Trajectory Analysis of Speech using Continuous-State Hidden Markov Models

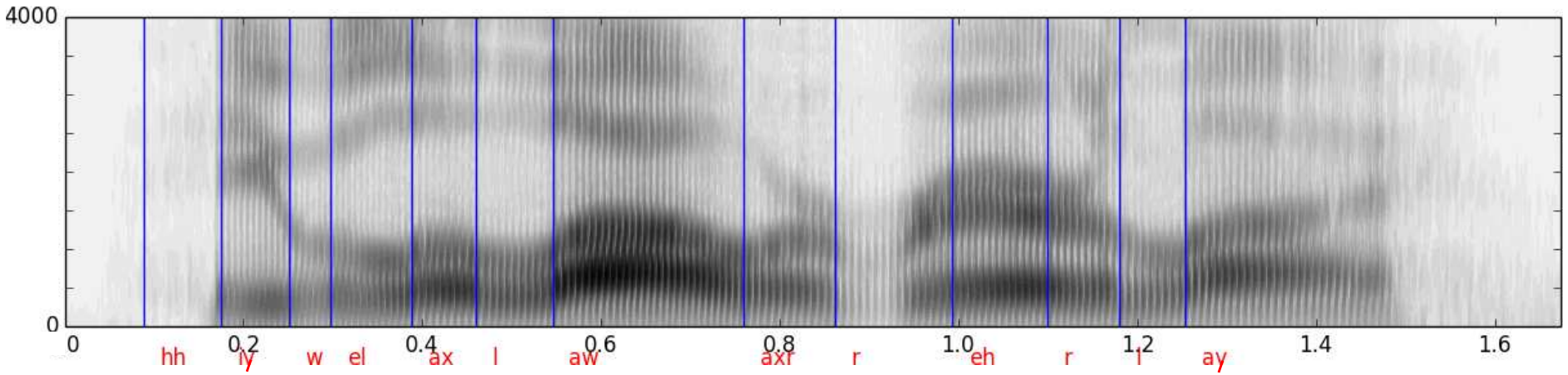
P. Weber, S. M. Houghton, C. J. Champion,
M. J. Russell and P. Jančovič

University of Birmingham, UK

ICASSP, 7 May 2014

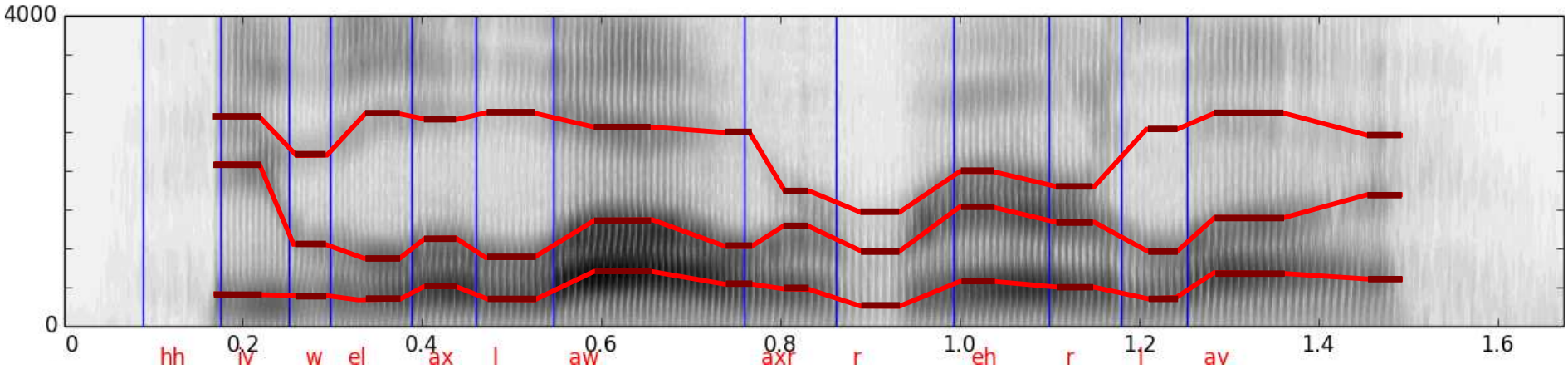
Continuously Changing Smooth Trajectories

Speech is not a set of discrete states
– rather a series of smooth, continuous transitions.



"he will allow a rare lie"

HMS Linear Dwell-Transition Model

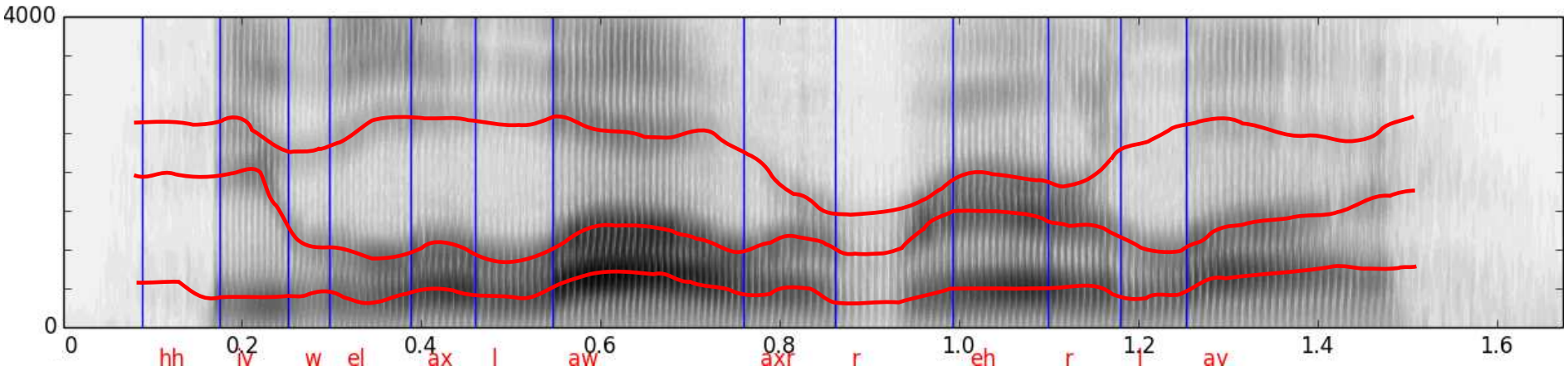


Holmes, Mattingley, Shearme, 'Speech Synthesis by Rule', (1964) (HMS).

- A sequence of stationary periods linked by smooth transitions.
- Piece-wise linear approximation.
- Dwell (articulator) target frequencies, transitions.

CS-HMM: Recovery of Phoneme Sequences

Given outputs generated according to the HMS Model, fit a continuous sequence of trajectories to them ...



- ... and thus recover the sequence of phonemes.
- Continuous-State HMM algorithm.

CS-HMM: Assumptions

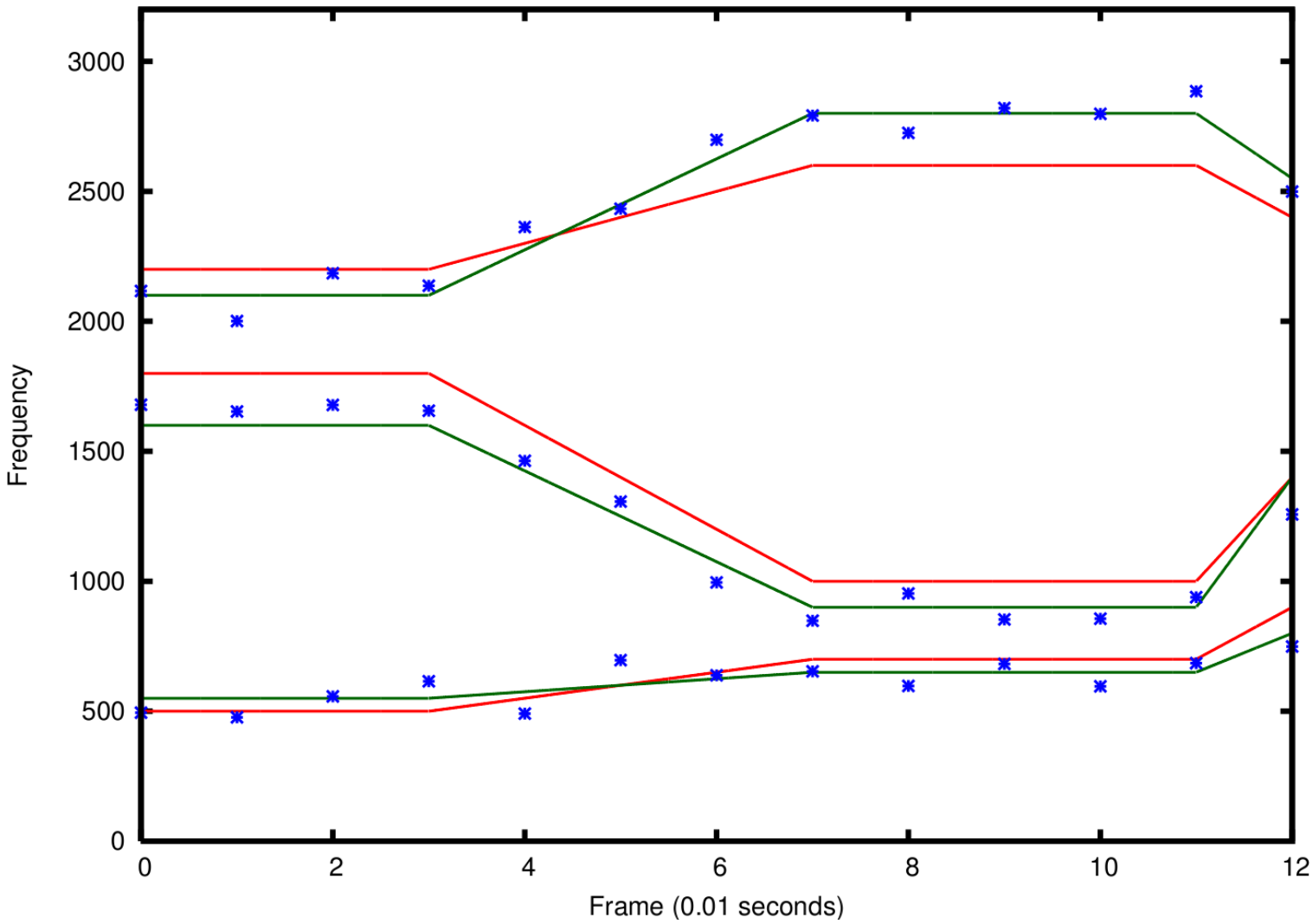
Speech is modelled as a sequence of constant dwells and linear transitions.
Dwell represents target sound (e.g. phoneme):

- Persists for $t_\phi \sim D_\phi$ time steps.
- Canonical phoneme frequency targets, $\mathbf{f}_\phi$.
- Actual frequency targets: noisy realisations, $\mathbf{f}_t \sim \mathcal{N}_d(\mathbf{f}_\phi, \mathrm{A})$.
- Realisation variance allows for systematic departures from the mean *(e.g. speaker dependence if that is not modelled separately)*.
- Noisy observations, $\mathbf{y}_t \sim \mathcal{N}_d(\mathbf{f}_t, \mathrm{E})$.

Transition represents smooth movement of articulators between dwells:

- Persists for $t_T \sim D_T$ time steps.
- Frequencies transition linearly between realised targets.
- Noisy observations about the linear transition.

CS-HMM: Assumptions



Red: canonical targets, **Green: realised targets**, **Blue: observations**.

CS-HMM: State

Continuous component

- $\mathbf{x}$: realised target frequencies at time t (dwells and transitions),
- $\mathbf{s}$: slopes at time t (transitions),

Discrete component

- identify whether in dwell or transition,
- h: time steps in current dwell/transition,
- identity of current phoneme,
- Phonetic history.

CS-HMM: Hypothesis

Probability information about an infinite set of states,

- Alpha value $\alpha_{t-1}(\mathbf{x})$:
  - sum of probabilities over all paths leading to state at $t-1\ldots$
  - $\ldots$ consistent with discrete history$\ldots$
  - $\ldots$ given path transitions and observations.
- Parametric form: scaled Gaussian,

$$\alpha_{t-1}(\mathbf{x}) = K_{t-1}\mathcal{N}_d(\mathbf{x}-\mu, \mathrm{P}),$$
$$\text{where} \quad \mathcal{N}_d(\mathbf{x}, \mathrm{P}) = (2\pi)^{-d/2}|\mathrm{P}|^{1/2}\exp\{-\tfrac{1}{2}\mathbf{x}^T\mathrm{P}\mathbf{x}\}.$$

μ and P are the mean and precision (1/variance) of the distribution over the current continuous state.
K_t is the sum of probabilities of paths consistent with the hypothesis.
d is the dimension of the space.

CS-HMM: 'Training'

Learn the parameters in the system, e.g.

1. Per-phoneme canonical target frequencies $\mathbf{f}_\phi$ ($\approx 40 \times 3$).
2. Target frequency co-variance matrix A (6).
3. Observation co-variance matrix E (6).
4. Timing model (can be anything).

CS-HMM: Recovery

1. Assume dwell start: Initialise one hypothesis per phoneme

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

2. Step through dwell. Observe $\mathbf{y}_t$, assumed drawn from $\mathcal{N}_d(\mathbf{x}, \mathrm{E})$.
3. Update hypothesis to take account of observation

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

where

$$\begin{aligned}\mathrm{P}_t &= \mathrm{P}_{t-1} + \mathrm{E}, \\ \mu_t &= \mathrm{P}_t^{-1}(\mathrm{P}_{t-1}\mu_{t-1} + \mathrm{E}\mathbf{y}_t), \\ \mathrm{K}_t &= \mathrm{K}_{t-1}\,\mathcal{N}_d\left(\mathbf{y}_t - \mu_{t-1}, (\mathrm{P}_{t-1}^{-1} + \mathrm{E}^{-1})^{-1}\right).\end{aligned}$$

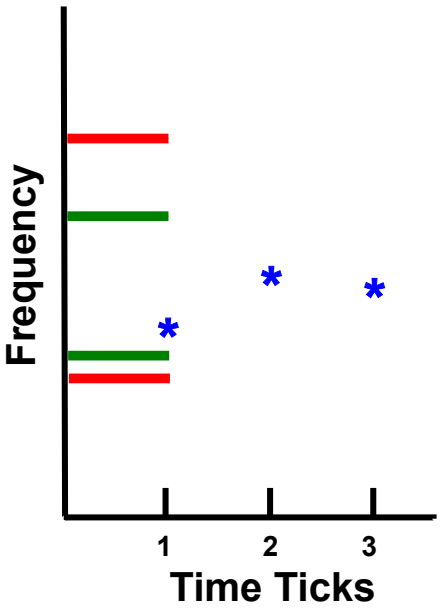
CS-HMM Recovery

Similar (but more complicated) formulae allow us to step through a transition and to move between dwells and transitions.

- At each tick during a dwell we have a choice between continuing the dwell and entering a transition.
- At each tick during a transition we have a choice between continuing the transition and entering a dwell for each phoneme in the inventory.
- Hence we branch on hypotheses at each step.
- And to keep their number within bounds, we threshold on K_t.
- For a full development of the mathematics, see the papers on our website (`www.birmingham.ac.uk/srbs`).

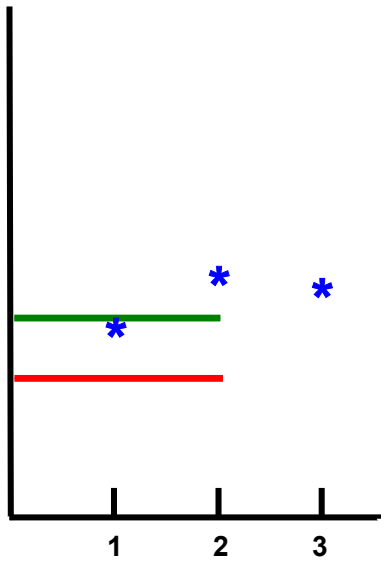
CS-HMM Process Illustration

t1: First Observation



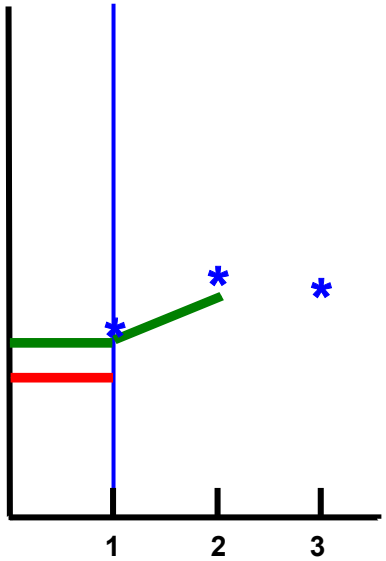
- Showing just one trajectory.
- Only two phonemes in inventory.

t2: extend dwell



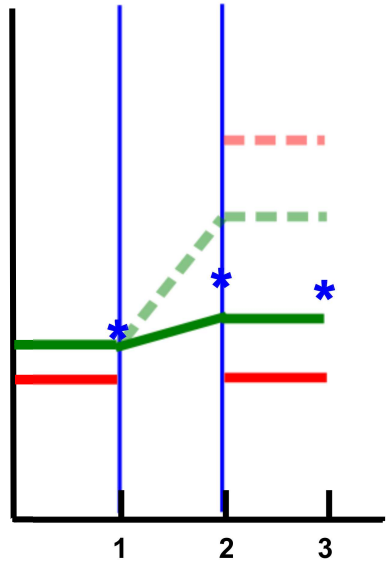
- Concentrating on one hypothesis.
- Realised target adjusts.

t2: new transition



- Alternative hypothesis: transition.

t3: new dwell



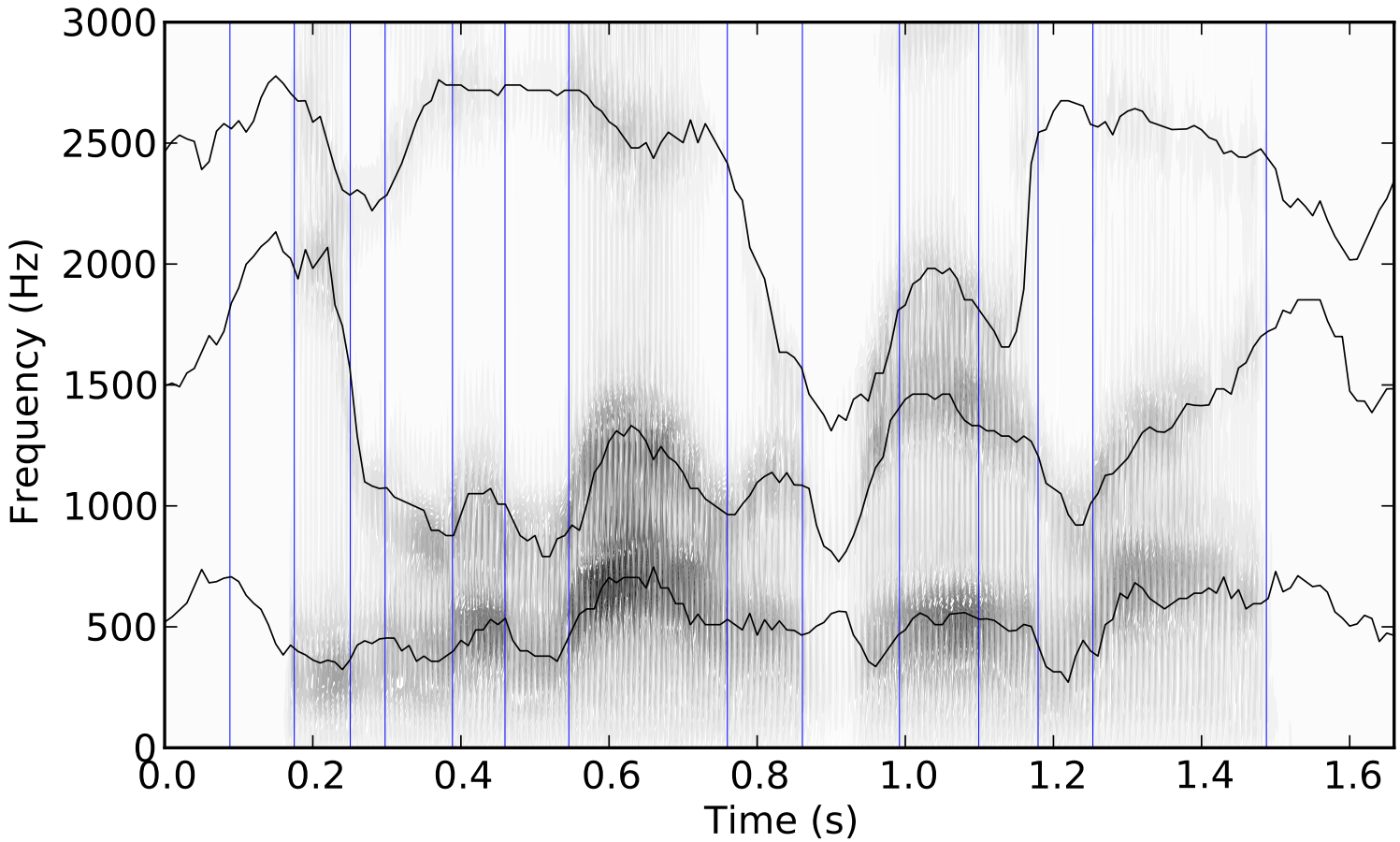
- New hypotheses for each phoneme target.

Experimentation

Limited experimentation to illustrate the technique.

- Recover TIMIT phoneme sequences.
- Vocal Tract Resonances [Deng *et al.*, 2006].
- Learn and recognise single utterances.

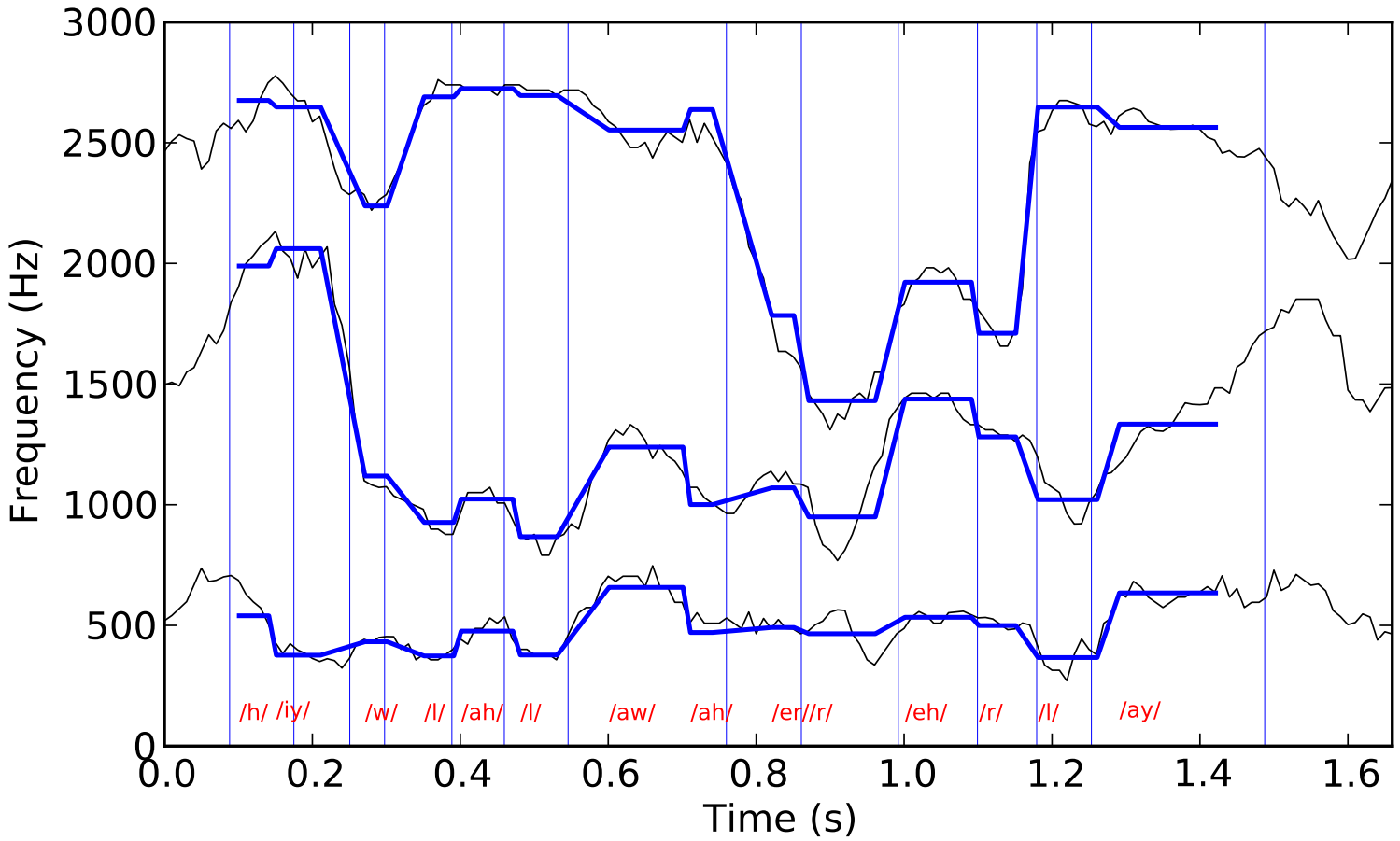
Recovery of Phoneme Sequences from TIMIT



Utterance test/dr2/mwew0/sx11: 'he will allow a rare lie'.

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

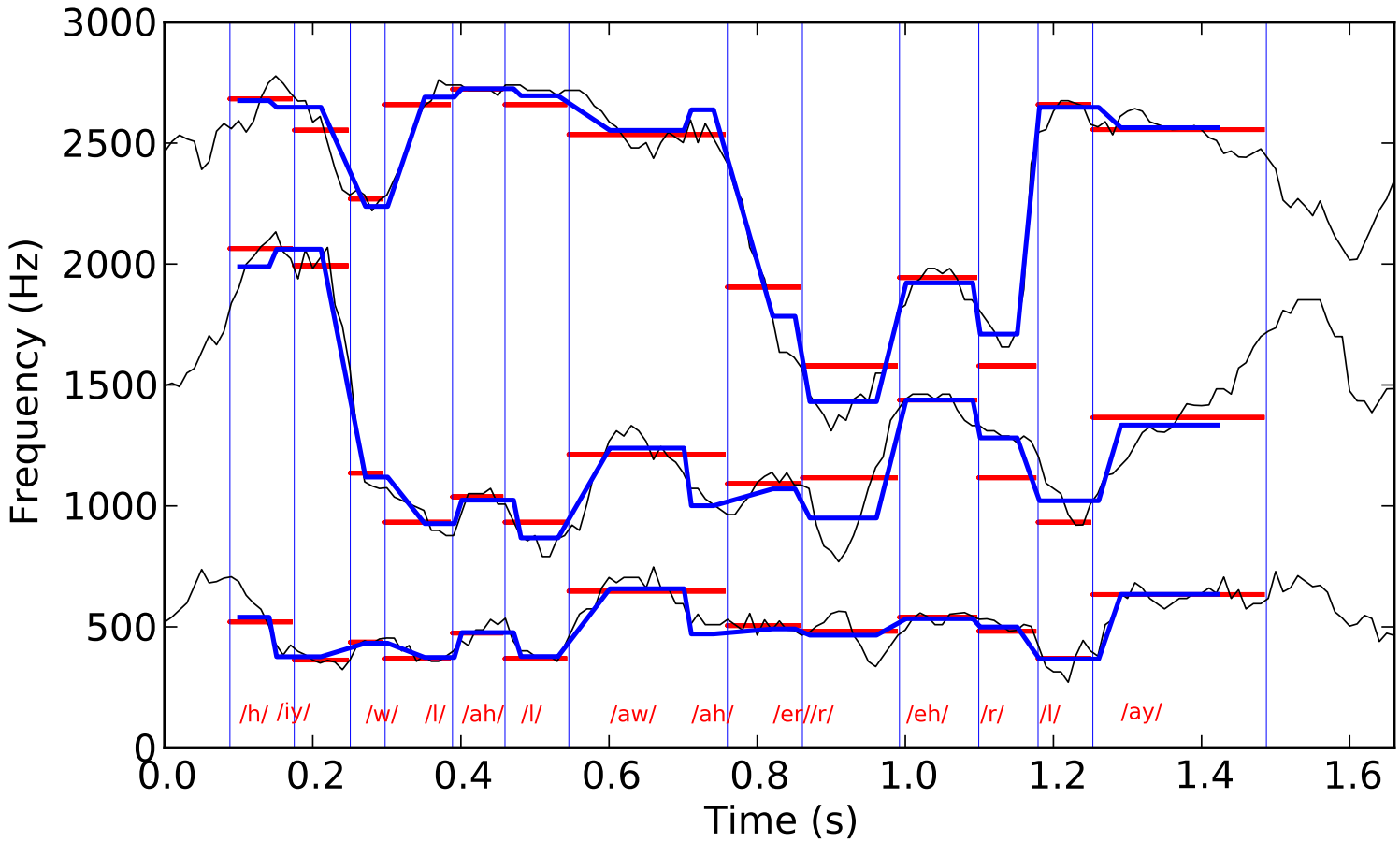
Recovery of Phoneme Sequences from TIMIT



Utterance test/dr2/mwew0/sx11: 'he will allow a rare lie'.

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/

Recovery of Phoneme Sequences from TIMIT



Utterance test/dr2/mwew0/sx11: 'he will allow a rare lie'.

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/

To Conclude

The CS-HMM algorithm is able to successfully recover phoneme sequences in a controlled environment

- using a more faithful model of speech,
- and a very limited parameter set.

Some developments are clear:

- accurate training of the phoneme inventory,
- automated parameter learning,
- handle systematic variation in speech,
- how does the model extend to non-sonorant speech?

Thank you.

Questions?

http://www.birmingham.ac.uk/SRbS/