TITLE: Magnetic resonance imaging of the brain and vocal tract: applications to the study of speech production and language learning.

Daniel Carey & Carolyn McGettigan

Department of Psychology, Royal Holloway, University of London, Egham UK

Abstract

The human vocal system is highly plastic, allowing for the flexible expression of language, mood and intentions. However, this plasticity is not stable throughout the life span, and it is well documented that adult learners encounter greater difficulty than children in acquiring the sounds of foreign languages. Researchers have used magnetic resonance imaging (MRI) to interrogate the neural substrates of vocal imitation and learning, and the correlates of individual differences in phonetic "talent". In parallel, a growing body of work using MR technology to directly image the vocal tract in real time during speech has offered primarily descriptive accounts of phonetic variation within and across languages. In this paper, we review the contribution of neural MRI to our understanding of vocal learning, and give an overview of vocal tract imaging and its potential to inform the field. We propose methods by which our understanding of speech production and learning could be advanced through the combined measurement of articulation and brain activity using MRI – specifically, we describe a novel paradigm, developed in our laboratory, that uses both MRI techniques to for the first time map directly between neural, articulatory and acoustic data in the investigation of vocalization. This non-invasive, multimodal imaging method could be used to track central and peripheral correlates of spoken language learning, and speech recovery in clinical settings, as well as provide insights into potential sites for targeted neural interventions.

Highlights

- Phonetic learning is a relatively understudied aspect of foreign language acquisition

- fMRI studies link fronto-parietal cortex (and other regions) to these processes
- Previous studies have used acoustic measures of speech as proxies for motor behavior
- Real-time MRI of the vocal tract offers a means to image phonetic learning *in vivo*
- Multivariate methods can unite vocal tract and brain MRI to probe speech learning

Keywords

Articulation, functional MRI, phonetic learning, individual differences, Representational Similarity Analyses

Acknowledgements

This work was funded by a Research Grant from the Economic and Social Research Council (Award Number: ES/L01257X/1).

1. Introduction

The sensorimotor plasticity of the human brain is essential to the acquisition of spoken language. When learning a second language, or L2, it has been shown that the age at which learning begins has a substantial impact on how native or natural spoken pronunciation of that language sounds (Flege, MacKay, & Meador, 1999; Flege, Munro, & MacKay, 1995). An increased knowledge of learning in the adult vocal system is, in general, critical to the wider understanding of human communicative behaviour, from the flexibility of self-expression in conversation to the recovery of speech and functional reorganization after brain injury. However, speech is one of the most complex actions we perform, with equally complex acoustic consequences. In order to understand its mechanistic underpinnings, we must identify adequate methodology with which to measure and link central neural processes, the actions of peripheral effector systems (i.e. the larynx and articulators) and the acoustic correlates of speech. Magnetic resonance imaging (MRI) offers the opportunity to achieve such a comprehensive account of vocal behavior. In this review, we examine the neuroimaging literature on vocal learning and sensorimotor adaptation, including individual differences in these processes, and point to some of the challenges of assessing speech performance. We provide an introduction to the MRI of vocal tract dynamics as a means of obtaining performance measures more proximal to the motor task of speech production. We outline the applications of vocal tract imaging techniques to date, and propose how these can be incorporated into cognitive neuroscience via a methodological approach in which acoustic, articulatory and neural data can be integrated via analyses of representational similarities.

2. Phonetic learning in the brain: functional and structural underpinnings identified with MRI

Humans are vocal learners, with a sophisticated capacity to volitionally inflect speech and vocalizations dependent on acoustic, linguistic and social contexts (McGettigan, 2015; Pisanski, Cartei, McGettigan, Raine, & Reby, 2016). Imitation of heard speech is largely instinctive to infant language learners, and as a task, is readily achievable for adult spoken language users – so much so that convergence on spoken pronunciation can occur in the absence of awareness (Kappes, Baumgaertner, Peschke, & Ziegler, 2009; Pardo, 2006a; Pardo, Gibbons, Suppes, & Krauss, 2012; Pardo & Jay, 2010; Pardo, 2006b). As part of the acquisition of a spoken language, the learner must engage in the rehearsal of spoken material, demanding the sensorimotor transformation of heard (or imagined) signals into motor plans for execution.

Studies of this basic process, and variability within it, thus bear great relevance for our understanding of phonetic learning.

Although many functional Magnetic Resonance Imaging (fMRI) studies have explored the neural correlates of language production, including investigations of bilingualism and language switching, relatively few have specifically addressed components pertaining to phonetic learning and flexibility. In general, these have reported variation in the structure and function of parts of the speech perception and production networks, related to differences in the production of non-native speech sounds due to learning/training, or individual variability in imitative skill (in monolingual and multilingual talkers). Here, we make distinct those studies that have explored phonetic learning in terms of speech perception and perceptual category formation from those that have focused on the imitation of perceptually discriminable stimuli (and hence are weighted toward audio-motor transformation and execution of speech). In the following sections, we restrict our discussion to the latter category.

2.1 Functional neuroimaging studies of phonetic learning and imitation:

The performance of overt speech imitation and phonological/phonetic manipulations typically involves the flexible manipulation of vocal output, including (but not limited to) the phonemes produced, the language engaged (i.e., L1 vs. L2), and the properties of the voice (e.g., pitch, accent, formant spacing). Neuroimaging studies consistently implicate regions of fronto-parietal cortex in these processes, typically including the opercular part of the left inferior frontal gyrus and the inferior parietal cortex (in particular, supramarginal gyrus; Golestani & Pallier, 2007; Moser et al., 2009; Peschke, Ziegler, Eisenberger, & Baumgaertner, 2012; Reiterer, Hu, Sumathi, & Singh, 2013; Reiterer et al., 2011; Simmonds, Wise, & Leech, 2011), as well as the anterior insula (McGettigan et al., 2013; Moser et al., 2009). These regions are major nodes in the dorsal pathway, described in several models of speech processing as having an important role in sensorimotor transformations for speech comprehension and production (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009; Scott & Johnsrude, 2003). Models of speech production implicate inferior frontal regions and adjacent premotor cortex in the representation of motor representations for learned and familiar articulations (ranging from phonemes to syllables and wordforms, depending on the model; Guenther & Vladusich, 2012; Hickok, 2012; Hickok, Houde, & Rong, 2011; Rauschecker & Scott, 2009; Tourville & Guenther, 2011). Most views posit that speech production is supported by a set of internal feedforward and feedback models, in which the system acts to minimize output errors by comparing the predicted and actual outcomes of speech in auditory and somatosensory cortices. The specific role of inferior parietal cortex varies slightly, with Guenther and colleagues proposing a direct somatosensory processing function, where Rauschecker & Scott (2009) suggest a more general "hub"-like role in matching feedforward signals from inferior frontal sites with sensory feedback signals from posterior temporal cortex. As inferior parietal cortex has been proposed as a putative substrate for the phonological store (Buchsbaum & D'Esposito, 2008; Jacquemot, Pallier, LeBihan, Dehaene, & Dupoux, 2003; Jacquemot & Scott, 2006), this is also considered as a possible role for areas such as the supramarginal gyrus in phonetic imitation (Reiterer et al., 2013; Reiterer et al., 2011). Several of the studies described in the coming sections report activation of the insula during imitation and learning. In the context of clinical speech pathology, a seminal lesion overlap study of apraxia of speech by Dronkers (1996) argued for a crucial role of the tip of the left precentral gyrus of the insula in the motor control of speech – however, its implication in functional neuroimaging studies of speech perception and production, and of language input and output processes more generally, has led others to suggest that the insula may act as a hub for information from speech and language processing areas (Oh, Duerden, & Pang, 2014).

Voluntary, on-demand changes in the sound of articulated speech – for example sounding more masculine or feminine (Cartei, Cowles, & Reby, 2012), or conveying particular personality traits (Hughes, Mogilski, & Harrison, 2014) – are relatively intuitive for human talkers. These behaviours can provide insights into the plasticity of speech systems (Pisanski et al., 2016). McGettigan and colleagues (2013) asked participants to perform spoken impressions of known talkers (from visual prompts) and found, in a comparison of altered versus normal speech, that the left anterior insula and pars opercularis were commonly engaged to emulate general accents/styles of speaking and specific vocal identities. In a similar study, participants performing overt phonological manipulations (i.e. modifying the prosodic or segmental content by producing inflected versions of heard words) showed increased activation of the inferior frontal gyrus (pars triangularis, and extending onto pars opercularis for segmental manipulations only) and the intraparietal sulcus (Peschke et al., 2012).

Garnier and colleagues aimed to investigate the difference between conscious imitation and unconscious phonetic convergence effects in the overt repetition of heard vowels (Garnier, Lamalle, & Sato, 2013). They reported that participants' speech showed significant correlations with target pitch for both overt imitation and phonetic convergence. In an ROI analysis of functional MRI responses, they further reported that the degree of acoustic imitation shown in an individual was positively related to activation of bilateral auditory cortex, left Wernicke's area, and bilateral inferior parietal regions including supramarginal gyrus. Another study found that the degree of participants' relatively unconscious matching to the duration of targets in a rapid speech shadowing task was correlated with activation in the right inferior parietal cortex (in the region of the supramarginal gyrus; Peschke, Ziegler, Kappes, & Baumgaertner, 2009).

Several studies have more directly modeled language learning and proficiency by testing imitation of non-native speech, and measuring time- and learning-dependent changes in neural activity (Moser et al., 2009; Segawa, Tourville, Beal, & Guenther, 2013; Simmonds, Leech, Iverson, & Wise, 2014). Moser and colleagues (2009) found overall increases in the neural response for imitation of heard non-native > English nonwords, where ROI analyses of the left inferior frontal gyrus and anterior insula showed marginal and significant decreases in activation throughout the course of the experiment, respectively; notably, an individual differences analysis revealed a significant correlation between the response of the left anterior insula and the amount of behavioural improvement during the scan. A study investigating the learning-related changes in the overt pronunciation of phonotactically illegal pseudowords in English (e.g. GVAZF) compared reading aloud of trained and novel items in fMRI (Segawa et al., 2013). Here, novel items generated greater activation in both frontal opercula, supplementary motor area (SMA), left superior temporal cortex and bilateral superior parietal lobule, as well as the globus pallidus. Simmonds and colleagues (2015) scanned participants before and after an intensive 1-week self-administered programme of foreign language pronunciation training. They reported a significant increase in the BOLD response during imitation of non-native (vs native) speech, in bilateral auditory and motor cortices, the basal ganglia, thalamus and cerebellum. However, while these authors also report activations in bilateral frontal operculum and inferior parietal lobule for this contrast, they attribute this engagement not to sensorimotor processes but rather to cingulo-opercular salience and fronto-parietal central executive networks, respectively. Activity of bilateral anterior striatum was enhanced during the initial phase of the first scanning session, and larger in the pre- versus post-training scans. The authors argue that, in line with evidence from songbirds, the striatum plays a key role in the acquisition of novel motor sequences for speech. Simmonds (Simmonds, 2015) has suggested that a vocal learning pathway involving the striatum facilitates native-like speech learning through the introduction of variability in speech production (akin to "babbling"), and that the reason for impaired ability to acquire good pronunciation in adulthood is due to a fall-off in the recruitment of this path in users of an overlearned native language. While striatal activation in their learning study was independent of performance measures (ratings of pronunciation accuracy by native speakers of the imitated languages), there was an indication that, in these adult monolingual speakers, variation in performance accuracy was instead associated with activation in the bilateral frontal opercula, as well as regions of left superior temporal cortex.

2.2 Individual differences: Phonetic "talent" in the adult brain

Given the large variability reported in the persistence of foreign accent in the L2 (Piske, MacKay, & Flege, 2001), there has been an interest in exploring vocal imitation in terms of the neural substrates of individual differences in "phonetic talent" (or "phonetic compliance", Delvaux, Huet, Piccaluga, & Harmegnies, 2014) as a relatively stable characteristic within an individual. In the main, these have revealed striking consistency with the results described above. For example, a study measuring individual differences in the immediate imitation of a non-native plosive demonstrated variation in the white-matter density of left anterior insula, and bilateral inferior parietal cortex (Golestani & Pallier, 2007).

Reiterer and colleagues have published several neuroimaging studies where the basis of the design has been to divide participants according to existing measures of pronunciation ability in a non-native language (Hu et al., 2013; Reiterer et al., 2011; Reiterer et al., 2013). From a large cohort of 140 "late" bilinguals (German participants who began learning English at around 10 years old), Reiterer and colleagues (2011) compared speech imitation (using Tamil, German and English) in the top and bottom 15% of performers (identified through a Hindi imitation pre-test). Here, low performers showed greater activation in left inferior frontal gyrus (including the opercular part), and in left and right inferior parietal cortex (including supramarginal gyrus), compared with high performers. Examining the correlation of activation strength with Hindi imitation ability, the authors reported that lower performance was associated with increased responses in left inferior frontal gyrus and supramarginal gyrus. In a study of voluntary foreign accent imitation during reading aloud, Reiterer and colleagues (2013) again reported increased signal for low performers in (left-dominant) fronto-parietal cortex, as well as the basal ganglia (with peaks in left and right caudate). Finally, Hu and colleagues (2013) reported group contrasts of functional activation during imitation of heard English and German sentences, on this occasion defining aptitude based on pronunciation of participants' familiar L2 (English). They found increased activity in the high performance group, in a wide range of brain regions including several elements of the fronto-parietal dorsal speech pathway (including left lateralized IFG and anterior insula), as well as bilateral SMA, left caudate, bilateral thalamus and right cerebellum.

2.3. Development of pronunciation expertise: L2 production in early vs. late bilinguals, and across the lifespan

When considering the challenge of phonetic flexibility for language learning, some insights can be drawn by looking at speakers of multiple languages, particularly with reference to the emergence of expertise in an L2. Simmonds and colleagues (Simmonds, Wise, Dhanjal, & Leech, 2011) examined late bilingual speakers whose common L2 was English. Comparing activations during spontaneous speech in L1 and L2, the authors found increased activation for L2 speech production in left inferior frontal gyrus, motor cortex, superior temporal and inferior parietal cortices, bilateral SMA, and the cerebellum, as well as bilateral globus pallidus and thalamus. The authors focus in particular on the engagement of planum temporale and the parietal operculum as evidence for closer monitoring of auditory and somatosensory outputs during the less automatic production of the L2. In a study of reading aloud, Berken and colleagues (Berken et al., 2015) compared responses to L2 (>L1) in early ("simultaneous": i.e., those who learned their two languages from birth) and late ("sequential") bilinguals. They found increased activation in sequential bilinguals in left inferior frontal gyrus and premotor cortex, compared with both simultaneous bilinguals and monolinguals speaking the same language; these regions also showed a positive correlation with age of acquisition of the L2. Again, the authors interpret the finding in terms of increased articulatory effort in emulating a native-like accent in the L2, where this is more difficult for those who have learned the language later in development.

While it is commonly accepted that, as demonstrated above, late acquisition of a second language often incurs greater processing effort and the presence of a persisting non-native accent, the developmental literature on *immediate* phonetic imitation tells a different story. Hashizume and colleagues (Hashizume et al., 2014) carried out a functional imaging study investigating the previously documented finding that initial attempts at imitation of a foreign language are more successful in older children. In a cross-sectional analysis, they found increased L2 > L1 differences in the left frontal operculum as children grow older – the authors interpret this effect in terms of the developmental maturation and specialization of this region for speech motor control. More broadly, the maturation of expressive language in children has been tied to widespread adaptations to cortical processes in diffuse brain networks, even for L1 alone. A recent meta-analysis showed that when controlling for differences in task performance, maturation of children's expressive language was associated with increased motor cortical activity, but decreases in left lateral and right medial prefrontal activity, along with decreases in posterior cingulate and extra-striate cortical activity (Weiss-Croft and Baldeweg, 2015). Recent functional imaging evidence has further suggested that linguistic complexity may serve to modulate the recruitment of lateral prefrontal regions in children during expressive language, but with decreased activation in these regions as complexity increases (in contrast to increasing activation for adults; Krishnan et al., 2015). Taken together, these findings suggest that developmental trajectories for speech production implicate an increasing specialization of cortical networks, with frontal regions in particular modulated by speech complexity. In this regard, the left inferior frontal gyrus appears to have a particularly important role with respect to L2 production.

2.4 Phonetic imitation and learning in the brain - Summary

Researchers have probed the neural correlates of phonetic learning and flexibility in a variety of ways: comparing volitional versus unconscious imitation of native speech, examining the timecourse of non-native phonetic learning (with and without training), comparing "phonetic talent" as a stable characteristic across individuals and groups of talkers, and studying L2 production in bilinguals and in development. Despite the diversity of approaches, there have been some relatively consistent findings in terms of the neurobiological substrates of vocal learning in speech. In the cortex, the frontal operculum, anterior insula and portions of the inferior parietal cortex have been implicated in learning-related changes in activation, overall modulation by sensorimotor task difficulty (non-native > native, L2 > L1) and variation according to inter-individual differences in task performance. These findings have been interpreted mainly in terms of motor planning and execution (inferior frontal cortex and insula) and sensorimotor transformations in speech (inferior parietal cortex), supporting the monitoring of somatosensory outcomes of unfamiliar speech and the engagement of novel articulatory routines (or the repurposing of existing routines into unfamiliar sequences).

However, none of the studies cited in this section provided direct insight into the articulatory gestures involved in speech production, and instead, reported measures of acoustic speech outputs. As it stands, the neuroscientific literature on phonetic learning for acquisition of an L2 therefore faces a crucial challenge. In speech, actions of the intercostal muscles, laryngeal muscles and/or the articulators in the upper vocal tract, must precede sounds. Yet, to date, we have no account of how variations in these actions – the fundaments of speech production – are represented in the brain. This is particularly pertinent when we consider specific conditions such as expressive aphasias or stammering, some of whose symptoms are marked by the absence of speech sounds, thus offering no possibility for acoustic analysis (e.g. speech "blocks" in stammering). Here, it is crucial to obtain insights into the underlying articulations and their relationship to neural function, if we are to have a better understanding of pathology and plasticity in neural speech production systems. In the next sections, we present real-time MRI (rtMRI) as a viable dependent variable for tracking phonetic performance in language learning, and propose methods by which vocal tract images, as well as acoustic productions, can be unified with brain activation data from functional MRI in order to more comprehensively describe the processes supporting vocal flexibility and pronunciation learning, as well as individual differences in speech production.

3. Real-time vocal tract MRI – a window onto the dynamics of speech articulation

Analysing vocal tract dynamics during speech presents a complex and multidimensional set of challenges for researchers. In addressing these challenges, a variety of real-time MR imaging techniques have been developed that allow the entire vocal tract to be imaged non-invasively, and afford millimetre spatial resolution alongside temporal precision at the requisite rates for capturing speech (e.g., 20 frames/s). Moreover, considerable advances in the offline analysis approaches available to quantify vocal tract dynamics have greatly improved our understanding of speech production, with applications to date in the fields of phonetics, phonology and speech pathology (see Section 3.3 below). A variety of other techniques have been employed in imaging speech, both for clinical and research purposes. Electropalatography (EPG), where a set of sensors within a retainer-like device are used to detect points of contact between the tongue and palate, has been widely employed in therapeutic speech interventions with developing populations (see Dagenais, 1995), and in characterising phonetic variation in tongue-palate contact both in typical developmental (Cheng et al., 2007) and in adults (Gibbon et al., 2007; McLeod, 2006). Electromagnetic articulography (EMA) has similarly been applied in clinical settings (for instance, treating apraxia of speech in an expressive aphasia; Katz et al., 1999), as well as in studies of speech rate in healthy adults (Goozée et al., 2003; 2005), and in characterizing inter-subject variation in vocal tract morphology and articulatory performance in phonetic context (Weirich & Fuchs, 2013; Guenther et al., 1999). EMA enables the visualisation of the position of several articulators (e.g., tongue tip or blade relative to the palate), and the tracking of articulatory temporal dynamics (Schoenle et al., 1987). Ultrasound has been employed widely in clinical settings, notably in the assessment of post-surgical clinical speech outcomes (e.g., Bressman et al., 2007), and can enable visualisation of tongue dynamics using 3D reconstructions (see Bressman, 2010). While these techniques offer clinical utility and the potential to visualise facets of articulation, each is limited with respect to real-time MRI. In particular, none of the techniques afford the spatial *and* temporal resolution across the entirety of the vocal tract (encompassing the lips, tongue, hard and soft palate, and larynx) that can be achieved with rtMRI. Moreover, EMA and EPG both require detector apparatus to be worn inside the mouth, a clear departure from naturalistic speech production; ultrasound similarly may entail disruptive effects during speech, due to detector contact against the lower jaw. Importantly, MR techniques also avoid many of the disadvantages of other imaging methods, such as Xray or Computed Tomography (i.e., exposure to radiation) (Bresch, Kim, Nayak, & Byrd, 2008; Narayanan, Nayak, Lee, & Byrd, 2004).

In this section, we provide a brief overview of some of the vocal tract MR imaging techniques currently in use, together with a review of advances in automated vocal tract measurement and quantification.

3.1 Vocal tract MR - sequences & acquisition parameters

Imaging the vocal tract entails an array of technical issues. Chief amongst these are ensuring adequate spatial signal-to-noise ratio (SNR) of the imaged volume during speech and obtaining adequate temporal precision of the imaging frame rate (Baer, Gore, Gracco, & Nye, 1991). A variety of MR sequences are now in use for vocal tract imaging, with each providing different extents of spatial and temporal resolution, as well as varying artifact sensitivity. Commonly used 2D acquisitions include Cartesian sequences (single-slice gradient echo, e.g., Silva & Teixeira, 2015a) spiral acquisitions (e.g., Bresch & Narayanan, 2009; Narayanan et al., 2004), and radial acquisitions (e.g., Niebergall et al., 2013). Approaches to imaging the vocal tract in 3D have typically employed multi-slice T_1w acquisitions (e.g., Badin et al., 2002; Baer et al., 1991). Obstacles to achieving high-quality MR images regularly emerge during vocalisation, and include magnetic susceptibility-induced image distortions at air-tissue boundaries, image blurring due to magnetic off-resonance effects, and non-speech head motion artifacts (see Bresch et al., 2008). While non-speech head motion can be mediated with padding and appropriate subject instruction, susceptibility and off-resonance artifacts largely reflect the image acquisition parameters. Such vocal tract image artifacts may be reduced by manipulation of a variety of parameters within each sequence type; for instance, Narayanan et al. (2004) used very short (< 3 ms) gradient readout durations, to reduce phase build-up and resulting off-resonance during spiral imaging. More recent advances in parallel imaging techniques (where images are reconstructed from multiple phased array coils in parallel) have greatly facilitated MR image acquisition, shortening scanning times with negligible losses in image quality (e.g., Kim, Narayanan & Nayak, 2010).

The type of sequence chosen for vocal tract imaging will largely depend on the research question to be addressed and concomitant technical requirements (e.g., frames per second, spatial resolution), together with considerations of the offline processing to be conducted. For instance, high image frame rates (33 frames/s) with good spatial resolution (3 x 3 mm^2) may be achieved via spiral imaging with sliding window reconstruction (where frames are sampled as soon as they are acquired; Hagedorn, Proctor, Goldstein, & Narayanan, 2011; Narayanan et al., 2004). A combination of high spatial resolution (1.5 x 1.5 mm^2) *and* fine temporal precision (30 frames/s) may be demanded when examining dynamics of small structures, such as the vocal folds; here, 3D radial sequences allied with parallel imaging may afford the most spatially and temporally precise results (Niebergall et al., 2013).

3.2 Vocal tract MR - automated segmentation & analyses

Early methods adopted for quantitative vocal tract analysis of MR images typically comprised the manual delineation of tissue boundaries (e.g., Baer et al., 1991). Further approaches have applied descriptive analyses of articulatory dynamics, based on visual inspection of articulator position on a frame-by-frame basis (e.g., Proctor, Bresch, Byrd, Nayak, & Narayanan, 2013). While suitable at the single subject level, such analyses at a group level are often time and labour intensive, and may be liable to inter-subject error. In recent years, computational advances have afforded robust and time-efficient automated methods for measurement of tissue boundaries, enabling the quantification of vocal tract dynamics across frames and in larger cohorts of subjects. Analysis approaches common to several automated frameworks involve tracking the movements of vocal tract tissue types (i.e., surfaces of the articulators) across frames. Such approaches begin with the assignment of landmark points by the observer (commonly the larynx, hard palate, alveolar ridge and lower lip), followed by imposition of a semi-polar, regularly spaced gridline system across the vocal tract (Kim, Kumar, Lee, & Narayanan, 2014; Proctor, Bone, Katsamanis, & Narayanan, 2010). This system of gridlines forms an approximately sigmoidal shape, covering the airway. Using intensity-based algorithms that map the change in image intensity between upper and lower vocal tract boundaries at each gridline (i.e., transition from tissue, to airway, back to tissue), the airway path may be estimated, and in turn, tissue boundary distance traces along the vocal tract can be measured (Proctor et al., 2010). Recent improvements to this procedure have adopted image pixel enhancement (to reduce grainy noise) prior to airway boundary estimation, and involve additional discrete stages where labial and laryngeal position are tracked across frames, independently of vocal tract tissue boundary measurement (Kim et al., 2014) (see Fig.1). Normative data provided within Kim et al. (2014) highlight the robustness of the toolbox algorithms across a range of speech phones. The method can be further enhanced by analyses of synchronous audio output of speech; improvements in available recording technology and MRI scanner noise cancellation algorithms now allow for direct comparison of speech output as produced in the scanner during rtMRI acquisition (e.g., Bresch, Nielsen, Nayak, & Narayanan, 2006).

More elaborate computational approaches have been developed recently, imposing specific model predictions concerning vocal tract morphology and articulator position (e.g., Bresch & Narayanan, 2009; Silva & Teixeira, 2015a). In one such approach, the model is trained *a priori* on a corpus of real-time images; vocal tract data from a given subject may then be segmented based on the model’s assumptions regarding vocal tract morphology (Silva & Teixeira, 2015b). This segmentation procedure based on model fitting yields a robust set of vocal tract tissue boundaries that can accommodate different vocal tract configurations (e.g., vowels), and parses the vocal tract into distinct articulator subclasses (i.e., lips, tongue, velum, pharynx, etc.). Positions of, and relative distances between, articulators can then be quantified according to a unit circle method that allows for representation of multiple distance vectors across the articulators (Silva & Teixeira, 2015b).

While the quantification methods described above vary in their instantiation and complexity, the goal of each is to index tissue position and relative or absolute vocal tract distances in an objective manner. Further approaches can involve analysing image pixels directly, such that intensity variation allows tissue movement to be captured. For instance, pixel intensity change methods have been used to identify constriction location in the vocal tract (Hagedorn et al., 2011). Approaches that enhance image signal-to-noise (SNR) prior to analysis may be useful in such cases, particularly in regions where MR signal intensity is low (e.g., the alveolar ridge). The adaptive averaging procedure of Scott et al. (Scott, Boubertakh, Birch, & Miquel, 2013) provides an elegant solution to such SNR problems. Using a region-specific pixel cross-correlation procedure, pixels of high similarity are identified across frames, and these frames are averaged together to improve SNR. In latter sections of this review, we demonstrate that this adaptive averaging technique may be used to create trial-by-trial summary vocal tract images, which can then be used within multivariate analysis approaches to probe neural fMRI data.

3.3 rtMRI and Speech - Applications of vocal tract image quantification

A major goal of quantifying vocal tract dynamics during articulation is to develop a more concrete set of parameters with which we can better understand speech production. Applying quantitative vocal tract analyses to speech can help to account for key differences in vocal tract movements; for instance, within and across utterance subtypes. Moreover, the potential to measure the vocal apparatus affords the opportunity to chart important changes in vocal tract morphology with age. We review some of these applications below.

In adopting a quantitative approach to vowel analysis, traces of tongue tissue boundaries during articulation can allow us to identify articulatory profiles specific to different vowel species. In line with phonetic measurements of German vowels, Niebergall and colleagues (2013) found that tongue blade and tip position tended to lie more anteriorly for vowels such as /i/, /e/ and /ɛ/ (phonetic front vowels), but were retracted for vowels like /u/ and /o/ (phonetic back vowels). More importantly, their results indicated that tongue position varied as a function of utterance complexity: tongue blade and tip were less retracted for these same back vowels when they were produced in sentence contexts, versus in isolation. These data suggest a direct insight from quantitative rtMRI tissue metrics into the effects of coarticulation (i.e., impact of consonants on following vowels) on tongue dynamics during vowel articulation.

In Figure 1, using segmentations of the vocal tract (via the method of Kim et al., 2014), we show additional applications of quantitative tissue tracking algorithms to indexing articulatory profiles of different vowels (illustrated in a single subject). Tissue boundary distance traces allow us to measure local changes in vocal tract aperture size at different points along the vocal tract (e.g., Fig. 1(i) charts the greater distance anteriorly for /a/ and /u/, relative to /i/). Moreover, tracking of lip dynamics across frames enables us to measure the greater labial tissue excursion for the rounded /u/, versus the unrounded /i/ and /a/ (see Fig. 1(ii)). In line with the work of Niebergall et al., these metrics afford further means by which articulatory dynamics may be indexed over multiple articulators, to provide quantitative insights into speech production differences across utterances.

Further approaches to measuring vocal tract dynamics have sought to compare linguistic and non-linguistic movements by capturing regional variation in vocal tract tissue position. Vasquez Miloro et al. (2014) compared dynamics of a series of vocal tract tissues (tongue hyoid muscles, laryngeal elevation) during production of a paralinguistic ‘effortful pitch glide’ (i.e., raising vocal pitch while contracting the pharynx, by producing a forceful /i/ vowel), and during swallowing. Notably, their results indicated no significant differences in the extent of tissue excursion between the effortful pitch glide and the non-linguistic swallowing condition. These data may suggest clinical utility of the effortful pitch glide, given the similar extents of tissue dynamics involved in both cases (for instance, using effortful pitch glide as a rehabilitation strategy for dysphagia; Vasquez Miloro et al., 2014). Such insights would not have been possible without real-time MR data and the derived movement metrics.

Real-time MRI may also prove insightful when exploring non-linguistic constituents of speech, such as emotion. For instance, Lee et al. (2006) explored articulatory dynamics with rtMRI in a single subject, as the subject produced happy, sad, angry, and neutral versions of sentences. Quantitative measurements of the real-time data revealed that vocal tract shape showed greater variability during angry speech, compared to the other conditions. Further, overall vocal tract length spanned a considerably larger range for happy versions of the sentences, versus the other conditions (likely due to changing laryngeal height). While limited to a single subject, these data nevertheless illustrate the potential to develop quantitative indices of non-linguistic aspects of speech articulation, such as emotional state, using rtMRI of the vocal tract.

The study of developmental change in the vocal tract with MRI is also greatly enhanced by quantitative analysis of vocal tract morphology. For instance, the extent to which growth of vocal tract structures is uniform or regionally variable between birth and early childhood was probed by Vorperian et al. (Vorperian et al., 2005), using linear distance traces for a range of vocal tract structures (e.g., mandible, larynx, tongue, hard and soft palate). Their results indicated a large-scale increase in total vocal tract length in early childhood, together with periods of relatively accelerated length growth during the first two years in specific vocal tract regions (including the hard palate). Further extension of these growth trends into puberty has been charted based on quantitative analysis of regionalised vocal tract distance traces, measured from MRI images (see Fitch & Giedd, 1999). Further evidence (from EMA) has suggested that variation in vocal tract morphology across subjects interacts with articulatory performance, such that subject-wise articulatory profiles may differ as a function of both vocal tract morphology and utterance contextual factors (Weirich & Fuchs, 2013; see also Guenther et al., 1999).

While the above studies suggest clear benefits of vocal tract imaging when addressing questions related to vocal tract morphology, and particularly speech dynamics across utterances, surprisingly little research has focused on plasticity or learning in speech production using rtMRI of the vocal tract. As we note in earlier sections, considerable challenges face language learners of all ages, and notably in adulthood, when mastering the articulatory dynamics of unfamiliar speech sounds. If we are to study speech as a truly multidimensional behaviour, then charting learning outcomes in terms of articulatory dynamics, speech acoustics *and* brain function affords a more holistic and insightful means of understanding speech plasticity. In the next section, we review novel approaches to combining articulatory, acoustic and neural indices of speech learning, to provide a unified account of plasticity in speech production.

4. MR imaging of the vocal tract and brain - New analysis approaches

Learning to produce the speech sounds of a new language involves a highly complex set of sensorimotor processes, both prior to and during speech output. These include the encoding of speech stimulus acoustics, transformation from a sensory to a motor target, and execution of an articulatory motor plan whilst monitoring one's on-going speech output. The complexity involved in speech learning obliges us to use analysis approaches that can probe the contribution of acoustic, neural *and* articulatory processes to learning outcomes. Further, the potential to explore individual differences that relate to success of vocal learning is enhanced if we adopt analyses that unite these multidimensional components of speech production.

As a first approach, vocal learning in a new language may be indexed directly from the vocal tract via rtMRI methods. Measurements of speech dynamics made from rtMRI data (using methods described in the preceding section) allow us to assay production behavior in a quantitative manner – e.g., by indexing changes in lip and laryngeal position, and profiles of tongue position, across utterances. Such measures can afford direct insights into whether the articulatory dynamics that underpin non-native vowel production (e.g., changes in tongue position, lip rounding, etc.) are acquired during learning.

In seeking to combine data from multiple modalities of speech imaging (e.g., rtMRI, fMRI), individual differences analyses can provide fruitful insights, particularly for learning of novel speech. Correlational analyses can allow us to test the relationships between neural activity during speech production and the attendant articulatory kinematics, as indexed with rtMRI. Such analyses may prove useful in instances where BOLD activation in fMRI of speech is predicted to vary according to systematic patterns of articulation. For instance, difficult-to-observe articulatory behavior, such as raising or lowering of the larynx during non-native speech pitch modulations, could be indexed with rtMRI; in turn, these indices of laryngeal dynamics could be used to predict individual differences in univariate BOLD activation, for fMRI data collected under the same conditions of novel vocal pitch behavior.

While correlational analyses may prove useful where articulatory behaviour is expected to vary systematically for a single articulator, the production of speech necessitates the combined involvement of many vocal tract effectors. Moreover, correlational analyses are necessarily limited to testing bivariate relationships. These considerations mean that elementary correlational analyses cannot capture the dynamism of the entire vocal tract at once in a single analysis. Further, bivariate analyses do not allow a parallel appraisal of the relationships between speech acoustics, articulatory dynamics, and neural activation. Here, we review a novel data-driven analysis framework that we have devised, which permits the combined analysis of speech acoustics, articulatory data (i.e., rtMRI data covering the entire vocal tract) and functional brain data (i.e., fMRI). Our approach utilises the multivariate statistical technique of Representational Similarity Analysis (RSA) to explore data in representational terms; this allows for qualitatively different data types (e.g., acoustic spectra, vocal tract images, fMRI data) to be compared and tested in a quantitative manner. We begin with a brief overview of RSA, and then propose the framework within which we employ the technique. Applications to the study of plasticity in speech production are considered, along with possible clinical implications of our framework.

4.1 Representational Similarity Analysis (RSA)

As proposed by Kriegeskorte and colleagues (Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013), RSA aims to identify patterns of similarity that emerge from data; this is achieved by cross-correlating data across all possible pairs of conditions present in a design. The relationships in RSA are calculated as Pearson Product moment correlations between the data in each of the different condition pairs; this yields a matrix of correlation coefficients, where each cell in the matrix reflects the relationship between the data for a given pair of conditions. Kriegeskorte et al. (2008) propose using *correlation distance* to quantify these relationships between conditions (i.e., 1 minus each correlation coefficient in the matrix). This matrix of correlation distance values is known as a Representational Dissimilarity Matrix (RDM; see Fig. 2). In the RDM, condition pairs where data are highly similar to each other have cell values close to 0, while conditions with more dissimilar data have cell values closer to 1 (i.e., inverse of a Pearson correlation). The RDM diagonal is 0, since each condition relates to itself identically.

As a technique, RSA is amodal, and therefore permits comparison of qualitatively different data (e.g., acoustic data and neural) that reflect the same conditions. This is possible because the comparison between different data types is made at the abstract *representational* level. We achieve this by comparing the RDMs that we derive from the different data types over the same experimental conditions. Critically, this comparison at the representational level can be quantified; that is, we can determine the statistical relationship between a RDM from one data type and another data type (e.g., using Spearman or Kendall's Tau correlations). It is this amodal, yet quantifiable, representation of similarity between conditions, *and across data types*, that makes RSA a powerful and insightful multivariate tool with which to study speech production in a multidimensional framework.

To date, RSA has been applied to questions surrounding speech perception (e.g., Evans & Davis, 2015), and higher-level language semantics (e.g., Devereux, Clarke, Marouchos, & Tyler, 2013). However, to our knowledge, no RSA study of speech so far has adopted a framework in which RDMs derived from speech acoustics and real-time images of speech articulation are used to probe neural representations within fMRI data. Moreover, little – if any – work has considered how each of these representations might vary with respect to plasticity of speech production. Below, we provide an overview of our analysis framework, which seeks to analyse these data types with RSA. We consider possible applications of this approach to studies of speech learning, as well as the potential clinical utility of the approach.

4.2 rtMRI and RSA: A combined analysis framework.

To assay acoustic and articulatory representations of speech, we adopt a data-driven RSA approach with respect to the category boundaries that define the conditions in our design. Here, the relationships that reflect these category boundaries are allowed to emerge directly from the data. This contrasts with conceptual approaches to RSA, which specify models where category boundaries are defined explicitly by the researcher (e.g., Evans & Davis, 2015). Figure 2 presents sample data for this approach, using the vowels /i/, /a/ and /u/; note that these data are intended as an illustration of the technique, and do not reflect the results from a specific study.

We begin by processing the vowel acoustic and articulatory data using standard Matlab routines. Vowel stimulus audio files are cropped about their midpoint, to create a set of audio files of equal duration per condition. The vowel stimulus acoustic spectra can then be extracted in Matlab, by taking the power spectral density matrix derived from the Goertzel-algorithm estimate of each cropped vowel's spectrogram. We can then then cross-correlate the spectra of the vowel stimuli on a token-wise basis (i.e., by comparing the PSD matrices across all possible stimulus pairs). We process real-time MR images of vowel articulation as follows. First, the series of image frames from the steady-state portion of the vowel in a single trial are averaged together, using the adaptive averaging method of Scott et al. (2013). For each trial per condition, we take the middle frame of the trial-averaged series, and then mask this image to restrict the field-of-view to the vocal tract (minimising non-vocal tract tissue, as much as possible; see Fig 2, top left). If averaging of image frames across scanning sessions is required, we apply an intensity-based rigid body registration before masking, to ameliorate effects of head motion.

Following the preparation above, the acoustic and articulatory data can be processed and analysed using the freely available RSA toolbox (http://www.mrc-cbu.cam.ac.uk/methods-and-resources/toolboxes/). Briefly, the toolbox separately reads each audio and rtMRI file into Matlab in matrix form; the matrices are then vectorised, and cross-correlated over each condition. RDMs for each data type can then be visualised easily, as in the case of the vowels presented here (see Fig. 2, top); note the close overall match of the RDM patterns for these vowels, despite us having constructed the RDMs from entirely different data (rtMRI and acoustic).

Central to our analysis framework are these acoustic and articulatory RDMs, which we use as models that we test against speech fMRI data from the corresponding conditions. In this way, we can ask whether the neural relationships across conditions within specific regions of the brain bear significant relation to the acoustic *and* articulatory relationships across those same conditions.

Several approaches can be used to address these questions. If a specific brain region-of-interest (ROI) exists where the functional representation of speech production conditions is expected to be homogenous, then a neural data RDM may be built by cross-correlating the condition-wise activity across the entirety of the ROI. This ‘whole ROI’ neural RDM can then be compared to the acoustic and articulatory RDMs separately, via Spearman correlation. Kriegeskorte et al. (2008) propose evaluating the correlation between RDMs as 1-correlation coefficient (i.e., ‘a dissimilarity of dissimilarity matrices’).

Alternatively, if homogenous neural representations are not guaranteed over the full extent of a speech production ROI, we can refine the spatial scale of the analysis using a RSA searchlight approach (Kriegeskorte, Goebel, & Bandettini, 2006). Here, we can use the acoustic and articulatory RDMs to probe the representational patterns that emerge from small ‘snapshots’ of neural data within an entire brain volume, or within a speech ROI. We move incrementally through the volume according to the searchlight size (e.g., a radius of several mm), determining a RDM from the activity in the voxels in each searchlight; we then compare this searchlight RDM to the acoustic and articulatory RDMs. This results in a set of maps for each subject, describing the local correlations between searchlight and acoustic/articulatory RDMs; second level group statistics can then be performed. This provides a spatially more refined means of targeting representational analyses, particularly for speech production, where a volume or ROI may comprise tens of thousands of voxels. While whole-brain RSA searchlight analyses are feasible and commonly used (e.g., Evans & Davis, 2015), they are computationally more intensive, and typically require more stringent corrections to control for type 1 error.

4.3 rtMRI and RSA: applications to speech plasticity.

In the current review, we present RDMs derived from vowel stimulus acoustics, and from real-time vocal tract images of a native English speaker imitating these same stimuli (Fig 2). In this case, imitation was confined to vowels that are native to English. We might then ask how the articulatory representational pattern would look like if subjects imitated unfamiliar non-native vowels as well as these native vowels. Moreover, we can consider whether the articulatory RDM would resemble the acoustic RDM derived from the native and non-native vowel stimuli. Most importantly, we could consider whether these acoustic and articulatory RDMs would map onto searchlight RDMs derived from fMRI data, for subjects imitating these native and non-native vowels. Finally, we could explore whether these RDMs derived from neural data would change over time, if learning were to occur within the MRI scanner.

We are currently addressing these questions via a learning paradigm that enables us to probe the representational basis of trained and untrained speech production. Of particular focus are differences in representation that may reflect stimulus versus articulatory models. Differences in the representational patterns from acoustics and vocal tract imaging may allow us to parcellate neural representations in distinct regions of the speech production network. Specifically, searchlight analyses of fMRI data for trained and untrained speech, using test RDM models built from rtMRI images as subjects imitate speech under these same conditions, can allow us to address these questions empirically. Further, separately considering fMRI scanning runs from the same imaging session may also allow for changes in representation with learning to be probed. Moreover, allied to analyses of speech acoustics, behavioural performance, and articulatory tissue dynamics, we can begin to examine whether learning success is mediated by individual differences in the neural representation of articulation and speech acoustics. RSA provides us with a highly flexible and powerful tool with which to probe these questions, so that we can better understand plasticity of speech production via a multidimensional framework.

Clinical applications of this framework could include the investigation of outcomes of speech therapy. For instance, Using rtMRI, Hagedorn and colleagues (Hagedorn, Proctor, Goldstein, Tempini, & Narayanan, 2012) were able to demonstrate that the nature of intrusion errors during a tongue-twister task was different in an apraxic patient compared with control participants. More intriguingly, they also showed that the patient frequently performed covert “rehearsal” articulations in a syllable production task that, due to a lack of phonation, would have been omitted from a standard acoustic assessment. Recent clinical investigations have suggested suitability of rtMRI to the study of velopharyngeal closure, particularly following paediatric surgical interventions for cleft palate (Scott et al., 2012; Sagar & Nimkin, 2015). Image-based approaches to capturing the degree of velar closure (angle of velar eminence - i.e., mid-sagittal angle between velum and uvula) hold utility in quantitatively diagnosing extent of velar insufficiency (Drissi et al., 2011). In combination with simultaneous audio recording, it may be possible to link this velar insufficiency to acoustic properties of speech deficits that arise (e.g., hypernasalisation) (Sagar & Nimkin, 2015). Such clinical assessment methods could be extended using the tissue tracking methods described in earlier sections of the present review (see 3.3); variability of tissue traces proximal to the velum may offer insight into the noisiness of articulatory dynamics in velar insufficiency. Following traumatic brain injury or stroke, patients are often treated using articulatory training strategies, as a means of improving speech production abilities. Our framework has the potential to explore both how clinical deficits in speech production manifest, and how they are ameliorated by therapy, based on quantifying real-time vocal tract images as patients speak. Further, where functional brain organisation of speech has been affected in patients, our framework could be used to explore the representation of speech production. By probing fMRI data with test models derived from patients' own articulations, it may be possible to account for differences in speech representation following damage, and perhaps plasticity of representation after therapy. Similarly, charting representations of speech production in particular brain regions may provide a means by which the targets for neural interventions (e.g., electro-stimulation, or surgical) can be determined.

Future challenges for our present framework will involve extending the study of speech representation to connected discourse, such as full sentences, or narrative speech. A key challenge here is devising models that adequately capture the spatio-temporal complexity of connected speech from rtMRI images, such that these can be used in RSA searchlights. For instance, 'frame-by-frame' models that are built based on a series of images (capturing articulator position for successive consonant clusters and vowels) may prove fruitful. In addition, as discussed above, individual differences analyses that ally univariate fMRI data with measures derived from real-time MRI may also prove insightful.

5. Summary

We have presented an overview of the state of knowledge on the neural correlates of phonetic imitation and learning, as revealed using functional neuroimaging techniques. We suggest that the understanding of spoken language learning can be enhanced through combining functional MRI with real-time MRI of the vocal tract, and using Representational Similarity Analysis to unite acoustic, articulatory and neural data sources. Our proposed method has a variety of potential applications for future research on the representation of speech, and on plasticity in vocal motor behaviours, within both clinical and non-clinical contexts. We have made some suggestions for how our method may be implemented, and hope that this article will inspire other researchers to pursue these and alternative multimodal approaches to the study of speech learning.

References

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., & Savariaux, C. (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, *30*(3), 533–553.

Baer, T., Gore, J. C., Gracco, L. C., & Nye, P. W. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of Acoustical Society of America*, *90*, 799–828.

Berken, J. A., Gracco, V. L., Chen, J.-K., Watkins, K. E., Baum, S., Callahan, M., & Klein, D. (2015). Neural Activation in Speech Production and Reading Aloud in Native and Non-Native Languages. *NeuroImage*, *112*, 208–217.

Bresch, E., Kim, Y., Nayak, K., & Byrd, D. (2008). Seeing speech : capturing vocal tract shaping using real-time magnetic resonance imaging. *IEEE Signal Processing Magazine, May,* 123–129.

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

Bresch, E., Nielsen, J., Nayak, K., & Narayanan, S. (2006). Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *The Journal of the Acoustical Society of America*, *120*(L), 1791–1794.

Bressman, T., Ackloo, E., Heng, C.L., & Irish, J.C. (2007). Quantitative three-dimensional ultrasound imaging of partically resected tongues. *Otolaryngology - Head and Neck Surgery, 136,* 799–805.

Bressmann, T. (2010). 2D and 3D ultrasound imaging of the tongue in normal and disordered speech. In B. Maassen & P.van Lieshout (eds.) *Speech Motor Control: New Developments in Basic and Applied Research* (3rd ed.), pp. 351–362.

Buchsbaum, B. R., & D’Esposito, M. (2008). The search for the phonological store: From loop to convolution. *Journal of Cognitive Neuroscience*, *20*(5), 762–778.

Cartei, V., Cowles, H. W., & Reby, D. (2012). Spontaneous Voice Gender Imitation Abilities in Adult Speakers. *Plos One*, *7*(2).

Cheng, H.Y., Murdoch, B.E., Goozée, J.V., & Scott, D. (2007). Electropalatographic assessment of tongue-to-palate contact patterns and variability in children, adolescents and adults. *Journal of Speech, Language and Hearing Research, 50,* 375–392.

Dagenais, P.A. (1995). Electropalatography in the treatment of articulation/phonological disorders. *Journal of Communication disorders, 28,* 303–329.

Delvaux, V., Huet, K., Piccaluga, M., & Harmegnies, B. (2014). Phonetic compliance : a proof-of-concept study. *Frontiers in Psychology, 5,* doi.org/10.3389/fpsyg.2014.01375

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational Similarity Analysis Reveals Commonalities and Differences in the Semantic Processing of Words and Objects. *Journal of Neuroscience*, *33*(48), 18906–18916.

Drissi, C., Mitrofanoff, M., Talandier, C., Falip, C., LeCouls, V., & Adamsbaum, C. (2011). Feasibility of dynamic MRI for evaluating velopharyngeal insufficiency in children. *European Radiology, 21,* 1462–1469.

Dronkers, N. F. (1996). A new brain region for coordinating speech articulation. *Nature*, *384*(6605), 159–161.

Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception : evidence from searchlight similarity analysis. *Cerebral Cortex, 25* (12), 4772-4788.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America*, *106*(3), 1511–1522.

Flege, J. E., MacKay, I. R., & Meador, D. (1999). Native Italian speakers’ perception and production of English vowels. *Journal of the Acoustical Society of America*, *106*(5), 2973–2987.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5 Pt 1), 3125–3134.

Garnier, M., Lamalle, L., & Sato, M. (2013). Neural correlates of phonetic convergence and speech imitation. *Frontiers in Psychology*, *4*, 1–16.

Gibbon, F.E., Yuen, I., Lee, A.S., & Adams, L. (2007). Normal adult speakers' tongue palate contact patterns for alveolar oral and nasal stops. *Adavances in Speech-Language Pathology, 9*(1), 82–89.

Golestani, N., & Pallier, C. (2007). Anatomical correlates of foreign speech sound production. *Cerebral Cortex*, *17*(4), 929–934.

Goozée, J.V., Lapointe, L.L., & Murdoch, B.E. (2003). Effects of speaking rate on EMA-derived lingual kinematics: a preliminary investigation. *Clinical linguistics and phonetics, 17*(4-5), 375–381.

Goozée, J.V., Stephenson, D.K., Murdoch, B.E., Darnell, R.E., & Lapointe, L. (2005). Lingual kinematic strategies used to increase speech rate: comparison between younger and older adults. *Clinical linguistics and phonetics, 19*(4), 319–334.

Guenther, F. H., & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, *25*(5), 408–422.

Guenther, F.H., Espy-Wilson, C.Y., Boyce, S.E., Matthies, M.L., Zandipour, M., & Perkell, J.S. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *Journal of the Acoustical Society of America, 105*(5), 2854–2865.

Hagedorn, C., Proctor, M., Goldstein, L., & Narayanan, S. (2011). Automatic analysis of constriction location in singleton and geminate consonant articulation using real-time magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *130*(4), 2548.

Hagedorn, C., Proctor, M., Goldstein, L., Tempini, M. L. G., & Narayanan, S. S. (2012). Characterizing covert articulation in apraxic speech using real-time MRI. In *13th Annual Conference of the International Speech Communication Association 2012* (pp. 1050–1053).

Hashizume, H., Taki, Y., Sassa, Y., Thyreau, B., Asano, M., Asano, K., … Sugiura, M. (2014). Developmental Changes in Brain Activation Involved in the Production of Novel Speech Sounds in Children. *Human Brain Mapping, 4089*, 4079–4089.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13*(2), 135–145.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron*, *69*(3), 407–422.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., & Reiterer, S. M. (2013). Language aptitude for pronunciation in advanced second language (L2) Learners: Behavioural predictors and neural substrates. *Brain and Language*, *127*(3), 366–376.

Hughes, S. M., Mogilski, J. K., & Harrison, M. A. (2014). The Perception and Parameters of Intentional Voice Manipulation. *Journal of Nonverbal Behavior*, *38*(1), 107–127.

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: A functional magnetic resonance imaging study. *Journal of Neuroscience*, *23*(29), 9541–9546.

Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Science*, *10*(11), 480–486.

Kappes, J., Baumgaertner, A., Peschke, C., & Ziegler, W. (2009). Unintended imitation in nonword repetition. *Brain and Language*, *111*(3), 140–151.

Katz, W.F., Bharadwaq, S.V., & Carstens, B. (1999). Electromagnetic articulography treatment for an adult with Broca's aphasia and apraxia of speech. Journal of Speech, Language, and Hearing Research, 42, 1355–1366.

Kim, J., Kumar, N., Lee, S., & Narayanan, S. (2014). Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. *Proceedings of the 10th International Seminar on Speech Production (ISSP)*, (i), 222–225.

Kim, Y.-C., Narayanan, S. S., & Nayak, K. S. (2009). Accelerated 3D MRI of vocal tract shaping using compressed sensing and parallel imaging. *IEEE International Conference on Acoustics, Speech and Signal Processing.* doi: 10.1109/ICASSP.2009.4959498

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 1–28. doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the USA*, *103*(10), 3863–3868.

Krishnan, S., Leech, R., Mercure, E., Lloyd-Fox, S., & Dick, F. (2015). Convergent and divergent fMRI responses in children and adults to increasing language production demands. *Cerebral Cortex, 25*(10), 3261-3277.

Lee, S., Bresch, E., Adams, J., Kazemzadeh, A., & Narayanan, S. S. (2006). A study of emotional speech articulation using a fast magnetic resonance imaging technique. Proceedings of the International Society for Speech (*INTERSPEECH* 2006), 1792–1795.

McLeod, S. (2006). Australian adults' production of /n/: an EPG investigation. *Clinical linguistics and Phonetics, 20*(2-3), 99–107.

McGettigan, C. (2015). The social life of voices: studying the neural bases for the expression and perception of the self and others during spoken communication. *Frontiers in Human Neuroscience*, *9*, 1–4. doi.org/10.3389/fnhum.2015.00129

McGettigan, C., Eisner, F., Agnew, Z. K., Manly, T., Wisbey, D., & Scott, S. K. (2013). T'ain't what you say, it's the way that you say it – left insula and inferior frontal cortex work in interaction with superior temporal regions to control the performance of vocal impersonations. *Journal of Cognitive Neuroscience, 25*(11), 1875–1886.

Moser, D., Fridriksson, J., Bonilha, L., Healy, E. W., Baylis, G., Baker, J. M., & Rorden, C. (2009). Neural recruitment for the production of native and novel speech sounds. *Neuroimage*, *46*(2), 549–557.

Narayanan, S., Nayak, K., Lee, S. B., & Byrd, D. (2004). An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America*, *115*(4), 1771–1776.

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. *Magnetic Resonance in Medicine*, *69*(2), 477–485.

Oh, A., Duerden, E. G., & Pang, E. W. (2014). Brain & Language The role of the insula in speech and language processing. *Brain and Language*, *135*, 96–103.

Pardo, J. S. (2006a). Expressing Oneself in Conversational Interaction. In E. Morsella (ed.), *Expressing One's Self: Communication, Cognition and Identity* (pp. 183-196). Hove: Psychology

Press/Taylor Francis.

Pardo, J. S. (2006b). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190–197.

Pardo, J. S., & Jay, I. C. (2010). Conversational role influences speech imitation. *Attention Perception & Psychophysics*, *72*(8), 2254–2264.

Peschke, C., Ziegler, W., Eisenberger, J., & Baumgaertner, A. (2012). Phonological manipulation between speech perception and production activates a parieto-frontal circuit. *Neuroimage*, *59*(1), 788–799.

Peschke, C., Ziegler, W., Kappes, J., & Baumgaertner, A. (2009). Auditory-motor integration during fast repetition: The neuronal correlates of shadowing. *Neuroimage*, *47*(1), 392–402.

Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice modulation: A window in the origins of human vocal control? *Trends in cognitive sciences*, *20*(4), 304-318.

Piske, T., MacKay, I. R. a., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics*, *29*(2), 191–215.

Proctor, M., Bresch, E., Byrd, D., Nayak, K., & Narayanan, S. (2013). Paralinguistic mechanisms of production in human “beatboxing”: A real-time magnetic resonance imaging study. *Journal of the Acoustical Society of America*, *133*(2), 1043–1054.

Proctor, M. I., Bone, D., Katsamanis, N., & Narayanan, S. (2010). Rapid Semi-automatic Segmentation of Real-time Magnetic Resonance Images for Parametric Vocal Tract Analysis. In *11th Annual Conference of the International Speech Communication Association 2010* (pp. 1576–1579).

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724.

Reiterer, S., Hu, X., Erb, M., Rota, G., Nardo, D., Grodd, W., … Ackermann, H. (2011). Individual differences in audio-vocal speech imitation aptitude in late bilinguals: functional neuro-imaging and brain morphology. *Frontiers in Psychology*, *2*, doi.org/10.3389/fpsyg.2011.00271.

Reiterer, S. M., Hu, X., Erb, M., Rota, G., Nardo, D., Grodd, W., … Ackermann, H. (2011). Individual differences in audio-vocal speech imitation aptitude in late bilinguals: functional neuro-imaging and brain morphology. *Frontiers in Psychology*, *2*, doi: 10.3389/fpsyg.2011.00271.

Reiterer, S. M., Hu, X., Sumathi, T. A., & Singh, N. C. (2013). Are you a good mimic? Neuro-acoustic signatures for speech imitation ability. *Frontiers in Psychology*, *4*, doi.org/10.3389/fpsyg.2013.00782

Sagar, P., & Nimkin, K. (2015). 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. *Paediatric Radiology, 45,* 217–227.

Schoenle, P.W., Graebe, K., Wenig, P., Hoehne, J., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: use of laternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language, 31,* 26–35.

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. *The British Journal of Radiology, 85,* e1083–e1092.

Scott, A. D., Boubertakh, R., Birch, M. J., & Miquel, M. E. (2013). Adaptive Averaging Applied to Dynamic Imaging of the Soft Palate, *874*, 865–874. http://doi.org/10.1002/mrm.24503

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends Neurosci*, *26*(2), 100–107.

Segawa, J. A., Tourville, J. A., Beal, D. S., & Guenther, F. H. (2013). The neural correlates of speech motor sequence learning. *Journal of Cognitive Neuroscience*, 1–10.

Silva, S., & Teixeira, A. (2015a). Quantitative systematic analysis of vocal tract data. *Computer Speech & Language*, *36,* 307–329.

Silva, S., & Teixeira, A. (2015b). Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Computer Speech & Language*, *33*(1), 25–46.

Simmonds, A. J. (2015). A hypothesis on improving foreign accents by optimizing variability in vocal learning brain circuits. *Frontiers in Human Neuroscience*, *9*, doi.org/10.3389/fnhum.2015.00606

Simmonds, A. J., Leech, R., Iverson, P., & Wise, R. J. S. (2014). The response of the anterior striatum during adult human vocal learning. *Journal of Neurophysiology*, *112*(4), 792–801.

Simmonds, A. J., Wise, R. J. S., Dhanjal, N. S., & Leech, R. (2011). A comparison of sensory-motor activity during speech in first and second languages. *Journal of Neurophysiology*, *106*(1), 470–478.

Simmonds, A. J., Wise, R. J. S., & Leech, R. (2011). Two tongues, one brain: imaging bilingual speech production. *Frontiers in Psychology*, *2*, doi.org/10.3389/fpsyg.2011.00166

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, *26*(7), 952–981.

Weirich, M., & Fuchs, A. (2013). Palatal morphology can influence speaker specific realizations of phonemic contrasts. *Journal of Speech, Language, and Hearing Research, 56,* S1894–S1908.

Weiss-Croft, L.J., & Baldeweg, T. (2015). Maturation of language networks in children: a systematic review of 22 years of functional MRI. *NeuroImage, 123,* 269–281.

Vasquez Miloro, K., Pearson, W. G., & Langmore, S. (2014). Effortful Pitch Glide: A Potential New Exercie Evaluated by Dynamic MRI. *Journal of Speech, Language and Hearing Research*, *57*, 1243–1250.

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. *Journal of Acoustical Society of America*, *117*(1), 338–350.

Figures

Figure 1: Overview of the real-time vocal tract imaging framework and its application to functional neuroimaging.

(a) (i) Quantitative vocal tract distance traces for vowels /i/, /a/ and /u/. Panels display mean traces (±SD) averaged over multiple productions of each vowel by a single subject. Distance measurements are plotted according to intervals specified by a regular gridline system (here, at inter-gridline intervals of 2mm), progressing from the lips (left) to the larynx (right). Distance traces are extracted on a trial-wise basis as the point-to-point 2D Euclidean distance between the upper and lower tissue boundary traces (see inset: green and red traces, respectively). **(ii) Quantitative lip co-ordinate tracking.** Lip and laryngeal position are extracted based on tissue boundary tracking algorithms that determine pixel intensity change within a given lip/larynx search distance (see inset). Lip co-ordinates (averaged over multiple trials) are plotted against image frames per vowel (co-ordinates reflecting lip retraction fall at the top of the plot, and protrusion at the bottom). Larynx co-ordinates are not plotted for the present data.

Figure 2: Representation Similarity Analysis (RSA) of real-time vocal tract images and speech acoustics, as applied to searchlights of fMRI data. *Upper left:* Trial-wise images of the articulators during steady state phonation after adaptive averaging; images are masked to restrict field-of-view to the vocal tract. Masked images are then processed with the RSA toolbox (see text), yielding the cross-condition Representational Dissimilarity Matrix (RDM) to the right. Upper right: Vowel stimulus acoustics are processed based on the power spectral density matrix for each stimulus (see text). Using a similar procedure to the vocal tract images, the cross-condition acoustic RDM is constructed. *Bottom:* Activation maps for functional MRI data are specified for each condition of interest (vs. rest). Across these vowel conditions, searchlight RDMs for fMRI data are compared to RDMs built from vocal tract images of vowel articulation or from measurement of vowel stimulus acoustic spectra. For example, the vocal tract RDM can be used to searchlight somatomotor regions during production of these same vowels – see (b) & (c); the stimulus acoustic RDMs can searchlight superior temporal regions during vowel perception, as in (d).