited by insufficient tissue contrast and muscle visualization, and by ionizing radiation dose, which is especially important to limit in children. Beer et al. (2004) used RT-MRI at 6 frames per second (fps) to study seven patients with suspected VPI, and compared results to VF as the reference standard. Their main result was that qualitative interpretation of the RT-MRI images had excellent agreement with VF, showing that classification of VPI is possible using RT-MRI. After this early study, several other groups have improved image quality (e.g. by increasing temporal resolution to 21.4 fps, Bae et al., 2011) and shown its feasibility in children, an important VPI patient group (Drissi et al., 2011).

RT-MRI also has potential to be used in speech disorders to better understand different conditions, aid in diagnosis and refine treatment. Apraxia is a condition where neurological dysfunction leads to impaired motor planning of the complex spatiotemporal patterns needed for speech production. Hagedorn et al. (2017) showed that RT-MRI shows important new aspects of apraxic speech that are not visible using traditional methods, specifically covert (silent) intrusion errors that are not picked up by audio recordings and greater variability between repetitions of the same speech task (see Fig. 6). This pilot study shows the potential of RT-MRI to better understand speech disorders. Furthermore, RT-MRI may be used to specifically tailor treatment programs for each patient individually.

In summary, RT-MRI has several exciting potential clinical applications, promising improved patient monitoring and care. The main advantages of RT-MRI compared to currently used imaging modalities include the excellent soft tissue contrast, free choice of imaging plane, and absence of ionizing radiation. The main obstacles to be overcome are to establish cost-effectiveness, and to ensure widespread availability of high-quality RT-MRI sequences. Furthermore, there is a need to develop reliable software tools for reproducible quantitative measurements, which need to be held to a high standard for accuracy and reproducibility (Kessler et al., 2015) to ensure reliable diagnosis and evaluation.

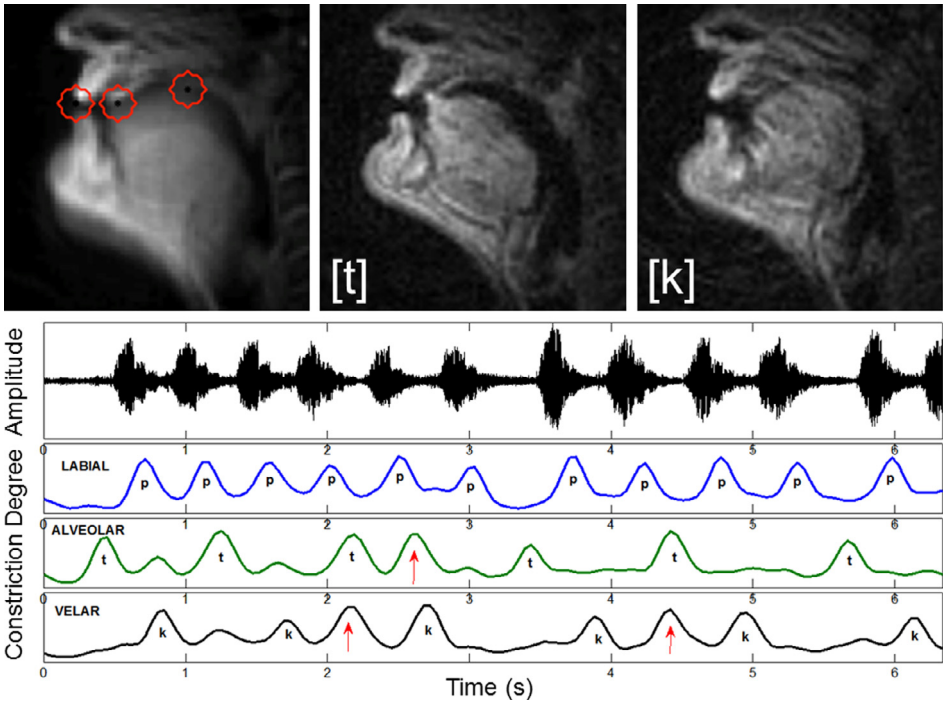

Fig. 6. *ROI analysis of apraxic speech.* Top Left: location of regions of interest: labial, alveolar and velar; Top Center: production of alveolar stop /t/; Top Right: production of velar stop /k/; Bottom: articulator traces from each ROI reveal that, for this subject, apraxic speech is characterized by pervasive gestural intrusions (red arrows), not always evident in the acoustic speech signal because they occur simultaneously with stop target gestures (Hagedorn et al., 2017). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

9. Discussion and outlook

In this review, we surveyed a range of methods that have been employed to analyze data obtained through real-time magnetic resonance imaging (RT-MRI) of human vocal production. We have grouped these methods into four main classes based on three main factors – the type of data processing method, abstractness or specificity of output representation, and the amount of prior or auxiliary information required. These four classes of analysis are based on: (i) the whole-image, (ii) regions-of-interest (ROI), (iii) analysis grids; and (iv) tissue contours. This typology of analysis classes provides a unified way of understanding the various methods deployed in the field. Contextualizing different techniques within this typology helps us understand the motivations and constraints which have influenced the choice and development of these methods in the landscape of linguistic, speech science and clinical applications and real-time MRI technology research (see Tables 2 and 3).

It is important to note that there is no standard recipe or "one size fits all" analysis method for a given dataset. The choice of analysis method(s) depends on multiple factors – first and foremost, the specific research goal of the image analysis in question. While one can apply a wide variety of techniques drawn from the computer vision and pattern recognition community to such data, one should see how such methods and analysis tie back to the specific linguistic, clinical or technological research questions that require a better understanding of the dynamics of vocal tract shaping and the interplay between vocal tract structure and function. Other factors include the prior assumptions of different processing methods and how the data were acquired. This last point is particularly relevant, as it determines data characteristics and limitations such as the spatial resolution, temporal resolution, image artifact and noise patterns, image size and overall data quality for the speech task in question, each of which informs subsequent image analysis (for a more in-depth discussion, see Lingala et al., 2016). The inherent trade-off between spatial and temporal resolution means that in order to capture the rapid movement of articulators, such as those observed in Tamil retroflex and tap consonants, for example, one might need to sacrifice spatial resolution, which in turn affects

subsequent image analysis. Such trade-off choices, along with noise and artifacts can also influence how clearly we can capture other smaller vocal tract structures, such as the velum (Sutton et al., 2010) and epiglottis.

We have further identified certain key issues which much be considered when performing RT-MRI analysis, including reproducibility of analyses across speakers, datasets and studies. Important factors affecting reproducibility include MRI system variability, human error and reliability issues, and speaker-specific variability. Morphological factors, such as the size of the head, vocal tract length and the degree of head movement and rotation, in particular, varies from speaker to speaker and requires attention depending on the analysis in question, especially if one is interested in making generalizations across speakers. Noise-robustness is yet another factor to take into account while designing and implementing analysis techniques, and is crucial to ensuring reproducibility and generalizability of the analyses. Yet another important consideration is the current lack of a gold standard or set of standards for collecting, processing and interpreting RT-MRI data. This could include benchmark datasets for interpreting images, or for different analysis techniques such as image segmentation. This will require a joint effort by the entire RT-MRI analysis community, including the extended community of radiologists/medical practitioners, electrical/computer engineers and speech scientists/linguists.

We will also need such an interdisciplinary effort in order to tackle the many research directions that exist in this still nascent field. We have argued in this paper for a combination of engineering data-driven and linguistic/clinical knowledge-driven methods for RT-MRI image analysis. Going forward, one could use image analysis techniques to inform data acquisition (for instance, using information from a previously acquired high resolution 3D image to inform the choice of subsequent 2D imaging planes in real time). One could use an image atlas, obtained by a high-resolution 3D scan, in conjunction with 2D images to obtain better inferences from the data. Additionally, one could combine such image analyses with the analysis of data from other modalities, such as the speech signal, EMMA, or XRMB, in order to build richer models and understanding of speech production dynamics. Further, such RT-MRI video analysis and processing can inform subsequent higher-level modeling that in turn can provide an understanding of linguistic and/or paralinguistic aspects of speech production. Finally, the wealth of data emerging from real-time MRI, and related modalities, enable new modeling possibilities afforded by contemporary machine learning advances; for instance, deriving insights through end-to-end mapping between acquired data and representations of linguistic and cognitive importance.

Acknowledgments

The authors were supported by NIH grants R01-DC007124 and R01-DC03172, NSF grant #1514544, ARC grant DE150100318, the USC Imaging Sciences Center, the LAC-USC hospital and the USC Center for High Performance Computing and Communications (HPCC). We also thank the USC Ming Hsieh Institute for generously sponsoring the MRI Speech Summit Workshop in 2014, where many of the ideas presented in this paper first crystallized.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.csl.2018.04.002.

References

Arens, R., McDonough, J.M., Costarino, A.T., Mahboubi, S., Tayag-Kier, C.E., Maislin, G., Schwab, R.J., Pack, A.I., 2001. Magnetic resonance imaging of the upper airway structure of children with obstructive sleep apnea syndrome. Am. J. Respir. Crit. Care Med. 164 (4), 698–703. doi: 10.1164/ajrccm.164.4.2101127.

Asadiabadi, S., Erzin, E., 2017. Vocal tract airway tissue boundary tracking for rtMRI using shape and appearance priors. In: Proceedings of the 2017 INTERSPEECH, pp. 636–640.

Atal, B., 1983. Efficient coding of LPC parameters by temporal decomposition. In: Proceedings of the 1983 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'83, 8. IEEE, pp. 81–84.

Badin, P., Bailly, G., Raybaudi, M., Segebarth, C., 1998. A three-dimensional linear articulatory model based on MRI data. In: Proceedings of the Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis.

Bae, Y., Kuehn, D.P., Conway, C.A., Sutton, B.P., 2011. Real-time magnetic resonance imaging of velopharyngeal activities with simultaneous speech recordings. Cleft Palate-Craniofac. J. 48 (6), 695–707. doi: 10.1597/09-158.

Baer, T., Gore, J.C., Gracco, L.C., Nye, P.W., 1991. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels. J. Acoust. Soc. Am. 90 (2), 799–828.

Beautemps, D., Badin, P., Laboissiére, R., 1995. Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: a new model for vowels and fricative consonants based on experimental data. Speech Commun. 16, 27–47.

Beer, A.J., Hellerhoff, P., Zimmermann, A., Mady, K., Sader, R., Rummeny, E.J., Hannig, C., 2004. Dynamic near-real-time magnetic resonance imaging for analyzing the velopharyngeal closure in comparison with videofluoroscopy. J. Magn. Reson. Imaging 20 (5), 791–797. doi: 10.1002/jmri.20197.

Birkholz, P., 2013. Modeling consonant-vowel coarticulation for articulatory speech synthesis. PLoS One 8 (4), e60603.

Birkholz, P., Kröger, B.J., 2006. Vocal tract model adaptation using magnetic resonance imaging. In: Proceedings of the Seventh International Seminar on Speech Production (ISSP 2006), pp. 493–500.

Bresch, E., Katsamanis, A., Goldstein, L., Narayanan, S., 2010. Statistical multi-stream modeling of real-time MRI articulatory speech data. In: Proceedings of the Eleventh Annual Conference of the International Speech Communication Association.

Bresch, E., Narayanan, S., 2009. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. IEEE Trans. Med. Imaging 28 (3), 323–338.

Browman, C., Goldstein, L., 1995. Dynamics and articulatory phonology. In: van Gelder, T., Port, B. (Eds.), Mind as Motion: Explorations in the Dynamics of Cognition. MIT Press, Cambridge, MA, pp. 175–193.

Burdumy, M., 2016. Dynamic Imaging of Singers Using Magnetic Resonance Tomography. Albert-Ludwigs-Universität Freiburg im Breisgau Ph.D. thesis.

Byrd, D., Saltzman, E., 2003. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. J. Phon. 31 (2), 149–180.

Byrd, D., Tobin, S., Bresch, E., Narayanan, S., 2009. Timing effects of syllable structure and stress on nasals: a real-time MRI examination. J. Phon. 37, 97–110.

Carey, D., Miquel, M.E., Evans, B.G., Adank, P., McGettigan, C., 2017. Vocal tract images reveal neural representations of sensorimotor transformation during speech imitation. Cereb. Cortex 27 (5), 3064–3079.

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

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

Chi, L., Comyn, F.L., Mitra, N., Reilly, M.P., Wan, F., Maislin, G., Chmiewski, L., Thorne-FitzGerald, M.D., Victor, U.N., Pack, A.I., Schwab, R.J., 2011. Identification of craniofacial risk factors for obstructive sleep apnoea using three-dimensional MRI. Eur. Respir. J. 38 (2), 348–358. doi: 10.1183/09031936.00119210.

Cootes, T., Taylor, C., Cooper, D., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image Underst. 61, 38–59.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23 (6), 681–685.

Delvaux, V., Metens, T., Soquet, A., 2002. French nasal vowels: acoustic and articulatory properties. In: Hansen, J.H.L., Pellom, B. (Eds.), Proceedings of the International Conference on Spoken Language Processing, pp. 53–56.

Demolin, D., George, M., Lecuit, V., Metens, T., Soquet, A., Raeymaekers, H., 1997. Coarticulation and articulatory compensations studied by dynamic MRI. In: Proceedings of the 1997 EUROSPEECH.

Demolin, D., Hassid, S., Metens, T., Soquet, A., 2002. Real-time MRI and articulatory coordination in speech. C. R. Biol. 325 (4), 547–556.

Demolin, D., Metens, T., Soquet, A., 2000. Real time MRI and articulatory coordinations in vowels. In: Proceedings of the Fifth Speech Production Seminar, pp. 86–93.

Deng, L., 1998. A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. Speech Commun. 24 (4), 299–323.

Deng, L., Ramsay, G., Sun, D., 1997. Production models as a structural basis for automatic speech recognition. Speech Commun. 22 (2), 93–111.

Ding, C.H., Li, T., Jordan, M.I., 2010. Convex and semi-nonnegative matrix factorizations. IEEE Trans. Pattern Anal. Mach. Intell. 32 (1), 45–55.

Drissi, C., Mitrofanoff, M., Talandier, C., Falip, C., Le Couls, V., Adamsbaum, C., 2011. Feasibility of dynamic MRI for evaluating velopharyngeal insufficiency in children. Eur. Radiol. 21 (7), 1462–1469. doi: 10.1007/s00330-011-2069-7.

Echternach, M., Burk, F., Burdumy, M., Traser, L., Richter, B., 2016. Morphometric differences of vocal tract articulators in different loudness conditions in singing. PLoS One 11 (4), e0153792. doi: 10.1371/journal.pone.0153792.

Eide, E., Gish, H., 1996. A parametric approach to vocal tract length normalization. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96, 1, pp. 346–348. doi: 10.1109/ICASSP.1996.541103.

Engwall, O., 2003. A revisit to the application of MRI to the analysis of speech production − testing our assumptions. In: Proceedings of the Sixth International Conference on Speech Production, pp. 43–48.

Engwall, O., 2004. From real-time MRI to 3D tongue movements. In: Proceedings of the 2004 International Conference on Speech Communication and Technology.

Engwall, O., Badin, P., 1999. Collecting and analysing two and three-dimensional MRI data for Swedish. KTH STL-QPSR 3 (4), 011–038.

Eryildirim, A., Berger, M.-O., 2011. A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. In: Proceedings of the Nineteenth European Conference on Signal Processing, pp. 61–65.

Fitch, W.T., Giedd, J., 1999. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. J. Acoust. Soc. Am. 106, 1511–1522. September 1999. doi: 10.1121/1.427148.

Frankel, J., King, S., 2001. ASR-articulatory speech recognition. In: Proceedings of the Seventh European Conference on Speech Communication and Technology.

Freitas, A.C., Wylezinska, M., Birch, M.J., Petersen, S.E., Miquel, M.E., 2016. Comparison of cartesian and non-cartesian real-time MRI sequences at 1.5 T to assess velar motion and velopharyngeal closure during speech. PLoS One 11 (4), e0153322.

Fu, M., Barlaz, M.S., Holtrop, J.L., Perry, J.L., Kuehn, D.P., Shosted, R.K., Liang, Z.-P., Sutton, B.P., 2017. High-frame-rate full-vocal-tract 3D dynamic speech imaging. Magnet. Reson. Med. 77 (4), 1619–1629. doi: 10.1002/mrm.26248.

Fu, M., Zhao, B., Carignan, C., Shosted, R.K., Perry, J.L., Kuehn, D.P., Liang, Z.-P., Sutton, B.P., 2015. High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. Magnet. Reson. Med. 73 (5), 1820–1832. doi: 10.1002/mrm.25302.

Ghosh, P., Narayanan, S., 2011a. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. J. Acoust. Soc. Am. 130 (4), EL251–EL257.

Ghosh, P., Narayanan, S., 2011b. A subject-independent acoustic-to-articulatory inversion. In: Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4624–4627.

Greenwood, A.R., Goodyear, C.C., Martin, P.A., 1992. Measurements of vocal tract shapes using magnetic resonance imaging. IEE Proc. I – Commun. Speech Vis. 139 (6), 553–560.

Hagedorn, C., Lammert, A., Bassily, M., Zu, Y., Sinha, U., Goldstein, L., Narayanan, S.S., 2014. Characterizing post-glossectomy speech using real-time MRI. In: Proceedings of the 2014 International Seminar on Speech Production. Cologne, Germany, pp. 170–173.

Hagedorn, C., Proctor, M., Goldstein, L., 2011. Automatic analysis of singleton and geminate consonant articulation using real-time magnetic resonance imaging. In: Proceedings of the International Conference on Speech Communication and Technology. Florence, Italy, pp. 409–412.

Hagedorn, C., Proctor, M., Goldstein, L., Wilson, S.M., Miller, B., Gorno-Tempini, M.L., Narayanan, S.S., 2017. Characterizing articulation in apraxic speech using real-time magnetic resonance imaging. J. Speech Lang Hear. Res. 60 (4), 877–891.

Hardcastle, W.J., 1972. The use of electropalatography in phonetic research. Phonetica 25 (4), 197–215.

Harshman, R., Ladefoged, P., Goldstein, L., 1977. Factor analysis of tongue shapes. J. Acoust. Soc. Am. 62 (3), 693–707.

Hart, C., Giszter, S., 2010. A neural basis for motor primitives in the spinal cord. J. Neurosci. 30 (4), 1322–1336.

Heinz, J.M., Stevens, K.N., 1964. On the derivation of area functions and acoustic spectra from cinéradiographic films of speech. J. Acoust. Soc. Am. 36 (5), 1037–1038.

Hewer, A., Steiner, I., Wuhrer, S., 2014. A hybrid approach to 3D tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation. In: Proceedings of the 2014 INTERSPEECH, pp. 418–421.

Iltis, P.W., Frahm, J., Voit, D., Joseph, A.A., Schoonderwaldt, E., Altenmüller, E., 2015. High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players. Quant. Imaging Med. Surg. 5 (3), 374.

Israel, A., Proctor, M., Goldstein, L., Iskarous, K., Narayanan, S.S., 2012. Emphatic segments and emphasis spread in Lebanese Arabic: a real-time magnetic resonance imaging study. In: Proceedings of the 2012 International Conference on Speech Communication and Technology. Portland, OR.

Jolliffe, I.I., 2002. Principal Component Analysis. Springer-Verlag, New York.

Jung, T.-P., Krishnamurthy, A.K., Ahalt, S.C., Beckman, M.E., Lee, S.-H., 1996. Deriving gestural scores from articulator-movement records using weighted temporal decomposition. IEEE Trans. Speech Audio Process. 4 (1), 2–18.

Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: active contour models. Int. J. Comput. Vis. 1 (4), 321–331.

Katsamanis, A., Bresch, E., Ramanarayanan, V., Narayanan, S., 2011. Validating RT-MRI based articulatory representations via articulatory recognition. In: Proceedings of the International Conference on Speech Communication and Technology. Florence, Italy.

Kessler, L.G., Barnhart, H.X., Buckler, A.J., Choudhury, K.R., Kondratovich, M.V., Toledano, A., Guimaraes, A.R., Filice, R., Zhang, Z., Sullivan, D.C., 2015. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. Stat. Methods Med. Res. 24 (1), 9–26. doi: 10.1177/0962280214537333.

Kim, J., Kumar, N., Lee, S., Narayanan, S., 2014. Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. In: Proceedings of the 2014 International Seminar on Speech Production, pp. 222–225.

Kim, Y.-C., Proctor, M.I., Narayanan, S.S., Nayak, K.S., 2012. Improved imaging of lingual articulation using real-time multislice MRI. J. Magn. Reson. Imaging 35 (4), 943–948.

Kröger, B.J., Birkholz, P., 2009. Articulatory synthesis of speech and singing: State of the art and suggestions for future research. In: Esposito, A., Hussain, A., Marinaro, M., Martone, R. (Eds.), 2009. Multimodal Signals: Cognitive and Algorithmic Issues. Vol. 5398, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, pp. 306–319.

Kröger, B.J., Birkholz, P., 2007. A gesture-based concept for speech movement control in articulatory speech synthesis. Verbal and Nonverbal Communication Behaviours. Springer, pp. 174–189.

Labrunie, M., Badin, P., Voit, D., Josep, A., Lamalle, L., Vilain, C., Boë, L.-J., Frahm, J., 2016. Tracking contours of orofacial articulators from real-time MRI of speech. In: Proceedings of the International Conference on Speech Communication and Technology, pp. 470–474.

Ladefoged, P., Anthony, J.F.K., Riley, C., 1971. Direct Measurement of the Vocal Tract. J. Acoust. Soc. Am. 49 (1A), 104. https://asa.scitation.org/doi/abs/10.1121/1.1975547.

Lammert, A., Goldstein, L., Narayanan, S., Iskarous, K., 2013a. Statistical methods for estimation of direct and differential kinematics of the vocal tract. Speech Commun. 55 (1), 147–161.

Lammert, A., Proctor, M., Goldstein, L., Pouplier, M., Narayanan, S., 2011. Automatic identification of stable modes and fluctuations in a repetitive task using real-time MRI. In: Proceedings of the International Seminar on Speech Production. Montreal, Canada.

Lammert, A., Proctor, M., Narayanan, S., 2013b. Interspeaker variability in hard palate morphology and vowel production. J. Speech Lang. Hear. Res. 56 (6), S1924–S1933.

Lammert, A., Proctor, M.I., Narayanan, S.S., 2010. Data-driven analysis of realtime vocal tract MRI using correlated image regions. In: Proceedings of the International Conference on Speech Communication and Technology. Makuhari, Japan, pp. 1572–1575.

Lammert, A.C., Narayanan, S.S., 2015. On short-time estimation of vocal tract length from formant frequencies. PLoS One 10 (7), e0132193.

Lee, L., Attias, H., Deng, L., 2003. Variational inference and learning for segmental switching state space models of hidden speech dynamics. In: Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), 1. IEEE, pp. I–872.

Lee, L., Rose, R., 1998. A frequency warping approach to speaker normalization. IEEE Trans. Speech Audio Process. 6 (1), 49–60. doi: 10.1109/89.650310.

Lee, Y., Goldstein, L., Narayanan, S., 2015. Systematic variation in the articulation of the Korean liquid across prosodic positions. In: Proceedings of the International Congress on Phonetic Sciences. Glasgow.

Li, C., Xu, C., Gui, C., Fox, M.D., 2010. Distance regularized level set evolution and its application to image segmentation. IEEE Trans. Image Process. 19 (12), 3243–3254.

Li, M., Kim, J., Lammert, A., Ghosh, P.K., Ramanarayanan, V., Narayanan, S., 2016. Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. 36, 196–211.

Li, T., Ding, C., 2006. The relationships among various nonnegative matrix factorization methods for clustering. In: Proceedings of the Sixth International Conference on Data Mining, ICDM'06. IEEE, pp. 362–371.

Ling, Z.-H., Richmond, K., Yamagishi, J., 2013. Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression. IEEE Trans. Audio Speech Lang. Proc. 21 (1), 207–219.

Ling, Z.-H., Richmond, K., Yamagishi, J., Wang, R.-H., 2009. Integrating articulatory features into HMM-based parametric speech synthesis. IEEE Trans. Audio Speech Lang. Proc. 17 (6), 1171–1185.

Lingala, S.G., Sutton, B.P., Miquel, M.E., Nayak, K.S., 2016. Recommendations for real-time speech MRI. J. Magn. Reson. Imaging 43 (1), 28–44.

Harandi, N.M., Abugharbieh, R., Fels, S., 2015. 3D segmentation of the tongue in MRI: a minimally interactive model-based approach. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. 3 (4), 178–188.

Ma, J., Deng, L., 2004. Target-directed mixture dynamic models for spontaneous speech recognition. IEEE Trans. Speech Audio Process. 12 (1), 47–58.

Mády, K., Sader, R., Beer, A., Hoole, P., Zimmermann, A., Hannig, C., 2003. Consonant articulation in glossectomee speech evaluated by dynamic MRI. In: Proceedings of the Fifteenth International Congress of Phonetic Sciences (ICPhS-15), pp. 3233–3236.

Maeda, S., 1979. An articulatory model of the tongue based on a statistical analysis. J. Acoust. Soc. Am. 65 (S1), S22.

Maeda, S., 1990. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. Speech Prod. Speech Model. Part of the NATO ASI Series book series (ASID, volume 55) 131–149.

Mcdermott, E., Nakamura, A., 2006. Production-oriented models for speech recognition. IEICE Trans. Inf. Syst. 89 (3), 1006–1014.

McGowan, R., 1994. Knowledge from speech production used in speech technology: Articulatory synthesis. Haskins Laboratories Status Report on Speech Research SR-117/118, 25–29.

Mermelstein, P., 1973. Articulatory model for the study of speech production. J. Acoust. Soc. Am. 53 (4), 1070–1082.

Metze, F., Waibel, A., 2002. A flexible stream architecture for ASR using articulatory features. In: Proceedings of the Seventh International Conference on Spoken Language Processing.

Mussa-Ivaldi, F., Gantchev, N., Gantchev, G., 1999. Motor primitives, force-fields and the equilibrium point theory. In: Drinov, M. (Ed.), From Basic Motor Control to Functional Recovery. Academic Publishing House, Sofia, Bulgaria, pp. 392–398.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., Byrd, D., 2004. An approach to real-time magnetic resonance imaging for speech production. J. Acoust. Soc. Am. 115, 1771.

Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.-C., Zhu, Y., Goldstein, L., et al., 2014. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). J. Acoust. Soc. Am. 136 (3), 1307–1311.

Niebergall, A., Zhang, S., Kunay, E., Keydana, G., Job, M., Uecker, M., Frahm, J., 2013. Real-time MRI of speaking at a resolution of 33 ms: undersampled radial flash with nonlinear inverse reconstruction. Magnet. Reson. Med. 69 (2), 477–485. doi: 10.1002/mrm.24276.

Öhman, S.E.G., 1967. Numerical model of coarticulation. J. Acoust. Soc. Am. 41 (2), 310–320.

Olthoff, A., Zhang, S., Schweizer, R., Frahm, J., 2014. On the physiology of normal swallowing as revealed by magnetic resonance imaging in real time. Gastroenterol. Res. Pract. 2014, 1–10. doi: 10.1155/2014/493174.

Ostry, D., Gribble, P., Gracco, V., 1996. Coarticulation of jaw movements in speech production: is context sensitivity in speech kinematics centrally planned? J. Neurosci. 16 (4), 1570–1579.

Perkell, J.S., Cohen, M.H., Svirsky, M.A., Matthies, M.L., Garabieta, I., Jackson, M.T., 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. J. Acoust. Soc. Am. 92 (6), 3078–3096.

Prasad, A., Ghosh, P.K., 2016. Information theoretic optimal vocal tract region selection from real time magnetic resonance images for broad phonetic class recognition. Comput. Speech Lang. 39, 108–128. doi: 10.1016/j.csl.2016.03.003.

Proctor, M., Bresch, E., Byrd, D., Nayak, K., Narayanan, S., 2013a. Paralinguistic mechanisms of production in human beatboxing: a real-time magnetic resonance imaging study. J. Acoust. Soc. Am. 133 (2), 1043–1054.

Proctor, M., Goldstein, L., Byrd, D., Bresch, E., Narayanan, S., 2009. Articulatory comparison of Tamil liquids and stops using real-time magnetic resonance imaging. J. Acoust. Soc. Am. 125 (4), 2568.

Proctor, M., Goldstein, L., Lammert, A., Byrd, D., Toutios, A., Narayanan, S.S., 2013b. Velic coordination in French Nasals: a realtime magnetic resonance imaging study. In: Proceedings of the International Conference on Speech Communication and Technology, pp. 577–581.

Proctor, M., Lammert, A., Goldstein, L., Narayanan, S., 2010a. Temporal analysis of articulatory speech errors using direct image analysis of real-time magnetic resonance imaging. J. Acoust. Soc. Am. 128 (4), 2289. doi: 10.1121/1.3508036.

Proctor, M., Lo, C., Narayanan, S., 2015. Articulation of English vowels in running speech: a real-time MRI study. In: Proceedings of the 2015 International Congress on Phonetic Sciences. Glasgow.

Proctor, M., Walker, R., 2012. Articulatory bases of English liquids. In: Parker, S. (Ed.), The Sonority Controversy. De Gruyter, Berlin, pp. 285–312.

Proctor, M., Zhu, Y., Lammert, A., Toutios, A., Sands, B., Hummel, U., Narayanan, S., 2016. Lingual consonant production in Khoekhoe: a real-time MRI study. In: Shah, S., Brenzinger, M. (Eds.), Proceedings of the Fifth International Symposium in Memory of Henry Honken and Mathias Schladt, Riezlern/Kleinwalsertal – Khoisan Languages and Linguistics. Rüdiger Köppe Verlag, Köln, pp. 337–366.

Proctor, M.I., Bone, D., Narayanan, S.S., 2010b. Rapid semi-automatic segmentation of real-time Magnetic Resonance Images for parametric vocal tract analysis. In: Proceedings of the International Conference on Speech Communication and Technology. Makuhari, Japan, pp. 1576–1579.

Proctor, M.I., Lu, L.H., Zhu, Y., Goldstein, L., Narayanan, S.S., 2012. Articulation of Mandarin Sibilants: a multi-plane realtime MRI study. In: Proceedings of the 2012 Speech Science & Technology.

Raeesy, Z., Rueda, S., Udupa, J.K., Coleman, J., 2013. Automatic segmentation of vocal tract MR images. In: Proceedings of the Tenth IEEE International Symposium on Biomedical Imaging. IEEE, pp. 1328–1331.

Rahim, M.G., Goodyear, C.C., Kleijn, W.B., Schroeter, J., Sondhi, M.M., 1993. On the use of neural networks in articulatory speech synthesis. J. Acoust. Soc. Am. 93 (2), 1109–1121.

Ramanarayanan, V., Bresch, E., Byrd, D., Goldstein, L., Narayanan, S.S., 2009. Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation. J. Acoust. Soc. Am. 126 (5), EL160–EL165.

Ramanarayanan, V., Byrd, D., Goldstein, L., Narayanan, S., 2010. Investigating articulatory setting-pauses, ready position, and rest-using real-time MRI. In: Proceedings of the Eleventh Annual Conference of the International Speech Communication Association.

Ramanarayanan, V., Ghosh, P.K., Lammert, A., Narayanan, S.S., 2012. Exploiting speech production information for automatic speech and speaker modeling and recognition-possibilities and new opportunities. In: Proceedings of the 2012 Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, pp. 1–6.

Ramanarayanan, V., Goldstein, L., Byrd, D., Narayanan, S.S., 2013a. An investigation of articulatory setting using real-time magnetic resonance imaging. J. Acoust. Soc. Am. 134 (1), 510–519.

Ramanarayanan, V., Goldstein, L., Narayanan, S.S., 2013b. Spatio-temporal articulatory movement primitives during speech production: extraction, interpretation, and validation. J. Acoust. Soc. Am. 134 (2), 1378–1394.

Ramanarayanan, V., Katsamanis, A., Narayanan, S., 2011. Automatic data-driven learning of articulatory primitives from real-time MRI data using convolutive NMF with sparseness constraints. In: Proceedings of the Twelfth Annual Conference of the International Speech Communication Association.

Ramanarayanan, V., Lammert, A., Goldstein, L., Narayanan, S., 2014. Are articulatory settings mechanically advantageous for speech motor control? PLoS One 9 (8), 1–8.

Ramanarayanan, V., Van Segbroeck, M., Narayanan, S.S., 2016. Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. 36, 330–346.

Rose, R., Schroeter, J., Sondhi, M., 1996. The potential role of speech production models in automatic speech recognition. J. Acoust. Soc. Am. 99, 1699.

Sagar, P., Nimkin, K., 2014. Feasibility study to assess clinical applications of 3-T cine MRI coupled with synchronous audio recording during speech in evaluation of velopharyngeal insufficiency in children. Pediatric Radiol. 45 (2), 217–227. doi: 10.1007/s00247-014-3141-7.

Sampaio, R.D.A., Jackowski, M.P., 2017. Vocal tract morphology using real-time magnetic resonance imaging. In: Proceedings of the Thirtieth SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 359–366. doi: 10.1109/SIBGRAPI.2017.54.

Scott, A.D., Boubertakh, R., Birch, M.J., Miquel, M.E., 2012. Towards clinical assessment of velopharyngeal closure using MRI: evaluation of real-time MRI sequences at 1.5 and 3T. Br. J. Radiol. 85 (1019), 1083–1092. doi: 10.1259/bjr/32938996.

Scott, A.D., Boubertakh, R., Birch, M.J., Miquel, M.E., 2013. Adaptive averaging applied to dynamic imaging of the soft palate. Magnet. Reson. Med. 70 (3), 865–874.

Shosted, R.K., Sutton, B.P., Benmamoun, A., 2012. Using magnetic resonance to image the pharynx during Arabic speech: Static and dynamic aspects. In: Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, pp. 2182–2185.

Silva, S., Teixeira, A., 2015. Unsupervised segmentation of the vocal tract from real-time MRI sequences. Comput. Speech Lang. 33 (1), 25–46.

Silva, S., Teixeira, A., 2016. Quantitative systematic analysis of vocal tract data. Comput. Speech Lang. 36, 307–329. doi: 10.1016/j.csl.2015.05.004.

Silva, S.S., Teixeira, A.J.S., Oliveira, C., Martins, P., 2013. Segmentation and analysis of vocal tract from midsagittal real-time MRI. In: Proceedings of the 2013 ICIAR.

Singh, A.P., Gordon, G.J., 2008. A unified view of matrix factorization models. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 358–373.

Smith, C., 2014. Complex tongue shaping in lateral liquid production without constriction-based goals. In: Proceedings of the International Conference on Speech Communication and Technology. Cologne, Germany, pp. 413–416.

Sosnik, R., Hauptmann, B., Karni, A., Flash, T., 2004. When practice leads to co-articulation: the evolution of geometrically defined movement primitives. Exp. Brain Res. 156 (4), 422–438.

Stone, M., Davis, E.P., 1995. A head and transducer support system for making ultrasound images of tongue/jaw movement. J. Acoust. Soc. Am. 98 (6), 3107–3112.

Stone, M., Davis, E.P., Douglas, A.S., Aiver, M.N., Gullapalli, R., Levine, W.S., Lundberg, A.J., 2001. Modeling tongue surface contours from cine-MRI images. J. Speech Lang. Hear. Res. 44 (5), 1026–1040.

Strang, G., 2006. Linear Algebra and its Applications. Thomson, Brooks/Cole.

Subtelny, J.D., Oya, N., Subtelny, J.D., 1972. Cineradiographic study of sibilants. Folia Phoniatr. 24 (1), 30–49.

Sutton, B.P., Conway, C.A., Bae, Y., Seethamraju, R., Kuehn, D.P., 2010. Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3T. J. Magn. Reson. Imaging 32 (5), 1228–1237. doi: 10.1002/jmri.22369.

Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., Shosted, R., 2012. Real-time MRI for portuguese. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), Computational Processing of the Portuguese Language. Springer, Berlin/Heidelberg, pp. 306–317.

Tiede, M., Masaki, S., Vatikiotis-Bateson, E., 2000. Contrasts in speech articulation observed in sitting and supine conditions. In: Proceedings of the Fifth Seminar on Speech Production. Kloster Seeon, Bavaria, pp. 25–28.

Tilsen, S., Spincemaille, P., Xu, B., Doerschuk, P., Luh, W.-M., Feldman, E., Wang, Y., 2016. Anticipatory posturing of the vocal tract reveals dissociation of speech movement plans from linguistic units. PLoS One 11 (1), e0146813.

Toda, T., Black, A.W., Tokuda, K., 2004. Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis. In: Proceedings of the Fifth ISCA Workshop on Speech Synthesis.

Töger, J., Lim, Y., Lingala, S.G., Narayanan, S.S., Nayak, K.S., 2016. Sensitivity of quantitative RT-MRI metrics of vocal tract dynamics to image reconstruction settings. In: Proceedings of the INTERSPEECH 2016, pp. 165–169. doi: 10.21437/Interspeech.2016-168.

Vaz, C., Toutios, A., Narayanan, S., 2016. Convex hull convolutive non-negative matrix factorization for uncovering temporal patterns in multivariate time-series data. In: Proceedings of the INTERSPEECH 2016, pp. 963–967.

Vijay Kumar, K., Shankar, V., Santosham, R., 2012. Assessment of swallowing and its disorders: a dynamic MRI study. Eur. J. Radiol. 82 (2), 215–219. doi: 10.1016/j.ejrad.2012.09.010.

Vorperian, H.K., Kent, R.D., Lindstrom, M.J., Kalina, C.M., Gentry, L.R., Yandell, B.S., 2005. Development of vocal tract length during early childhood – a magnetic resonance imaging study. J. Acoust. Soc. Am. 117 (1), 338–350. doi: 10.1121/1.1835958.

Welch, K.C., Foster, G.D., Ritter, C.T., Wadden, T.A., Arens, R., Maislin, G., Schwab, R.J., 2002. A novel volumetric magnetic resonance imaging paradigm to study upper airway anatomy. Sleep 25 (5), 532–542.

Welling, L., Ney, H., Kanthak, S., 2002. Speaker adaptive modeling by vocal tract normalization. IEEE Trans. Speech Audio Process. 10, 415–426.

Westbury, J., Milenkovic, P., Weismer, G., Kent, R., 1990. X-ray microbeam speech production database. J. Acoust. Soc. Am. 88, S56.

Whalen, D., Iskarous, K., Tiede, M., Ostry, D., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E., Hailey, D., 2005. The Haskins optically corrected ultrasound system (Hocus). J. Speech Lang. Hear. Res. 48 (3), 543.

Wrench, A., 2000. A multi-channel/multi-speaker articulatory database for continuous speech recognition research. In: Proceedings of the 2000 Workshop on Phonetics and Phonology in ASR.

Yehia, H., Tiede, M., 1997. A parametric three-dimensional model of the vocal-tract based on MRI data. In: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-97, 3. IEEE, pp. 1619–1622.

Zhang, D., Yang, M., Tao, J., Wang, Y., Liu, B., Bukhari, D., 2016. Extraction of tongue contour in real-time magnetic resonance imaging sequences. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 937–941.

Zhang, S., Olthoff, A., Frahm, J., 2012. Real-time magnetic resonance imaging of normal swallowing. J. Magn. Reson. Imaging 35 (6), 1372–1379. doi: 10.1002/jmri.23591.

Zu, Y., Narayanan, S.S., Kim, Y.-C., Nayak, K., Bronson-Lowe, C., Villegas, B., Ouyoung, M., Sinha, U.K., 2013. Evaluation of swallow function after tongue cancer treatment using real-time magnetic resonance imaging. JAMA Otolaryngol. Head Neck Surg. 139 (12), 1312–1319. doi: 10.1001/jamaoto.2013.5444.