Comparing Articulatory Images: An MRI / Ultrasound Tongue Image Database

Joanne Cleland[1], Alan A. Wrench[2], James M. Scobbie[1], Scott Semple[3]

[1]CASL (Clinical Audiology, Speech and Language) Research Centre, Queen Margaret University, Musselburgh, EH21 6UU, Scotland, UK

[2]Articulate Instruments Ltd, Queen Margaret University Campus, Musselburgh, EH21 6UU, Scotland, UK

[3]CRIC (Clinical Research Imaging Centre), Queen's Medical Research Institute, Edinburgh, EH16 4TJ, Scotland, UK

jcleland@qmu.ac.uk

Abstract. *We report the development of a database that will contain paired ultrasound and MRI of tongue movements and shapes from 12 adults, illustrated with pilot data from one speaker. The primary purpose of the database will be to evaluate the informational content of ultrasound tongue images on the basis of the richer articulatory structures visible with MRI, and to provide tongue shape information that can later be incorporated into an image processing algorithm to enhance ultrasound tongue images. Ultrasound is an increasingly popular technique for studying speech production since it provides a real-time image of tongue movements. Its potential as a visual-feedback speech therapy tool has been recognised but has not yet been exploited to any great extent. In part this is because obstruents like /t/ /k/ /tʃ/, which are important targets for therapy, have tongue shapes in both canonical and common error productions which ultrasound displays rather poorly compared to the more easily-imaged vowels, glides and liquids. By enhancing ultrasound images in real time with information based on our corpus, we aim to create images which we hypothesise will A) be more easily understood by children for clinical feedback B) extend the range and utility of ultrasound generally.*

1. Introduction

1.1 Background

Ultrasound has become an increasingly popular technique for studying speech production. It provides a real-time image of most of the surface of the tongue plus partial information about other structures such as the tongue's internal musculature, position of the hyoid and mandible and, when the tongue is pressed against it, the hard palate. It is safe, non-invasive and relatively cheap to use. It has, however, several serious drawbacks. Firstly, information about the tip of the tongue, a key active articulator, is often lost when the tip is raised and extended forward due to a sublingual airspace. Secondly, the image suffers from artefacts such as double edges,

discontinuities, reflections and general poor image quality (Stone, 2005) making the image difficult to interpret. Finally, since the hard palate and pharynx are often not visible during speech there is no obvious or consistent coordinate space for measurement (Scobbie et al., 2011).

Clinically, few studies have used ultrasound as a visual feedback technique for people with speech disorders; perhaps because of difficulties clients may have interpreting the image. Those studies that have used ultrasound have done so with adolescents or adults who are likely to process the cognitive skills to interpret ultrasound images more easily (for example, Bernhardt, 2005). However, there is some evidence that children with Speech Sound Disorders (SSDs) may benefit from visual feedback techniques (Michi et al., 2003) which allow them to see and modify real-time images of their own articulations. This suggests that enhancing the ultrasound image to make it easier for children to interpret could have real clinical benefit, and therefore be an important component of future feedback-based treatment systems. Bernhardt and colleagues have primarily used ultrasound for the remediation of /r/, vowels and /s/. In the UK the consonants /t d k g s/ plus /tʃ/ and /ʃ/ are frequent targets for therapy (though seldom /r/): however, the ultrasound image is often poor for these segments.

A new three year ULTRAX project (EPSRC, EP/I027696/1) aims to apply a tongue model to this problem, making use of explicit sequence-based optimization for dynamic tracking (and smoothing) through time. That is, by using tongue contour models from other imaging techniques, such as MRI, it may be possible to enhance the ultrasound image in real-time so as to provide effective visual feedback for these segments.

1.2 Using MRI to model tongue contours

"Enhancing" 2D ultrasound images can mean many things, but two main areas spring to mind when ultrasound images are compared to MRI images. One, not the topic here, is the addition of non-lingual articulatory structures, such as the lips or rear pharyngeal wall, velum, larynx (and perhaps non-articulatory but useful reference structures, such as the cranium and facial shape). The other relates to improving the existing tongue contour itself. Enhancement could deal with ambiguities, distortions or artefacts affecting the surface echo data; attenuation of surface echoes due to the angle of the tongue relative to the probe; missing data at the tongue root or tip due to masking by the hyoid or mandible; missing data due to a sublingual cavity. Some of these problems arise due to contact between tongue and palate, some when the tongue surface is nearly parallel to the ultrasound echo-pulse beams, perhaps in the root, or in the tip and blade during retroflexion.

Such problems might be especially important when dealing with disordered speech where articulations may occur outside of the normal range of places of articulation, even when the target is normally visible on ultrasound. For example, alveolar targets may be substituted with linguolabial productions. Magnetic Resonance Imaging (MRI) of static productions of speech sounds can provide a clear and detailed 2D image of the tongue in the midsagittal plane and indeed the whole of the vocal tract, during all sorts of segmental productions. By providing information that cannot

normally be seen with ultrasound, a matched dataset allows validation and enhancement of ultrasound images.

This paper will detail a protocol for a matched MRI-ultrasound database of tongue contours, which systematically samples a wide range of possible tongue shapes that may occur in both typical and disordered speakers. There are three key aims: 1. To use MRI to illustrate a full range of possible tongue shapes, including tongue-shapes normally only associated with disordered speech. 2. To validate ultrasound images by illustrating similar tongue shapes with both techniques. 3. To provide a corpus of MRI data which can be used to develop an algorithm for enhancing ultrasound images.

Previous attempts to match ultrasound and MRI images have met with difficulty (Lee & Stone, 2010). Frame rate and slice thickness differ significantly between MRI and ultrasound, posing a challenge in comparing the same point in an articulation across imaging techniques. We aim to overcome some of these difficulties by emulating MRI conditions as closely as possible when collecting ultrasound data.

2. Method

2.1 Participants

Data presented here is from a pilot recording of a trained adult male phonetician. The full database will use comparable participants. This means we can design materials to sample as many different tongue shapes as possible without relying only on language-specific inventories. Likewise, it allows us to model disordered speech shapes without recruiting a wide set of speakers with heterogeneous speech disorders. The trained speakers are also more able to replicate their productions on different occasions.

2.2 Materials

The main goal is to sample different places of articulation, with a close constriction, with some variety of secondary articulation, plus a range of approximants and vowels. The secondary goal is to sample some double articulations and variation in constriction.

For the full study stops and fricatives at most places of articulation detailed on both the IPA (IPA, 2005) and ExtIPA (ICPLA, 2002) will be sampled. Voiced approximants and nasals were sampled if they occurred in English or were considered easier to produce. Voiceless stops and fricatives were preferred over their voiced cognates. In order to emulate double articulations that occur in conversational speech, some clusters were included. To allow for the effects of coarticulation on tongue shape each consonant was sampled in four contexts: sustained, [aCa], [iCi] and [oCo]. Table 1 overleaf details a subset of these consonant materials used for the pilot recording. Further, cardinal vowels [i,e,ɛ,a,ɑ,ɔ,ʌ,o,u,ʉ,ə] were sampled in [pVp] (i.e. pseudo-word or real word) and in sustained contexts. All vowels were recorded in the pilot. In addition, some non-speech sounds and vocal-tract shapes were included (Table 1).

	Type	Sustained	a_a	i_i	o_o
	Dental	θ:	aθa	iθi	oθo
		t:	ata	iti	oto
		s:	asa	isi	oso
		ɹ:	aɹa	iɹi	oɹo
	Alveolar	ɬ:	aɬ	iɬ	oɬ
	Post-alveolar	ʃ:	aʃa	iʃi	oʃo
	Retroflex	ʈ:	aʈa	iʈi	oʈo
	Palatal	ç:	aça	içi	oço
	Velar	k:	aka	iki	oko
			akla	ikli	oklo
	Clusters		akɹa	ikɹi	okɹo
Pseudo-disordered (relative to English)	**Linguolabial**	l̼:	al̼a	il̼i	ol̼o
	Uvular	q:	aqa	iqi	oqo
	Uvular	χ:	aχa	iχi	oχo
	Pharyngeal	ħ:	aħa		
	Super-retroflex	ʇ:	aʇa		
	Click	ŋ!:	aŋ!a	iŋ!i	oŋ!o
	Alveolar trill	r:	ara	iri	oro
	Alveolar ejective	t':	at'a	it'i	ot'o
Non-speech		swallow	palate contact	cough	extreme protrusion
Double articulations	**Alveolar-velar**		at car	eat key	oat co
	Velar-alveolar		pack tap	peak tea	oak toe
Variants	**Alveolar [ts]**		pat sap	peat see	oat so
	Lateral release		at lap	eat lee	oat low

Table 1: Pilot study consonant and non-speech materials

2.3 MRI pilot recording procedure

MRI was acquired using a Siemens Verio 3T scanner (Siemens Medical, Germany). We aimed to get the sharpest image from as short a per-token image capture time as possible. For a given sample rate, the smaller the voxel size, the higher the spatial resolution but the weaker the signal-to-noise (SNR) ratio. Therefore increasing spatial resolution can actually result in loss of detail. SNR (and image quality) can be improved by increasing the repetition time (TR) between MRI excitation pulses but a long acquisition period increases the likelihood of the tongue moving slightly, causing motion blur, which again can result in loss of detail. SNR and image quality can also be improved by increasing the slice thickness. Again, if the slice is too thick then bluring will occur as structures vary with depth. Taking all those conflicting constraints into consideration, our optimised settings for MRI acquisition are as in Table 2.

The speaker sustained the segment of interest (i.e. the consonant in the VCV sequences, the vowel in the [pVp] sequences, the double articulation in the clusters, the stop in the variants) for between one and two seconds. Prompts were delivered via video-goggles and the speaker waited for the noise of the scanner to commence before

speaking. Each MRI sequence (a row in Table 1) lasted 14secs. Simultaneous audio recordings were made using a FOMRI-III™ Fiber Optic Microphone, which consists of two separate phase and amplitude matched microphones, arranged orthogonally, which capture speech and MRI noise. Adaptive noise cancellation is then applied to these signals.

Slice Thickness	8mm
Voxel size	1.71875 x 1.71875mm
Initial Image size (whole head)	192 x 192 mm
Cropped image size (vocal tract)	100 x 75 mm
Final interpolated image size (Lancoz3)	640 x 480 mm
Echo Time (TE)	52ms
Repetition time (TR)	400ms

Table 2: MRI settings

2.4 Ultrasound pilot recording procedure

Previous studies (e.g. Engwell, 2006) have suggested that speech produced in the supine position (required for MRI) differs due to a change in the direction of gravitational force on the speaker's body. In particular, a slight superior and posterior displacement of the tongue root in supine position is evident (Wrench, Cleland & Scobbie, 2011). To provide the best chance of being able to match ultrasound tongue contours with those collected in the MRI we therefore collected ultrasound data in supine position. Materials and instructions to sustain speech sounds were identical to those used for the MRI.

Ultrasound data was acquired using an Ultrasonix SonixRP machine remotely controlled via Ethernet from a PC running Articulate Assistant Advanced™ software (Articulate Instruments Ltd, 2010). The echo return data was recorded at 121fps with 69 beam-formed echo pulses evenly spread over a 135 degree field of view (FoV). A hardware pulse was generated by the SonixRP at the instant that each complete set of 69 echo pulses had been recorded. This synchronization pulse sequence was recorded on a multichannel analogue acquisition system at 22,050Hz along with the acoustic speech signal. The pulses were then detected in a post processing operation allowing each ultrasound frame to be accurately time tagged. A standard graphical interpolation is performed on the raw data to convert it to an image for analysis in AAA, similar to the image processing that is normally carried out within the ultrasound scanner. The depth setting was 80mm and the echo return vectors had 412 discrete samples (providing approximately 5 pixels per mm). The transducer frequency was 5MHz providing an axial resolution of approximately 0.9mm. Recordings were made in a sound-treated studio. The speaker was fitted with a headset to stabilize the ultrasound probe. Synchronous 60Hz de-interlaced NTSC video from a headset-mounted micro-camera, imaging a profile of the nose was used to verify that there was no movement of the probe relative to the head during speech.

2.5 Annotation and analysis

MRI DICOM files were converted to AVI, synchronized with the audio recordings and imported into AAA. For each segment a key frame was chosen that represented the midpoint of the segment. The slow frame rate of MRI (400ms) often provided only one or two frames free of motion blur. A spline was fitted to the tongue contour, using AAA. The spline starts at the anterior region where the underside of the tongue meets the floor of the mouth (lingual frenulum) ends where the root of the tongue meets the epiglottis (valecula). For retroflex articulations a separate spline was added for the floor of the mouth. A single hard palate spline was added for reference. Splines were then exported and overlaid for comparison. Figure 1 (left) shows a typical contour superimposed on the MRI image. Figure 1 also shows a comparison between ultrasound and MRI for similar segments.

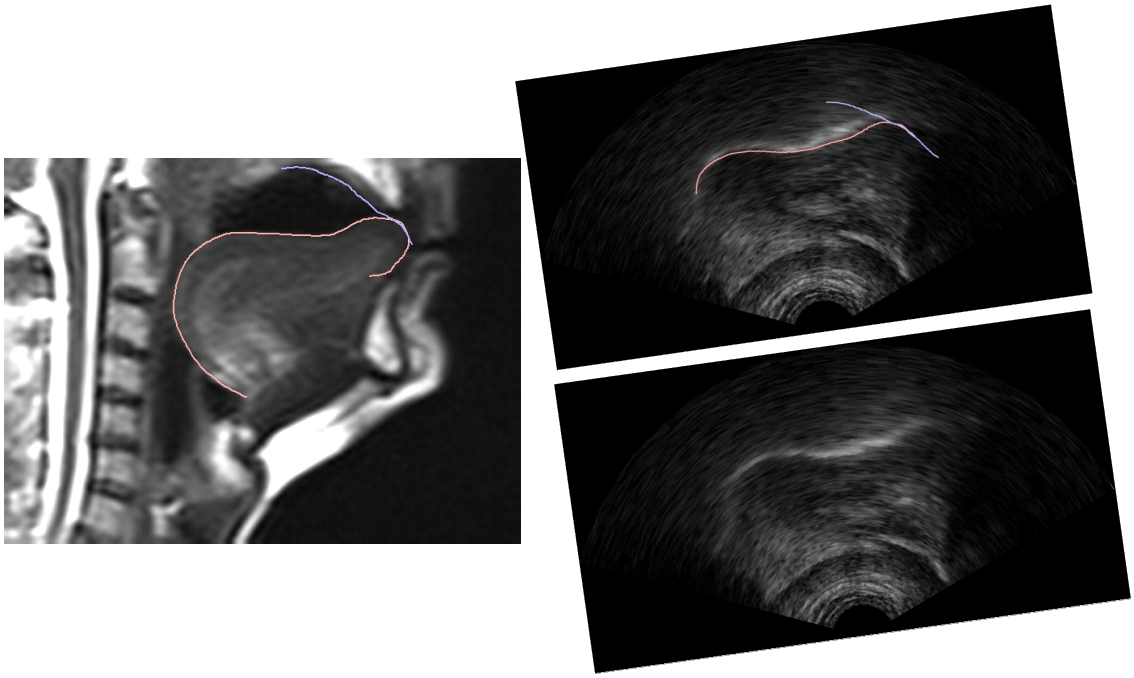

Figure 1: MRI of [ɬ] (left) with tongue and palate splines and ultrasound of [ɬ] (right) with (top) and without (bottom) tongue and palate splines

3. Results

The MRI tongue contours corresponding to tokens in Table 1, are plotted in three groups in order to facilitate comparison. Figure 2 shows sample contours for single tokens of English consonants [t,s,ɹ,ʃ,ç,k,ɬ] sustained in isolation. These exemplify normal English tongue shapes. Figures 3a and 3b exemplify the types of unfamiliar and extreme tongue contours which may be found in disordered speech (and some languages other than English), A range of vowel contexts are included to show coarticulatory variation.

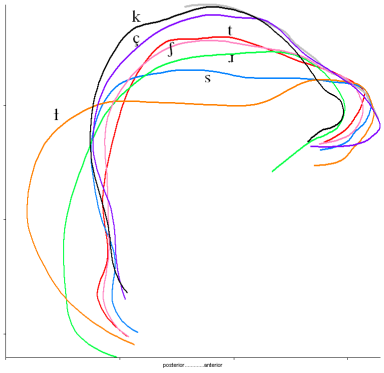

Figure 2: MRI tracings of tongue contours for selected consonants

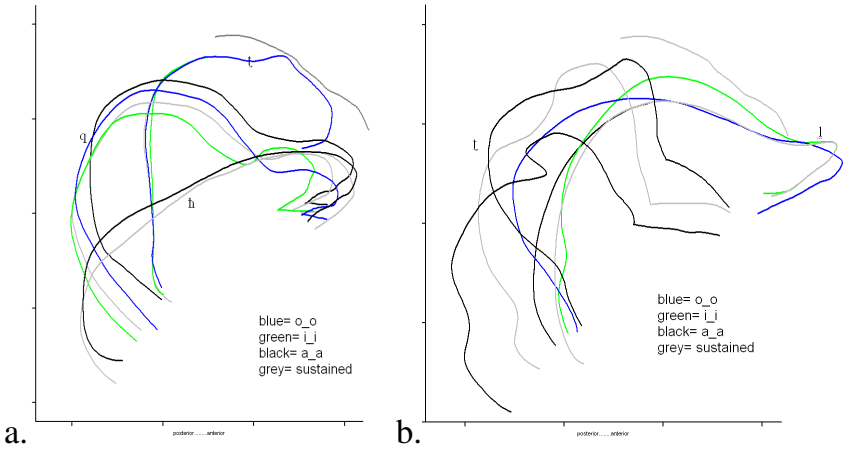

Figure 3. MRI tracings of: **a**. Non-English segments [ɢ ʈ ʜ]. **b**. ExtIPA segments (linguolablial lateral [l̼] and super-retroflex [ʈ̠]).

4. Discussion and Conclusions

Previous studies using MRI have not detailed a full range of possible tongue shapes relevant to disordered speech, choosing instead to focus on individual languages. In this pilot study we document an extended range of possible tongue shapes. This range of shapes will be used to extract the principal components that describe a midsagittal tongue contour and to validate tongue contours fitted to the corresponding ultrasound images. Comparing Figure 2 with Figure 3 clearly demonstrates that models of articulation based only on typical speakers (especially only of English) will be inadequate at capturing the range of tongue shapes which may be encountered in the Speech Therapy clinic. While we acknowledge that the use of trained phoneticians, rather than disordered speakers may under-estimate the full range of disordered tongue shapes (including in speakers with atypical anatomy), the protocol described here represents an achievable way of replicating disordered speech without having to record vast numbers of heterogeneous speakers with speech disorders.

While some segments (e.g. vowels) do image adequately with ultrasound, others

(e.g. consonants such as [t̪ , l̪]) do not. Figure 1 highlights some limitations of ultrasound: some of the root, some of the tip and the entire underside of the tongue is missing. However, by exploiting shape information derived from MRI images, we hope to improve on simple edge detection for the fitting of contours to ultrasound images with the ultimate aim of doing this in real-time. Ultrasound is cheap, non-invasive, real-time, quiet, has a high sampling rate, and can be combined with EPG and other articulatory instruments, all of which make it suitable as a speech therapy aid. We hope to retain these key advantages while enhancing the images.

Acknowledgments

This work was supported by an EPSRC grant (EP/I027696/1). Thanks to our ULTRAX project colleagues Steve Renals and Korin Richmond. Scott Semple is supported by the British Heart Foundation Centre of Research Excellence Award. Thanks to Steve Cowen for technical assistance and Annette Cooper for MRI data acquisition.

References

Articulate Instruments Ltd. *Articulate Assistant Advanced Ultrasound Module User Manual, Revision 2.12*, [manual]. Author, Edinburgh, 2010.

Bernhardt, B., Gick, B., Bacsfalvi, P., and Adler-Bock, M. Ultrasound in speech therapy with adolescents and adults. *Clinical Linguistics & Phonetics,* 19: 605-617, 2005.

Engwall, O. Assessing Magnetic Resonance Imaging Measurements: Effects of Sustenation, Gravitation, and Coarticulation. In: Harrington, J., Tabain, M., editors, *Speech Production: Models, Phonetic Processes, and Techniques*. Hove: Psychology Press, 301-314, 2006.

ICPLA (International Clinical Phonetics and Linguistics Association). *ExtIPA Symbols for Disordered Speech,* 2002.

IPA (The International Phonetic Association). *The International Phonetic Alphabet,* 2005.

Michi K-I, Yamashita Y, Imai S, Suzuki N and Yoshida H. Role of visual feedback treatment for defective /s/ sounds in patients with cleft palate. *Journal of Speech and Hearing Research*, 36: 277–285, 1993.

Lee, J. and Stone, M. *Overlaying Ultrasound to MRI Sequences*. Paper presented at Ultrafest V (March, 2010) retrieved May 19th, 2011 from http://www.haskins.yale.edu/conferences/UltrafestV/abstracts.html

Scobbie, J.M., Lawson, E., Cowen, S. Cleland, J. and Wrench, A.A. A common co-ordinate system for mid-sagittal articulatory measurement. *Proceedings of Interspeech*, *Florence,* 2011in press.

Stone, M. A Guide to Analysing Tongue Motion from Ultrasound Images. *Clinical Linguistics and Phonetics,* 19: 455-501, 2005.

Wrench, A.A., Cleland, J. and Scobbie, J.M. An Ultrasound Protocol for Comparing Tongue Contours: Upright vs. Supine. *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong*, 2011 in press.