Estimating vocal tract length from formant frequency data using a physical model and a latent variable factor analysis. P61

Richard E. Turner, Thomas C. Walters, and Roy D. Patterson

ret26@cam.ac.uk
tcw24@cam.ac.uk
rdp1@cam.ac.uk

1. Introduction
- Formant frequencies can be estimated from individual voices.
- This poster shows how we can summarise the formant information in speaker size [vocal tract length (VTL)] which we propose as a tracking variable for speech recognition.

2. VTL influences in the data of Peterson and Barney (1952)
- Peterson and Barney (1952) recorded two repititions 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 voice.
- The formant frequencies have been converted into wavelengths because the focus of this poster is VTL.
- Id Gaussian distributions are fitted to each vowel cluster. A probability contour (an ellipse) is plotted at 1 std along each axis (Fig. 1).
- VTL accounts for 90% of the intra-vowel variability, but there is a consistent bias.
- Investigate VTL shape variability and noise.

3. VT shape variability is linear
- Fitch and Gielis (1999) used MBI to record the VTL dimmensions, heights and weights of 53 females and 76 males of different ages.
- VT shape – the ratios of VT sections to the total VTL – vary: the height and weight of 53 females and 76 males of different ages.
- Two factors account for 90% of the intra-vowel variability.

4. Formant correlations are linear
- We investigated the correlations between formants by plotting the 50 pairs of formants from the Peterson and Barney study (Fig. 4).
- Bayesian methods show each pair is best described by a linear model
- This is not surprising for standing wave resonances 
- However it is surprising for the Helmholtz resonances (typified by wavelengths much greater than 4 times the VTL of the speaker).

5. Measurement noise is important
- There is a known problem in extracting the first formant of sounds.
- Peterson and Barney used an unoptimised method to extract the formants and 20% were defined by only one pitch harmonic.
- The noise in a formant frequency measurement is therefore 1/3-1/4 of the pitch, but this has been ignored in previous studies.
- This biases formant ratios and the vowel clusters (Figs. 5 and 6).
- Measurement noise should be incorporated into the model.

6. An information-theoretic model and an application
- We developed a model of VTL and shape variability, formant physics, and measurement error.
- Assumptions:
  1. distribution of VTLS in the population is approximately Gaussian
  2. each formant of each vowel has a wavelength which is linearly dependent on the effective length (L) of the VTL
  3. the effective lengths are linearly related to size (a) of the individual ($\beta_3$)
  4. Gaussian noise ($\eta_3$) is present in each formant measurement making different contributions to each formant ($\eta_3=\eta_3^{\text{VTL}}+\eta_3^{\text{true}}+\eta_3^{\text{Gaussian}}$)
- This is a factor analysis model with a single latent factor – the size of the person – which causes the correlations in the formant wavelengths.
- The size of the person is encoded into formant wavelengths.
- Machine learning (Bayesian) methods can be used to decode this message in an optimal way (the EM algorithm).
- We found that formant scaling is much more uniform than previously thought.

7. Conclusions
- We have shown:
  1. VT shape variability is non-uniform but depends linearly on size
  2. Vowel formants are linearly correlated
  3. We developed a statistical model to allow for the possibility of formant error measurement. The model was fitted to the formant data using machine learning techniques.
  4. It indicates formant scaling is much more uniform than previously thought.
- Finally we presented a Bayesian algorithm for estimating the VTL of a speaker from formant frequency measurements.

References

Acknowledgements
Research supported by the U.K. Medical Research Council (G9901257, G9900369)