Speaker & Speech Recognition

by

Dr Philip Jackson

Centre for Vision, Speech & Signal Processing (CVSSP)
Department of Electronic Engineering

UNIVERSITY OF SURREY

www.ee.surrey.ac.uk/Teaching/Courses/eem.ssr

Module overview

- Speaker Recognition
  - Use of speech as a biometric
  - 6 hours lecture by Prof Josef Kittler

- Speech Recognition
  - Automatic transcription of spoken language into text
  - 21 hours lecture + 3 labs by Dr Philip Jackson

Contents of speech recognition

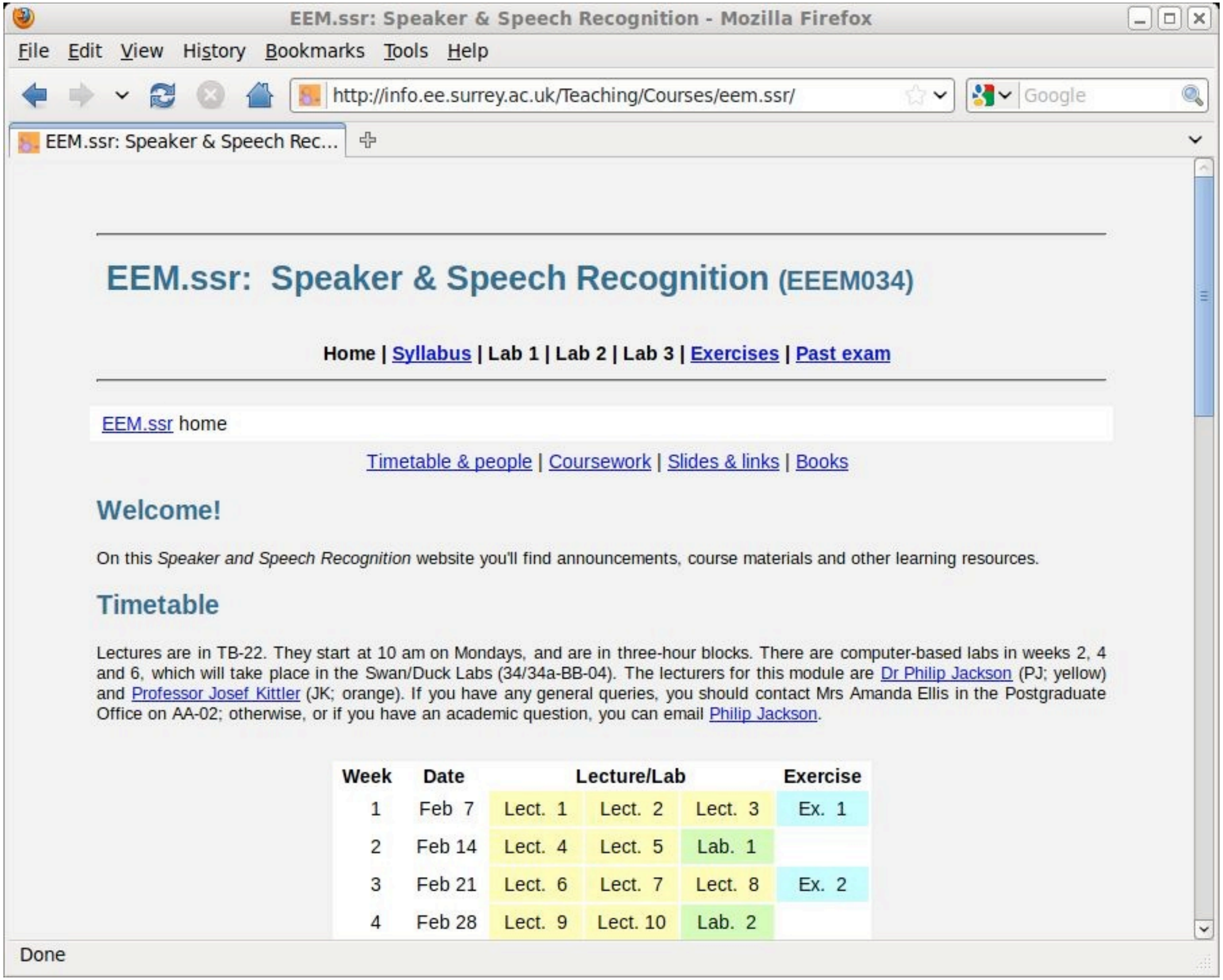
- Introduction to automatic speech recognition (ASR)
  - Speech production and vocal tract acoustics
- Speech as spoken language
  - Phonetics, syntax and language modeling
- Machine processing of speech for recognition
  - Speech patterns and feature extraction
- Statistical modeling of speech
  - Hidden Markov models
- Advanced topics in ASR
  - Speaker adaptation, noise robustness

Getting the most out of the course

- Preparation
  - Expect to research topics prior to class, complete homework
- Minimise disruption in class
  - Arrive on time, phone on silent, no eating
- Interaction
  - Contribute in class, ask questions of general concern, stop me if a problem arises
- Making notes
  - Bring pens and paper, add comments/sketches, date and file
- Learning continues
  - Books, slides, exercises and past exam papers available via the module website:

    http://www.ee.surrey.ac.uk/Teaching/Courses/eem.ssr/

Module web site



EEM.ssr: Speaker & Speech Recognition - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://info.ee.surrey.ac.uk/Teaching/Courses/eem.ssr/
Google
EEM.ssr: Speaker & Speech Rec...
Done

EEM.ssr: Speaker & Speech Recognition (EEEM034)

Home | Syllabus | Lab 1 | Lab 2 | Lab 3 | Exercises | Past exam

EEM.ssr home

Timetable & people | Coursework | Slides & links | Books

Welcome!

On this *Speaker and Speech Recognition* website you'll find announcements, course materials and other learning resources.

Timetable

Lectures are in TB-22. They start at 10 am on Mondays, and are in three-hour blocks. There are computer-based labs in weeks 2, 4 and 6, which will take place in the Swan/Duck Labs (34/34a-BB-04). The lecturers for this module are Dr Philip Jackson (PJ; yellow) and Professor Josef Kittler (JK; orange). If you have any general queries, you should contact Mrs Amanda Ellis in the Postgraduate Office on AA-02; otherwise, or if you have an academic question, you can email Philip Jackson.

Week	Date	Lecture/Lab			Exercise
1	Feb 7	Lect. 1	Lect. 2	Lect. 3	Ex. 1
2	Feb 14	Lect. 4	Lect. 5	Lab. 1	
3	Feb 21	Lect. 6	Lect. 7	Lect. 8	Ex. 2
4	Feb 28	Lect. 9	Lect. 10	Lab. 2	

UNIVERSITY OF SURREY

Module assessment

- Exam (**60%**):
  - 2 hour written paper, answer 3 out of 4 questions

- Coursework (**40%**):
  - 3 computer-based lab assignments
  - presented in weeks 2, 4 and 8

- Outcomes:
  - Develop an understanding of spoken language processing
  - Derive fundamentals of statistical machine learning
  - Construct your own program to recognize words

Coursework

- Lab 1: Speech enrolment
  - Issued in Week 2
  - Deadline Week 4

- Lab 2: Feature extraction
  - Issued in Week 3
  - Deadline Week 6

- Lab 3: HMM training and recognition
  - Issued in Week 6
  - Deadline Week 10

Module pre-requisites

- Digital sampling of audio signals
  - quantisation and aliasing, discrete Fourier transform (DFT)
  - short-time Fourier transform (STFT) and spectrograms
- Z domain for modeling discrete-time linear systems
  - autoregressive (all-pole) transfer function
  - linear predictive coding (LPC) and autocorrelation method for computing coefficients (Levinson-Durbin)
- Cepstral/homomorphic analysis
  - calculation of the cepstrum
  - relation of cepstral coefficients to log spectral envelope
- Speech science
  - Source-filter theory of speech production
  - Critical bands and non-linear operations in sound perception
  - Speech processing technologies and applications

Introduction to Speech Recognition

What is speech recognition?

- The task of speech recognition is to decode the acoustical signal into the sequence of words
- It forms part of spoken language understanding:
  - automatic speech recognition
  - natural language understanding
- **Automatic speech recognition** (ASR) converts spoken words to machine-readable form
  - e.g., from audio signal to commands/text
- Natural language understanding seeks a higher cognitive interpretation:
  - i.e., structure, meaning and even intention

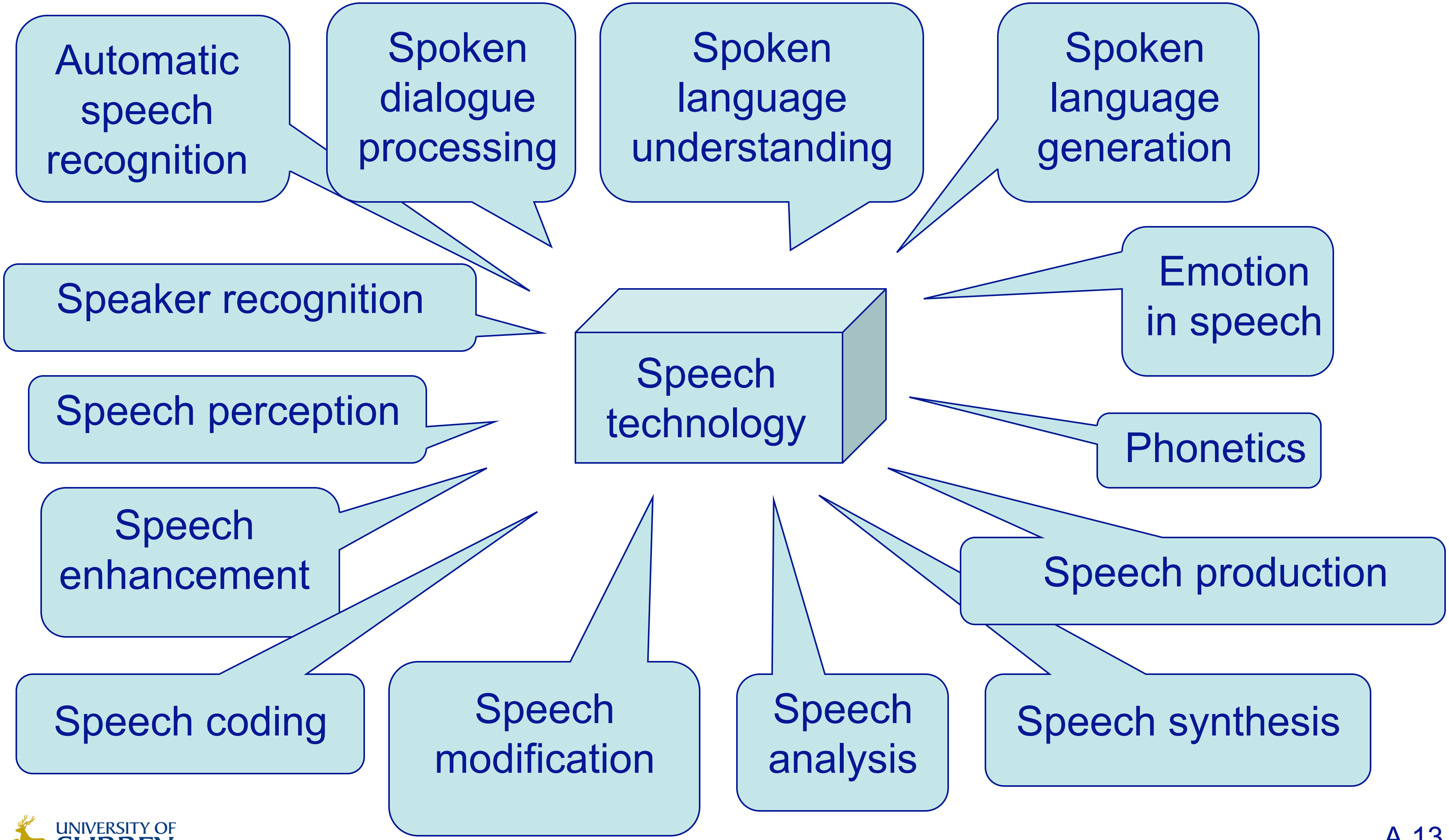
Why is it important?

- Business/desktop applications
  - Dictation/diarisation/voice indexing
  - Voice command
- Voice enabled services/mobile applications
  - Server-based systems via voice user interface (VUI)
  - Voice search and browsing
- Games and interactive entertainment
- Education
  - First and foreign language acquisition
- Speech therapy and rehabilitation
- Hearing assistance and widening access
  - Subtitling

It comes as naturally as breathing…

- Humankind’s preferred modality
- Natural language is good for interacting with complex systems
- Hands-free
- Eyes-free
- Small footprint
- No specialist training required

World of speech technologies



UNIVERSITY OF SURREY

What makes speech recognition challenging?

- The dream and reality
  - Intelligent machines?
  - Size of vocabulary: 50, 1000, 20000 words
  - Speaker -dependent/-independent ASR
- Discovering our ignorance
  - How does the ear work?
  - How is information encoded in an acoustic signal?
  - How do the auditory cortex and the brain process sounds to decode an acoustic message?
- Circumventing our ignorance
  - Ad-hoc rules vs. pattern matching
  - Probabilistic approaches using statistical models
  - Artificial neural networks and machine learning techniques

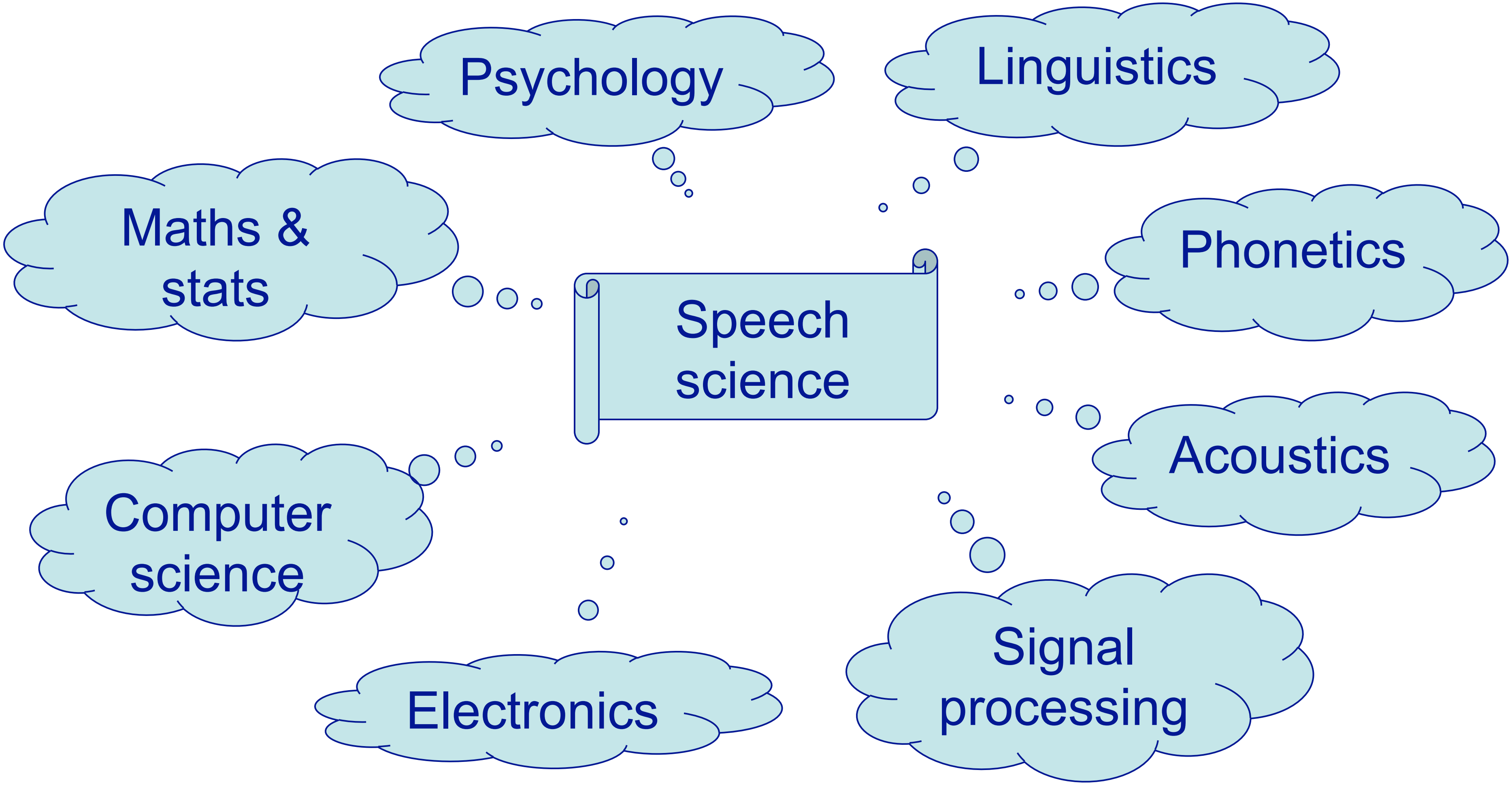
Factors affecting the difficulty

- Speaker dependency
  - Acoustic variability from one person to the next
- Vocabulary size
  - Dealing with thousands of word templates
- Isolated words vs. continuous speech
  - Reduced pronunciation in spontaneous speech
- Language constraints and knowledge sources
  - Language is alive, fluid and constantly changing
- Acoustic ambiguity
  - Many word segments are easy to confuse
- Noise robustness
  - Normal listening conditions include noise, acoustic reflections and even other voices

What can ASR do for you?

- Simple data entry
  - yes/no
  - credit card details
- Appliance control
  - voice dialling
  - domotics
  - aircraft cockpit direct voice input
- VUI for text processing
  - dictation and word processing
  - email and SMS input
- Telephone services
  - call-centre routing
  - form filling
- Mobile, online and on-demand services
  - Voice-enabled applications
  - web browsing
  - content-based spoken audio search
  - spoken language translation
- Other
  –

What do we need to study to understand speech?



Speech recognition summary

- Dream and reality
  - Speech-to-text machines
  - Vocabulary size and flexibility traded for recognition accuracy
- Incomplete specification
  - Of language, of the background noise and acoustic environment,
    of the human ear and auditory processing, and
    of how the brain extracts meaning from speech
- An engineering solution
  - Use statistical pattern matching techniques
  - Most successful based on Hidden Markov Models
  - Employ large databases for training
  - Research continues to explore alternatives, e.g., HMM/ANN hybrids, trajectory HMM, dynamic Bayesian networks