**Summary of Vocal Tract Modelling Papers**

Anthony White

Daniel Tan Masters – Vocal Tract Modelling: Magnetic Resonance Imaging (MRI) vs. Acoustic Reflectometry (AR)

It is known that as an individual ages, audible changes in voice characteristic and quality occurs. These differences can be caused many different factors, including the psychological state of the speakers and the physical changes to the vocal tract structure as the speaker ages. This study sets out to investigate modern imaging techniques suitable for obtaining 3D structural vocal tract data.

The most common imaging techniques include the X-ray, the CT scan, the MRI and the ultrasound. X-ray and CT scanning can have some radiation risks, whereas the MRI and ultrasound have no known side effects. Drawback to MRI is it required the subject to lie on their back, which can compromise the vocal tract structure.

AR is chosen as a comparison to the MRI. The acoustic reflectometer is a non-invasive acoustic device capable of detecting and recording the structural data of a cavity. Fast acquisition time and subject can be upright. The aim of this thesis is to compare vocal tract cross-sectional areas obtained from the same speakers as well as the resonances deduced from these area functions via the two different methods and determine the merits of the AR process.

**Speech Production**

In the process of speech production, sounds are produced by the movement of air in the cavities within the vocal tract. Speech sounds are divided into two types: vowels and consonants.

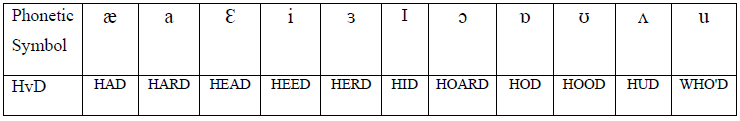
* **Vowels:** Vocal tract is unobstructed and there is no restriction on airflow. Vowels are voiced speech sounds, resulting from the air pulses originating from the glottis being resonated in the vocal tract. For the voiced speech sounds, the vocal tract remains relatively unchanged through the duration of the vocalisation, meaning that the structure is time invariant.
* **Consonants:** Harder to define, as they include semi vowels and laminar flow, but are mostly produced with constrictions in the vocal tract, causing an obstruction to air flow.

Components of a voice production system include the lungs and upper respiratory tract, which is made up of the nasal cavity, pharynx and larynx. The lungs provide the source of airflow for the duration of the voice production. For the production if a vowel sound, air exits the lungs, through the trachea, and passes through the vocal folds in the larynx causing them to vibrate. Depending on the geometry of the vocal tract, the natural frequencies of vibration within the chamber varies, and thus amplifies and attenuates different components of the sonic pulse. This directly reflects the resulting sonic vibrations which reaches and radiates from the lips.

To produce a different sound, the geometric structure of the vocal tract needs to be changed to provide different natural frequencies of vibration. This can be achieved with the movement of the articulators such as the lips, the jaw, the tongue and the velum.

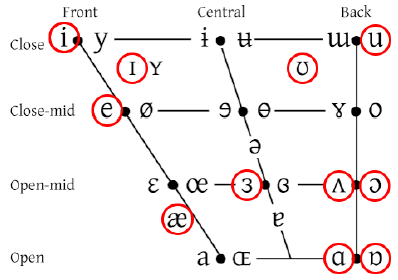
**Vowels**

Vowels are suitable for vocal tract studies involving static imaging techniques, due to their time invariant nature during the vocalisation of the target sound. A pure vowel sound is known as a monophthong. Table below shows an example of some monophthongs.



**Phonetic Vowel Space**

The phonetic vowel space is a visual representation of the auditory quality of different vowels and their relation to the positioning of the articulators specific to each vowel. For instance, [i], as in 'HEED', sounds higher than [Ɛ], as in 'HEAD', and is therefore defined as a 'higher' vowel. This can be represented with the articulatory quadrilateral, shown below.



* **Front vowels:** Tongue positioned towards the front of the mouth. This decreases the amount of space in the oral region of the tract and increases the volume of the pharyngeal region.
* **Back vowels:** Tongue is positioned towards the back of the mouth when back vowels are articulated. This causes a constriction to occur in the pharyngeal region of the vocal tract and increases the volume within the oral cavity.
* **High/closed vowels:** Defined by the amount of jaw opening during the articulation and how close the tongue is to the roof of the oral cavity. For high vowels, the jaw opening is small, and the tongue is positioned near the roof of the oral cavity, causing a constriction in that region.
* **Low/open vowels:** Articulated with a large jaw opening, and the tongue is positioned towards the bottom of the oral cavity. This generally creates a large oral cavity with constrictions in the pharyngeal region.

**Acoustic Analysis of Speech**

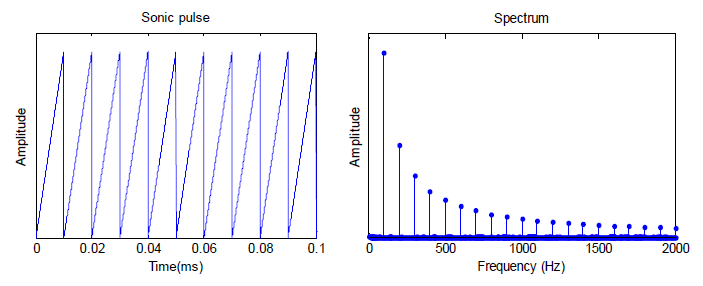
Used to investigate the physical properties of speech signals, such as pitch, loudness, formant/spectral information. It is possible, through acoustic analysis, to model the geometric structure of the acoustic chamber. This can also be performed in reverse to obtain the speech characteristics from an acoustic chamber of a given geometry.

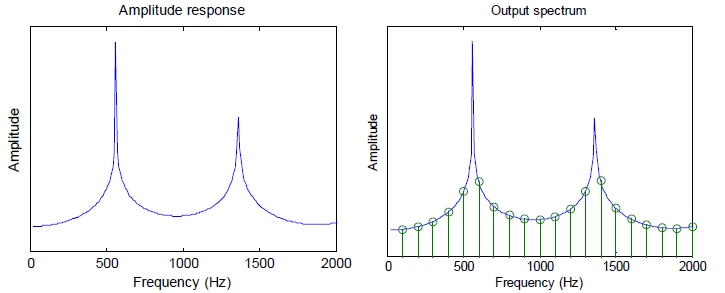
**Formants and Resonances**

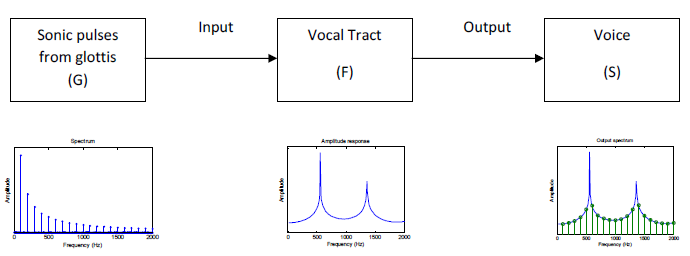
Formants are the dominant frequencies present within a speech signal, which can be obtained via spectral analysis on a speech signal. Resonances are the calculated values which describe the behaviours of the acoustic chamber.

**Source-Filter Model**

The source-filter model can be used to represent speech, where the ‘source’ are the sonic pulses produced by the vibrating glottis, which is filtered by the vocal tract to produce the output, which is the speech emitted from the lips. The input, produced from the glottis, can be approximated by a sawtooth wave, shown below.

 The geometry of the vocal tract determines the natural resonance frequencies of its chambers, allowing the amplification of certain frequencies and the attenuation of others. In the source filter model, the vocal tract behaves like a filter with a transfer function that depends on its geometry. Once the spectrum of the input signal and transfer function of the filter have been obtained, it is possible to obtain the output spectrum, which describes the sound radiating from the lips.





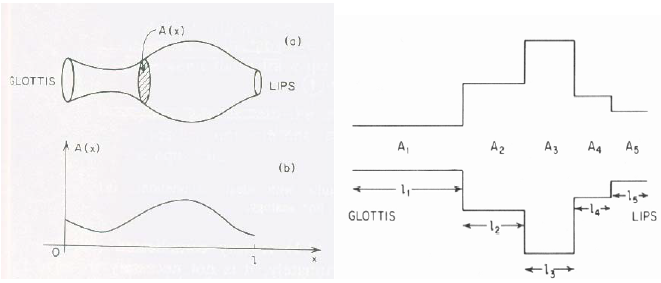
**Vocal Tract as an Acoustic Chamber**

Acoustic chambers have fundamental frequencies specific to its geometric structure, which are determined by a number of factors, such as how sound reflects off the walls of the chamber or how much resistance is met by the propagating sound waves. These factors affect the amplification/dampening of different frequencies in the chamber. A simplified form of the vocal tract is an approximately 17cm tube of uniform cross-sectional area, which is open at one end and closed at the other. Sound waves of specific frequencies will resonate by being reflected at the end of the tube back towards the source. These frequencies are given by

where is the instance of the formant, is the speed of sound in meters per second and is the vocal tract length in meters. The uniform tube model does not take into account the varying levels of constriction on the vocal tract.

**Lossless Acoustic Tube Model**

The lossless acoustic tube is an improved model, which takes into account the varying cross-sectional area along the tube. This can be approximated by the model on the right, which has an area for each section of the vocal tract. It is obvious that more sections result in a more accurate approximation.



Note that when modelling the vocal tract, a large number of sections may result in an unstable filter function. The vocal tract can be represented as a series of concatenated lossless acoustic tubes if certain assumptions are made;

* Vocal tract shape is time-invariant.
* No energy losses of the sound waves from friction, heat conduction or wall vibrations.
* Structural features such as the piriform sinuses and the ventricular appendix do not contribute to the resonance of the vocal tract chamber.
* Wave plane propagation only valid for frequencies below 4 kHz.
* Inability to model anti-resonances.

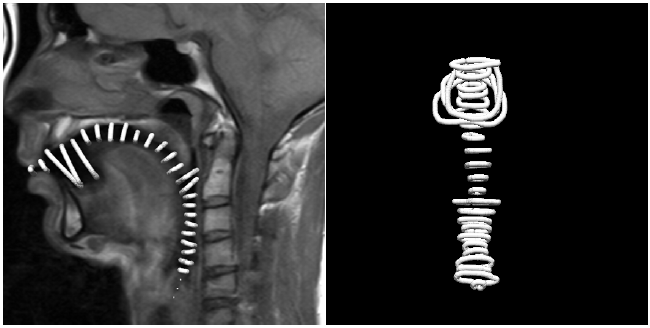
By using this model to represent vocal tract structural information it is possible, by the use of linear predictive coding, to obtain the transfer function representing the tract, which will be an all-pole model.

**Modelling the Vocal Tract**

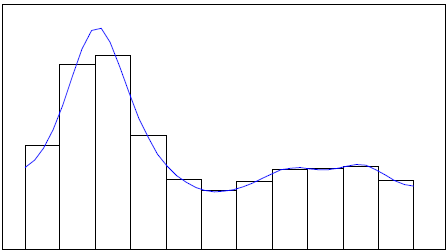
The process of modelling the vocal tract is as follows:

* Collect vocal tract shape data via MRI/AR.
* Process data into cross sectional area functions.
* Obtain acoustic tubes.
* Apply LPC method to obtain spectrum formants.

Output of AR measurements are already in the form of cross-sectional areas. MRI data, on the other hand, must be converted before cross-sectional areas can be found. Software tool CMGUI is used to do this. A wire frame model of a vocal tract in CMGUI is shown below.



When this model has been obtained, the vocal tract can be modelled using the lossless acoustic tube model (discussed previously). The radius of each tube is shown in a graph below.

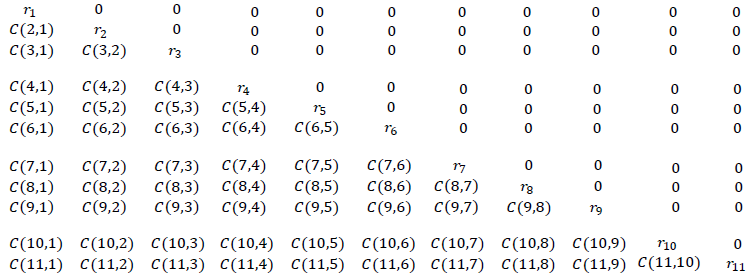


**Linear Predictive Coding Method**

Reflection coefficients are used to describe the behaviour of the environment with respect to a travelling wave. In this case, these coefficients describe the relationship of the sound being passed through a series of acoustic tubes and how much sound is reflected back towards the source. Reflection coefficients are given by

The LPC coefficients are calculated using a recursive autocorrelation technique,

The autocorrelation matrix is shown below. Is this a Toeplitz matrix?



The spectrum of the vocal tract is described by the transfer function given by

Jenny Sahng Part IV Project - Demographic trends within vocal tract area functions and vocal tract resonances for multiple speakers

Previous work by Catherine: 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. Used principal component analysis and linear predictive coding.

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.

**Vowels, Resonances and Formants**

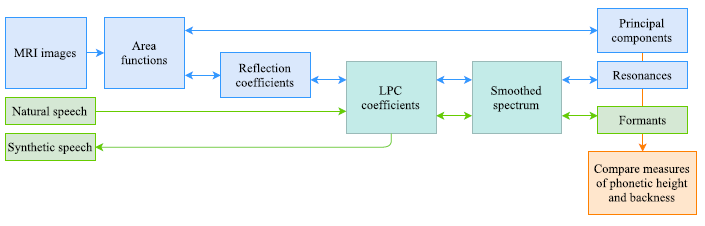
Vowels are speech sounds which resonate at a certain frequency, with air vibrating through the vocal tract in a uniform, laminar manner. Vowel analysis tends to focus on formants, since these frequency peaks are what distinguishes one vowel from another during speech.

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.

**Resonances and Formants**

From the area functions, the resonant frequencies of the vocal tract can be estimated via LPC. The resonant frequencies can be compared to the formant frequencies which can be heard in a recorded speech signal.

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’). Compared to formants, resonances are a more direct reflection of the vocal tract shape.



**Vowel Quadrilateral**

Look at previous paper. Two main chambers of the vocal tract are the oral cavity and the pharyngeal cavity:

* Oral cavity: space above and before the highest part of the tongue, before the uvula.
* Pharyngeal cavity: space between the tongue and vertical back wall of the mouth, from the uvula and the glottis.

As these cavities change size, the resonances of vocal tract change accordingly, as do the formants. There is a well-established correlation between first and second formants (F1 and F2) with first and second resonances in terms of phonetic height and backness. When F1 and F2 are plotted on an x/y axis, we obtain the vowel quadrilateral.

**Effect of age, gender and accent**

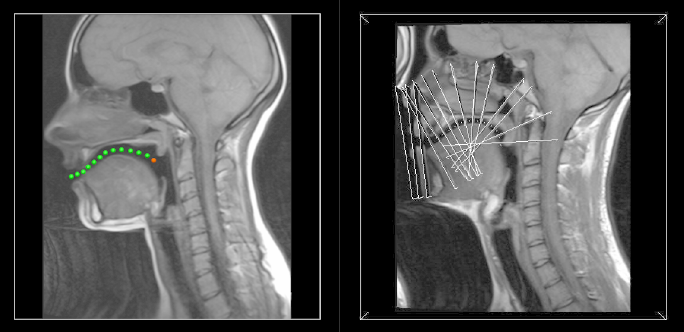
As we age, the tissues in the mouth relax, the mucous membranes dry and thin, and fine motor control may become compromised. Vocal chords also undergo atrophy, resulting in changes in pitch, volume, and importantly, resonance.

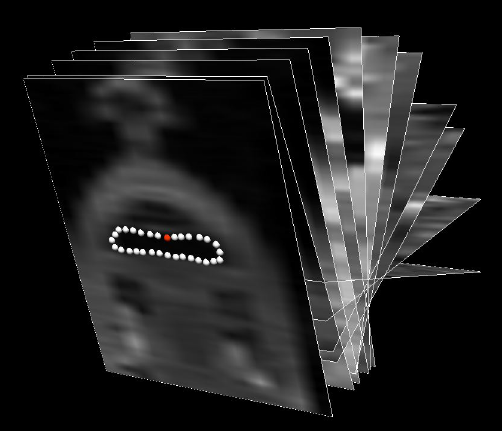
The effect of gender on vowel sounds is relatively simple. It is well established that human male voices are on average twice as deep as those of females, despite males only being 10% taller and 20% heavier on average. This disproportionate decrease in resonant frequency is likely to originate from the increased thickness and length of male vocal chords due to testosterone, as well as the slightly larger size of the vocal cavities.

**Area Functions from MRI Data**

Used CMGUI to generate cross-sectional slices along the curve of the vocal tract from sagittal MRI images, from the lips to the glottis. Data points were then manually placed around the boundary of the oral and pharyngeal cavities 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).

Image below shows data points defining the curvature of the oral cavity in CMGUI. Right image shows slices orthogonal to the user-defined curvature of the oral cavity. Tell me more about how CMGUI works?





* Plotted mean area functions to visualise trends in area function data
* Quantifying variability via a few methods, such as ‘performing correlations on the principal components of the same sets of data which had been processed a second time’. Pearson product-moment correlations between the principal components (PCs) of the area functions.

**Principal Component Analysis (PCA)**

Used to reduce the number of dimensions in the data frame of area functions. 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). The PCs encode the dimensions along which there is the most variance, in decreasing order (PC1 aligns with dimension with most variance, PC2 second most etc.). Many studies have shown that the first two principal components suffice in capturing up to 90% of the variance in the original data.

**Resonance Analysis**

Use LPC to model the sound production system via the source-filter model (all-pole model),

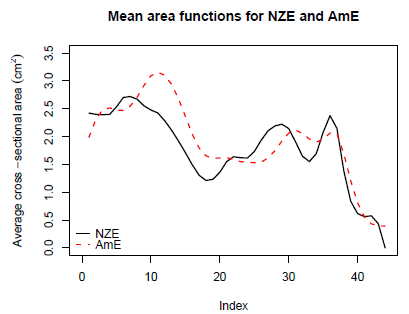
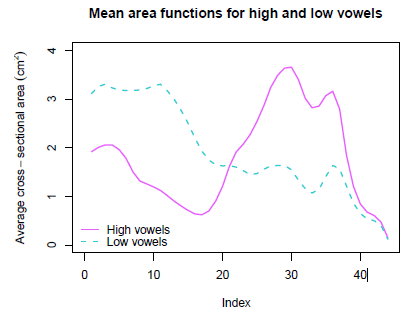
where is the input (source) signal and is the output signal. and are their coefficients which encode the properties of the vocal tract filter (its resonances and anti-resonances).

Once LPC coefficients are calculated, they can be related to the area function via a set of reflection coefficients . 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 hypothetical sampling frequencies () 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 determined by the number of cross-sectional areas () provided, the length of the speaker’s vocal tract (), as well as the speed of sound in air ().

**Results – Area Functions**

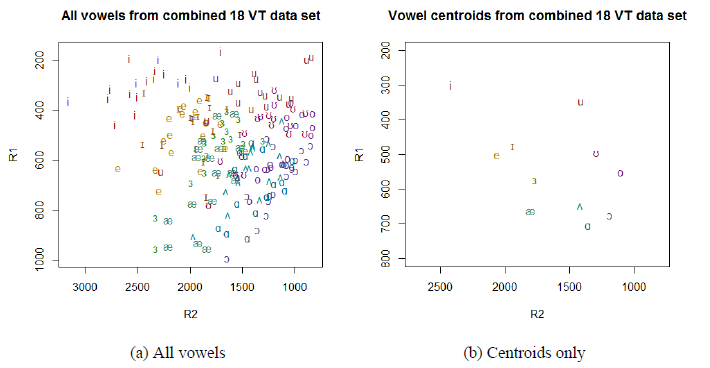
Below shows average area functions. Index – 0 is near lips, 45 is near glottis. Comparison between high/low vowels and between NZ/American English.

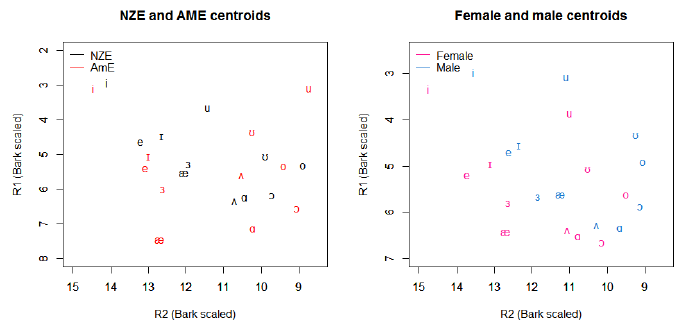
****

Principal component analysis was also performed. PC1 and PC2 of the NZE data set accounted for 60.1% of the variance.

**Vowel plots (R1-R2 plane)**

The resonances appear to have separated the vowels out by their phonetic height and backness.

****

****

**Discussion**

* **High vowels**: large cross-sectional area in pharyngeal region, and small area in the oral cavity. High vowels require tongue to be raised, narrowing gap between tongue and roof of mouth. Tongue back and root are pulled up and forward, increasing space in pharyngeal region.
* **Low vowels**: small cross-sectional area in pharyngeal region, and large area in the oral cavity. Tongue is flattened as the jaw drops, increasing volume of oral cavity and decreasing the volume of the pharyngeal cavity as the tongue positions itself further towards the back of the throat.
* NZE vowels appear to be more ‘fronted’ than GenAm vowels. Front vowels have smaller oral cavities and larger pharyngeal cavities. This is reflected in the corresponding mean area functions. NZE vowels are also ‘higher’ on average.
* Female speakers, particular speakers of GenAm, had overall higher vowels than men. 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.
* 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.
* NZE vowels on PC1-PC2 plane show good separation of individual vowels by phonetic height and backness.
* Results support hypothesis that PCA is able to capture vowel qualities such as phonetic height and backness, while ignoring individual speaker differences. Opens the door to analysing changes in vowel quality using combined, bulk data sets as opposed to analysing each speaker as a case study
* 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. 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.
* **Resonance analysis:** GenAm vowels are further back compared to NZE. GenAm bounded by corner vowels /i, ae, ə, u/. NZE bounded by /i, ə, o/ as cardinal vowels, and /u/ is high-central rather than high-back. Corner vowels?
  + **‘herd’:** NZE achieves this by rounding the lips resulting in lower F1/F2/R1/R2, GenAm achieves this by rhoticising (prominently pronouncing the ‘r’).
  + Male speakers have lower resonance values, likely reflecting larger vocal tract size.
  + Resonances calculated via LPC are reliable for measuring vowel quality and vocal tract shape.

Helen Elizabeth Searle - Refinement and proof of a measurement system for extracting length, area and volume data from MRI Images

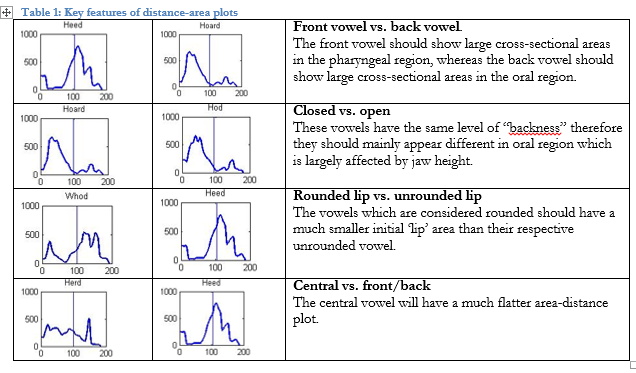
Interested in investigating changes in the vocal tract, especially the pharyngeal region, due to aging.

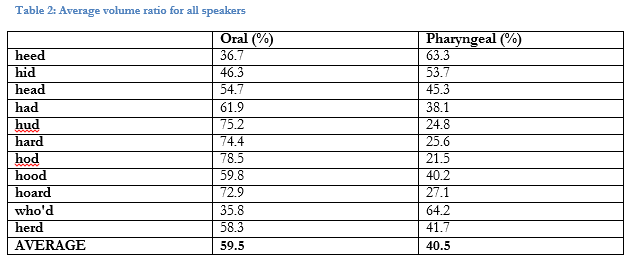
**Previous work**

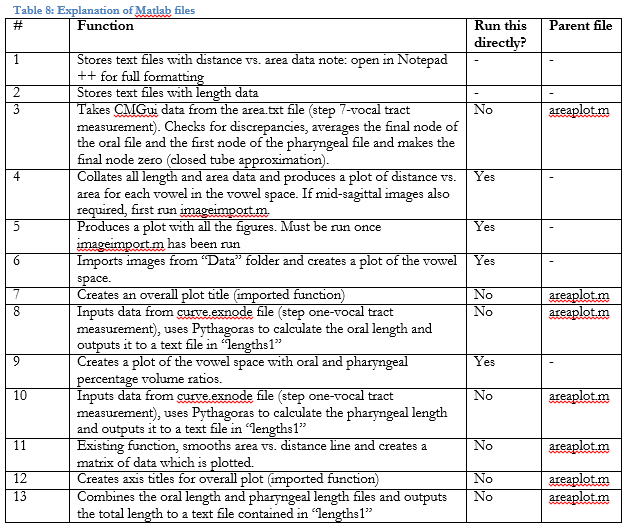
* Kalyan Chilukuri: Developed method for calculating cross-sectional area of the vocal tract from MRI images. Teeth and the lips were two areas which required further investigation as these had a significant impact on the accuracy of the area calculation (summer intern).
* Justine Hui: Measured vocal tract via acoustic reflectometry. Noted difficulty in correctly identifying the velar-pharyngeal port.

**Important Points**

* Appendix A is a good, detailed literature review.
* Appendix D provides method for identifying oral and pharyngeal cavities.







(Just in case I end up using this)

