Phonology **34** (2017). Supplementary materials

The phonetic specification of contour tones: evidence from the Mandarin rising tone

Edward Flemming
Massachusetts Institute of Technology

Hyesun Cho
Dankook University

Supplementary materials

Appendix: Estimating the parameters of the weighted-constraint model

1 Modelling errors

We estimated parameter values by Maximum Likelihood Estimation, i.e. our estimates are the values that maximise the probability of the data, given the model. To assign probabilities to data it is necessary to make assumptions about the probability of observing deviations from the model predictions, i.e. errors. We assume that errors in L, H and M are normally distributed, but not necessarily independent. For example, greater error in modelling H might tend to be accompanied by greater error in modelling L. In a model of a single variable with normally distributed errors, the maximum-likelihood parameter estimates minimise the sum of the squared errors in the fitted values (Myung 2003: 95). The current approach is similar, but the squared errors are weighted according to the variances and covariances of the errors in the three variables being modelled.

More precisely, the ith data point is a vector $\mathbf{y}_i$ of the values of L, H and M from a single utterance, and the model predicts a vector $\hat{\mathbf{y}}_i$ of fitted L, H and M values based on the observed segment durations. The deviations from these predictions, $\mathbf{y}_i - \hat{\mathbf{y}}_i$, are modelled as being distributed according to a multivariate normal distribution with mean equal to 0 on all three dimensions. The shape of the multivariate normal distribution is specified by its covariance matrix, Σ, a 3×3 matrix specifying the variances of the errors on each of the three dimensions, L, H and M, and the covariances between errors on different dimensions.

The probability density of one error vector, $\mathbf{y}_i - \hat{\mathbf{y}}_i$, is then given by the multivariate normal density function with mean 0 and covariance Σ (e.g. Johnson & Wichern 2007: 150). We refer to probability density rather than probability because L, H and M are continuous variables, and can thus take on an infinite range of values, so the probability of any specific value is zero. However, the probability of the value of a variable falling in a specified interval is non-zero, and is given by the area under the probability density function of that variable over that interval. Given the simplifying assumption that the observations are independent, the joint probability density of the whole data set is then the product of these individual probability densities, as shown in (8) (Johnson & Wichern 2007: 168), where $\mathbf{y}_i$ is the ith of n observation vectors, $\hat{\mathbf{y}}_i$ is the corresponding fitted value derived from the model, given the observed segment durations, and p is the number of variables, i.e. 3.

(8) $$\prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\ exp(-(\mathbf{y}_i - \hat{\mathbf{y}}_i)'\Sigma^{-1}(\mathbf{y}_i - \hat{\mathbf{y}}_i)/2)$$

If we change the parameters of the model then the fitted values $\hat{\mathbf{y}}_i$ change, and consequently so does the joint probability density of the data. The function that maps model parameter values onto the joint probability density for a fixed set of data is referred to as the likelihood function. The maximum-likelihood estimate of the parameter values is the set of values that maximises the likelihood function.

2 Estimating the error covariance matrix

However, the model parameter values are not sufficient to calculate likelihoods using (8) – we also need to know the error covariance matrix, Σ. But Σ is not known, and thus must be estimated from the data as well. To estimate both Σ and the other model parameters, we adopted the iterative procedure in (9), as in Iteratively Reweighted Least Squares (Goldstein 1986, Carroll & Ruppert 1988: 13ff).

(9) Starting from an initial estimate of Σ:
  a. find model parameters that maximise likelihood, given Σ;
  b. calculate an improved estimate of Σ from the residuals of the fitted values derived from those model parameters;
  c. iterate steps (a)–(b) until the parameter estimates converge.

To obtain an initial estimate of Σ, we need preliminary estimates of the errors in fitted values $\hat{\mathbf{y}}_i$, but we obviously cannot use the tone model to derive these estimates, since the estimate of Σ is needed first in order to fit the tone model. Instead we derive an initial estimate of Σ by fitting linear models to predict L, H and M from syllable and interval duration, and calculating the variances and covariances of the errors of these models. Syllable and interval durations are the only external (or 'exogenous') variables in the tone model, so they provide the only basis for preliminary models of L, H and M from which to estimate errors and thus error covariances. Linear models provide good estimates of L and H, and a linear model of M is adequate for present purposes.

3 Optimisation algorithm for maximising the likelihood function

With a preliminary estimate of Σ in place, numerical optimisation algorithms were used to search for the parameter values that maximise the likelihood of the data (step (a) of the iterative procedure in (9) above). In fact, we maximise the log of the data likelihood, because this is equivalent but computationally simpler. Maximising log-likelihood is not easy, because there are local maxima in the likelihood function (cf. Myung 2003: 94). We tried several optimisation algorithms, but the most successful was Generalised Simulated Annealing, a stochastic optimisation algorithm implemented in the R package GenSA (Xiang *et al.* 2013).

All numerical optimisation algorithms depend on good starting values for parameters in order to converge on optimal values (at least in a reasonable amount of time). Starting values were obtained by deriving expressions from the tone model that specify predicted relationships between L, H and M, and fitting those expressions to the data using non-linear least squares (nls; R Core Team 2016). These expressions are derived by taking partial derivatives of the cost function in (6) with respect to each variable (L, H, M), and setting them equal to zero, since the gradient of the cost function is zero at its minimum (cf. Flemming 2001: 20ff). For example, this procedure derives the expression in (7a), relating M to rise duration, $H-L$.

These expressions cannot be used directly to estimate the model parameters, because they relate predicted variables (e.g. L, H and M in (7a) are all outputs of the tone model), and because the same model parameters show up in more than one expression (e.g. T_M appears in the expressions for each of L, H and M), so fitting them separately would yield multiple estimates of the same parameter. However, fitting these expressions to the data provides good starting values for the parameters, averaging the multiple estimates of T_M.

The resulting starting values were:

(10) A_L 53% of syllable duration
A_H 80% of V-to-V interval duration
T_M 75 Hz
w_S 30,526
T_S 0.37 Hz/ms
w_H 0.5
w_L 0.19

GenSA was used to search in the ranges in (11) through 27,000 steps.

(11) A_L 40–70% of syllable duration
A_H 70–100% of V-to-V interval duration
T_M 60–90 Hz
w_S 1000–70,000
T_S 0.1–0.5 Hz/ms
w_H 0.1–1
w_L 0.1–1

The procedure was then iterated with an improved estimate of Σ (steps (b) and (c) of the procedure described in (9) above), but this resulted in no change in parameter estimates (the largest change was less than 0.005%).

Fitting the expanded model with pitch-range parameters, reported in §4.5, required three iterations between estimation of parameters and the covariance matrix to converge. The Nelder-Mead algorithm (implemented in the R function Optim) was used to estimate model parameters after the first iteration because the optimum was close to the starting point, and Nelder-Mead was much faster than GenSA under these conditions.

ADDITIONAL REFERENCES

Carroll, Raymond J. & David Ruppert (1988). *Transformation and weighting in regression*. New York: Chapman & Hall.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**. 43–56.

Johnson, Richard A. & Dean W. Wichern (2007). *Applied multivariate statistical analysis*. 6th edn. Upper Saddle River, NJ: Pearson Prentice Hall.

Xiang, Yang, Sylvain Gubian, Brian Suomela & Julia Hoeng (2013). Generalized simulated annealing for global optimization: the GenSA package. *The R Journal* **5:1**. 13–28.