Trajectory Analysis of Speech using Continuous-State Hidden Markov Models

Phil Weber, Steve M. Houghton, Colin J. Champion, Martin J. Russell and Peter Jančovič

School of EESE, University of Birmingham

Speech Recognition by Synthesis

Aim: develop a more faithful model of speech, useful for recognition, that uses a minimal set of parameters and accounts for the smooth variation found in real speech.

'Standard' discrete-state HMM GMM/DNN models

- assume generation of speech from discrete states,
- succeed due to model complexity and data availability,
- do not use/improve understanding of the speech signal,
- do not account for continuous, smooth nature of speech.

Our Method: Continuous-State Hidden Markov Model to recover the underlying sequence of phonemes from measurements of smoothly varying acoustic features, according to the 'HMS' model of speech.

The Holmes, Mattingley, Shearme (HMS) Model

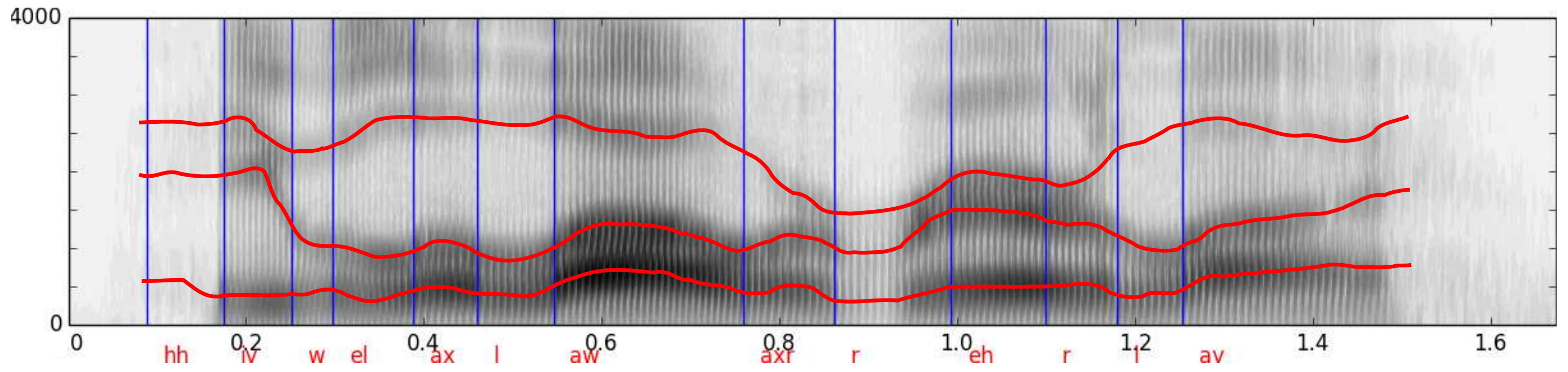
Developed for speech synthesis but also proposed for recognition. Speech is modelled by

- smooth trajectories (in a suitable space),
- piece-wise linear approximation, and
- alternating stationary periods connected by smooth transitions,

corresonding to slow, continuous movement of human articulators between target positions for the various speech sounds.

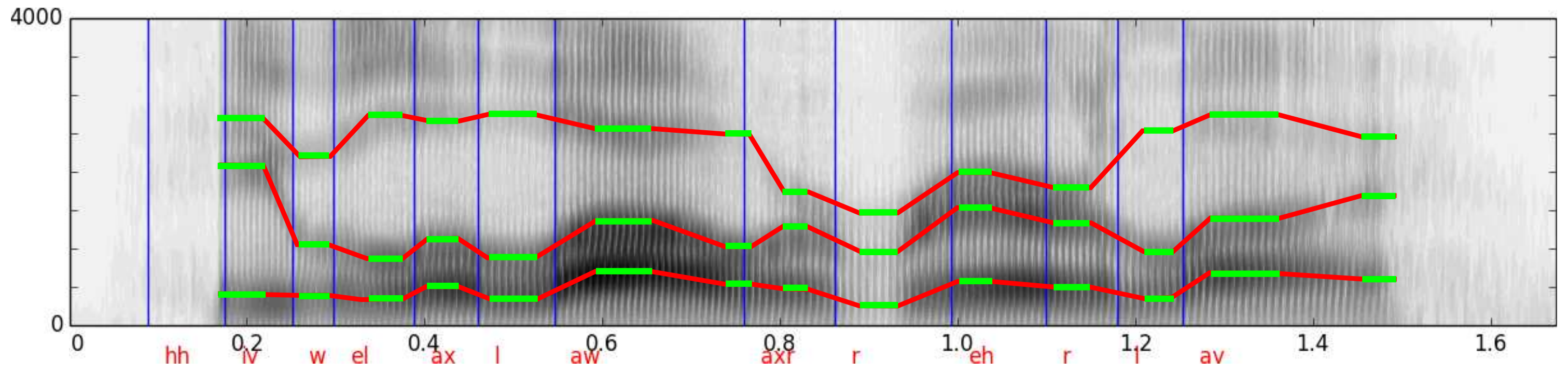
Research Outline

Given outputs generated according to the HMS Model,



such as smooth Vocal Tract Resonances, or formants,

fit a continuous sequence of trajectories,



and recover the sequence of phonemes.

We use a continuous state (CS-HMM) algorithm.

CS-HMM Model

Assume an HMS model of speech:

- canonical frequency targets, $\boldsymbol{f}_\phi$, noisy realisations, $\boldsymbol{f}_t \sim \mathcal{N}_d(\boldsymbol{f}_\phi, \mathrm{A})$,
- noisy observations during dwells, $\boldsymbol{y}_t \sim \mathcal{N}_d(\boldsymbol{f}_t, \mathrm{E})$,
- noisy observations around linear transitions,
- a dwell/transition timing model.

A State contains continuous and discrete components

- $\boldsymbol{x}$ (realised target frequencies – dwell/transition),
- $\boldsymbol{s}$ (slopes – transitions), and
- identifies current phase (dwell/transition), phoneme identity, 'ticks' in phase, and phonetic history.

A hypothesis contains

- probability information about an infinite set of states,
- in parametric form (scaled Gaussian):

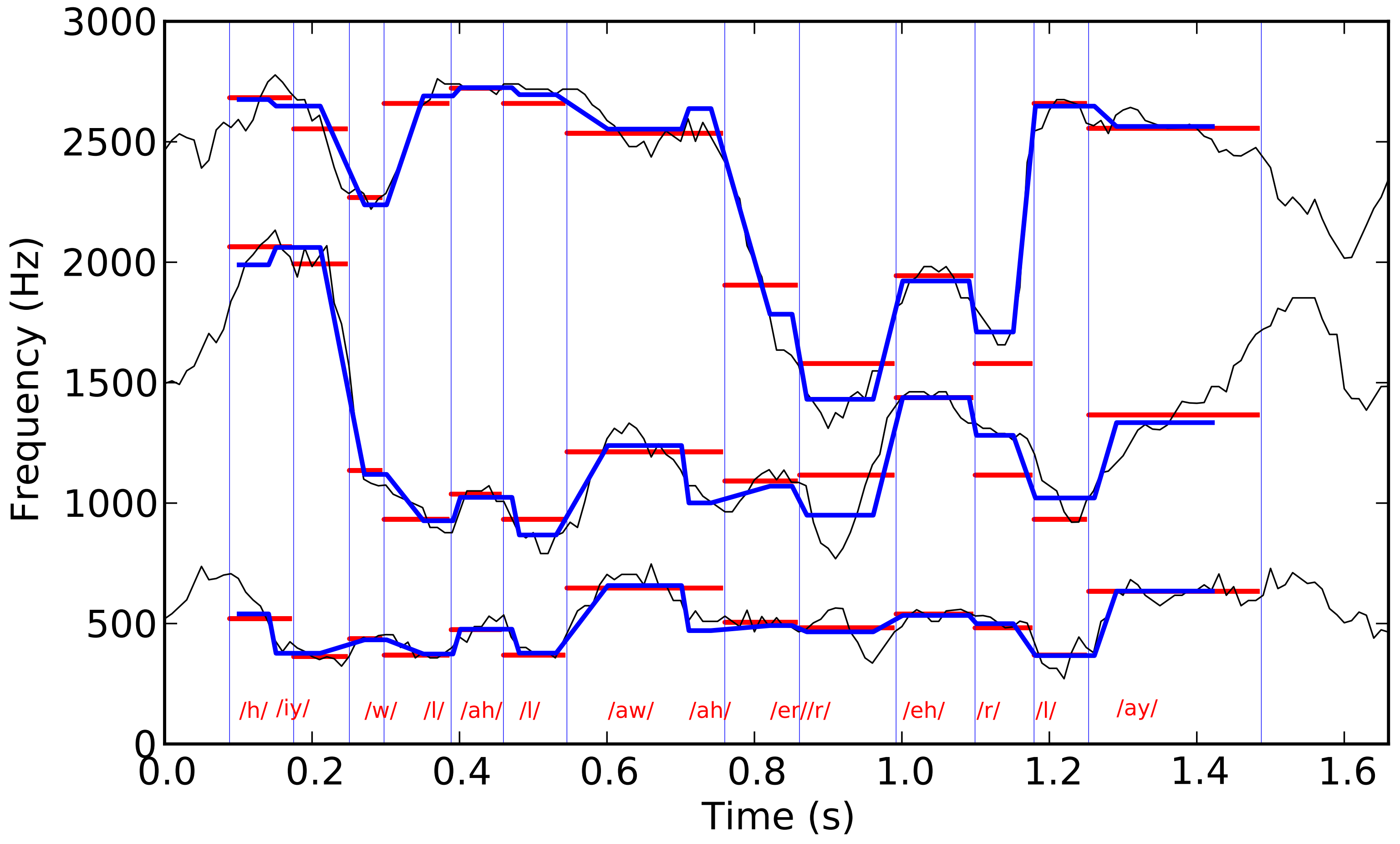
$$\alpha_{t-1}(\boldsymbol{x}) = K_{t-1}\mathcal{N}_d(\boldsymbol{x} - \boldsymbol{\mu}, \mathrm{P}),$$

μ and P: mean and precision of distribution over state.
K_t: sum of probabilities of paths consistent with the hypothesis.

A Minimal Number of Parameters to Train

$\approx 40 \times 3$ phoneme target frequencies, target frequency and observation covariance matrix(es), and 'any' timing model.

Example: Recovery of TIMIT Utterance



Transcription:	/hh iy w l ah l aw er r eh r l ay/
Recovery:	/hh iy w l ah l aw **ah** er r eh r l ay/

Initial (controlled) experimentation to prove the algorithms.
Changepoints have been found by the algorithm.
Errors in recovery map to underlying phenomena.

CS-HMM: Recovery

① Assume dwell start: Initialise one hypothesis per phoneme

$$\alpha_0(\boldsymbol{x}) = \mathcal{N}_d(\boldsymbol{x} - \boldsymbol{f}_\phi, \mathrm{A}).$$

② Step through dwell. Observe $\boldsymbol{y}_t$, assumed drawn from $\mathcal{N}_d(\boldsymbol{x}, \mathrm{E})$.

③ Update hypothesis to take account of observation

$$\begin{aligned}\alpha_t(\boldsymbol{x}) &= K_{t-1}\mathcal{N}_d(\boldsymbol{x} - \boldsymbol{\mu}_{t-1}, \mathrm{P}_{t-1})\mathcal{N}_d(\boldsymbol{y}_t - \boldsymbol{x}, \mathrm{E}) \\ &= K_t\mathcal{N}_d(\boldsymbol{x} - \boldsymbol{\mu}_t, \mathrm{P}_t).\end{aligned}$$

where

$$\mathrm{P}_t = \mathrm{P}_{t-1} + \mathrm{E}, \qquad \boldsymbol{\mu}_t = \mathrm{P}_t^{-1}(\mathrm{P}_{t-1}\boldsymbol{\mu}_{t-1} + \mathrm{E}\boldsymbol{y}_t),$$

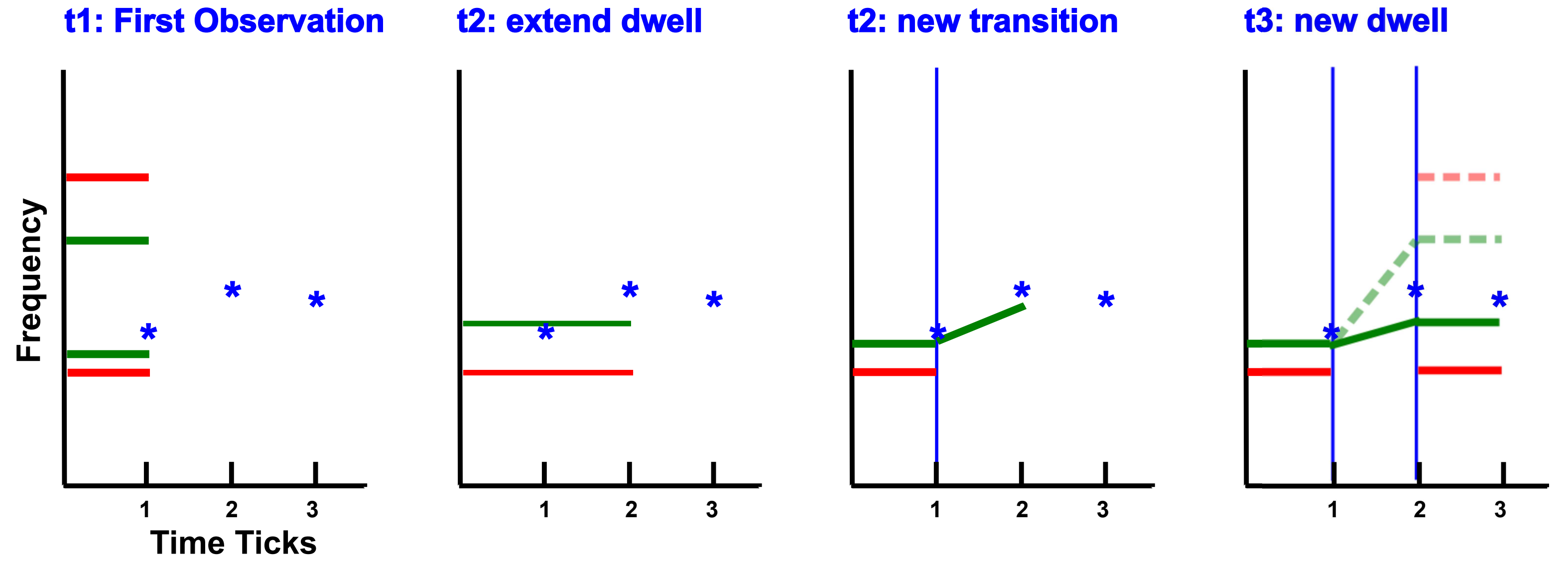
$$\mathrm{K}_t = \mathrm{K}_{t-1}\mathcal{N}_d\Big(\boldsymbol{y}_t - \boldsymbol{\mu}_{t-1}, (\mathrm{P}_{t-1}^{-1} + \mathrm{E}^{-1})^{-1}\Big).$$

④ Realised frequency is a compromise between canonical and observed frequencies.

⑤ Next step: choice to extend the dwell, or enter transition, so we split the hypotheses.

⑥ Hypotheses in transition: choice to continue, or split off hypothesis for each phoneme in the inventory.

⑦ Threshold on K_t to manage the hypothesis list.



- Showing just one trajectory.
- Only two phonemes in inventory.

- Concentrating on one hypothesis.
- Realised target adjusts.

- Alternative hypothesis: transition.

- New hypotheses for each phoneme target.

References

P. Weber, S. M. Houghton, C. J. Champion, M. J. Russell and P. Jančovič, *"Trajectory Analysis of Speech using Continuous State Hidden Markov Models"*, In Proc. *ICASSP*, 2014.

C. J. Champion and S. M. Houghton, *"Application of Continuous State Hidden Markov Models to a Classical Problem in Speech Recognition"*, Computer Speech and Language, *(submitted)*, 2014.

J. N. Holmes, I. G. Mattingly and J. N. Shearme, *"Speech Synthesis by Rule"*, Language and speech, 7, 127-143, 1964.

L. Deng *et al.*, *"A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing Acoustics"*, In Proc. *ICASSP*, 2006.