A COMPARISON OF TWO TECHNIQUES THAT MEASURE VOCAL TRACT SHAPE

Catherine I. Watson[1], C. William Thorpe[2], Xiao Bo Lu[2]
[1]Department of Electrical and Computer Engineering
[2]Bioengineering Institute, The University of Auckland

This study compares vocal tract shapes of four vowels for a single speaker estimated via analysis of magnetic resonance images and acoustic reflectometer measurements. The first and second resonances of the vocal tract shapes from the two different methods are compared to the first and second formants obtained from an acoustic analysis of the speech signal. It is demonstrated that speech production is compromised by the mouthpiece used for the acoustic reflectometer, and as such this tool is not useful for studying articulatory phonetics.

INTRODUCTION

The sounds of speech are produced by movements of the speech organs and their effect on the air flow through the vocal tract. By changing the position of the articulators (i.e. tongue, jaw, lips), and the nature of the acoustic air flow through the vocal tract we produce speech. However, the speech production mechanism is hidden, and it is non-trivial to ascertain the precise configuration corresponding to different speech sounds. Acoustic phoneticians are interested in determining differences in production for various environmental effects such as accent, aging, and pathology, and utilise acoustic analysis techniques that relate specific spectral features of the speech signal to features of the vocal tract shape. However, such analysis relies on acoustic models of speech production to solve the inverse problem, and does not always result in reliable estimates of vocal tract shape.

The fundamental acoustic model of speech production is the source filter model [1], in which the acoustic energy source is separated from the time-varying filter which imparts a specific spectral shape to the speech sound. In vowel production the acoustic source consists of a quasi-periodic train of pulses of air emitted by the vibrating vocal folds, and the filter can be well modelled by small number of resonances corresponding to the resonating cavities in the vocal tract. The spectrum of a (sustained) vowel sound is therefore a line spectrum (with spacing equal to the vibration frequency of the vocal folds) with several distinct peaks. These peaks are termed formants and it has been shown that the first two or three formant frequencies collectively determine the identity of the vowel sound [1, 2]. Since the resonances (and consequently the formants) depend on the vocal tract shape, they will be affected by factors such as size, effects of aging, and manner of articulation (how the vocal organs are moved during speech). Unfortunately – especially for high pitched voices – the relatively wide spacing between the spectral lines means that there can be insufficient information to determine uniquely the centre frequency and bandwidths of the resonances and therefore a unique vocal tract shape. Thus, other methods of determining the vocal tract configuration are of interest if differences in its shape are to be investigated.

Researchers have obtained measurements of the vocal tract shape by various imaging techniques including X-rays (e.g. [1]), computer–tomography (C-T) (e.g. [3]), and magnetic resonance (MR) (e.g. [4]). The latter two approaches enable 3-D shapes of the vocal tract to be constructed through post-processing of the images. All these methods involve expensive equipment and well trained operators. With X-rays and C-T scans there is also some risk to the subjects if they are exposed to repeated measurements. For a large scale study on speech production, all the above factors make it difficult to obtain comprehensive data from a large number of subjects using these techniques.

It is also possible to deduce the vocal tract shape using a technique called acoustic reflectometry (AR) [5]. This measurement technique is used for determining the cross-sectional area of ducts. It has been adapted for diagnostic measurements of upper respiratory airways, and has been previously used in studies of the vocal tract shape (e.g. [6, 7]). The technique involves transmitting pulse-like signals through a wave tube and into the vocal tract. The pulses are partly reflected when they encounter physical obstructions or changes in the cross-section of the tract. Analysis of the reflected waves gives the impulse response of the tract, from which the cross sectional area of the tract can be calculated (see [5] and [8] for a more in depth discussion of the technique). Acoustic reflectometry is easy to perform, the equipment is cheap in comparison to the former approaches, and it has no known side effects on the subjects. This makes it a potentially ideal instrument for a large scale study relating vocal tract shapes to specific speech features.

Once we have obtained the vocal tract shape (by MR images, AR, or any other measurement technique), it is a routine process to calculate the vocal tract resonances corresponding to that shape [9]. We can therefore compare the shapes obtained by different measurement techniques with the spectral patterns expected for different speech sounds (as determined by direct measurement of the acoustic output).

Because of their geometrical accuracy, MR image studies are the "gold standard" for determining physiological

structure, but as mentioned the cost is prohibitive when considering a large scale study. AR is appealing to use in a large scale study due to the low costs associated with collecting the data. However, questions arise as to how it compares to the MR image approach; does it give accurate enough information about the vocal tract shape; and can we assess the effects of the speech produced by such a shape? To date there has been no acoustic phonetic study done comparing vocal tract shapes calculated via MR images and AR. The purpose of this study is to do that comparison, and also to contrast the vocal tract resonances calculated from these shapes to formants from recorded speech.

METHOD

The study involves three different data sets collected from a single male speaker of New Zealand English (NZE). With this speaker we did an analysis of 3-D MR images of the vocal tract, an analysis of the cross sectional area of the vocal tract obtained from AR and an acoustic analysis of the speech. Four vowels were studied /iː, aː, ɔː, ɜː/, the NZE vowels in the words "heed", "hard", "hoard" and "heard" respectively.

2.1 Magnetic Resonance Imaging Analysis

The MR images were acquired with a 1.5T Siemens Magnetom Avanto MRI scanner. The scanning parameters were: T1-weighted image; parallel sagittal planes; 7 mm slice thickness; no gaps between slices; 200x250 mm field of view; 1660 ms repetition time; 9.4 ms echo time; 1 mm resolution; 20 slices and a total scanning time of 21 sec. MR images were collected in a supine position with the head supported to prevent movement. Images were obtained for all four vowels. Since the MR images produce a three dimensional image of a single vocal tract shape, the subject had to maintain a sustained production of each vowel for the entire scan time of 21 seconds. For this reason these vowels were pronounced in isolation, rather within a word context. The subject's background in voice science meant he was able to ensure the articulator positions were appropriate for each vowel. The MR images were stored on the computer and can be viewed as DICOM images.

To determine the vocal tract area, cross-sections of the vocal tract were obtained at 15 points along the mid-line of the vocal tract. All image processing was performed using the CMGUI image processing and analysis software (http://www.cmiss.org/cmgui). A centre-line was constructed through the visible vocal tract on the mid-sagittal plane. Next, a smooth line was fitted through these points with a cubic Hermite spline having 15 equally spaced nodes. At each node a plane was constructed perpendicular to the centre-line, as illustrated in Figure 1 (left), and the 3-D image stack resampled onto each plane, thereby producing a sequence of images that cut the vocal tract perpendicularly throughout its length. The boundary of the vocal tract was then manually marked on each of the planes (see Figure 1(right)). Finally, a smoothing spline was fit to these data and the internal area computed at each of the planes. This sequence of measurements forms a 1-D vocal tract area function. This results in a discrete model of the vocal tract, approximating the varying area of the tract by a series of concatenated tubes of uniform thickness and varying cross-sectional area. From this the resonant frequencies were calculated from custom functions based on the standard linear prediction model of speech (e.g. [9]). All functions were implemented in R (http://www.r-project.org/).

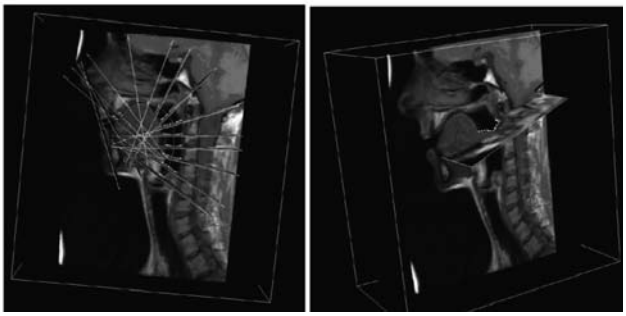

Figure 1: Illustration of how the cross section areas of the vocal tract were determined from the MR images: Slices computed perpendicular to the vocal-tract midline (left), and the cross-sectional area is obtained by manually locating points on the edge of the vocal tract cavity on each slice (right).

2.2 Acoustic Reflectometery

The AR vocal tract profiles were acquired by the ECCOVISION Acoustic reflectometer. It provides a non-invasive assessment of the cross-sectional area profile of the oral and pharyngeal spaces down to the larynx. Subjects place the wavetube in their mouth, position their articulators for the target vowel, and hold that position for two to three seconds whilst a series of sonic pulses are sent down the vocal tract and measurement takes place. The subjects are required to seal their lips tightly around the mouthpiece to prevent acoustic leaks of the sonic pulses. In addition the vocal folds need to be closed during the measurement (achieved by gently blocking expiratory airflow). Vowels can not be voiced during the measurement since the glottal excitation interferes with the measurement pulses, and therefore subjects do not receive any aural feedback on the production of their vowels.

To aid the subject to get the correct tongue placement for the vowels (the jaw placement was compromised due to the wavetube) we collected the data in a specific way. Firstly, we collected speech recordings immediately before the AR data (see section 2.3 for more details about this process). Secondly, although only four vowels were investigated in this study, we collected AR data for nine of the eleven monophthongs in New Zealand English. Two of the vowels, /ʌ/ (as in "hud") and /ʊ/ (as in "hood"), were excluded on the grounds that they have been shown to differ primarily in duration, but not in quality, with /a:/ and /ɔ:/ respectively [10]. The vowel order the data were collected was also important - each consecutive vowel was both an articulatory and acoustic neighbour, e.g. /e/ (as in "head") was recorded after /i:/,and /æ/ (as in "had") was recorded after /e/. Thirdly, a series of four separate vocal tract measurements were obtained for each vowel, and after each measurement, the vocal tract profiles were checked to ensure consistency and

a clear glottal closure. Any flawed data were rejected, and the measurements were retaken. All the measurements were done by a trained research assistant.

A collection of custom functions have been developed in R which allow the data from the reflectometer to be visualised and processed. Using this software, the start and end of the vocal tract (i.e. lips and glottis) were manually identified, and resonances calculated from the resulting vocal tract shape using the same algorithms as for the MRI data (See section 2.1). For the AR data the vocal tract was subdivided into 11 tube segments of uniform length.

2.3 Formant analysis of Speech

We recorded the subject's speech in an acoustically isolated 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. We collected citation form speech of nine words "heed", "head", "had", "hard", "hod", "hoard", "who'd, "herd" and "hid", four of which were used in this study. Five tokens of each vowel were recorded, with the order of the vowels randomised within each repetition. The speech data were transferred to the computer and the vowel portions of each word phonetically labelled using the EMU speech database system (http://emu.sourceforge.net/). The first three formant centre frequencies and their bandwidths were calculated (the settings were 12^{th} order linear predictive coding analysis, cosine window, 49-ms frame size, and 5-ms frame shift). All formant tracks were visually checked, and tracking errors were corrected. For each vowel, the target was manually identified. The acoustic vowel target is presumed to be the section of the vowel that is least influenced by phonetic context effects. The criterion for identifying the vowel targets varies between the different vowels (see [10] for more details). The formant values were extracted at the vowel targets and analysed in R/EMU.

RESULTS

The mid-sagittal images of the vocal tract from the subject when producing sustained productions of the three NZE point vowels /i:/, /a:/, and /ɔ:/, and the central vowel /ɜ:/ can be seen in Figure 2. The vocal tract is the black region which is bounded by the lips and the vocal folds. Note the markedly different dimensions of the vocal tract for each vowel configuration. For each of the point vowels, the tongue tip, jaw opening, and tongue body respectively are essentially at their articulation extremities. For /i:/ the greatest point of narrowing in the vocal tract is at the hard palate; the tongue surface is close to the roof of the mouth, as far forward as the alveolar ridge; and the jaw opening is very small. For /a:/ the jaw is at its most open and the tongue body is further back in the vocal tract than /i:/. For /ɔ:/ the tongue body is even further back than for /a:/ with the constriction location (greatest point of narrowing due to the tongue) at the pharynx; and the jaw opening is similar to that for the production of /i:/. For the central vowel /ɜ:/ there is an almost uniform cross-sectional area along the length of the vocal tract, unlike for the other three vowels where there are distinct wide and narrow sections of the tract.

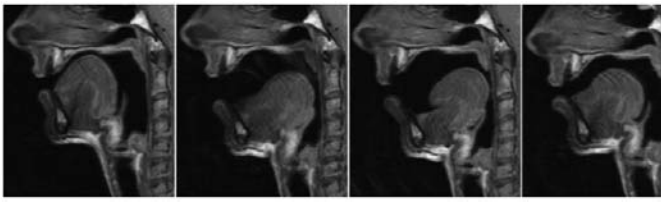

Figure 2. Mid sagittal MRI images of a male speaker doing a sustained production of: (left to right) /i:/, /a:/, /ɔ:/ and /ɜ:/ vowels.

Figure 3 shows the cross-sectional areas of the vocal tract for the four vowels /i:, a:, ɔ:, ɜ:/ obtained from AR (top four plots), and the MR images (bottom four plots). The cross-sectional areas obtained from the MR data for the four vowels are consistent with the mid-sagittal MR images in Figure 3. As expected, where there is a small constriction in the vocal tract in Figure 3, there is a corresponding small cross-sectional area, and where the vocal tract is wide there is a large cross-sectional area.

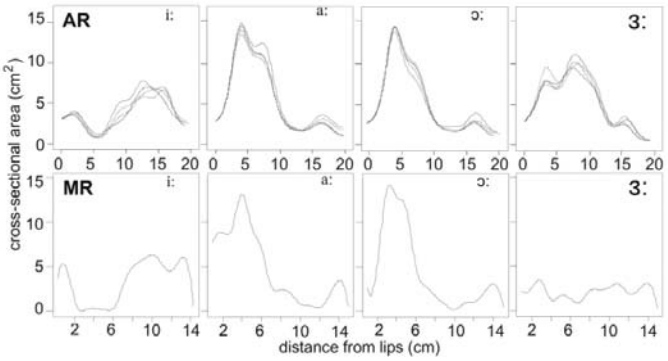

Figure 3: The cross-sectional area function of the vocal tracts for the four vowels indicated at the top, obtained from the AR (upper) and the MR (lower) data.

We obtained four readings for each vowel from AR, and all four readings have been plotted. The consistency in the vocal tract shape across repetitions is noteworthy, considering that the vowels could not be voiced during the measurement of the vocal tract. The cross-sectional area shapes obtained from the both the MR images and AR are similar for /i:, a:, ɔ:/ except around the lips. This is because the wavetube used in AR fixes the jaw position to the width of the wavetube, whereas the subject is free to move their jaw to any position for the MR measurements. There was a difference in the area functions for /ɜ:/, with the shape determined from the AR analysis having an unexpected large cavity in the front of the mouth. It may be that the fixed placement of the lips for the AR measurements interfered with the ability to correctly place the articulators, although more data from other subjects using AR will need to be analysed before that can be determined.

The two methods differed substantially in the measurement of the vocal tract length. For the MRI data the vocal tract length varied between 16.2 cm (for i: and ɜ:) and 17.8 cm (for a:). For the AR data the lengths varied between 19 cm for /i:/ and 20.6 cm for /ɔ:/.

Figure 4(a) plots the mean values of the first and second formants (F1 and F2) of the four vowels on a traditional F1 vs. F2 plot. The formant values for our speaker are typical for an NZE speaker (c.f. [10]). F1 was lowest for /i:/, and highest for /a:/. F2 is lowest for /ɔ:/ and highest for /i:/. The /ɜ:/ vowel (and the remaining NZE monophthongs) falls within the space between the vowels /i:, a:, ɔ:/.

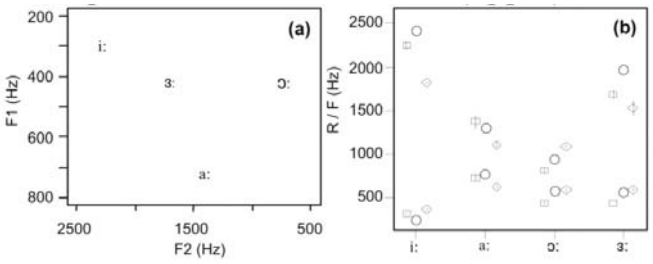

Figure 4: (a) The mean formant frequencies of the four vowels in the study on an F1 vs. F2 plot (note how the scales of the F1 and F2 axis have been reversed, thus enabling a direct comparison between the acoustic and articulatory spaces). (b) For the four vowels in the study, the mean first and second formants (□), the derived first and second resonances from the MRI data (○) and AR data (◊). The vertical lines indicate the standard deviation of the multiple measurements.

Figure 4(b) contrasts the mean F1 and F2 of the recorded speech tokens with the first and second vocal tract resonances (R1 and R2) calculated from the MR data and AR data for each of the four vowels. The AR resonance values are also means of the four measurements, but the MR resonances were calculated from the single image for each vowel. For /i:, a:, ɔ:/ there is a good match between R1 and R2 from the MR method and the first two formants. However the values are more extreme. For /ɜ:/ however the R2 value is much higher than expected. The AR data follow similar patterns to the formants for /i:/ and /a:/, i.e. /i:/ has the lowest R1, and highest R2, and /a:/ has the highest R1, however there is an issue with R1 and R2 values for /ɜ:/ and /ɔ:/. For both vowels, the R1 and R2 values are higher than expected. Also the R1 and R2 values for /a:/ and /ɔ:/ are very similar

DISCUSSION

The necessity to form a seal with the lips on the wavetube in AR means that the vocal tract shape is necessarily compromised for almost all speech sounds. For example /a:/ and /ɔ:/ differ mainly in jaw opening (see the images in Figure 2), which has the effect of changing both F1 and F2 for the two vowels substantially (see Figure 4(a)). However, the use of the wavetube fixes the jaw position and thereby removes this point of difference between the vowels – consequently in the AR data the derived vocal tract shapes for these two vowels are similar (see Figure 3), as are the R1 and R2 values (see Figure 4(b)). The fixed jaw opening imposed by the wavetube is also the reason why the R1 and R2 ranges for the AR measurements are much more constrained than for the MR data.

There is some suggestion in the data that the pharyngeal portion of the vocal tract is reasonably comparable between the AR and MR derived vocal tract shapes, at least for three of the four vowels. However the difference in the vocal tract length measurement between the two techniques is of some concern. There are a number of possible reasons for this difference. Firstly it may have been due to the position the subject was in whilst the measurements were taken. The AR data were collected whilst the subject was sitting holding the wave tube whereas the MR data were collected whilst the subject was lying supine. We subsequently repeated the AR measurements on these four vowels in the supine position. The vocal tract lengths for /i:/ and /a:/ remained about the same but for /ɜ:/and /ɔ:/ the mean lengths were reduced 1cm and 2 cm respectively. This is possibly due to a raised larynx in the supine position, however the change did not account for all the differences between the AR and MR derived vocal tract lengths.

In a previous comparison between X-ray data and MRI data [4], it was found that the MR data tended to underestimate the vocal tract length. The underestimation was attributed to the post processing method used to obtain the vocal tract shape. However we used a different method to get the shape, so it is unlikely this is the reason. Another consideration is that with the MR imaging, the determination of the precise vocal tract end-point at the lips is difficult because the opening at the lips is not a plane but is curved with some parts of the boundary being effectively longer than others, depending on the vowel (i.e. for /i:/: the corners of the mouth are retracted relative to the front). Whilst this may be a factor, it is important to note that the MRI analysis yielded vocal tract lengths in the expected region of 17 cm [1], whereas the AR derived lengths were longer than expected. In another study of vocal tract lengths measured using the AR technique [7], the vocal tract lengths were also around 17 cm. In that study however the subjects had the vocal tract in a rest position, not a speech like shape.

For all vowels, R1 and R2 from the MR data matched the F1 and F2 from the speech data much better than R1 and R2 from the AR data. For AR, the necessity to form a seal around the wavetube compromised the subject's ability to put his articulators in the appropriate position to say the vowel. But it is also notable that for both the AR and MR data the R1 and R2 values for /ɜ:/ and /ɔ:/ were higher than expected when looking at the overall vowel space (e.g. Figure 4(a)). Both vowels are produced with rounded lips (the lips are protruded and puckered). Epps and colleagues [11] measured R1 and R2 values for Australian English monophthong vowels using a different technique, and also found that the R1 and R2 values for the lip rounded vowels (such as /ɔ:/, /ʊ/)[1] were higher than the would be expected from acoustic formants of Australian English monophthongs (e.g. see [10]). A possible explanation is that the lip rounding is affecting the (acoustic) formants by some mechanism (such as a radiation effect) that is not part of the actual vocal tract resonance.

[1]but note /ɜ:/ is not produced with lip rounding in Australian English

CONCLUSIONS

We have compared two measurement methods which enable us to study the vocal tract shape, and compared the resonances obtained from these shapes with formants obtained from acoustic speech recordings. As expected there is reasonable agreement with the first and second formants and the first and second resonances from the vocal tracts measured from the MR data. There was also considerable agreement between the MRI and AR data for the vocal tract shapes of the three of the four vowel studies in the pharyngeal region. However the calculated vocal tract resonances from the AR data are not able to be compared meaningfully to the formant data from recorded speech. The wavetube used in the AR technique appears to compromise the speaker's ability to produce a meaningful vocal tract shape for the vowels where the mouth opening does not closely match the wavetube size. Further, the inability to vocalise whilst the measurement is being taken is also a methodological difficulty. Whilst the data were collected in a very specific manner, which created an optimal environment to get the correct articulator placement, the above two factors mean that speech production data across all vowels can not be collected using the AR technique. Consequently, it seems that acoustic reflectometry has limited used as an articulatory phonetic tool.

REFERENCES

[1] G. Fant, *The Acoustic Theory of Speech Production*, Mouton, The Hague (1960)

[2] M. Joos, "Acoustic phonetics" *Language* **24**, 1–136 (1948)

[3] J. Sundberg, C. Johansson, H. Wilbrand, and C. Yteerbergh, "From sagittal distance to area: a study of transverse vocal tract cross-sectional area" *Phonetica* **44**, 76-90 (1987)

[4] T. Baer, J.C. Gore, L.C. Gracco, and P.W. Nye, "Analysis of Vocal Tract Shape and Dimension Using MRI: Vowels", *J Acoust Soc. Am.* **90**(1), 799-828 (1991)

[5] J.J. Fredberg, M.E. Wohl, G.M. Glass and H.L Dorkin, "Airway area by acoustic reflections measured at the mouth", *Journal of Applied Physiology*, **48**(5), 749-758 (1980)

[6] S.A. Xue, J. Jiang, E. Lin, R. Glassenberg, and P.B. Mueller, "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] S.A. Xue and G.P. Hao, "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)

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

[9] J.D. Markel and A.H. Gray, Jr, *Linear Prediction of Speech*, Springer-Verlag, New York (1976)

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

[11] J. Epps, J.R. Smith and J. Wolfe ''A novel instrument to measure acoustic resonances of the vocal tract during speech'', *Measurement Science and Technology* **8**, 1112-1121 (1997)