20 July 2011
revised 28 July, 18 Aug

Recognition by rule 3:

Application of Continuous State Hidden Markov Models to a Classical Problem in Speech Recognition

Colin Champion

Summary

This paper is my third attempt to present correct equations for HMS decoding.

It sets out formally an algorithm for determining the changepoints (and optionally the phonetic correlates) of formants generated under a simplified version of the Holmes-Mattingly-Shearme model and measured subject to error.

Recognition by Rule 3:

Application of Continuous State Hidden Markov Models to a Classical Problem in Speech Recognition

Introduction

1. In 1964 John Holmes, Ignatius Mattingley and John Shearme (*HMS*) proposed a model for speech synthesis which they also hoped would also prove fruitful for recognition. Concentrating on vowels and similar sounds, they modelled each phoneme as having a set of formant targets, and the speech signal as produced by dwelling at the phoneme targets for a certain duration and then following linear formant trajectories to the next one. The lowest formant at one time moves to the lowest at the next, and so forth, so that formants never cross. A negative dwell time for elided phonemes allows the formants to change direction towards a successor phoneme before reaching their target.

2. Difficulties confronting their programme were encountered in the lack of good methods for tracking formants from audio and the absence of suitable algorithms for recognising the phoneme sequence even given a perfect set of formant recoveries.

3. The subsequent development and success of HMM algorithms for speech recognition has left the HMS programme in the shade. However it has often been observed that the HMS model enshrines structural properties which current HMMs can capture only by expanding their parameter space. In particular the HMS model accounts for the entire trajectory between two targets from the properties of its end points whereas conventional HMMs need the transition region to be divided into discrete units each modelled as a separate sound.

4. This had led to a schism between ‘speech scientists’ who are aware of the weaknesses of conventional models, and engineers who alone have viable systems. It has not been generally noticed that the algorithms underlying HMM systems provide the ideal framework for solving the second problem encountered by the HMS programme: that of determining the sequence of phonemes from the formant tracks. This may be called the ‘HMS decoding problem’.

5. The standard algorithms for processing HMMs are the Baum-Welch alpha, beta, and gamma calculations, the Baum-Welch reestimation procedure, and dynamic programming. In most treatments they are described as working on discrete state spaces, but in all cases they may be extended to continuous state spaces given suitable assumptions of normality. In this paper I shall show how the alpha pass can be used to find the optimal phoneme sequence corresponding to a set of formant tracks. I shall analyse a simplified version of the HMS decoding problem which is sufficient to illustrate the algorithm, omitting details which would add nothing of principle. At first I shall exclude phonemes from the model completely, assuming no more of formant targets than that they come from a random population. I shall limit the model to 3 formants and assume that all dwell times are zero, sacrificing the confirmation which comes from sustained vowels but also saving us from the difficulties caused by elided phonemes. I shall assume that the transition time from one target to the next comes from a known distribution and is independent of the phonetic content. I shall assume that we have formant observations at a discrete sequence of times, and that these are distributed about the true frequencies according to a gaussian error. This last

assumption must be slightly counterfactual since observations of formant frequencies which are negative or incorrectly sorted are not deemed impossible. I also treat the errors as independent, which will usually be the most significant error in the assumptions.

6. As a notational convention I shall write the normal pdf in the form

$$n(\underset{\sim}{x}, p) \;=\; (2\pi)^{-\dim/2}|p|^{1/2}\exp\{-\tfrac{1}{2}\underset{\sim}{x}^{\mathsf{T}} p\, \underset{\sim}{x}\} \tag{1}$$

where *p*, the *precision*, is the inverse of the covariance matrix.

Alpha-pass for scoring hypotheses

7. The first step of the algorithm will be to determine the *changepoints*, ie. the times at which which formants reach their targets: these are the points at which formant tracks change direction. An assumed set of changepoints will be evaluated by summing the probabilities of all formant tracks consistent with them, and multiplying the result by the prior probability of the implied transition durations.

8. The sum over formant tracks is a Baum-Welch alpha pass. We shall let the state be the ordered pair $(\underset{\sim}{x}, \underset{\sim}{s})$ where $\underset{\sim}{x}$ is the vector of formant frequencies at the most recent changepoint and $\underset{\sim}{s}$ is the vector of their slopes. A more natural parametrisation would work with the pair $(\underset{\sim}{x}_0, \underset{\sim}{x}_1)$ of formant sets at each end of a trajectory. This would be easier to implement but would require us to know the transition duration in order to extend a hypothesis beyond the time of $\underset{\sim}{x}_0$.

9. Let us write α_t for the sum over all formant tracks from time 0 to time *t* of the product of the probability of the formant track with the probability of the observations given the track. Let us assume that

$$\alpha_{t-1}(\underset{\sim}{x}, \underset{\sim}{s}) \;=\; k\, n((\underset{\sim}{x}, \underset{\sim}{s}) - \underset{\sim}{\mu}, p) \tag{2}$$

so that the alphas at time *t*-1 take the form of a scaled normal distribution with a mean $\underset{\sim}{\mu}$ (which may be broken down as $(\underset{\sim}{\mu}_{\underset{\sim}{x}}, \underset{\sim}{\mu}_{\underset{\sim}{s}})$) and a 6-by-6 precision matrix *p*. If at time *t* we are *h* time increments beyond the most recent changepoint, then the formant frequencies at *t* will be $\underset{\sim}{x} + h\underset{\sim}{s}$ and I shall assume that the measured frequencies $\underset{\sim}{y}$ are normally distributed about this point with precision *e*. Notice that *e* is 3-dimensional, and that we are mixing terms with different dimensionalities in our formulae.

10. If we are not at a changepoint, then

$$\begin{aligned} \alpha_t(\underset{\sim}{x}, \underset{\sim}{s}) \;&=\; k\, n((\underset{\sim}{x}, \underset{\sim}{s}) - \underset{\sim}{\mu}, p)\, n(\underset{\sim}{y} - (\underset{\sim}{x} + h\underset{\sim}{s}), e) \\ &=\; k'\, n((\underset{\sim}{x}, \underset{\sim}{s}) - \underset{\sim}{\mu}', p') \end{aligned} \tag{3}$$

where

$$p' \;=\; p + \begin{pmatrix} e & he \\ he & h^2 e \end{pmatrix}, \tag{4}$$

$$\underset{\sim}{\mu}' \;=\; \{(\underset{\sim}{y}e, h\underset{\sim}{y}e) + \underset{\sim}{\mu}p\}(p')^{-1}\text{, and} \tag{5}$$

$$k' \;=\; \frac{k\,|p|^{1/2}|e|^{1/2}}{(2\pi)^{3/2}|p'|^{1/2}}\exp\{-\tfrac{1}{2}[\underset{\sim}{y}^{\mathsf{T}} e\underset{\sim}{y} + \underset{\sim}{\mu}^{\mathsf{T}} p\underset{\sim}{\mu} - \underset{\sim}{\mu}'^{\mathsf{T}} p'\underset{\sim}{\mu}']\} \tag{6}$$

where these formulae are obtained by completing the square, and their details are left as an exercise for the reader.

11. Now suppose we come to a changepoint time *t*. The above formulae give us the alpha value for a state comprising the target frequencies $\underset{\sim}{x}$ at the <u>previous</u> changepoint and the slope $\underset{\sim}{s}$ of the trajectory from it. We want to rewrite the state in terms of the new target $\underset{\sim}{x}' = \underset{\sim}{x} + h\underset{\sim}{s}$: the sum of path probabilities leading to this target is

$$\begin{aligned} & \int \mathrm{d}\underset{\sim}{s}\ \alpha_t(\underset{\sim}{x}' - h\underset{\sim}{s}, \underset{\sim}{s}) \\ = {} & k \int \mathrm{d}\underset{\sim}{s}\ n((\underset{\sim}{x}' - h\underset{\sim}{s}, \underset{\sim}{s}) - \underset{\sim}{\mu}, p) \\ = {} & k\ n(\underset{\sim}{x}' - \underset{\sim}{\mu}_{\underset{\sim}{x}} - r_{\underset{\sim}{x}\underset{\sim}{x}}^{-1} r_{\underset{\sim}{x}\underset{\sim}{s}} \underset{\sim}{\mu}_{\underset{\sim}{s}}, r_{\underset{\sim}{x}\underset{\sim}{x}}) \end{aligned} \qquad (7)$$

where

$$q = h^2 p_{\underset{\sim}{x}\underset{\sim}{x}} - h(p_{\underset{\sim}{x}\underset{\sim}{s}} + p_{\underset{\sim}{s}\underset{\sim}{x}}) + p_{\underset{\sim}{s}\underset{\sim}{s}} \qquad (8)$$

and

$$r = p - p \begin{pmatrix} h^2 q^{-1} & -hq^{-1} \\ -hq^{-1} & q^{-1} \end{pmatrix} p \qquad (9)$$

and $p_{\underset{\sim}{x}\underset{\sim}{x}}$, $r_{\underset{\sim}{s}\underset{\sim}{x}}$ etc. denote the quadrants of *p* and *r*.

12. Equation (7) is obtained by a second (more delicate) exercise in completing the square, making use of the facts that $|p| = |q||r_{\underset{\sim}{x}\underset{\sim}{x}}|$ and $r_{\underset{\sim}{s}\underset{\sim}{s}} = r_{\underset{\sim}{s}\underset{\sim}{x}} r_{\underset{\sim}{x}\underset{\sim}{x}}^{-1} r_{\underset{\sim}{x}\underset{\sim}{s}}$. Further simplification may be possible.

Branching algorithm for recovering transition times

13. Now let us start with alpha values for the initial frame in the form (1): this is a single hypothesis accounting for the first observation. We may step through the data frame by frame, at each point looping over hypotheses for time *t*–1 and extending each hypothesis in each of the two possible ways to time *t*. The first extension continues the formant tracks and assumes that *t* is not a changepoint, so that the *k*, $\underset{\sim}{\mu}$ and *p* associated with the hypothesis are revised according to (3). The second extension continues the formant tracks in the same way, again applying (3), but then assumes that *t* is a changepoint and accordingly marginalises the distribution on the new target following (7). The resulting distribution on target alone is extended to a joint distribution with slope by bringing in the prior mean (presumably 0) and precision of slope vectors.

14. Thus we find that if the initial alphas take the form (1), then they take the same form at all subsequent times.

15. The likelihood of any assumed sequence of changepoints is the product of its prior probability with the sum over alpha values assigned by this algorithm, and this sum is simply the scale-factor *k*. The number of possible hypotheses doubles at each step, but various well-known thresholding techniques can be used to preserve the most probable hypotheses and discard the rest.

Phonetic modelling of formant targets

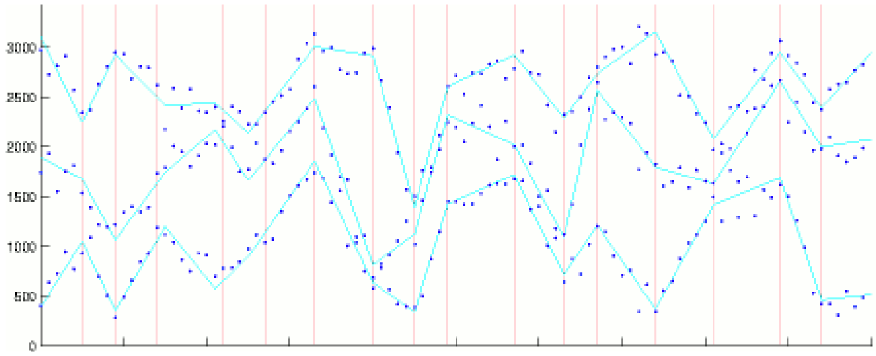
16. Now, instead of having formant targets chosen randomly from 3-dimensional space, let us assume that the phonetic repertory consists of a set of phonemes, and that the formant target for phoneme φ comes from a normal distribution whose mean and precision are $\underset{\sim}{\theta}_{\varphi}$ and a_{φ}. Then the same branching process may be pursued as before, but making the additional assumption of the identity of the phoneme at every postulated changepoint. We therefore multiply the expression (7) by the extra term

$$n(\underset{\sim}{x}' - \underset{\sim}{\theta}_{\varphi}, a_{\varphi}) \tag{10}$$

for each φ, requiring another square to be completed (very simple in this case); and at the same time we multiply the prior probability of the sequence of transition times by the prior probability of the assumed phoneme sequence (which may come from a language model of arbitrary complexity).

Illustration

17. For the following example, synthetic formant tracks were generated as shown by the pale blue lines; observations shown by the dark blue dots were made at each time with a standard error of 125Hz, and these were passed to a program which sought to recover the changepoints: the results are shown by the pink lines. There are a couple of errors, but human judgement unassisted by the pale blue lines would be unable to do as well.



18. The program which produced these results was less than 150 lines long.

Trajectories in the log domain

19. The model I have assumed so far is of trajectories along which frequencies are piecewise linear functions of time, and in which measurement errors and formant targets for phonemes are normally distributed in frequency space. An alternative would regard log frequencies as following linear trajectories.

20. Whether this would be a more accurate model of speech I am unqualified to say. As a model of measurement errors it would probably be slightly less accurate, but it would have the virtue of being restricted to positive frequencies.

21. In a logarithmic space the correction for vocal tract length is an additive term which is constant for a speaker. It could be an additional component of state. This is a tidier way of dealing with vocal tract normalisation than the method currently favoured.