Statistical Analysis of Linguistic Data

Pitch information is an integral part of human language; however, a comprehensive quantitative model for pitch can be a challenge to formulate due to the large number of effects and interactions between effects that lie behind the human voice's production of pitch, and the very nature of the data being a contour rather than a point. Pitch can be characterised in a number of ways, and in particular the fundamental frequency F0 is of interest. This project will look at extending work on F0 using functional principal component analysis (see Aston et al, JRSSC, in press) to massive data sets where both estimation and prediction are of interest. Extensive data is available for Mandarin Chinese, particularly for read text rather than natural speaking. This project would primarily be interested in working out methods to apply FDA in a large data set with many thousands or even tens of thousands of functions. This will prove both a computational and theoretical problem, and there is extensive scope for both aspects to be considered. **The primary aim for this project is to produce a model for the data which is both linguistically informative and also has predictive power.** A certain amount of this project will be to build the data structures needed from the "raw" data (which has already been partially achieved) although this should be fairly simple with some computing background. In addition, careful choices of basis functions and methods will be needed to ensure the results correspond to meaningful linguistic output.

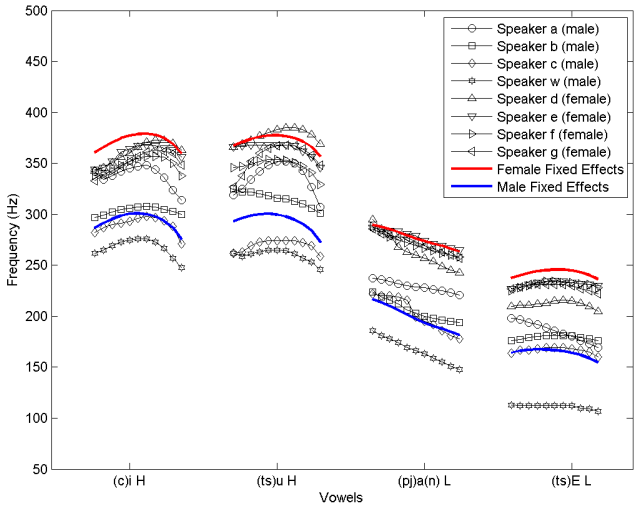

Figure 1: The graph shows the statistical model's estimated speech trajectories (using mixed effect models and functional data analysis, a technique for treating data as continuous curves in a function space rather than points) for males (blue) and females (red) when saying the word "riverbank" in the dialect Qiang, spoken in the Sichuan Province of China. Also on the graph are the actual speech recordings of 8 native speakers of the language, who seem to match well with the model. One aim of the project will be to extend this model to Mandarin Chinese

Abilities Required: Some knowledge of statistics will be useful. In addition, the project will require large scale computation, so some knowledge of MATLAB or R or other numerical language (include C/C++)

would be very useful. No prior knowledge of linguistics is required.

PhD Opportunities: This project very naturally leads into a PhD project where the data will be multiple dimensional (where not only the fundamental frequency but others are also considered). In addition, providing a valid theoretical and computational basis for cross language comparison would be challenging and very interesting extensions of the project. Note: This PhD project is also being offered to PhD applicants from outside Warwick as well.

References

J. A. D. Aston, J.-M. Chiou and J. E. Evans (2010) Linguistic Pitch Analysis using Functional Principal Component Mixed Effect Models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, in press.