An analysis of the size information in classical formant data: Peterson and Barney (1952) revisited

Richard E. Turner and Roy D. Patterson †

† Centre for the Neural Basis of Hearing, Department of Physiology, University of Cambridge

Downing Street, Cambridge, CB3 2EG, UK.

E-mail: †ret26@mrc-cbu-cam.ac.uk, rdp1@mrc-cbu-cam.ac.uk

URL: http://www.mrc-cbu.cam.ac.uk/cnbh/index.html

Abstract Irino and Patterson (2002) have suggested the Mellin Transform as a model for vocal tract normalisation in the auditory system. In this report, we reanalyse the classical formant data reported by Peterson and Barney (1952) to see if it supports the normalisation hypothesis. The vowel formant data are clustered, quantitatively, using very general assumptions about speaker-variability. These clusters allow us to assess the degree to which vowel formant variability is attributable to changes in vocal tract length (VTL). The width of clusters associated with men, women and children within a given vowel cluster motivated consideration of a natural space in which to analyse scaled frequency components of sounds. By recasting Peterson and Barney's data into this new representation we are able to quantify the utility of scale normalisation.

Keywords Mellin transform, Vowel normalisation, Scale

1. Introduction

Vocal fold size information and vocal tract length information are encoded in spoken sounds and account for the percepts of pitch and scale. Larger people have longer, more massive vocal folds and consequently speak with a lower pitch. Models show that VTL changes scale the resonances of the tract in a reciprocal manner (Flanagan, 1972). Consequently, formant frequencies are lower in larger people. Both aspects of speech sounds – pitch and vocal tract length - provide cues to speaker size, whilst vocal tract shape defines the vowel spoken. Since VTL differences are a major source of inter-speaker variability, Irino and Patterson suggest that a scale normalisation tool (such as the Mellin Image), which segregates VTL from VT shape information, would be of great benefit to the auditory system. One of the aims of this study is to quantify the importance of VTL variability in spoken vowel sounds using Peterson and Barney's classic work on formant frequencies (Potter & Steinberg, 1950; Peterson & Barney, 1952). Peterson and Barney recorded two repetitions of 10 vowel sounds from 76 men, women and children. From each of these recordings they extracted the frequency of the first three formants and the pitch of the vowel using a spectrogram. They also played the vowel sounds to 70 listeners and recorded the recognition rates.

2. Method

Peterson and Barney measured three formants and we describe their vowel data in terms of a three dimensional formant-frequency space. It is useful to delineate the regions occupied by particular vowels in this space. It was with this goal in mind that Peterson and Barney clustered their vowels, drawing 'closed loops for each vowel … arbitrarily to enclose most of the points'. Modern day computing methods afford us with a more quantitative procedure. Three dimensional Gaussian distributions, with arbitrary orientation in the formant-space, have been fitted to each of the vowel clusters; firstly for all speakers in combination, and secondly for men, women and children separately. The clusters are best visualized by plotting a surface of constant probability. This surface is an ellipsoid centered on the mean of the data. The clusters produced in this study (see Figure 1) have a well defined shape (based on very general assumptions about speaker variability) and size (drawn to one standard deviation in extent in each direction and therefore to enclose 30% of the data points).

3. Results and Discussion

The separation of the vowel clusters is much more prominent in the f_1/f_2 plane than in the f_1/f_3 and f_2/f_3 planes (see Figure 1), suggesting there is sufficient information in the first two formants to enable recognition of most vowels. This is consistent with the finding that

synthetic vowels composed of two damped sinusoids at the positions of the first two formants are readily recognised (Patterson et al., 2000). It is also clear that two distinct 'vowel slabs' appear in this space; they have been mathematically expressed by Broad and Wakita (1977) and are described by Miller (1989). Finally, note that the size, shape and orientation of the fitted clusters reflect many of the established ideas about vowel formant frequencies, vindicating the clustering methodology.

3.1. Scale Variation

A VTL change scales the VT resonant frequencies proportionately. In vowel space, this transformation scales the magnitude of the vowel vector whilst keeping the spatial orientation fixed. So, scaled sounds lie on straight lines passing through the origin of the vowel vector space, as illustrated in Figure 2. To analyse the variability we

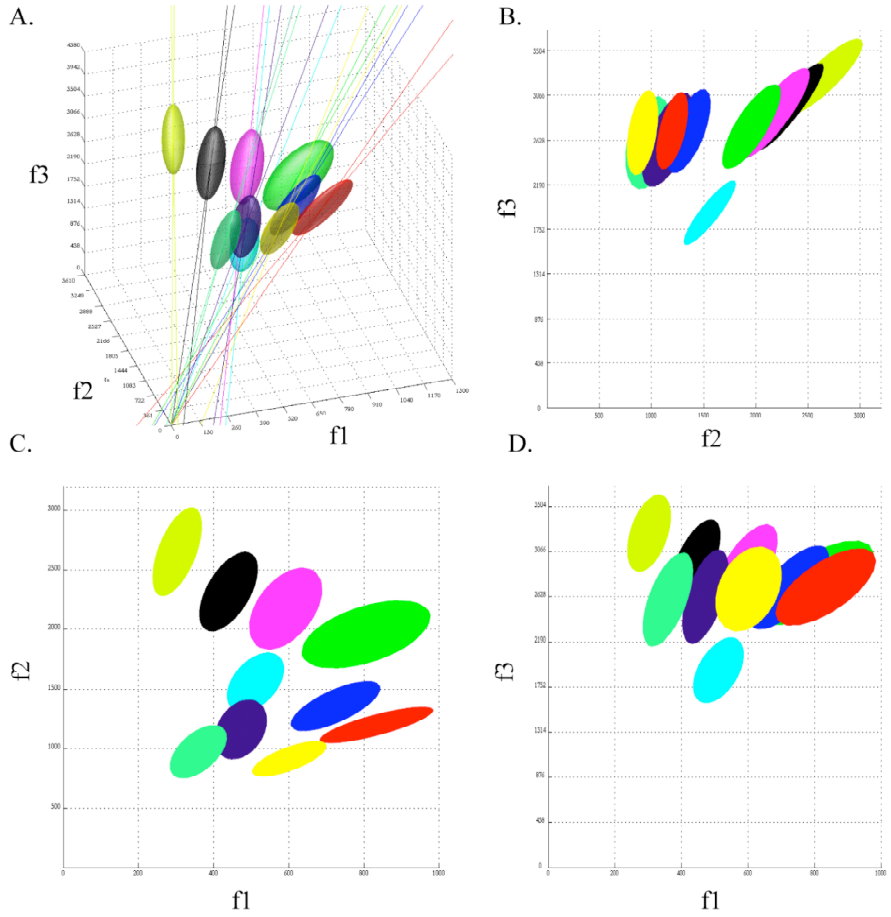

Figure 1: Maximum likelihood Gaussian cluster contours plotted at one standard deviation in each direction (to enclose 30% of cluster points). A. 3d perspective, two lines are drawn for each cluster, one showing the major axis of the ellipsoid, the other the predicted scaling line (which passes through the origin and the mean of the data). The angle between these two lines is ∏. All clusters are elongated toward the predicted direction. B. f_2 / f_3 plane. C f_1 / f_2 plane. D. f_1 / f_3 plane.

Table 1. The angle, ||, between predicted and actual lines allows model comparison. The flattening measure is defined as:

$$Flattening = \frac{X \square \left(Y^2 + Z^2\right)^{1/2}}{X}$$

where X is the longest (major) axis and Z & Y the shorter (minor) axes. A Flattening coefficient of zero corresponds to a sphere, a value of one to an infinitely thin line. All clusters are highly elongated; they are more 'line-like' than 'sphere-like'.

Vowel	hod	who'd	hud	heed	head	heard	hid	hawed	hood	had
∏ ± 0.5 degrees	8	2	6	13	3	7	1	5	7	9
Flattening	0.72	0.73	0.76	0.74	0.81	0.78	0.82	0.8	0.78	0.67

consider two hypotheses concerning the vowel clusters: 1) Scale is the major source of variability for vowel sounds of a given type. 2) Scale is not a major source of variability. These hypotheses can be evaluated by considering the angle, //, the major axis of the ellipsoidal makes with the scaling line through the vowel cluster mean. Table 1 shows the values of //.

If scale is a dominant source of variation in vowel sounds, we expect the individual vowel vectors to lie on straight lines through the origin. We do, however, expect some variability in vowel formant frequencies between individuals which is independent of the scale variation. This will cause the vowels to have some distribution around the aforementioned line. The scaling hypothesis predicts that the major axis of the fitted ellipsoids will point at the origin of the space if scale is the most important source of variability. We approximate the probability distribution of // as a single tailed Gaussian, normalized between 0 and 90 degrees. Using these assumptions, the variance can be chosen to maximise the probability of the data. Alternatively, if scale is not a major source of variability, the cluster orientation will be random, and the angle // will take any angle (between 0 and 90 degrees) with equal probability. This is a uniform probability distribution. We use a Bayesian approach to compare these two hypotheses (Mackay, 2003). For the scaling hypothesis we find the variance to be 7.0 degrees and the scaling hypothesis is 8 powers of ten more likely than the uniform model. Sampling simulations confirm this result. The data is consistent with the hypothesis that scale variations are the major source of variability between spoken vowels of the same class.

3.2. A Natural Scale Space

For a given vowel, the data for men, women and children cluster at points progressively further out from the origin of the formant space. Vocal tract morphological differences (Yang & Kasuya, 1995) do not affect vowel cluster means significantly. However, the variances of the clusters increase from male (small variance) to female (medium variance) to child (large variance). This can be explained from the perspective of scale. Consider a simple recognition process which uses a vector as a template for a given vowel. The template vector points in particular direction, but there is variability in the way we all speak and so the /a/ vectors for different people will not be identical. The system must recognise vowels independent of this variability, and to do this it must define some region around the template

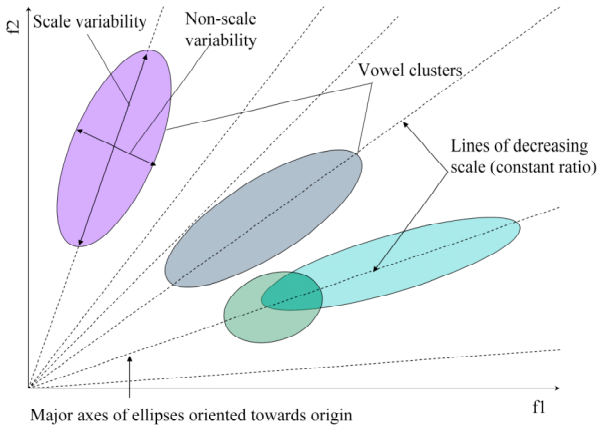

Figure 2: Illustration of expected vowel distribution in f_1/f_2 space. Vowels will form clusters around the scale trajectories. We might expect the variability due to scale to be much larger than the other sources.

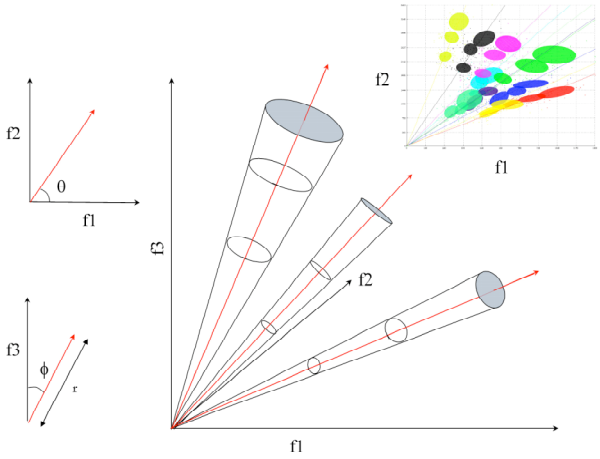

Figure 3: Centre: Predicted loci of vowel clusters. Left: Spherical polar representation. Top right: f_1/f_2 plot for the ten vowels showing clusters for each of the three speaker types. 1. Men (nearest the origin), 2. Women (in the centre), 3. Children (furthest from the origin).

vector in which /a/ vectors typically lie. If statistical fluctuations around the template vector are random, these classification volumes will form cones in formant space as illustrated in Figure 3. The cones get broader as we move further from the origin. Children can therefore afford to have more (absolute) variability in their vowel formants and still have them correctly recognised. They have more of the formant frequency space available to them. Clearly not all points in this cone are physically realisable – people's sizes vary within limits. Thus, we expect the vowel clusters to assume a more 'egg-like' shape; a combination, if you like, of our new cone and the

previous ellipsoidal distributions. As the symmetry of the ellipsoid matches the symmetry of an egg-like cluster, the results of the previous analysis remain valid.

These observations motivated us to use an alternate coordinate system to describe the formant space. This new space may not be utilised perceptually, but it facilitates understanding of the problems facing the auditory system. As Figure 3 illustrates, we can describe a formant vector, using spherical polar coordinates, with two angles (the direction, ∏ and ∐) and one distance (the magnitude, *r*). In this way, the vowel type information (angle) can be segregated from the VTL information (magnitude). By plotting the spherical polar coordinate values for the vowels on Cartesian axes, we now change the metric of the space, warping spheres centred on the origin into flat sheets which all have the same area. We have constructed the discrete analogue of a stabilized wavelet Mellin transform (Irino and Patterson, 2002) which segregates the scale information (vowel-vector magnitude) from the vowel type information (vowel-vector angle). The representation can be used to simulate scale normalisation as it might occur in the auditory system. It can also assist us in quantifying the importance of scale in sounds. The method can be generalized to any dimension.

4. Results: Clusters in the new space

In Figure 4 we see the results of this metric change. In the ∐-∏ plane (the two angular dimensions) vowel clusters are segregated. Clusters are well defined by their ∏ value (the angle relating to the f_1/f_2 plane). Two 'clusters of clusters' (Miller, 1989; Broad and Wakita,

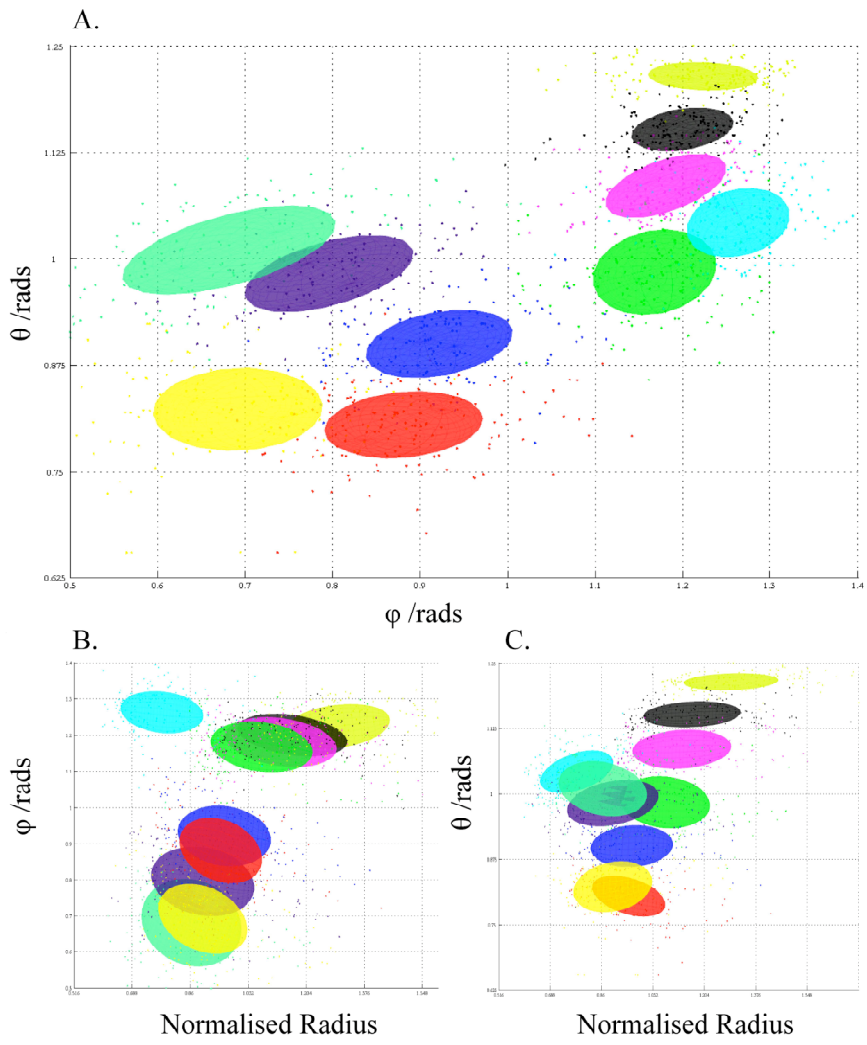

Figure 4: Spherical Polar Plots: The angles are denoted ∐ and ∏, the third axis is normalised scale (a rescaled version of the magnitude vector). Ellipsoids, generated from fitted Gaussian distributions, are shown for each cluster. The length of the clusters in the scale dimension should be identical if scaling of all cavities is uniform. The average length is found to be (0.29+/-0.04) indicating that scaling is uniform to a good approximation.

1977) are seen as determined by the angle ∐. The third dimension – normalised scale – is seen to carry very little vowel-type information, as expected.

This new representation can be assessed through application of a vowel classification process to both of our vowel spaces. Randomly selecting a vowel from the Peterson and Barney data set, we attempt classification using the remaining vowels as a library. We use the library to generate new Gaussian probability distributions which best describe the library data. The vowel we wish to classify is classified according to which cluster it is most probable that it arose from. The same procedure can be repeated for all vowels in the set and the percentage of successful classifications recorded. The $f_1/f_2/f_3$ space representation allowed 84% of the vowels to be correctly classified. Using just the two angle coordinates of the alternate space – ∐ and ∏ – a recognition rate of 79% was recorded. This suggests that the majority of the vowel-type information can been segregated using these two coordinates. We have reduced the dimensionality of the space. Giving the classifier access to the scale information increased recognition rates to 86%. These results should be contrasted with the success of Peterson and Barney's listeners. They managed 91% accuracy in their recognition tests.

5. Discussion

The classification procedure could undoubtedly be improved, for example, with the dynamic classification method of Miller (1989). However the results compare favourably with other studies (Syrdal and Gopal, 1985). We have also ignored level, duration, envelope and other properties of sounds known to aid vowel classification (Hillenbrand & Nearey, 1999; Hillenbrand et al., 1995). However, it is not the purpose of this study to recognise vowels but to illustrate the effect of normalisation on recognition. The classification results show that we have successfully segregated much of the vowel-type information, and they suggest that we have found a natural space in which to describe formant information.

6. Conclusions

We have quantitatively clustered the Peterson and Barney formant data using very general ideas about speaker variability. It has been shown that the orientation of the clusters is consistent with the hypothesis that scale is the largest source of variation in spoken vowel sounds. A scaling hypothesis is 8 powers of ten more likely than a model that posits no major source of vowel variability. Vowel clusters have been further broken down to show the volumes occupied by men, women and children in the frequency space. The size and position of these volumes together with scale considerations motivated us to consider an alternate space in which vowels can be represented. We have shown, by means of a simple classifier, that the new space segregates scale and vowel-type information into orthogonal coordinates and improves the recognition process. This further indicates the importance of scale variability in speech and suggests that a scale-normalisation tool, such as the Mellin transform, would aid speaker independent vowel recognition.

References

[1] Broad, D.J. and Wakita, H. (**1977**). "Piecewise-planar Representation of Vowel Formant Frequencies", *Journal of the Acoustical Society of America* **62**, 1467–1473.

[2] Flanagan, J. (**1972**). *Speech Analysis and Synthesis and Perception*, (Springer, New York).

[3] Hillenbrand, J.M., Getty, L.A., Clark, M.J. and Wheeler K., (**1995**). "Acoustic characteristics of American English vowels", The Journal of the Acoustical Society of America **97**, 3099-3111.

[4] Hillenbrand, J. M. and Nearey, T. M. (**1999**). "Identification of resynthesized /hVd/ utterances: Effects of formant contour", *Journal of the Acoustical Society of America* **105**, 3509-3523.

[5] Irino, T., Patterson, R. (**2002**). "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The Stabilised Wavelet Mellin Transform," *Speech Communication*, **36** (3-4), 181-203.

[6] Mackay, D. J. (**2003**). *Information Theory, Inference and Learning Algorithms*, Cambridge University Press.

[7] Miller, J. D. (**1989**). "Auditory-perceptual interpretation of the vowel", *Journal of the Acoustical Society of America,* **85**, 2114-2133.

[8] Patterson, R.D., Uppenkamp, S., Norris, D. Marslen-Wilson, W., Johnsrude, I. and Williams, E. (**2000**). Phonological processing in the auditory system: A new class of stimuli and advances in fMRI techniques. In: *Proceedings of the 6th International Conference of Spoken Language Processing (ICSLP)* Beijing, China, Vol. II. 1-4.

[9] Peterson, G. E. and Barney, H. I. (**1952**). "Control methods used in the study of vowels", *Journal of the Acoustical Society of America,* **24**, 75-184.

[10] Potter, R. K. and Steinberg, J. C. (**1950**). "Toward the Specification of Speech", *Journal of the Acoustical Society of America,* **22**, 807-820.

[11] Syrdal, A. K. and Gopal, H. S. (**1985**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels", *Journal of the Acoustical Society of America,* **79**, 1086-1100.

[12] Yang, C. and Kasuya, H. (**1995**). "Dimension Differences in the Vocal Tract Shape Measured From MR Images Across Boy, Female and Male Subjects", *Journal of the Acoustical Society of America* (Jpn) **16**, 41-44.