



# THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

### The magnetic resonance imaging subset of the mngu0 articulatory corpus

**Citation for published version:**

Steiner, I, Richmond, K, Marshall, I & Gray, C 2012, 'The magnetic resonance imaging subset of the mngu0 articulatory corpus' Journal of the Acoustical Society of America, vol 131, no. 2, pp. EL106-EL111. DOI: 10.1121/1.3675459

**Digital Object Identifier (DOI):**

[10.1121/1.3675459](https://doi.org/10.1121/1.3675459)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Publisher's PDF, also known as Version of record

**Published In:**

Journal of the Acoustical Society of America

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# The magnetic resonance imaging subset of the mngu0 articulatory corpus

**Ingmar Steiner**

*INRIA/LORIA Speech Group, Bat. C, 615 Rue du Jardin Botanique, 54600 Villers-lès-Nancy, France  
ingmar.steiner@inria.fr*

**Korin Richmond**

*Centre for Speech Technology Research, University of Edinburgh, Informatics Forum,  
10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom  
korin@cstr.ed.ac.uk*

**Ian Marshall**

*Medical Physics & Medical Engineering, University of Edinburgh, Chancellor's Building,  
49 Little France Crescent, Edinburgh, EH16 4SB, United Kingdom  
ian.marshall@ed.ac.uk*

**Calum D. Gray**

*Clinical Research Imaging Centre, University of Edinburgh, Queen's Medical Research Institute,  
47 Little France Crescent, Edinburgh, EH16 4TJ, United Kingdom  
calum.gray@ed.ac.uk*

**Abstract:** This paper announces the availability of the magnetic resonance imaging (MRI) subset of the mngu0 corpus, a collection of articulatory speech data from one speaker containing different modalities. This subset comprises volumetric MRI scans of the speaker's vocal tract during sustained production of vowels and consonants, as well as dynamic mid-sagittal scans of repetitive consonant–vowel (CV) syllable production. For reference, high-quality acoustic recordings of the speech material are also available. The raw data are made freely available for research purposes.

© 2012 Acoustical Society of America

**PACS numbers:** 43.70.Aj, 43.70.Jt [AL]

**Date Received:** November 28, 2011      **Date Accepted:** December 13, 2011

## 1. Introduction

Technology applications that use speech data, for example, text-to-speech (TTS) synthesis and automatic speech recognition (ASR), focus almost exclusively on acoustics. However, the acoustic signal produced by a human speaker crucially depends on the shape of the speaker's vocal tract and the movements of articulators such as the tongue or lips. Techniques such as motion capture or medical imaging can provide valuable articulatory data to supplement the acoustic speech signal. Applications such as TTS synthesis or ASR can exploit such data to improve their modeling of human speech production.

However, recording articulatory data is unfortunately not as straightforward as recording an acoustic signal. Specialist facilities and expertise are typically required, which make articulatory recordings more expensive. Indeed, the recording process itself can often be tricky, with practical complications that may also increase the burden on the subject. Consequently, the few articulatory corpora, which have been made freely available (e.g. [Munhall \*et al.\*, 1995](#); [Westbury, 1994](#); [Wrench, 2000](#); [Narayanan \*et al.\*, 2011](#)), have been well received and extensively used by the research community. Each of these contains articulatory and acoustic data for a range of speakers, but the data for any individual speaker may be insufficient for certain applications.

The mngu0 articulatory corpus addresses this shortfall and provides a large amount of articulatory data from a single speaker of British English. The corpus consists of multiple subsets of data acquired in different modalities, including a large number of sentences recorded using video and electromagnetic articulography (EMA), dental casts, and magnetic resonance imaging (MRI) scans of the speaker's vocal tract. Taken together, they provide both high-speed (200 Hz) articulatory movement data during running speech, and the speaker's vocal tract geometry in three dimensions, and offer a valuable resource to the speech community. Although the EMA portion of the mngu0 corpus has previously been described and made publicly available (Richmond *et al.*, 2011), the purpose of *this* paper is to announce the availability of the corresponding MRI data for this speaker and describe the details of its acquisition.

## 2. MRI data

The purpose of MRI scanning the mngu0 speaker was two-fold. First, we wanted to capture the three-dimensional (3D) geometry of the speaker's vocal tract, as well as representative configurations of the speaker's articulators for producing a range of speech sounds. A series of volumetric MRI scans was performed for this purpose. Second, we wanted to investigate and capture the effects of coarticulation, for which a set of dynamic MRI scans was performed. Overall, therefore, by adapting the procedure described in Birkholz and Kröger (2006), a prompt list was designed that consisted of sustained vowels and consonants, as well as dynamic vowel–consonant–vowel transitions, elicited by repetitive production of consonant–vowel (CV) syllables.

The scanner used for this study was a GE Medical Systems Signa HDx 1.5T. The speaker was placed in the scanner in the supine position and fitted with a head-and-neck radio frequency coil. All of the scans were completed within one 120 min session, with a short break between the sustained and dynamic scans.

The mngu0 subject is a trained, professional speaker. As well as general advantages in terms of the level of performance obtained, this was beneficial for the MRI scans in particular, as he was able to sustain production of vowels for 20 s on average, and of repetitive utterances for approximately 10 to 15 s. The speaker was therefore capable of producing each prompt over the entire duration of a scan.

As Fig. 1 shows, the region of interest (ROI) is clearly visible in the resulting image data, extending from the lips to the rear wall of the pharynx in anterior–posterior direction, from the larynx to the nasal cavity in inferior–superior direction, and (in the volumetric scans) laterally between the mandibular joints.

Figure 1 also exhibits an aliasing artifact; a segment of the back of the speaker's head and neck appears at the anterior edge of the images, overlapping with the speaker's nose. This occurs when the imaged anatomy extends outside the field of view (FOV). But as we wanted to gain good coverage of the vocal tract and the aliasing does not impact image quality in the ROI, we chose this configuration and the aliasing can safely be ignored.

Due to the acoustic conditions within the scanning chamber and the lack of a noise-canceling fiber-optic microphone, no simultaneous acoustic recordings were possible during the MRI session. For this reason, a separate acoustic recording session was conducted as well (see Sec. 3).

### 2.1 Volumetric scans

Static, 3D scans of 13 sustained vowels and 15 sustained consonants were acquired with a fast gradient echo sequence having 26 sagittal slices of 4 mm thickness, repetition time (TR) 51 ms, echo time (TE) 3 ms, flip angle 30°, and FOV 280 mm reconstructed as 256 × 256 pixels.

The prompts are listed in Table 1(a). For each scan, the speaker produced the corresponding prompt, maintaining audible production of the target phone during the entire acquisition.



Fig. 1. Cutaway volume rendering of raw volumetric data for [a].

Figure 1 and Mm. 1 illustrate one raw volumetric scan [a], whereas Fig. 2 and Mm. 2 show a 3D vocal tract mesh extracted from that scan, along with a surface rendering of the face and head (the aliasing artifact has been removed).

**Mm. 1.** Cutaway volume rendering of volumetric [a] scan (360° horizontal rotation, static clip plane). (Quick-Time movie, 1.8 MB.)

**Mm. 2.** 3D vocal tract extracted from volumetric [a] scan, with surface rendering (360° horizontal rotation). (QuickTime movie, 1.8 MB.)

## 2.2 Dynamic mid-sagittal scans

Dynamic scans of 16 consonants in three vocalic contexts [a, i, u] were acquired with a fast gradient echo sequence similar to that mentioned previously, but with a single midline sagittal slice of thickness 10 mm, and TR 4 ms and TE 2 ms, enabling 40 consecutive time frames to be acquired in 10 s.

The prompts are listed in Table 1(b). The speaker produced each prompt as repetitive CV syllables, synchronizing production of the target consonants to the scans by timing their stable phase with the noise emitted by the MRI scanner.

An example of the dynamic data for nasals [m, n, ŋ] is displayed in Fig. 3. Each panel shows an overlay of 30 MRI frames, providing an “averaged” image that offers an enhanced view of the articulators.

## 2.3 Dental reconstruction

As the teeth are invisible in MRI,<sup>1</sup> a final volumetric scan was acquired using blueberry juice [which has favorable nuclear magnetic resonance (NMR) properties, due to its high manganese content] to distinguish oral cavity from teeth. This produced a negative 3D scan of the teeth for subsequent dental reconstruction.

For this dental scan, the speaker lay prone in the scanner, and filled his mouth with the juice by sucking it from a bottle through a flexible tube. Although this did produce images clearly showing the teeth, there was no time left in the session for a

Table 1. Prompt lists for the MRI scanning session. The orthographic prompts are emphasized, with the underlined letter(s) corresponding to the target phone; the target phone itself is given in IPA notation.

(a) Sustained production <sup>a</sup>			
<u>hi</u> t	ɪ	<u>fi</u> n	f
<u>pe</u> t	ɛ	<u>thi</u> n	θ
<u>ha</u> t	æ	<u>si</u> n	s
<u>ho</u> t	ɒ	<u>shi</u> n	ʃ
<u>hu</u> t	ʌ	<u>mo</u> ck	m
<u>pu</u> t	ʊ	<u>knoc</u> k	n
<u>hea</u> t	i	<u>thi</u> ng	ŋ
<u>hoo</u> t	u	<u>ri</u> ng	ɹ
<u>hur</u> t	ɜ	<u>lo</u> ng	l
<u>har</u> t	ɑ	<u>lo</u> ch	x
<u>ou</u> ght	ɔ	<u>sl</u> ee <u>p</u>	l
<u>ab</u> out	ə	⟨p⟩	p
<u>the</u> re	ɛɹ	⟨t⟩	t
		⟨k⟩	k
		<u>ba</u> ll	l
(b) Dynamic production <sup>b</sup>			
<u>ap</u> a	<u>ip</u> i	<u>up</u> u	p
<u>at</u> a	<u>it</u> i	<u>ut</u> u	t
<u>ak</u> a	<u>ik</u> i	<u>uk</u> u	k
<u>af</u> a	<u>if</u> i	<u>uf</u> u	f
<u>ath</u> a	<u>ith</u> i	<u>uth</u> u	θ
<u>as</u> a	<u>is</u> i	<u>us</u> u	s
<u>ash</u> a	<u>ish</u> i	<u>ush</u> u	ʃ
<u>ar</u> a	<u>ir</u> i	<u>ur</u> u	r
<u>al</u> a	<u>il</u> i	<u>ul</u> u	l
<u>am</u> a	<u>im</u> i	<u>um</u> u	m
<u>an</u> a	<u>in</u> i	<u>un</u> u	n
<u>ang</u> a	<u>ing</u> i	<u>ung</u> u	ŋ
<u>ach</u> a	<u>ich</u> i	<u>uch</u> u	x
<u>ar</u> a	<u>ir</u> i	<u>ur</u> u	ɹ
<u>aw</u> a	<u>iw</u> i	<u>uw</u> u	w
<u>ay</u> a	<u>iy</u> i	<u>uy</u> u	j

<sup>a</sup>Vowels are shown on the left, consonants on the right. The prompts ⟨p,t,k⟩ represent the speaker holding the occlusion of the corresponding stop.

<sup>b</sup>Each prompt represents an individual scan, yielding the vocal tract configuration for the target phone in the vocalic context of [a, i, u], respectively.

long acquisition, and therefore the spatial resolution of the dental scan is no higher than that of the other volumetric scans.

### 3. Acoustic reference recordings

To compensate for not being able to simultaneously record the acoustic signal, and to give the speaker opportunity to familiarize himself with the prompt list and the general procedure in an informal, nonclinical environment, high-quality acoustic reference recordings were made on the day before the MRI session.

The acoustic recording session took place in a sound-proofed room at the Informatics Forum, University of Edinburgh. The prompts were recorded using a DPA Type 4035 microphone mounted on a headset. The microphone signal was captured directly to hard disk using an EDIROL UA-25 audio interface connected to a

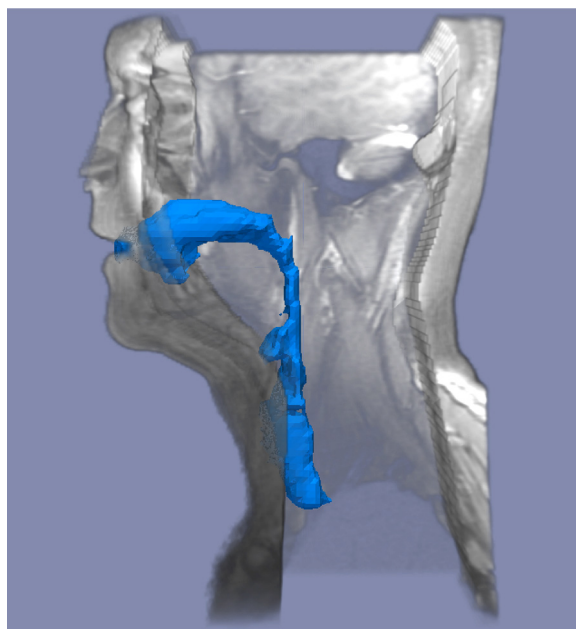


Fig. 2. (Color online) 3D vocal tract extracted from [a] scan, shown with surface rendering.

laptop computer. The recordings were made with a 96 kHz sampling rate at 24 bit quantization.

The speaker read out the prompt list twice, once standing upright, and again in the supine position. This was done to allow comparison in the acoustic domain

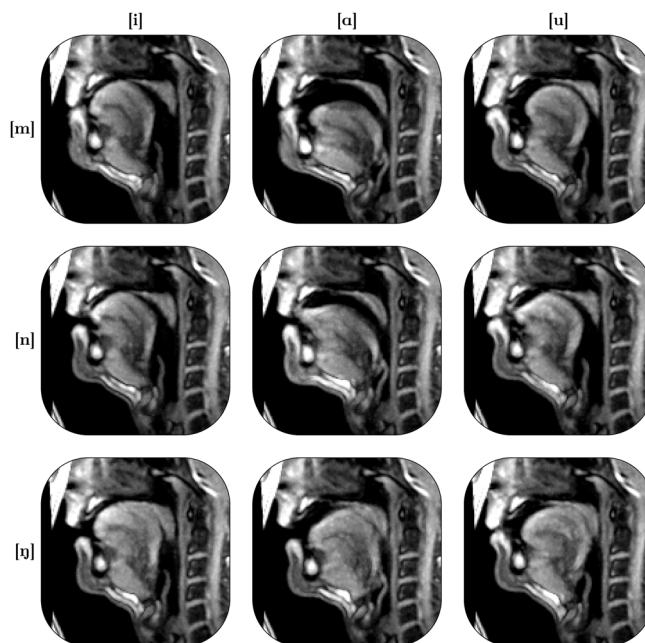


Fig. 3. Overlay of 30 mid-sagittal frames of [m, n, ŋ] (rows) dynamically produced in vocalic context [i, a u] (columns). The critical articulators for each consonant (lips, tongue tip, and tongue dorsum, respectively) achieve occlusion, whereas the tongue body assumes a different target shape in each vocalic context. The velum is lowered in all conditions, allowing the speaker to sustain production of the nasal.



between these two postures. The supine recordings were made to ensure that the audio matched the articulatory configuration during the MRI scans, where gravity and posture can influence articulation (e.g., Kitamura *et al.*, 2005).

The acoustic speech data has been manually segmented into prompts.

#### 4. Distribution

The raw data from the MRI session has been made anonymous to protect the privacy of the speaker, but is otherwise unmodified from the DICOM files produced at the MRI facility.

The volumetric scans and dynamic scans will be made freely available for research purposes, as separate downloads at a dedicated website, <http://mngu0.org/>. The acoustic data and the corresponding segmentation data will likewise be made available.

#### Acknowledgments

Imaging was carried out at the Brain Research Imaging Centre, Edinburgh (<http://www.bric.ed.ac.uk>), Division of Clinical Neurosciences, University of Edinburgh, Western General Hospital, Edinburgh, a core area of the Wellcome Trust Clinical Research Facility and part of the SINAPSE collaboration (<http://www.sinapse.ac.uk>). This work was supported by Marie Curie Early Stage Training Site “EdSST” (MEST-CT-2005-020568) and EPSRC Grant No. EP/E027741/1 (“ProbTTS”).

<sup>1</sup>Their NMR properties make them nearly indistinguishable from the surrounding air.

#### References and links

- Birkholz, P., and Kröger, B. J. (2006). “Vocal tract model adaptation using magnetic resonance imaging,” in *Proceedings of the 7th International Seminar on Speech Production*.
- Kitamura, T., Takemoto, H., Honda, K., Shimada, Y., Fujimoto, I., Syakudo, Y., Masaki, S., Kuroda, K., Oku-Uchi, N., and Senda, M. (2005). “Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner,” *Acoust. Sci. Technol.* **26**, 465–468.
- Munhall, K. G., Vatikiotis-Bateson, E., and Tohkura, Y. (1995). “X-ray film database for speech research,” *J. Acoust. Soc. Am.* **98**, 1222–1224.
- Narayanan, S., Bresch, E., Ghosh, P. K., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A., Proctor, M., Ramanarayanan, V., and Zhu, Y. (2011). “A multimodal real-time MRI articulatory corpus for speech research,” in *Proceedings of Interspeech*, pp. 837–840.
- Richmond, K., Hoole, P., and King, S. (2011). “Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus,” in *Proceedings of Interspeech*, pp. 1505–1508.
- Westbury, J. R. (1994). *X-Ray Microbeam Speech Production Database User's Handbook Version 1.0* (University of Wisconsin Press, Madison, WI).
- Wrench, A. A. (2000). “A multi-channel/multi-speaker articulatory database for continuous speech recognition research,” *PHONUS* **5**, 1–14.