

Biomedical Physics & Engineering Express



TOPICAL REVIEW

Research in methodologies for modelling the oral cavity

OPEN ACCESS

RECEIVED

15 August 2023

REVISED

26 January 2024

ACCEPTED FOR PUBLICATION

13 February 2024

PUBLISHED

18 March 2024

Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



Muhammad Suhaib Shahid^{1,*} , Andrew P French^{1,2} , Michel F Valstar¹ and Gleb E Yakubov^{2,*}

¹ School of Computer Science, University of Nottingham, NG8 1BB, United Kingdom

² School of Biosciences, University of Nottingham, LE12 5RD, United Kingdom

* Authors to whom any correspondence should be addressed.

E-mail: Muhammad.Shahid@Nottingham.ac.uk and Gleb.Yakubov@Nottingham.ac.uk

Keywords: AI in the physical world, clinical phonetics, vocal tract research

Abstract

The paper aims to explore the current state of understanding surrounding *in silico* oral modelling. This involves exploring methodologies, technologies and approaches pertaining to the modelling of the whole oral cavity; both internally and externally visible structures that may be relevant or appropriate to oral actions. Such a model could be referred to as a ‘complete model’ which includes consideration of a full set of facial features (i.e. not only mouth) as well as synergistic stimuli such as audio and facial thermal data. 3D modelling technologies capable of accurately and efficiently capturing a complete representation of the mouth for an individual have broad applications in the study of oral actions, due to their cost-effectiveness and time efficiency. This review delves into the field of clinical phonetics to classify oral actions pertaining to both speech and non-speech movements, identifying how the various vocal organs play a role in the articulatory and masticatory process. Vitaly, it provides a summation of 12 articulatory recording methods, forming a tool to be used by researchers in identifying which method of recording is appropriate for their work. After addressing the cost and resource-intensive limitations of existing methods, a new system of modelling is proposed that leverages external to internal correlation modelling techniques to create a more efficient models of the oral cavity. The vision is that the outcomes will be applicable to a broad spectrum of oral functions related to physiology, health and wellbeing, including speech, oral processing of foods as well as dental health. The applications may span from speech correction, designing foods for the aging population, whilst in the dental field we would be able to gain information about patient’s oral actions that would become part of creating a personalised dental treatment plan.

1. Introduction

Technologies capable of creating 3-Dimensional facial models (based on some form of inputted data, videos, pictures etc.) are often limited to only the external view, seldom providing manipulable cross-sections with observable internal structures. Existing technologies that can help in producing real time models of the mouth, such as Electromagnetic Articulography and Electropalatography, are limited in their use (Kochetov 2020a; Rebernik *et al* 2021); a consequence of their resource and cost intensive running cost and inability to encapsulate the movements of all articulators.

Although the problem statement centres around oral actions, the review is limited not only to the mouth. It explores elements of medical and computer science

fields that have demonstrated the use of ideas, approaches, and methodologies capable of addressing the problem statement. A survey of literature surrounding mouth specific movements and structures is a vital addition to forming an understanding of what a complete 3D oral model would consist of. Logically, the next step would be knowing how these structures relate to the actions the mouth performs i.e. speaking, chewing/swallowing, breathing etc. The paper also takes an in-depth look at existing methods used to capture oral movements during action. These are a variety of 2D and 3D approaches that, to varying degrees, are currently used in helping to visualise oral movements.

Mapping the movements of the mouth is made difficult by the mouth’s complex and deformable 3-dimensional structure, and featuring as it does

external and internal elements. The complex movements of the mouth are the result of the interaction between multiple elements, including both soft tissues (such as the tongue and velum) and hard structures (such as the jaw and teeth). One such approach to addressing this problem is by forming predictive models that can accept observations of (easily accessible) external movements and create a predicted internal structure. Such techniques have been spearheaded within the medical field of radiology and will be explored in regards to their potential for application in oral modelling.

To fully comprehend the intricacies of oral actions, it is essential to closely examine the movements that occur within the vocal tract and their relationship to oral physiology.

2. Articulatory phonetics

Before we can explore the Computer Science and Clinical Phonetics fields that address the problem of oral modelling and speech analysis, it is important to understand the articulatory phonetics that govern the production of speech. This section defines relevant terminologies pertaining to this field, and identifies the ‘external’ and ‘internal’ articulators.

The first pertinent definition is that of articulatory phonetics itself; Articulatory Phonetics, a subfield of phonetics, can be defined as a field of phonetics that focuses on the study of how speech sounds are produced in different languages by examining the movements and positions of the vocal organs, also known as articulators (Keating 2001).

The definition touches upon another closely linked process worth defining: Articulation. Articulation is the means by which speech is formed, through the movements of vocal organs called articulators. In phonetics, articulation has been defined as the movement and/or positioning of the vocal organs (such as the tongue, lips, and jaw) during speech production. These movements and positions influence the shape and configuration of the vocal tract, which in turn affects the quality and characteristics of the resulting speech sounds (Ladefoged and Johnson 2015).

Speech recruits the use of various vocal organs, known as articulators. During this articulatory process some of these vocal organs may be externally visible. These include the upper and lower lips, as well as the teeth and at certain times the tongue when it is protruding through the lips and teeth; these are the ‘external articulators’. However, most of the oral cavity inside the mouth cannot be seen externally and thus form what are referred to as ‘internal articulators’.

Throughout this paper, we will also refer to the vocal tract. The following is a definition of the term:

‘vocal tract is the term used to refer to the entire speech apparatus, with the larynx as the central

element which subdivides the apparatus into lower and upper regions’ (Ball 2021).

2.1. The articulators

Before we dive into the complexities of speech production and oral movements, it’s important to familiarise ourselves with the various articulators involved in the process. These include the lips, jaw, tongue, teeth, hard palate, velum, and larynx. The external articulators primarily consist of the jaws and lips, while the internal ones include the hard palate, velum, and larynx. It’s worth noting that the role of the teeth and lips in articulation can vary depending on the specific sentence being spoken.

3. Speech production

Speech production (SP) is the process by which words are spoken. This may seem to be the same as the previously described ‘Articulation’, but there is a difference. Speech production involves the physical creation of speech sounds, as well as hearing, perception, and information processing in the nervous system and brain. The process is complex and involves a feedback loop to ensure the speech produced is meaningful (Docio-Fernandez and García Mateo 2015).

In other words, SP is the complete process by which initial thoughts are translated to speech; articulation is just one part of SP. There are three main stages to the process: initiation, phonation, and articulation (there is also an additional fourth, coordination). (Ball 2021) provides a detailed overview of the SP process, exploring in detail these three stages. Below is provided a short summation of each stage, paying closer attention to systems that engage the vocal tract (rather than other systems i.e. lungs).

3.1. Initiation

As the name suggests, this initial state illustrates the beginning of the speech process. The previously mentioned definition of articulation was defined as a process that modifies an air stream to produce the sounds of speech; the initiation of speech is the method by which humans generate the air pushed upwards through the vocal tract. This method of air generation is known as the airstream mechanism and can initiate from three points of the body: the lungs, the velum, and the glottis.

Airstream mechanisms beginning from the lungs are controlled by the pulmonic system and are thus called Pulmonic airstreams. Contractions of the ribcage, controlled by the diaphragm, work to fill the lungs with air to then be released through to the vocal tract. Velaric airstreams are the redirected flow of air produced by the lungs into the oral or nasal cavity, this is a task completed by the velum. The raising and lowering of the velum dictates normal breathing or production of nasal sounds. Glottalic airstream mechanisms

control the movement of air by action of the glottis. The opening and closing of the glottis form an upward or downward movement of air, to then reach the second point of articulation (further down the vocal tract). Most sounds produced on a glottic air stream are ejectives, such that sounds are formed by air being pushed out through the mouth and nose (also referred to as an egressive airstream).

3.2. Phonation

Phonation is the secondary stage of vocal sound and speech production, a process by which the previously mentioned pulmonic egressive airstreams undergo pressure changes induced by the motion of two vocal folds situated in the larynx. Movements of the cartilage structure surrounding the larynx open and close a triangular-like space between the vocal folds that allow the passage (or restriction) of air; this space is called the glottis, an opening crucial in forming vowels and other consonants.

Consonant sounds produced within common speech are a result of two main vocal fold configurations. The first are ‘voiced’ consonants. These are created when the vocal folds are held together and vibrating, thus creating a narrower glottal aperture; an example of such a word would be ‘broom’. The second, ‘voiceless’ consonants, are a result of a larger glottal opening with an example of such a word being ‘hat’.

3.3. Articulation

This is our final state of interest. Expanding upon the previous definition provided, articulation refers to the shaping of the resultant airstream, generated and altered during the initiation and phonation stages; at this point the articulators are configured to form the desired labialisations.

Table 1 classifies the passive (rigid) or active (mobile) motion of articulators, and the IPA symbol of which voiced or voiceless consonants they create.

Additionally, are provided examples of consonant types, and sample voiced/voiceless fricatives.

It is worth noting that in some speech, two simultaneous primary places of articulation can occur. This is called double articulation. For example, labial-velar consonants are doubly articulated and engage the use of both the velum and lips.

Now that the articulatory process has been discussed, speech itself can be defined. Simply put, this can be explained as the use of vocal organs to generate speech. However, a more formal definition can put it as:

‘...movements or movement plans that produce as their end result acoustic patterns that accord with the phonetic structure of a language.’ (Kent 2015).

Speech taxonomies, that is defining the various speech behaviours, are generally a well-researched sphere. To provide an idea of what these constitute, some have been listed below (Kent 2015):

- Emotional speech: Speech that expresses an emotion such as anger, sadness, happiness, or fear; sometimes contrasted with neutral speech
- Empty speech: Speech that is semantically void (e.g., comprising automatisms, vague circumlocutions, or single words)
- Exaggerated (overarticulated) speech: Speech produced with unusually large ranges of articulatory movement and/or force; similar to hyperspeech but with more deliberate and extensive movements
- Nonsensical speech (nonsense): Speech that does not convey meaning, usually because it involves phonetic sequences that do not conform to the words in a given language

Table 1. A table providing information regarding the articulatory structure, including sample consonants and voiced/voiceless fricatives.

Articulator	Passive (rigid) or Active (mobile)	Example of consonant type	Voiced Fricative	Voiceless Fricative
Lips	Active	<i>Bilabial</i> —sounds made when both lips are engaged	The consonant [p]—in P alm	The consonant [m]—in p al M
Teeth	Passive	<i>Labiodental</i> —sounds made when lower lip contacts upper teeth.	The consonant [f]—in F ar	The consonant [v]—in N ever
Hard Palate	Passive	<i>Palatal</i> —sounds made when body of tongue contacts hard palate.	The consonant [j]—in U niversity or Y oung	N/A
Velum	Active	<i>Velar</i> —sounds made when the back of the tongue touches the velum.	The consonant [d]—in D og	The consonant [k]—in K ing
Glottis /Larynx	Active	<i>Glottal</i> —sounds made using the glottis as primary articulation.	The consonant [h]—in H ind	The consonant [h]—in H igh
Tongue	Mobile	<i>Retroflex</i> —sound made when the tongue has a flat, concave or curled shape; articulated between alveolar ridge and hard palate.	The consonant [ɭ]—in R est	The consonant [ʂ]—in Swedish word f o R S (meaning ‘rapids’)

4. Nonspeech oral movements

Both verbal and nonverbal actions, are governed by the craniofacial and masticatory musculatures of the face; more specifically these include movements pertaining to speech, facial expressions, biting, chewing, ventilation, and swallowing (Kent 2015). This section will now review the nonspeech elements of oral action.

4.1. Nonspeech oral movements

Kent (2015) reviews a vast array of literature to collate definitions and propose taxonomies for both speech, and non-speech oral movements (NSOMs). Although definitions and taxonomies for the oral process can vary, the paper provides clear descriptions of the movements themselves; thus, this evaluation of NSOMs refers back to Kent (2015) often. The narrative review defines NSOMs as:

‘Motor acts performed by various parts of the speech musculature to accomplish specified movement or postural goals that are not sufficient in themselves to have phonetic identity’

In essence, NSOM’s cover a vast range of orofacial movements that are performed alone or with other movements for varying purposes; governing these movements are the articulators and facial muscles. Alongside speech, facial muscles serve two main non-speech functions, chewing and facial expressions (Westbrook *et al* 2022). The following sections will explore these movements.

4.1.1. Mastication

The chewing process, also referred to as mastication, is a motor activity intended for processing food in preparation for swallowing. The complex process involves the action of the suprahyoidal muscles, craniofacial musculature, vocal organs, and even saliva (van der Bilt *et al* 2006). The process is complex in the sense that the movements for mastication are formed by multiple interacting parts. Although the chewing process can be explored in much detail, this review is primarily interested in how processes engage the articulators and will thus primarily focus on such literature and taxonomy pertaining to the vocal organs.

As mentioned, mastication aims to break down and crush food to be mixed with saliva and moved to the back of the throat for deglutition (swallowing). The ‘muscles of mastication’ consist of the muscle groups: temporalis, masseter, medial pterygoid, and lateral pterygoid.

However, it is key to note that the chewing process involves more than just the ‘muscles of mastication’. Neurological control of the jaw and other muscles, individual anatomy and even the types of food being processed govern the cycle of mastication adopted;

with certain foods having a longer/shorter cycle (Soboļeva *et al* 2005).

4.1.2. Facial expressions and other NSOMs

Alongside mastication, the process of facial expression generation is one of the main NSOMs surrounding ‘all things oral action’ that we are trying to unfold. Certain facial expressions generated adopt the use of identified articulators or oral motor systems: this includes facial expression such as smiling and surprise, as well as lip pursing, jaw opening, and tongue protrusion (Kent 2015). However, others draw on the use of non-identified systems: these include actions such as coughing, laughing, and blowing.

At times facial expression may be a consequence of another movement. Coughing, for example engages muscle systems including the respiratory system (among others); during which process the distinct ‘coughing facial expression’ is produced.

Kent (2015) provides a table of proposed speech-like and non-speech movements, categorised into the muscle systems they employ and their general function. Below are identified some of these movements, the full classification of which can viewed within their paper:

Oral only: Licking, Sucking, Smiling,

Respiratory: Subglottal air pressure control, Prolonged expiration

Respiratory and laryngeal: Grunting, Moaning, Crying

Oral and respiratory: Panting, Blowing, Sighing, Whistling

Oral, Laryngeal and Respiratory: Coughing, Laughing

Additionally, certain NSOMs produce an audible output as a result of the action i.e. coughing, panting, moaning, laughing.

5. Capturing of articulatory actions

Having now identified the various articulators that form the speech process, it is just as important to realise how these structures and motions can be recorded. Recording and quantifying the movements of the articulators is a difficult task. Depending on the needs of the experiment/research, any specific methodology can be desirable. There are currently in use various technologies capable of capturing the movements of the vocal tract, each one addressing the five vocal organs to a varying degree.

Table 2 provides a modified extract from Kochev (2020a). It lists the 12 methods considered here and then indicates the individual capabilities of each of these systems. The final column of the table, titled ‘MRI scan highlighting the articulators recorded’ shows the relevant articulators highlighted in different colours. Note, that the method may not necessarily

Table 2. A chart illustrating the capabilities of each articulatory recording system, with MRI images highlighting the articulators each method captures. Images sourced from (Lim *et al* 2021).

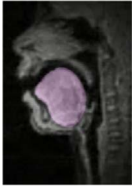

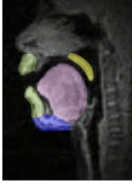
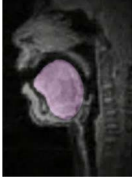
Methods	Oral Gestures			Nasal Gestures Velum	Laryngeal Gestures Larynx	MRI scan highlighting the articulators recorded.
	Lips	Tongue	Jaw			
Electro-palato-graphy (EPG)		X				
Ultrasound		X			X	
Electro-magnetic Articulography (EMA)	X	X	X	X		
Static Plataography		X				

Table 2. (Continued.)

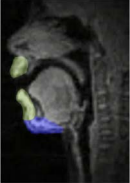







Methods	Oral Gestures			Nasal Gestures Velum	Laryngeal Gestures Larynx	MRI scan highlighting the articulators recorded.
	Lips	Tongue	Jaw			
Video and Optical Tracking	X		X			
X-ray microbeam	X	X	X			
Electro-glottography (EGG)				X	X	
Endoscopy and Photo-glottography		X		X	X	

Table 2. (Continued.)

Methods	Oral Gestures			Nasal Gestures Velum	Laryngeal Gestures Larynx	MRI scan high-lighting the articulators recorded.
	Lips	Tongue	Jaw			
Oral Airflow/pressure	X	X		X	X	
Nasal Airflow					X	
Real Time Magnetic Resonance Imaging (rtMRI)	X	X	X	X	X	
X-rays	X	X	X	X	X	

engage the use of MRI, rather the image aims to only display the relevant organs to the reader.

The methodologies chosen are often subject to financial constraints, as well as access to machinery and trained staff. Furthermore, certain research methodologies require the collection of both auditory and articulatory data. In the case of Electromagnetic Articulography (EMA), this becomes problematic since the acoustics are changed when sensors are attached to the tongue and around the mouth (Meenakshi *et al* 2014).

Kochetov (2020a) reviewed 379 full research articles published between 2000–2019, to find out which methods of articulatory recording have been used most during this 20-year period. The survey included papers published in the field of Language and Speech, Phonetics, Phonology and more. The results showed that around 60% of these experiments used Electropalatography (EPG), Ultrasound, and EMA techniques for articulatory recording. The other 40% was occupied by the remaining techniques, with MRI notably only taking up 6% of used methods, despite being a technique that can quantify the movements of all the articulators, the whole vocal tract.

Electromagnetic Articulography is the most used, however it is also important to consider that often techniques are used in conjunction with one another.

This paper provides a summation of the various techniques mentioned. For each of the recording methods, below can be found a brief overview of the approach, along with examples of use, overall safety of the technique, sampling rate, audio compatibility, cost, availability of data sets and also a recent review that has been completed; a review that took place after Kochetov's review (2021 or later). The section aims to

offer the reader an oversight to the current use of various techniques and also provide a base from which a researcher can select a method suited to their needs.

In some instances, including cases where the technique has not been in use for quite some time, a recent review, cost, or example of a dataset have not been found. It is worth noting that factors such as cost and safety overview are presented for a casual comparison, but a more in-depth, up-to-date investigation would be required by the researcher before adopting a technique.

5.1. Electropalatography (EPG)

Electropalatography is a technique introduced in 1970, used to identify the tongue and hard palate location during articulation; the technique's ability to record dynamic speech features further allows for the detection of sound production (Mat Zin *et al* 2021). During the process, a custom-built artificial plate is placed within the speaker's mouth, and subsequently clipped on to the individual's upper palate. The palate is lined with a grid of electrodes, capable of registering the contact taking place between the tongue and the roof of the mouth (Verhoeven *et al* 2019). Detailed in table 3, the technique allows for quantifying where and how the tongue touches the roof of the mouth during speech. The EPG is can also be used to analyse contact patterns during real-time speech generation. (Hardcastle *et al* 1989). With every consonant uttered, a unique contact pattern is produced on the hard palate. This can be used to identify the sound produced during speech, with the location of the tongue and hard palate being detected by electrode sensors present on the artificial palate.

Table 3. A table covering nine different factors surrounding the use of electropalatography in capturing oral movements.

Advantages	This technique is suitable for children and individuals with disabilities who find it difficult to remain still.
Disadvantages	The retainer-like contraption placed against the hard palate means the technology is unsuitable for individuals who already use dental prosthetics. Data is limited to the oral gestures of the tongue. Provides no information about the location of the tongue when not in contact with the hard palate. Method is also invasive, as plate has to be placed in the mouth.
Examples of use	Wood (2010) - used EPG on individuals with Down Syndrome and found that they can continue to improve their speech production and intelligibility as they progress from adolescence to adulthood.
Overall safety	Material used in developing EPG is nontoxic (Mat Zin <i>et al</i> 2021). The artificial palates are made from acrylic resin, silver electrodes and copper wire; material that is FDA approved and widely used in dental applications (such as dentures and retainers and EMG).
Sampling Rate	The linguopalatal contact is tracked dynamically, typically taking samples every 10 milliseconds (Kochetov 2020a).
Audio Compatibility	Yes.
Cost per session	Custom-made EPG plate costing around £495 (Kochetov 2020a).
Recent Review	Mat Zin <i>et al</i> (2021), 'The technology of tongue and hard palate contact detection: a review'.
Examples of Available datasets	EPG data from two female speakers of Central Arrernte. Both subjects recorded uttering the same words using two different sorts of palates (Tabain 2011).

Table 4. A table covering nine different factors surrounding the use of ultrasound in capturing oral movements.

Advantages	With availability of smaller portable ultrasound systems, this technique is affordable and accessible to researchers. In the past, a small sample rate has meant that short articulations were not captured, or in poor quality. This problem has since been reduced with the introduction of higher frame rate devices.
Disadvantages	The observation of the motion at the tip of the tongue is missed when the tongue is raised or extended forward (Cleland <i>et al</i> 2011). The results of the technique can experience double edges, reflections, and general poor quality images generated (Stone 2005).
Examples of use	Bennett <i>et al</i> (2017) presents an ultrasound analysis of the secondary palatalisation constant in Irish, analysing data from 5 different Irish speakers.
Overall safety	'Ultrasound, however, is becoming cheaper, is safe, is easy to set up and use, and is able to provide real-time images of the whole tongue during speech.' (Wilson 2014).
Sampling Rate	Ultrasounds with frequencies up to 10 MHz are usually used in medical practice (Reda <i>et al</i> 2021).
Audio Compatibility	Yes
Cost per session	£500
Recent Review	Al-hammuri <i>et al</i> (2022), titled, 'Tongue Contour Tracking and Segmentation in Lingual Ultrasound for Speech Recognition: A Review'
Examples of Available datasets	UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions (Eshky <i>et al</i> 2018). Comparing articulatory images: An MRI / Ultrasound Tongue Image database (Cleland <i>et al</i> 2011).

5.2. Ultrasound

Ultrasound is an imaging technique introduced in the 1960s (Kelsey *et al* 1969). It has since become the second most popular method of articulator recording over the past 20 years, as shown in table 4 (Kochetov 2020a). The technique uses a transducer probe, capable of emitting a high frequency sound wave. When held against the neck, the thin beam projected from the probe travels through the tongue tissue and is reflected back to the transducer, forming a 2D image of the tongue. Ultrasounds inability to image bone or air means it does not allow for the visualisation of the palate, jaw, or rear pharyngeal wall; making it suitable only for imaging the tongue in speech research applications (Bliss *et al* 2018,

Kochetov 2020a). Although most ultrasound machines are stationary devices situated in hospitals, mobile USB probes are now being used more often, that make the recording process more convenient and accessible.

5.3. Electromagnetic articulography (EMA)

Electromagnetic Articulography is a point tracking technique (Mennen *et al* 2010), during which a series of sensors placed on target articulators record real-time movements in 3D (table 5). Later developments in the approach have led to its capability of taking five dimensional recordings, collecting three cartesian coordinates and two angular coordinates (Hoole and Zierdt 2010), therefore capturing information in

Table 5. A table covering nine different factors surrounding the use of electromagnetic articulography in capturing oral movements.

Advantages	Data collected within the oral cavity has high spatial accuracy and temporal resolution, thus producing fairly accurate information on articulatory gestures. EMA allows for the measuring of multiple articulators at once (Rebernik <i>et al</i> 2021).
Disadvantages	Sensor positioning is limited to the anterior oral tract, with velum tracking not possible without causing significant discomfort to subjects (Rebernik <i>et al</i> 2021). Sensors cannot be placed too close to each other without disturbing measurement accuracy. Method does not allow for high-quality simultaneous recording of auditory data since sensors attached to the tongue change acoustics (Meenakshi <i>et al</i> 2014); however, it does afford some speech production with 'moderate interference (Hasegawa-Johnson 1998)' (Dromey <i>et al</i> 2018)
Examples of use	Hoke <i>et al</i> (2019) used EMA to investigate the effects denture adhesives have in minimising denture displacement while chewing. They successfully used EMA to demonstrate that the use of denture adhesives statistically reduces the likelihood of denture micro movements.
Overall safety	Articulographs are generally considered safe to use (Hasegawa-Johnson 1998).
Sampling Rate	The NDI Wave and NDI Vox articulographs have a maximum sampling rate of 400 samples/s and can track 16 channels simultaneously (upto 16 sensors can be used). The AG500 can record 200 samples/s in 12 channels, while the AG501 can record 1250 samples/s of up to 24 channels (Ji <i>et al</i> 2014, Sigona <i>et al</i> 2018, Rebernik <i>et al</i> 2021)
Audio Compatibility	Yes.
Cost per session	Unkown
Recent Review	Rebernik <i>et al</i> (2021), titled 'A review of data collection practices using electromagnetic articulography'.
Examples of Available datasets	The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data (Ji <i>et al</i> 2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC) (Narayanan <i>et al</i> 2014).

greater detail. EMA is one of the few methods that illustrate oral gestures continuously, as opposed to technologies such as EPG that only illustrate the motion of the tongue when in contact with the palate. This makes it possible to record multiple articulators simultaneously and thus observe inter-articulatory behaviour (Rebernik *et al* 2021). The sensor pads placed in the mouth are small, taking only around 10-minutes for an adult to become adapted (Dromey *et al* 2018).

5.4. Static palatography

Static palatography is technique developed in the 19th century, used to study constant articulation (Kochetov 2020a). During this process, the tongue is painted black using an edible paint-like material to record the contact it makes makes the roof of the mouth during articulation (Anderson 2008). A mirror is inserted into the subjects mouth and a photo or video is taken, to show the location of paint traces on the hard palate (post articulation). As indicated in table 6, the simplicity of the technique makes it perhaps one of the most accessible methods of articulation visualisation.

Table 6. A table covering nine different factors surrounding the use of static palatography in capturing oral movements.

Advantages	The technique is cheap and portable and allows users to gather information about the tongue and the hard palate (Kochetov 2020a).
Disadvantages	The technique is very primitive, it cannot record constant tongue motion, and is rather invasive. The data collection is also very time consuming, and researchers are limited in the utterances they can record.
Examples of use	Butcher <i>et al</i> (2004) used static palatography to examine the articulations and acoustics of Australian Aboriginal languages, when compared to English.
Overall safety	The use of an edible black paint makes it safe to use, however the insertion of a mirror into one's mouth poses certain risks.
Sampling Rate	N/A
Audio	No
Compatibility	
Cost per session	Minimal cost, only of camera, mirror and cheap paint.
Recent Review	No recent review found
Examples of Available datasets	Not Available

5.5. Video and optical tracking

Video and optical tracking is a simple, non-invasive method to record the movements of a patient's lips, jaw, and to some extent the tongue. When coupled with the uses of a mirror, the method also allows us to see a side view profile of the individuals mouth during speech. Displacement of the visible articulators are used to understand lip configurations. Additionally, the technique can be used in conjunction with other articulatory recording systems, such as ultrasound and EMA. Further details are highlighted in table 7.

Table 7. A table covering nine different factors surrounding the use of video and optical tracking in capturing oral movements.

Advantages	The technique, due to its simplicity, is not field limited. Video recording (with audio) can be performed anywhere and yield good results, given the experiment is performed well with appropriate equipment.
Disadvantages	It is only able to show the movements of external articulators and provide no direct information as to what is going on inside the mouth.
Examples of use	(Wrench and Balch-Tomes, 2022) investigates the use of pose estimation to perform markerless estimation of speech articulators; using hand-labelled camera images.
Overall safety	The technique only uses a video camera to record the mouth, and is therefore considered safe.
Sampling Rate	'The frame rate of a typical camcorder/video camera is 24–30 fps' (Kochetov 2020a); however high speed cameras can record around 125fps
Audio	Yes, audio can be recording alongside video
Compatibility	when using a camera.
Cost per session	Cost is only that on a standard camera or camera phone. Around £150–200
Recent Review	No recent review found
Examples of Available datasets	Not Available

5.6. X-ray microbeam

X-ray microbeam (XRMB) is a computer-controlled point tracking system that uses a narrow (0.4 mm in diameter) x-ray beam to locate and track the movements of gold pellets attached to the target organ (in this case, target articulator); these include the lips, jaw, and tongue (Barlow and Stumm 2009). Serving as a reference point, two additional gold pellets are attached to the bridge of the nose. As presented in table 8, the scanned images produce a shadow, detected on a sodium iodide crystal detector, which is then transmitted to a computer that allows us to study the movements of the articulators.

Table 8. A table covering nine different factors surrounding the use of x-ray microbeam in capturing oral movements.

Advantages	The technique can be combined with cine-fluorography to examine the displacement of the tongue during articulation of deaf subject (Barlow and Stumm 2009).
Disadvantages	The technique has a complex configuration which makes it difficult to use, and additionally produces some x-ray radiation exposure to subjects.
Examples of use	(Whalen <i>et al</i> 2018) compares the variability in articulation and acoustics for 32 speakers, from the Westbury XRMB database (Westbury <i>et al</i> 1990).
Overall safety	X-ray microbeam, although safer than x-ray, exposes patients to some radiation. Thus, would not be considered safe (especially with alternative options present).
Sampling Rate	145 Hz
Audio	Unclear
Compatibility	
Cost per session	Unknown
Recent Review	No recent review found
Examples of Available datasets	X-ray microbeam speech Production database, consisting of 57 English speakers (Westbury <i>et al</i> 1990)

5.7. Electroglottography (EGG)

Electroglottography is a low-cost and non-invasive method to indirectly observe vocal fold vibratory activity during laryngeal voice production (Herbst 2020). The technique works by attaching two electrodes to a participant's thyroid cartilage (Adam's apple area),

Table 9. A table covering nine different factors surrounding the use of electroglottography in capturing oral movements.

Advantages	EGG is a non-invasive process that has little influence on the process of articulation. The weak high frequency signal generated during the process does not have any damaging effect on tissue, nor does it cause any significant uncomfortable sensation.
Disadvantages	EGG does not provide any information pertaining to the degree of glottal opening, only that it is open or closed. Additionally, there is the difficulty in determining the difference between laryngeal and laryngopharyngeal constrictions. Brunelle <i>et al</i> (2010) further documented problems with reliable detection of signals for individuals with smaller larynges (Kochetov 2020b).
Examples of use	Study uses 3D ultrasound and electroglottography to analyse the speech produced by 6 voice actors, intended to help in understanding which muscles are under volitional control.
Overall safety	Considered safe: 'An electroglottograph such as the laryngograph is a safe and non-threatening instrument that was well accepted by most children.' (Cheyne <i>et al</i> 1999).
Sampling Rate	Sampled at 22.05–44.1 kHz, as with audio recordings (Kochetov 2020b).

Table 9. (Continued.)

Audio	Yes,
Compatibility	
Cost per session	~\$850
Recent Review	Herbst (2020), titled 'Electroglottography—An Update' review recent developments in the field of EGG. The paper covers a span of 25 years, where it summarises some earlier contributions and developments since the last review as completed.
Examples of Available datasets	The CMU Arctic speech databases—consists of 3 speakers of English, phonetically balanced sentences with EGG recordings. (Kominek and Black 2004).

and passing through each side a high-frequency, low amperage current. Moments when the folds are closed allows the free passage of this current, where as moments when the folds are open prevents the flow (Kochetov 2020b); thus producing a graph of open and closed instances (a cycle). The technique shows the movements of the velum and larynx. Further details are highlighted in table 9.

5.8. Endoscopy and photoglottography

This section defines two techniques that both pertain to study of glottal activity; the techniques can quantify the movements of the tongue, velum and larynx. As detailed in table 10, endoscopy works by inserting a laryngoscope down a patient's throat, to observe glottal activity. The laryngoscope is a thin tube, with attached to it a video camera and light. Since the presence of the endoscope hinders one's ability to speak, the technique is limited to only allowing the study of certain sounds, such as prolonged vowels. Photoglottography (PGG) is a system developed by Sonesson (Chi *et al* 2021), also used in studying glottal behaviours. This method is rather non-invasive

Table 10. A table covering nine different factors surrounding the use of endoscopy and photoglottography in capturing oral movements.

Advantages	PGG is a relatively non-invasive method to capture glottal activity, and can be combined with other methods, such as airflow and air pressure (Kim <i>et al</i> 2018, Kochetov 2020b).
Disadvantages	Endoscopy is an invasive technique that can at times warrant the need for anaesthesia, and thus experiments involving this method often have very few participants (Kochetov 2020b). Additionally, the method requires the use of qualified medical personnel to run the experiment which means its availability is low and cost is high.
Examples of use	(Moisik <i>et al</i> 2014)—used laryngoscopy and laryngeal ultrasound to examine Mandarin tone production.
Overall safety	The technique is generally considered safe, with plenty of surrounding literature and common use in health care.

Table 10. (Continued.)

Sampling Rate	Cameras used with endoscopes typically record videos at a standard frame rate (about 30 frames per second [fps]) (Kochetov 2020b). For PGG, the sampling rate is higher, over 8 kHz.
Audio Compatibility	Only in PGG, evident from Chi <i>et al</i> (2021) with the creation of a portable PGG capable of recording audio.
Cost per session	~£1000
Recent Review	Through not a review, Chi <i>et al</i> (2021) titled 'Portable Photoglottography for Monitoring Vocal Fold Vibrations in Speech Production' explored a portable PGG unit that made it possible to record PGG and audio samples with relative ease.
Examples of Available datasets	None Available

(compared to the former). It works by shining an external light down the oesophagus, and using external sensors placed on the skin surface below the glottis. These sensors detect changes in light intensity, and therefore provide an indirect image of glottal width.

5.9. Oral airflow/pressure

Techniques in this section observe the study of oral airflow and air pressure, called pneumotachography.

Table 11. A table covering nine different factors surrounding the use of oral airflow/pressure in capturing oral movements.

Advantages	The technique allows for simultaneous audio recording and is generally not too invasive; for a technique that can record the movement of four articulators.
Disadvantages	Hirshkowitz and Kryger (2017) reported that the mask fitting around the mouth can be an uncomfortable experience for the patient. Additionally, it was unclear what the cost of such a system would be, this is also suggestive of the rarity of such a machine and thus, high cost.
Examples of use	(Kim <i>et al</i> 2018) - the paper uses PGG and intra-oral air pressure to understand the speech mechanisms and laryngeal features involved in the Korean Language.
Overall safety	There is little literature surrounding the safety of the technique. However, such mask-based systems are in common use within the medical field.
Sampling Rate	'Signals are digitized at a very high sampling rate (e.g., 1.375 kHz)...' (Kochetov 2020b).
Audio Compatibility	Yes
Cost per session	Unknown
Recent Review	No recent review found
Examples of Available datasets	None available

The technique can record the movements of the lips, tongue, velum and the larynx. Data regarding oral air flow is collected using a mask that is placed around the patient's mouth, the individual speaks into the mask while holding it against their mouth; this process can be uncomfortable (Hirshkowitz and Kryger 2017). Alongside audio, the system records speech air flow measured by the volume of air that leaves the mouth within a certain period. Intraoral air pressure can be monitored by using a small tube attached to the mask that is inserted into the patient's mouth (Kochetov 2020b). Further details are highlighted in table 11.

5.10. Nasal airflow

Nasalance is a subsequent method following on from the previous that allows us to measure nasal air flow. The technique uses two microphones positioned between the nose and the upper lip to measure the amplitude related to the air released by the nasal tract and air emitted by oral tract. The nasal air flow provides a rough measure of velum height. The device used for the recordings (such as the Nasometer II 6450) is held up to the mouth by the patient (Kochetov 2020b). Further details are highlighted in table 12.

Table 12. A table covering nine different factors surrounding the use of nasal airflow in capturing oral movements.

Advantages	Compared to other laryngeal recording techniques, such as Endoscopy and pneumotachography, Nasalance is the least invasive, that only uses external instruments.
Disadvantages	Interpretation of Nasalance data is not as simple as the other two methods, and requires normalisation: taking into account background air and noise.
Examples of use	(Echternach <i>et al</i> 2021)—the paper uses Nasalance to study its effects on vocal fold oscillation patterns during classical singing.
Overall safety	With its common use with both younger and elderly population it can be thought that it is relatively safe.
Sampling Rate	'The sampling rate for such recordings is high, as is usually the case for audio recordings (e.g., 44.1 kHz).' (Kochetov 2020b).
Audio Compatibility	Yes
Cost per session	Unknown
Recent Review	No recent review found
Examples of Available datasets	None available

5.11. Real time magnetic resource imaging (rtMRI)

Real time Magnetic Resonance Imaging enables the examination of the entire vocal tract during speech production, illustrating and quantifying articulator choreography in space and time (Ramanarayanan *et al* 2018). Explored in table 13, the unique technique can provide dynamic information of an individuals mid sagittal (or other) planes of interest, capturing not only labial and jaw motion but also velum, pharynx, and larynx. (Narayanan *et al* 2014). MRI machines can produce scans in any orientation of interest, though most commonly they are done in the sagittal plane (Kochetov 2020b). This side-on view allows us to view both internal and external articulators in motion,

Table 13. A table covering nine different factors surrounding the use of real time magnetic resource imaging in capturing oral movements.

Advantages	The technique can provide a complete view of the oral cavity. Technique has a relatively non-invasive means of imaging, with the ability to image in 3D or several arbitrary 2D planes. Despite a low sampling rate, advances in parallel imaging and sparse reconstruction have aided with increasing image resolution (Ramanarayanan <i>et al</i> 2018).
Disadvantages	The cost and maintenance of the machines are considerable. With most MRI machines being in hospitals, researchers require a close collaboration with these institutes to take MRI recording for articulatory studies. An additional consideration is the effects the spine position adopted by subjects has on one's ability to articulate. However, a x-ray microbeam study conducted concluded that the spine position has minimal effect on speech production (Tiede <i>et al</i> 2000, Ramanarayanan <i>et al</i> 2018).
Examples of use	Iribar <i>et al</i> (2019) presented an experimental articulatory characterisation of vocalizations in Basque, using MRI midsagittal images.
Overall safety	'clear advantage over other methods with respect to patient safety, relative non-invasiveness of imaging...'(Ramanarayanan <i>et al</i> 2018).
Sampling Rate	5 to less than 100 Hz (Narayanan <i>et al</i> 2014).
Audio	Yes.
Compatibility	
Cost per session	Unclear.
Recent Review	Kennerley <i>et al</i> (2021), titled 'Real-time magnetic resonance imaging: mechanics of oral and facial function' is an extract from the 'British Journal of Oral and Maxillofacial Surgery'.
Examples of Available datasets	RT-MRI with synchronized audio 3D volumetric MRI Static T2w MRI—American English—75 speakers (Lim <i>et al</i> 2021). RT-MRI with synchronized audio 3D volumetric MRI—French—2 speakers (Dourous <i>et al</i> 2019).

making it possible to deduce a potential correlation within their movements. RtMRI scans can also be coupled with EMA readings (Kim *et al* 2014). This creates a system that simultaneously illustrates the movement of the articulators in EMA and rtMRI scans.

5.12. X-rays

X-ray imaging is one of the two techniques capable of capturing the movements of the whole vocal tract. The research of articulatory movements is concerned with a form of medical imaging called 'Projectional Radiography', which produces 2D x-ray images; during which process, X-ray beams are passed through tissue and recorded on a special detector plate. The method is a resource intensive technique that must generally be performed by a radiographer, in a setting adapted with shielding for medical personnel (table 14). Due to the health hazards of intensive x-ray use, the technique is no longer in use and new data sets are not being created. Most work done is based on data collected from the 1970s, digitalised by Munhall *et al* (1995).

Table 14. A table covering nine different factors surrounding the use of x-rays in capturing oral movements.

Advantages	The technique can provide a complete view of the oral cavity. The technique is non-invasive (although not particularly safe) and has a very high frame rate.
Disadvantages	The denser structure of bones, compared to the tissue of vocal organs, can obstruct the view of parts of the tongue in the images produced. Additionally, the greatest disadvantage, is the hazard to health the process poses (Kochetov 2020b); something that cannot be justified when other methods are available.
Examples of use	(Sock <i>et al</i> 2011)—Used an x-ray database to develop and present processing tools that used for projects on speech production.
Overall safety	X-ray imaging can be said to be the dangerous technique out of all listed. X-rays can cause DNA mutations in living organisms, as it is known to induce DNA strand breaks (Immel <i>et al</i> 2016). This can lead to later development of cancer and it is thus considered a carcinogen.
Sampling Rate	30×10^3 Hz to 30×10^4 Hz
Audio	Yes, evident from creation of dataset synchronised with audio (Sock <i>et al</i> 2011).
Compatibility	
Cost per session	~£101
Recent Review	No recent review found
Examples of Available datasets	(Munhall <i>et al</i> 1995): Consists of 25 films (totalling 55 min) of x-ray footage converted from film collected in the 1970s. The data set contains a total of 14 Canadian English and French speakers.

6. Discussion

6.1. Growing interest in vocal tract MRI and overarching problems of existing methods

So far, we have discussed the existing techniques that allow us to quantify, to varying degrees of detail, the movements of the mouth's external and internal structure (as well as other activities happening in the vocal tract) during the articulatory and masticatory processes. These methodologies each have their own limitations (and likewise, advantages) that they present, something researchers take into consideration during the selection process.

The 2020s decade saw a growing interest in real-time vocal tract MRI owing to the distinctive multi-articulator capabilities it offers. Currently, four publicly accessible datasets have been released (Douroso *et al* 2019, Scholes and Skipper 2020, Lim *et al* 2021, Ruthven *et al* 2021). Each dataset is accompanied by transparent protocols detailing the procedures during data collection, specifying the MRI machine utilised, and the specific coil configurations employed.

Moreover, datasets are available upon request (Birkholz *et al* 2020, Dediu *et al* 2022, Isaieva *et al* 2023), and the associated protocols, even when the dataset is not immediately available, serve as valuable guidelines for others to gather high-quality data (Lim *et al* 2023, Wu *et al* 2023). Notably, there is a uptick in the utilisation of machine learning techniques in tasks related to vocal tract MRI (Ribeiro *et al* 2022, Laprie *et al* 2023, Ruthven *et al* 2023). Additionally, a toolkit for assessing vocal tract shape has been developed (Belyk *et al* 2023).

As the interest in Real-time Magnetic Resonance Imaging (rtMRI) continues to surge, researchers are continually driven to seek novel and innovative solutions for advancing speech analysis research.

Although recent evidence highlights the promising potential of rtMRI, it is not immune to significant limitations that are shared by the majority of techniques used to study the vocal tract. They are usually too expensive or invasive, and quite often both; those that bypass these constraints are limited to only one or two articulators. They are not practical for consumer purposes, and companies often cannot spend the amount of money that is required in performing data collection for systems such as MRI and EMA. Furthermore, individuals are not interested in partaking in a time consuming, and at times invasive process. As a result, collection of primary data is limited. This is a problem previously identified by a University of Southern California study group specialising in speech production and articulation (SPAN). They are working on bridging this gap by creating open source MRI datasets aimed at fuelling the development of applications and ideas inspired by AI and machine learning methods (Lim *et al* 2021).

In order to aid a wider research community, low cost, non-invasive and time efficient systems and

methodologies are required. One such approach can be inspired by the use of internal-external correlation modelling. There is very little research in applying this technique in creating solutions for the oral cavity, but we have seen similar and relevant approaches applied to other parts of the body. Below we will address and review these works that link the external and internal. They are relatively non-invasive techniques that predict internal structures, based on external observations.

6.2. Internal-external correlation modelling

Internal-external correlation models are a method to estimate the motion and presence/location of an internal object, based on its external view. Although the use of such an approach has not yet been fully explored in oral modelling, it can be found being used in other organ modelling systems.

Chen *et al* (2018) explore the development of a local topology preserved non-rigid point matching algorithm, used in creating an internal-external correlation model for internal action mapping with applications in lung cancer radiotherapy treatment. Organs and tumours in the thoracic region go through significant respiration-induced motion - translation, rotation and deformation. This motion can be utilised to accurately track both tumours and surrounding organs at risk. This is done by registering the vector fields, which describe the motion between internal and external components. They are acquired by individually aligning the meshes of internal organs and external surfaces from the images via the developed algorithm.

Several other studies have also demonstrated the feasibility of finding correlations between internal and external motions, detected by respiratory surrogates. Fayad *et al* (2011) aimed at assessing motion correlation between a patient's external surface and internal anatomical landmarks. They concluded that it is possible to reduce variability and associated errors in respiratory motion synchronisation and motion modelling process by capturing in real-time the motion of the complete external patient surface as well as choosing the area of the surface that correlates best with the internal motion.

Martin *et al* (2013) presented a novel method to build a surrogate driven motion model of a tumour using a Dental Cone Beam scan, without the need of markers. The method was shown to extract tumour motion from a variety of lung cancer patients, with tumours present in different location within the cavity. By tracking the movement of an external reference point in real time, doctors can use this model to guide treatments that are synchronized with the tumour's motion. The model is created just before each treatment session to account for any changes in the tumour's position. This method also helps doctors better understand the shape and movement of the tumour before delivering precise radiotherapy treatments. The method involves two steps. First, the

tumour area is highlighted in the CT scan images. Then, the model is created based on the movement of an external reference point. In tests using simulated data, the average difference between the estimated and actual tumour positions was reduced to just 1 millimetre. When applied to real patient data, the average difference between the estimated and clinically-identified tumour positions was less than 2.5 millimetres in both up-down and sideways directions.

6.3. Modelling the Interrelationship between the face and vocal tract

Applying this approach to our problem requires us to first define the internal and external components. The external component is that of the face, and the internal is the vocal tract. The face during articulation (or indeed mastication) can be captured using a RGB recording camera. These videos or pictures can be captured from several views, including the forward and side on views. Figure 1 shows a still frame from a video where the participant shown utters the phrase 'Miss black thought about the lap'. It includes the simultaneous capturing of the coronal (frontal) and sagittal (longitudinal) planes.

The internal view can be represented with either of the articulatory recording techniques mentioned in section 4. Depending on a researcher's specific requirements they could choose any of the 12 methods. Ideally the chosen technique would be one that involves the use of as many articulators as possible, to maximise the learned traits from the two modalities. Out of all the currently viable techniques, real time MRI is the single option that can record all the articulators as well as provide a view of the entire vocal tract. Due to this very reason, rtMRI stands out as one of the most suitable techniques to represent the internal view. Figure 2 below shows the MRI view for the external view frame of figure 1.

Continuing with the two modalities mentioned, the task here would then be to find the correlation between the representation depicted in a RGB external camera view, and the MRI internal vocal tract view. To simplify the explanation, the problem can be expressed as follows:

Variables:

M = Input from internal view (rtMRI)

I = Input form external view (Camera)

Y = Oral actions

D = Interrelationship between M and I

D' = Interrelationship between M and I over time.

For any given frame pair, we can state D to be the relationship between M and I , represented as:

$$D = \{M, I\}$$

We can additionally observe the relationship between M and I in the context of a whole oral action, represented through multiple consecutive frames (a video). A whole sentence in the form of a video would include temporal information present over multiple frames. With Y as oral actions and D' as the

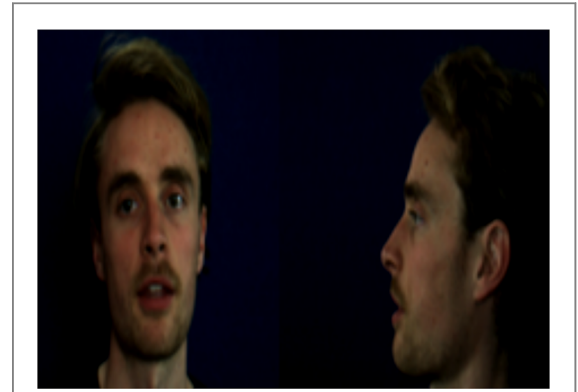


Figure 1. Image illustrating the coronal (left) and sagittal (right) view of the face (Scholes and Skipper 2020), (Reproduced from Scholes and Skipper (2020) under CC BY license).

relationship between M and I over time. D' can be represented as:

$$D' = \{M, I, Y\}.$$

When determining D or D' in various combinations of internal or external modes of representation, certain factors should be taken into account. For each specific articulation, there exists an absolute ground truth for both the internal and external views, which corresponds to the real-time movement of the articulators. However, the choice of modality used to represent either of the two views is limited by the constraints of the specific technique employed. For instance, in the case of MRI, factors such as pixel resolution, frame rate, or the fidelity of MRI signal deconvolution can impact the representation and potentially alter the ground truth view, depending on the MRI machine being used. Therefore, any interrelationship between the two views must consider the fact that the captured modality represents an interpretation of the absolute ground truth. It is therefore crucial to acknowledge that the modality employed to capture the views is subject to specific constraints, potentially leading to variations in the ground truth. Any connections drawn between the two views should consider the interpretive nature of the captured modality.

As far as our findings indicate, (Scholes and Skipper 2020) is the only work aimed at investigating the link between facial and vocal tract movements during speech production. They formed a unique dataset that consists of paired, temporarily aligned videos of both the face (captured in the front on, side and 45-degree angles) and sagittal MRI view during 10 different utterances. Using this aligned cross modal dataset they applied principal component analysis (PCA) to demonstrate that the MR images sequences can be reconstructed with high fidelity using videos of only the external face. The PCA worked by capturing dynamic regions of the vocal tract, such as the tongue and lips, while ignoring static areas with little movement though an utterance (such as brain/spinal cord). MR sequences could then be reconstructed by projecting the video input data into the MR PCA space

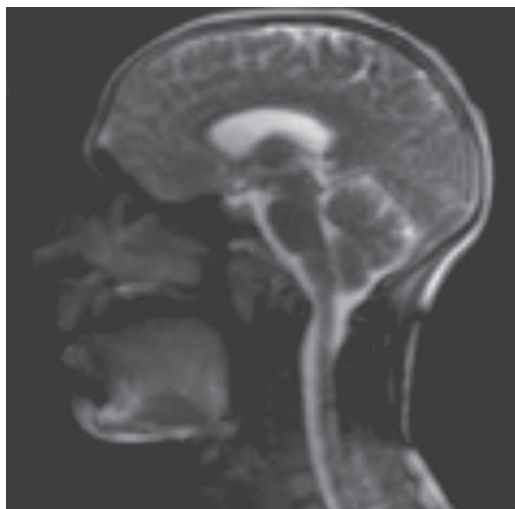


Figure 2. Image illustrating a MRI sagittal view, showing all the articulators forming the vocal tract. (Scholes and Skipper 2020), (Reproduced from Scholes and Skipper (2020) under CC BY license).

generated; the opposite was done for the generation of the external view from the internal view. Resultant reconstructed MR sequences (from video input) were very similar to the original sequences. Their work revealed that there is sufficient information in the face to recover vocal tract shape during speech, for set utterances, and likewise reconstruct sequences from either of the two imaging domains. However, it does not bypass the need for acquiring the MRI sequence itself. To reconstruct either of the sequences, the PCA space of the opposing modality is still required; so, to create a reconstructed MRI sequence, the original MRI is still needed. A solution apt at addressing our problem statement must be able to produce the MRI sequence of an external video without having its specific corresponding PCA space. In other words, be able to produce an internal representation of the vocal tract without needing a specific matching external view. This generative approach would be a result of the learned correlation between the face and vocal tract during articulation.

A solution to address this problem could potentially be found in computer vision/deep learning approaches, a field centred around learning the characteristics of a dataset to then predict and interpret visual information. By leveraging the power of neural networks and advanced algorithms, computer vision and deep learning can analyse images or videos, extract meaningful features, and make novel predictions based on the learned patterns.

6.4. Deep learning approaches for image synthesis

In the previous section we briefly discussed the main drawback of traditional statistical correlation modelling approaches when it comes to image synthesis, the need of both imaging domains for an inference. Deep learning, a subset of machine learning, is well posed to provide a solution to this computer vision problem.

Computer vision is a field of AI aimed at enabling computers to derive meaningful information from visual stimuli. These tasks can range from low-level edge detection to a high-level task such as complete scene understanding. Over the last decade, impressive developments in computer vision have come because of advancements in deep learning.

Deep learning is a machine learning method used in training artificial neural networks. With the growing availability of large scale datasets and ever increasing processing power of computers, researchers are apt in developing pattern recognition models, for use in many fields including medical imaging (Esteva *et al* 2021).

Deep learning involves training artificial neural networks to learn patterns and make predictions. As the availability of extensive datasets and the processing power of computers continue to expand, researchers have been able to develop highly effective pattern recognition models. These models find applications in various domains, including medical imaging, as demonstrated by Esteva *et al* (2021).

The backbone of deep learning in computer vision is the Convolutional Neural Network (CNN). These are specifically designed to analyse visual data by mimicking the human visual system. CNNs consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers. Convolutional layers perform feature extraction by applying a set of filters or kernels to the input image. Each filter captures different visual patterns, such as edges, textures, or shapes, and convolving this with the input image produces a feature map. The use of shared weights and local receptive fields in convolutional layers enables the network to learn hierarchical representations of the input data. Pooling layers reduce the spatial dimensions of the feature maps by down sampling them. This helps in creating more robust features by discarding irrelevant spatial information and retaining important features. Common pooling techniques include max pooling and average pooling. Fully connected layers are responsible for making predictions based on the features extracted by the previous layers. These layers connect every neuron from the previous layer to every neuron in the current layer, enabling the network to learn complex relationships between features and make accurate predictions.

The problem we wish to address is how a dataset of paired MRI and external view images can be used to synthesis MRI views of the face, from patterns recognised between the two modalities. This falls into the field of Image-to-image translation. Image-to-image translation is the process by which an image from one modality is transformed into another, with the aim of learning the relationship between the input and output image. This is a deep learning task (often addressed using generative adversarial networks). Such models are best trained with datasets of paired and aligned images. The concept of image-to-image translation

allows for various applications, such as converting images from one style to another (e.g., grayscale to color), transforming images across different modalities (e.g., day to night), or synthesizing realistic images from rough sketches. It's important to note that successful image-to-image translation relies on having a well-prepared and representative dataset for training, as well as careful selection and design of the deep learning architecture and loss functions. Additionally, understanding the limitations of the approach, such as potential artifacts or biases in the synthesized images, is crucial for ensuring the reliability and accuracy of the generated data.

A recent survey of Cross-Modality Synthesis was done by Xie *et al* (2022), that comprehensively approaches this complex task from different perspectives, including the level of supervision, loss function, range of modality and downstream tasks. The downstream task in this case would be the use of the MRI generated image.

6.5. Semantic segmentation of the articulators

Exploring how to extract meaningful information regarding articulatory movements from both original and generated MR sequences is a vital step towards further realising how rtMRI can be used in articulatory research. Knowing the relative positions of each of the vocal organs in a given frame will allow for clearer image understanding; thus improving the fidelity of any generated outcomes. For this process, researchers employ segmentation techniques to analyse vocal tract MRIs, enabling a comprehensive evaluation of the vocal tract's structure and function during speech.

Image segmentation (or more specifically Medical Image Segmentation) is a process used in identifying meaningful regions and structures within a medical image, a process through which a desired object (vocal organ) is extracted from a medical image (2D or 3D) (Li *et al* 2021). The modality of acquiring the medical image can be through systems such as CT, MRI, X-ray and more.

The process in our use case involves precisely delineating the various anatomical components, including the tongue, lips, jaw, and velum, within the vocal tract. Accurate segmentation facilitates the extraction of quantitative measurements and geometric data about the vocal tract's regions. These measurements offer insights into speech production biomechanics, aiding in the understanding of speech disorders, language development, and treatment efficacy.

Segmentation approaches include manual delineation, semi-automated algorithms, and deep learning-based methods. Manual delineation involves experts manually tracing boundaries, ensuring precise results but requiring significant time and effort. Semi-automated algorithms assist by providing initial segmentations that can be refined manually. Deep learning techniques, employing convolutional neural

networks, automatically recognise and segment vocal tract structures, reducing time and effort.

In the case of the medical field, this is often for planning and guiding operations as well as measuring the outcome of therapeutic procedures (Kapur *et al* 2014).

During image segmentation, the various sections of the target image are delineated and are given labels. To put this into perspective, we can observe an example of an annotated (delineated) still image by Ruthven *et al* (2021). Figure 3 shows an MRI view of the vocal tract to the left, alongside the same image annotated with each colour representing a different articulator.

Several segmentation technologies are available that can perform image segmentation utilising machine learning and deep neural networks. These approaches require ground truth data to train models to accurately segmenting new, unseen images. However, the delineation process is widely acknowledged as highly complex, and as a result, a significant portion of the annotation process continues to be performed manually (Wallner *et al* 2019).

Segmentation techniques used (also to create ground truth data), can be broadly divided into two categories: intensity-based segmentation and shape-based segmentation. Each of these two methods have various semi- and fully-automatic segmentation algorithms. The following section will present examples of these and discuss some of the algorithms in use, in relation to the techniques adopted in the segmentation of rtMRI images of the face (as in figure 3).

6.5.1. Intensity-based segmentation

Intensity-based segmentation (IBS) relies on the principle that voxels within the target object, such as an organ, possess a distinct grey value (intensity) different from their surrounding structures. Even if this disparity is subtle and imperceptible to the human eye, models can effectively discern these differences. However, medical images often exhibit a wide range of grey scale distribution within the target object itself, which poses challenges in accurately distinguishing voxel intensities. IBS models encompass various techniques, including thresholding, clustering, deep learning, watershed, and graph-cut, each with its own advantages and applications.

Thresholding-based segmentation is particularly effective when applied to images with high voxel contrast compared to their surroundings. This technique is well-suited for imaging bony structures and their surrounding tissues in CT scans, where there is a significant contrast in voxel intensities. Clustering is an unsupervised learning method that groups voxels within an image based on their similarities, without the need for ground truth data. It can identify clusters of voxels with similar characteristics (e.g. intensity), aiding in the segmentation process. Region-growing-based segmentation is an iterative process initiated by selecting a single seed point within the target object

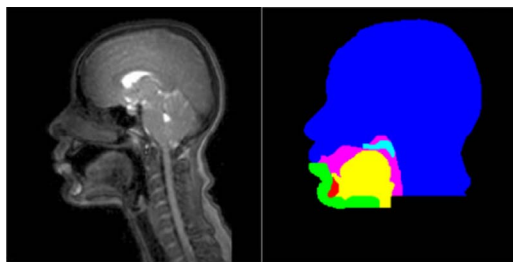


Figure 3. A still frame of a 2D rtMRI recording (left) alongside a manually annotated segmented form of the image, illustrating each articulator as an instance (right). (Reproduced from Ruthven *et al* (2021) under CC BY 4.0 Licence).

manually. From this seed, the region grows and expands until it encompasses the entire target object. The underlying assumption is that voxels within the same object are similar, allowing the algorithm to determine when to stop the region's expansion.

In summary, intensity-based segmentation techniques offer various approaches for segmenting medical images. By leveraging concepts like thresholding, clustering, deep learning, and region-growing, these techniques enable the identification and differentiation of voxels belonging to the target object, despite the challenges posed by the wide range of voxel intensities within the image.

6.5.2. Shape-based segmentation

In shaped-based segmentation (SBS), the outline of the target object is roughly known in advance, such that segmentation can be completed by identifying a particular shape. In the case of the mouth this could be likened to, for example, identifying the positioning of the upper lip. Such methods explicitly use prior knowledge of a target shape, such that the target shape is learned from a group of pre-annotated shape templates. These techniques include statistical shape models, statistical appearance models, and atlas based segmentation. These pre-annotated images, also known as template images, may limit shape variations and differences, as they may not necessarily be present in the target images; poorly annotated images can reduce segmentation quality.

Statistical shape models (SSM) work by mathematically describing the geometric shape of the target object. The variations in the target object are learned through shape templates (annotated images). This three-stage process involves: the construction of shape templates, SSM creation from shape templates, and adapting SSM to new image.

A good SMM will use a large set of shape templates to allow the model to learn shape variations and variabilities that occur. These shape templates are formed by manually annotating medical images. Once the model has been trained it should ideally be able to identify the target shape when segmenting, thus allowing it to be applied to new images.

Statistical appearance modelling (SAM) works on a similar principle to SSM, but additionally incorporating the appearance of a shape. This includes the colour and 'texture' (e.g. represented by voxel intensity) of the target object.

Atlas-based segmentation is an SBS technique that that can segment images without the need of well-defined delineations between regions and pixel intensities. The approach utilises reference images and corresponding segmentation templates (atlases) to form transformation matrixes, enabling reference images to be registered with the new image itself. The atlas is able to provide an approximate location of the object position in an image and this information therefore allows the model to localise the object within the new image, and further distinguishes between the object of interest and its surrounding.

Active contour modelling, another SBS, differs from the previously mentioned models. It does not necessarily require such training templates. In this case the algorithm utilises the contour present within the image to form a delineation. This form of modelling can be seen in use within photo editing software, for example the Lasso tool in Photoshop. Through an iterative process, a user places several marks around a target object present in an image, which the model then connects based on the contour around the shape. This does however mean that the initial contour must be provided by a user manually, most of the time, for the contour to be then found automatically.

6.5.3. Deep learning semantic segmentation approaches

The previous two subsections primarily went over earlier segmentation algorithms, approaches that are still currently in use. As section earlier touched upon, for the past several years deep learning-based approaches have paved a way for a new generation of image segmentation models with outstanding performance improvements (Minaee *et al* 2020).

A deep learning segmentation pipeline typically consists of dataset preparation, network architecture selection, training, validation, and inference. First, a labelled dataset is created, comprising input images and corresponding ground truth annotations that define the desired segmentation. A suitable deep neural network architecture, such as U-Net (Ronneberger *et al* 2015) or Mask R-CNN (Kaiming *et al* 2017), is then chosen or designed specifically for the segmentation task. The network is trained using the labelled dataset, with its parameters optimized iteratively to minimise a chosen loss function, such as pixel-wise cross-entropy or Dice coefficient, which quantifies the dissimilarity between predicted segmentations and the ground truth.

Validation is performed using a separate dataset to assess the network's performance and guide any necessary fine-tuning. Once trained, the network is ready for inference, where it takes unseen input images

and produces segmentation predictions by applying the learned patterns and features.

Deep learning-based segmentation has demonstrated remarkable capabilities in various fields, including medical imaging, object detection, and semantic segmentation. Its ability to automatically learn relevant features from large datasets has significantly advanced the accuracy and efficiency of segmentation tasks, leading to important applications in computer vision research and real-world applications.

Here, we focus on a few different type of network architectures used specifically for medical image segmentation. Not focusing too much on other factors such as type of learning or loss functions, the following section provides a low level overview of a few of these networks, listing both 2D and 3D architectures.

Originally proposed in 2015, U-net is a convolutional neural network (CNN) developed for 2D biomedical image segmentation (Ronneberger *et al* 2015). The CNN has a modified architecture, adopting a symmetrical structure and skip connections aimed at allowing for optimal model training on medical image datasets. The networks popularity can be deduced from its ability to learn segmentation in an end-to-end setting, ability to precisely localise and distinguish borders and work well with very few annotated images. Currently, U-net has become the standard for most medical image segmentation tasks and the backbone from which several other popular architectures are structured (Wang *et al* 2022).

Most recently, a complete segmentation of the vocal tract was done to delineate 4 different articulators and the vocal tract (Ruthven *et al* 2021). A dataset of five participants was used, each subject counting from numbers one through to ten in British English whilst in a RtMRI machine. Between the five participants, there were a total of 392 MR images (or frames) which were segmented by a radiologist. The paper successfully presented an automatic method to fully segment multiple groups of articulators and the vocal tract using a U-net like framework and additionally provide a novel clinically relevant metric for assessing the accuracy of vocal tract and articulator segmentations. Although generalisability was noted to be good, the work stated that the model performed less favourably in preserving airway gaps between articulators, especially in the case of soft palate closures in instances where the ground truth data suggested the space was open. Larger classes provided better dice coefficient and general Hausdorff distances than those that were smaller, as could be expected. Future work requires addressing these factors and potentially using a larger range of vocal tract configurations. It is additionally worth noting that the model's applicability to recordings taken from other MRI machines is unclear, and it is likely it will not perform well for images not taken in the same MRI machine as the paper used.

Medical image data produced, being either CT or MRI are often taken in 3D. To take advantage of these

high-dimensional data sets, Çiçek *et al* (2016) furthered the u-net architecture to be applied to 3D data, proposing an architecture apt at performing segmentation directly, named 3D U-net. However, due to computational limitations posed when using such a dataset, the number of down-sampling steps had to be reduced, resulting in a model with a reduced segmentation accuracy.

V-Net worked around this problem by employing residual connections to create a deeper network with more down-sampling steps. Although the network performance did improve, the 3D segmentation network, and others developed after, face an underlying issue surrounding the need to high computational power and GPU memory, often not available during the training process.

7. Conclusion

To conclude, we have thoroughly explored the various oral actions the vocal tract goes through and how this pertains to the articulatory function. A detailed summation of the methods of articulatory recording has been provided, addressing their advantages and limitations, as well as other metrics relevant to their use. Cross domain image-to-image translation seems viable, constraints surrounding datasets can be worked around, whilst the use of image segmentation shows promising applications for processing downstream tasks. The overarching problems associated with these systems have been talked about in the discussion section, with the subsequent sub-sections providing the foundations for spear heading solutions viable in addressing problems related to speech analysis and speech correction, mastication and, more broadly, oral processing. Throughout the paper, particularly in reference to the 12 modelling techniques, we have seen the potential clinical significance of a technique capable of modelling the complete mouth. In its simplest form, a viable way to model the complete mouth will see down steam applications in speech correction and designing foods for the aging population. In the dental field we would be able to gain information about patient's oral actions that would become part of creating a personalised dental treatment plan. In its initial state, image-to-image translation holds the potential to facilitate seamless transitions between diverse MRI weightings. Given the variability of MRI machines in hospitals, leveraging this technology could prove instrumental in enhancing the fidelity of vocal tract MRI frames.

Acknowledgments


This research was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) and Haleon PLC (formerly known as GSK Consumer Healthcare), BB/V509553/1. Drs Shridevi Pandit,

David Bradshaw and Maria-Teresa Addison (Haleon PLC) are gratefully acknowledged for their insights and help with conceptualising this research.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Muhammad Suhaib Shahid  <https://orcid.org/0000-0003-2431-7697>

Andrew P French  <https://orcid.org/0000-0002-8313-2898>

Michel F Valstar  <https://orcid.org/0000-0003-2414-161X>

Gleb E Yakubov  <https://orcid.org/0000-0001-5420-9422>

References

- Al-hammuri K, Gebali F, Thirumarai Chelvan I and Kanan A 2022 Tongue contour tracking and segmentation in lingual ultrasound for speech recognition: a review *Diagnostics* **12** 2811
- Anderson V B 2008 Static palatography for language fieldwork *Language Documentation & Conservation* **2** 1–27
- Ball M J 2021 ed M J Ball (Routledge, Taylor & Francis Group) (<https://doi.org/10.4324/9780429320903>)
- Barlow S M and Stumm S 2009 Speech production: adult *Encyclopedia of Neuroscience* ed L R Squire (Academic) 247–54
- Belyk M, Carignan C and McGettigan C 2023 An open-source toolbox for measuring vocal tract shape from real-time magnetic resonance images *Behav. Res. Methods* **55**
- Bennett R, Chiosain M, Padgett J and McGuire G 2017 An ultrasound study of Connemara Irish palatalization and velarization *Journal of the International Phonetic Association* **48** 1–44
- Birkholz P, Kürbis S, Stone S, Häsner P, Blandin R and Fleischer M 2020 Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties *Sci. Data* **7** 255
- Bliss H, Abel J and Gick B 2018 Computer-assisted visual articulation feedback in L2 pronunciation instruction: a review *JSLP* **4** 129–53
- Brunelle M, Nguyễn D D and Nguyễn K H 2010 A laryngographic and laryngoscopic study of northern vietnamese tones *PHO* **67** 147–69
- Butcher A, Butcher A, Tabain M and Tabain M 2004 On the back of the tongue: dorsal sounds in australian languages *PHO* **61** 22–52
- Chen H, Zhong Z, Yang Y, Chen J, Zhou L, Zhen X and Gu X 2018 Internal motion estimation by internal-external motion modeling for lung cancer radiotherapy *Sci. Rep.* **8** 3677
- Cheyne H A, Nuss R C and Hillman R E 1999 Electrolaryngography in the pediatric population *Archives of Otolaryngology–Head & Neck Surgery* **125** 1105–8
- Chi Y, Honda K and Wei J 2021 Portable photoglottography for monitoring vocal fold vibrations in speech production, in: ICASSP 2021 - 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP) Presented at the ICASSP 2021 - 2021 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) 6438–42
- Çiçek Ö, Abdulkadir A, Lienkamp S S, Brox T and Ronneberger O 2016 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation *Medical Image Computing and Computer-Assisted Intervention* (https://doi.org/10.1007/978-3-319-46723-8_49)
- Cleland J, Wrench A A, Scobbie J M and Semple S 2011 9th International Seminar on Speech Production 2011, ISSP 2011 Comparing articulatory images: an MRI / Ultrasound Tongue Image database
- Dediu D, Jennings E M, van't Ent D, Moisiuk S R, Di Pisa G, Schulze J, de Geus E J C, den Braber A, Dolan C V and Boomsma D I 2022 The heritability of vocal tract structures estimated from structural MRI in a large cohort of Dutch twins *Hum Genet* **141** 1905–23
- Docio-Fernandez L and García Mateo C 2015 Speech production *Encyclopedia of Biometrics* ed S Z Li and A K Jain (Springer) 1493–8
- Douros I K, Felblinger J, Frahm J, Isaieva K, Joseph A, Laprie Y, Odille F, Tsukanova A, Voit D and Vuissoz P-A 2019 A Multimodal Real-Time MRI articulatory corpus of french for speech research *INTERSPEECH 2019 - 20th Annual Conf. of the Int. Speech Communication Association Graz, Austria* (<https://doi.org/10.21437/Interspeech.2019-1700>)
- Dromey C, Hunter E and Nissen S L 2018 Speech adaptation to kinematic recording sensors: perceptual and acoustic findings *Journal of Speech, Language, and Hearing Research* **61** 593–603
- Echternach M, Högerle C, Köberlein M, Schlegel P, Döllinger M, Richter B and Kainz M-A 2021 The effect of nasalance on vocal fold oscillation patterns during the male passaggio *Journal of Voice* **35** 500.e9–00.e16
- Eshky A, Ribeiro M S, Cleland J, Richmond K, Roxburgh Z, Scobbie J and Wrench A 2018 UltraSuite: a repository of ultrasound and acoustic data from child speech therapy sessions *Interspeech* **2018** 1888–92
- Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, Liu Y, Topol E, Dean J and Socher R 2021 Deep learning-enabled medical computer vision *npj Digit. Med.* **4** 1–9
- Fayad H, Pan T, Clement J F and Visvikis D 2011 Technical note: correlation of respiratory motion between external patient surface and internal anatomical landmarks *Med. Phys.* **38** 3157–64
- Hardcastle W, Jones W, Knight C, Trudgeon A and Calder G 1989 New developments in electropalatography: A state-of-the-art report *Clinical Linguistics & Phonetics* **3** 1–38
- Hasegawa-Johnson M 1998 Electromagnetic exposure safety of the Carstens Aerticulograph AG100 *J. Acoust. Soc. Am.* **104** 2529–32
- Herbst C T 2020 Electrolaryngography—An update *Journal of Voice* **34** 503–26
- Hirshkowitz M and Kryger M 2017 Chapter 164 - monitoring techniques for evaluating suspected sleep-related breathing disorders *Principles and Practice of Sleep Medicine* ed M Kryger, T Roth and W C Dement (Elsevier) 6th edn 1598–609.e3
- Hoke P, Tiede M, Grender J, Klukowska M, Peters J and Carr G 2019 Using electromagnetic articulography to measure denture micromovement during chewing with and without denture adhesive *Journal of Prosthodontics* **28** e252–8
- Hoole P and Zierdt A 2010 Five-dimensional articulography *Speech Motor Control: New Developments in Basic and Applied Research* ed B Maassen and P van Lieshout (Oxford University Press) (<https://doi.org/10.1093/acprof:oso/9780199235797.003.0020>)
- Immel A et al 2016 Effect of x-ray irradiation on ancient DNA in sub-fossil bones—Guidelines for safe x-ray imaging *Sci. Rep.* **6** 32969
- Iribar A, Pagola R M, Túrrez I, Arroyo J L G, Zapirain B G and Ruiz I O 2019 Parameters of tongue shape of /n/ and /l/ in Basque *Journal of the International Phonetic Association* **49** 207–21
- Isaieva K, Odille F, Laprie Y, Drouot G, Felblinger J and Vuissoz P-A 2023 Super-resolved dynamic 3D reconstruction of the vocal tract during natural speech *J. Imaging* **9** 233
- Ji A, Berry J J and Johnson M T 2014 The electromagnetic articulography mandarin accented english (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data 2014 *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*

- (ICASSP). Presented at the 2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) 7719–23
- Kapur T, Egger J, Jagadeesan J, Toews M and Wells W 2014 Registration and segmentation for image-guided therapy *Intraoperative Imaging and Image-Guided Therapy* (Springer) 79–91
- Keating P A 2001 Phonetics: Articulatory *International Encyclopedia of the Social & Behavioral Sciences* ed N J Smelser and P B Baltes (Pergamon) 11381–6
- Kelsey C A, Minifie F D and Hixon T J 1969 Applications of ultrasound in speech research *J. Speech Hear. Res.* **12** 564–75
- Kennerly A J, Mitchell D A, Sebald A and Watson I 2021 Real-time magnetic resonance imaging: mechanics of oral and facial function *British Journal of Oral and Maxillofacial Surgery* **60** 596–603
- Kaiming H, Gkioxari G, Piotr D and Ross G 2017 Mask R-CNN 2017 IEEE International Conference on Computer Vision (ICCV) 2980–8
- Kent R D 2015 Nonspeech oral movements and oral motor disorders: a narrative review *Am. J. Speech Lang Pathol.* **24** 763–89
- Kim H, Maeda S, Honda K and Crevier-Buchman L 2018 The mechanism and representation of Korean three-way phonation contrast: external photoglottography, intra-oral air pressure, airflow, and acoustic data *PHO* **75** 57–84
- Kim J, Lammert A C, Ghosh P K and Narayanan S S 2014 Co-registration of speech production datasets from electromagnetic articulography and real-time magnetic resonance imaging *J. Acoust. Soc. Am.* **135** EL115–21
- Kochetov A 2020a Research methods in articulatory phonetics I: Introduction and studying oral gestures *Language and Linguistics Compass* **14** e12368
- Kochetov A 2020b Research methods in articulatory phonetics II: Studying other gestures and recent trends *Language and Linguistics Compass* **14** e12371
- Kominek J and Black A W 2004 The CMU ARCTIC speech databases *5th ISCA Speech Synthesis Workshop* In 5th ISCA Speech Synthesis Workshop 223–4
- Ladefoged P and Johnson K 2015 *A Course in Phonetics* (Wadsworth Publishing)
- Laprie Y, Ribeiro V, Isaeva K, Leclere J, Felblinger J and Vuissoz P-A 2023 Modeling the temporal evolution of the vocal tract shape with deep learning *20th Int. Congress on Phonetic Sciences* (<https://inria.hal.science/hal-04209848>)
- Li J, Erdt M, Janoos F, Chang T and Egger J 2021 1 - Medical image segmentation in oral-maxillofacial surgery *Computer-Aided Oral and Maxillofacial Surgery* ed J Egger and X Chen (Academic) 1–27
- Lim Y et al 2021 A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images *Sci. Data* **8** 187
- Lim Y, Kumar P and Nayak K S 2023 Speech production real-time MRI at 0.55 T *Magn. Reson. Med.* **91** 337–43
- Martin J, McClelland J, Yip C, Thomas C, Hartill C, Ahmad S, O'Brien R, Meir I, Landau D and Hawkes D 2013 Building motion models of lung tumours from cone-beam CT for radiotherapy applications *Phys. Med. Biol.* **58** 1809–22
- Mat Zin S, Md Rasib S Z, Suhaimi F M and Mariatti M 2021 The technology of tongue and hard palate contact detection: a review *Biomed. Eng. Online* **20** 17
- Meenakshi N, Yarra C, Yamini B K and Ghosh P K 2014 Comparison of speech quality with and without sensors in electromagnetic articulograph AG 501 recording *Annual Conf. of the Int. Speech Communication Association, Interspeech* 935–9
- Mennen I, Scobbie J M, de Leeuw E, Schaeffler S and Schaeffler F 2010 Measuring language-specific phonetic settings *Second Language Research* **26** 13–41
- Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N and Terzopoulos D 2020 Image segmentation using deep learning: a survey *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44** 3523–42
- Moisik S, Lin H and Esling J 2014 A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS) *Journal of the International Phonetic Association* **44** 21–58
- Munhall K G, Vatikiotis-Bateson E and Tohkura Y 1995 X-ray film database for speech research *J. Acoust. Soc. Am.* **98** 1222–4
- Narayanan S et al 2014 Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC) *J. Acoust. Soc. Am.* **136** 1307–11
- Ramanarayanan V, Tilsen S, Proctor M, Töger J, Goldstein L, Nayak K S and Narayanan S 2018 Analysis of speech production real-time MRI *Comput. Speech Lang.* **52** 1–22
- Rebernik T, Jacobi J, Jonkers R, Noiray A and Wieling M 2021 A review of data collection practices using electromagnetic articulography *Laboratory Phonology* **12** 1–42
- Reda R, Zanza A, Cicconetti A, Bhandi S, Miccoli G, Gambarini G and Di Nardo D 2021 Ultrasound imaging in dentistry: a literature overview *Journal of Imaging* **7** 238
- Ribeiro V, Isaeva K, Leclere J, Vuissoz P-A and Laprie Y 2022 Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated *Speech Commun.* **141** 1–13
- Ronneberger O, Fischer P and Brox T 2015 U-Net: convolutional networks for biomedical image segmentation *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (https://doi.org/10.1007/978-3-319-24574-4_28)
- Ruthven M, Miquel M E and King A P 2021 Deep-learning-based segmentation of the vocal tract and articulators in real-time magnetic resonance images of speech *Comput. Methods Programs Biomed.* **198** 105814
- Ruthven M, Miquel M E and King A P 2023 A segmentation-informed deep learning framework to register dynamic two-dimensional magnetic resonance images of the vocal tract during speech *Biomed. Signal Process. Control* **80** 104290
- Scholes C and Skipper J I 2020 The inter-relationship between the face and vocal-tract configuration during audio-visual speech *Proceedings of the National Academy of Sciences* **177** 32791–8
- Sigona F, Stella M, Stella A, Bernardini P, Gili Fivela B and Grimaldi M 2018 Assessing the position tracking reliability of Carstens' AG500 and AG501 electromagnetic articulographs during constrained movements and speech tasks *Speech Commun.* **104** 73–88
- Soboļeva U, Lauriņa L and Slaidiņa A 2005 The masticatory system—an overview *Stomatologija* **7** 77–80 PMID: 16340271
- Sock R et al 2011 An X-ray database, tools and procedures for the study of speech production *ISSP 2011 - 9th International Seminar on Speech Production. Montréal (Canada)* 41–8 (<https://hal.archives-ouvertes.fr/hal-00610297>)
- Stone M 2005 A guide to analysing tongue motion from ultrasound images *Clin Linguist Phon.* **19** 455–501
- Tabain M 2011 Electropalatography data from Central Arrernte: a comparison of the new Articulate palate with the standard Reading palate *Journal of the International Phonetic Association* **41** 343–67
- Tiede M K, Masaki S and Vatikiotis-Bateson E 2000 Contrasts in speech articulation observed in sitting and supine conditions *Proc. of the 5th Seminar on Speech Production* 25–8
- van der Bilt A, Engelen L, Pereira L J, van der Glas H W and Abbink J H 2006 Oral physiology and mastication *Physiology & Behavior, Making Sense of Food* **89** 22–7
- Verhoeven J, Miller N R, Daems L and Reyes-Aldasoro C C 2019 Visualisation and analysis of speech production with electropalatography *Journal of Imaging* **5** 40
- Wallner J, Schwaiger M, Hochegger K, Gsaxner C, Zemmann W and Egger J 2019 A review on multipatform evaluations of semi-automatic open-source based image segmentation for cranio-maxillofacial surgery *Comput. Methods Programs Biomed.* **182** 105102
- Wang R, Lei T, Cui R, Zhang B, Meng H and Nandi A K 2022 Medical image segmentation using deep learning: a survey *IET Image Proc.* **16** 1243–67
- Westbrook K E, Nessel T A, Hohman M H and Varacallo M 2022 Anatomy, head and neck, facial muscles *StatPearls* (StatPearls Publishing)
- Westbury J, Milenkovic P, Weismer G and Kent R 1990 X-ray microbeam speech production database *J. Acoust. Soc. Am.* **88** S56–56

- Whalen D H, Chen W-R, Tiede M and Nam H 2018 Variability of articulator positions and formants across nine English vowels *Journal of Phonetics* **68** 1–14
- Wilson I 2014 Using ultrasound for teaching and researching articulation *Acoust. Sci. Technol.* **35** 285–9
- Wood S 2010 Electropalatography in the assessment and treatment of speech difficulties in children with down syndrome *Down Syndrome Research and Practice* **12** 98–102
- Wrench A and Balch-Tomes J 2022 Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut *Sensors* **22** 1133
- Wu P, Li T, Lu Y, Zhang Y, Lian J, Black A W, Goldstein L, Watanabe S and Anumanchipalli G K 2023 arXiv:2307.02471 Deep Speech Synthesis from MRI-Based Articulatory Representations
- Xie G, Wang J, Huang Y, Zheng Y, Lu K, Jin Y and Zheng F 2022 Cross-modality neuroimage synthesis: a survey **56** 80