

# **Overview of Spoken Language Systems**



**Dr. Roberto Togneri**  
**CIIPS (SLSR Group)**  
**E&E Eng, UWA**

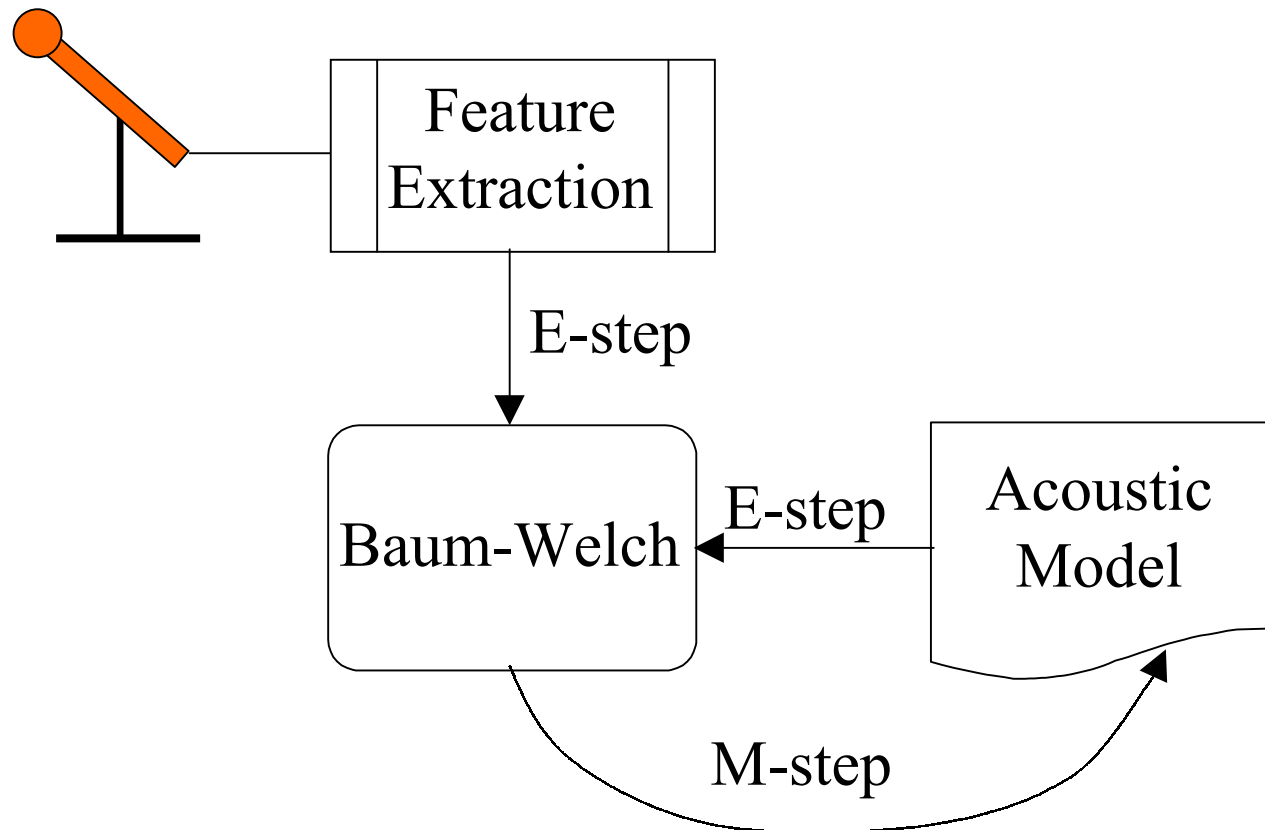
# Contents



- ASR/SLS Schematic Overview
- Feature Extraction
- Acoustic Modelling
- Language Modelling and Understanding
- A New Approach to Acoustic Modelling
- New Approach: Formulation
- New Approach: Evaluation Results
- Conclusions and Future Work

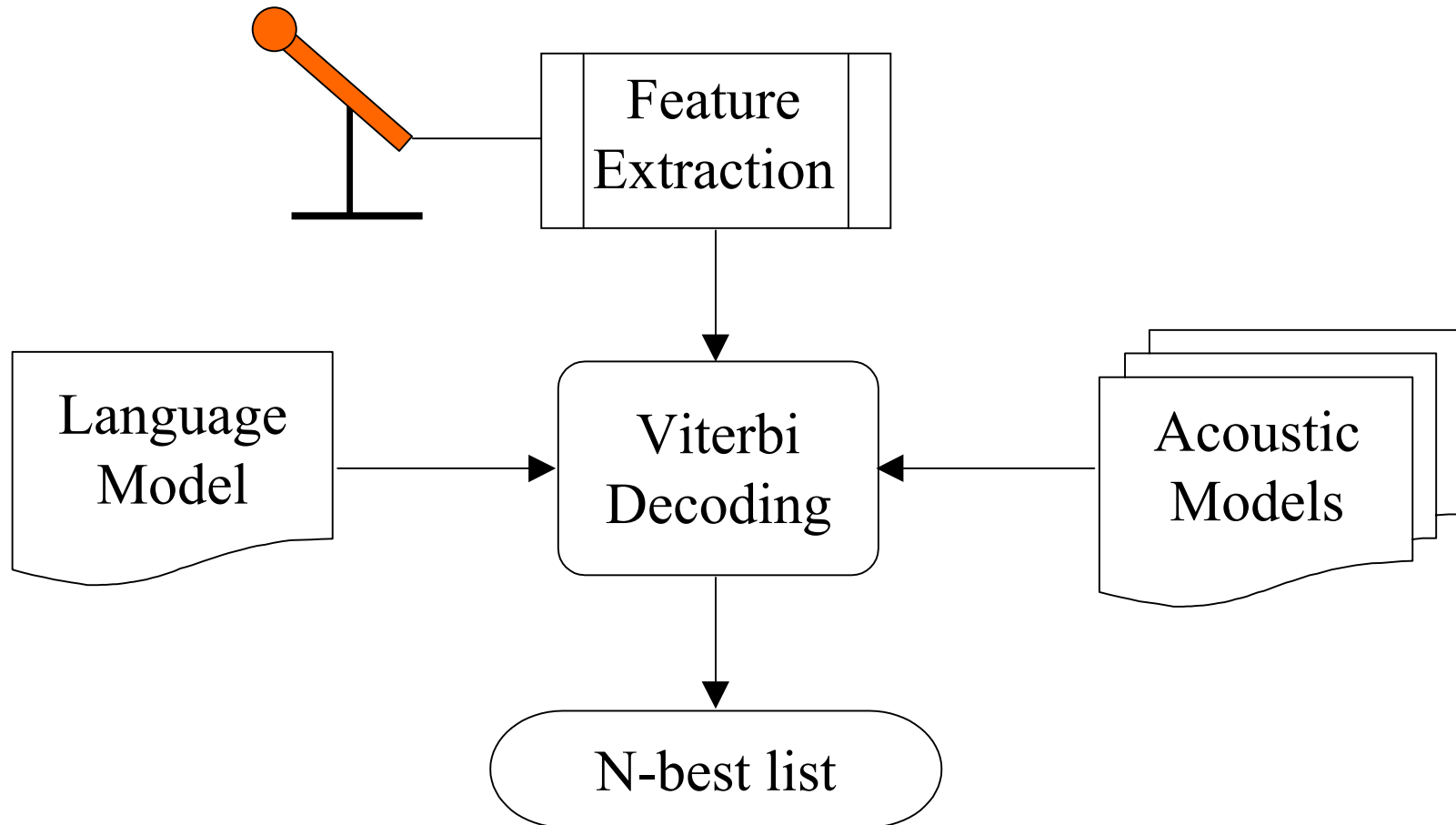
# ASR Schematic Overview

## ■ Training/Estimation (EM Algorithm)

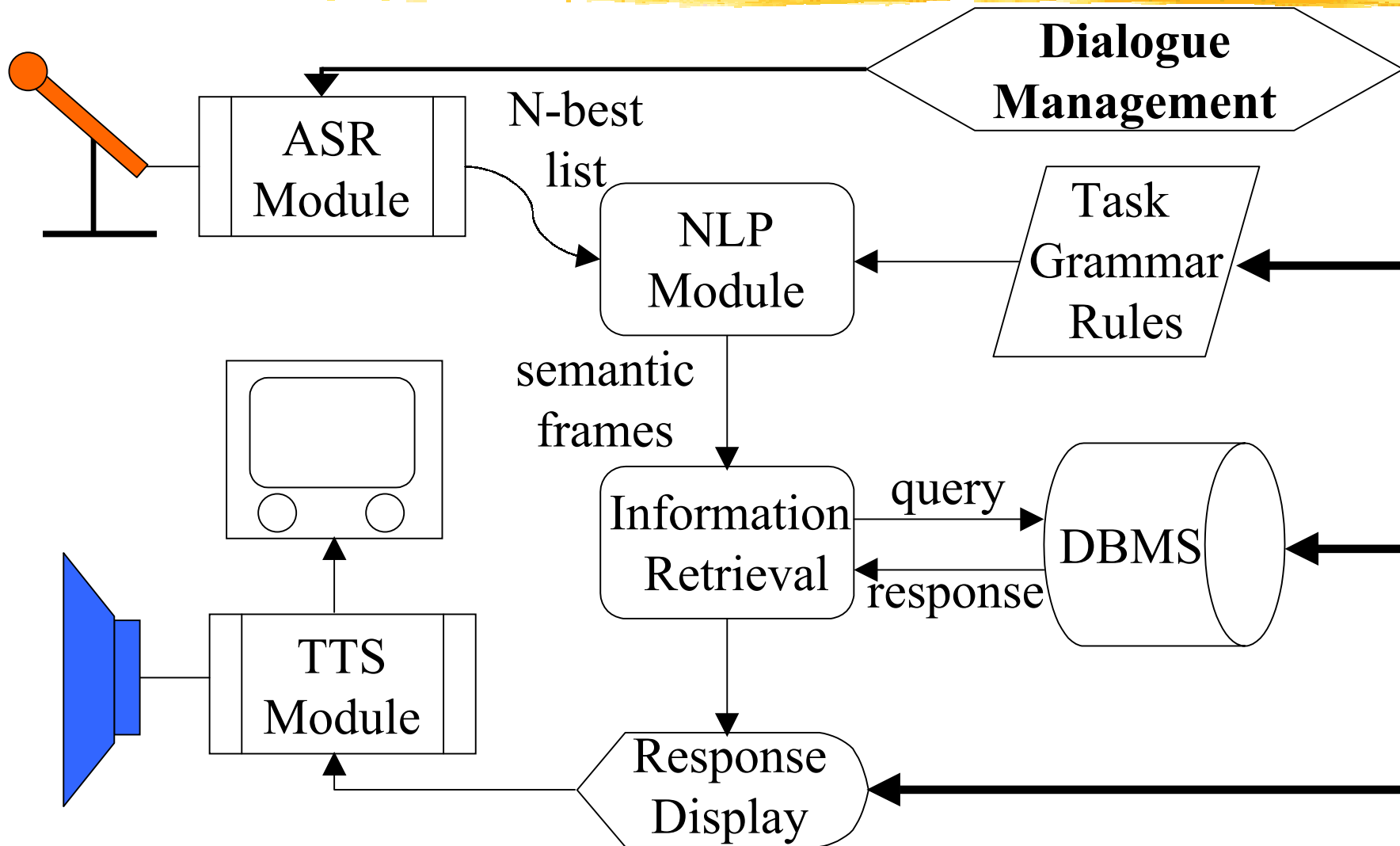


# ASR Schematic Overview

## ■ Testing/Decoding (Viterbi)



# SLS Schematic Overview



# Feature Extraction



- Linear Predictive Coding (LPC)
  - models speech as an AR process
  - all-pole model of vocal tract response
- FFT Filter Bank (FBANK)
  - 256-point FFT magnitude “binned” into a smaller number of intervals ( $\sim 12$ )
  - mel-Spaced FBANK by selecting intervals according to human pitch perception
    - smaller intervals at lower frequencies to emulate enhanced human perception of lower frequencies

# Feature Extraction

- Mel Frequency Cepstral Co-efficients (MFCC)
  - $cc = \text{FFT}^{-1}(\log|\text{FFT}(s)|)$ 
    - $\text{FFT}(s) \rightarrow$  Mel-spaced FFT filter bank
  - 0<sup>th</sup> cc represents the energy or gain
  - low-order cc represents vocal tract response
  - high-order cc represents pitch response
    - remove unwanted pitch response by “liftering”
  - Use DCT for  $\text{FFT}^{-1}$ 
    - DCT basis functions  $\approx$  KLT basis functions
    - $\therefore$  cc's are highly decorrelated  $\rightarrow$  use diagonal **C**

# Robust Feature Extraction

- MFCC with Cepstral Mean Subtraction
  - convolutional channel distortion/noise is additive in the cepstral domain
  - assume channel distortion is time-invariant
  - subtract running time-average from cc
    - removes speech spectral mean
    - removes constant channel distortion
  - very simple and effective
    - removes time-invariant convolutional noise
  - idea of modulation spectrum → Rasta-PLP

# Robust Feature Extraction



## ■ Spectral Subtraction

- estimate noise spectrum statistics in silence
- remove estimate of noise from noisy speech
- not very effective
  - noise statistics change with time
  - speech is degraded
- BUT estimation of acoustic model parameter assumes “noisy” data anyway
  - except that it is only really good for zero-mean WGN

# HMM Acoustic Modelling

## ■ HMMs for Speech

- characteristics of speech change with time, but there are regions of identical characterisation
  - finite-state model (MM)
  - distribution of acoustic observations per state (HMM)
  - e.g. "cat" starts in state /k/, then /a/ and ends in /t/
- topology
  - left-to-right with self-loop and skip states
  - self-loop for slower rate of speaking
  - skip state for faster rate of speaking

# HMM Acoustic Modelling



## ■ HMM Models

- one model for each word in the vocabulary
  - | unwieldy with large vocabulary systems
  - | models intra-word context
- one model for each phone
  - | ~ 40 phonemes for English
  - | context-independency (CI) of model limits the applicability to co-articulated speech
- tri-phone model with left and right contexts
  - | context-dependent (CD) modelling
  - | up to  $\sim 40^3$  possible models

# HMM Parameter Estimation

- ML estimation by EM (Baum-Welch)
  - Maximise  $L(\text{Observation Sequence}|\text{Model})$ 
    - | requires consideration of all possible paths
  - E-step
    - | calculate the forward and backward probabilities
    - | derive expectations of state sojourn and transitions
  - M-step
    - | re-estimate state transition probabilities directly from expectation of state transitions
    - | re-estimate state distribution sample mean and covariances using sojourn expectation of observations in state

# HMM Model Decoding



## ■ Viterbi Algorithm

- Score a given unknown observation sequence against all models
  - | derive likelihood of best state sequence
- Record of best path up to time  $t$  for each state
  - | unwieldy with large vocabulary systems
- Fast decoding for LVCSR
  - | prune paths which fall below a set threshold
  - | use a language model to weight score against unlikely words

# Language Modelling

- Use to reduce Viterbi search space
- Types
  - Task Grammar
    - | finite-state automata or network (FSA/FSN)
    - | defines which words will follow the current word
    - | may not be well-defined for most tasks
  - Bi-gram Models
    - | statistical analysis of word-pairs
    - | weight probability from current to next word
  - Others
    - | tri-gram models, N-grams, class-based N-grams

# HMM Model Adaptation

- The “mismatch” problem
  - Training and testing in different environments
    - “insufficient” training data to cover all cases
      - different speakers
      - different background noise characteristics
      - different microphone
- Bayesian Adaptation (MAP estimation)
  - prior: trained HMM
  - data: observations from test environment
  - re-estimated HMM is adapted to environment

# N-Best List Generation



- What is an N-best list?
  - Viterbi decoding keeps track of the N best paths and returns N possible word-level transcriptions (N-best list)
  - More computationally and memory demanding
- Why an N-best list?
  - Only useful with language understanding
    - N-best list provides more information for analysis
    - NLP semantic analysis can identify the correct transcription which may not be the 1-best

# Language Understanding

## ■ Parsing and Semantic Frames

- parse into words and relational phrases
- create a semantic frame

- “What are the flights from Sydney to Auckland on Thursday morning”

```
<departing-city = Sydney>,  
<destination-city = Auckland>,  
<command-type = flight schedule>
```

## ■ Form Database Query

- translate semantic frame to SQL query

## ■ Generate A Response

## ■ Dialogue Management

# R&D Activity



- The hotter, the more active
  - Feature Extraction for plain ASR (cool)
  - Feature Extraction and Acoustic Modelling for Robustness and Adaptation (hot)
  - Acoustic Modelling (warm but being re-ignited)
  - Language Modelling and Fast Decoding (warming up)
  - Language Understanding and Dialogue Management Systems (incandescent)
  - Applications of SLS Technology (thermonuclear)

# Why a New Approach?



- Problem with HMM-based LVCSR Systems
  - blind data-driven approach based on statistical “ignorance” modelling
  - gigantic numbers of parameters trained with huge amounts of data
    - 3 million parameters trained with 3 Gb of data
  - speech recognition is becoming less and less constrained in both task grammar and environment
    - even more parameters with even more data!

# Hidden Dynamic Models



## ■ HDM Fundamental Idea

- use of internal, hidden speech dynamics more closely based on speech data generation
  - hidden dynamic can represent articulatory or vocal tract resonance (VTR) dynamics
- context-dependence inherent in the model structure
  - continuity condition of hidden dynamic
- compact parameter set
  - less parameters, less training data and improved generalisation

# Model Framework

- Non-linear switched target-directed hidden dynamic state-space model

$$Z_{k+1} = \Phi_j Z_k + (I - \Phi_j) T_j + w_k$$

$$O_k = h_j(Z_k) + v_k$$

- | Z: VTR dynamic
- O: observations (MFCC)
- $T_j$ : target for phone j
- $\Phi_j$ : time-constant for phone j
- $h_j(Z)$ : non-linear dynamic to observation mapping
- $w_k$ : dynamic process "noise"
- $v_k$ : observation noise

# Model Framework

- T describes VTR of phone
  - | formants in the case of voiced sounds
- $\Phi$  describes “speaking style” of phone
  - | target “under-shoot” in conversational speech
- Z is continuous from phone  $j$  to  $j+1$ 
  - | inherent context-dependence modelling
- $h(Z)$  represents a mapping from the VTR dynamic to the observations
  - | 3-dim VTR mapped to 12-dim MFCC
  - | use MLP or RBF non-linear mapping function
    - W: ANN weights

# Model Parameter Estimation

- ML via EM (Deng & Ma, 1999)
  - E-step: use EKF to estimate  $Z$  given  $(O, \Theta)$ 
    - $\Theta$  is the parameter set  $\{\Phi, T, W\}$  to be estimated
    - E-step computes  $E\{\log L(Z, O | \Theta) | O, \underline{\Theta}\}$
  - M-step: maximise joint log likelihood
    - choose  $\Theta$  to maximise  $E\{\log L(Z, O | \Theta) | O, \underline{\Theta}\}$ 
      - $\underline{\Theta}$  : next estimate
    - yields non-linear equations in  $\{\Phi, T\}$ 
      - Newton-Rhapson method, generalised form of EM, etc.
    - $W$  estimated by MLP back-prop
      - $\underline{Z}$ : input ;  $O$ : desired output

# Model Parameter Estimation

- RLS via EKF (Togneri & Deng, 2000)
  - use EKF for joint state  $\{Z\}$  and parameter  $\{\Phi, T, W\}$  estimation
    - parameters estimated with each iteration
  - joint state and parameter estimation via EKF is not new
    - control literature: (Ljung, 1979)
    - RNN training via EKF: (Feldkamp & Puskorius, 1998)
  - successfully used on synthetic data
    - for estimating  $\{\Phi, T\}$ , but requires additional constraints for unique estimation of  $\{\Phi, T, W\}$

# Model Evaluation



- N-best rescoring paradigm
  - CD tri-phone HMM is used to provide N-best list and phone time-alignments (via Viterbi)
  - model likelihood for each transcription is calculated and the best sentence hypothesis is chosen
  - obviates the need to calculate transcription based on fast decoding and language models

# Model Evaluation Results

- WS'97 SWB Corpus (Deng & Ma, 1999)
  - Training Data:
    - | 30 minutes from single male
  - Test Data:
    - | 1241 utterances
    - | 23 male speakers (excluding training speaker)
  - 5-best + reference
    - | conventional HMM: 44% WER
    - | HDM (HMM segmts): 32% WER
    - | HDM (hand segmts): 23% WER

# On-going Work



## ■ Estimation

- Currently phone boundary segments are obtained from an “off-line” HMM
  - investigate optimal segmentation strategies
- $h(Z)$  non-linear mapping is not very satisfactory
  - one  $h(Z)$  per phone or one global  $h(Z)$ ?
  - piecewise-linear? linear? non-linear?
- more effective constraints on parameters
  - avoid non-unique solutions
- EM algorithm .vs. EKF algorithm

# On-going Work



- Decoding
  - fast decoding using optimal segmentation
  - integration with language model
- Other
  - speaker adaptation
  - parameter tying and clustering

# Conclusions



- Conventional HMM
  - highly refined modelling framework
  - works well for highly constrained LVCSR
  - poor performance on mildly constrained LVCSR (almost 50% WER on the SWB corpus!)
  - superficial model structure does not allow intelligent adaptation

# Conclusions



## ■ New HDM

- still requires a lot of development work
- crude HDM already outperforms polished HMM on the SWB corpus using N-best rescoring
- potential for intelligent (i.e. less data needed) adaptation to new speakers and environments
- more powerful modelling structure
  - good: fewer parameters, better generalisation
  - bad: more complex and powerful strategies needed

# Planned Sequels



- **HTK Workshop (next week!)**
  - Walk through the Entropic HTK Tutorial on building a continuous speech recognition system using all the “tricks in the book”
- **Switched State-space Parameter Estimation**
  - Details of the EM and EKF estimation theory, algorithms, results and issues
- **Relational Grammar for ATIS Task**
  - Discussion of conceptual relational grammar approach for NLP of the ATIS task