# Speech recognition

## Novel approaches

School of informatics

School of Philosophy, Psychology and Language Sciences

contact: Simon.King@ed.ac.uk

*Centre for Speech Technology Research* *a joint research centre of the College of Science & Engineering and the College of Humanities and Social Science*

## Theory

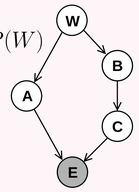
## Speech recognition

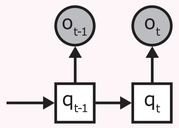
## Other tasks

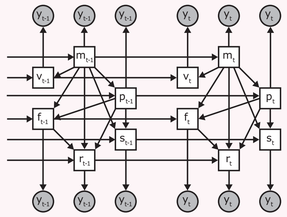
### Models

#### Dynamic Bayesian Networks

$$P(A,B,C,E,W) = P(E|A,C)P(C|B)P(B|W)P(A|W)P(W)$$



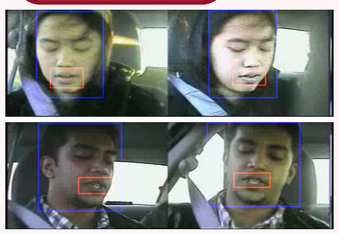



DBNs allow us to model underlying structure in speech. Linguistic knowledge guides our choice of network structures: the variables in the network have specific linguistic meanings.

#### Structure learning

Too many network structures to try! As well as using linguistic intuition, we have methods for *learning* optimal structure.

#### Audiovisual

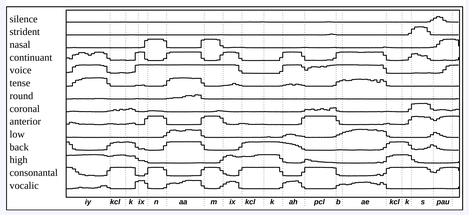


#### EMA

Electromagnetic articulograph records position of tongue, lips and velum during speech. Insights into speech production. Informs models of speech.

#### Phonetic features

Phonetic features are a **factored** representation of speech: they have multiple streams such as the manner of articulation (vowel, fricative, nasal,...) or the place in the mouth where the sound is articulated.




#### Multilingual

#### PFs

More universal than phonemes.

#### Graphemes

Avoid writing lexicon for each new language.

#### Tandem

Train neural net on some other language(s) that have more data.

#### Audio search

Speech recognisers have fixed vocabularies – how can we search for words not known to the system? *Cannot pre-recognise the audio into words.* Instead, label audio with lattices of sub-word units.

If we use phonemes, then have to guess pronunciation of search term: this is error prone.

**Why not use graphemes instead?**

### Sub-word units

#### Learnt units

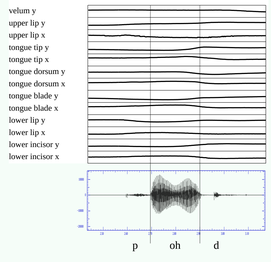
Lots of evidence to suggest that phonemes are not the ideal unit for speech recognition. As well as using other units (see right), we are attempting to use machine learning to discover an optimal unit inventory.

#### Graphemes

Avoid the problems of creating a lexicon. Poor linguistic motivation, yet they seem to work! Why are phonemes no better than graphemes?

### Features

#### Articulatory measurements




#### Posterior-based features

Harness power of neural networks (large input context windows, discriminative training,...), and conventional Hidden Markov Models (statistical modelling, adaptation, parameter sharing,...): use the net to perform a classification task – from speech into phoneme posterior probabilities – then derive features from that distribution and feed to an HMM. **But is phoneme classification the best task?**