THE UNIVERSITY
OF AUCKLAND
FACULTY OF ENGINEERING

ENGSCI 700A/B

DEPARTMENT OF ENGINEERING SCIENCE

Demographic trends within vocal tract area functions and vocal tract resonances for multiple speakers

Author:
Jenny Sahng

Supervisors:
Dr Catherine Watson
Dr Richard Clarke

September, 2016

Abstract

Speech technology is a growing field in our increasingly multicultural, multilingual society. Such tools require accurate models of speech production, which are developed using data of the vocal tract's anatomy and acoustics. Previous work by Dr Catherine Watson explored an alternative method of gathering vowel data through the use of MRI images. Area functions (measures of cross-sectional area by distance from the lips to the glottis) were extracted from MRI images of the vocal tract from multiple speakers. Using principal component analysis and linear predictive coding, two alternative measures of vowel height and backness were found which showed strong correlations with speech recording formants.

This project extended these methods by performing a variety of statistical analyses and validation tests on a larger set of MRI-derived area functions, consisting of 18 speakers of different age, gender and accent. From these area functions, the two most significant principal components and calculated resonances were found to account for the majority of the variance in the data, and also appeared to map to vowel height and backness, validating the results found by Watson. Strong inter-speaker correlations between these values also showed that useful phonetic information can be extracted from MRI images, and qualitative demographic trends which reflected known vowel behaviours were observed in the combined data sets. In addition, improvements were made to the image processing pipeline, a set of R scripts and functions were written for performing these analyses, and several MRI image sets were processed into area functions.

These results show that this methodology is able to extract meaningful, quantitative measurements of vowel quality and vocal tract shape from MRI images. By allowing speakers of different demographics to be analysed together, groups trends can be revealed while individual speaker differences are minimised. Exploring and understanding these variations in the underlying physio-acoustic system of human speech is the key to unlocking more accurate and intuitive speech technologies.

Acknowledgements

First and foremost, I would like to express my gratitude towards my wonderful supervisors **Dr Catherine Watson** and **Dr Richard Clarke** for their endless support. I truly appreciated the calm but firm guidance you both gave me as I navigated this learning experience. Your patience, expertise and good humour definitely defined this project, and I felt like I was in good hands from the first day. Also, the philosophical tangents during our meetings were always very enjoyable.

Catherine, thank you for your introducing me to the field of acoustic phonetics, a wonderful bridge between linguistics and computer science. Your previous work and experience in the field formed key foundations for my project, and the way you shared your passion for the research was infectious. To Richard, your advice on how to best communicate this research in the context of a Part IV Project was truly invaluable. I found your reassuring yet realistic perspective of these projects very useful as I prepared the products of my research. Above all, thank you for your kind guidance and patience as I worked my way through this research process.

A big thanks to **Xiao Bo Lu** and **Peter Bier** who, alongside Dr Catherine Watson, helped to develop the method and write the Perl, MATLAB and CMGUI scripts used in this project for extracting phonetic information from MRI images. This image processing was the foundation of all the analyses in this project. I would also like to thank **Daniel Tan**, **Kalyan Chilukuri** and **Helen Searle** whose previous undergraduate and postgraduate research work contributed to the analysis of this data set and the refinement of this system of MRI image processing. I also thank Dr Catherine Watson again for her work in gathering this MRI data set, under the ethics approval granted by The University of Auckland Human Participants Ethics Committee (Reference number 2010/136, granted 12/05/2010).

I would also like to thank **Dr Richard Christie** and doctoral candidate **Stephen Waite** for assisting me in using CMGUI during the initial data collection phase of this project.

Lastly, I would like to thank **The Department of Engineering Science** and all its members over the last five years. You truly make this department the melting pot of friendship and learning that it is.

Contents

Notations & Definitions

Abbreviations

VT	vocal tract
NZE	New Zealand English
GenAm	General American English
PC	Principal Component
R1, R2...	Resonance 1, Resonance 2...

11 Monophthongs of English

Table 1: The 12 New Zealand English speakers of this study were asked to articulate and to hold /hVd/ words during MRI imaging, where ‘V’ was one of the 11 NZE monophthongs.

Vowel	**/hVd/ word**
iː	heed
ɪ	hid
e	head
ɜː	herd
æ	had
ɑː	hard
ɒ	hod
ʌ	hud
ɔː	hoard
ʊ	hood
uː	who’d

Chapter 1

Introduction

> Language is the most massive and inclusive art we know, a mountainous and anonymous work of unconscious generations.
>
> *Edward Sapir*

1.1 Speech Technology

Speech is bafflingly complex. *Speech production* requires fine neurological and motor processes acting in concert, exploiting various physiological structures to transform airflow into precise sound waves. And yet, every human being is capable of producing fluent streams of mutually understandable utterances. *Speech recognition* is equally astounding. Every day, we navigate endless variables such as tone, pitch, accent, and context with subconscious ease. We are able to easily understand a friend in a room full of noisy conversations; we instantly recognise speech in a language we know; we generally have no trouble understanding someone with a stammer or a lisp. Humans possess a formidable knack for producing and understanding these subtle, precise sound bites.

Unfortunately for computers, it isn't so easy. However, speech signal processing and the voice science is a thriving area of research today in response to the growing role of speech technology in our world [20, 29]. Improving speech recognition allows us to better utilise the vast corpus of digitised human language available to us via the Internet, which in turn provides useful data for improving speech production models for analysis and synthesis. The applications are all around us, from speech aids for the voice- or hearing-impaired; virtual assistants like Siri or a GPS unit; even in forensic speech science where voice recognition is used as evidence in court or to build suspect profiles.

This project focuses on quantifying age, gender and accent trends in vowel sounds by analysing

MRI images of the *vocal tract* (VT)[1]. This method of extracting meaningful acoustic and phonetic data from anatomical MRI images provides a way of bridging the gap between anatomy and acoustics when building speech models. A speech recognition or synthesis system which accounts for such demographic factors is likely to produce more accurate, realistic results.

1.2 Vowels, Resonances and Formants

1.2.1 The acoustics and phonetics of vowels

Unlike consonants, vowels are speech sounds which resonate at a certain frequency, with air vibrating through the vocal tract in a uniform, laminar manner[2]. Vowel analysis has traditionally focused on the *formants*[3] from speech recordings, since these frequency peaks are the key variables which distinguish one vowel from another during speech. Alternatively, the resonant frequencies of our vocal tract (i.e. mouth and throat) may be estimated from the vocal tract profile measured with acoustic pulse reflectometry (APR), but the device used in this method requires the lips to be in a fixed, unnatural position which affects the length of the vocal tract, and therefore its acoustic properties [11, 44]. More recently, vocal tract shapes have been obtained with imagine techniques such as X-ray, computed tomography and magnetic resonance imaging (MRI), capturing images of a participant during *phonation*[4]. Of these, MRI is considered the safest as there is no risk to the participants from being over-exposed to ionising radiation [44].

Vowel sounds are made by positioning our tongue in precise positions around our vocal tract. The shape of the space created in our mouth by our tongue's contour when uttering a particular vowel is called the *vocal tract shape* of the vowel. It is made up of two parameters: phonetic height and phonetic backness, that is, how high or low and front or back your tongue is when articulating the vowel. This shape can be recorded by imaging sections of the vocal tract during speech production, and plotting the cross-sectional area of the air space between the tongue and the roof of the mouth, against the distance of the slice from the lips. We call this plot the *area function* - the cross-sectional area of the vocal tract at a given distance from the lips.

1.2.2 Resonances and Formants

From these area functions, the *resonances* (or *resonant frequencies*) of the vocal tract can be estimated by applying the technique of linear predictive coding (LPC) [22, 12] to a linear speech

[1] In articulatory phonetics, the vocal tract is essentially the mouth and throat, from the lips to the glottis where the vocal folds are situated.

[2] The most common consonants tend to be characterised by turbulent airflow and the indistinct spread of energy across the frequency spectrum as opposed to clearly defined resonant frequencies.

[3] These are the resonant frequency peaks in a natural speech recording which has been transformed into the spectral/Fourier domain.

[4] The utterance of a speech sound.

production model [8, 22] (also known as the acoustic tube model [4, 39]). LPC is a time-domain formant extraction technique in speech signal processing, similar to a fast Fourier transform but resulting in a smoother frequency spectrum which allows more precise identification of peaks on the frequency spectrum [21]. However, this process can be reversed to obtain frequency peaks from a given area function [3, 39, 40, 21]. A summary of the mathematical justification behind how the acoustic tube model and LPC work together to estimate resonances from area functions is provided in the Methods section (Section 2.4).

Once the resonant frequencies are calculated, we may compare these against the *formant* frequencies that we hear in a recorded speech signal of the same vowel. The main difference between these two terms is that resonances refer only to the resonant frequency peaks resulting from a pure periodic input (which approximates a vowel sound, made with vibrations of the vocal folds) being transformed by shape of the vocal tract (vocal tract filter). Formants, on the other hand, are the result of the combined effect of not only the vocal tract filter, but also the glottal filter (the effect of one's vocal folds on the shape of the waveform produced, which will be an impulse train as the vocal folds open and close [30]) and lip radiation (whether the lips are rounded, as in 'who', or spread, as in 'heed'). Therefore, though similar, formants and resonances are conceptually and physically different measures. Furthermore, resonances are a more direct reflection of vocal tract shape.

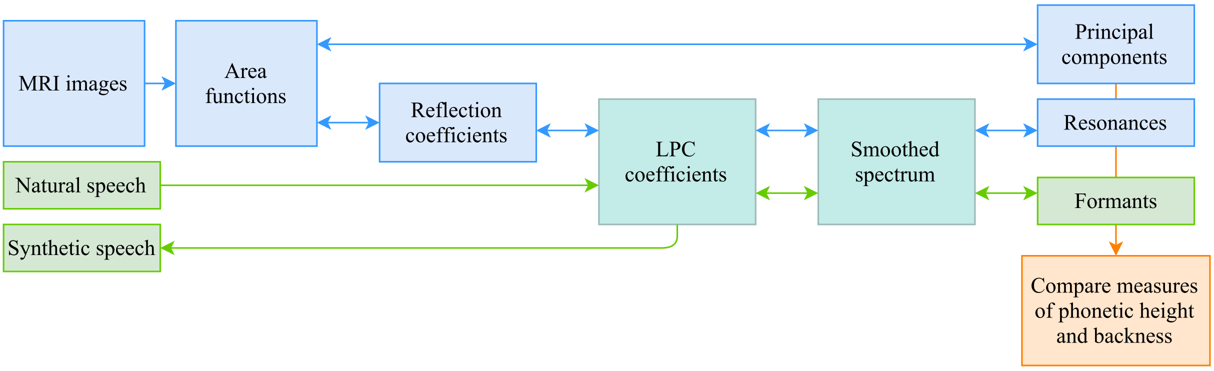


Figure 1.1: Relationships between different measures and representations of vowel quality (adapted from [12, p. 240])

1.2.3 The Vowel Quadrilateral

The two main chambers of the vocal tract are the *oral cavity* (the space above and before the highest point of the tongue, before the uvula) and the *pharyngeal cavity* (the space between the tongue and the vertical back wall of the mouth, from the uvula to the glottis). As these cavities change size depending vocal tract shape (i.e. tongue position)[5], the resonances of vocal tract change accordingly, as do the formants.

There is a well-established correlation between the first and second formants, and the first and

[5] The smaller a space, the higher the resonant frequencies would be and the opposite for a larger space.

second resonances, with phonetic height and backness, respectively. Since vowels are distinguished by their specific height and backness, formants are therefore a widely accepted representation of a vowel, commonly used in analyses of vowel change and accent comparison. When the the first and second formants (F1 and F2) are plotted on the y and x axes, we obtain the vowel quadrilateral, a representation of all the vowel positions in the vocal tract (Figure 1.2). New Zealand English has 11 of these simple vowels, which are called *monophthongs*.

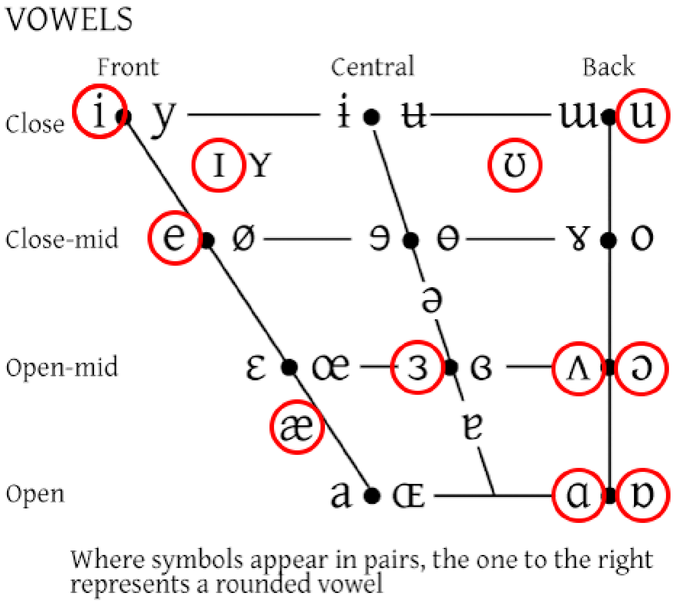


Figure 1.2: Vowel quadrilateral showing positions of all the vowels in the International Phonetic Alphabet, with the 11 NZE monophthongs circles [2]

1.2.4 Effects of age, gender and accent

The effect of age on vowel sounds is complex, given that vowels are so reliant on the resonant frequencies, and therefore the anatomy of the mouth. As we age, the tissues in the mouth relax, the mucous membranes dry and thin, and fine motor control may become compromised [1]. Vocal chords also undergo atrophy, resulting in changes in pitch, volume, and importantly, resonance. Though these changes are likely to be subtle, it will be interesting to see if they are significant enough to be picked up through an analysis of the vowel spaces.

On the other hand, the effect of gender on vowel sounds is relatively simple. It is well established that human male voices are on average twice as deep as those of females, despite males only being 10% taller and 20% heavier on average [9]. This disproportionate decrease in resonant frequency is likely to originate from the increased thickness and length of male vocal chords due to testosterone, as well as the slightly larger size of the vocal cavities. In terms of the vowel space, there should be no relative shift between the vowels purely due to the speaker's gender when all else is kept constant, as the difference is simply a drop in the resonant frequencies of their vocal cavities and vocal chords.

Lastly, we expect comparisons between vowel spaces across different accents to produce some interesting results, as accents are largely defined by the quality of their vowels[6], and a comparison of vowel spaces may be a way of quantifying these differences.

1.3 Previous Research

Previously, Dr Catherine Watson had gathered MRI images of the vocal tract during phonation for 12 speakers of New Zealand English (NZE). Area functions (cross-sectional area of air space in mouth, plotted against distance) for 5 of these speakers (2 sets each, see Appendix A for details) were extracted from MRI images and transformed using principal component analysis (explained in Section 2.3). Linear predictive coding was used to calculate resonances from the area functions, and the mappings between vocal tract area functions, vocal tract resonances and speech formants were described [41]. The concept of analysing these area functions as a combined data set, rather than treating each speaker as an individual case study, was new to the field. The study showed that 1) area functions from MRI images are a valid source of phonetic data and 2) that combining such data from different speakers ignores individual speaker differences (such as tone, volume or pitch of someone's voice), potentially allowing the analysis of group trends (such as the impact of aging, gender, accent) on the bulk data.

1.3.1 PCA on VT Shape

When principal component analysis was performed on the area functions for all vowels and speakers combined, Watson found that the first two principal components (PCs) accounted for an average 78% of the variance for each speaker, and appeared to correspond to backness and height respectively when the vowels were plotted on PC2-PC1 planes. Watson also found very strong, significant correlations between the first two PCs of each speaker treated individually, with mean absolute values of correlation of 0.9 and 0.76 respectively. Figure H.1 shows the same vowels from all the speakers clustered together on the PC2-PC1 plane (each speaker's vowels are in different colours). From this, we can say that the variation accounted for by the first two PCs is due to difference in vowel quality, namely phonetic height and backness, instead of individual speaker variation which Watson suspected may be encoded in the higher order PCs.

1.3.2 Resonances from VT Shape

The first four resonance frequencies were then estimated from the area functions using Linear Predictive Coding, and R1-R2 plots using the first and second resonances were made. The Bark

[6] As well as differences in consonants

frequency scale[7] was used to normalise the frequency axes[8]. Lobanov normalisation[9] was also applied to data for each speaker as further neutralisation of individual speaker differences was required. From the R1/R2 plot (Figure H.2), we saw that the vowel distribution was very similar to that of the PCA plots. R1 and R2 appeared to correlate with phonetic height and backness respectively.

1.3.3 Formants from Speech Recording

The first and second formants were extracted from speech recordings of the participants and used to create an F1/F2 plot, the definition of a vowel quadrilateral. Figure H.3 (again normalised using the Bark scale and Lobanov normalisation) showed great similarities to the vowel distribution seen in the PC and resonance plots derived from the transformed area data. It was concluded that principal components and resonances derived from MRI image data are suitable representations of vowel position.

1.4 Project Aims

Following on from Watson's study of area functions from MRI images [41], my Part IV Project aims to extend this analysis by:

1. Further validating the feasibility of using combined vocal tract area data from MRI images *across multiple speakers* so that group trends on vowel sounds can be studied. This will be done by performing the above analyses on a larger dataset of 12 speakers.
2. Repeating the analysis on a General American English (GenAm) dataset to quantify the effect of accent.
3. Validating our methods of processing MRI images for phonetic data.

[7] A frequency scale on which equal distances correspond with perceptually equal distances.

[8] This is because the range for R2 is far greater than that of R1.

[9] This algorithm scales speaker's formant values as proportion of speaker's maximum formant frequency, allowing better comparison between speakers with different average pitches.

Chapter 2

Methods

To list the tools used for this project, MRI images were processed using CMGUI, Perl and MATLAB. All further analyses on the area functions extracted from the MRI images, as well as the American English (GenAm) area functions from Story [33], were carried out in R [36]. To preserve the anonymity of the participants, speaker labels 'VT' and 'SP' have been used, for NZE and GenAm data respectively.

2.1 Area Functions from MRI Images

The initial step in this project was to extract area functions from the two speakers in the MRI data set who were yet to be analysed, VT04 and VT07. The method and scripts for extracting the vocal tract shape from these MRI images were provided by Dr. Catherine Watson[1]. In order to extract the area functions from the MRI images, CMGUI[2] (version 2.7.0) was used to generate cross-sectional slices[3] along the curve of the vocal tract from the sagittal MRI images, from the lips to the glottis. Data points were then manually placed around the boundary of the oral and pharyngeal cavities (The air space between the tongue and the upper/back surface of the mouth and throat) on these slices, to outline the cross-sectional area be measured on each slice. A Perl script is then used to calculate the cross-sectional areas on each slice, and a MATLAB script used to calculate the distance of each slice along the vocal tract curve from the lips. An area function is produced when these cross-sectional areas (y-axis) are plotted against the distance from the lips (x-axis). Similar methods employed in other studies have shown that area functions extracted from MRI and X-ray images are able to capture vowel qualities and match with acoustically measured vowel qualities (formants) [16, 44, 48]. This process was completed initially for one

[1] See Acknowledgements for further details on those who contributed to the development of this procedure.

[2] The interactive front-end of the software CMISS, which stands for Continuum Mechanics, Image Analysis, Signal Processing and System Identification.

[3] These are initially coronal at the mouth, and become increasingly transverse as the vocal tract curves down the throat capturing the cross-sectional area of the cavity.

set of vowels for VT04 and VT07[4], but further sets were later analysed as needed[5]. Please see Appendix B for a schematic explaining how these sets of data are grouped.

2.1.1 Acquisition of MRI images

Previous to this project, MRI images of vocal tracts of 12 participants (henceforth referred to as 'speakers') were captured using 1.5T and 3.0T Siemens Magnetom Avanto MRI scanners [41]. For each speaker, 13 saggital sections of the head were taken from jaw edge to jaw edge, with 6 mm separations between each scan. The resolution of the images was 1 mm with a field of view of approximately 200 × 250 mm. The participants were all speakers of New Zealand English (NZE), with a range of age and genders (see Appendix C). They were instructed to articulate and hold one vowel at a time for 15 seconds while the scans were taken. The vowels which were captured were the 11 NZE monophthongs (simple vowels) /iː, ɪ, e, ɜː, æ, ɑː, ɒ, ʌ, ɔ, ʊ, u/, uttered in /hVd/ syllable frames[6]. The sequence of 13 parallel saggital scans was repeated twice for each of the 11 NZE monophthongs, creating two sets for each speaker.

2.1.2 Defining the vocal tract in CMGUI

Defining the curve of the vocal tract

These MRI images were then read into CMGUI using the provided com files. For each individual vowel, the following procedure was carried out twice - once for the oral cavity, and once for the pharyngeal cavity. The mid-saggital section from the vowel's 13-slice set (Slice 7) would be displayed in the CMGUI graphics window with the command file `snake.com`. The user would then place data points manually along the centre of the vocal tract to define the curve along which cross-sectional slices would be generated. Figure 2.1 shows the curvature of the oral cavity from lips to uvula. A separate curve would be made for the pharyngeal cavity, from the uvula to the glottis, as each cavity is treated completely separately until the area functions are combined in the final step. The command `create_curve` would generate 15 equispaced data points along this user-defined trajectory, and the coordinates of these equispaced data points automatically saved to a file called `curve.exnode`.

[4] Only Set 1 vowels were processed to begin with, since only one set of vowels per speaker would be needed for inter-speaker PCA and resonance correlation analyses, and some of the speakers did not have a second set of data available.

[5] Set 2 of VT11 was needed for intraspeaker correlations (Section 3.2.5 and the Set 1 cardinal vowels from VT04 and VT11 for methodology validation (Section 2.2)

[6] The 'V' in /hVd/ is a variable which would be replaced by the vowel in question, with a voiceless glottal fricative /h/ and a voiced alveolar stop /d/ on either side. Such vowel frames are popular in studying phonetics for the neutralising effect that they have on the vowel, which allows a 'purer' vowel sound for analysis. Examples of vowels in an hVd frame would include *had*, *hid* and *head*. See Appendix B for the full list.

Creating orthogonal slices

The output `curve.exnode` were fed into `CreateSlices.pl` which automatically generated CMGUI `.exnode` and `.exelem` files. These files represented a set of 15 planes at each of the 15 data points, orthogonal to the curve (Figure 2.1). The command file `create_slices.com` then produced a sequence of images progressing through vocal tract orthogonally throughout its length, by interpolating between the 13 original MRI sections and projecting the resulting 3D volume texture onto the orthogonal planes. These slices were displayed on the orthogonal slices along with the mid-saggital section using `data_point_placement.com`.

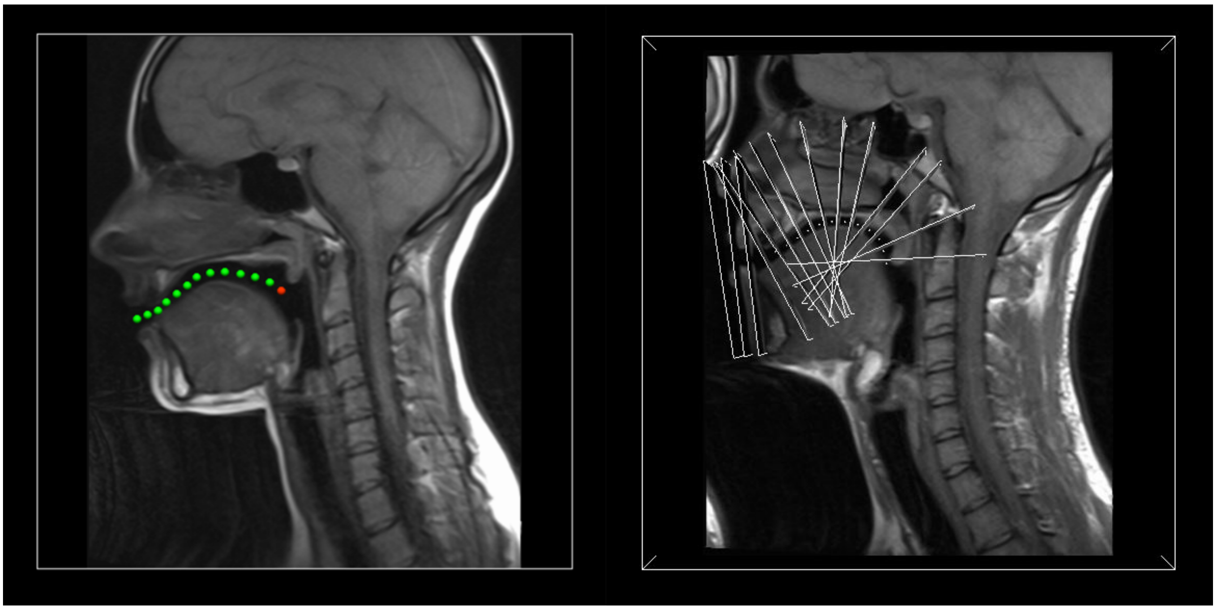


Figure 2.1: Right: Data points defining the curvature of the oral cavity in CMGUI's graphics window. Left: Slices orthogonal to the user-defined curvature of the oral cavity are displayed on a mid-saggital MRI section.

Outlining vocal tract cross-sections

From here, the user outlined the vocal tract boundary on each slice, one at a time, marking this boundary with data points (Figure 2.1.2). The vocal tract appears as an area of lower signal on the MRI image in the centre of the slice, usually symmetrical in shape. Once all 15 slices are annotated with data points, the function `write_to_file` saves the data point coordinates to `.exnode` files for display, and `.exdata` files for the next step in area function calculation. As a final review step before calculating area functions, all of the slices and their associated data points can be viewed using `data_point_viewer.com` with the function `read_from_file`.

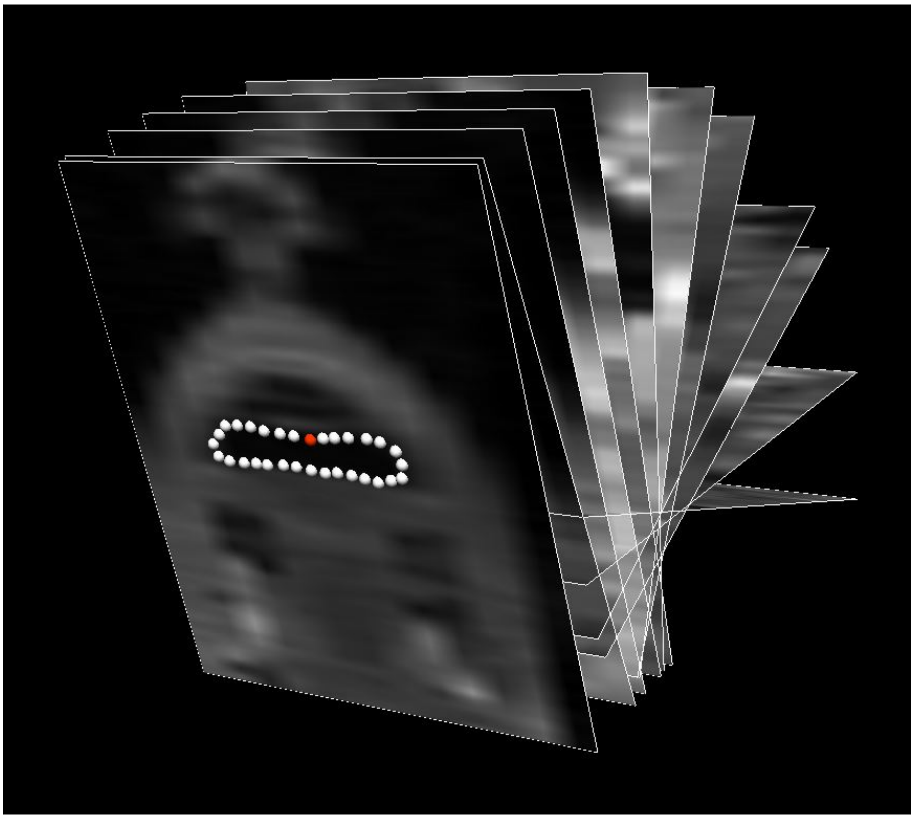


Figure 2.2: The boundary of the oral cavity on one of the planes being marked up by data points. The cross-sectional slice at the front shows a cross-section of the tongue, upper lip, and near the top, the beginning of the nasal cavity.

2.1.3 Calculating area and distance

The `.exdata` files are called by the script `calculate_area.perl`, which outputs 15 values in a .txt file: the cross-sectional areas from the data points on each of the 15 planes (in mm^2). The above process was repeated for the pharyngeal cavity, and then the whole process for both cavities again for all eleven vowels in a speaker's set. The MATLAB script `areacalc.m` can then be run to combine the two `area.txt` file for each vowel (oral and pharyngeal) and to calculate the corresponding distances of the slices from the lips. The result are two columns for distance from lips and cross-sectional areas with 29 data points each[7], printed to individual .txt files named after each vowel (e.g. `had.txt`, `hard.txt`) in the folder `distance_area/`. Area functions may also be visualised using `areaplot.m`, which lay out each vowel's area function in its approximate position on a vowel quadrilateral (see Figures C.1 and C.2 in Appendix C). Area functions for each speaker grouped by vowel height can also be seen in Appendix D.

2.1.4 Reading in area functions to R

These area function .txt files from VT04 and VT07 were read in R along with those of all the other vocal tracts into a single data frame on R. This gave a combined data frame of 132 area

[7] The middle value is an average of the last value from the oral cavity and the first value from the pharyngeal cavity, hence the 29 data points.

functions (11 vowels per speaker × 12 speakers) for the NZE data. Each row contained the area function from X1 to X29, linearly interpolated with equispaced area values along the vocal tract, alongside labels indicating speaker number, set number (Set 1 or 2, if applicable) and vowel from which the area function was extracted[8]. Interpolation was required due to the different distance step sizes between the oral and pharyngeal regions. To confirm that the interpolated area functions follow the raw data closely, Figure E.2 compares the raw area function data with uneven distance steps[9] with the linear equispaced interpolation points. Despite a slight lag which develops halfway through the function, the shape of the area functions are maintained. Also, two speakers were plotted to show that despite VT04's smaller amplitude relative to VT07, the same general VT shape is seen for the same vowel.

For the American English data, the area functions of 6 GenAm speakers published in Story [33] (Tables XI to XVI) were copied and pasted into Excel and saved as comma-separated values. I then wrote a script to read these .csv files into a data frame on R which was compatible with the data frame made for the NZE data, to allow them to be combined. Plots of these area functions are available in the cited article [33].

2.1.5 Mean area functions

To quickly visualise any trends that may be present in the area function data, mean area functions were calculated on R. The means area values (dependent variable) were plotted as our combined data frame of 11 NZE speakers and 6 GenAm speakers was divided into vowel types (high, low, front, back) and demographics groups of speaker accent, gender and age. The areas were not normalised, to keep differences in vocal tract size intact for analysis, although our NZE area function values were converted from mm^2 to cm^2 to match the units of the GenAm. The corresponding distances from the lips (independent variable) were not used in the plots, to normalise for vocal tract length.

2.2 Methodology Validation & Critique

Area functions derived from MRI images are the foundations on which this project is based, and their validity as a source of phonetic information is a key aim of this project. As such, it was crucial to validate the data extracted from the images and to minimise opportunities for error to be introduced in the image processing pipeline. Alongside the tests described below which were designed and carried out specifically for this reason, I also comment on the validity of my results and methods in each the analyses to follow.

[8] Again, see Appendix B for a clearer picture of these groupings

[9] This is due to different lengths of the oral and pharyngeal cavities, each of which are divided evenly into 15 cross-sectional area slices causing slightly different distances between each slice between the regions.

2.2.1 Quantifying variability in image processing

This procedure of extracting area functions from MRI images was validated across several stages. First, I sought to minimise the error introduced in the manual step of marking up the vocal tract boundaries on CMGUI. This was done by carefully reading and following notes from previous researchers who had already marked up some of the speakers' data[31, 7, 6], and displaying their marked up vocal tract boundaries on CMGUI using `data_point_viewer.com` (see Section 2.1.2 ”Outlining vocal tract cross-sections”). Since my final data set of 12 NZE speakers consisted of data marked up by three different researchers, it was important that the strategies we used to deal with imaging artifacts or blurry features were kept consistent since all of the area functions would be analysed together.

However, variability was inevitable in the way a user manually places the data points on the cross-sectional slices. This variability was quantified by performing correlations on the principal components of the same sets of data which had been processed a second time, to see if the area functions coming from the repeated procedures correlated well with the original area functions. I selected the four cardinal vowels from VT04 Set 1 and VT11 Set 1 (Had, Heed, Hod and Who'd) to repeat, and performed Pearson product-moment correlations between the principal components (PCs) of the area functions on R (more information on these statistical techniques in Section 2.3). These two sets were chosen since VT04 Set 1 had been processed by me the first time through, and VT11 Set 1 had been processed by Helen [31]. Repetitions on these two sets could show variability for the same user (VT04 Set 1), and between different users (VT11 Set 1)[10]. Given that these area functions are derived from the same data, any deviation from 100% correlation was an indication of the level of variability inherent in this extraction procedure.

The results in Table E.1 showed that the variances accounted for by each principal component were very similar, and the PC-rotated data of the two repetitions within each set were very highly correlated. Having different users mark up the images appeared to introduce no extra variability, even despite imaging artifacts (VT11). This confirms that this methodology is highly repeatable even for different users, and that area functions derived by different users can be combined and analysed together. On a site note, the proportions of variance accounted for by the first three principal components (almost adding up to 100%) showed that principal component analysis (PCA) is also a valid way of reducing the dimensionality of combined area function data. In addition to these results, we later find that correlations even between the area functions from *different* speakers (completely different data) are very strong, allowing us to draw conclusions from any trends that we see from the PCA analyses (Figure 3.6, Section 4.2.2).

[10] VT11 was also a good choice for checking variability between different users as it was something of a worst case scenario. There were imaging artifacts in the oral cavity region of VT11's images, which had the potential to introduce extra variation as Helen and I may have treated the obscured features differently.

2.2.2 Limitations of MRI image processing

There are several quirks to the CMGUI Scene Editor and Graphics Window which must be noted before processing the MRI images, such as data points needing to be placed in a consecutive clockwise or anticlockwise direction, and data points failing to generate on the right half of a slice[11]. Further, many of the interpolated slices showed rather unclear, blurry and often asymmetrical structures. Bony structures such as teeth give low signals due to their low water content, often intersecting with the dark areas of air indicating the vocal tract. Physiological assumptions should be used in these cases, such as assuming symmetric, elliptic vocal tract cross-sections where appropriate (see [31] for more details).

It is for these reasons that an automated method of outlining the vocal tract in these sections would be difficult to achieve, even if could improve repeatability by eliminating the only remaining manual (and very time-consuming) process in the pipeline. Increasing the number of sagittal slices taken in the raw data set may help with these ambiguities; however, MRI data acquisition is a slow and rather strenuous process for participants, who must hold completely still while articulating these vowel sounds for long periods of time. Given the previously discussed repeatability of manually marking up these vocal track boundaries, we can conclude that the benefits of a researcher being able to judge the features in each plane on a case by case basis outweigh the human errors they introduce and the effort it would take to teach a machine to make these judgements with the same level of accuracy.

One other limitation to the CMGUI step of the image processing pipeline was the error from the large step sizes during the cross-sectioning of the vocal tract. Ideally, there would be infinite slices which create a smooth volume shape defining the vocal tract. However, this is impractical since every slice must be manually marked up. The large steps between slices mean that some of the volume and shape information is lost (Figure E.1). The trade-off between fidelity and the manual processing time required is a difficult one.

2.2.3 Improvements made to image processing pipeline

At the beginning of this project, I was given a set of resources from previous research periods including raw MRI data, MATLAB and R scripts, and reports from previous research students [31, 7, 6]. The data of the 12 speakers was split among the different periods, with five of the speakers having had both sets of data analysed by Daniel, and the rest of the speakers bar VT04 and VT07 having one set per vowel analysed by Helen. All of this data was consolidated into one directory which contained a consistent file structure for clarity, and to allow automatic reading of files in my R scripts.

Processing the raw MRI images for VT04 and VT07 involved several tedious steps, such as

[11] Instead, the user must create the data point on the left half of the slice and drag it over to the desired position. However, this issue disappeared when using the same CMGUI executable on the Department of Engineering Science computers.

copying and pasting multiple files across several directories and renaming certain files and lines of code to match the vowel or speaker at hand. To speed up this process, I wrote a batch file which could be pasted into a Windows Command Prompt (`cmd.exe`), punctuated by steps where tasks had to be carried out on CMGUI. This sped up the process considerably, leaving only the absolutely necessary steps for manual processing.

Several text files containing R commands and console outputs were provided by Dr Catherine Watson. The existing code was reworked to accommodate a variable data set with any number of speakers, sets and vowels. The functions were also simplified in their number of inputs, and generalised to improve versatility and re-usability.

2.3 Principal Component Analysis on Area Functions

In order to reduce the number of dimensions in our data frame of area functions, principal component analysis (PCA) was used via the `prcomp` function in R [36]. PCA is a statistical technique for replacing variables with much fewer, linear combinations of the original variables which retain the majority of the original data's variance called principal components (PCs) [15, 26, 13]. These principal components encode the dimensions along which there is the most variance, in decreasing order - the first principal component (PC1) aligns with the dimension along which there is the most variance in the data, the second principal component in the second-most-varied dimension and so on. PCA has often been used on tongue contours and vocal tract shapes in phonetics research for its apparent ability to decompose vocal tract data into the main features of the vowels they represent, namely phonetic height and backness[12] [24], given that these should intuitively be the greatest variations in vocal tract shape for different vowel sounds. Many studies have shown that the first two principal components suffice in capturing up to 90% of the variance in the original data, the first appearing to encode phonetic backness and the second encoding phonetic height [34, 16, 44, 48, 24].

R functions and scripts were written by me to automate the analyses described below. For the NZE data, the first and last values of the 29-point area functions were omitted before as the corresponding MRI image slices were usually poorly defined. The first and last slices, the lips and the glottis, were clear, consistent anatomical landmarks for defining the start and end of the area functions. However, they were difficult to trace in their cross-sectional slice since the first only showed the outermost edge of the lips, and the glottis is blurred by the vibrating vocal folds. The GenAm data from Story [33] had 44 data points, none of which were omitted.

[12] These are simplified terms for complex and nuanced tongue movements termed ‘front-raising’ and ‘back-raising’ [18], but our definitions will suffice for these analyses.

2.3.1 Vowel plots on PC1-PC2 planes

PCA was performed on the *combined* set of NZE vowels (132 area functions in total, 11 vowels from each of the 12 speakers), which were first normalised vowel-specifically by expressing all cross-sectional areas as a proportion of the maximum cross-sectional area for that vowel [13]. Each vowel was plotted by their first two principal components on PC1-PC2 planes (PC1 on the x-axis, PC2 on the y-axis) using the `eplot` function in the EmuR package [45] to assess whether these principal components separate out the vowels by their height and backness.

2.3.2 Inter-speaker PCA correlations and demographic trends

To analyse trends within different demographic groups, PCA was performed on each speaker's data set separately (11 area functions/vowels each), and a Pearson product-moment correlation performed between PC1, PC2 and PC3 of each speaker (hence 'inter-speaker' PCA correlations) [24, 34, 41]. GenAm vowel data from Story [33] was also included, which each had 44 data points, requiring us to interpolate the 29-point NZE area functions to the same number. The correlation estimates and their p-values were written into data frames and output to a .csv for viewing in Excel.

2.3.3 Intra-speaker PCA correlations

In an attempt to quantify individual speaker effects, the same PCA correlation analysis was carried out between vowel sets from the *same* speaker, when there were two sets of raw MRI data available. We had access to two sets of raw MRI data for each speaker, with some exceptions (Set 2 may have been missing or incomplete, like only having second MRI images of certain vowels, having less than 13 sagittal slices for the second set. See Appendix B for full list.). Former Masters student Daniel Tan had already processed both Set 1 and Set 2 for five speakers (VT03, VT05, VT08, VT09, VT10), whose intra-speaker correlations were published in Watson [7, 41]. Given the time-consuming nature of processing these images and the gaps in the data where full second sets were not available, it was decided that a second MRI image set would only be processed for VT11 to add to the intra-speaker correlation analysis, selected for its consistently high inter-speaker correlations.

2.3.4 Validation of code

To validate the R code that I had written to carry out these analyses, I compared my results with the ones from Watson which used the same set of data [7, 41]. I attempted to reproduce

[13] Each area function (i.e. each vowel) was divided by the maximum value in that particular area function. This gave better separation of vowels by height and backness on the PC1-PC2 plots.

the inter- and intra-speaker correlations and proportions of variance accounted for by the first three principal components using the same 2 sets from the 5 speakers. Intra- and inter-speaker correlation values were all similar in that they were extremely strong, but the values were not an exact match[14]. The same was true for the proportions of variances (Table 2.1). The reasons behind this discrepancy is unclear - it may be that Watson [41] did not linearly interpolate the area functions when reading them in from the .txt files which may have introduced slight variations[15]. Overall, the variances and correlations were deemed close enough to approve my methods.

Table 2.1: Variance (%) accounted for by principal components in combined data sets

Variance	**PC1**	**PC2**	**PC3**	**PC1+PC2**
Watson [41]	39.6	20.8	10.9	60.4
5VT 2 Set	41.7	20.43	10.44	61.5
12VT 1 Set	39.8	20.3	9.4	60.1

2.4 Resonance Analysis

The following is a summary of the method used to calculate resonances from area functions, as explained by Harrington & Cassidy [12, p. 216-244]. LPC is a time domain technique which models a signal y as a linear combination of delayed values weighted with coefficients $\alpha_1 ... \alpha_p$ (where p is the order of the LPC model) plus some error signal ϵ_n [12]. We aim to find a set of coefficients that minimise ϵ_n^2.

$$y_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + ... + \alpha_p y_{n-p} + \epsilon_n \tag{2.1}$$

As we hold this thought, let us consider a model of vowel sound production, a periodic source (the opening and closing of the vocal folds in the glottis) moving through a vocal tract filter (the effect of the vocal tract shape on periodic sound waves from the vibrating vocal folds). This is called the source-filter model of speech [5, 8], and it is described by the convolution in Equation 2.2 where x and y are the input (source) and output signals respectively, and a and b are their coefficients which encode the properties of the vocal tract filter (its resonances and anti-resonances).

[14] For intra-speaker correlations, compare values along the diagonal in Table H.2 with values from Figure F.6 for the speakers with the speaker label 'SP'. For inter-speaker correlations, compare the off-diagonal values in Table H.2 with values from Figure 3.6. All speakers that were also used in the Watson publication [41] are indicated on the participants list (Table A.1) with speaker labels 'SP'.

[15] The fact that Watson [41] did not omit the last data point was taken into account; the data point correponding to the glottis which was omitted in all my other analyses was added back in.

$$y * a = b * x \tag{2.2}$$

For the vowels that we are analysing, which are considered to be purely periodic sounds, $b = 1$. The convolution in the reduced equation can now be expanded as a linear combination of scaled delayed values [12, p. 158], which when re-arranged, bear a remarkable resemblance to Equation 2.1. We can now clearly see that for a speech signal whose spectrum consists of resonances (as is the case with vowels in this all-pole model of speech production), we can directly relate the p-order LPC coefficients $(/alph_n)$ with the k coefficients of the vocal tract filter (a_n).

$$y_n = -a_1 y_{n-1} - a_2 y_{n-2} - ...a_k y_{n-k} + x_n \tag{2.3}$$

The only term left to equate is the error signal ϵ and the source signal x. Fortunately, the LPC procedure of fitting a set of coefficients which minimise ϵ^2 flattened the spectrum of ϵ, while an impulse train (which also has a flat spectrum) is an appropriate model of vowel sounds (given the periodic motion of the vocal folds). From here, there are several techniques for calculating the LPC coefficients, out of which the recursive method (which we use in our analyses) is explained in Harrington [12, p. 220].

Once the LPC coefficients have been calculated, they can be related directly to the area function via a set of reflection coefficients m_i. When the vocal tract is modelled as a set of adjacent, lossless tubes of equal width, the reflection coefficients are simply a ratio between the areas of these tubes which encode the reflection at these boundaries. The equation for this is given in Harrington & Cassidy [12, p. 233]

Following the methods described above, the resonances each of the area functions in our data set (both NZE and GenAm) were calculated based on a standard linear prediction model of speech. Three R functions were provided for calculating reflection coefficients (RCs) from area functions, converting RCs to LPC coefficients using a recursive algorithm [21], and finally modelling a frequency spectrum using the LPC coefficients [41, 44]. These functions were incorporated into a script which would output a spectrum corresponding to each of the area functions[16], and pick out the first n resonances. The hypothetical sampling frequencies (f_s) of each spectrum were also calculated (the highest frequency in the spectrum) to be able to scale the frequency bins for each spectrum according to the speaker's vocal tract length. The sampling frequency can be calculated using the relationship in Equation 2.4, determined by the number of cross-sectional areas (M) provided, the length of the speaker's vocal tract (L), as well as the speed of sound in air (c).

$$M = \frac{2Lf_s}{c} \tag{2.4}$$

[16] Area functions were not normalised since the absolute values of areas are irrelevant as the reflection coefficients are obtained by ratios between adjacent cylinder areas.

Assuming that $\frac{2L}{c} = \frac{2\times 17\mathrm{cm}}{34000\mathrm{cm\,s^{-}1}} = \frac{1}{1000}\mathrm{s}$, the equation can be simplified to $M = \frac{f_s}{1000}$. A sampling frequency of around 8000 Hz would be recommended to capture at least the first three resonances [12, p. 239]. This means that a minimum of 8 cross-sectional areas (M) must be provided (9 data points). Furthermore, an LPC-smoothed spectrum can only represent up to half the number of peaks as the number of LPC coefficients provided (= the number of cross-sectional areas, M) [12, p. 227]. In our case where we are analysing only the first three resonances, this required us to have, at minimum, the larger of 6 (3 peaks $\times$ 2 LPC coefficients per peak) or 8 ($f_s/1000$) areas [17]. Once the number of data points was determined and the frequency bins scaled, the spectra from each area function were superimposed onto their respective frequency scales and the first three resonant peaks identified. Each vowel was plotted on an R1-R2 plane in a similar manner to the PC1-PC2 plots (Section 2.3), and inter-speaker correlations made between R1, R2 and R3[18].

2.4.1 Limitations of the acoustic tube model

The lossless acoustic tube model on which this method is based is "at best a gross approximation to the actual cross-sectional area of the vocal tract and at worst [providing] misleading and inaccurate information" [32, 12]. The reasons given by Harrington & Cassidy [12] are:

1. Unaccounted wave energy losses to heat and vibration through vocal tract walls (some solutions provided by Wakita [40]).
2. The assumption of plane wave propagation, only valid for frequencies below 4000 Hz (some consonants may exceed this, but vowel formants are well below this value).
3. The inability to model anti-resonances (these mostly affect nasal and lateral consonants [28, 17, 25] which we are not studying).

The effects of these limitations on the accuracy of our model for our purposes of analysing MRI-derived area functions vowel sounds appear to be minimal[19].

[17] 9 data points in total as a minimum. We found that increasing the number of data points from this minimum did not affect our method's accuracy in identifying first three resonant frequencies. Eventually, we decided on 16 data points, interpolated from 29 data points for NZE or 44 in the case of GenAm.

[18] Intra-speaker resonance correlations would have been possible for NZE speakers who had two sets of images. However, given the absence of meaningful correlations between higher order principal components in the intra-speaker PC correlations, any trends or individual speaker effects appearing in these derived resonances was unlikely. It was also decided that this analysis would not be needed as a validation step, since the other analyses showed such strong inter-speaker correlations and R1-R2 plots.

[19] Another disadvantage to this method is that an area function generated from a speech signal via LPC may include with it errors introduced by the fact that the LPC coefficients encompass not only the effects of the vocal tract filter, but also that of of the glottal filter as well as lip radiation [12, p. 219]. Therefore, the area function generated is affected by factors other than the resonances of the vocal tract shape. Fortunately, as we are calculating resonances *from* area functions which are anatomically derived, the effect of vocal tract shape (the vocal tract filter) is completely isolated for analysing.

Chapter 3

Results

3.1 Area Functions

Please note that due to page constraints, many figures have been included in the appendices. These will be referenced where appropriate, with labels beginning with an alphabetical letter followed by the figure number.

3.1.1 Area functions of VT04 and VT07

Two complete sets of area functions consisting of 11 vowels each were added to the existing data set from VT04 and VT07 respectively (Figures C.1 and C.2 in Appendix C). The area functions for these two speakers appeared to be smaller in amplitude when compared with area functions produced by the other speakers (Figure E.2; also see [31]). This was somewhat expected, since it was noted by Dr Catherine Watson as she was supervising the MRI acquisition that these two participants did not enunciate or open their mouths as clearly as the other speakers. For this reason, these two data sets had been left until last to analyse.

Nevertheless, their shapes follow the general trends for each vowel. For example, when looking at top row of area functions for VT04 (Figure C.1), we see that for the front high vowel 'heed', the cross-sectional area is very low in the oral cavity at the front of the mouth (before the blue line), but increases to a peak in the pharyngeal cavity (after the blue line). This indicates that the tongue is lifted towards the roof at the front of the mouth, leaving little air space in the oral cavity, but allowing the pharyngeal cavity area to increase as the tongue back is pulled forward, as one would expect for a front high vowel. On the other end of the row, we have the high back vowels such as 'hoard' and 'hood', which show the opposite shape with larger oral cavities (peak in front of the blue line) and smaller pharyngeal cavities as the tongue arches back and fills up the space at the back of the vocal tract.

3.1.2 Mean area functions

The mean area functions shown in Figures 3.1 to 3.4 are plotted with the area nearest to the lips at index 0, then progressing along the vocal tract towards the glottis. The first half of the indices (approximately index 0 to 22) indicate the oral cavity areas, while the second half represents the pharyngeal cavity. Firstly, Figure 3.1 clearly shows the mean cross-sectional area down the vocal tract is much lower in the front of the mouth for high vowels than it is for low vowels, and vice versa in the pharyngeal region. The mean area functions of our NZE and GenAm data show similar average cross-sectional areas in the pharyngeal cavity, but a higher peak in the oral cavity for GenAm (Figure 3.2). For gender (Figure 3.3), surprisingly, female cross-sectional areas appear to be on average larger throughout the vocal tract until the very end of the pharyngeal cavity when the two area functions merge. This effect decreased significantly when VT12 was removed from the female data set. In the case of age groups (Figure 3.4), the middle aged speakers seemed to have larger overall cross-sectional areas in both the oral and pharyngeal cavities. There was little difference between the other groups, but for the large peak in the pharyngeal cavity for senior speakers. All demographic groups appear to follow a similar bimodal shape, with the first peak in the oral cavity being larger than the second peak in the pharyngeal cavity.

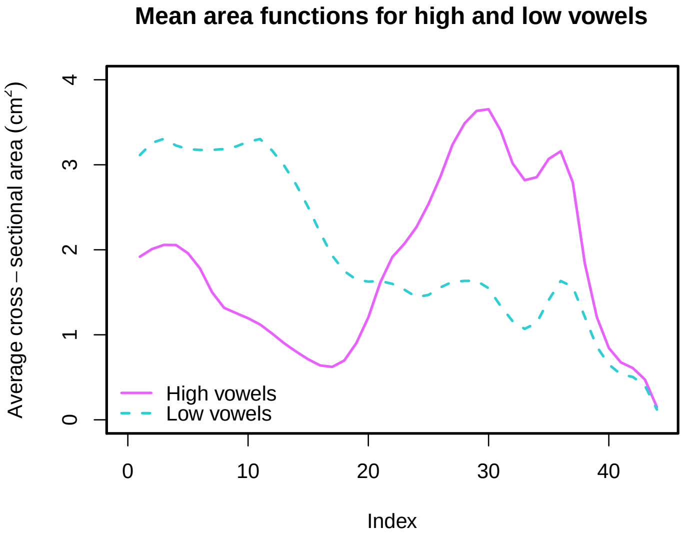


Figure 3.1: Mean area functions for high vowels (heed, hid, who'd) and low vowels (had, hard, head, hod)

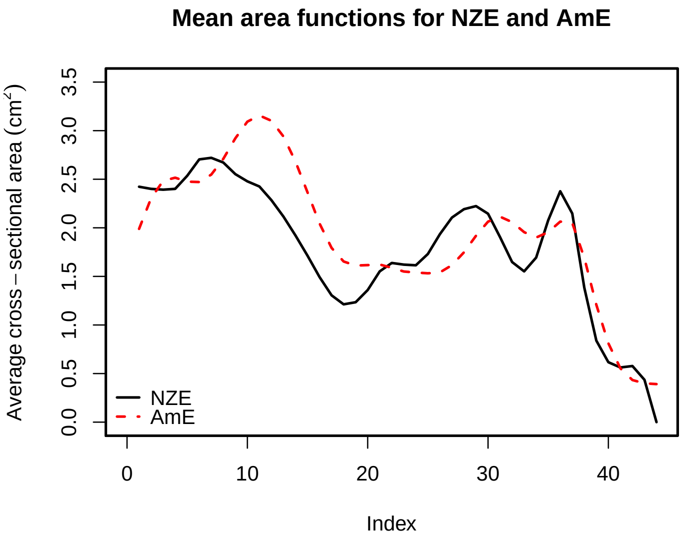


Figure 3.2: Mean area functions for NZE and GenAm

3.2 Principal Component Analysis on Combined Area Functions

3.2.1 Vowel plots on PC1-PC2 planes

Figure F.1 shows all of the vowels across all 12 NZE speakers[1] plotted by the first and second principal components of their area functions. The same vowels appear to cluster together somewhat. Plotting each vowel's centroid on a PC1-PC2 plane shows great similarity with the centroid plot in Watson [41] (Figure 3.5). In Figure 3.5a, we see each vowel centroid accurately separated out by height and backness as in a standard vowel quadrilateral (Figure 1.2), with relatively little error[2]. However, there are several instances where the vowel *heights* appear to be out of order, such as the /ɔː/ vowel appearing higher than the /ʊ/ and /uː/ vowels, /ʌ/ being lower than both /ɑː/ and /ɒ/, and /ɜː/ being lower than æ and ɪ. This may suggest that PC2 is a weaker predictor of vowel height relative to PC1's accurate performance in separating out the vowels by backness.

[1] A total of 132 area functions, with 1 set of 11 vowels coming from each of the 12 speakers

[2] This would be the case if the x-axis was flipped. We did not flip the axis here to visually show the current plot's similarity to Figure 3.5b. The order of the axes/the sign on PCs are insignificant, as PCA only captures the *direction* of greatest variance.

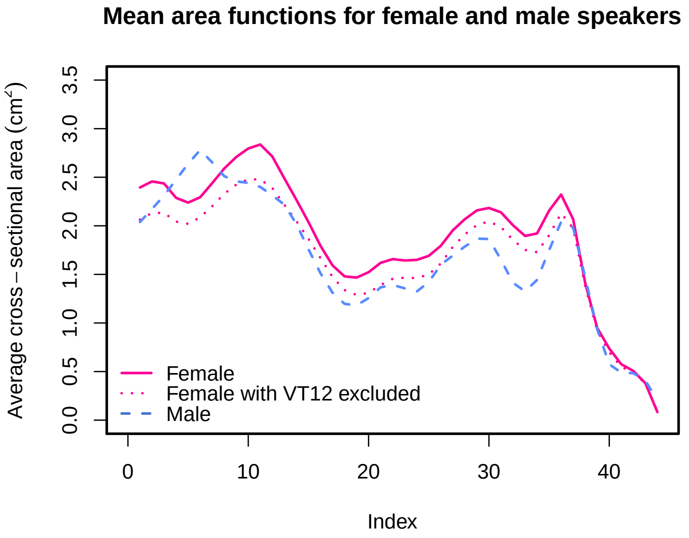
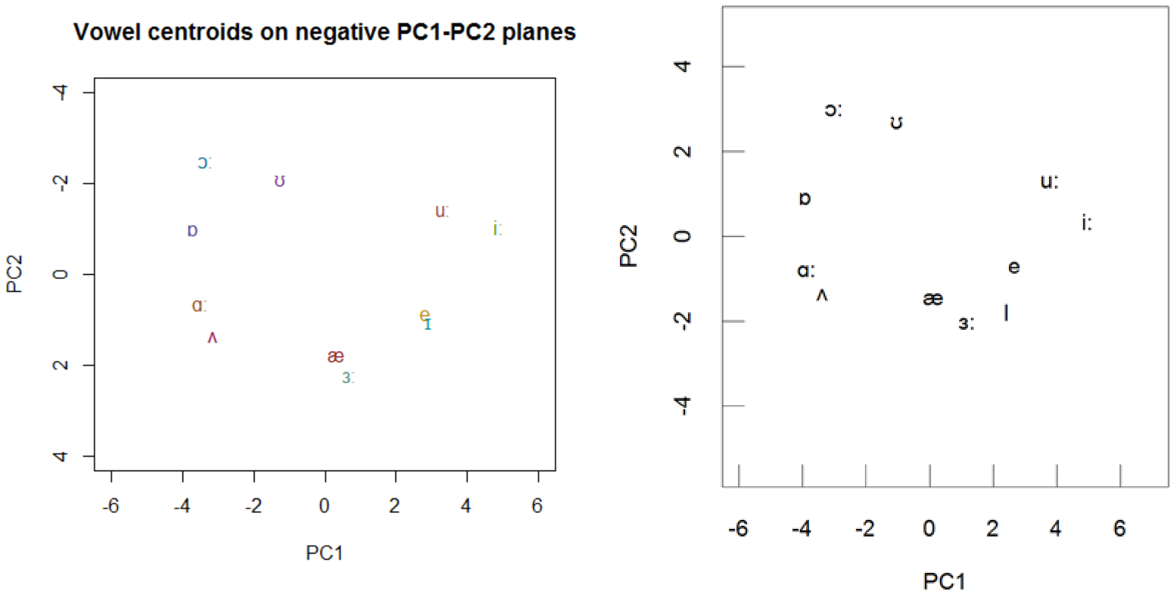


Figure 3.3: Mean area functions for female and male speakers



(a) Centroids from NZE dataset (12 VTs, 1 set each)

(b) Centroids from Watson (5 VTs, 2 sets each)

Figure 3.5: Comparison of vowel centroid PC plots from combined NZE data and Watson [41]

3.2.2 Proportions of variance

The first two principal components (PC1 and PC2) of the NZE data set accounted for 60.1% of the variance, virtually the same proportion captured in the original study done with a combined data set of just 5 speakers with 2 sets each (Table 2.1) [41]. Figure F.2 shows that significance of the PCs drop significantly until PC3, where it begins to plateau at very low variance proportions.

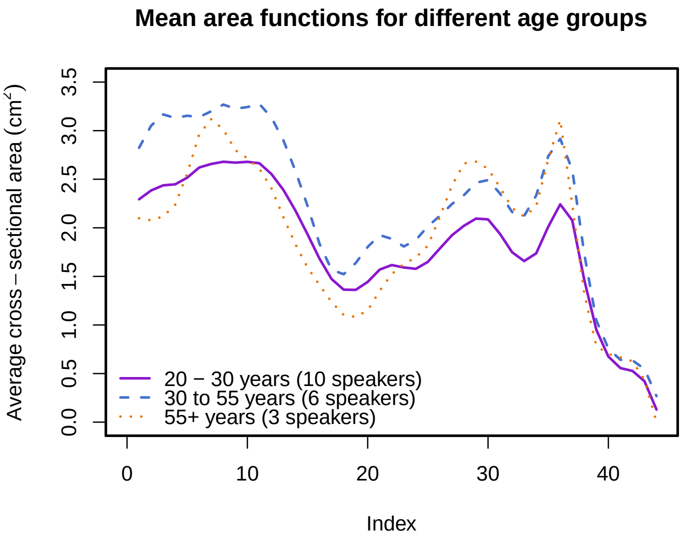


Figure 3.4: Mean area functions for different age groups

3.2.3 Validating individual speaker data

In order to validate these results, principal component analysis was performed on each individual speaker's data (11 vowels each) to check for consistency between the speakers' area functions. Firstly, the proportion of variances for each individual speaker was assessed. The results in Figure F.3 show that for all of the speakers except VT06, the variances are still largely accounted for by PC1 and PC2 alone. In the majority of cases, the two most significant principal components accounted for over 70% of the total variance (orange line). This also verified that VT04 and VT07's data which I processed contained meaningful information that is consistent with the other speakers, despite initial doubts about their validity due to their area functions being relatively small in amplitude.

3.2.4 Inter-speaker PCA correlations and demographic trends

Figures 3.6 to I.1 are tables of correlation estimates between the first three principal components (PC1, PC2 and PC3) of each speaker's combined data (principal component analysis performed on the area functions of all 11 vowels for each speaker). Blue 'VT' speaker labels indicate NZE speakers, while red 'SX' labels indicate GenAm speakers with 'X' indicating gender. A black border highlights the correlations between NZE and GenA speakers. The correlation estimates (as well as their corresponding p-values in Appendix G) have been colour coded to better visualise any trends, and a color bar has been provided on each figure to show the values corresponding to the stages in the gradient scale[3].

[3] The absolute values of correlations have been taken for ease of color coding and interpretation, since the signs on principal components do not have any effect on the variance that they capture. In any case, all correlations for PC1 were all positive to begin with - negative correlations only tended to occur in the lesser principal components, especially PC3.

The results of the Pearson product-moment correlations showed generally very strong correlations between all speakers, averaging around 90% (Figure 3.6) with most p-values well under 0.001 (Figure G.1). The second principal components are moderately correlated, approximately 65% to 75% with a few completely uncorrelated pairings (Figure 3.7). These weaker correlations were also reflected in the p-values, roughly half of which were borderline (nearing 0.05) (Figure G.2). VT04, VT07 and SF2 appeared to have more instances of weak correlations (under 50%) than the other speakers. As for PC3, there is virtually no correlation between the speakers (Figures 3.7 and G.2). There were no visible trends of correlations being higher or lower within gender or age groups[4]. For accent groups, the average PC1 and PC2 correlation values were slightly lower between different accents (inside the black boxes) than within the same accent groups.

PC1	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	0.94	0.92	0.87	0.94	0.91	0.96	0.91	0.92	0.90	0.95	0.96	0.90	0.78	0.78	0.88	0.79	0.74
VT02		0.93	0.84	0.90	0.90	0.95	0.87	0.91	0.85	0.88	0.97	0.93	0.73	0.75	0.93	0.78	0.83
VT03			0.75	0.96	0.92	0.94	0.97	0.90	0.87	0.86	0.95	0.91	0.81	0.80	0.94	0.82	0.90
VT04				0.80	0.84	0.88	0.73	0.70	0.73	0.88	0.84	0.76	0.60	0.62	0.70	0.58	0.63
VT05					0.94	0.94	0.98	0.92	0.87	0.92	0.95	0.88	0.81	0.76	0.88	0.75	0.81
VT06						0.95	0.95	0.87	0.86	0.93	0.95	0.82	0.83	0.66	0.87	0.69	0.75
VT07							0.93	0.91	0.88	0.91	0.97	0.90	0.83	0.77	0.90	0.80	0.83
VT08								0.90	0.86	0.89	0.93	0.84	0.85	0.74	0.89	0.75	0.81
VT09									0.93	0.90	0.95	0.91	0.84	0.75	0.91	0.81	0.75
VT10										0.92	0.93	0.87	0.88	0.79	0.90	0.84	0.70
VT11											0.95	0.84	0.78	0.67	0.82	0.68	0.63
VT12												0.90	0.80	0.73	0.92	0.77	0.78
SF1													0.83	0.88	0.96	0.92	0.89
SF2														0.83	0.88	0.90	0.75
SF3															0.87	0.97	0.84
SM1																0.92	0.90
SM2																	0.87

Figure 3.6: Correlations between the first principal components of speakers' area functions

PC2	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	0.81	0.79	0.17	0.81	0.80	0.05	0.71	0.79	0.70	0.83	0.73	0.51	0.41	0.64	0.70	0.58	0.61
VT02		0.93	0.53	0.90	0.67	0.34	0.87	0.79	0.91	0.91	0.94	0.46	0.40	0.50	0.80	0.69	0.56
VT03			0.56	0.94	0.64	0.33	0.93	0.83	0.97	0.88	0.92	0.63	0.43	0.68	0.82	0.70	0.65
VT04				0.60	0.11	0.78	0.55	0.36	0.61	0.63	0.65	0.01	0.37	0.14	0.51	0.17	0.18
VT05					0.59	0.42	0.94	0.79	0.89	0.92	0.94	0.53	0.62	0.70	0.90	0.70	0.65
VT06						0.28	0.47	0.73	0.52	0.58	0.55	0.67	0.04	0.47	0.50	0.68	0.42
VT07							0.42	0.08	0.30	0.39	0.41	0.07	0.51	0.11	0.51	0.27	0.41
VT08								0.64	0.92	0.84	0.88	0.43	0.64	0.63	0.81	0.69	0.62
VT09									0.80	0.78	0.77	0.74	0.25	0.74	0.71	0.67	0.61
VT10										0.86	0.91	0.53	0.42	0.62	0.74	0.63	0.54
VT11											0.89	0.33	0.42	0.48	0.74	0.51	0.45
VT12												0.49	0.51	0.59	0.89	0.68	0.56
SF1													0.13	0.82	0.62	0.78	0.74
SF2														0.61	0.68	0.66	0.60
SF3															0.77	0.78	0.85
SM1																0.87	0.82
SM2																	0.90

Figure 3.7: Correlations between the second principal components of speakers' area functions.

[4] For details about each speaker, see participants list in Appendix A.

Effects of age

Being in the same age group appeared to have no effect on the correlation between one's area function principal components. However, research suggests that aging may have an effect on the vocal tract cavities [46], such as enlargement of oral cavity volume [47] and lesser muscle tone causing collapse of the pharyngeal cavity as participants lie supine in the MRI scanner [23]. In order to further explore at any characteristics specific to the senior speaker group (VT01, VT02, VT11; the GenAm did not have any speakers in the 'senior' category), the same inter-speaker correlation analysis was performed the oral and pharyngeal regions separately.

Overall, there were no age-related trends observed. Correlations were neither consistently higher nor lower within age groups, for either cavity, for PC1 (age groups are indicated by the same colour in Figures F.4 and F.5; p-values in Appendix G), PC2 or PC3 (not pictured). There was also little difference between the PC1 correlations for the oral and pharyngeal cavities[5] [6]. For PC2 however, the oral cavity data showed much higher correlations than the pharyngeal, with median correlations of 59% and 23% respectively. Since PC2 tends to encode phonetic height which is mostly related to the oral cavity (jaw opening, tongue lowering), this result is expected.

3.2.5 Intra-speaker PCA correlations

In order to contribute to the data which had two sets of MRI images available, I processed Set 2 of VT11 and added the area functions to the existing pool of 5 speakers who already had two full sets of vowel images each, processed and available for analysis [7, 41]. Intra-speaker correlations were carried out between Set 1 and Set 2 for each speaker. Figure F.6[7] shows that the two sets of area functions for each speaker were mostly well correlated for PC1, PC2, and even PC3 for some speakers, with significant p-values to match[8]. However, for VT10 and VT11, there is low correlation between its two vowel sets for PC2. Between higher order PCs, correlations became varied. When these values were compared against the PC1 intra-speaker correlations in Watson's Interspeech paper, the values were very close, being all very highly correlated (see Figure H.2) [41].

[5] Correlations were only very slightly lower in the pharyngeal cavity data. The average correlation between oral cavity data was 86%, compared to 81% for pharyngeal data (medians of 90% and 84% respectively)

[6] VT10 appears to have significantly lower correlations with the other speakers in the oral cavity. This may be attributed to the fact that during MRI acquisition for VT10, a phonetician was not available to supervise, which may have resulted in slightly different vocal tract shapes. These sorts of anomalies are likely to show up in the oral cavity where the tongue has the greatest freedom, rather than the pharyngeal cavity.

[7] The alternative 'SP' speaker label refers to the labels in Watson [41] for easy of comparison.

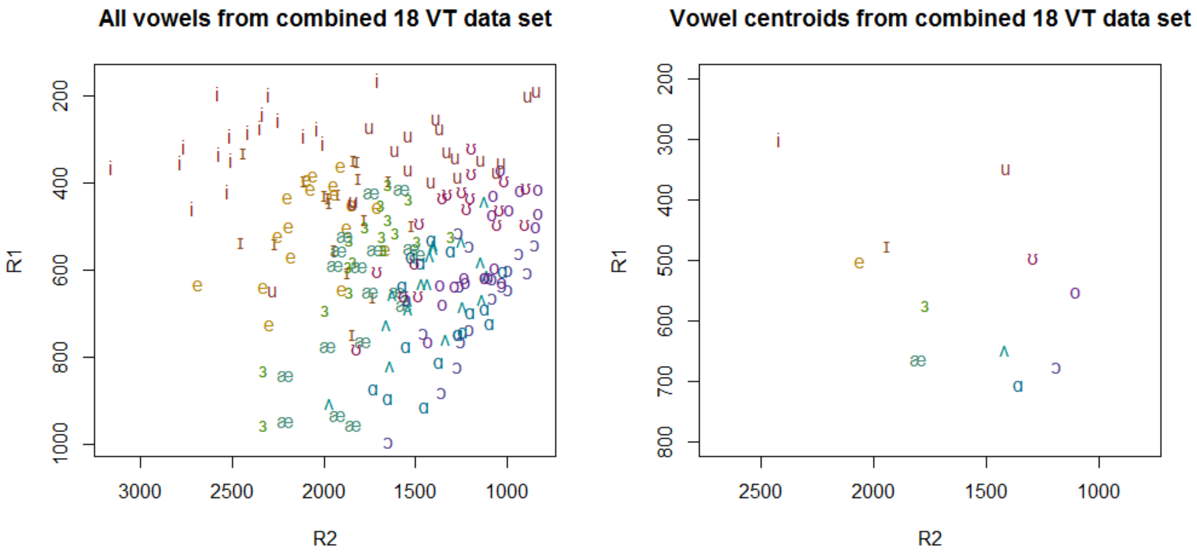
[8] See Appendix G for colour bars on the p-value colour-coding

3.3 Resonance Analysis

3.3.1 Comparison with Story (2005) formants

In order to validate my method of calculating the resonances from area functions, I compared a plot of the vowel centroids on an R1-R2 plane (including only the GenAm area functions) with plots created from the resonance (called 'calculated formants') and formant ('natural formants') values in Story [33]. Figure F.7 shows very close alignment between the three centroid plots. Since the area functions for my resonances were taken from the Story article, it would theoretically have been possible to match my resonance centroids exactly to coincide with Story's natural formants. However, given the small variables throughout the LPC process from area function to resonance which were not detailed in Story (such as the choice of speed of sound in air when calculating sampling frequencies or the method of identifying peaks in the spectrum), the fact that my R1-R2 plot is so close to Story's is an encouraging validation of the methodology.

3.3.2 Vowel plots on R1-R2 planes and demographic trends

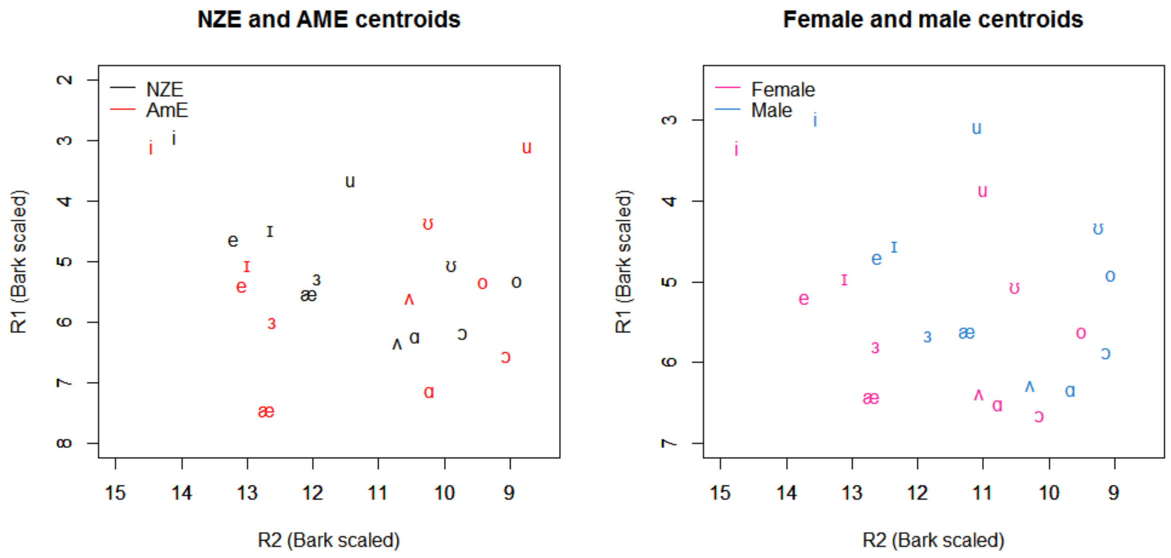


(a) All vowels

(b) Centroids only

Figure 3.8: R1-R2 plots for combined NZE and GenAm data set

Figures 3.8a and 3.8 show all the vowels from all speakers plotted according to their R1 and R2 values. Figure 3.8 shows only the centroids - the average resonance values - for each vowels across all speakers. The resonances appear to have separated the vowels out by their phonetic height and backness, as we see front vowels such as /i, e and æ/ towards the left edge of the plot, and the back vowels /ʊ, o/ further to the right. The highest value for R1 sits well below the lowest value for R2, which is to be expected of the first two resonances. The axes on the following plots were normalised using the Bark scale, a frequency scale on which equal distances correspond with perceptually equal distances, to even out the ranges of R1 and R2 [37]. The range for R1 is much smaller than that of R2, as exemplified in Figure 3.8 where the range of F1 is approximately 800 Hz, while the range of F2 is 2000 Hz. Bark scaling evens this out so values are not disproportionately weighted by F2 when performing 2D analyses such as measuring the Euclidean distance between vowels [43].



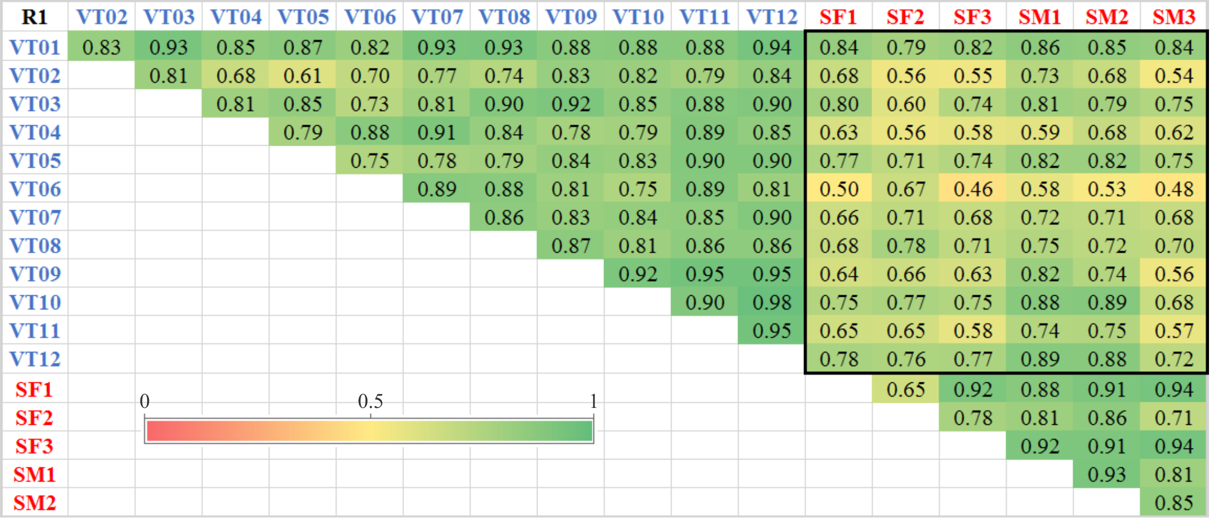
(a) Centroids of NZE and GenAm vowel resonances.(b) Centroids of female and male vowel resonances.

Figure 3.9: R1-R2 plots for combined NZE and GenAm data set

In the plot of vowel centroids divided by accent (Figure 3.9a), we see the ‘who’d’ vowel (u) positioned much further back for GenAm than in NZE. Also, despite our reputation for lowering our ‘hid’ vowels (ɪ, as in ‘fush’n’chups’), the R1-R2 plot shows our ‘hid’ vowel being positions higher than GenAm that of, and even GenAm’s ɜ vowel. Another interesting feature is the way that the NZE ‘herd’ vowel (ɜ) appears to be significantly higher and more back than that of GenAm. As for gender, male speakers tended to have lower first and second resonances (Figure 3.9b; note the flipped, Bark-scaled x and y axes). Both plots show an overall resemblance to the classic vowel quadrilateral.

3.3.3 Inter-speaker resonance correlations

R1 and R2 showed very strong correlations across the board, which one would expect since the speakers are saying the same vowels. R3 shows a mixture of weak to strong correlations, but in the inter-accent region (black border), most of the correlations are very weak. In the R1 correlation table, we can see that correlations also do decrease between different accents, and much more so than in R2.



R1	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	0.83	0.93	0.85	0.87	0.82	0.93	0.93	0.88	0.88	0.88	0.94	0.84	0.79	0.82	0.86	0.85	0.84
VT02		0.81	0.68	0.61	0.70	0.77	0.74	0.83	0.82	0.79	0.84	0.68	0.56	0.55	0.73	0.68	0.54
VT03			0.81	0.85	0.73	0.81	0.90	0.92	0.85	0.88	0.90	0.80	0.60	0.74	0.81	0.79	0.75
VT04				0.79	0.88	0.91	0.84	0.78	0.79	0.89	0.85	0.63	0.56	0.58	0.59	0.68	0.62
VT05					0.75	0.78	0.79	0.84	0.83	0.90	0.90	0.77	0.71	0.74	0.82	0.82	0.75
VT06						0.89	0.88	0.81	0.75	0.89	0.81	0.50	0.67	0.46	0.58	0.53	0.48
VT07							0.86	0.83	0.84	0.85	0.90	0.66	0.71	0.68	0.72	0.71	0.68
VT08								0.87	0.81	0.86	0.86	0.68	0.78	0.71	0.75	0.72	0.70
VT09									0.92	0.95	0.95	0.64	0.66	0.63	0.82	0.74	0.56
VT10										0.90	0.98	0.75	0.77	0.75	0.88	0.89	0.68
VT11											0.95	0.65	0.65	0.58	0.74	0.75	0.57
VT12												0.78	0.76	0.77	0.89	0.88	0.72
SF1													0.65	0.92	0.88	0.91	0.94
SF2														0.78	0.81	0.86	0.71
SF3															0.92	0.91	0.94
SM1																0.93	0.81
SM2																	0.85

Figure 3.10: Correlations between the first resonant frequencies of speakers' area functions

Chapter 4

Discussion

One of the overall aims of this project was to assess the validity of analysing MRI-derived vocal tract data together as a combined data set, and to see the extent to which group trends such as the effects of age, gender and accent could be observed from the principal component and linear predictive coding resonance techniques employed. The validity of these methods is important as it has the potential to provide us with a novel method of studying vowel sounds, particularly with individual speaker differences eliminated. It became clear that the reliability of any conclusions drawn from these MRI-based analyses would be dependent on the the quality of our MRI data and area function extraction process. Consequently, adding in extra analyses to further explore and to validate findings, particularly quirks and irregularities in the data, became routine.

Once again, to clarify abbreviations, all speakers will be referred to by their speaker labels (VT for New Zealand English, NZE; SP for General American English, GenAm) to preserve anonymity. PC refers to principal component, while R (as in R1, R2) refers to resonances.

4.1 Mean Area Functions

Figure 3.1 demonstrates how our area functions can effectively capture vowel quality. High vowels are seen with a significant peak in cross-sectional area in the pharyngeal region, while low vowels peak in the oral cavity on average. This reflects how high vowels require the tongue to be raised high in the mouth, narrowing the gap between the tongue and the roof of the mouth and therefore reducing the cross-sectional area in that region. As a consequence of the tongue lifting up (and usually also forward slightly), the tongue back and root are pulled up and forward, increasing the space in the pharyngeal cavity which is seen as the peak in the pharyngeal region. The opposite is true for low vowels: the tongue is flattened to the floor of the mouth as the jaw drops, increasing the volume of the oral cavity and decreasing the volume of the pharyngeal cavity as the tongue positions itself further towards the back of the throat.

The mean area function for GenAm showed a higher oral cavity peak when compared with NZE (Figure 3.2). Front vowels have smaller oral cavities and larger pharyngeal cavities (the opposite is true for back vowels), so this may be a reflection of how NZE vowels tend to be more fronted than GenAm vowels [add references]. This was also seen in plots of NZE and GenAm resonance centroids (Figure 3.9a, where the vowels which would normally be considered ‘back’ or ‘low’ such as /u/ or /æ/ appear to be shifted to a higher, more front position. Thus, it follows that the average NZE vowel is higher/more front than that of GenAm, and therefore gives a mean area function that reflects the smaller cross-sectional area in the oral cavity.

In Figure 3.3, it was at first surprising to see that the mean female area function had higher cross-sectional areas throughout the majority of the vocal tract compared to men. Median area functions were plotted to see whether any outlying area values were skewing the averages, but the resulting plot was very similar. However, when VT12 was removed from the female data set, a dramatic drop in the mean area functions for females was seen. VT12, being an opera singer, had outstandingly varied area functions (see Figure D.1) and had enunciated the vowels very clearly during MRI acquisition. Exploratory PC1-PC2 plots of vowel centroids separating out male and female speakers for NZE and GenAm were also made, which showed that female speakers, particular speakers of GenAm, had overall higher vowels than men (not pictured). This is reflected in the drop in cross-sectional area at the front of the oral cavity for women, as higher vowels take up more of the space in the oral cavity.

As for the age groups (Figure 3.4), it may be that the larger pharyngeal cavity for the two older age groups is supporting the studies which do suggest widening of the pharyngeal cavity with age [47] [find some other references for this]. However for this plot, it is likely that these mean area functions are not reliable representations which we may draw generalisations from, since the age groups were not fairly balanced. For example, the 55+ years group only consisted of 3 speakers, all of whom were NZE speakers. The overall higher areas in the middle-aged group may also be attribute to the fact that of the six speakers included, 4 were male who are likely to have larger vocal tracts, and one of the females an opera singer whose area functions had relatively higher amplitudes. On the other hand, the 20 to 30 years group had more females than males. These unbalanced factors are likely to have an effect on the overall area function means which are produced. Fortunately, in the cases of gender and of accent, the participants’ pool had a fairly even divide of the other two factors so that group differences could be cancelled out[1].

Despite mean area functions being a crude measure of vocal tract shape, these gross demographic trends [Also link to results from other analyses to see if they match up]. Also, the two-peak shape of every plot with a small dip at the second peak is a clear reflection of the vocal tract anatomy. It is likely that the first peak drops down to a local minima at the uvula, then expands again into the pharyngeal cavity, drops slightly again to accommodate the epiglottis, then expands again slightly before closing at the glottis.

[1] We tried normalising the area functions by dividing by the maximum cross-sectional area of each speaker, in an attempt to eliminate vocal tract size differences (vocal tract length has already been normalised as the areas have been plotted with indices as x-values) to uncover other trends. However, virtually no difference to the trends was seen.

4.2 Vowel Plots and Correlations from PCA

4.2.1 Vowel plots on PC1-PC2 planes

The NZE vowel plots on the PC1-PC2 planes showed fairly good separation of the individual vowels by height and backness. The way that points of the same vowel clustered together in Figure F.1 shows that the PCA is able to extract information on vowel quality from MRI-derived area functions, even from a combined, diverse data set that blends data from different speakers which has been processed by different researchers[2]. Furthermore, given our centroid plot's close resemblance to the centroid plot from Watson and the standard vowel quadrilateral (once the x-axis has been flipped) (Figure 3.5b), the addition of 7 new speakers to Watson's data set did not hinder the PCA's ability to extract meaningful vowel height and backness information from MRI images. This was supported by the fact that despite throwing 7 new speakers, with a variety of ages and genders, into the mix, the proportions of variance accounted for by the first two principal components barely dropped at all (Table 2.1), indicating that the principal components are robustly capturing meaningful vowel-related information. The data on which these results are based was validated by checking that the majority of the variances in each speaker's data set were being accounted for by the first and second principal components. Despite reservations about the quality of the VT04 and VT07 data, the results of this verification showed that there was no significant difference between the area functions of these two speakers and the rest of the data set (Figure F.3), validating our choice to include them in the combined speaker analyses.

4.2.2 Inter-speaker PCA correlations and demographic trends

A correlation analysis between the first principal components of the speakers also showed very strong evidence ($p < 0.01$) for strong correlation (above 70%) between individual speakers (Figure 3.6). Other than providing powerful evidence that there is little random noise varying the area data derived from these MRI images, it also supports the notion that PC1 and PC2 are suitable surrogates for phonetic backness and height. This is further supported by the fact that the aforementioned PC1-PC2 centroid plot (Figure 3.5a) which separated out the vowels by their height and backness quite well. The slightly lower correlations for PC2 may be related to PC2's lesser ability to separate out the vowels by their heights in the centroid plots.

Again, little difference was seen between VT04 and VT07's correlations and those of the other speakers, except slightly weaker correlations in PC2[3]. This further supports the hypothesis that PCA is able to capture vowel qualities, namely phonetic height and backness, while ignoring individual speaker differences. This opens up the door to analysing changes in vowel quality

[2] P1-P2 plots for the GenAm area functions were not able to be made due to the time constraints of the project. However, a number of comparisons were made between the GenAm and NZE data sets in the form of PC and resonance correlations as described in the following sections.

[3] These are likely due to the participant speaking with relatively subtle mouth movements during the MRI acquisition, as noted by Watson and evidenced by the smaller amplitudes in their area plots (Appendix C).

using combined, bulk data sets as opposed to analysing each speaker as a case study, as has been the case for much of the area function-based vowel studies to date. It allows us to uncover general trends within populations, using PC1 and PC2 as proxies for vowel height and backness.

Accepting that PC1 and PC2 encode vowel height and backness, the lack of any age and gender trends in the inter-speaker correlations was to be expected. There is no reason for there to be a difference between the vowels of males and females; speakers of different gender with the same accent should articulate their vowels in the same manner, and thus have the same vocal tract shapes. As for speakers of different ages, it was speculated that differences may be seen in the PCA correlations of the individual oral and pharyngeal cavities. However, our inconclusive results now add to the body of literature that is conflicted on the effects of aging on the oral and pharyngeal [46, 47, 23]. Sociolinguistically, it may be possible that older speakers have more British-sounding accents as a result of immigration and/or the vowel changes which have occurred to NZE over the past century [10, 38, 42]. But given that even differences between GenAm and NZE were barely captured in the inter-speaker PCA correlations, it would be highly unlikely for this method to pick up on the lingering remnants of past British accents in NZE speakers.

On that note, slightly lower correlations were seen between speakers from accent groups compared to correlations within accent groups. However, this may be attributed to the inevitable small differences in the methods of image processing and area function extraction used by us and by Story [35, 33][4].

From the extremely low correlation estimates seen for PC3, it appears that noise dominates the higher order principal components. However, as discovered while quantifying the variability between users who process the MRI images (section 2.2), the method of deriving this data itself is holds at least this much noise due to the manual, approximate way that the cross-sectional areas are defined. Therefore, it is not clear at this stage whether there are more dimensions of vowel or speaker difference to be discovered in the higher order principal components upon elimination of this variability.

4.2.3 Intra-speaker PCA correlations

Intra-speaker correlations are another way of validating the area function data and the assumptions of PC1 and PC2 encoding vowel properties. By checking that the first and second PCs of a particular speaker correlate, we confirm that the variance captured by PC1 and PC2 is fairly consistent and repeatable. Intra-speaker correlations might also provide insight into individual speaker effects. Given that PC3 generally does not correlate *between* speakers (Figure I.1), any correlations seen in the higher orders *within* speakers may hint towards properties in the area

[4] This may be validated by obtaining MRI data from one GenAm speaker, processing the MRI data using the same methods that we used, and seeing whether their inter-speaker correlations with Story's speakers show lower correlations. However, due to the potential for individual speaker differences, it would take several GenAm speakers contributing to know that any drops in correlation within the accent group are due to method differences rather than individual speaker differences.

functions that are specific to each speaker.

The more repetitions we have available of each speaker, the stronger the results of this analysis can be. Therefore, it was regrettable that I was not able to include the available repetition sets for VT04, VT06 and VT07 due to time constraints - this would be recommended if one wishes to use this analysis in the future. Also, though perhaps difficult and time-consuming to acquire, more MRI image sets for the same speaker would help with seeing whether individual speaker differences are encoded in the higher order principal components. With only two sets each, it is currently difficult to say whether higher order correlations or lack thereof are a product of chance or an underlying effect.

The weaker intra-speaker correlations in VT10 and VT11 are likely to be due to genuine differences in the vocal tract shape of the speaker between the two sets, rather than errors introduced in the image processing (Figure F.6). The quantification of variability between users in Section 2.2.1 showed that the process of extracting area functions from MRI images is extremely repeatable even for different users and in the presence of imaging artifacts. It is far more likely that the speaker simply articulated the vowels slightly differently a second time around.

To explore whether this imaging artifact was a factor in the lower correlations for VT11, we isolated the pharyngeal region and performed the same intra-speaker correlation analysis. We had expected the intra-speaker correlations for VT11 to increase if we only analysed the pharyngeal region, since only the oral cavity was affected by the imaging artifact. Instead, we found that isolating the pharyngeal region mostly decreased intra-speaker correlations (Figure I.2). To check this whether this was simply due to the smaller size of the truncated area functions amplifying the effect of noise, the oral region was also isolated (Figure I.3). PC1 and PC2 correlations were mostly just as high as the full area functions, with the exception of VT10 which dropped in its PC1 correlation significantly in the oral cavity. This would suggest that the slight decrease in correlation when only the pharyngeal cavity was analysed is not the result of a smaller data set, but truly reflective of the variation in the pharyngeal cavity (or the image processing of the pharyngeal cavity), counter to expectations.

Alternatively, it may be that the area function of the pharyngeal region is only able to capture one dimension along which vowel property may change, such as backness. This would make anatomical sense, since the pharyngeal cavity can only really narrow as the tongue's back and root shift forward and back; there are no other dimensions along which the pharyngeal cavity's area function can change. Conversely, the size and shape of the oral cavity can encode backness and height due to the dexterity of the tongue's tip and front. Thus, PC1 may encode the majority of the variances in the pharyngeal cavity, while PC2 is relegated to the smaller variances like the other higher order PCs. This was supported by a quick analysis of variance. In PCA on just the pharyngeal cavity (for VT11), PC1 and PC2 encoded 61% and 13% of the variance respectively, while for just the oral cavity, these values are 59% and 23%, the variance shared more evenly between PC1 and PC2. Despite this probe not explaining the reason behind VT11's low correlations, these results have shown yet another facet of how PCA on area functions seems to give intuitive results relating to the changes in vocal tract shape.

4.3 Resonance analysis

In the R1-R2 plot of NZE and GenAm (Figure 3.9a), the American /u/ vowel was much further back than the NZE /u/ vowel. In fact, GenAm tends to have vowels which are generally further back than NZE [27]. The GenAm very quadrilateratl, spread vowel space, bounded by the corners vowels /i, ae, ə, u/ is a characteristic difference between American and New Zealand vowel plots [14], while NZE vowel spaces tend to be more triangular with /i, ə, o/ as their cardinal vowels, and /u/ being high-central rather than high-back [41]. It was also noted that the NZE 'herd' vowel was higher and more back than that of GenAm. This may be explained by the way that NZE speakers tends to round their lips when achieving this vowel (the lips pushing out, pouting slightly), while GenAm speakers distinguish this vowel by instead rhoticising (pronouncing the 'r' sound prominently). Lip rounding tends to decrease F1 and F2 [19, p. 191] as it extends the effective length of the vocal tract. The same decreases can be expected for R1 and R2, and this very effect is seen in the 'herd' vowel on the NZE vowel space which is is plotted at lower R1 and R2 values (which is shown as higher and more back on the flipped axes).

In the R1-R2 plots of gender, male speakers had lower average resonance values than the female speakers. It is highly likely that this is the effect of larger vocal tract sizes in men [23, 9], which would decrease the resonant frequencies in the vocal tract cavities. Overall, the strongly separated vowels combined with the vowels' adherence to the classic vowel quadrilateral shape indicated that the resonances calculated from area functions using the linear predictive coding technique are also a reliable measure of vowel quality as well as vocal tract shape, and may also be combined for different speakers of different demographics to observe group trends. However, the trends described here are strictly qualitative. A more rigorous analysis with a larger set of speaker data will be needed to derive reliable, quantitative measurement of trends.

In the inter-speaker resonance correlations, we saw correlations decrease in the inter-accent region more significantly for R1 than for R2. As R1 encodes phonetic height which mostly affects the oral cavity (as discussed in Section 4.2.3), it is likely that this is where most of the accent differences take place, hence showing this difference in R1. In some ways, that difference - the fact that R1 showed bigger drops in correlation between accents than R2 - supports the idea that the drop in correlation does indicate accent differences, rather than method variation. This is because if it had been variation in the methods that caused less correlated area functions for NZE and GenAm, one could argue that R2 should have dropped by the same proportional amount. However, as is clear in Figures 3.10 and F.9, this was not the case. The average correlation value between speakers with the same accent for R1 was 84.8%, while that of R2 was 90.3%. Between speakers with different accents (black border region on correlation tables), the average R1 correlation dropped to 70.4% (a decrease of 16.9%) while R2 correlations only dropped to 83.8% (-7.1%). Therefore, there must have been an additional factor driving the extra decrease in R1 correlations, which may be the differences in oral cavity VT shapes due to accent. As for trends in age and gender, there are no immediate signs of any trends in these demographic groups. It may be that we are reaching the limits of the information we can infer from these correlation analyses.

Chapter 5

Conclusions

Vowel studies using MRI-derived area functions rarely analyse multiple speaker data together, forgoing the possibility of observing group trends in vocal tract shapes. Watson's paper [41] was one of the first studies to combine MRI-derived area function data from speakers of different ages and genders and to distill this down to key phonetic information such as phonetic height and backness with the use of principal component analysis and linear predictive coding. This project further validated this methodology with an expanded 18 speaker data set, including the addition of another accent group. The principal component analysis and linear predictive resonances analysis gave several key results.

- Mean area functions were reflective of the vocal tract shapes expected for demographic and vowel subsets.
- Adding 7 new speakers to the newly combined New Zealand English (NZE) data set did not reduce the variance accounted for by the first and second principal components, suggesting that PC1 and PC2 are capturing vowel quality information rather than individual speaker differences.
- Inter-speaker correlations between first two principal components and first two resonances showed very strong correlations, again supporting the idea that these values are able to capture phonetic height and backness.
- Resonance plots were highly reflective of accent trends between NZE and General American English (GenAm).

In addition, the numerous validation tests performed throughout the project gave mostly positive results, confirming the significance of these observed trends. Overall, MRI-derived area functions are a useful alternative method for analysing vocal tract shape, as they are free from the effects of the glottal filter or lip radiation which are inherent in speech formants. Among other anatomical methods of obtaining vocal tract shape (acoustic pulse reflectometry, X-ray or computed tomography), this MRI process is generally the least obstructive and safest for the participants involved, albeit time-consuming in parts. There is much scope for this methodology to be expanded with different techniques for transforming and interpreting the area functions, or with different forms of vocal tract information such as volume functions or tongue contours.

Appendices

Appendix A

Participant List

Although a sample size of 12 is a considerable size in these kinds of voice studies [35, 33, 40, 44], there were several inconsistencies in the dataset which may have had some minor effects on these results. Firstly, not all of the speakers were supervised by a phonetician during the MRI procedure, who would give corrections on their head position or their articulation of vowels. For VT08 and VT10 who were not supervised, this may have contributed to MRI imaging artifacts (e.g. head not being centred) or inaccurate representations of vowels. In future, these may be eliminated by having one phonetically trained person supervise all participants, and encouraging the completion of full vowel sets.

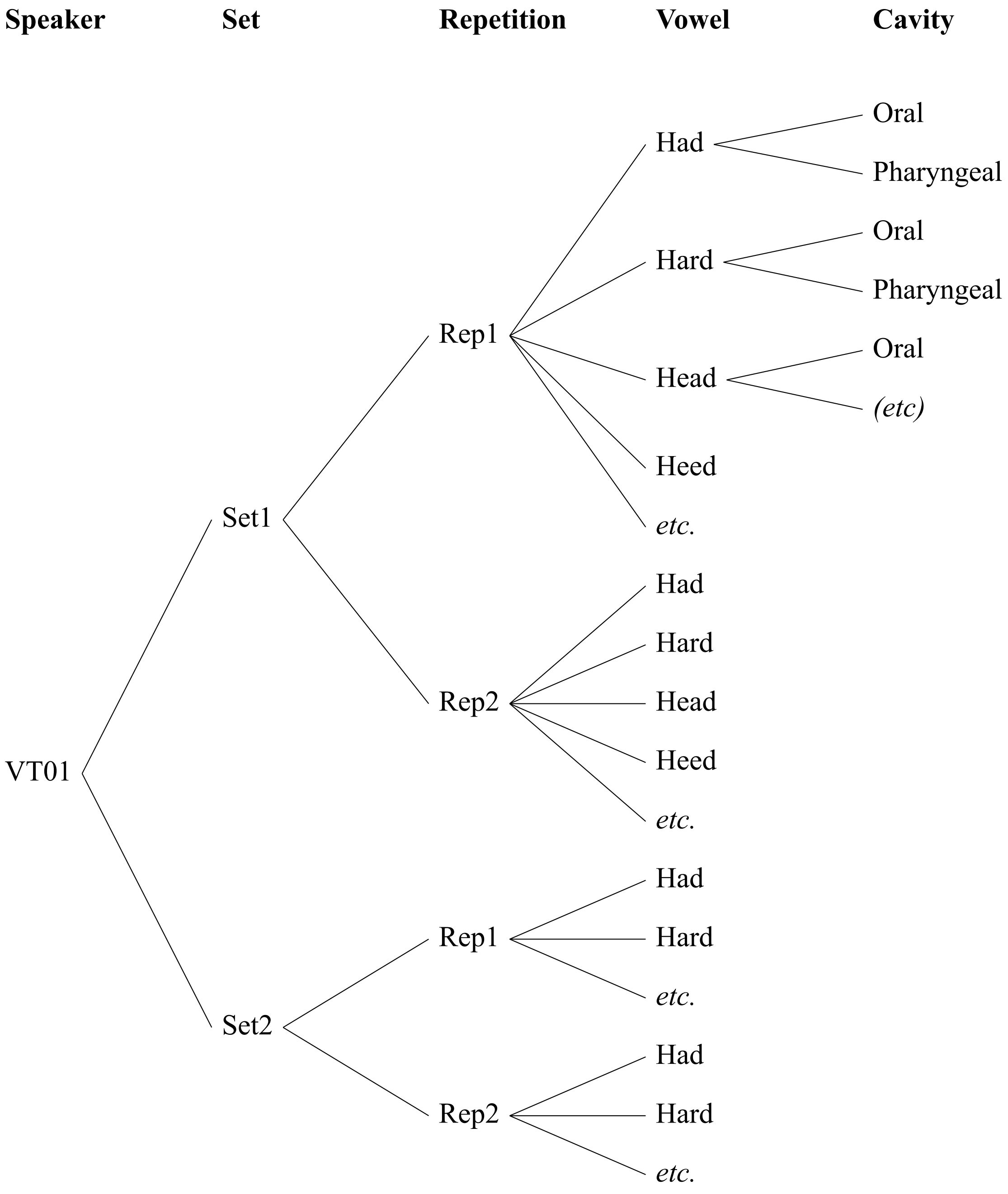
Table A.1: Speakers listed according to their vocal tract numbers used throughout this report (e.g. VT01) and their attributes. The column 'Label in [41]' refers to the labels used in Watson (2014) to refer to the same speaker's data set.

VT	**Date of birth**	**Year of data**	**Age in data**	**Age group**	**Gender**	**Label in [41]**	**Notes**
1	1945	2010	65	senior	m		
2	1945	2010	65	senior	f		Australian origin.
3	1985	2010	25	young	m	SP01	
4	1987	2010	23	young	m		
5	1987	2010	23	young	m	SP02	
6	1985	2010	25	young	f		
7	1986	2010	24	young	f		Spoke unclearly during MRI.
8	1964	2009	45	middle	f	SP03	Not supervised by phonetician during MRI.
9	1964	2009	45	middle	m	SP04	
10	1985	2010	25	young	m	SP05	Not supervised by phonetician during MRI.
11	1938	2005	67	senior	f		Imaging artifacts due to metal plate in mouth.
12	1958	2011	53	middle	f		

Appendix B

MRI Data Hierarchy

This schematic shows the MRI image data provided for **one** speaker, and the labels that we use to refer to each data set. This same hierarchy would be repeated for all twelve speakers in the NZE data set. In words, each speaker had up to two sets of data. In each set would have up to two repetitions, where the same MRI images were processed at a different time, and therefore produced slightly different area functions due to the human error introduced in the manual process of marking up the MRI images (Section 2.1.2). Each repetition has up to 11 vowels, and each vowel is divided into its Oral and Pharyngeal regions. The area function process described in Section 2.1 is repeated for each of these cavities. Some of the speakers only had one set of MRI images available, and only VT04 Set 1 and VT11 Set 1 had their area function analysis repeated twice (with only the four cardinal vowels, Had, Heed, Hod, and Who'd).


Speaker
Set
Repetition
Vowel
Cavity
Had
Oral
Pharyngeal
Hard
Oral
Pharyngeal
Rep1
Head
Oral
(etc)
Heed
Set1
etc.
Had
Hard
Rep2
Head
VT01
Heed
etc.
Had
Rep1
Hard
etc.
Set2
Had
Rep2
Hard
etc.

Appendix C

VT04 and VT07 Area Function Plots

Area functions for all 11 vowels in the VT04 and VT07 data sets have been plotted in each frame. For each area function (the smaller frames within), the x-axis indicates vocal tract length (i.e. the distance from the lips) and the y-axis shows the cross-sectional area, both in millimetres. A vertical blue line indicates the boundary between the oral and pharyngeal cavities. The corresponding mid-saggital MRI image from which the area function was derived is shown to the right of each area function. These pairs are placed in the positions that the vowels would normally be situated on a vowel quadrilateral (Figure 1.2).

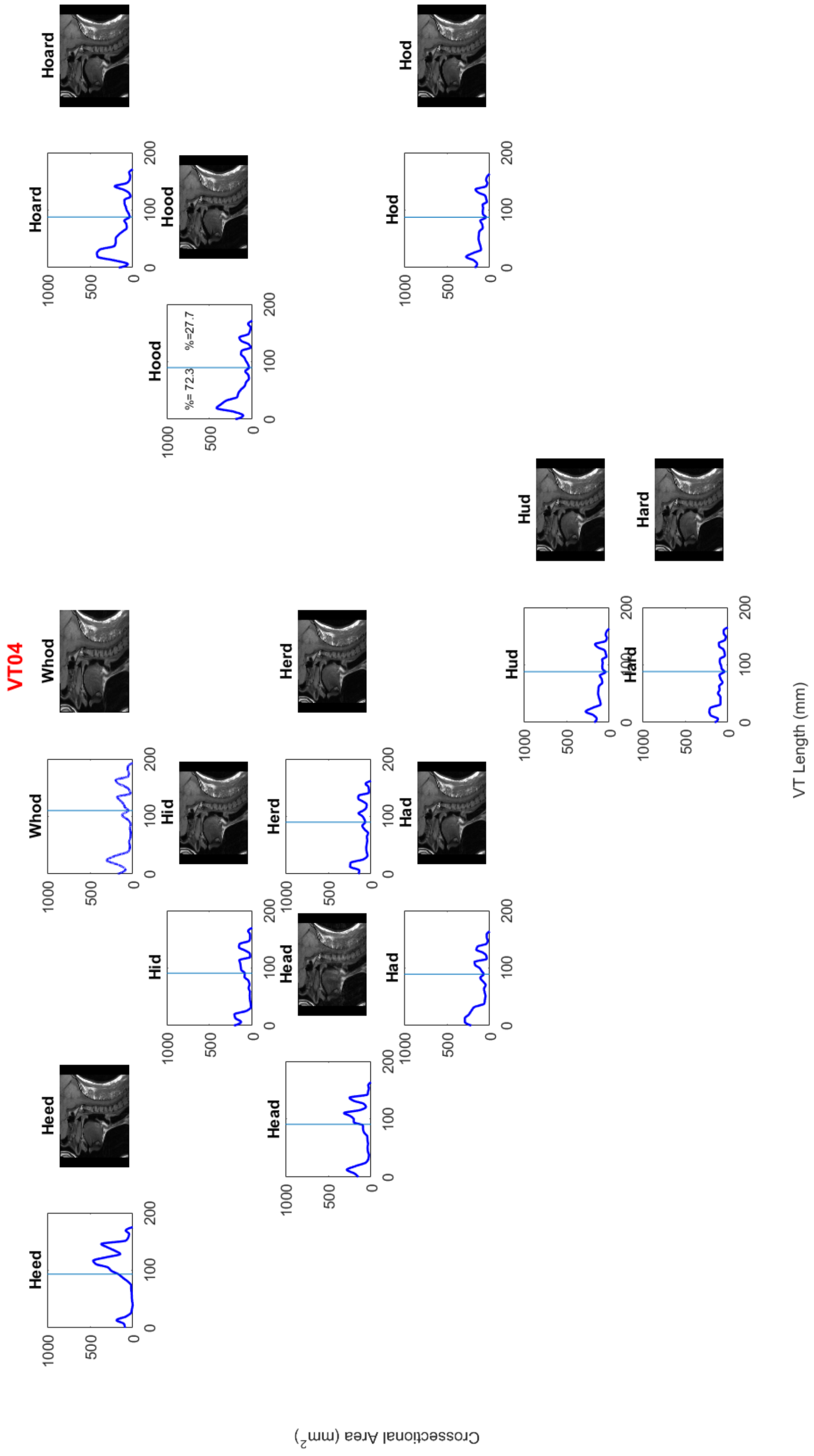


Figure C.1: Area functions of Vocal Tract 4

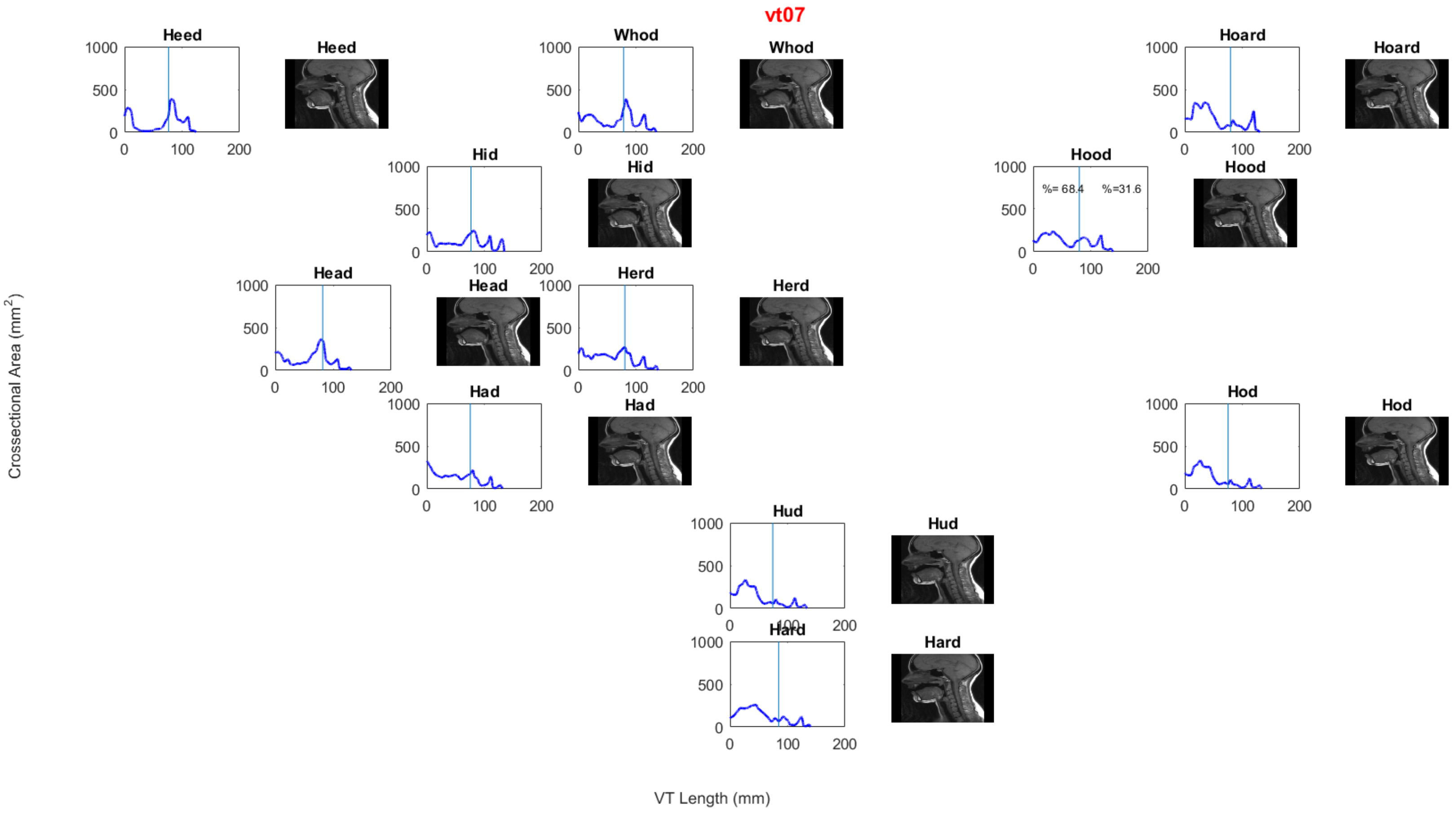


Figure C.2: Area functions of Vocal Tract 7

Appendix D

Area Functions by Speaker and Vowel Type

Figure D.1 shows plots of all the area functions from all 12 NZE speakers. Note the y-axis ranges for these plots, as they are not all the same - clearly, some speakers have much larger cross-sectional areas than others (compare VT12, an opera singer, against VT07 for example). The x-axis is fixed from 0 to 200 mm, which shows the differing vocal tract lengths for the different speakers. Women tend to have shorter vocal tract lengths (proportionate to their overall anatomy), which is shown in female speakers like VT06 and VT07 who had vocal tract lengths of around 165 mm, while male speakers like VT01, VT04 and VT09 had vocal tract lengths which exceeded 200 mm. Figures D.2 and D.3 show plots of only the high (heed, hid, who'd) and low vowels (had, hard, head, hod) respectively. The area functions clearly show

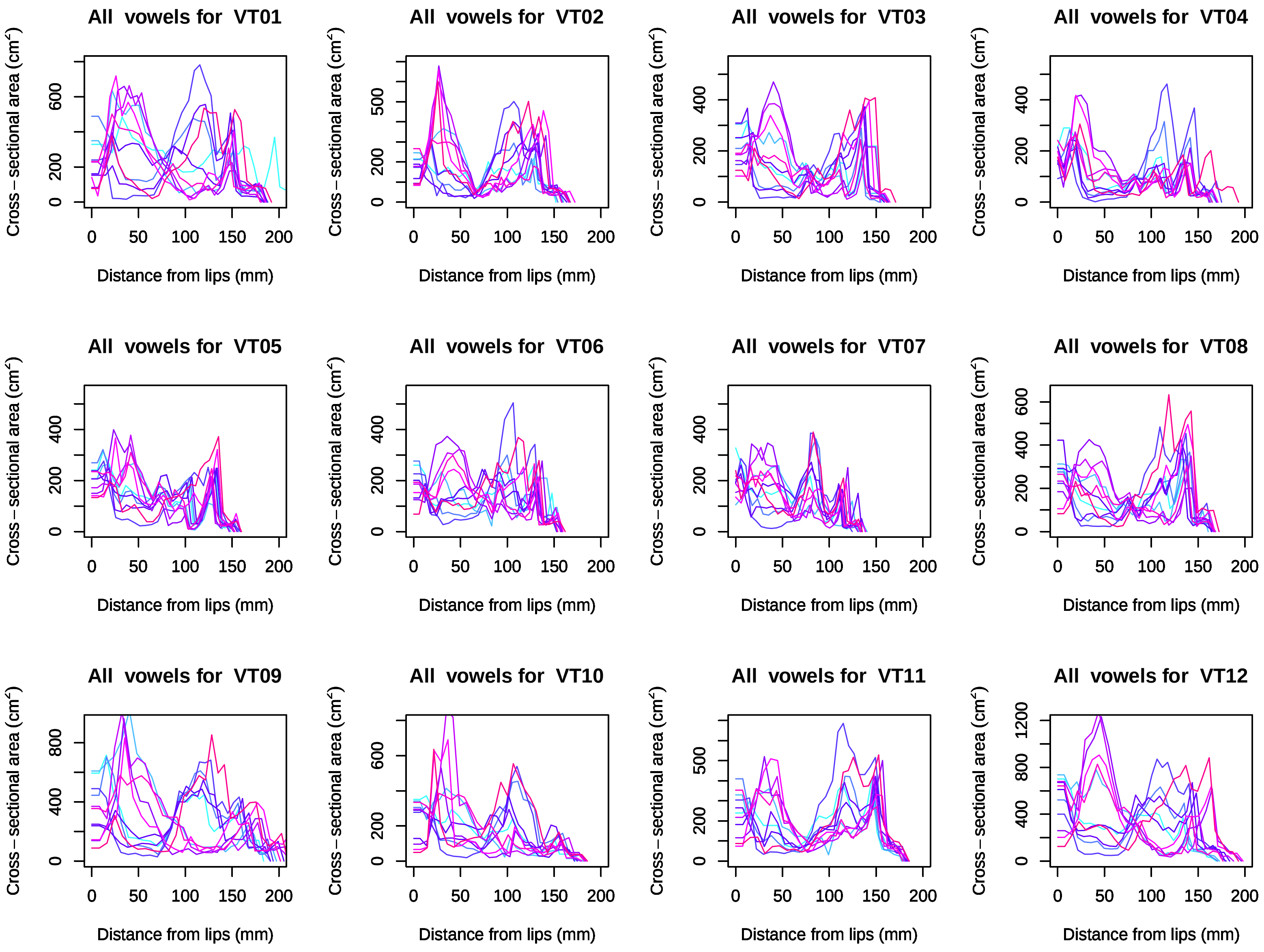


Figure D.1: All area functions for 12 speakers of NZE

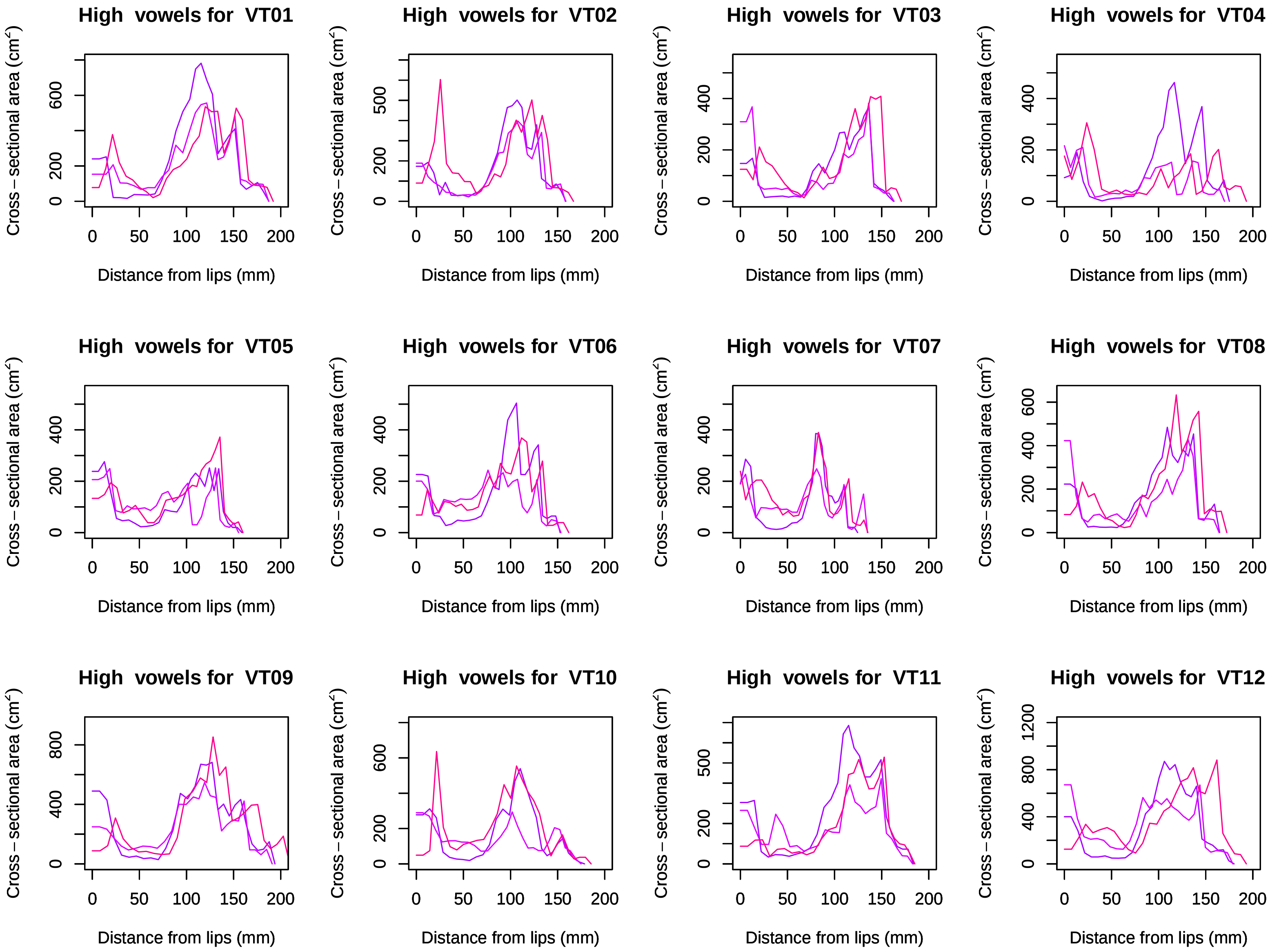


Figure D.2: Area functions of high vowels (heed, hid, who'd) for 12 speakers of NZE

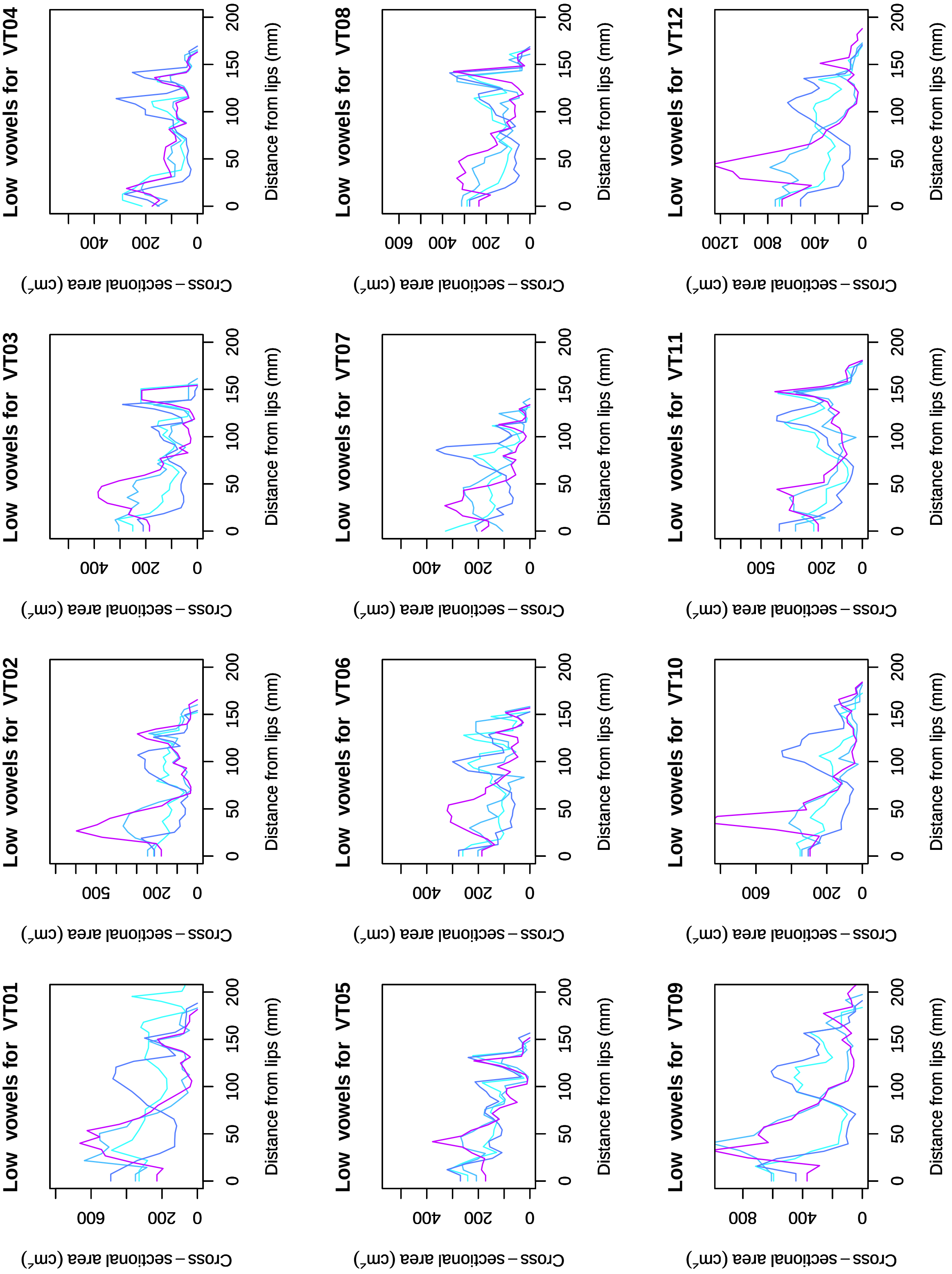


Figure D.3: Area functions of low vowels (had, hard, head, hod) for 12 speakers of NZE

Appendix E

Additional Figures for Validation

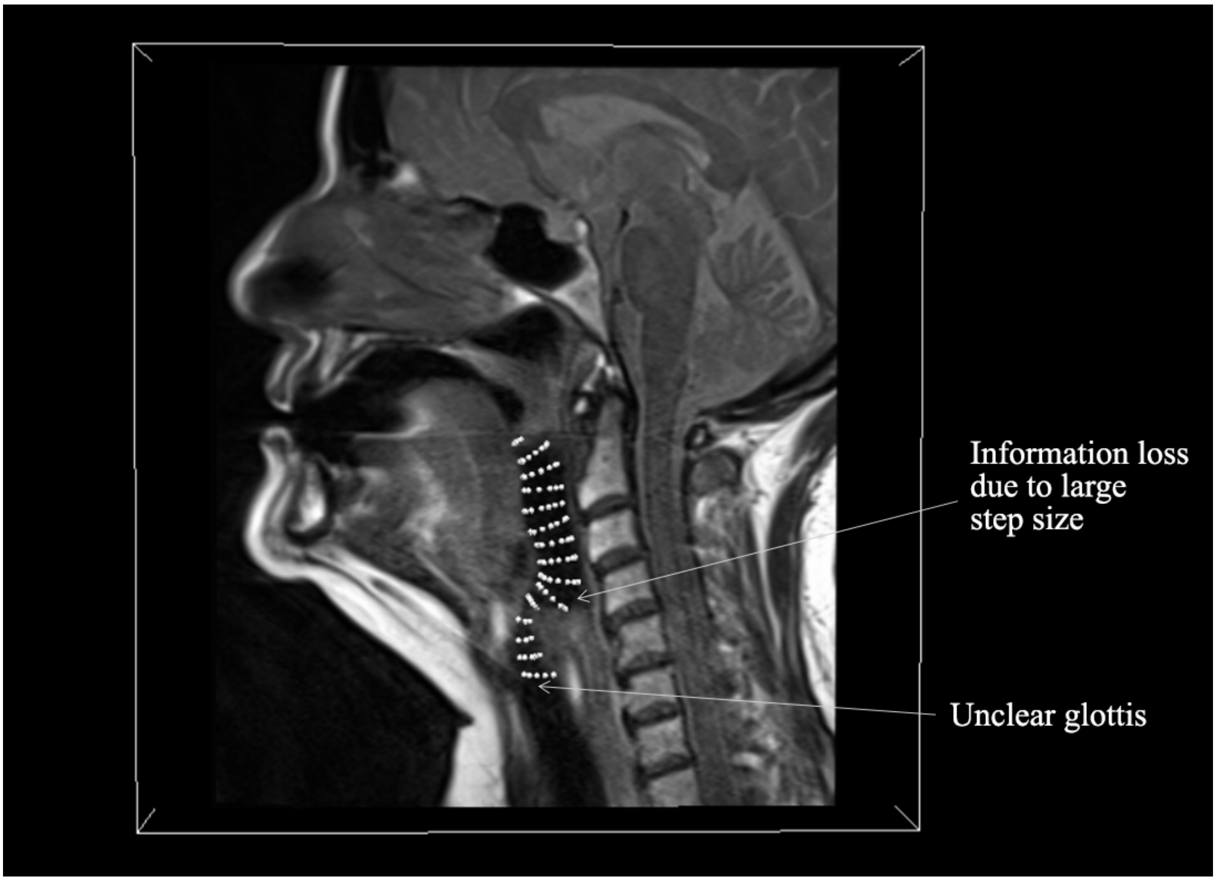


Figure E.1: Mid-saggital section of the vocal tract on CMGUI with data points placed on 15 cross-sectional slices in the pharyngeal cavity.

=

Table E.1: Variability in MRI image processing method between repetitions done by same and different users.

		VT04 Set 1		VT11 Set 2	
Repetition		1	2	1	2
Processed by		Jenny	Jenny	Helen	Jenny
% Variance accounted	PC1	0.68860	0.62363	0.6629	0.7630
	PC2	0.20229	0.24102	0.26746	0.19022
	PC3	0.10911	0.13535	0.06964	0.04677
Correlation (top) & P-value (bottom)	PC1	0.9866		0.9942	
		0.0134		0.0058	
	PC2	0.9915		0.9981	
		0.00852		0.0019	
	PC3	0.9840		0.9949	
		0.0160		0.0051	

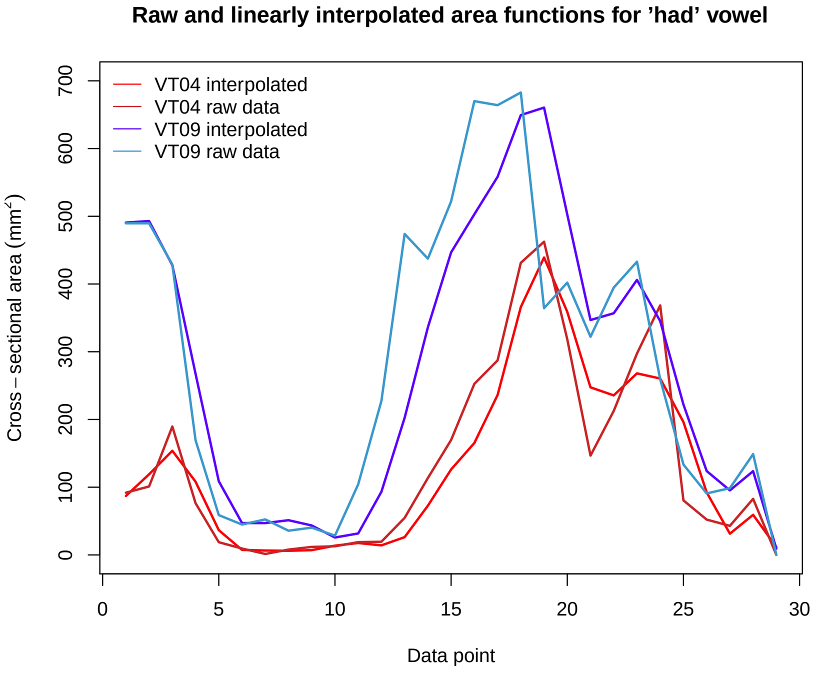


Figure E.2: Raw and interpolated area functions for the VT04 and VT09 ‘had’ vowel.

Appendix F

Additional Results

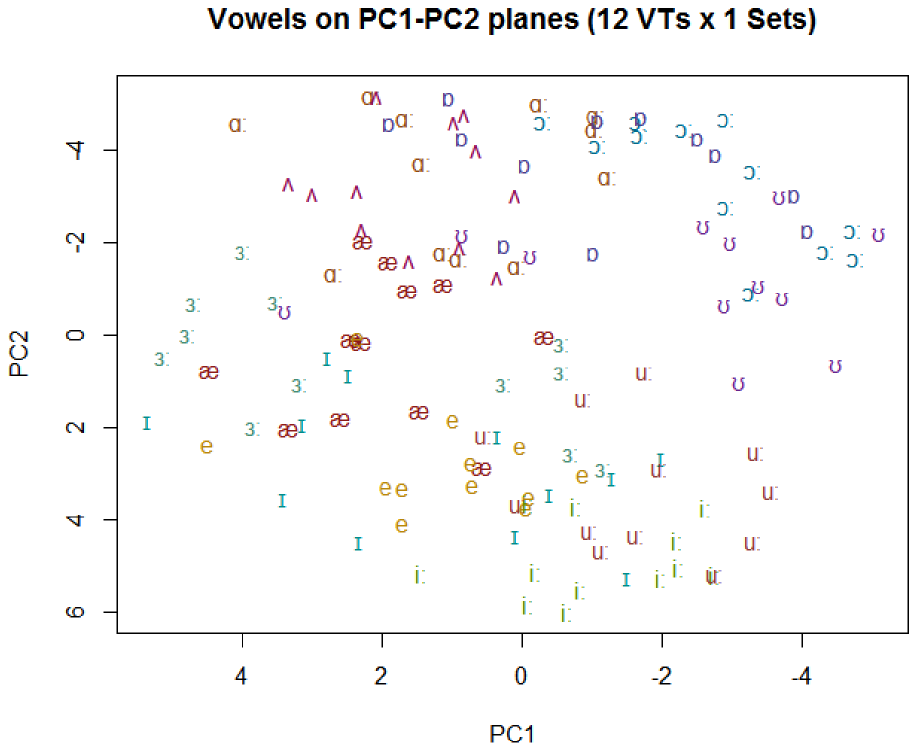


Figure F.1: All vowels for 12 NZE speakers plotted on PC1-PC2 plane.

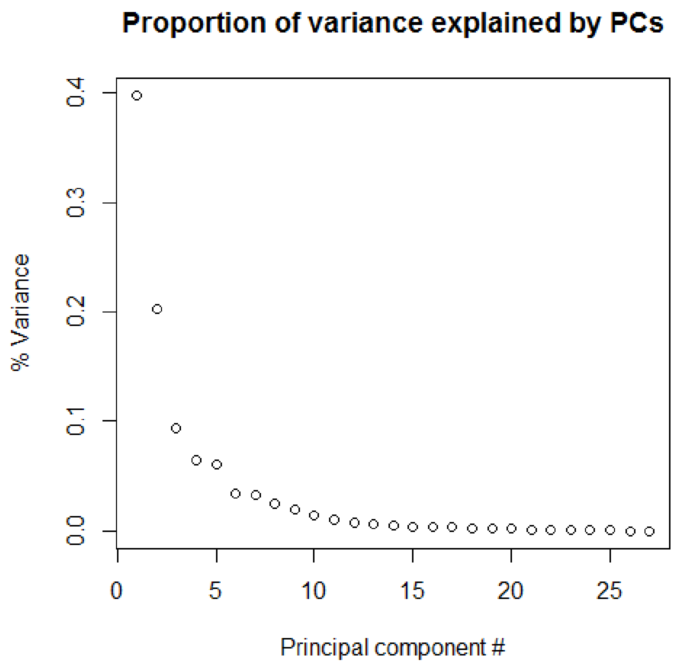


Figure F.2: Variance accounted for by principal components in NZE data set.

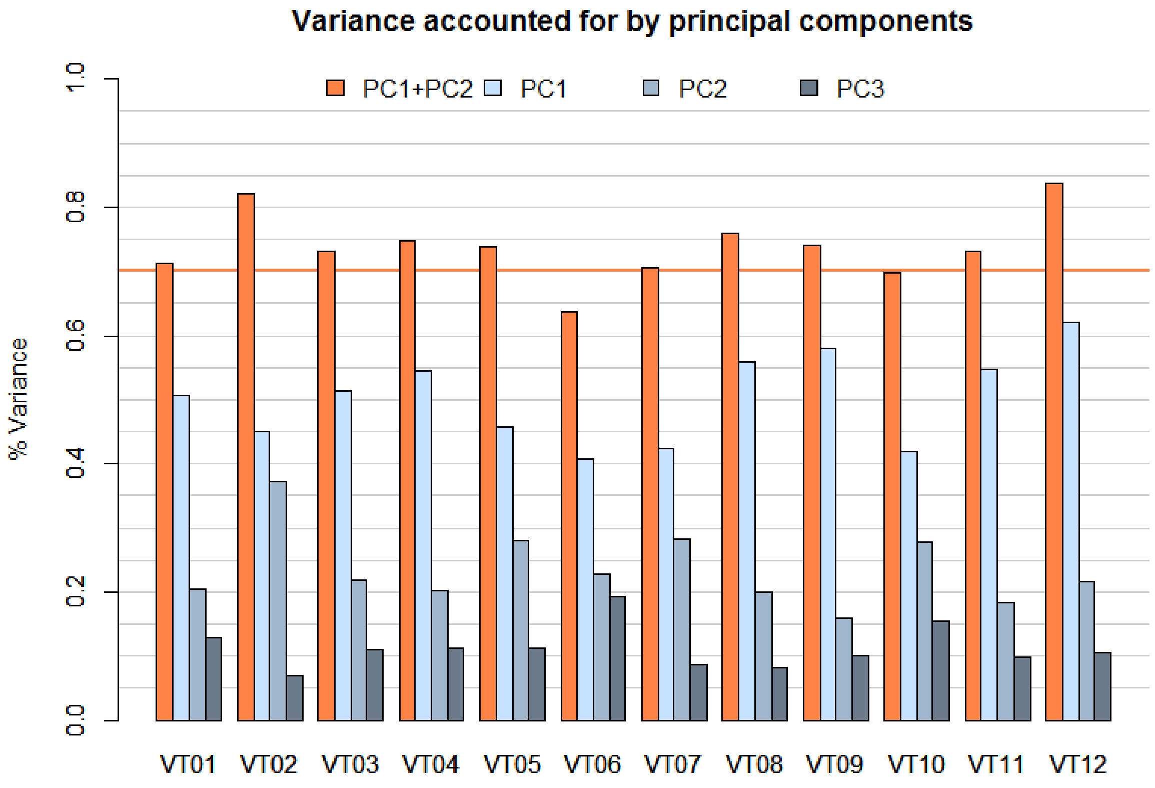
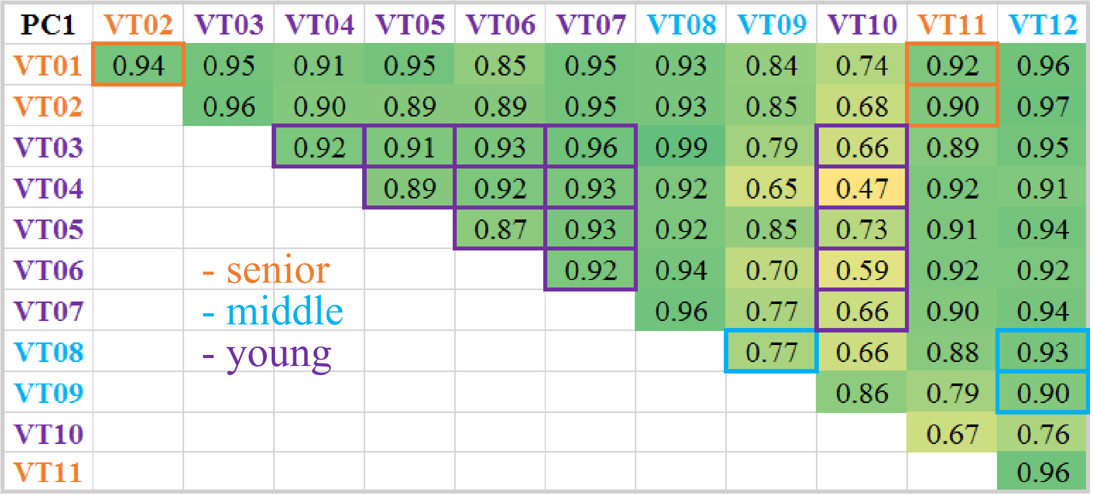
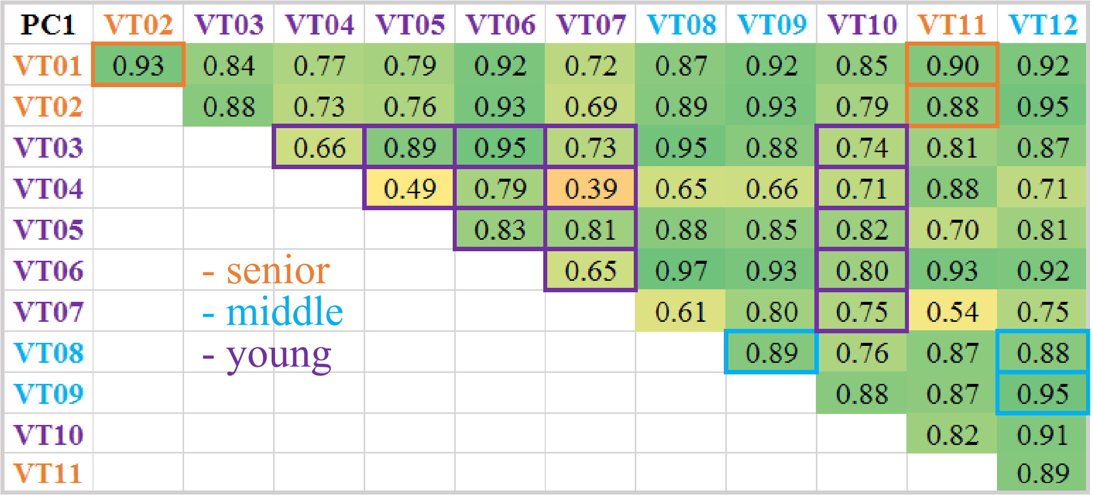


Figure F.3: Percentages of total variance in each speaker's area functions coming from first three principal components.



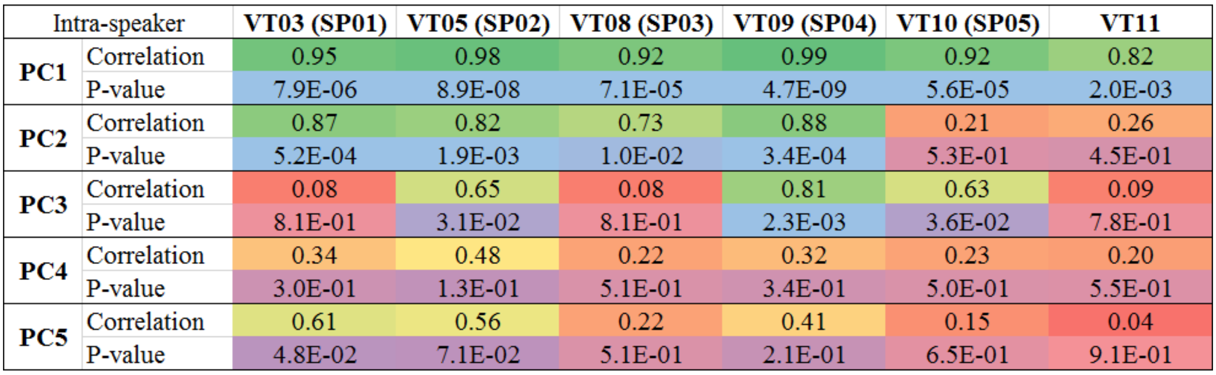
PC1	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12
VT01	0.94	0.95	0.91	0.95	0.85	0.95	0.93	0.84	0.74	0.92	0.96
VT02		0.96	0.90	0.89	0.89	0.95	0.93	0.85	0.68	0.90	0.97
VT03			0.92	0.91	0.93	0.96	0.99	0.79	0.66	0.89	0.95
VT04				0.89	0.92	0.93	0.92	0.65	0.47	0.92	0.91
VT05					0.87	0.93	0.92	0.85	0.73	0.91	0.94
VT06		- senior				0.92	0.94	0.70	0.59	0.92	0.92
VT07		- middle					0.96	0.77	0.66	0.90	0.94
VT08		- young						0.77	0.66	0.88	0.93
VT09									0.86	0.79	0.90
VT10										0.67	0.76
VT11											0.96

Figure F.4: Inter-speaker PC1 correlations between area functions of the oral cavity only.



PC1	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12
VT01	0.93	0.84	0.77	0.79	0.92	0.72	0.87	0.92	0.85	0.90	0.92
VT02		0.88	0.73	0.76	0.93	0.69	0.89	0.93	0.79	0.88	0.95
VT03			0.66	0.89	0.95	0.73	0.95	0.88	0.74	0.81	0.87
VT04				0.49	0.79	0.39	0.65	0.66	0.71	0.88	0.71
VT05					0.83	0.81	0.88	0.85	0.82	0.70	0.81
VT06		- senior				0.65	0.97	0.93	0.80	0.93	0.92
VT07		- middle					0.61	0.80	0.75	0.54	0.75
VT08		- young						0.89	0.76	0.87	0.88
VT09									0.88	0.87	0.95
VT10										0.82	0.91
VT11											0.89

Figure F.5: Inter-speaker PC1 correlations between area functions of the pharyngeal cavity only.



Intra-speaker		VT03 (SP01)	VT05 (SP02)	VT08 (SP03)	VT09 (SP04)	VT10 (SP05)	VT11
PC1	Correlation	0.95	0.98	0.92	0.99	0.92	0.82
	P-value	7.9E-06	8.9E-08	7.1E-05	4.7E-09	5.6E-05	2.0E-03
PC2	Correlation	0.87	0.82	0.73	0.88	0.21	0.26
	P-value	5.2E-04	1.9E-03	1.0E-02	3.4E-04	5.3E-01	4.5E-01
PC3	Correlation	0.08	0.65	0.08	0.81	0.63	0.09
	P-value	8.1E-01	3.1E-02	8.1E-01	2.3E-03	3.6E-02	7.8E-01
PC4	Correlation	0.34	0.48	0.22	0.32	0.23	0.20
	P-value	3.0E-01	1.3E-01	5.1E-01	3.4E-01	5.0E-01	5.5E-01
PC5	Correlation	0.61	0.56	0.22	0.41	0.15	0.04
	P-value	4.8E-02	7.1E-02	5.1E-01	2.1E-01	6.5E-01	9.1E-01

Figure F.6: Intra-speaker correlations for principal components 1 to 5.

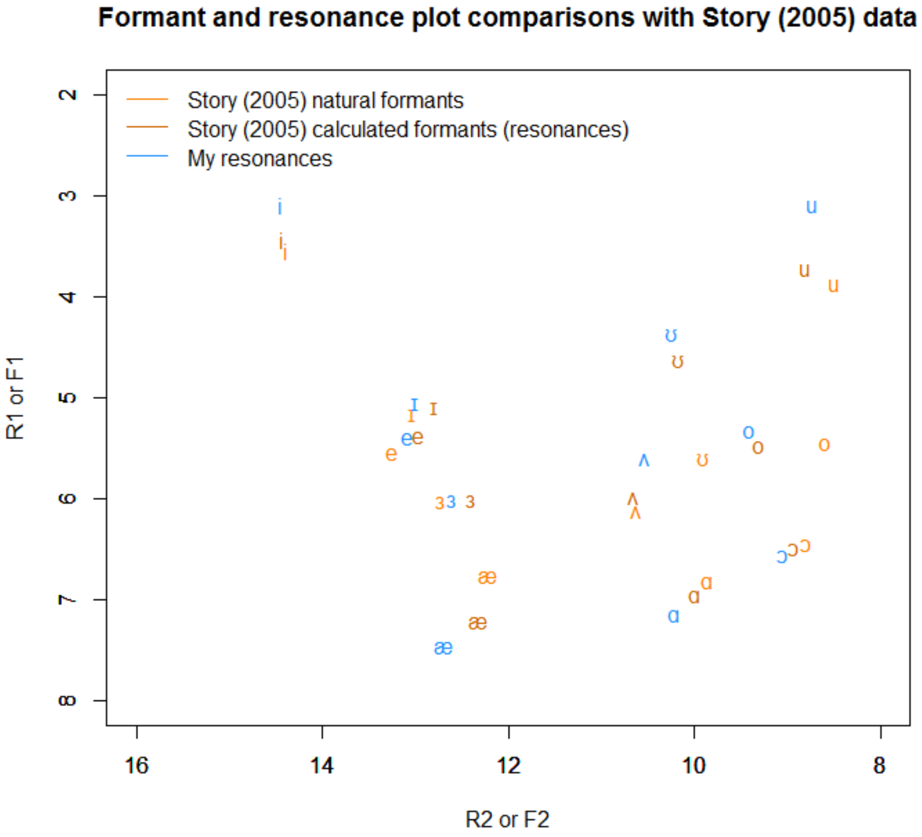


Figure F.7: Centroids plots comparing resonances derived from Story data with resonances and natural formants reported in Story [33].

R3	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	0.25	0.12	0.17	0.14	0.18	0.21	0.32	0.03	0.13	0.03	0.01	0.22	0.02	0.24	0.31	0.39	0.37
VT02		0.54	0.14	0.41	0.62	0.17	0.16	0.40	0.01	0.18	0.40	0.15	0.04	0.36	0.23	0.33	0.27
VT03			0.70	0.67	0.59	0.20	0.75	0.76	0.33	0.72	0.80	0.09	0.05	0.17	0.15	0.08	0.09
VT04				0.33	0.34	0.16	0.58	0.76	0.51	0.45	0.74	0.05	0.20	0.31	0.46	0.32	0.32
VT05					0.73	0.70	0.52	0.57	0.08	0.81	0.74	0.12	0.20	0.16	0.29	0.02	0.02
VT06						0.68	0.64	0.59	0.22	0.60	0.82	0.24	0.33	0.25	0.25	0.10	0.13
VT07							0.26	0.28	0.00	0.46	0.57	0.23	0.15	0.21	0.07	0.08	0.05
VT08								0.53	0.25	0.61	0.79	0.14	0.34	0.33	0.34	0.10	0.16
VT09									0.65	0.70	0.74	0.28	0.29	0.19	0.33	0.08	0.12
VT10										0.54	0.34	0.49	0.18	0.47	0.61	0.22	0.37
VT11											0.74	0.46	0.16	0.00	0.07	0.05	0.35
VT12												0.26	0.22	0.09	0.04	0.17	0.04
SF1													0.25	0.22	0.12	0.60	0.72
SF2														0.48	0.52	0.02	0.15
SF3															0.84	0.05	0.07
SM1																0.39	0.17
SM2																	0.72

Figure F.8: Inter-speaker correlations between the third resonant frequencies of speakers' area functions.

R2	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	0.93	0.86	0.84	0.73	0.85	0.72	0.90	0.96	0.89	0.88	0.94	0.81	0.70	0.82	0.77	0.83	0.78
VT02		0.94	0.93	0.80	0.90	0.76	0.95	0.95	0.88	0.84	0.95	0.91	0.82	0.93	0.90	0.90	0.89
VT03			0.91	0.87	0.89	0.81	0.93	0.90	0.86	0.87	0.89	0.82	0.76	0.87	0.82	0.82	0.84
VT04				0.85	0.93	0.87	0.96	0.91	0.92	0.89	0.96	0.91	0.88	0.94	0.91	0.92	0.92
VT05					0.94	0.92	0.88	0.87	0.77	0.86	0.86	0.69	0.79	0.81	0.70	0.70	0.82
VT06						0.95	0.93	0.94	0.88	0.91	0.95	0.80	0.85	0.85	0.81	0.85	0.86
VT07							0.85	0.85	0.87	0.93	0.86	0.66	0.85	0.75	0.70	0.74	0.78
VT08								0.97	0.91	0.91	0.97	0.91	0.87	0.94	0.91	0.92	0.95
VT09									0.92	0.92	0.99	0.86	0.84	0.90	0.85	0.88	0.89
VT10										0.95	0.94	0.84	0.87	0.86	0.84	0.88	0.85
VT11											0.92	0.73	0.78	0.79	0.73	0.77	0.78
VT12												0.91	0.87	0.94	0.89	0.92	0.92
SF1													0.84	0.96	0.97	0.97	0.95
SF2														0.91	0.91	0.92	0.91
SF3															0.94	0.93	0.96
SM1																0.99	0.96
SM2																	0.95

Figure F.9: Inter-speaker correlations between the second resonant frequencies of speakers' area functions.

Appendix G

Correlation P-values

PC1	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	1E-5	5E-5	4E-4	2E-5	1E-4	4E-6	1E-4	6E-5	1E-4	6E-6	2E-6	2E-4	4E-3	5E-3	4E-4	7E-3	1E-2
VT02		3E-5	1E-3	1E-4	2E-4	6E-6	4E-4	9E-5	9E-4	3E-4	1E-6	3E-5	1E-2	8E-3	4E-5	8E-3	1E-3
VT03			7E-3	2E-6	6E-5	1E-5	1E-6	1E-4	5E-4	7E-4	1E-5	1E-4	2E-3	3E-3	1E-5	4E-3	2E-4
VT04				3E-3	1E-3	3E-4	1E-2	2E-2	1E-2	4E-4	1E-3	7E-3	5E-2	4E-2	2E-2	8E-2	4E-2
VT05					1E-5	1E-5	8E-8	6E-5	5E-4	5E-5	1E-5	4E-4	2E-3	7E-3	3E-4	1E-2	3E-3
VT06						8E-6	5E-6	5E-4	6E-4	3E-5	8E-6	2E-3	2E-3	3E-2	4E-4	3E-2	7E-3
VT07							4E-5	1E-4	3E-4	9E-5	7E-7	2E-4	2E-3	5E-3	1E-4	6E-3	1E-3
VT08								2E-4	7E-4	3E-4	4E-5	1E-3	1E-3	9E-3	2E-4	1E-2	3E-3
VT09									4E-5	2E-4	6E-6	1E-4	1E-3	8E-3	9E-5	4E-3	8E-3
VT10										8E-5	3E-5	5E-4	4E-4	4E-3	2E-4	3E-3	2E-2
VT11											1E-5	1E-3	5E-3	2E-2	2E-3	3E-2	4E-2
VT12												2E-4	3E-3	1E-2	6E-5	1E-2	5E-3
SF1													1E-3	3E-4	2E-6	2E-4	2E-4
SF2														1E-3	4E-4	4E-4	8E-3
SF3															6E-4	4E-6	1E-3
SM1																2E-4	2E-4
SM2																	1E-3

0.001 0.05 1

Figure G.1: P-values for inter-speaker correlations between first principal components.

PC2	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	0.002	0.004	0.616	0.002	0.003	0.887	0.015	0.004	0.017	0.002	0.011	0.107	0.206	0.036	0.016	0.077	0.047
VT02		4E-05	0.094	1E-04	0.023	0.312	5E-04	0.004	1E-04	1E-04	1E-05	0.151	0.227	0.114	0.003	0.029	0.073
VT03			0.073	2E-05	0.034	0.329	4E-05	0.002	9E-07	3E-04	6E-05	0.039	0.187	0.022	0.002	0.025	0.031
VT04				0.052	0.737	0.004	0.077	0.281	0.046	0.036	0.03	0.983	0.265	0.678	0.111	0.642	0.598
VT05					0.058	0.195	2E-05	0.004	2E-04	6E-05	2E-05	0.095	0.041	0.017	1E-04	0.023	0.03
VT06						0.397	0.15	0.01	0.101	0.062	0.078	0.023	0.9	0.141	0.121	0.03	0.197
VT07							0.199	0.822	0.37	0.232	0.215	0.833	0.111	0.757	0.106	0.455	0.216
VT08								0.033	8E-05	0.001	4E-04	0.182	0.035	0.037	0.002	0.027	0.042
VT09									0.003	0.004	0.005	0.009	0.455	0.009	0.014	0.034	0.045
VT10										7E-04	1E-04	0.095	0.193	0.043	0.009	0.049	0.084
VT11											3E-04	0.329	0.198	0.139	0.009	0.129	0.162
VT12												0.125	0.11	0.055	2E-04	0.03	0.072
SF1													0.693	0.002	0.042	0.008	0.009
SF2														0.048	0.021	0.037	0.052
SF3															0.005	0.008	9E-04
SM1																9E-04	0.002
SM2																	5E-04

0.001 0.05 1

Figure G.2: P-values for inter-speaker correlations between second principal components.

PC3	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	0.477	0.529	0.456	0.535	0.468	0.392	0.280	0.651	0.472	0.775	0.290	0.628	0.021	0.987	0.028	0.869	0.166
VT02		0.681	0.351	0.443	0.864	0.944	0.918	0.329	0.059	0.614	0.137	0.350	0.528	0.190	0.020	0.963	0.433
VT03			0.331	0.751	0.703	0.550	0.554	0.290	0.120	0.710	0.277	0.726	0.149	0.346	0.495	0.114	0.461
VT04				0.407	0.342	0.205	0.786	0.090	0.371	0.385	0.539	0.018	0.096	0.241	0.423	0.274	0.355
VT05					0.136	0.330	0.803	0.031	0.079	0.271	0.385	0.502	0.971	0.378	0.751	0.562	0.087
VT06						0.322	0.993	0.254	0.648	0.196	0.382	0.015	0.562	0.042	0.680	0.946	0.029
VT07							0.604	0.970	0.144	0.120	0.580	0.091	0.955	0.395	0.496	0.514	0.725
VT08								0.364	0.393	0.097	0.278	0.933	0.761	0.540	0.638	0.744	0.481
VT09									0.133	0.243	0.143	0.752	0.366	0.933	0.487	0.786	0.653
VT10										0.302	0.909	0.473	0.300	0.115	0.376	0.258	0.801
VT11											0.053	0.992	0.707	0.559	0.501	0.582	0.482
VT12												0.377	0.881	0.503	0.031	0.907	0.130
SF1													0.457	0.011	0.243	0.060	0.013
SF2														0.696	0.380	0.861	0.327
SF3															0.153	0.055	0.007
SM1																0.505	0.020
SM2																	0.120

0.001 0.05 1

Figure G.3: P-values for inter-speaker correlations between third principal components.

R1	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	0.001	3E-05	8E-04	4E-04	0.002	4E-05	4E-05	4E-04	3E-04	4E-04	2E-05	0.001	0.004	0.002	7E-04	0.002	0.001
VT02		0.002	0.022	0.048	0.017	0.005	0.01	0.001	0.002	0.004	0.001	0.022	0.074	0.08	0.01	0.03	0.085
VT03			0.002	1E-03	0.011	0.003	2E-04	7E-05	0.001	3E-04	1E-04	0.003	0.049	0.009	0.003	0.007	0.008
VT04				0.004	4E-04	9E-05	0.001	0.004	0.004	2E-04	9E-04	0.038	0.07	0.062	0.056	0.029	0.043
VT05					0.008	0.005	0.004	0.001	0.002	2E-04	2E-04	0.006	0.014	0.009	0.002	0.004	0.008
VT06						3E-04	4E-04	0.002	0.007	2E-04	0.002	0.118	0.024	0.153	0.063	0.117	0.132
VT07							6E-04	0.002	0.001	9E-04	2E-04	0.026	0.015	0.021	0.013	0.021	0.02
VT08								5E-04	0.002	7E-04	6E-04	0.022	0.004	0.015	0.008	0.02	0.016
VT09									6E-05	1E-05	7E-06	0.034	0.027	0.036	0.002	0.014	0.075
VT10										1E-04	1E-07	0.008	0.006	0.008	4E-04	6E-04	0.022
VT11											1E-05	0.029	0.032	0.061	0.009	0.013	0.067
VT12												0.005	0.006	0.006	3E-04	8E-04	0.013
SF1													0.03	7E-05	3E-04	2E-04	2E-05
SF2														0.005	0.002	0.001	0.015
SF3															8E-05	2E-04	2E-05
SM1																1E-04	0.002
SM2																	0.002

0.001 0.05 1

Figure G.4: P-values for inter-speaker correlations between first resonant frequencies.

R2	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	3E-05	6E-04	0.001	0.01	8E-04	0.012	2E-04	4E-06	3E-04	3E-04	2E-05	0.002	0.017	0.002	0.005	0.003	0.005
VT02		1E-05	3E-05	0.003	2E-04	0.007	1E-05	7E-06	3E-04	0.001	5E-06	1E-04	0.002	4E-05	2E-04	4E-04	3E-04
VT03			1E-04	5E-04	3E-04	0.002	3E-05	1E-04	7E-04	5E-04	2E-04	0.002	0.006	4E-04	0.002	0.004	0.001
VT04				8E-04	3E-05	5E-04	5E-06	8E-05	7E-05	2E-04	4E-06	1E-04	3E-04	2E-05	1E-04	2E-04	5E-05
VT05					2E-05	5E-05	4E-04	4E-04	0.005	7E-04	7E-04	0.018	0.004	0.002	0.017	0.023	0.002
VT06						1E-05	3E-05	1E-05	3E-04	9E-05	1E-05	0.003	9E-04	8E-04	0.003	0.002	7E-04
VT07							1E-03	9E-04	5E-04	5E-05	8E-04	0.027	1E-03	0.007	0.016	0.015	0.005
VT08								9E-07	1E-04	1E-04	4E-07	1E-04	5E-04	2E-05	8E-05	1E-04	1E-05
VT09									5E-05	5E-05	2E-08	7E-04	0.001	2E-04	9E-04	8E-04	2E-04
VT10										6E-06	1E-05	0.001	5E-04	6E-04	0.001	8E-04	0.001
VT11											6E-05	0.012	0.005	0.004	0.011	0.009	0.005
VT12												1E-04	5E-04	2E-05	2E-04	2E-04	6E-05
SF1													0.001	2E-06	2E-06	3E-06	8E-06
SF2														1E-04	1E-04	1E-04	9E-05
SF3															2E-05	1E-04	3E-06
SM1																1E-07	3E-06
SM2																	3E-05

0.001 0.05 1

Figure G.5: P-values for inter-speaker correlations between second resonant frequencies.

R3	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	0.458	0.721	0.612	0.678	0.605	0.543	0.338	0.938	0.708	0.926	0.973	0.509	0.959	0.481	0.357	0.268	0.257
VT02		0.089	0.685	0.211	0.041	0.621	0.629	0.226	0.975	0.595	0.228	0.659	0.911	0.284	0.499	0.35	0.428
VT03			0.016	0.025	0.057	0.551	0.008	0.006	0.316	0.013	0.003	0.801	0.891	0.613	0.669	0.836	0.792
VT04				0.318	0.304	0.647	0.064	0.006	0.113	0.168	0.009	0.879	0.564	0.351	0.159	0.37	0.345
VT05					0.01	0.017	0.099	0.065	0.81	0.002	0.009	0.734	0.55	0.635	0.387	0.965	0.948
VT06						0.02	0.034	0.058	0.512	0.05	0.002	0.471	0.316	0.462	0.455	0.775	0.708
VT07							0.444	0.398	0.995	0.15	0.068	0.488	0.656	0.528	0.849	0.821	0.875
VT08								0.095	0.452	0.048	0.004	0.673	0.307	0.322	0.304	0.782	0.64
VT09									0.03	0.017	0.01	0.405	0.384	0.573	0.323	0.821	0.722
VT10										0.084	0.302	0.13	0.603	0.14	0.046	0.547	0.267
VT11											0.009	0.154	0.649	0.997	0.847	0.886	0.288
VT12												0.431	0.515	0.787	0.896	0.639	0.899
SF1													0.45	0.525	0.721	0.066	0.012
SF2														0.133	0.099	0.949	0.657
SF3															0.001	0.883	0.843
SM1																0.269	0.621
SM2																	0.02

0.001 0.05 1

Figure G.6: P-values for inter-speaker correlations between third resonant frequencies.

Appendix H

Results from Watson (2014)

Watson's paper [41] which was a precursor to this project analysed 5 speakers who each had 2 sets of 11 vowels processed and analysed. The intra-speaker correlation values were not exactly the same as the values along the diagonal of Figure H.2 when the principal component analysis was carried out on the same data. To test whether the discrepancy was due to not normalising the area functions first, the data was vowel-specifically normalised sets. However, the result correlations were still almost, but not exactly the same as Watson's values. This small discrepancy may be due to the fact that the first and last values of the area functions going into the PCA were omitted. These values correspond to the front of the lips and the glottis respectively, where the MRI images are usually poorly defined. This is because of the side boundaries of the lips not being pictured because of the way our lips protrude slightly, and because of the vibrations of the vocal folds at the glottis, making these first and last cross-sectional area values somewhat dubious. These were not omitted in [41], which may have led to slightly different PC vectors.

See Appendix A to see which speakers appear in both this project and in Watson [41].

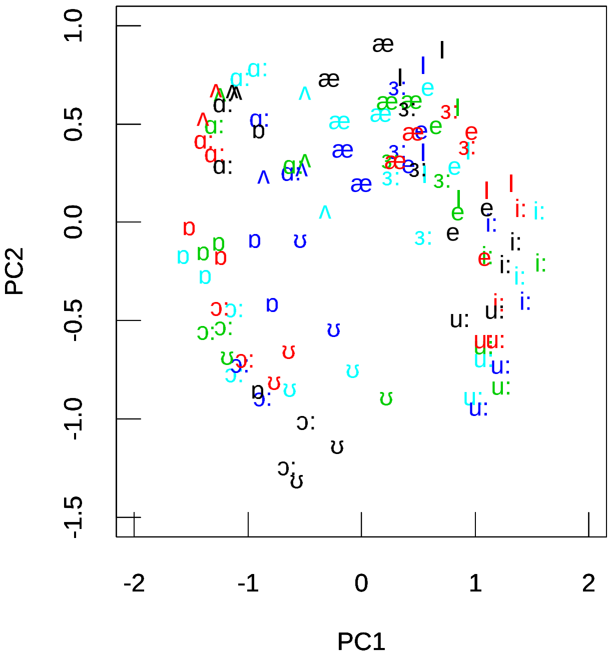


Figure H.1: Plot of the vowel data on a PC2/PC1 plane

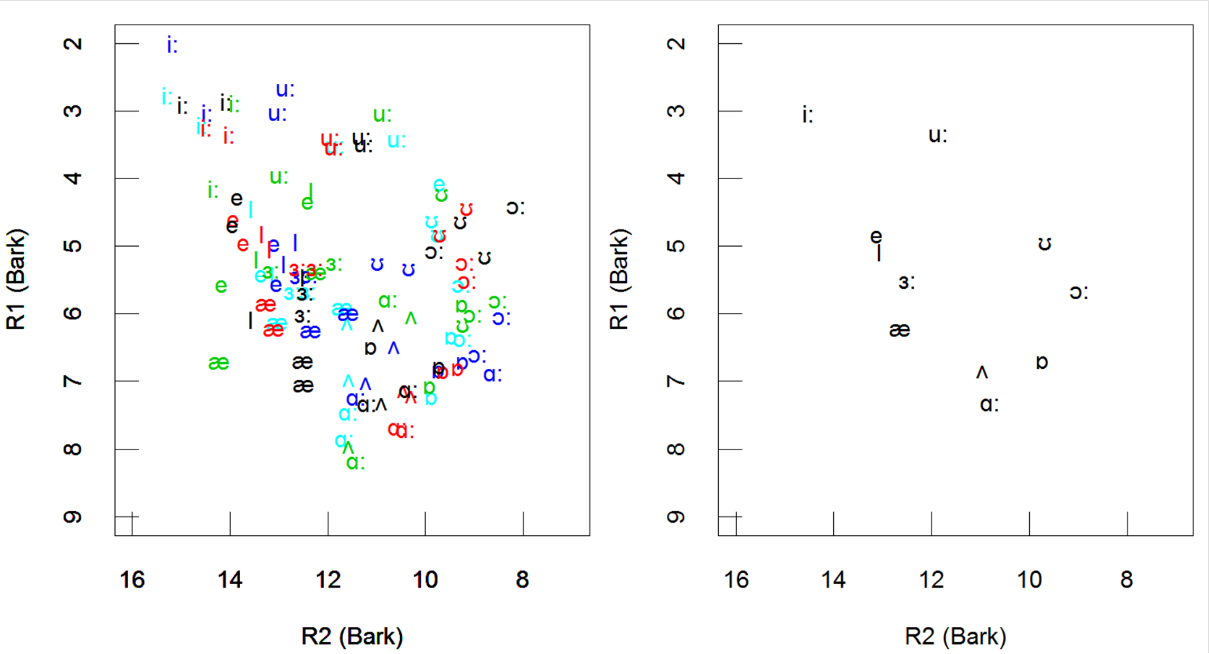


Figure H.2: Plot of the vowel data on a R1/R2 plane. Left: Combined dataset with different colours for each speaker. Right: Centroid values for each vowel.

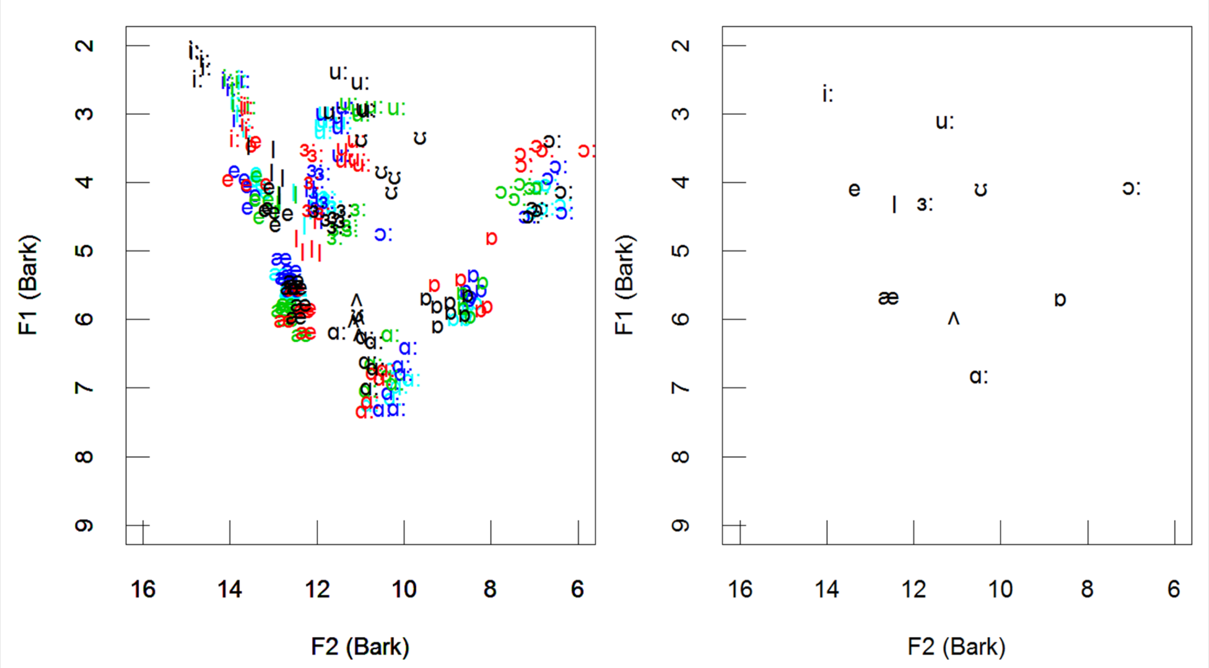


Figure H.3: Combined & centroid plots of vowel data on F1/F2 plane.

Table H.1: Variance accounted for by the first three principal components for each speaker.

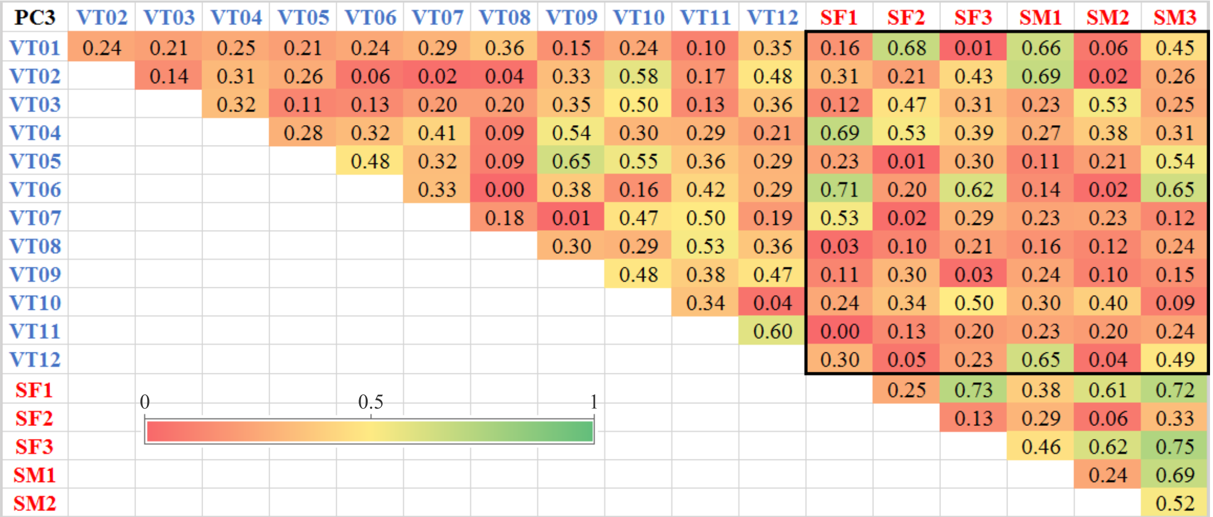
Speaker	Variance accounted by PC (%)		
	PC1	PC2	PC3
SP01	52.5	18.1	10.1
SP02	53.0	23.0	8.8
SP03	60.2	14.8	8.5
SP04	73.1	12.8	6.8
SP05	48.1	33.0	10.0

Table H.2: Correlations between PC1s from area functions of five speakers.

Speaker	SP01	SP02	SP03	SP04	SP05
SP01	**-0.95**	0.97	-0.93	-0.81	0.71
SP02		**0.94**	-0.97	-0.87	0.77
SP03			**0.93**	0.90	0.84
SP04				**0.98**	-0.96
SP05					**-0.95**

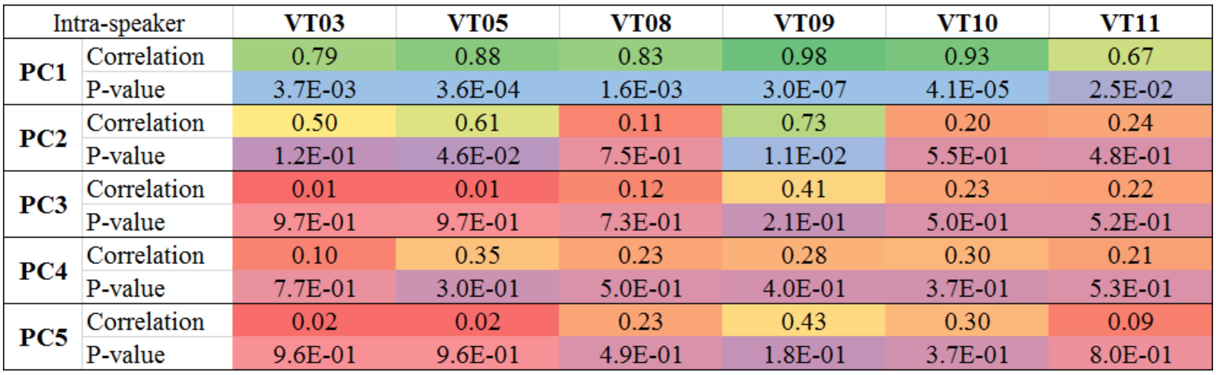
Appendix I

Additional Correlations



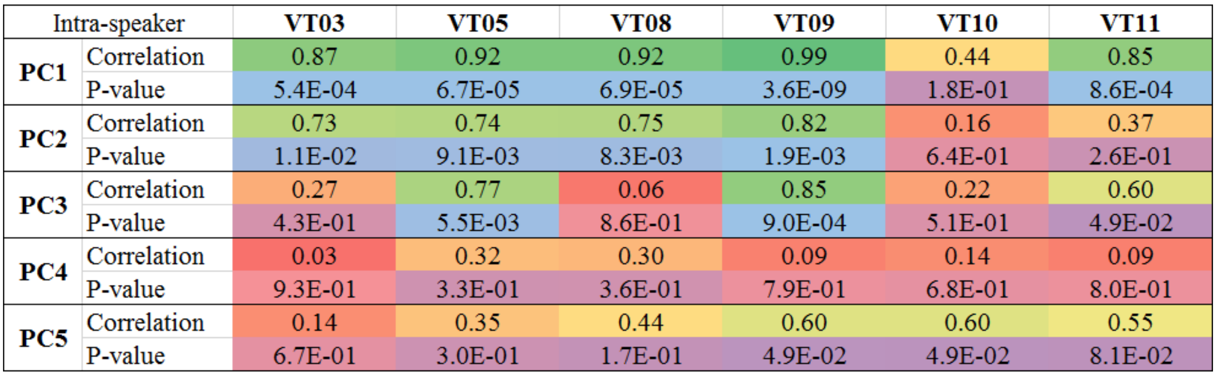
PC3	VT02	VT03	VT04	VT05	VT06	VT07	VT08	VT09	VT10	VT11	VT12	SF1	SF2	SF3	SM1	SM2	SM3
VT01	0.24	0.21	0.25	0.21	0.24	0.29	0.36	0.15	0.24	0.10	0.35	0.16	0.68	0.01	0.66	0.06	0.45
VT02		0.14	0.31	0.26	0.06	0.02	0.04	0.33	0.58	0.17	0.48	0.31	0.21	0.43	0.69	0.02	0.26
VT03			0.32	0.11	0.13	0.20	0.20	0.35	0.50	0.13	0.36	0.12	0.47	0.31	0.23	0.53	0.25
VT04				0.28	0.32	0.41	0.09	0.54	0.30	0.29	0.21	0.69	0.53	0.39	0.27	0.38	0.31
VT05					0.48	0.32	0.09	0.65	0.55	0.36	0.29	0.23	0.01	0.30	0.11	0.21	0.54
VT06						0.33	0.00	0.38	0.16	0.42	0.29	0.71	0.20	0.62	0.14	0.02	0.65
VT07							0.18	0.01	0.47	0.50	0.19	0.53	0.02	0.29	0.23	0.23	0.12
VT08								0.30	0.29	0.53	0.36	0.03	0.10	0.21	0.16	0.12	0.24
VT09									0.48	0.38	0.47	0.11	0.30	0.03	0.24	0.10	0.15
VT10										0.34	0.04	0.24	0.34	0.50	0.30	0.40	0.09
VT11											0.60	0.00	0.13	0.20	0.23	0.20	0.24
VT12												0.30	0.05	0.23	0.65	0.04	0.49
SF1													0.25	0.73	0.38	0.61	0.72
SF2														0.13	0.29	0.06	0.33
SF3															0.46	0.62	0.75
SM1																0.24	0.69
SM2																	0.52

Figure I.1: Inter-speaker correlations between the third principal components of speakers' area functions.



Intra-speaker		VT03	VT05	VT08	VT09	VT10	VT11
PC1	Correlation	0.79	0.88	0.83	0.98	0.93	0.67
	P-value	3.7E-03	3.6E-04	1.6E-03	3.0E-07	4.1E-05	2.5E-02
PC2	Correlation	0.50	0.61	0.11	0.73	0.20	0.24
	P-value	1.2E-01	4.6E-02	7.5E-01	1.1E-02	5.5E-01	4.8E-01
PC3	Correlation	0.01	0.01	0.12	0.41	0.23	0.22
	P-value	9.7E-01	9.7E-01	7.3E-01	2.1E-01	5.0E-01	5.2E-01
PC4	Correlation	0.10	0.35	0.23	0.28	0.30	0.21
	P-value	7.7E-01	3.0E-01	5.0E-01	4.0E-01	3.7E-01	5.3E-01
PC5	Correlation	0.02	0.02	0.23	0.43	0.30	0.09
	P-value	9.6E-01	9.6E-01	4.9E-01	1.8E-01	3.7E-01	8.0E-01

Figure I.2: Intra-speaker correlations between area functions in pharyngeal cavity only.



Intra-speaker		**VT03**	**VT05**	**VT08**	**VT09**	**VT10**	**VT11**
PC1	Correlation	0.87	0.92	0.92	0.99	0.44	0.85
	P-value	5.4E-04	6.7E-05	6.9E-05	3.6E-09	1.8E-01	8.6E-04
PC2	Correlation	0.73	0.74	0.75	0.82	0.16	0.37
	P-value	1.1E-02	9.1E-03	8.3E-03	1.9E-03	6.4E-01	2.6E-01
PC3	Correlation	0.27	0.77	0.06	0.85	0.22	0.60
	P-value	4.3E-01	5.5E-03	8.6E-01	9.0E-04	5.1E-01	4.9E-02
PC4	Correlation	0.03	0.32	0.30	0.09	0.14	0.09
	P-value	9.3E-01	3.3E-01	3.6E-01	7.9E-01	6.8E-01	8.0E-01
PC5	Correlation	0.14	0.35	0.44	0.60	0.60	0.55
	P-value	6.7E-01	3.0E-01	1.7E-01	4.9E-02	4.9E-02	8.1E-02

Figure I.3: Intra-speaker correlations between area functions in oral cavity only.

References

[1] American Academy of Otolaryngology–Head and Neck Surgery. The voice and aging. `http://www.entnet.org/content/voice-and-aging`, 2015.

[2] I. P. Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[3] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655, 1971.

[4] R. Carré. From an acoustic tube to speech production. *Speech communication*, 42(2):227–240, 2004.

[5] T. Chiba and M. Kajiyama. *The vowel: Its nature and structure*. Tokyo-Kaiseikan, 1941.

[6] V. L. K. Chilukuri. Practical work report, 2011.

[7] H. Fan. Comparison of vocal tract shape modelling methods: MRI vs. AR. Master's thesis, University of Auckland, New Zealand, 2012.

[8] G. Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter, 1971.

[9] A. A. Ghazanfar and D. Rendall. Evolution of human vocal production. *Current Biology*, 18(11):R457–R460, 2008.

[10] E. Gordon, L. Campbell, J. Hay, M. Maclagan, A. Sudbury, and P. Trudgill. *New Zealand English: its origins and evolution*. Cambridge University Press, 2004.

[11] C. Gray. *Acoustic Pulse Reflectometry for Measurement of the Vocal Tract with Application in Voice Synthesis*. PhD thesis, The University of Edinburgh, 2005.

[12] J. Harrington and S. Cassidy. *Techniques in speech acoustics*, volume 8. Springer Science & Business Media, 1999.

[13] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[14] E. Jacewicz and R. A. Fox. Acoustics of regionally accented speech. *Acoustics Today*, 12(2):31–38, 2016.

[15] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[16] B. J. Kröger, R. Winkler, C. Mooshammer, and B. Pompino-Marschall. Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results. In *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, pages 333–336, 2000.

[17] J. Kyttä. Influence of the nose on the acoustic pattern of nasal sounds. *Acta Oto-Laryngologica*, 69(sup263):95–98, 1970.

[18] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice. Generating vocal tract shapes from formant frequencies. *The Journal of the Acoustical Society of America*, 64(4):1027–1035, 1978.

[19] N. J. Lass. *Principles of experimental phonetics*. Mosby Incorporated, 1996.

[20] K. Livescu, F. Rudzicz, E. Fosler-Lussier, M. Hasegawa-Johnson, and J. Bilmes. Speech production in speech technologies: Introduction to the csl special issue. *Computer Speech & Language*, 36:165–172, 2016.

[21] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[22] J. D. Markel and A. J. Gray. *Linear prediction of speech*, volume 12. Springer Science & Business Media, 2013.

[23] S. Martin, R. Mathur, I. Marshall, and N. Douglas. The effect of age, sex, obesity and posture on upper airway size. *European Respiratory Journal*, 10(9):2087–2090, 1997.

[24] P. Mokhtari, T. Kitamura, H. Takemoto, and K. Honda. Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients. *Journal of Phonetics*, 35(1):20–39, 2007.

[25] J. J. Ohala and M. Ohala. The phonetics of nasal phonology: Theorems and data. *Nasals, nasalization, and the velum*, 5:225–249, 1993.

[26] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[27] G. E. Peterson and H. L. Barney. Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2):175–184, 1952.

[28] G. E. Peterson and J. E. Shoup. The elements of an acoustic phonetic theory. *Journal of Speech, Language, and Hearing Research*, 9(1):68–99, 1966.

[29] S. Petrov. Announcing syntaxnet: The world's most accurate parser goes open source. `https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html`, 2016.

[30] P. Rubin and E. Vatikiotis-Bateson. Measuring and modeling speech production. In *Animal acoustic communication*, pages 251–290. Springer, 1998.

[31] H. Searle. Summer Studentship: Vocal Tract Measurement - Refinement and proof of a measurement system for extracting length, area and volume data from MRI Images, 2012.

[32] M. Sondhi. Estimation of vocal-tract areas: The need for acoustical measurements. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(3):268–273, 1979.

[33] B. H. Story. Synergistic modes of vocal tract articulation for american english vowelsa). *The Journal of the Acoustical Society of America*, 118(6):3834–3859, 2005.

[34] B. H. Story and I. R. Titze. Parameterization of vocal tract area functions by empirical orthogonal modes. *Journal of Phonetics*, 26(3):223–260, 1998.

[35] B. H. Story, I. R. Titze, and E. A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 100(1):537–554, 1996.

[36] R. C. Team" et al. *R: A language and environment for statistical computing*, 2013.

[37] H. Traunmüller. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88(1):97–100, 1990.

[38] P. Trudgill, E. Gordon, G. Lewis, and M. Maclagan. Determinism in new-dialect formation and the genesis of new zealand english. *Journal of Linguistics*, 36(02):299–318, 2000.

[39] H. Wakita. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Transactions on Audio and Electroacoustics*, 21(5):417–427, 1973.

[40] H. Wakita. Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(3):281–285, 1979.

[41] C. I. Watson. Mappings between vocal tract area functions, vocal tract resonances and speech formants for multiple speakers. In *INTERSPEECH*, pages 1993–1997, 2014.

[42] C. I. Watson, M. Maclagan, and J. Harrington. Acoustic evidence for vowel change in new zealand english. *Language variation and change*, 12(01):51–68, 2000.

[43] C. I. Watson, M. A. Maclagan, J. King, R. Harlow, and P. J. Keegan. Sound change in māori and the influence of new zealand english. *Journal of the International Phonetic Association*, 46(2):185–218, 2016.

[44] C. I. Watson, C. W. Thorpe, and X. B. Lu. A comparison of two techniques that measure vocal tract shape. *Acoustics Australia*, 37(1):p7–11, 2009.

[45] R. Winkelmann, K. Jaensch, S. Cassidy, and J. Harrington. *emuR: Main Package of the EMU Speech Database Management System*, 2016. R package version 0.1.9.

[46] 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. *Logopedics Phoniatrics Vocology*, 24(3):132–137, 1999.

[47] S. A. Xue and G. J. Hao. Changes in the human vocal tract due to aging and the acoustic correlates of speech productiona pilot study. *Journal of Speech, Language, and Hearing Research*, 46(3):689–701, 2003.

[48] H. Yehia and M. Tiede. A parametric three-dimensional model of the vocal-tract based on mri data. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 3, pages 1619–1622. IEEE, 1997.