Two Short Studies in Vocal Tract Measurements

Catherine I. Watson[1], *Chung Ting Justine Hui*[1]

[1] Department of Electrical and Computer Engineering, The University of Auckland, Auckland, New Zealand

c.watson@auckland.ac.nz ,chui018@aucklanduni.ac.nz

Abstract

Vocal tract cross-sections using acoustic reflectometry are obtained from 5 young (20-25 yrs) and 5 middle-aged New Zealand English speakers from 9 monophthongs. Formant data, from hVd words, were also collected from the speakers. Whilst the vowel spaces for the two groups were similar, the pharyngeal volumes of the middle-aged group were significantly greater than the young group. In a second study mid-saggital vocal tract cross-sections from MRI scans are presented for 10 monophthongs produced by two speakers of NZE, and are compared to their formant spaces.

Index Terms: vocal tract, acoustic reflectometry, MRI

1. Introduction

We have a long term interest in modeling the impact of aging on speech production. The fundamental acoustic model we use, when describing speech production is the source filter model [1]. We are interested in both the source (see[2]) and the filter. Our long term aim is to understand how changes to the physiology of the vocal tract and larynx of adults impact on the acoustic features of speech, in particular vowel formants.

It is possible to work out the frequency response of the vocal tract, and consequently vowel formants, if we have knowledge of its shape. Researchers have obtained the cross-sectional area by various imaging techniques including x-rays, ultra-sound, and magnetic resonance. It is also possible to deduce the area using a technique called acoustic reflectometry, whereby periodic pulse signals are transmitted through a wave tube down the vocal tract and the airway's cross sectional area is deduced from the reflected waves.

The advantage of Acoustic Reflectometry (AR) over Magnetic Resonance Images (MRI) is that for the former the data collection method is unobtrusive, low cost, highly specialized staff (and expensive) are not required to obtain the data, and it is very quick. In addition participants are in a supine position for MRI, whereas for AR they can be in any position. In [7] we compared vocal tract cross-sectional areas obtained from MRI scans and AR measures for four different vowels from a single speaker. We found there was agreement in the pharyngeal regions of the two measures in general. However in AR the mouth piece of the wave tube, meant the speaker was unable to vary the opening of their mouth. Thus the cross-sectional measures of the oral region were not representative of true speech production. We therefore concluded that Acoustic Reflectometry could be used to investigate the pharyngeal region in vowel production, but not the overall vocal tract.

As a consequence of [7] we have decided to build vocal tract models from MRI data rather than via AR. We are currently embarking on a study collecting MRI data from 10 New Zealand English speakers, some of whom we also have AR data. This paper presents the findings of two short studies that were completed as a consequence of the larger ongoing study on modeling aging in speech production. The first study investigated whether aged related differences can be seen in the pharyngeal region of 10 speakers using AR. The second study presents the first collection of MRI mid-saggital vocal tract images of the New Zealand English monophthongs.

2. Vocal Tract Analysis using Acoustic Reflectometry.

The ECCOVISION Acoustic reflectometer provides a non-invasive assessment of the cross-sectional area profile of the oral and pharyngeal spaces down to the larynx (i.e. vocal tract). The system employs acoustic reflection technique, where a periodic pulse signal is generated and transmitted through the wave tube down the vocal tract. The pulse is then partly reflected when the vocal tract changes cross-section. The reflected wave is finally used to calculate the cross sectional area of the airways[5]. The main application for AR is for measurement of the upper respiratory airways. However Xue [6] has used it to look at vocal tract shapes, when the vocal tract was at rest. He found age related differences in the pharyngeal region, but no length related effects.

We compared vowel production from five young adults (20-25yrs) and five middle aged adults (40-50yrs). These two groups are called Y and M respectively. All participants are speakers of New Zealand English. We obtained both speech recordings and AR measurements.

2.1. Speech Data

Speech was recorded in a sound booth (Whisper Room MLD8484E) directly on to a Marantz PMD670 Solid State Recorder at a sampling rate of 20 kHz, using a Shure SM58 Microphone. For each speaker we collected citation form speech of hVd words for nine monophthongs. We collected five versions of each hVd word, which were presented to the speaker in a random manner. The monophthongs /ʌ/ and /ʊ/ were not recorded. At the time these were deemed unnecessary as according to [4] these two vowels can really be considered to differ from /ɐː/ and /ɔː/ in length only. Hence the resulting vocal tracts would essentially be the same (in hindsight this was not such a good idea).

The speech data were transferred to the computer and phonetically labeled using the EMU speech database system (http://emu.sourceforge.net/). The first three formant center frequencies and their bandwidths were calculated. All formant tracks were checked, and corrections were made if necessary. For each vowel the target was manually identified. (see [4] for more details). The formant values were extracted at the vowel targets and analysed in R/EMU (http://www.r-project.org/).

2.2. AR data Acquisition.

The AR measures were obtained at the same time as the speech recordings in early 2008. Obtaining the AR measures of the vocal tract is a slightly complicated task, which takes a little time for the participants to get accustomed to. Participants place the wavetube in their mouth, place their articulators in the correction position to produce the target vowel and hold that position for two to three seconds whilst the measurement takes place. The vocal folds need to be closed during the measurement, (via a gentle Valsalva maneuver). No voicing is possible during measurement as it interfered with the sonic pulses produced in the wavetube. Thus speakers did not receive any aural feedback on the production of their vowels.

To ensure the participants were familiar with the vowels the sound recordings were made first, then the participants used the reflectometer. For each vowel type four productions were made and measured. After each measurement the vocal tract profiles were checked to ensure that the shapes were reasonable, in some cases the participant had to redo the productions. For AR data collection the participants were sitting.

2.3. Processing the Data from the AR.

To process and analyse the data from AR a special purpose research platform has been developed in R with the graphics done with a Tk plugin. The AR measures the cross-sectional area for 30 cm along the vocal tract/airways, although we are only interested in the vocal tract for this study. For each reading, the start and end of the vocal tract (lips and glottis) are automatically identified. The platform looks at the region between 15cm and 20 cm and sets the glottis to be the minimum cross sectional area in that region. If the glottis is in fact outside this range the end value can be changed. The lips by definition are set to be -0.5cm. Figure 1 is a snapshot from the vocal tract analysis platform showing the full data from one of the AR readings, and the automatically identified start and end of the vocal tract (vertical lines). In the study all start and end points of the vocal tract were hand checked, and changes were made if necessary.

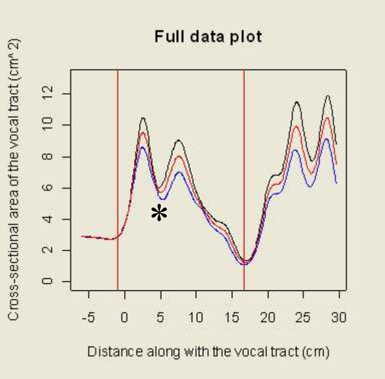

Figure 1 *A snapshot from the vocal tract analysis platform showing the full data from one of the AR readings, and the automatically identified start and end of the vocal tract (vertical lines).*

Once the start and end of the vocal tract have been verified, it is then possible to calculate the resonances of the vocal tract for that particular shape. It is also possible to then change the vocal tract shape and observe the changes in the resonance, although this functionality is not part of the current study (NB see [7] for difference between resonances and formants).

In this study our focus was on the pharyngeal volume. The volumes were calculated by integrating the cross-sectional areas. The pharyngeal cavity is the region between the velopharyngeal port and the glottis. This point was located in the data by automatically searching for the end of the first "valley" from the lips. When the vocal tract shape does not have this "valley", the point where the slope is the flattest after the first "hill" is taken as the velopharyngeal port (* marks the velopharygeal port for the data in Figure 1).

All pharyngeal volumes were normalized to enable comparisons between speakers. This was achieved by calculating the ratio of the pharyngeal volume to the total vocal tract volume and multiplying the ratio by 100.

2.4. Results

Figure 2 gives the vocal tract cross-sectional areas for one of the participants in the study. The area is plotted from the lips to the glottis. The *x* axis is the distance from the lips, and the *y* axis is the cross-sectional area. For each vowel we obtained four separate readings. Whilst there is a degree of variability in the shapes for each vowel, over all the results are quite consistent. The front vowels have large pharyngeal areas, and the back vowel small pharyngeal areas, as would be expected. Note the mouthpiece of the wave tube means the cross-sectional area at the lips is always the same.

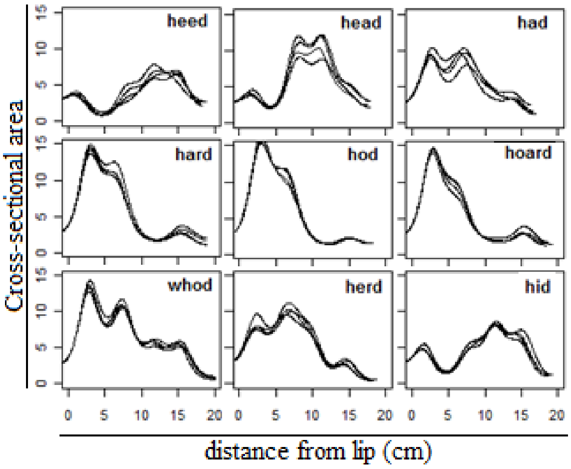

Figure 2: *AR cross-sectional vocal tract areas for nine monophthongs obtained from one speaker. The nine vowels are given as hVd keywords on the plots.*

Table 1 *Mean normalized Pharyngeal volumes expressed as percentages for the Y and M groups for each vowel, significance between the two groups is indicated by * ($p<0.05$) according to t- tests.*

	iː	ɛ*	æ	ɐː	ɒ	ɔ	uː	ɜː*	ɪ
Y	71	65	63	40	34	32	69	64	72
M	75	76	65	45	40	38	63	72	76

The normalized pharyngeal volumes were calculated (see section 2.3 for explanation how) for all speakers. Then the mean normalized volumes for each of the speaker groups were calculated. The results are given in Table 1. A two-way ANOVA was performed with speaker group and vowel type as main effects. There was a significant difference for group ($F=22.5,p=0.0$), and vowel ($F=140.3,p=0.0$), and a significant group vowel interaction ($F=2.9,p>0.01$). The mean pharyngeal volumes of the M (Middle-aged) group are greater than the Y (Young) group for all vowels except /uː/. On a vowel by vowel basis post-hoc t-tests revealed this difference was significant for /ɛ/ and /ɜː/.

Figure 3 shows the formant plots for the nine vowels from the two speaker groups for the male speakers (four speakers in each group). Two-way ANOVAs were performed on the male data, with group and vowel being the independent variable, and F1 and F2 being the dependent variables respectively. There was a significant difference for group for both F1 ($F=11.1,p<0.01$) and F2($F=62.6,p=0$, and also a significant group vowel interaction for F1 ($F=3.3,p<0.01$) and F2 ($F=8.4,p=0$). When comparing the Y with the M group post-hoc t-tests revealed F1 was significantly greater for the Y group for /æ / and F2 was significantly less for /i:,ɛ,æ,u:ɜ/. The female data behaved in a similar way, but since there was only one speaker per group, there was too few data to analyse.

2.5. Discussion

The results of the study showed that the pharyngeal volumes for the middle-aged group significantly larger than the younger group. The larger pharyngeal volume for the M group would lead to a larger F2 value, which is what we found for /i:,ɛ,æ,u:ɜ/ for the male speakers. More speakers are needed to see how robust is this finding .The pharyngeal volume finding reflects the age related differences found in [6], where elderly females had larger pharyngeal volumes than the middle aged females. Although in [8] Xue *et al* compared elderly and young adults, and found significant differences in overall volume, but not in the pharyngeal volume. In both these studies the vocal tract volume was measured for the vocal tract in a rest position, not during vowel production.

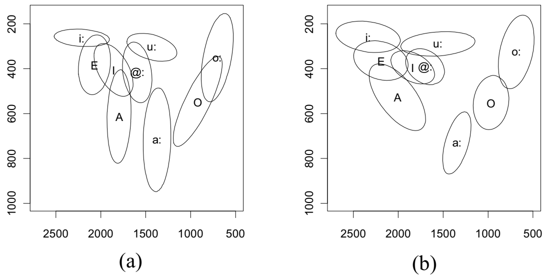

Figure 3: *The vowel spaces of (a) Group Y and (b) Group M. . For the vowel plots in this paper we have used Machine Readable Phonetic Symbols.*

While the AR analysis is yielding promising results, it is difficult to accurately identify the velopharyngeal port from the cross-sectional areas for all readings. This may mean our pharyngeal volumes may not be accurate. We have begun collecting MRI scans of the vocal tract, which. will enable us to build 3D models of the vocal tract and examine in more detail the volumes including pharyngeal volumes. In addition to the young group, and middle aged groups we also collecting data from the over 65 years age group, the aging effects are expected to be more pronounced.

3. Vocal Tract Analysis using MRI

MRI has been used in vocal tract research for over twenty years (see for example [3]). The MRI scans acquire multiple images, from which it is possible to reconstruct static 3-D vocal tract shapes. These models have not yet been built. However from a phonetic point of view, the mid-saggital scan of the vocal tract for different vowels provides an interesting insight into the positions of articulators during vowel production. It is this data we will present here.

3.1. MR image acquisition.

The MR images were acquired with a 1.5T Siemens Magnetom Avanto MRI scanner in 2010. Images have so far been obtained from eight participants, seven of who are NZE speakers. The images from two of the NZE participants are presented here. Participant 1 was from the M group and Participant 2 was from the Y group (see Section 2).The scans were done in parallel sagittal planes, each slice was 6mm, with no gaps and the images were T1-weighted. The remaining scan details differed for the two participants, and are given in Table 2. To acquire the images the participants had to sustain their pronunciation of vowels for a long period of time. In the pilot study the scan time was 24 second, however this was deemed too long and since the scan provided images of the entire head region, a lot of the images were not of use in a vocal tract study. For the remaining experiment the MRI scan was limited to saggital scans of the head only in the jaw area, and this reduced the scan time to 15 seconds. Participants were instructed to say hVd words, and hold their pronunciation of the vowel throughout the duration of the scan. For each vowel images of two separate productions were obtained.

Table 2: *MRI Scanning Parameters*

Participant	Field of View	Repetition time	Echo Time	Resolution
1	199*250	2090	9.4ms	1mm
2	211*260	1340	8.7ms	1mm

3.2. Results

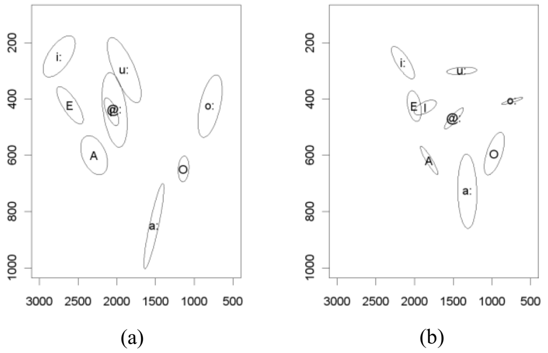

Figure 4: Vowels spaces of the two speakers from the MRI study (a) participant 1 (b) participant 2

Formant data were available for both participants, as they were also part of the AR study (see section 2.1 for details on speech analysis). Their ellipse plots in F1/F2 space can be seen in Figure 4. Both participants have the typical vowel triangle of NZE, with the point vowels being /i: ɐ: ɔ:/. Both speakers have raised / ɛ æ ɜ: ɔ:/ , the fronted / ɐ: u:/, and the retracted /I/ , typical of NZE [4].

Figure 5 shows the mid-saggital vocal tract images from the two participants. Only one image of the two productions for each vowel is shown, however for all speakers there was a lot of consistency between each of the productions. There is also quite a bit of consistency between participants vocal tract shapes for each vowel (see Figure 5), and further the tongue positions are exactly where they are expected. The tongue is at its front most position for /i:/, and its back most for / ɔ:/ and /ɒ/. The tongue body for /u:/ is closer to the front of the mouth than the back. Images were collected for /ʌ/, but as they are the same as those for /ɐ:/, they were not included in Figure 5 due to lack of space. In order to maintain long phonations, participants were instructed to not to phonate loudly. This may have impacted on the lack of jaw opening for /ɐ:/, which was not expected. The subtle changes in the vocal tract shape between /i: ɛ æ/ are interesting given the difference between these vowels are quite salient. However the differences may be more noticeable when looking at the images adjacent to the mid-saggital cross section. The large variations in the pharyngeal cavity between the front and back vowels are interesting, and the vocal tract shape for /ɐ:/ is more similar with the back vowels than the front vowels. The similarity in the vocal tract shape between /I/ and /ɜ:/ is noticeable, with both patterning more with the high front vowels. Lip rounding for NZE /ɜ:/ would only be noticeable from the images adjacent to the mid-saggital cross-section.

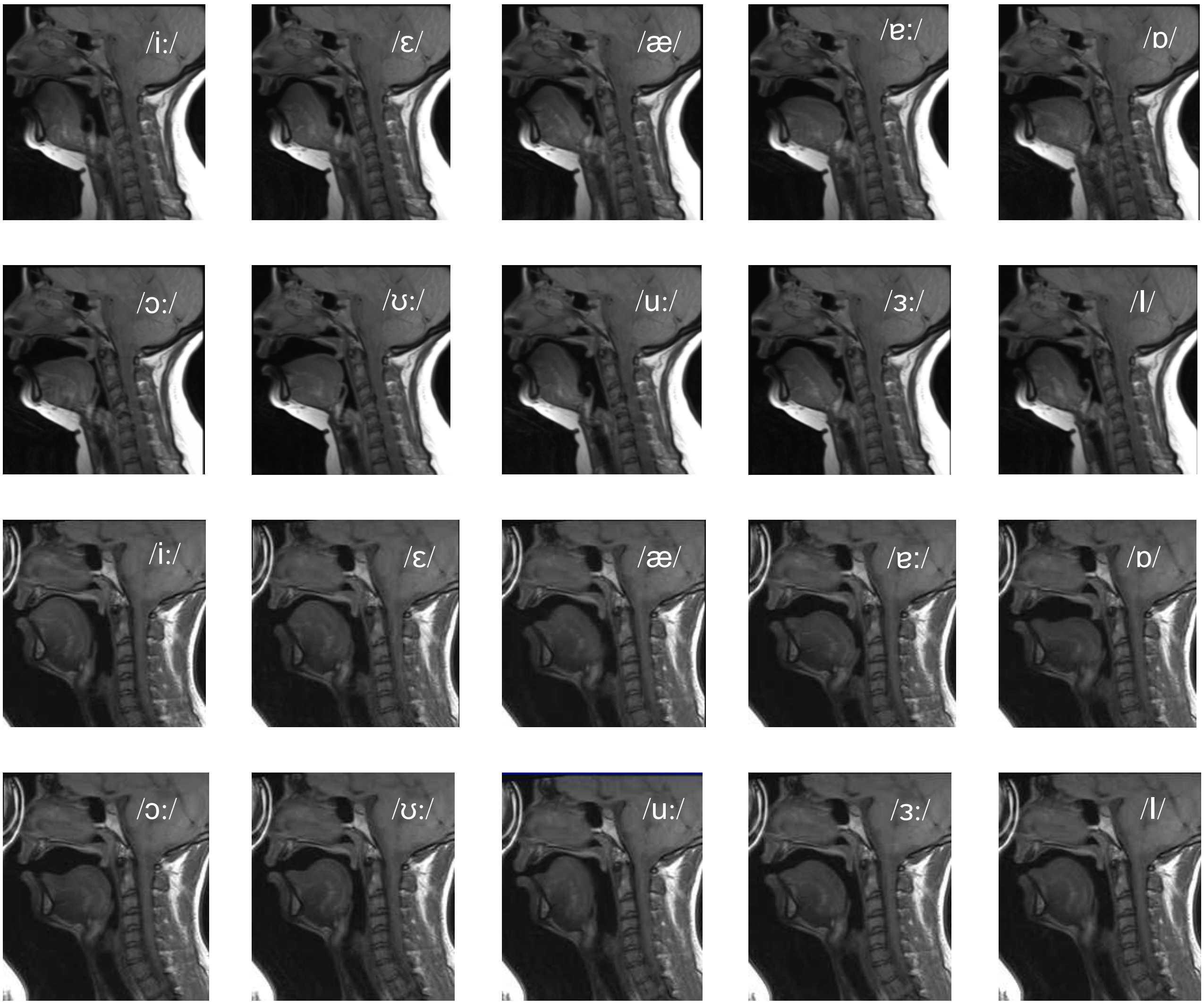

Figure 5 *MRI of Two speakers of New Zealand English, showing the mid-saggital cross-sections of the vocal tract for 10 of the 11 monophthongs for (top two row) participant 1and (bottom two rows)participant 2*

Note how the vowels with the larger pharyngeal volumes, compared to the oral volumes (i.e. the front vowels) are the ones with the higher F2 values.

4. Conclusions

Two different methods of obtaining measurements of the vocal tract have been presented. The first study obtained measures of the vocal tract cross-sectional areas using AR. The results suggest that as speakers age there is an increase of their pharyngeal volumes relative to their oral cavity volumes. The expected increase in F2 was observed in the Male data, where the F2 for the front vowels was significantly greater the middle-aged group (cf. the young group) This finding will be investigated more thoroughly by collecting data from more speakers and building more accurate 3-D models of the vocal tract from recently collected MRI data. Using a subset of this data, the second study presented the first published set of mid-saggital vocal tract cross-sections, showing the production of NZE monophthongs. The data was from two speakers.

5. Acknowledgements

This project has been supported by two University of Auckland research grants, and three Electrical and Computer Engineering studentships. The authors thank S. Yeung, J. Wang, T. Judkins and D. Eason; for contributions to the data collection and analysis platform, the anonymous participants of the MRI and AR studies and the anonymous referees.

6. References

[1] Fant, G., Acoustic Theory of Speech Production, Mouton, 1960.

[2] Bier,S.D. and Watson C.I. A platform for modeling aging in the adult vocal folds., SST 2010, Melbourne, this volume, (2010)

[3] Story, B.H. "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002 ", J. Acoust Soc Am, 123(1), 327- 335, (2008)

[4] Watson, C.I., Harrington J., and Evans Z., "An acoustic comparison between New Zealand and Australian English vowels", Australian Journal of Linguistics 18, 185–207. (1998)

[5] Sondhi M. M and. Gopinath B, "Determination of Vocal-Tract Shape from Impulse Response at the Lips", J. Acoust Soc Am, 49(6), 1867- 1873, (1971)

[6] Xue S.A.,. Jiang J,. Lin E,. Glassenberg R, and. Mueller P.B, "Age-related changes in human vocal tract configurations and the effects on speakers' vowel formant frequencies: a pilot study", Log. Phon Vocol, 24, 132-137 (1999)

[7] Watson, C.I., Thorpe, C.W., Lu, X.B. 'A Comparison Of Two Techniques That Measure Vocal Tract Shape', Acoustics Australia, 37, (1), p7-11. (2009)

[8] Xue S.A and. Hao G.P, "Changes in the Human Vocal Tract Due to Aging and the Acoustic Correlates of Speech Production: A Pilot Study." Journal of Speech, Language and Hearing research 46(3), 689-701 (2003).