Inferring speaker size from speech formant data: A theoretical basis for vocal-tract length estimation

Richard Turner a)

Gatsby Computational Neuroscience Unit, 17 Alexandra House, Queen Square, London
WC1N 3AR
turner@gatsby.ucl.ac.uk

and

Roy D. Patterson

Centre for Neural Basis of Hearing, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, United Kingdom.
rdp1@cam.ac.uk

(Received 16 April 2007)

Running title: Estimating vocal tract length from formants
This paper investigates the theoretical basis for estimating vocal-tract length (VTL) from the formant frequencies of vowel sounds. Magnetic resonance imaging (MRI) of the vocal tract shows that the growth of the components of the vocal tract is highly non-uniform, so the formant ratios of vowels might be expected to vary considerably with speaker height. Moreover, analysis of the variability in formant frequencies suggests that there might well be another component of variability beyond vowel type and VTL. A statistical model was developed to describe the relationships between vowel type, VTL, formant frequency, and vocal cavity size. When the parameters of the model are learned from formant data, the model shows that VTL is the major source of variability after vowel type, and that the contribution due to developmental changes in oral-pharyngeal ratio is small relative to measurement noise. This suggests that speakers adjust the shape of the vocal tract as they grow to maintain specific formant-frequency ratios for individual vowels, and as a result, it should be a straightforward matter to estimate VTL from vowel sounds, given vowel type.

PACS numbers: 43.72.Ar

a) Corresponding author
I. INTRODUCTION

People are surprisingly good at segregating and tracking the speech of a specific individual in a multi-speaker environment. One of the more stable speaker characteristics is vocal tract length (VTL), and recent research indicates that the auditory system can assess the acoustic scale information in vowels and syllables (Smith et al., 2005; Smith and Patterson, 2005; Ives et al., 2006). This suggests that speech recognition systems would benefit from the inclusion of a pre-processor that employs some form of VTL estimation on a syllable-by-syllable basis, which provides a stream of VTL values for use as a tracking variable in multi-speaker environments (Turner et al, 2006). There is a problem, however, with the estimation of formant frequencies which makes it difficult to assess the contribution of vowel type, VTL, and other sources of variability such as oral-pharyngeal length ratio.

This paper describes the development of a statistical model of formant ratio theory (FRT) to describe the relationships between vowel type, VTL, formant frequency, and vocal cavity size. The paper begins with a deterministic, quantitative clustering analysis of Peterson and Barney’s (1952) classical formant-frequency data to determine the proportion of formant variability attributable to vowel type and VTL. The analysis suggests that there might be one more consistent and significant source of variability such as oral-pharyngeal ratio. The recent Magnetic Resonance Imaging data of Fitch and Giedd (1999) are then analyzed to determine the growth functions for the oral and pharyngeal cavities and the potential effect of the oral-pharyngeal ratio on formant frequencies. Finally, the deterministic version of FRT is extended to include the non-uniform growth observed in the oral-pharyngeal ratio and a more accurate representation of the distribution of vocal tract lengths in the population. This statistical, non-uniform version of FRT suggests that there are no major consistent sources of variability beyond vowel type and VTL in these data. This suggests that VTL is an excellent
candidate for a tracking parameter in multi-speaker environments, and that it should be possible to use non-uniform FRT to extract accurate VTL estimates from running speech.

II. VOWEL VARIABILITY AND FORMANT RATIO THEORY

This section presents a reanalysis of Peterson and Barney’s (1952) classic vowel formant data. The purpose is to quantify the proportions of inter-vowel and intra-vowel variability, and to assess the role of VTL in the intra-vowel variability. Briefly, the analysis reveals that about 80% of the total variability in formant frequencies is accounted for by vowel type, and VTL accounts for up to 90% of the remaining intra-vowel variability. This indicates the potential value of VTL normalization in speech recognition as noted previously by Welling and Ney (2004).

A. Vowel variability

1. Inter-vowel variability: vowel type

In their classic study, Peterson and Barney (1952) recorded two repetitions of 10 American vowels from 76 men, women and children and, from the spectrogram of each recording, they extracted the frequency of the first three formants and the pitch of the vowel. When the data were plotted in if F1-F2 space, the tokens of each vowel were found to cluster into relatively well defined regions that Peterson and Barney delimited with hand-drawn ellipses (their Fig. 8). In order to quantify the analysis, we have fitted three-dimensional Gaussian distributions to the F1-F2-F3 values of all of the tokens in each vowel cluster. The contours of constant probability associated with this distribution are ellipsoids; the contour
associated with one standard deviation along each of the axes has been plotted for each of the
10 vowels in Fig. 1. The formant frequencies have been converted into their corresponding
wavelengths ($\lambda_1, \lambda_2, \lambda_3$) because the focus of this paper is VTL and the analysis is more direct
when presented in terms of wavelengths.

The positions of the ellipsoids in wavelength space reveal the established observations
concerning inter-vowel variability: 1) There is virtually no overlap between the ellipsoids in
this space. 2) The separation between the clusters is significantly greater in the $\lambda_1$-$\lambda_2$ plane
than in the $\lambda_2$-$\lambda_3$ plane or the $\lambda_1$-$\lambda_3$ plane, indicating that the first two formants carry most of
the vowel-type information. 3) The back vowels and front vowels occupy different planes in
wavelength space due to the relatively high, and roughly constant, second formant of front
vowels (Broad and Wakita, 1977). The analysis shows that inter-vowel variability accounts
for about 80% of the total formant variability in the Peterson and Barney data.

2. Intra-vowel variability: vocal tract length

The intra-vowel variability is largely summarized by the eccentricity of the ellipsoid, its
orientation, and its distance from the origin. With regard to eccentricity, in each case, one of
the principal axes of the ellipsoid is much longer than the other two. This is largely because
the vocal tract increases in length as a child grows up. The eccentricity can be quantified with
the aid of a principal components analysis (PCA), and it shows that approximately 90% of the
intra-vowel variability lies in the direction of the major axis of the ellipsoid. With regard to
the orientation, the ellipsoids all point towards the origin of the space, as illustrated by the
lines in Figure 1; they show the extension of the major axis of each ellipsoid in the direction
of the origin (given by the major-eigenvector of the covariance matrix). Together the
eccentricity and orientation of the ellipsoids indicate that, within each vowel cluster, the
formant ratios are approximately constant. If the formant ratios were precisely constant for all
of the tokens of a vowel within a cluster, then the major axis of each ellipsoid would go through the origin of formant-wavelength space precisely. With regard to the distance of the clusters from the origin of the space, there is a trend for the major axis of the ellipsoid to increase with the distance of the centroid from the origin (for example, compare /uw/ with /ae/). It was these observations about eccentricity, orientation and distance that originally provided the basis for Formant Ratio Theory (Potter and Steinberg, 1950; and Miller, 1989).

The purpose of this section is to assess the degree to which this hypothesis accounts for intra-vowel variability, and the significance of any remaining discrepancies. If vowel type and VTL account for virtually all of the variability in the formant frequency data, then it is a relatively easy matter to estimate VTL given vowel type.

B. Formant ratio theory

Formant Ratio Theory was originally presented as a perceptual observation; Lloyd (1890) reported that the auditory system is more sensitive to changes in formant ratios than to changes in the absolute values of the formant frequencies (see Miller, 1989, for a review). With the invention of the spectrogram, it became clear that, for a given vowel, the formant ratios for men and women and children were approximately constant (Potter and Steinberg, 1950; Peterson, 1961). As a result, formant ratios have been used to normalize vowels in models of human speech recognition (Miller, 1989) and to develop a transform that improves the performance of computer speech recognizers (Cohen, 1993; Irino and Patterson, 2002; Umesh et al, 2002; Welling and Ney, 2004).

The mathematical form for the strict version of FRT is

\[
\tilde{x}^k = a_j \langle \tilde{x}^k \rangle, \tag{1}
\]
where $k$ is vowel type, $j$ is the speaker, and $\mathbf{\lambda}_j^k$ is a three-component, formant-wavelength vector derived from vowel $k$ of speaker $j$. $\langle \mathbf{\lambda}_k \rangle$ is the average vector of formant-wavelengths for vowel $k$ in the population. $a_j$ is the individual’s VTL divided by the average VTL and it is known as the uniform-scale factor. FRT is very simple; there is a single parameter for each formant and a single variable – the relative VTL of the speaker – that relates the formants of the speaker to the mean of the population. This is the uniform scaling model, and if it were true, taking ratios of formants would lead to exact vocal tract normalization. Peterson and Barney were not able to measure the VTL of their speakers; indeed, it is very difficult (Fitch and Giedd, 1999). However, FRT predicts how VTL is manifested in the ratios of formant frequencies, and so, it can predict the form of the vowel ellipsoids; that is, the orientation of each vowel cluster is determined by the direction of the principal component of variability, and FRT predicts this will be along a uniform scaling line where formant ratios are constant. The eccentricity is partly determined by the variability of VTL in the population and the distance of the ellipsoid from the origin of the space. In fact, it is simple to show that the relative lengths of the ellipsoids along their major axes are predicted to depend entirely on their relative distances from the origin of the space according to

$$
\sigma^k = \frac{\langle \mathbf{\lambda}_k \rangle}{\langle \mathbf{\lambda}_k \rangle} = \sigma_a \left\| \langle \mathbf{\lambda}_k \rangle \right\| 
$$

where $\sigma^k$ is the magnitude of the principal component of the vowel cluster (which is equivalent to the length of the ellipsoid along the major axis), $\sigma_a$ is the standard deviation of the scale factors in the population, and $\left\| \langle \mathbf{\lambda}_k \rangle \right\|$ is the magnitude of the vowel-cluster mean, or the distance from the origin of the space to the centre of the ellipsoid.

These two predictions can be used to confirm that VTL is the largest source of intra-vowel variability and to assess the accuracy of the formant ratio hypothesis. For example, the
angle formed between the major axis of each ellipsoid and the uniform scaling line for that ellipsoid indicates how uniform the scaling is. The angles are presented in Table I for the ten vowels in the Peterson and Barney database along with the proportion of the intra-vowel variance accounted for by the principal component. The angles are surprisingly small and they indicate that VTL accounts for about 90% of the variability not attributable to vowel-type. Here, then, is a quantitative basis for FRT from the classic data of Peterson and Barney (1952).

C. Residual variability: oral-pharyngeal ratio or measurement noise?

Although the analysis of formant variability indicates that FRT is largely correct, a detailed examination of Figs 1 and 2 shows that when the main axes of the ellipsoids are extended toward the origin, they converge on a point just to the side of the origin, which suggests that there might be one more factor making a small, but consistent, contribution to formant frequency. A clue to the form of the remaining variability is provided in Fig. 2 by the observation that the centroids of the sub-clusters for men, women and children are relatively widely separated on the uniform scaling lines. This means that VTL variability is greater between speaker groups than within speaker groups – an observation that has recently been confirmed by Gonzalez (2004). Moreover, the principal axes of these sub-clusters are more closely aligned with the first-formant axis than the uniform scaling line. This suggests that within speaker sub-clusters, there may be another consistent source of variability which is only revealed in conditions where VTL variability is small.

There are several candidates for the source of this effect: Fant (1966; 1975) suggested that speaker variability in formant ratios arises, at least partly, because the pharynx is proportionately larger in men than in women and children. He proposed a non-uniform
scaling procedure with separate scale factors for each formant of each vowel. The MRI data of Fitch and Giedd (1999) (Section III) confirm that the pharynx is proportionately larger in men, but this does not immediately indicate how the scale factors would be affected by VTL. Subsequently, Umesh et al. (2002) showed that Fant’s scale factors could be averaged across vowels to form a single non-uniform scaling function that describes the scale factor as a function of formant frequency. In both cases, the implication is that there is one main latent variable in this system which is speaker size, but that this variable affects different formants in different ways, necessitating extra parameters to be added to the uniform-scaling model of FRT.

D. Summary

A clustering analysis, equivalent to a principal components analysis, has been applied to the classic formant data of Peterson and Barney (1952). The results show that 80% of vowel formant variability is accounted for by vowel type. The centroids of the vowel clusters and their distance from the origin represent the inter-vowel variability. Although Peterson and Barney were not able to measure VTL, this latent variable is evident through the correlations it causes between the formants of vowels. The analysis shows that up to 90% of the remaining intra-vowel variability is attributable to VTL. Whereas vowel type determines which cluster the vowel formants will lie in, VTL determines where the vowel will lie within the cluster, so, for example, a child’s vowel will be closer to the origin those from a man. These results demonstrate quantitatively that FRT captures the majority of the intra-cluster variability in formant data, and this, in turn, suggests that it should be possible to estimate VTL from vowel-sized segments of continuous speech. At the same time, however, the analysis suggests that there might be another consistent factor in the data such as the change in oral-pharyngeal ratio with age, which might limit the accuracy of VTL estimation.
III. NON-UNIFORM GROWTH OF THE VOCAL TRACT

Fitch and Giedd (1999) have recently used Magnetic Resonance Imaging (MRI) to examine the growth of the vocal tract as children mature into adults, and their data make it possible to evaluate the anatomical basis for the non-uniform FRT hypotheses of Fant (1975), Umesh et al. (2002) and others. Briefly, the analysis shows that the pharyngeal cavity does, indeed, grow at a significantly greater rate than the oral cavity. The analysis also shows, however, that the growth rates are linearly related to VTL, which suggests that the non-uniformity of the growth need not complicate the modelling unduly.

Fitch and Giedd (1999) used Magnetic Resonance Imaging (MRI) to measure the lengths of the vocal tract cavities in 129 men, women and children ranging in age from 2.8 to 25 years. They recorded each subject’s age, height and weight, but they did not record samples of their speech. The measurements were made with subjects in the nasal breathing posture, and care was taken to exclude those who were overweight or whose families had a history of language or developmental problems. Figure 3 shows VTL as a function of height for all of the males (triangles) and all of the females (circles) separately; VTL is essentially a linear function of height in both cases. There are proportionately more men at the tallest heights, but the two populations fall on the same line. It is also the case that the vocal tract grows proportionately slower than height, because the head is proportionately larger than the body in children, but proportionality is the same for both groups. The growth rate is 0.67mm /cm.

Figure 4 shows the relative lengths of the oral and pharyngeal portions of the vocal tract as a function of VTL, separately for males (triangles) and females (circles). The figure shows that the length of the oral cavity decreases, and the length of the pharyngeal cavity increases,
relative to VTL, as VTL increases. This is because the size of the oral cavity is largely
determined by the size of the head which decreases as a proportion of body height as the
person grows up. Note, however, that the changes are linear in these coordinates and that for
a given VTL, there is no difference between males and females in terms of the proportions
of the cavities. This suggests that the model of VT growth need not be excessively complex.

IV. NON-UNIFORM FORMANT RATIO THEORY

The fact that the growth functions for the oral and pharyngeal cavities are linear (Section
III), and the fact that the size-related component of the formant frequency function is linear
(Section II.B), suggest that it might be possible to estimate VTL using a relatively simple
‘latent variable’ model of formant frequencies which incorporates (a) the non-uniformity of
vocal tract growth, (b) the relationship between VTL and formant frequency, and (c) formant-
specific terms for the measurement noise.

A. Modelling vocal tract length

The model developed in this section contains a latent, or hidden, variable that represents
VTL, or equivalently, a general growth factor of the system. The VTL of a speaker will
increase with age throughout childhood and then become approximately fixed in adulthood.
It will also depend on the sex of the speaker beyond about age 12, when VTL becomes
somewhat longer in males relative to their height. Broadly speaking then, we might expect
the distribution of this latent variable to be multimodal in the population with clusters
corresponding to men, women and children.
It is assumed that the lengths of the various cavities and components of the VT are linearly related to the latent variable via the average length of the cavity, or component, and a weighting factor, which can be thought of as reflecting the growth rate of the cavity or component (see Eq. 3). For cavities like the pharynx, whose proportion changes with growth, the dependence is strong and the weighting factor is large. For components like the lips, whose proportion changes little with growth, the dependence is weak and the weighting factor is small. The weighting factors enable us to construct a model of the vocal tract where the growth is non-uniform but, nevertheless, a linear function of height. Mathematically, the model is

$$L^k = \langle L^k \rangle + a \frac{dL^k}{da} ,$$  

(3)

where $\langle L^k \rangle$ is the average VTL for people articulating the vowel $k$, and $a$ is the relative VTL of the individual. $\frac{dL^k}{da}$ is a constant that does not depend on the individual.

The total length of the vocal tract is the sum of the individual cavity lengths and component lengths,

$$L = \sum_k L^k = \sum_k \left[ \langle L^k \rangle + a \frac{dL^k}{da} \right] = \langle L \rangle + a \frac{dL}{da} ,$$  

(4)

and so it is linearly dependent on $a$ as well. This relation can be substituted back into Eq. 3 to eliminate the latent VTL variable, $a$, and produce an expression for the ratio of a cavity or component’s length to the total length of the vocal tract, $L$.

$$\frac{L^k}{L} = \left[ \langle L^k \rangle - \langle L \rangle \frac{dL^k}{dL} \right] \frac{1}{L} + \frac{dL^k}{dL} ,$$  

(5)

Thus, in this relatively simple model of VT growth, the growth of the individual cavities and components is predicted to be linear when plotted against the reciprocal of $L$, which is what was observed in Figure 4. There is variability in the data that the model does not absorb,
but there does not appear to be any consistent deviation as a function of height of the sort that would warrant including a quadratic, or higher-order, term in the model. Briefly, this can be established quantitatively by fitting N\textsuperscript{th}-order polynomials to the data, and learning maximum-likelihood parameters and corresponding error-bars on these inferences. The linear model can then be compared to models of higher order, weighting the best-fit likelihoods of the more complicated models by penalty factors known as Occam factors, which depend on both prior knowledge and the error bars on the maximum-likelihood parameter estimates. In Bayesian statistics, this is a non-arbitrary form of hypothesis test (Mackay, 2003). In the current case, the linear model is found to be much more probable than models with higher-order terms. Indeed, the higher order approximations were found to have similar linear terms and only small contributions from the higher terms within the range of the data. This, then, is the justification for the non-uniform, linear model of VTL variability, which can now be used to deconvolve the effect of vocal tract changes on vowel formant frequencies.

B. Non-uniform formant ratio theory

The next step is to convert the model of vocal tract variability into a model of formant variability that can capture the dependence of formant frequency on the relative size of the speaker, \(a\). Broadly speaking, the higher formants are well modelled as simple standing wave resonances, so they will have wavelengths which are linearly dependent on the length of the vocal tract for a given vowel and formant. However, a simple standing wave is not a good model of the first formant. For example, the wavelength of the first formant can be as much as eight times the length of the vocal tract (e.g. Fig. 1), which is twice the maximum length that would be expected for a standing wave. Fant (1966) has argued that the first formant is often a Helmholtz resonance, in which case, the relationship between the frequency of the
first formant (which is what is measured) and the growth of the vocal tract might be expected to be more complicated. However, we analysed the correlations between formants for all of the vowels using the Bayesian techniques described in the previous section, and we found that the relationships are well approximated by a linear relationship. As a consequence, any pair of formants in a vowel is linearly related; that is,

\[ \lambda^k_i = m^k_{im} \lambda^k_m + c^k_m. \]  

(6)

This means that a fairly simple model might be expected to capture the majority of the variability in Peterson and Barney’s data, so long as it incorporates the model of vocal-tract growth derived earlier in Section III.A. A straightforward approach, consistent with the data, is to describe each resonator in terms of an effective length that is a simple linear function of VTL, regardless of its physical complexity. That is, \( \lambda^k_i = n^k_i L^k_i \). Each of the effective lengths might be expected to develop in exactly the same way as the physical dimensions of the vocal tract (Eq. 3), in which case the predicted relationship between formant wavelengths is linear, exactly as observed previously in this section. This description can be generalized to the three component vowel-vectors;

\[ \lambda^k = c^k + a m^k \text{ where } c^k = n^k \circ \langle L^k \rangle \text{ and } m^k = n^k \circ \frac{d\langle L^k \rangle}{da}. \]  

(7)

The prediction of this model is that the vowel clusters will form on segments of lines oriented in the direction \( m^k \) with centroids at \( c^k \). If the growth rate of the vocal tract is uniform, then \( m^k \) and \( c^k \) are parallel and FRT is recovered as a simple limit. If the distribution of the VTL parameter, \( a \), is Gaussian then this model is equivalent to PCA and the analysis of Section II is recovered. However, as noted earlier, the distribution is not Gaussian; there are three distinct classes of speaker (men, women and children). Therefore a more sensible choice is a mixture of Gaussians, with a Gaussian component for each group.
Two versions of the statistical model were developed – distinguished by their assumptions concerning the source of the vowels in each vowel cluster. In the first, and simpler, version, the vowels in each cluster were treated as if they all came from different speakers, and thus the clusters can be fitted individually. In point of fact the vowels in the clusters are not independent with respect to VTL; each speaker contributes two tokens to each of the ten vowel clusters. The second version of the model incorporates this constraint, which, in turn, makes it possible to fit all the vowel clusters simultaneously. Although the inferred VTLs estimated with the second version of the model are almost certainly more accurate, the parameter values derived from the two models are very similar. Accordingly the discussion is restricted to the results from the second version of the model, and it is these values that are reported in Table I.

C. The variability of formant measurements

Having included the effects of vowel type and VTL, the question is whether the formant frequency data contain other consistent sources of variability, or whether the remaining variability is just due to measurement noise. In the latent variable model, it is necessary to put in an explicit term for the residual noise. When this model, with its noise term, is applied to the data the result is surprising; most of the remaining variability in the formant wavelength data is due to a consistent measurement error, and when the error is properly modelled, the uniform scaling model is observed to absorb most of the remaining variability. This indicates that, if there is another natural factor, then its effect is limited to a very small contribution.
The measurement error arises from the fact that it is difficult to estimate formant frequency values from a spectrogram, particularly for the first formant. LPC-based methods only guarantee accuracy of approximately a quarter of the glottal pulse rate (GPR) (Monsen and Engebretson, 1983; Vallabha and Tuller, 2002). Peterson and Barney’s method was less sophisticated; they used a simple weighted average of the harmonics, \( f_n \), in the neighbourhood of the formant (see Potter & Steinberg, 1950),

\[
\frac{\sum f_n \times a_n}{\sum a_n}.
\]

(8)

The method has similar restrictions on its accuracy as LPC methods; moreover, an analysis of the data shows that a curiously high proportion of the formant estimates (~ 20%) are integer multiples of the GPR as shown in Fig. 5. In such cases the formant frequency is assigned to a harmonic frequency, which it would only rarely be by chance, and suggests that many formants were defined by just one harmonic. It turns out that this consistent measurement error is the source of much of the remaining variability in the vowel formant data. The measurement noise is roughly the same in absolute terms for all of the formants and so, as a proportion, the effect is largest for the first formant and smallest for the third formant. In wavelength terms,

\[
\frac{\sigma_{\lambda}}{\lambda} = n \frac{f_n}{f}.
\]

(9)

Similarly the error in formant ratios is larger for ratios where the first formant is in the denominator and smaller for those where it is not. The model of the noise in the formant wavelengths, or for formant ratios has to include these effects.

In the deterministic FRT model (Section II.B), the fact that the noise was not explicitly modelled leads to the bias noted in Section II.C; the axis of the ellipse shifts away from the
uniform scaling line. The effect of the noise is illustrated schematically in Fig. 6. In the stochastic FRT model, there is an explicit term for the noise associated with each formant,

\[ \hat{\lambda}^k = \pm{\lambda}^k + \alpha m^k + \eta^k + \sum_{i=1}^{I} b_i n_i^k, \]  

(10)

where \( \eta^k \) is the formant-specific noise term. It is a vector of zero mean Gaussian noise with covariance given by: \( \Psi^k = \text{diag}\left(\frac{1}{\sigma_i}, \frac{1}{\sigma_z}, \frac{1}{\sigma_t}\right) \). A factor was also added to capture any other consistent source of natural variability in the data, \( b_i^k \). This allowed us to assess the relative contribution from natural factors and noise, which revealed that, in effect, there were no other natural factors.

The analysis is a modified version of Factor-Analysis (FA) (Roweis and Ghahramani, 1999), where the distribution over the latent variable, \( p(a) \), is a mixture of Gaussians rather than a single Gaussian. The mixture of Gaussians was used to represent \( p(b) \) as well, but in this case the divergence from a simple Gaussian was minimal.

D. Learning the parameters and inferring VTL in statistical, non-uniform FRT

In order to infer the VTL of an unknown speaker from their formant data, the values of the model’s parameters must learned from the data, namely, the statistics of the relative vocal tract lengths \( [p(a)] \), the noise \( [\Psi^k] \), and the factor loadings \( [m^k] \). The learning and inference is accomplished using Bayesian methods, in particular, the variational Expectation-Maximisation algorithm (EM) of Ghahramani and Hinton (1996). This algorithm repeatedly optimises a lower bound on the likelihood in two steps: in the Expectation (E) step the algorithm infers the VTLs of the speakers, given the current parameter estimates; in the Maximization (M) step it finds the most likely parameters given the inferred speaker sizes.
The iteration of the two steps typically converges in the region of the maximum-likelihood estimate for the parameters (Ghahramani and Hinton, 1996).

The algorithm is free to find several components of variability that can point in any direction; the striking result is that for each vowel there is one component that absorbs most of the variability (see Table I), and that it invariably points towards the origin of the space as would be predicted by FRT. The orientation of each component of the current analysis was compared to the orientation of the corresponding component derived in Section II. The results are present in the lower row of Table I; they show that, if measurement noise and other sources of natural variability are modelled statistically, the component of variability attributable to VTL becomes more uniform, while the residual noise decreases correspondingly. Moreover, the model learns that the average measurement error is 50 Hz which is consistent with the inherent inaccuracy in formant extraction described in Section IV.C.

In Section IV.B, it was assumed that the population of VTLs represented by the data in the Peterson and Barney study is actually comprised of three sub-populations (men, women, and children) each of which is Gaussian. The distribution of extracted VTLs is presented in Fig. 7 together with the fitted mixture of Gaussians. The distribution is clearly tri-modal, justifying the assumptions of the model.

V. SUMMARY AND CONCLUSIONS

A Principal Components Analysis was used to cluster the classical formant-frequency data of Peterson and Barney (1952) and provide ellipsoids showing the distribution of formant frequencies associated with each vowel and population subgroup. The analysis revealed that vowel type accounts for 90% of the variability in formant frequencies, and 80% of the
remaining variability is accounted for by VTL. Sufficient variability remained to support the hypothesis that there was at least one more consistent source of variability, such as developmental changes in oral-pharyngeal ratio. The MRI data of Fitch and Giedd (1999) were reanalysed to evaluate this hypothesis, and the analysis confirmed that the growth is non-uniform, with the pharyngeal cavity growing faster than the oral cavity. However, the growth functions are linear. What is more, the growth functions for men, women and children are all the same. Despite the non-uniform vocal tract growth, there is no commensurate change in the formant ratios for individual vowels. This means that the systematic variability in formant frequency data (at least the first three formants) is effectively divided between vowel type and VTL, and consequently suggests that speakers adjust the shape of the vocal tract as they grow to maintain specific formant-frequency ratios for individual vowels. As a result, it should be a straightforward matter to estimate VTL from vowel sounds, given vowel type.

The analysis of the formant frequency data (Peterson and Barney, 1952) and the MRI data (Fitch and Giedd, 1999) suggested that the model for estimating VTL should contain, (a) a latent variable to absorb the variability of all size related factors, (b) non-uniform growth functions for the oral and pharyngeal cavities, and (c) separate measurement-noise terms for each of the formants. A modified version of factor analysis was developed for this purpose and fitted to the data using variational-expectation maximization. The model can be used to infer the VTL of an unknown speaker from formant data given the vowel type. The use of statistical methods to model the measurement noise reveals that the vast majority of the variability not attributable to vowel type is associated with VTL, and if there are any other natural sources of systematic variability their contribution is small with respect to the error in formant frequency estimation.
ACKNOWLEDGEMENTS

The authors would like to thank Thomas Walters for helpful discussions and help with the figures. We would also like to thank Dr T. Fitch for kindly providing the individual data on the lengths of the parts of the vocal tract from their MRI data. The research was supported by the UK Medical Research Council (G0500221; G000369) and the European Office of Aerospace Research and Development (FA8655-05-1-3043).


Lloyd, R. J. (1890). “Speech sounds: Their nature and causation (I),” *Phonetica Studien* 3, 251-278.


**TABLE 1.** The proportion of variability accounted for by the first component, and the angle that component makes with the uniform scaling line for the principal component analysis (TOP) and the latent variable model (BOTTOM).

<table>
<thead>
<tr>
<th>Vowel</th>
<th>/aa/</th>
<th>/ae/</th>
<th>/ah/</th>
<th>/ao/</th>
<th>/eh/</th>
<th>/er/</th>
<th>/ih/</th>
<th>/iy/</th>
<th>/uh/</th>
<th>/uw/</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA: variability in direction of the PC</td>
<td>0.91</td>
<td>0.94</td>
<td>0.88</td>
<td>0.94</td>
<td>0.87</td>
<td>0.97</td>
<td>0.99</td>
<td>0.89</td>
<td>0.91</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>PCA: Angle / degrees</td>
<td>11</td>
<td>14</td>
<td>6.0</td>
<td>7.5</td>
<td>8.5</td>
<td>13</td>
<td>6.6</td>
<td>5.6</td>
<td>3.9</td>
<td>8.0</td>
<td>8.4</td>
</tr>
<tr>
<td>LVM: variability in direction of the PC</td>
<td>0.85</td>
<td>0.95</td>
<td>0.97</td>
<td>0.85</td>
<td>0.96</td>
<td>0.70</td>
<td>0.96</td>
<td>0.87</td>
<td>0.98</td>
<td>0.83</td>
<td>0.89</td>
</tr>
<tr>
<td>LVM: Angle / degrees</td>
<td>5.2</td>
<td>4.6</td>
<td>4.0</td>
<td>7.3</td>
<td>2.6</td>
<td>5.8</td>
<td>1.8</td>
<td>8.2</td>
<td>6.8</td>
<td>6.3</td>
<td>5.2</td>
</tr>
</tbody>
</table>
FIGURE CAPTIONS

FIG 1. Ellipsoids showing the distribution of formant wavelengths associated with each vowel in the classical vowel data of Peterson and Barney (1952). The ellipsoids were derived using principal components analysis; they represent two standard deviations from the mean of the vowel clusters (enclosing 90% of the data points). The lines show the orientation of the major axes of the ellipsoids; their extensions point toward the origin of the space, but there is a consistent, small bias away from the origin.

FIG 2. Ellipsoids showing the distribution of formant wavelengths associated with the population subgroups (men, women and children) for six vowels in the classical vowel data of Peterson and Barney (1952). The ellipsoids were derived using principal components analysis; they represent two standard deviations from the mean of the vowel clusters (enclosing 90% of the data points). The lines show the orientation of the major axes of the ellipsoids shown in Fig. 1; the sub-clusters of each vowel are well separated along the scaling line, but the principal axes of these sub-clusters are more closely aligned with the first-formant axis than the scaling line.

FIG 3. Vocal tract length as a function of height from the data of Fitch and Giedd (1999). The upper and lower faint trend lines are the fits to the males (squares) and females (circles), respectively. The solid trend line is the fit to the male and female data combined.

FIG 4. Growth functions for the oral cavity (upper cluster) and pharyngeal cavity (lower cluster) from the data of Fitch and Giedd (1999). The abscissa is the reciprocal of VTL which orders the subjects according to size from left to right across the cluster. The ordinate is the relative length of the cavity. The trend lines show linear fits to the two clusters. The data of the males (squares) and females (circles) cluster along the same line in both cases.
FIG 5. Histogram of formant frequencies normalized to glottal pulse rate to illustrating the over-abundance of formant frequencies which are integer multiples of the glottal pulse rate (ie. GPR harmonics).

FIG 6. Illustration of how measurement noise can bias the orientation of the major axis of the ellipsoid derived with a traditional deterministic analysis of noisy data. The true data (circles) lie on a uniform scaling line (dotted line), but are corrupted by anisotropic measurement noise (squares). A traditional fit with principal components analysis derives a biased ellipse (dotted) with a principal axis which is biased away from the uniform scaling line.

FIG 7. Histogram of vocal tract lengths derived from the data of Peterson and Barney (1952) using non-uniform, statistical formant ratio theory. The mixture of Gaussians fit from the model is shown by the solid line along with the two Gaussians which make up the fit (dotted lines).