Mappings between vocal tract area functions, vocal tract resonances and speech formants for multiple speakers.

Catherine I. Watson
University of Auckland, New Zealand
c.watson@auckland.ac.nz

Abstract
This study looks at mappings between vocal tract area functions (obtained from MRI scans), vocal tract resonances, and speech formants for five New Zealand English (NZE) speakers. All eleven NZE monophthongs were investigated, for each speaker. Principal component (PC) analysis on the area functions of both the individual speakers and combined speaker set is performed. In all cases the first two PCs account for most of the variances in the data. The first two PCs of the vocal tract area accounted for between 71-86% of the variance across the data for the individual speakers, and each was highly correlated between speakers. Across the combined speaker set the first two PCs accounted for 60.5% of the variance, and remained related to phonetic height and backness. This can clearly be seen when the transformed area function data is compared to the first two vocal tract resonances and speech formants.

Index Terms: vocal tract areas, resonances and formants, Bark scale

1. Introduction
There is a long-established and deterministic link between vocal tract shapes and speech formants. Early studies on formant analysis established a strong correlation between the first two formants with phonetic height and backness respectively (see eg. [1-3]). It is possible to also calculate vocal tract (VT) resonances from the vocal tract shapes derived from the various imaging techniques and compare these to speech formants (see eg.[1,4-8 ] to name a few). As the differences between the estimated VT resonances and measured speech formants are better understood the modelling of the vocal tract is improved.

The advent of accessible imaging technology has enabled a large number of studies on vowels looking at the vocal tract shapes (eg. [4-8,10-12]). Various analysis techniques are used to extract vocal tract shape data from the images. This data is typically multi-dimensional and data-reduction techniques are used to reduce the dimensionality such as principal components analysis. It is well established that most of the variability in the vocal tract data can be accounted for by the first two principle components (e.g.[4,5,7,10,11]). Perrier [13] argued this is to be expected for tongue data, given the biomechanical constraints of the vocal tract. By logical extension this would also be expected for the vocal tract shape.

Most earlier studies only collected data from a single speaker (e.g. [4,7,8,10,11]). Since speech production in different individuals is quite idiosyncratic, the question arises whether two principal components is sufficient when considering vocal tract data from multiple speakers. Story (2005) [5] provided a very thorough analysis of vocal tract (VT) area functions derived from 6 speakers of American English. For the first time it was possible to compare multiple VT area functions from the same vowel, but different speakers. The VT area functions for the different vowels looked very similar across the speakers. This was supported by there being a very strong correlation between the first three principal components for each of the speaker’s data sets. However despite this strong correlation Story (2005) never combined the data from all the speakers, which would have enabled him to look at group trends, and establish whether the individual speaker differences in vocal tract shapes would be stronger than the vowel differences in principal component analysis.

In these vocal tract studies the frequencies of the speech formants and VT resonances are given in the linear frequency scale (eg. [4-8,10,11]). Since most of the vocal tract studies are performed on single speakers this has never been an issue. Even in [5] the formant and VT resonance data for each speaker were presented separately. When looking at VT resonances and speech formants from multiple speakers, it is better these are represented on the Bark scale, which reduces non-phonetic speaker specific information from the speech signal [14], and increases separation between the different vowels [15]. In addition, given the feedback between the speech production system and the auditory system, it makes sense to look at the formants and resonances on a perceptually based frequency scale such as the Bark scale [16].

The purpose of the present study is to investigate the VT area function data from multiple speakers and establish whether it is feasible to combine the area function data from all speakers when principle component analysis is performed, and whether this allows the speaker differences in vocal tract shapes to be separated from the consistent vowel-dependent shapes that determine vowel resonances.

2. Method

2.1. Participants
This study uses data obtained from five New Zealand English (NZE) speakers (four men, 1 woman). Three participants were aged in their twenties, two in their forties.

2.2. Data Collection and Preparation

MRI scans

The images used in this study were mainly obtained by the 1.5T Siemens Magnetom Avanto MRI scanner, although part way through the study the scanner was upgraded to a 3.0T version (see Table 1 for Speaker and Scanner details). Scans were performed to obtain images of parallel sagittal planes with 6 mm separations with 1 mm resolution, and the field of view was around 199x250 for each session (see Table 1). 13 slices were obtained for each target vowel. The scanning went from jaw edge to the jaw edge, and these were determined by a localizing scan at the start of the data collection process. The participants were instructed to say each
vowel within an hVd frame, sustaining the vowel for the duration of the scan. Each scan took 15 seconds. For each participant we obtained two scans for each of the 11 NZE monophthongs. The vowel order was the same for each participant, they started with /i:/ and went anti-clockwise around the vowel space, finishing with the non-peripheral vowels, going down in decreasing height. This sequence was repeated twice. Each participant was in a supine position in the MRI scanner, and was in there for around 30 minutes.

The images from the scans were subsequently used to obtain structural information of the vocal tract. The images were converted to bitmaps and read into the open source program CMGUI (http://www.cmiss.org/cmgui) to create a 3-dimensional texture block. This enabled points within the vocal tract to be identified. From the 3-d vocal tract models area 2-d area functions were then calculated. Each vocal tract volume was ultimately segmented into 29 cross-sectional areas. A brief description of the process is given below. Points along the centre line of the mid-sagittal section of the vocal tract are first identified by hand, and then an interpolation algorithm was used to ensure all points are spaced at equi-distances between the lips and glottis. At each point a plane was constructed perpendicular to the centre line. This produced a sequence of images that cut the vocal tract perpendicularly throughout its length. The boundary of the vocal tract was then manually marked on each of the planes. Finally, a smoothing spline was fit to these data and the internal area computed at each of the planes. Further details of this process can be found in [8,19]. All area functions were read into R [17] for further analysis.

Table 1. Summary of participants and MRI details

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Age</th>
<th>MRI scanner</th>
<th>Field of view (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP01</td>
<td>25</td>
<td>1.5T</td>
<td>211x260</td>
</tr>
<tr>
<td>SP02</td>
<td>23</td>
<td>3.0T</td>
<td>211x260</td>
</tr>
<tr>
<td>SP03</td>
<td>45</td>
<td>1.5T</td>
<td>199x250</td>
</tr>
<tr>
<td>SP04</td>
<td>45</td>
<td>3.0T</td>
<td>185x220</td>
</tr>
<tr>
<td>SP05</td>
<td>25</td>
<td>1.5T</td>
<td>191x240</td>
</tr>
</tbody>
</table>

Speech Recordings

Speech was recorded in a sound booth (Whisper Room MLD8484E) directly on to a Marantz PMD670 Solid State Recorder at a sampling rate of 20 kHz, using a Shure SM58 Microphone. For each speaker we collected citation form hVd words for the 11 NZE monophthongs. We collected five versions of each hVd word, presented to the speaker in a random manner via word lists. To avoid list effects the final hVd word in each list was a dummy one and was removed.

The speech data were transferred to the computer and phonetically labelled using EMU speech tools [18]. The first four formants were calculated in EMU. Formant tracks were checked, and corrections were made if necessary. The formant values were extracted at the vowel midpoints and analysed in R/EMU [17,18].

Data Analysis

Two analyses were performed on the vocal tract (VT) area functions. Firstly principal component (PC) analysis was performed on the area functions to investigate how well actual the vowels could be distinguished on vocal tract shape only. To account for different speaker size, all area functions were normalised. For each vocal-tract shape, the maximum cross-sectional area was identified and the cross-sectional areas were expressed in terms of their proportion to this maximum. Secondly the resonant frequencies of the area functions were calculated from custom functions based on the standard linear prediction model of speech (e.g. [8, 20]). To ensure consistency in the dimensions between the formant and resonance data only the first four unique resonances are considered for each vowel. All analysis was done in R[17].

3. Results


The first two PCs of the VT area accounted for between 71-86% of the variance across the data for each of the speakers, the mean of this variance was 78%. The variation for this variance across the speakers can be seen in Table 2. The higher order PCs accounted for very little of the variance, the values for PC3 are also listed in Table 2. This is consistent with other studies on single speakers (see eg. [4,5,7,9-11]).

Using Pearson’s product-moment correlation, the within-speaker correlation (there are two data sets for each speaker), and also the between speaker correlations for the first two PCs were calculated for each speaker (higher order PCs were not considered as they accounted for very little variances in the data). In all case the correlations were significant (p>>0.01), and very strong. The mean of the absolute value of the correlation for the first PC was 0.9, and for the second PC it was 0.76. The within and between speaker correlation for the first PC are given in Table 3. The values in this study are comparable with [5].

Table 2: For each speaker the percentage of total variation in area functions from the first three principal components.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Variance accounted by PC1 (%)</th>
<th>Variance accounted by PC2 (%)</th>
<th>Variance accounted by PC3 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP01</td>
<td>52.5</td>
<td>18.1</td>
<td>10.1</td>
</tr>
<tr>
<td>SP02</td>
<td>53.0</td>
<td>23.0</td>
<td>8.8</td>
</tr>
<tr>
<td>SP03</td>
<td>60.2</td>
<td>14.8</td>
<td>8.5</td>
</tr>
<tr>
<td>SP04</td>
<td>73.1</td>
<td>12.8</td>
<td>6.8</td>
</tr>
<tr>
<td>SP05</td>
<td>48.1</td>
<td>33.0</td>
<td>10.0</td>
</tr>
</tbody>
</table>

The strong correlation between the first and second principal components for the area functions for each speaker can also be seen in Figure 1, where the transformed area data for each speaker is superimposed on the data from the other speakers (a different colour for each speaker). It can be seen that vowel data for the different speakers is grouped in similar areas of the PC1-PC2 plane. The first PC tends to separate all the speakers data according to phonetic backness, and the second PC tends to separate all the speakers data according to

Table 3: Matrix of coefficient correlation quantifying the similarity of the first principal component from the area functions for all five speakers.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>SP01</th>
<th>SP02</th>
<th>SP03</th>
<th>SP04</th>
<th>SP05</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP01</td>
<td>-0.95</td>
<td>0.97</td>
<td>-0.93</td>
<td>-0.81</td>
<td>0.71</td>
</tr>
<tr>
<td>SP02</td>
<td>0.94</td>
<td>-0.97</td>
<td>-0.87</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>SP03</td>
<td>0.93</td>
<td>0.90</td>
<td>0.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP04</td>
<td>0.98</td>
<td>-0.96</td>
<td>-0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP05</td>
<td>-0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1994
phonetic height. It should be noted however that PC values for SP03 and SP04 (the red and green data in Figure 1) was multiplied by minus 1, as these values were a mirror image of the equivalent vowels for the other speakers. This is also implied in Table 3, where the PC1 values for these two speakers are strongly inversely correlated with the remaining three speakers.

Figure 1: The PC1-PC2 plane for the data from each of the speakers superimposed on each other. Different colours indicate different speakers.

3.2. Vocal tract shape: Collective results.

There was a high correlation across all speakers with the PC data, and a clear similarity for the vowel distribution for each of the speakers in the PC1-PC2 plane (Figure 1). This was expected from earlier studies. Next the PC analysis was performed on the entire data set containing all vowels from all 5 speakers. The fraction of variance explained with each PC is plotted in Figure 2. As before the first two PCs account for most of the variance. With this large combined set the first two PCs of the VT area functions accounted 60.4% of the total variance in the data; 39.6% and 20.8% respectively. PC3 only accounted for a further 10.9% of the variance. This is somewhat less than the results for the per-speaker PC analyses and the findings for single speakers in other studies [see eg. [4,5,7,9-11]].

Figure 2: The fraction of variance for each of the principal components in the combined data.

Figure 3, a plot of the PC transformed area functions on a PC1-PC2 plane, shows the relationship amongst the vowel area functions. The left plot shows all the data, the different colours indicating different speakers, the right plot just plots the vowel centroids averaged across speakers. As with the data from the individual speakers, there is a clear correspondence between this data and the traditional formant data, albeit it is reflected in the x plane. The first two principal components are clearly accounting for the variability due to the different vowels, not due to the different speakers. Phonetically the first PC tends to separate the vowels according to vowel backness, and the second PC separates according to vowel height. Therefore PC analysis of the combined data set of area functions transforms the data in a similar manner to the transformations of the area data from the single speakers. The speaker differences in the combined data set are likely being accounted for with the high order PCs.

Figure 3: Representation of the vowels in the PC1-PC2 plane. Each point is based on the means of the PC1 and PC2 for each vowel across all the speakers.

Inspection of the rotation vectors for the first two PCs (Figure 4) provides useful insight into what parts of the vocal tract are the most important to distinguish between the vowels. The first PC is determined from the difference between the oral and pharyngeal cavity, with a positive PC1 resulting from a larger oral than pharyngeal cavity (ie back vowels, as seen in Figure 2) Heuristically contrasting this rotation vector with the MRI data, the first PC gives the most weight to the area around the front of the tongue, and the area around the base of the tongue, just up from the epiglottis (middle of the pharyngeal region). The second PC gives the greatest weight to the region in the middle of the vocal tract, which is around the velopharyngeal port. The rotation vectors are similar to those obtained from the single speakers in this study (not shown here) and also for other studies obtained from single speakers (e.g. see [5, 11]), although in those instances there was also a notable weighting close to the glottis for PC2, which is not present in our data.

Figure 4: The first two principal components vectors from the area functions data.

3.3. Vocal Tract Resonances: Collective results

For each VT area function the first four resonance frequencies were identified from the LPC spectra calculated from the area functions. Plots of the data on a first resonance (R1) second resonance (R2) plane are given in Figure 5. Lobanov speaker normalization [18,21] was applied to the data for each speaker, on account of there being a single female speaker, along with the four male speakers in the data set. Although the Bark scale does reduce speaker differences, the data from the female
speaker was still noticeable shifted to the bottom left on the R1-R2 plane before the normalization was performed. The left plot gives all the data points on the R1-R2 plane, with the data from the different speakers indicated with a different colour. The orientation of the axis is the same as the traditional formant plots. The right plot gives the centroid data. All the resonance data is presented in the Bark frequency scale. These plots are very similar to the vowel distribution in the transformed area function plots in Figure 3, although R1, which is related to phonetic height, maps to PC2, and R2, which is related to phonetic backness, maps to –PC1.

Figure 5: The normalized vocal tract resonances on a R1-R2 plane, (left) all the data points, different colours for different speakers (right) centroid data.

3.4. Formants from Speech: Collective results.

The first four formant for all the vowels, from the recorded speech, were extracted. As with the resonance data, all the formant data was speaker normalized using the Lobanov algorithm [18,21]. Plots of the data on a traditional; F1-F2 plane of the resonances, and F1-F2 plane of the formants.

The normalized speech formant frequencies (Bark scaled) on a F1-F2 plane, (left) all the data points, different colours for different speakers (right) centroid data.

Figure 6: The normalized speech formant frequencies (Bark scaled) on a F1-F2 plane, (left) all the data points, different colours for different speakers (right) centroid data

The first four formant for all the vowels, from the recorded speech, were extracted. As with the resonance data, all the formant data was speaker normalized using the Lobanov algorithm [18,21]. Plots of the data on a traditional; first formant (F1) second formant (F2) plane are given in Figure 6. The left plot gives all the data points on the F1-F2 plane, with the data from the different speakers indicated with a different colour. The right plot gives the centroid data. All the formant data is presented in the Bark frequency scale. These plots are very similar to the vowel distribution in the transformed area to the resonance plots (Figure 5), as is expected. Thus F1 maps to PC2, and F2 maps to –PC1.

4. Discussion

This study shows that it is possible to combine the area function data from multiple speakers and get sensible results, comparable to the analysis from single speakers. The first two principal components (PC) accounted for 60.4% of the total variance of the combined data set, and the plots resembled rotated vowel spaces. The first PC placed the greatest weight on area around the front of the tongue, contrasted with the area around the base of the tongue, just up from the epiglottis (middle of the pharyngeal region). The second PC is giving the greatest weight to the region around the velopharyngeal port. It is the changes in these regions which separate the vowels. Phonetically, from the PC1-PC2 plane (Figure 3), we can see PC1 is related to vowel backness, and PC2 is related to vowel height. This relationship was also found in [11] and is very similar to the findings in [9,22] when they transformed tongue surfaces with PCs. Although the latter are not vocal tract shapes, the tongue surface is clearly strongly related to vocal tract shapes. The studies from [9,11] were only on a single speaker. However there were 6 speakers in [22], so those results are in keeping with this multi-speaker study.

The results reinforce the well-known facts that the first two speech formants and first two VT resonances also map to phonetic height and backness respectively. In this study we used the Bark Scale for the resonance and formant analysis. The difference between this study and earlier studies was here the area function data, vocal tract resonances, and speech formants were all from the combined data set, not from single speakers. The usage of the Bark scale did reduce the speaker differences in the vocal tract resonances and speech formants. However due to the fact there was data from both men and a woman, we were required to do further speaker normalization on the resonances and formants.

However gender did not impact on the generalities area function findings. When the data from only woman speaker was removed from the combined data set, first two PC accounted for 66.3% of the area function data. This is not a huge increase, and in fact if any of the 4 men were removed instead then the total variance accounted for remainder ranged between 63.4% and 70.3%.

5. Conclusions

In this study we have looked at the area function data from the vowels of five different NZE speakers. We found that the PC1 and PC2 values of the area functions for the different vowels were similar for each speaker and it was possible to combine them. As with the single speakers, the first two PCs accounted for most of the variability in the combined speaker set, and they are related to phonetic height and backness respectively. There is a strong mapping between the vowel spaces in the PC1-PC2 plane of the area functions, and those in the R1-R2 plane of the resonances, and F1-F2 plane of the formants. Thus it is possible to combine the area function data for multiple speakers, which will enable group trends, such as the impact of aging in adults, to be more easily studied.

6. Acknowledgements

Thanks to Daniel Hsiao-Tien Fan, Helen Searle, and Kalyan Chilukuri for their assistance in preparing the 3-d vocal tract data, William Thorpe for his helpful and insightful feedback and to the Faculty of Engineering, University of Auckland for funding the project.

7. References


