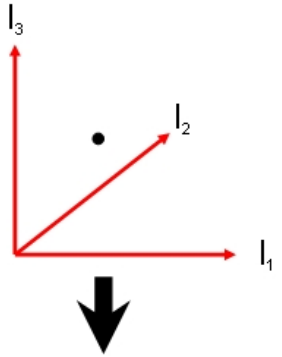
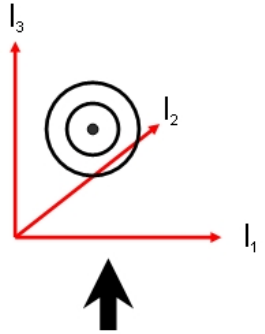
Vowel type, speaker size, and a bit of measurement error accounts for all the information in sustained vowels

Richard Turner (turner@gatsby.ucl.ac.uk)

Gatsby Computational Neuroscience Unit, 24/08/2007

Motivation





A Gross Approximation

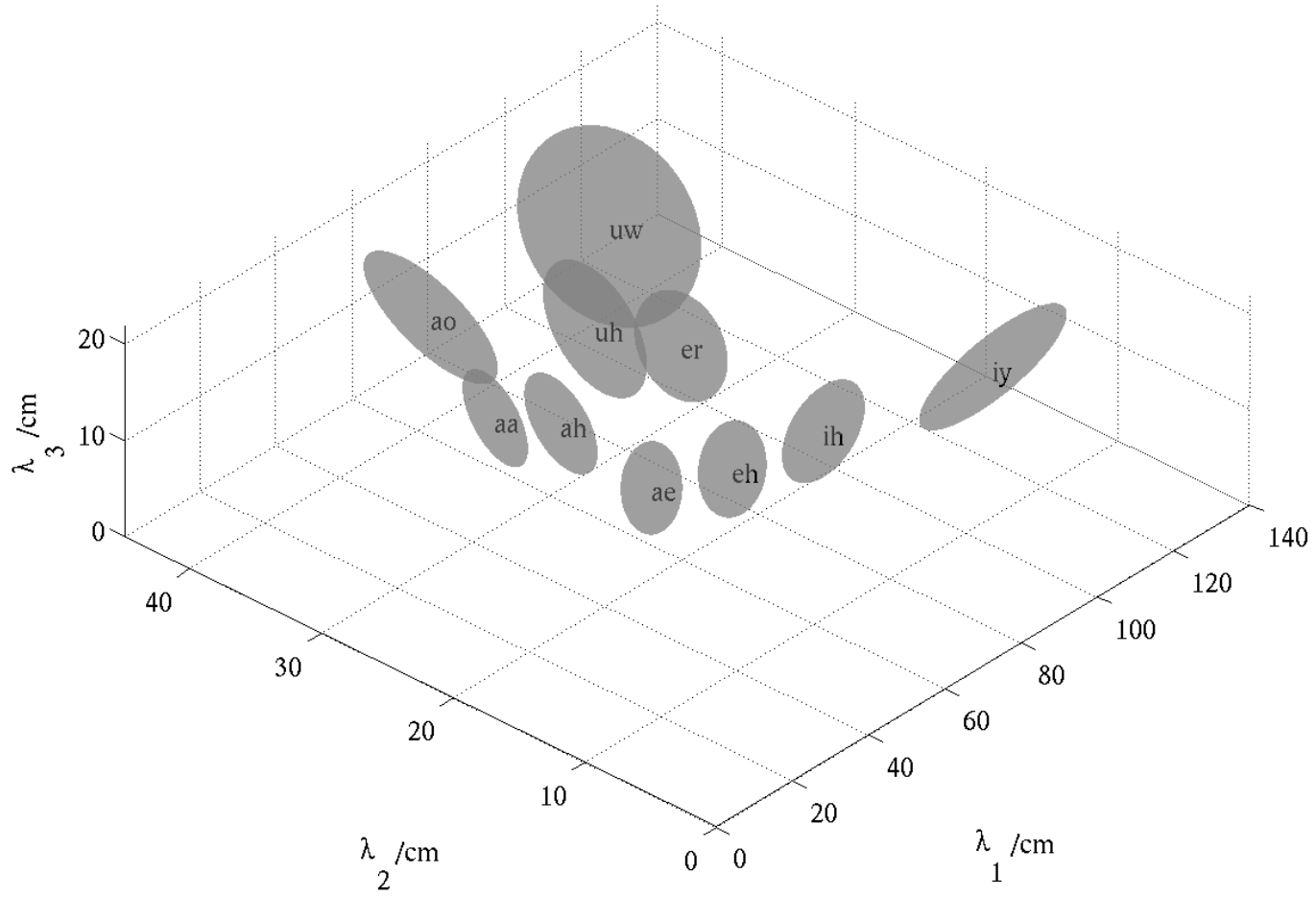
Three Approximations:

- **VTL growth**: All vocal tracts are the same shape. $l_i^n = a_{n,m} l_j^m$
- **Physics**: Formant wavelengths are all standing-wave-like. $\lambda_i^n = \alpha_i l_i^n$
- **Noise**: Formant frequencies can be recovered noiselessly.

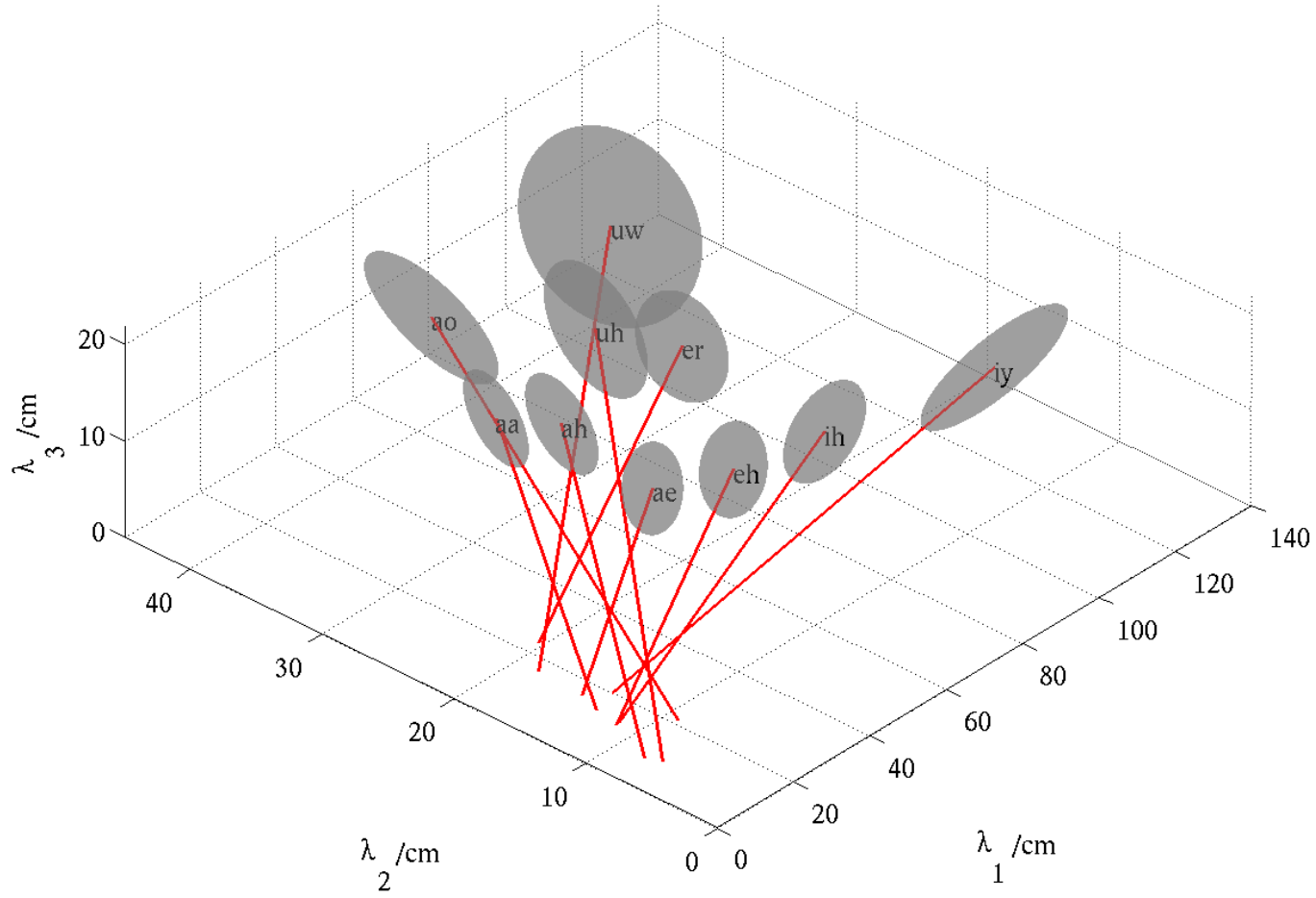
Two Predictions:

- Vowel clusters should be **pencils** pointing to the origin
- **Ratios of formants** should be speaker independent

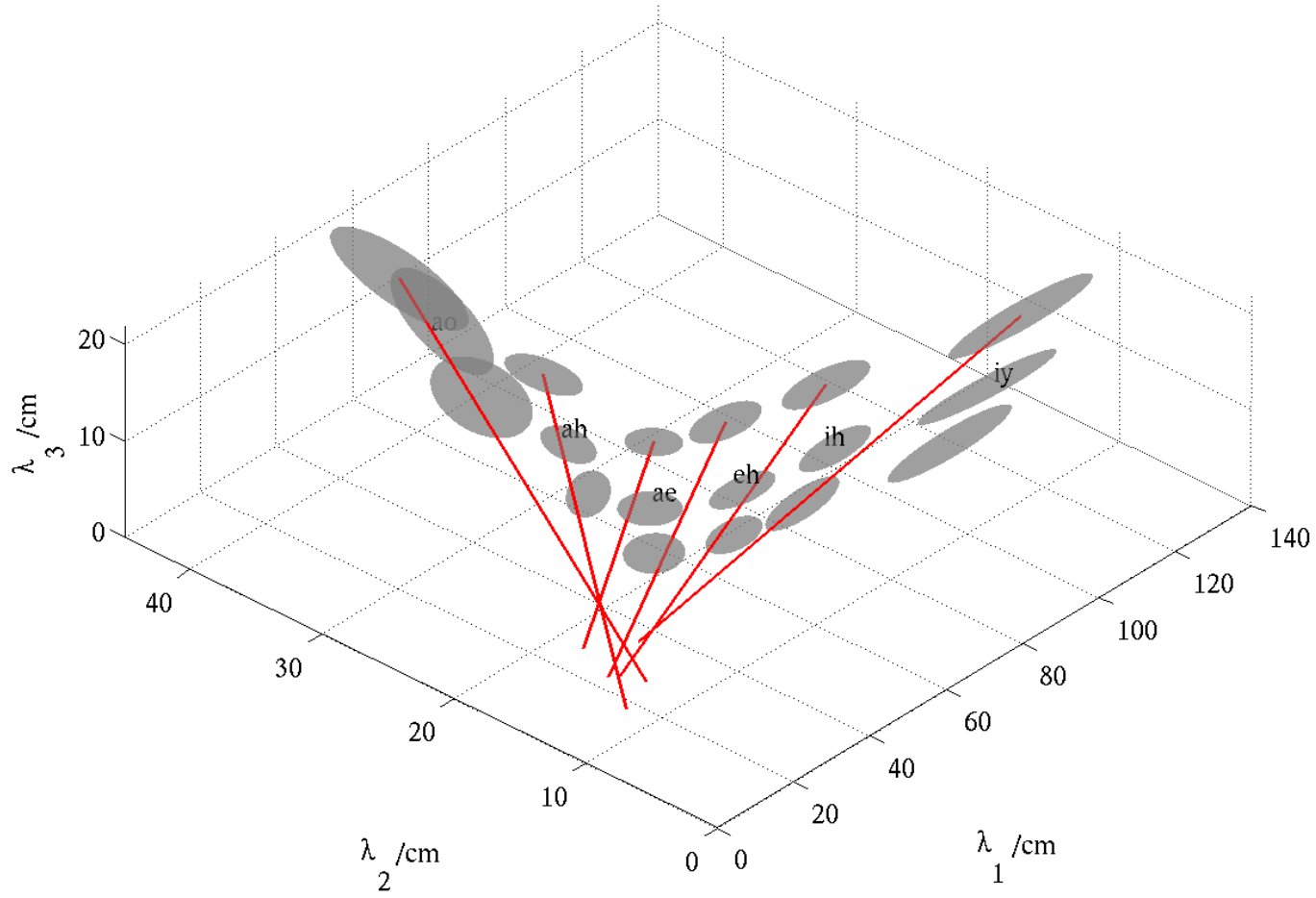
Clustering via PCA



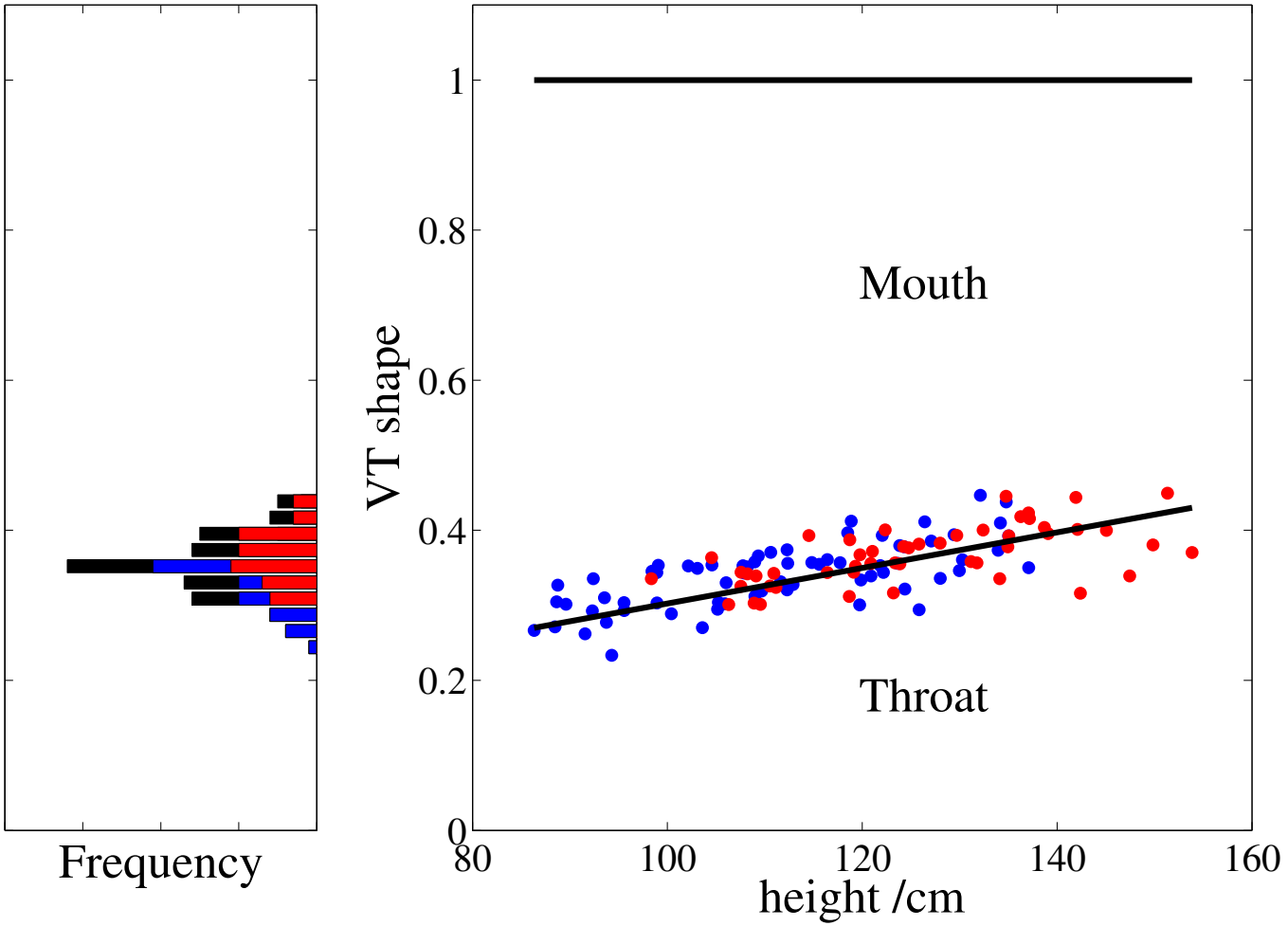
Clustering via PCA - $\langle\theta\rangle = 4.2$



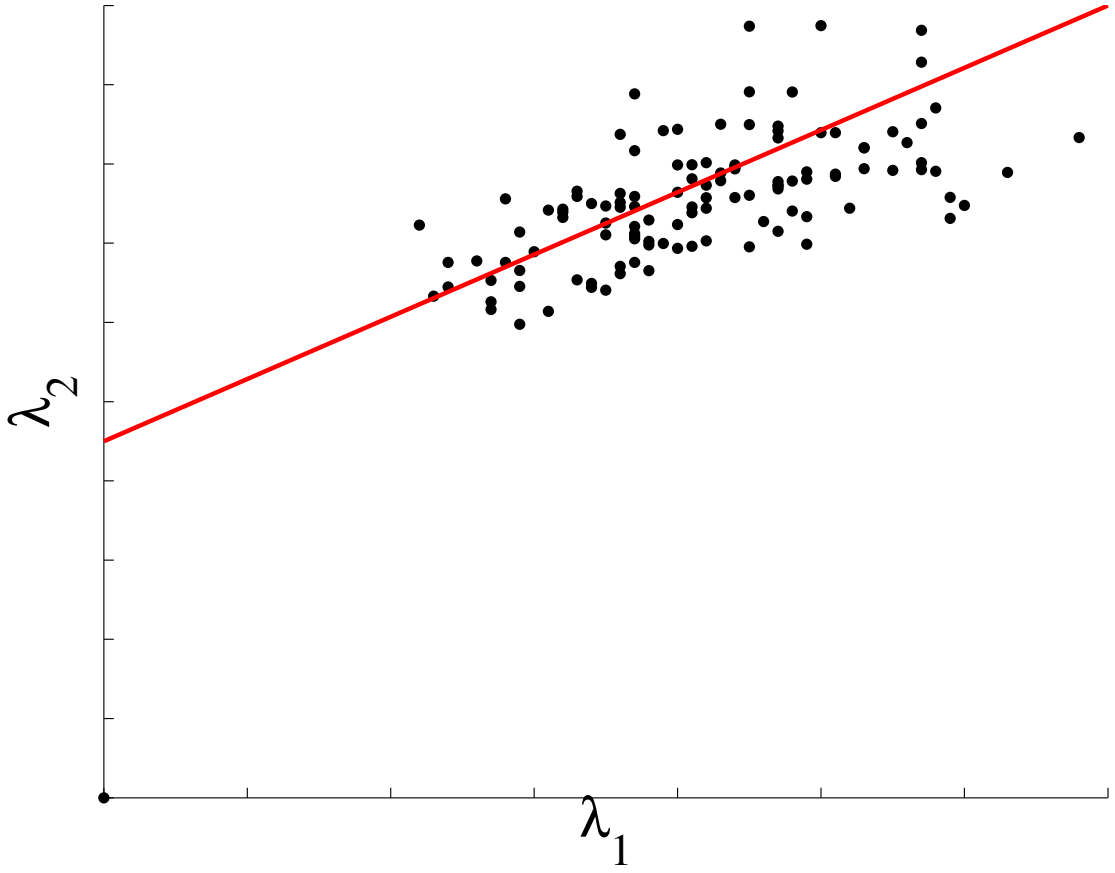
Clustering via PCA - $\langle\theta\rangle = 4.2$



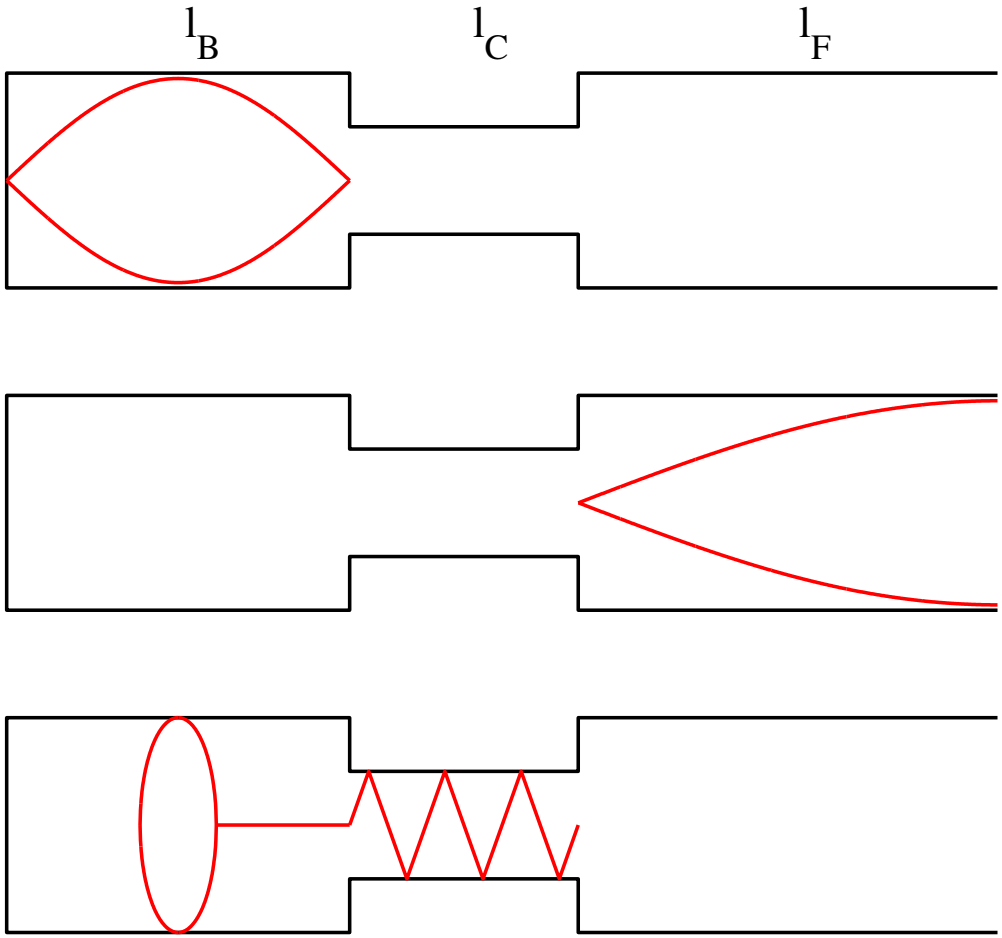
Are people scaled clones?



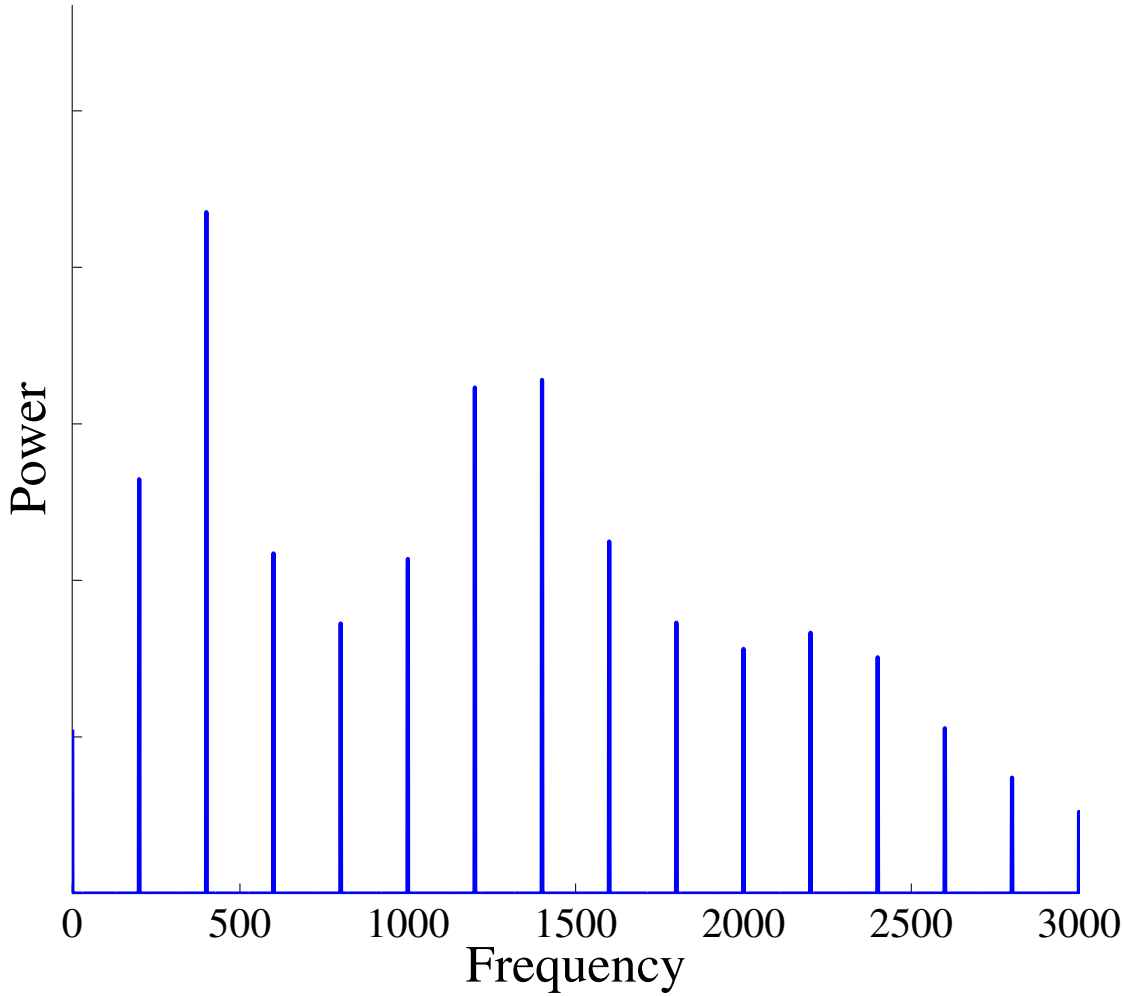
If people aren't clones, but the physics is simple



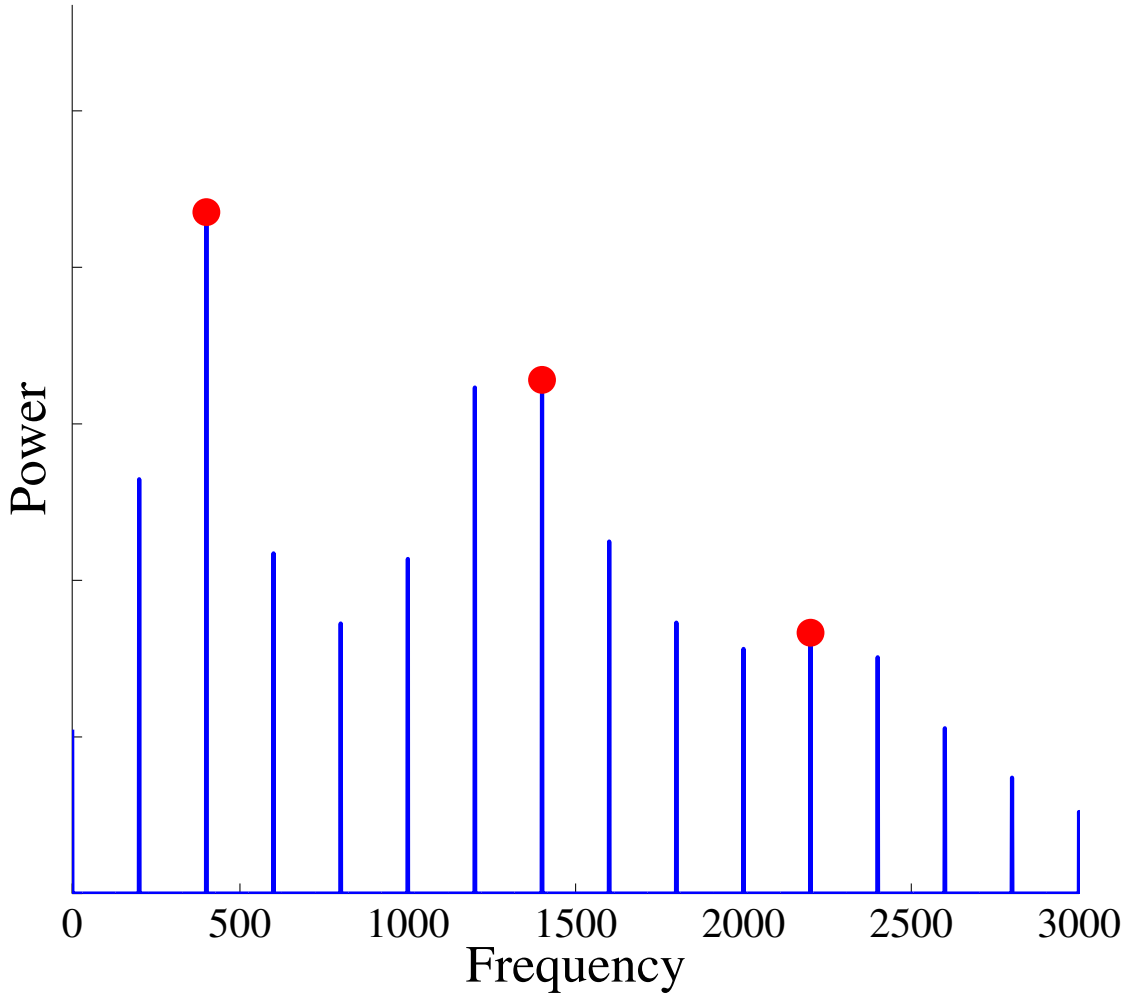
Is the physics standing-wave-like?



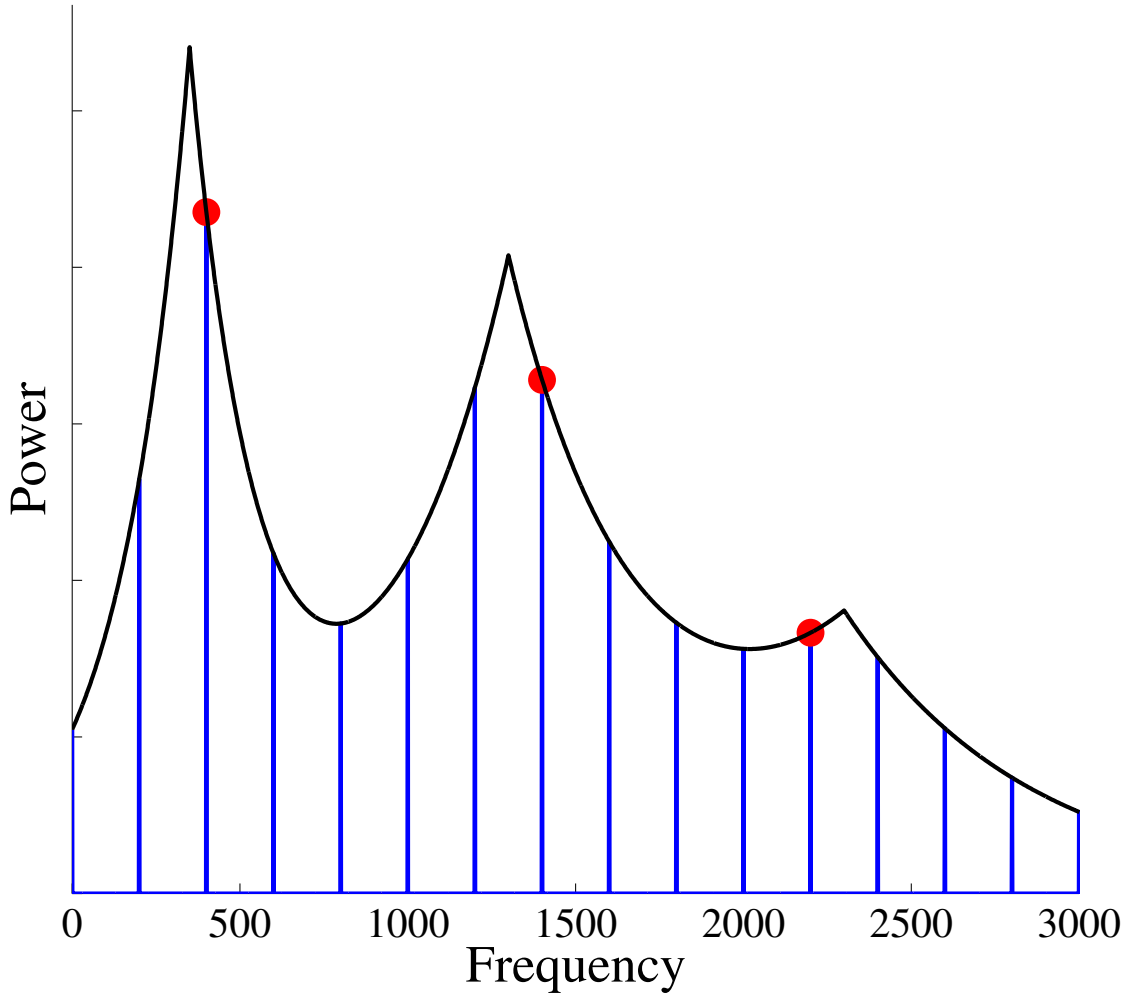
Are the measurements noiseless?



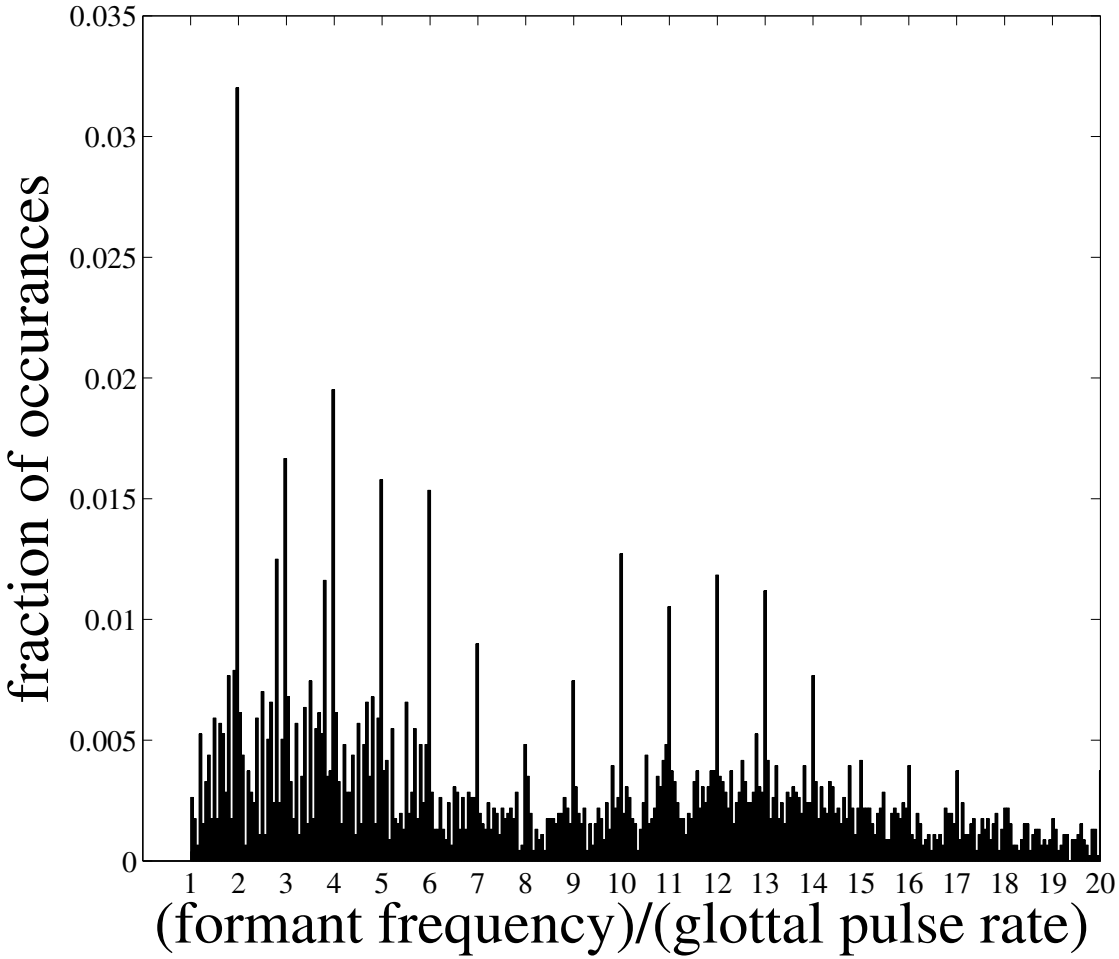
Are the measurements noiseless?



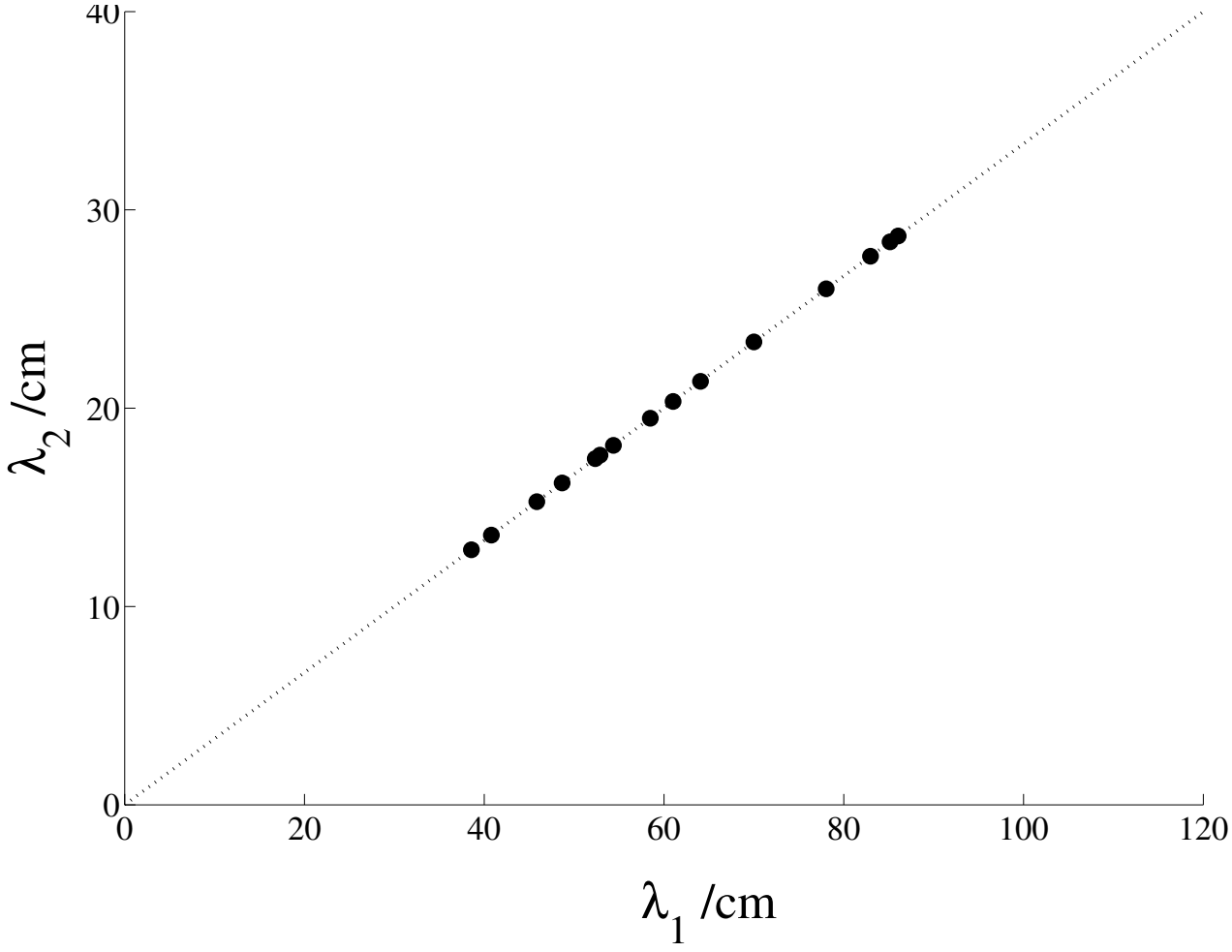
Are the measurements noiseless?



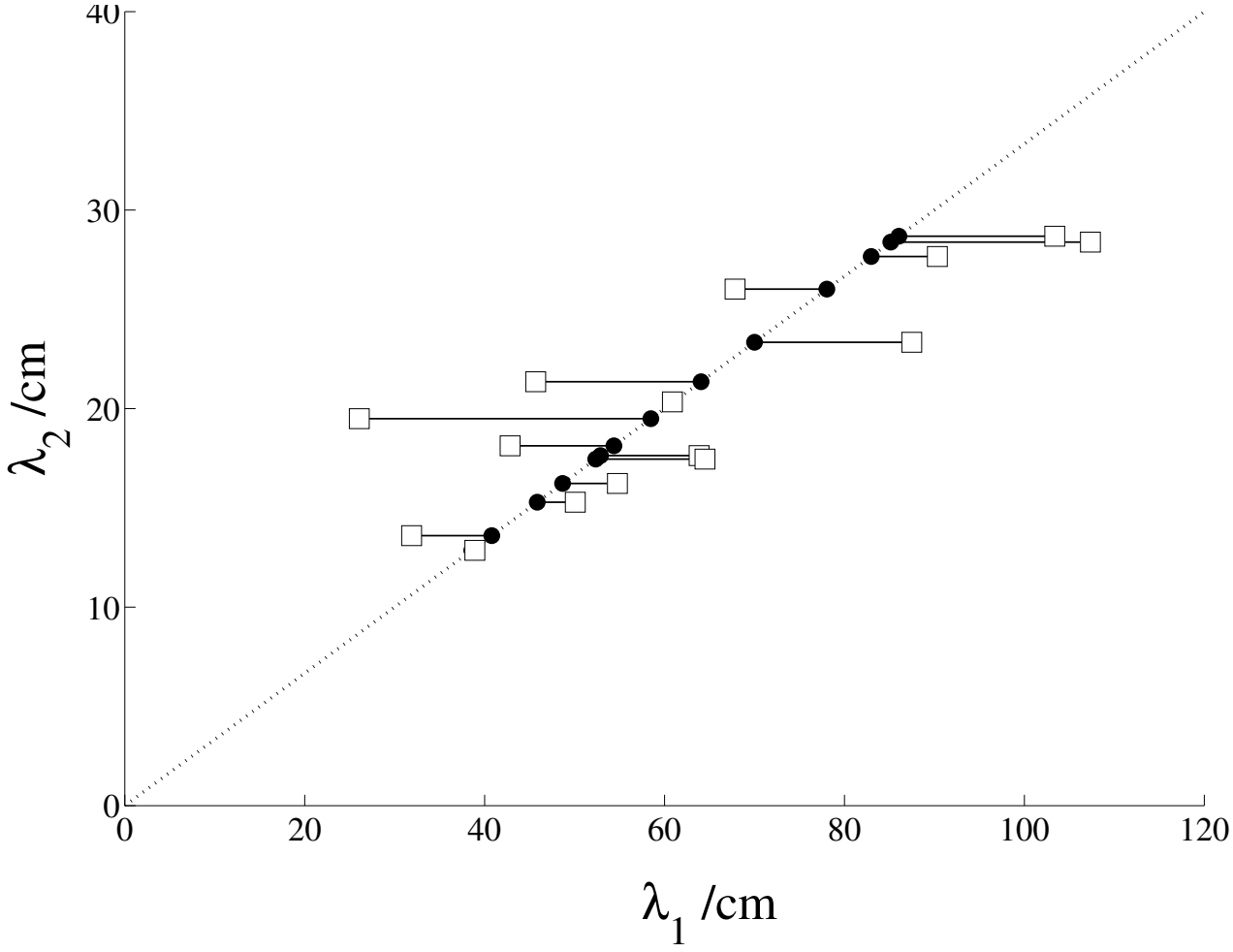
Is this true for the P&B data?



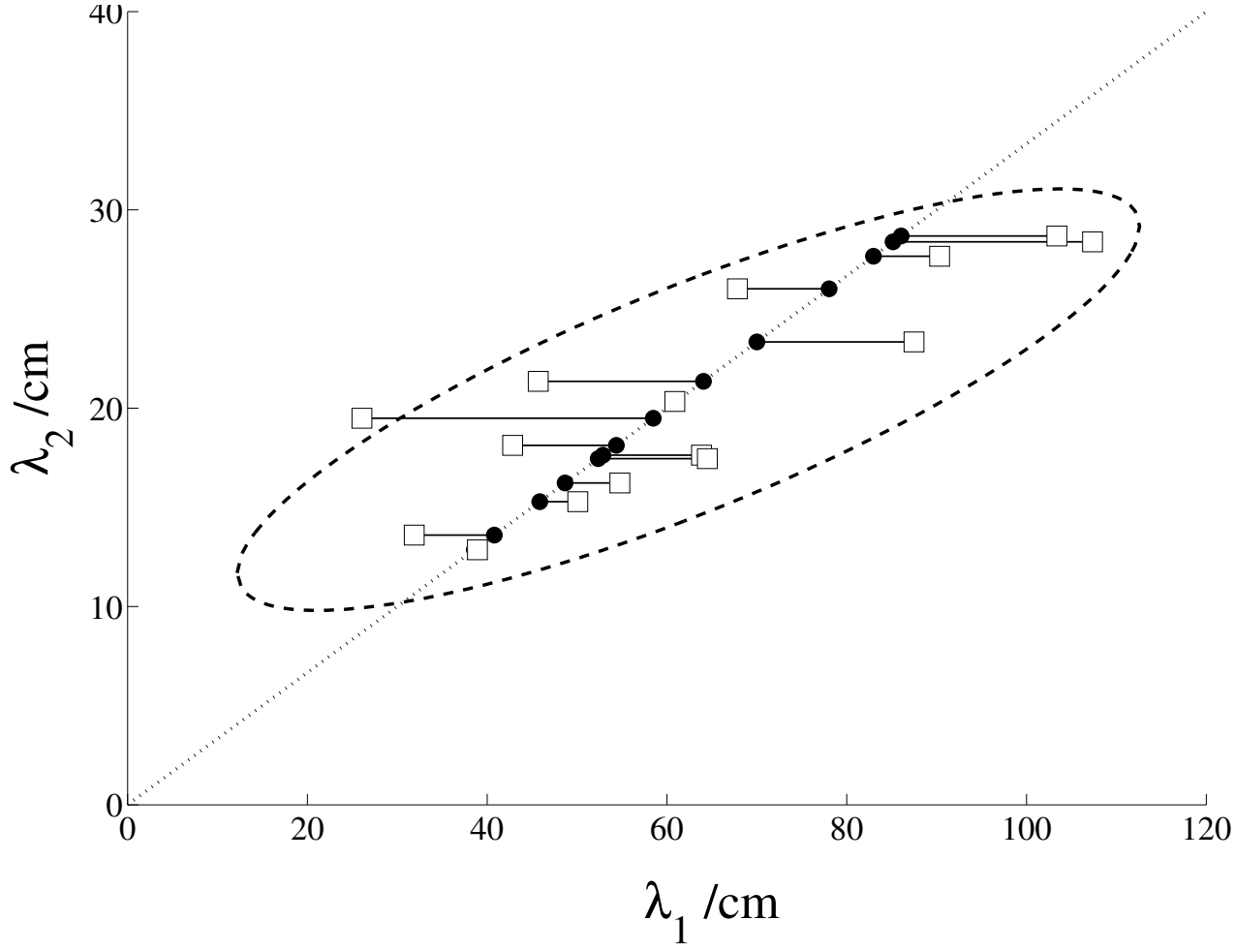
What effect does this noise have?



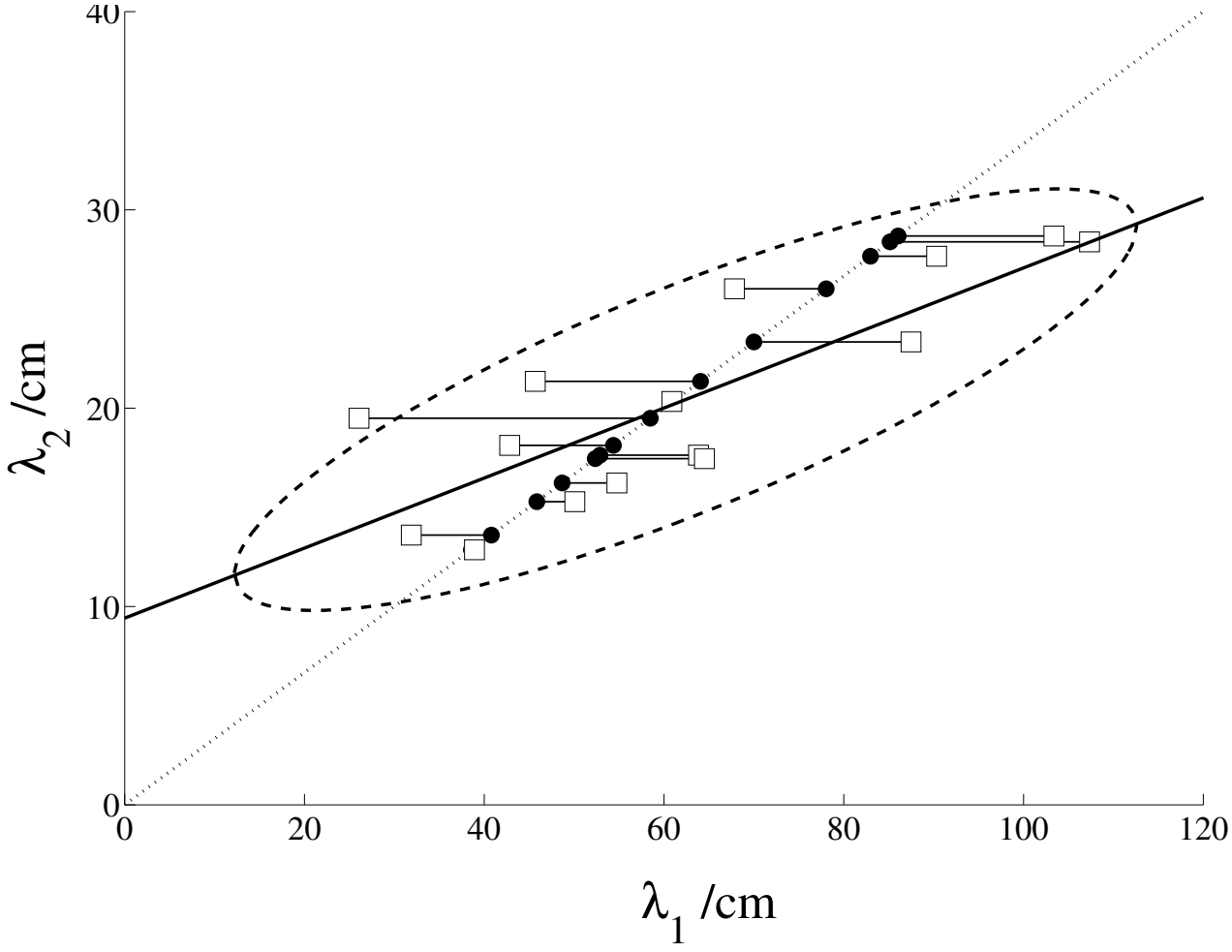
What effect does this noise have?



What effect does this noise have?



What effect does this noise have?



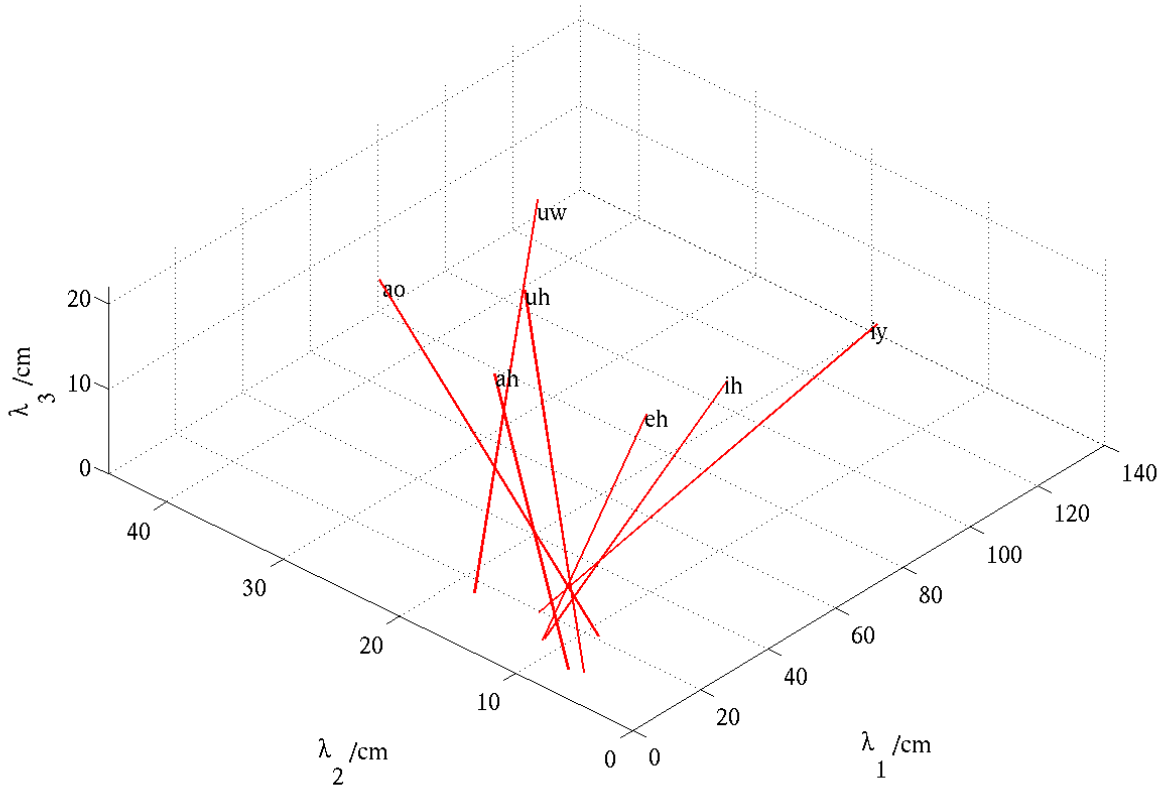
A new model - Noisy and VT Shape Varying

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix} a + \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix} + \sum_{i=1}^{I} \begin{pmatrix} h_{1,i} \\ h_{2,i} \\ h_{3,i} \end{pmatrix} a_i \qquad (1)$$

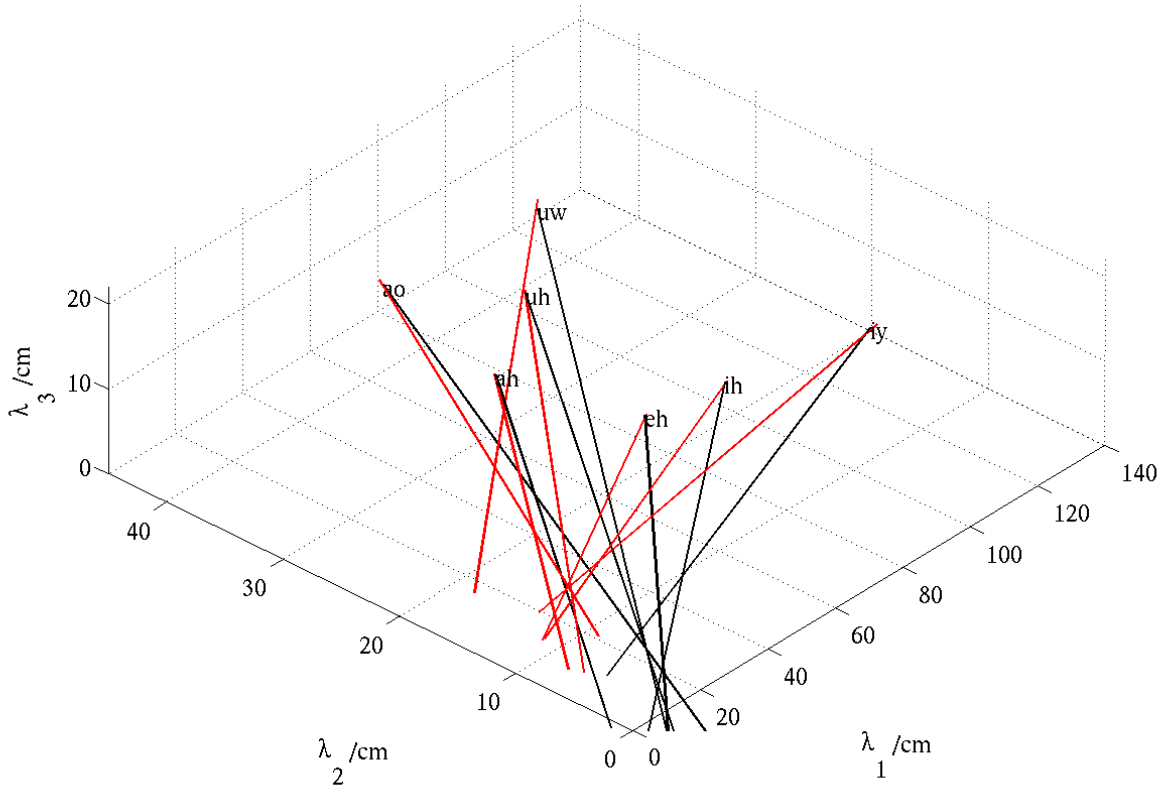
$$p(a) = \text{MOG} \qquad (2)$$

$$p(\epsilon) = \text{Norm}(0, \sigma^2) \qquad (3)$$

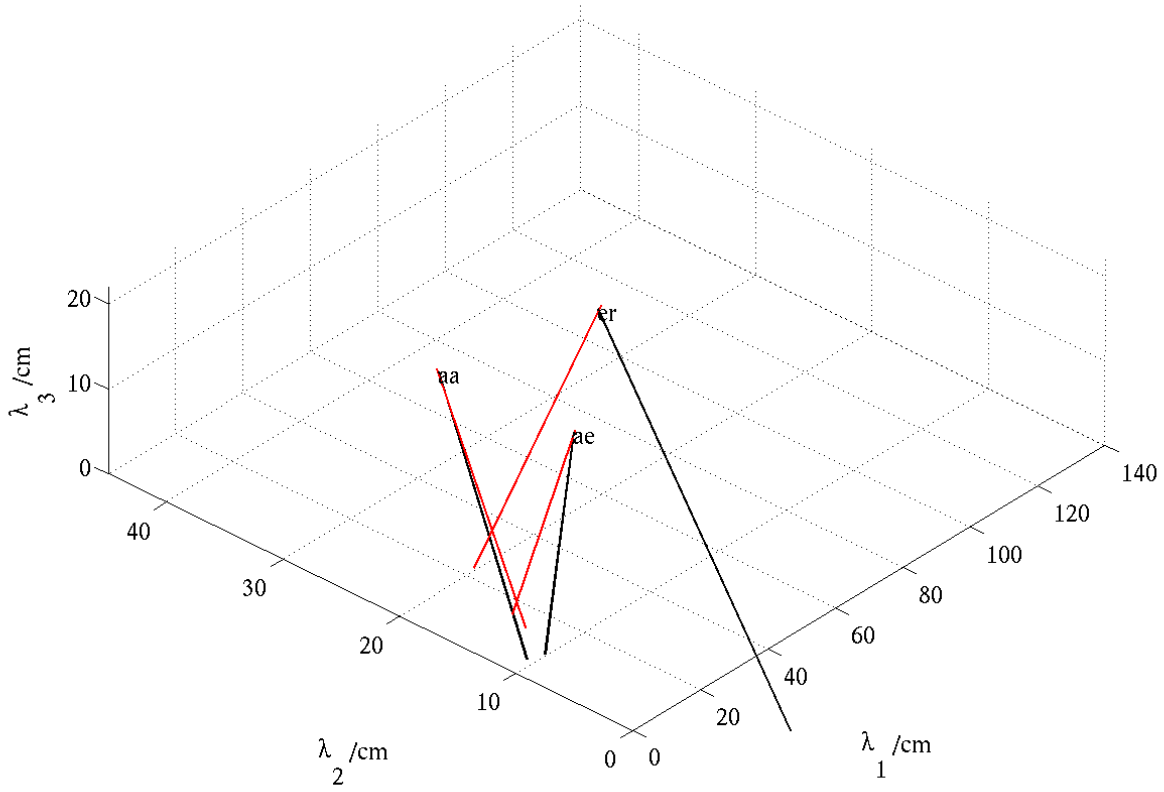
Results


uw
ao
uh
iy
ah
ih
eh
20
10
0
λ 3 /cm
40
30
20
10
0
λ 2 /cm
0
20
40
60
80
100
120
140
λ 1 /cm

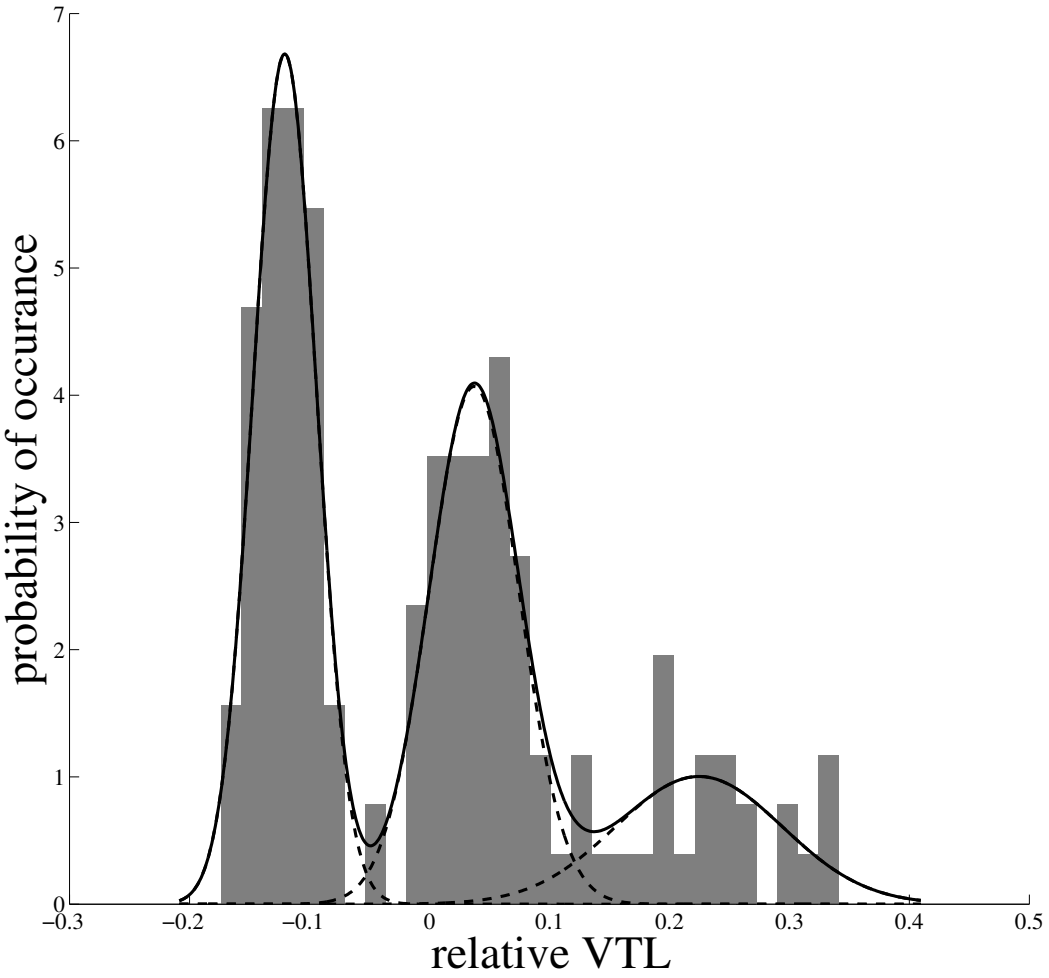
Results - $\langle\theta\rangle = 4.2$



Results - $\langle\theta\rangle = 4.2$



Results


probability of occurance
relative VTL
0
1
2
3
4
5
6
7
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5

Conclusion

- Formant Frequency data appear to contain a **considerable contribution from observation noise**.
- Once this noise is removed, the **remaining variability** is mainly in the form of a **uniform scaling**.
- Speakers appear to **compensate** for the non-uniform growth of the VT.