EPSRC
Engineering and Physical Sciences
Research Council

Natural
Speech
Technology
Edinburgh – Cambridge – Sheffield

Natural Speech Technology Programme Overview

Steve Renals
Centre for Speech Technology Research
University of Edinburgh

28 May 2015

http://www.natural-speech-technology.org

NST Programme Grant

- At the outset of NST we identified several weaknesses with speech technology systems
  - Fragile operation across domains
  - Synthesis and recognition developed independently
  - Reliance on supervised approaches, manually transcribed training data
  - Models for synthesis and recognition include relatively little speech knowledge
  - Models only weakly factor the underlying sources of variability
  - Systems react crudely (if at all) to the context / environment
- These weaknesses still drive our objectives

NST Technical Objectives

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

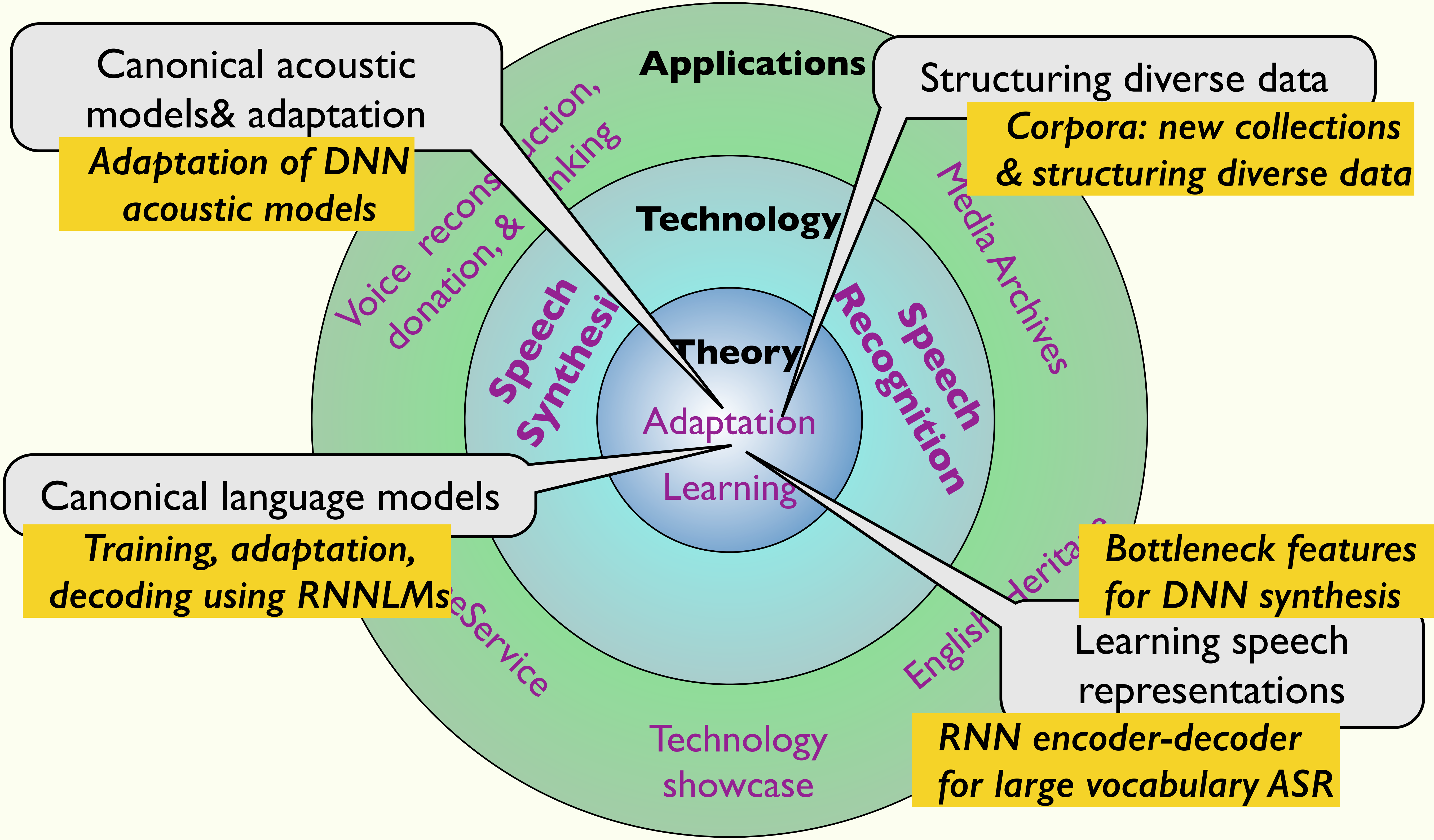
- **Learning and Adaptation**
  - learning to compactly represent speech and to adapt to new scenarios and speaking styles
- **Natural Speech Transcription**
  - Speech recognition systems that operate seamlessly across domain and acoustic environment
- **Natural Speech Synthesis**
  - Controllable synthesisers that learn from data, and can generate expressive conversational speech
- **Exemplar Applications**
  - prototype deployment in applications, focusing on health/social domain, media, and the needs of User Group stakeholders

NST
Highlights 2014-15

- Best paper awards at IEEE SLT-2015, IEEE ICASSP-2014
- Open source software – HTK, Kaldi, HTS
- Speech Recognition applications – BBC (NewsHack and MGB Challenge), Ericsson (Just-in-time ASR), MediaEval, Browsing oral history (English Heritage)
- Voice banking and reconstruction
- homeService
- Challenges and Evaluations – Spoofing challenge at

Learning and adaptation

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



Canonical acoustic models& adaptation

Adaptation of DNN acoustic models

Structuring diverse data

Corpora: new collections & structuring diverse data

Canonical language models

Training, adaptation, decoding using RNNLMs

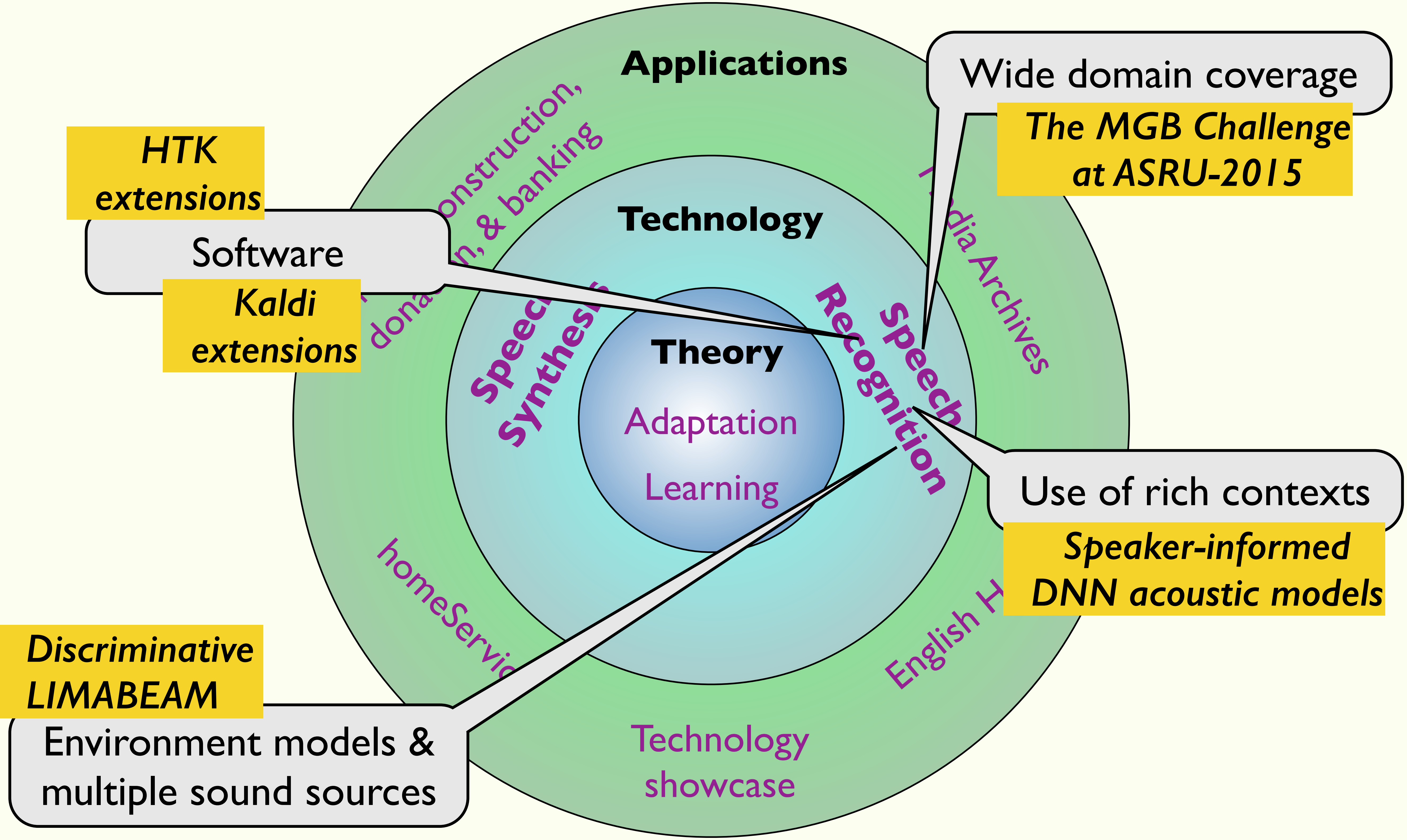
Bottleneck features for DNN synthesis

Learning speech representations

RNN encoder-decoder for large vocabulary ASR

Natural transcription

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



HTK extensions

Software

Kaldi extensions

Wide domain coverage

The MGB Challenge at ASRU-2015

Use of rich contexts

Speaker-informed DNN acoustic models

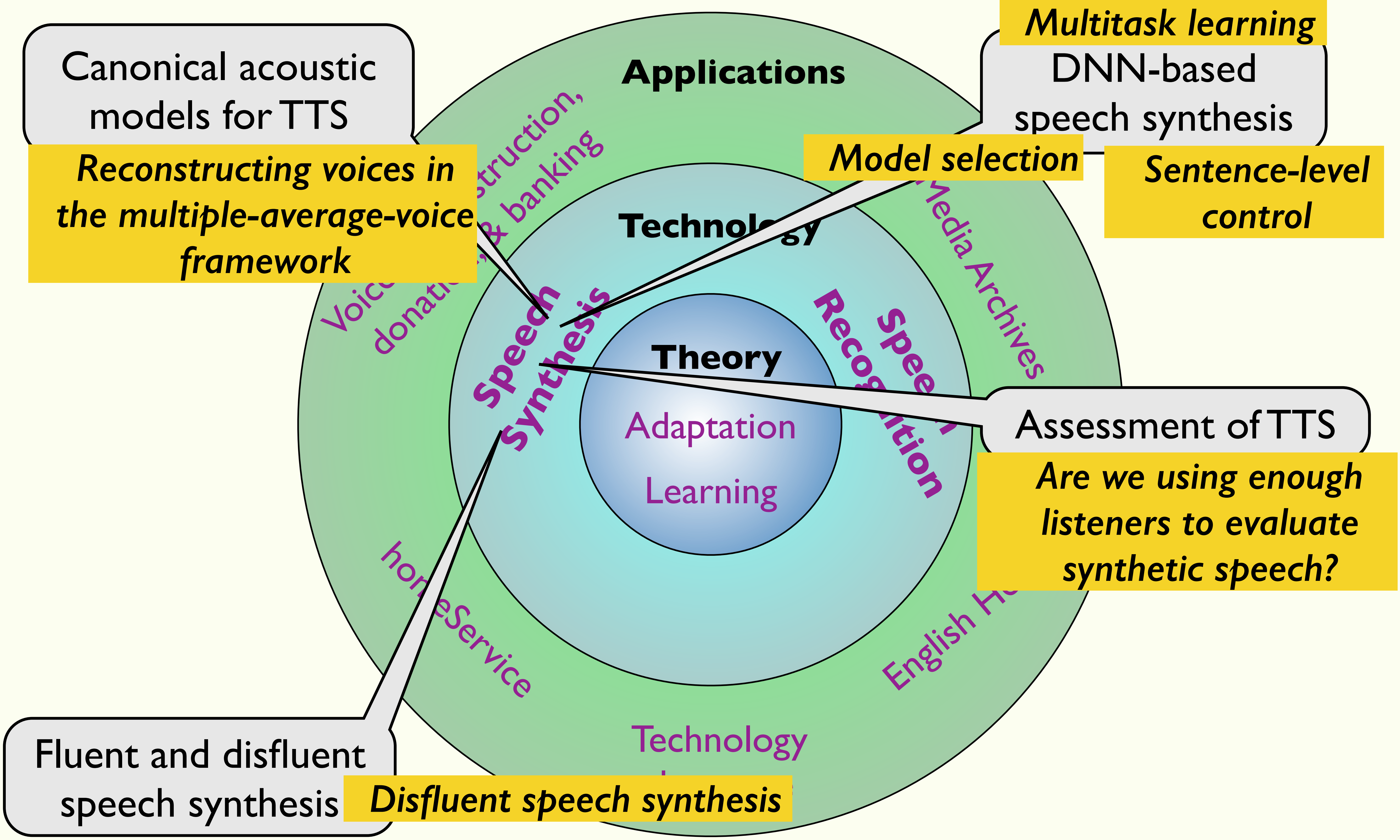
Discriminative LIMABEAM

Environment models & multiple sound sources

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Natural Synthesis



Canonical acoustic models for TTS

Reconstructing voices in the multiple-average-voice framework

Multitask learning

DNN-based speech synthesis

Model selection

Sentence-level control

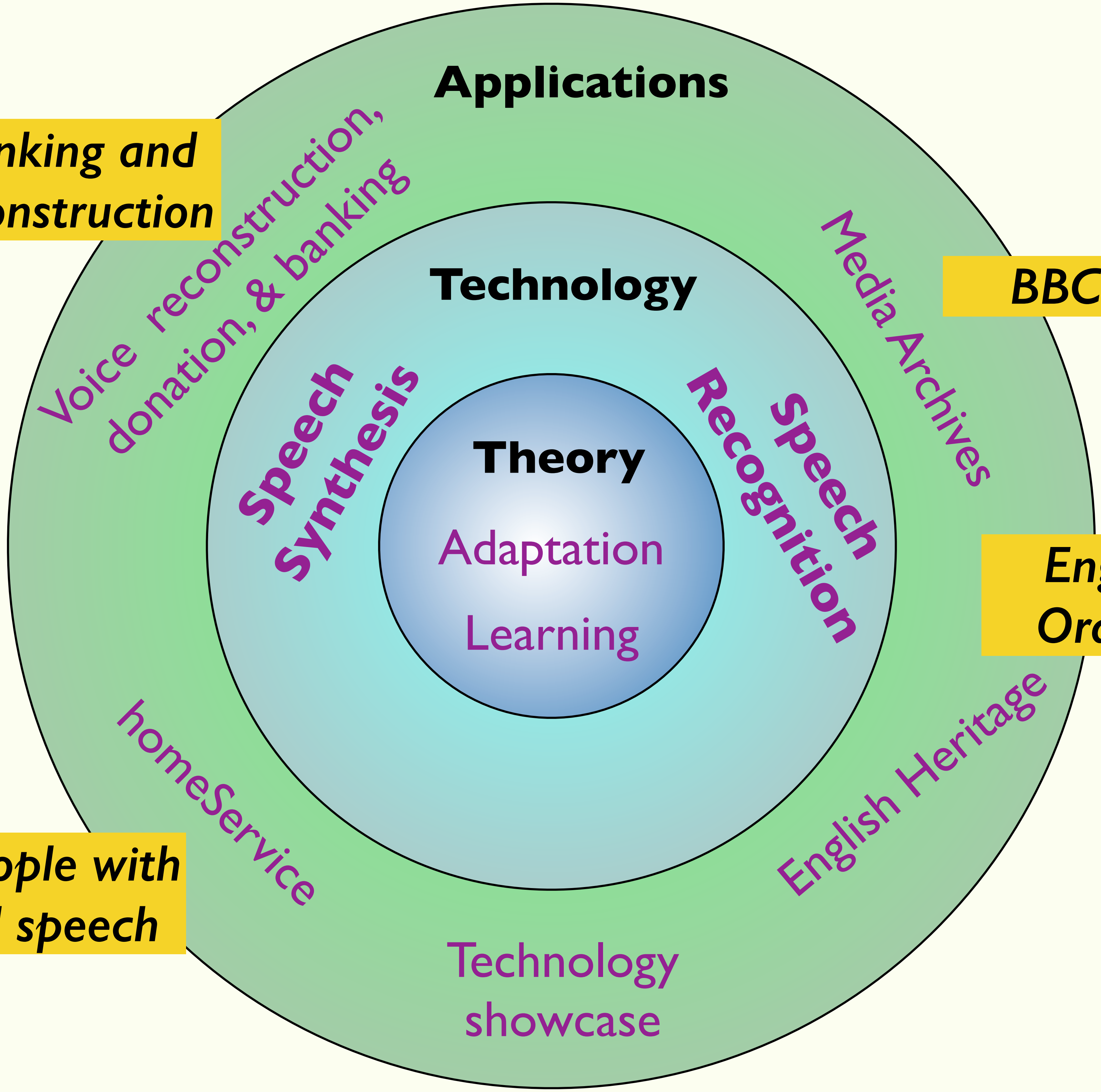
Assessment of TTS

Are we using enough listeners to evaluate synthetic speech?

Fluent and disfluent speech synthesis

Disfluent speech synthesis

Applications



Voice banking and voice reconstruction

BBC transcription

English Heritage - Oral History demo

ASR for people with disordered speech

HTK ANN Extensions

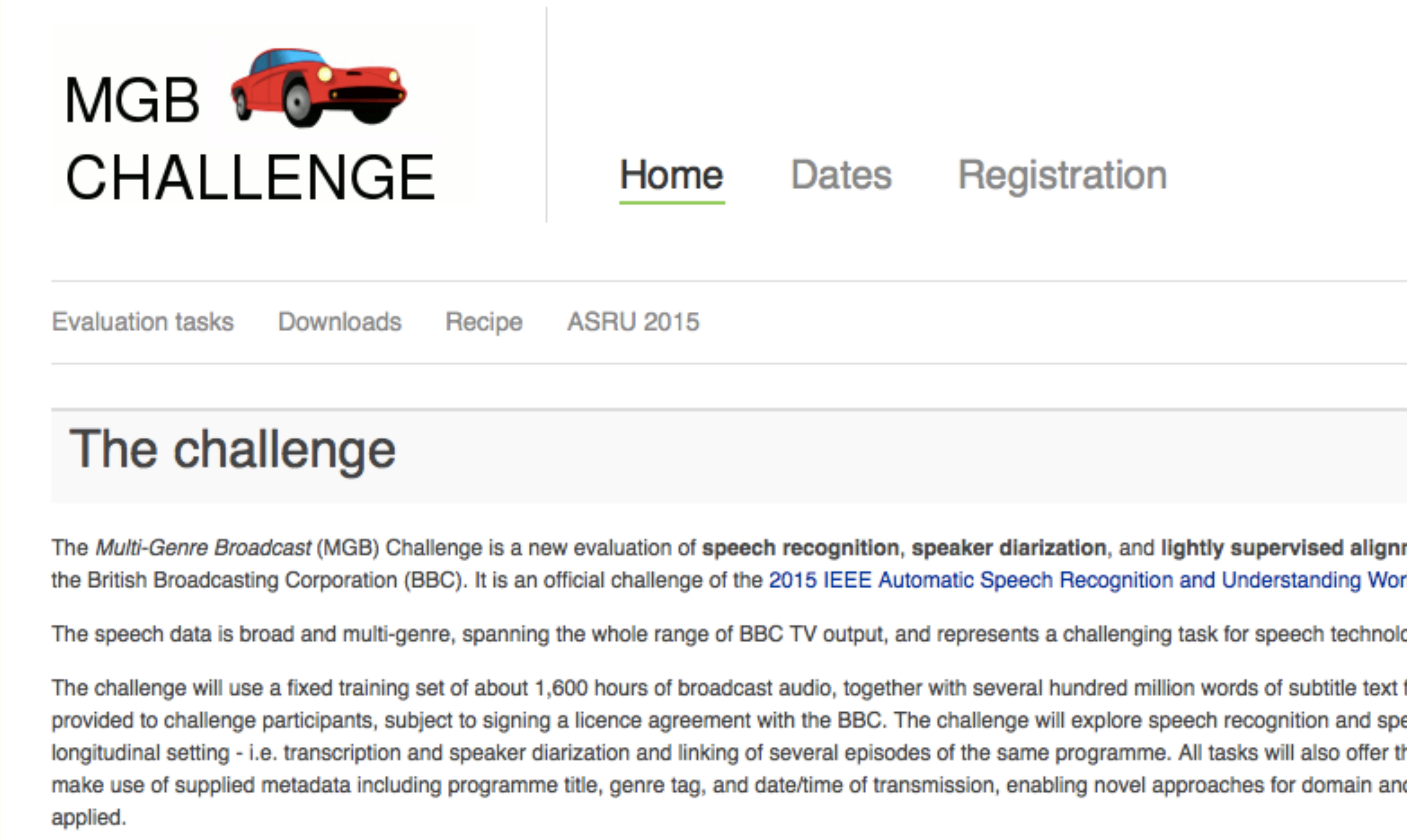
- HTK 3.5 will support ANNs, maintaining compatibility with most existing functions.
  - Minimises the effort to reuse previous source code and tool
  - Allows transfer of e.g. SI/SD input transforms, MPE/MMI sequence training
  - 64-bit compatible
- Generic extensions
  - Flexible input feature configurations
  - ANN structures can be any directed acyclic graph
  - Stochastic gradient descent supporting frame/sequence training
  - CPU/GPU math kernels for ANNs
- Decoders extended to support tandem/hybrid systems, system combination

HTK Language Model extensions

- HTK v3.5 support for decoding RNN language models
    - Lattice rescoring using RNNLMs
    - Class / Full word outputs, interpolation with n-grams
    - Similar functionality for feed-forward NN LMs
- RNNLM estimation enhancements
    - bunch mode GPU training
    - full/class output RNN LMs
    - NCE training
    - variance regularised training

MGB Challenge

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



The challenge

The *Multi-Genre Broadcast* (MGB) Challenge is a new evaluation of **speech recognition**, **speaker diarization**, and **lightly supervised alignr** the British Broadcasting Corporation (BBC). It is an official challenge of the 2015 IEEE Automatic Speech Recognition and Understanding Wor

The speech data is broad and multi-genre, spanning the whole range of BBC TV output, and represents a challenging task for speech technolo

The challenge will use a fixed training set of about 1,600 hours of broadcast audio, together with several hundred million words of subtitle text provided to challenge participants, subject to signing a licence agreement with the BBC. The challenge will explore speech recognition and spe longitudinal setting - i.e. transcription and speaker diarization and linking of several episodes of the same programme. All tasks will also offer th make use of supplied metadata including programme title, genre tag, and date/time of transmission, enabling novel approaches for domain and applied.

Spoofing Challenge

Natural
Speech
Technology
Edinburgh – Cambridge – Sheffield

ASVspoof 2015:

Automatic Speaker Verification Spoofing and Countermeasures Challenge

Introduction

Do you have a method/algorithm to discriminate between human and synthetic speech (generated from speech synthesis or voice conversion systems)? If so, you are invited to take part in the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenge.

Previously, both spoofing attacks and countermeasures have been developed with full knowledge of a particular speaker verification system used for vulnerability assessments. Similarly, countermeasures have been developed with full knowledge of the spoofing attack which they are designed to detect. This is clearly unrepresentative on the real use case scenario in which the specific attack, much less the specific algorithm, can never been known a priori. It is thus likely that the prior work has as much over-exaggerated the threat of spoofing as it has the performance of countermeasures.

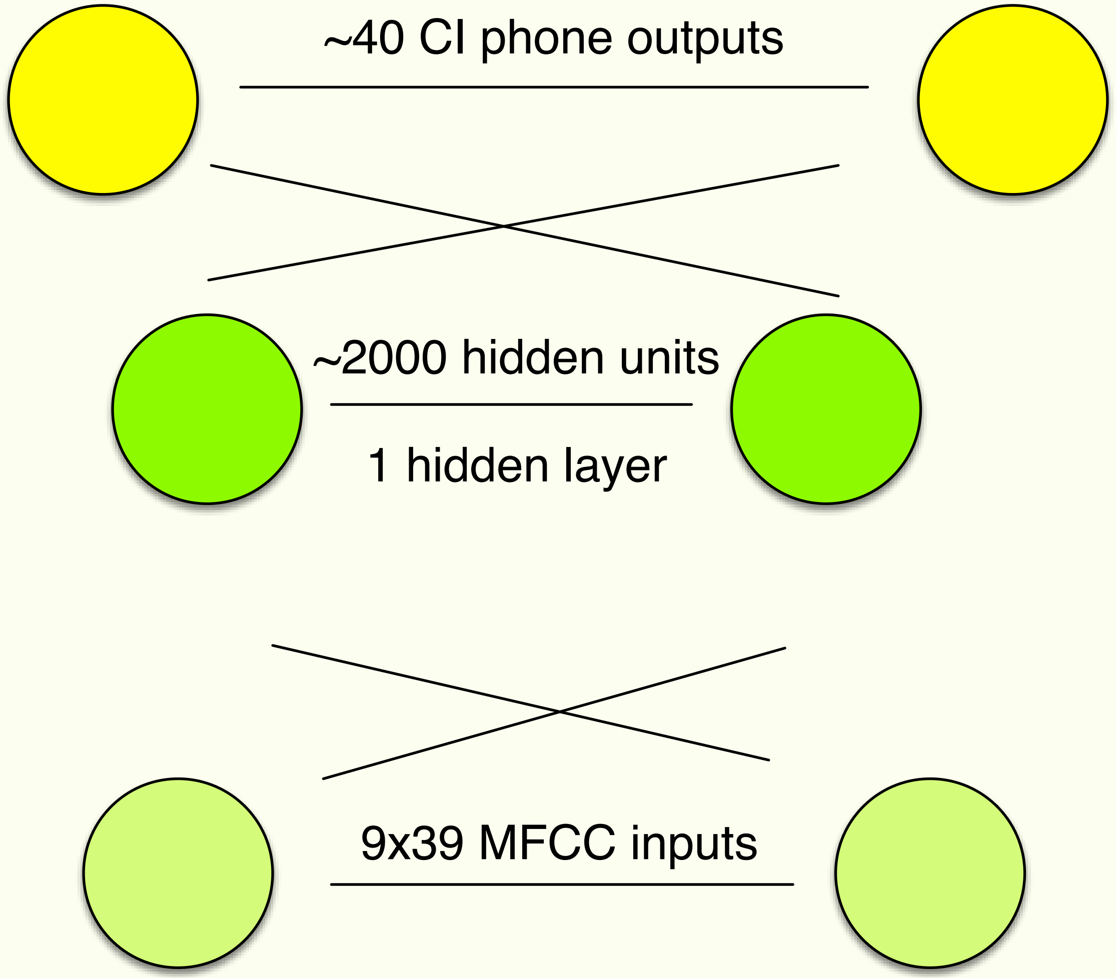
The ASVspoof challenge has been designed to help break this mould and to support, for the first time, independent assessments of vulnerabilities to spoofing and of countermeasure performance. While preventing as much as possible the inappropriate use of prior knowledge, the challenge aims to stimulate the development of generalised countermeasures with potential to detect varying and unforeseen spoofing attacks.

The first evaluation, ASVspoof 2015, is being held within the scope of a special session at INTERSPEECH 2015 and with a focus on spoofing detection. Participants are invited to submit spoofing detection results. You will be provide with a spoofing database along with a protocol for experiments. The spoofing database is generated from

The Rise
# of Neural Nets

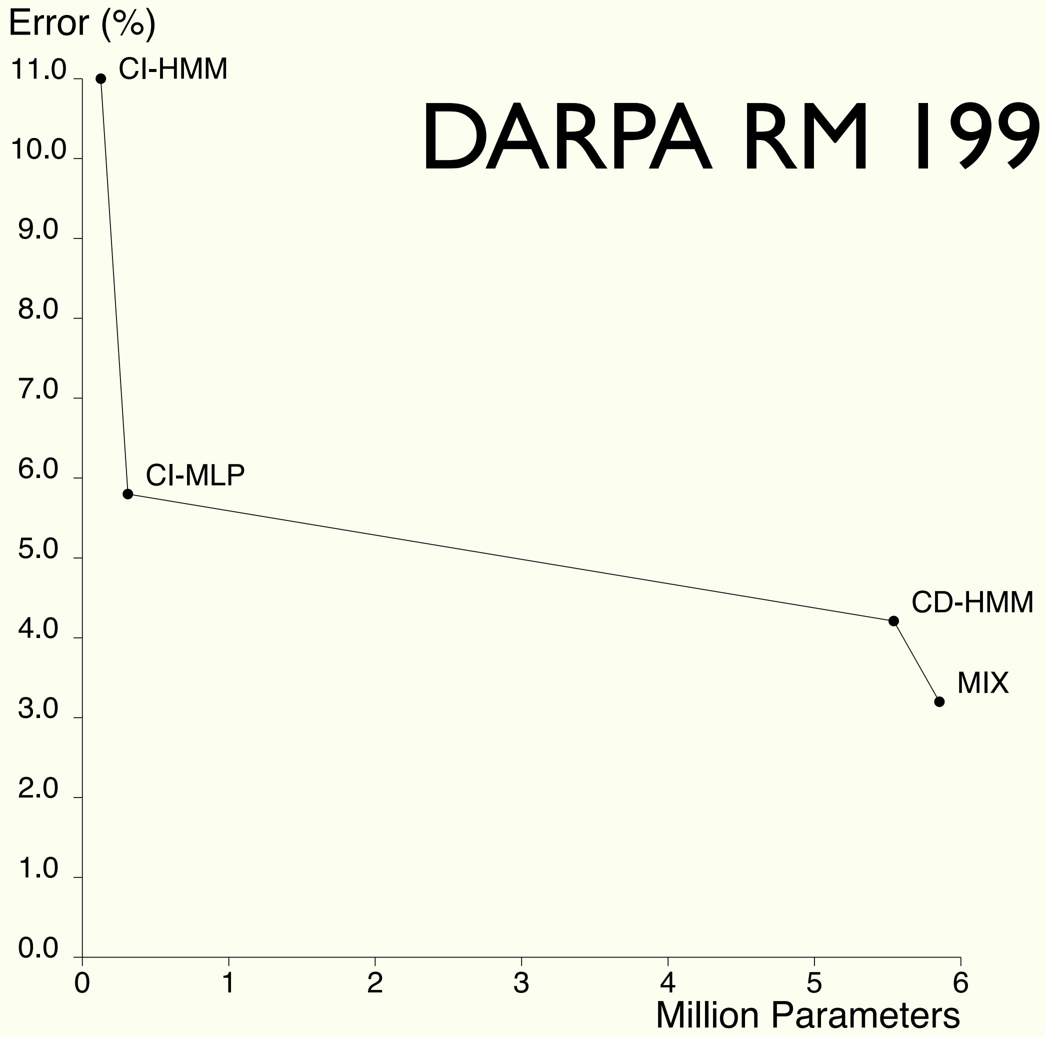
The Rise and Fall and Rise of Neural Nets

Neural network acoustic models (1990s)



Bourlard & Morgan, 1994

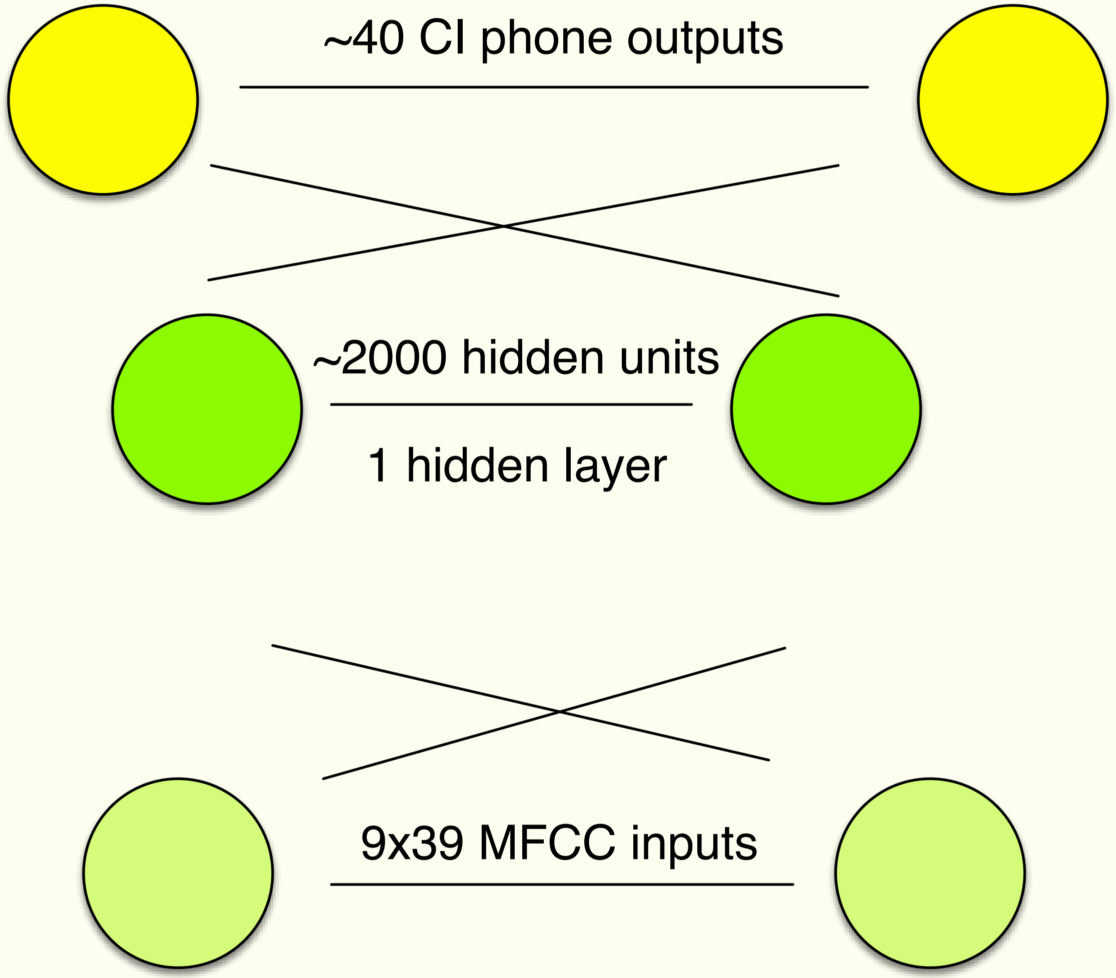
DARPA RM 1992



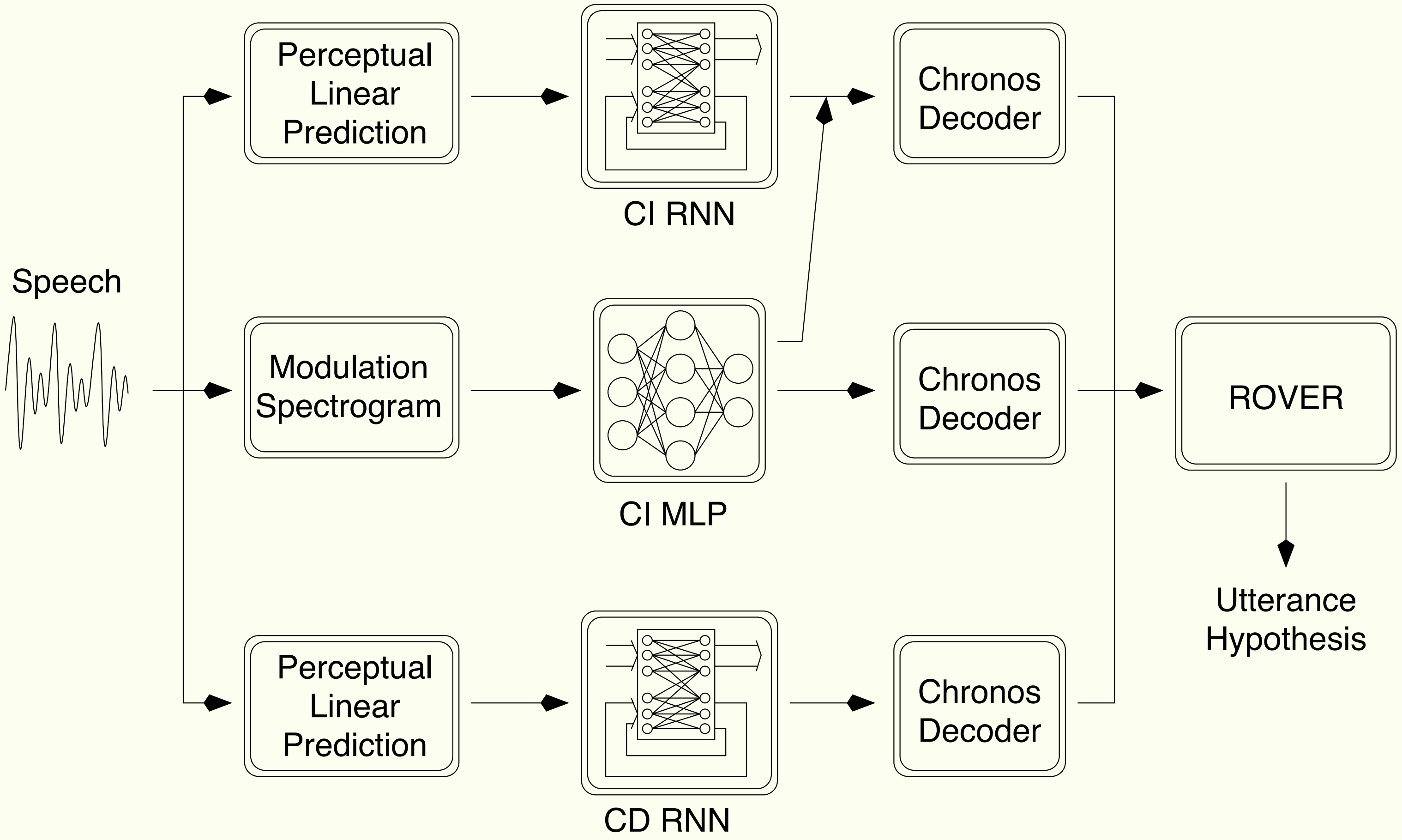
Renals, Morgan, Cohen & Franco, ICASSP 1992

Neural network acoustic models (1990s)

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



Bourlard & Morgan, 1994



Broadcast news 1998

20.8% WER

(best GMM-based system, 13.5%)

Cook, Christie, Ellis, Fosler-Lussier, Gotoh, Kingsbury, Morgan, Renals, Robinson, & Williams, DARPA, 1999

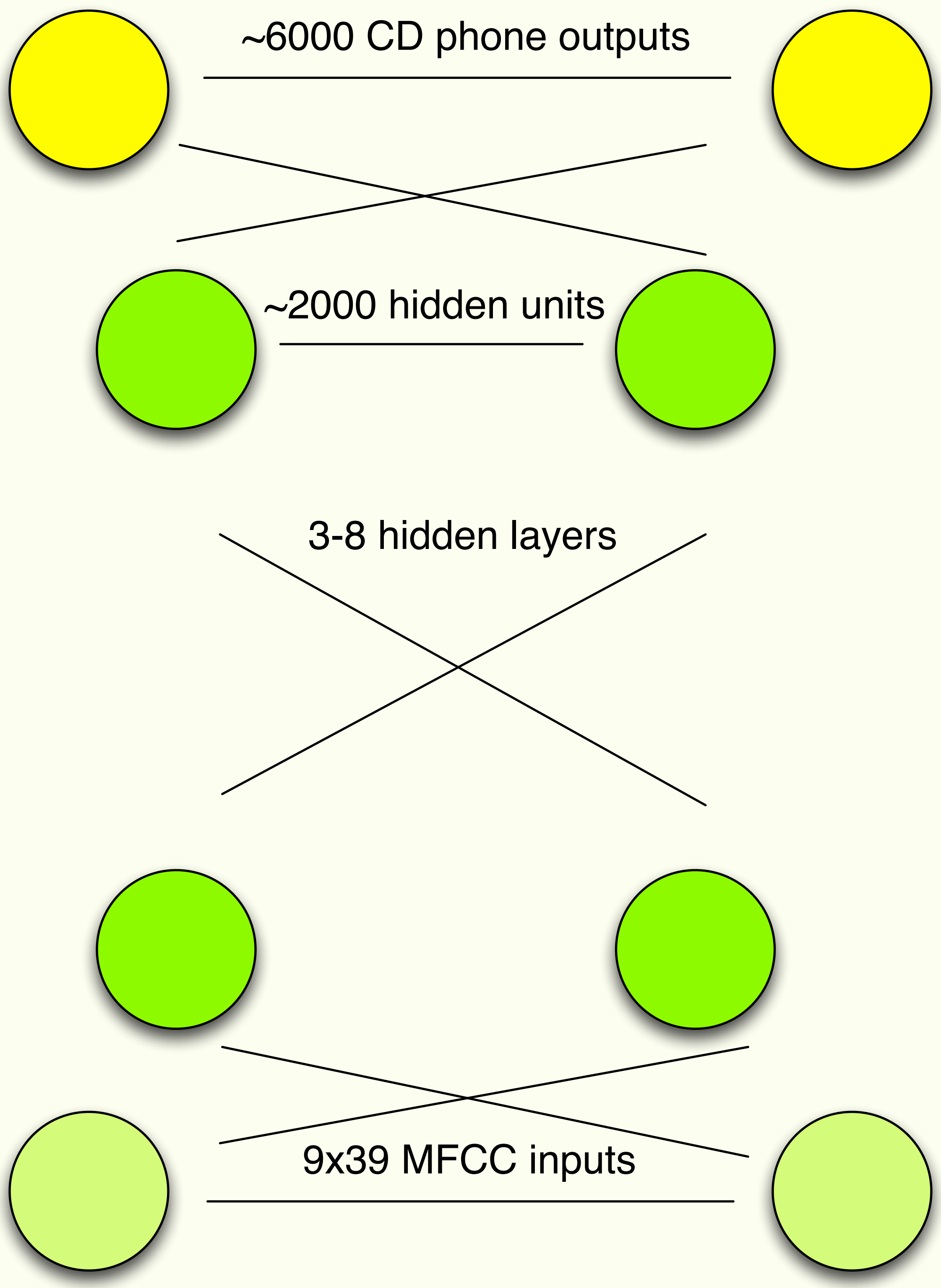
NN acoustic models
Limitations vs GMMs

- Computationally restricted to monophone outputs
  - CD-RNN factored over multiple networks – limited within-word context
- Training not easily parallelisable
  - experimental turnaround slower
  - systems less complex (fewer parameters)
    - RNN – <100k parameters
    - MLP – ~1M parameters
- Rapid adaptation hard (cf MLLR)

NN acoustic models
Benefits

- Fewer limitations on inputs
  - Correlated features
  - Multi-frame windows
- Discriminative training criteria (frame level and sequence level)
- Can be used to generate 'higher-level' features
  - tandem, posteriorgrams
  - bottleneck features

(Deep) neural network acoustic models (2010s)



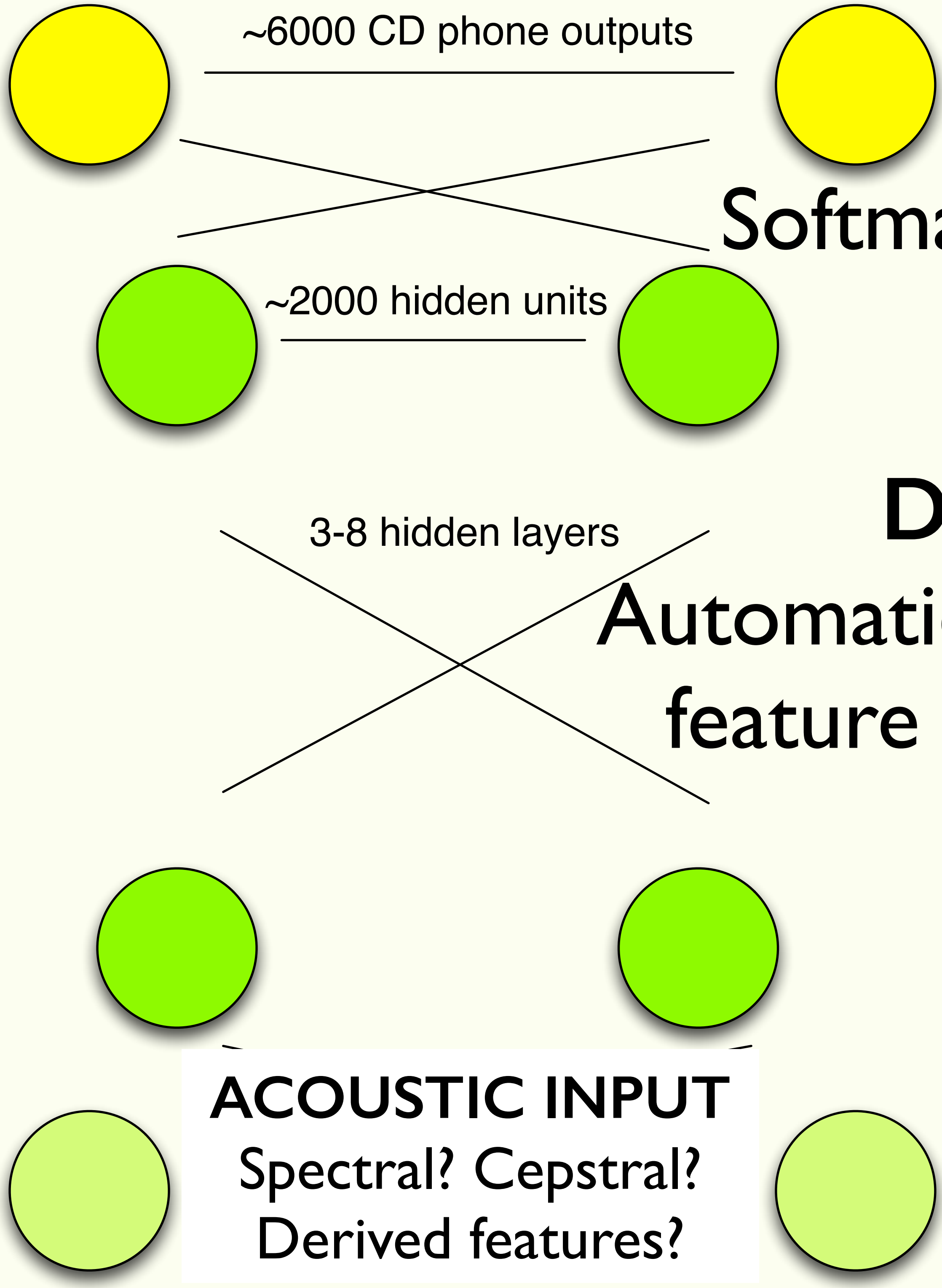
Dahl, Yu, Deng & Acero, IEEE TASLP 2012

Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath & Kingsbury, IEEE SP Mag 2012

(Deep) neural network acoustic models (2010s)

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



~6000 CD phone outputs

~2000 hidden units

3-8 hidden layers

ACOUSTIC INPUT

Spectral? Cepstral?
Derived features?

WIDE

Softmax output layer

DEEP

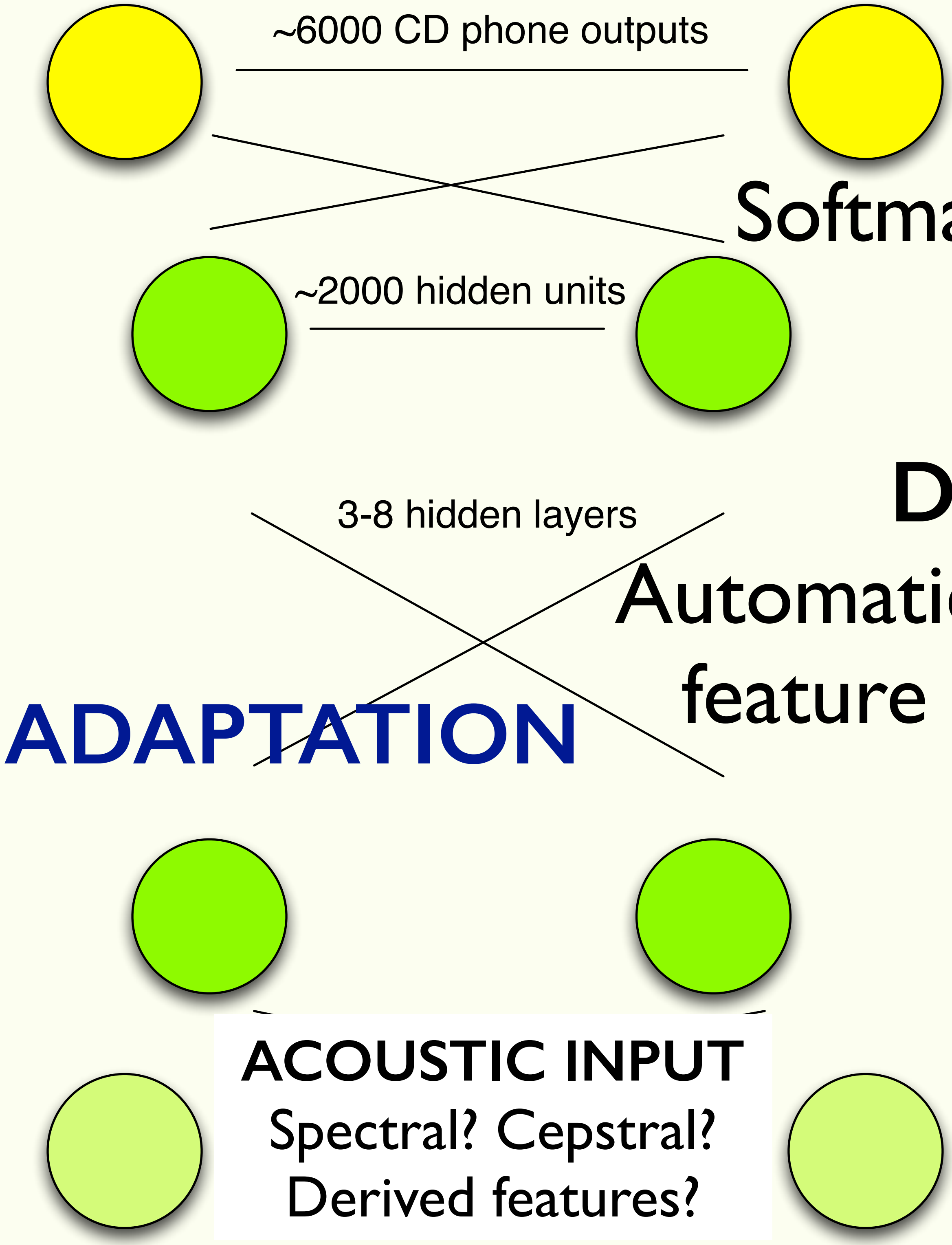
Automatically learned feature extraction

Dahl, Yu, Deng & Acero, IEEE TASLP 2012

Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath & Kingsbury, IEEE SP Mag 2012

(Deep) neural network acoustic models (2010s)

Natural
Speech
Technology
Edinburgh – Cambridge – Sheffield



WIDE

Softmax output layer

DEEP

Automatically learned feature extraction

ADAPTATION

ACOUSTIC INPUT
Spectral? Cepstral?
Derived features?

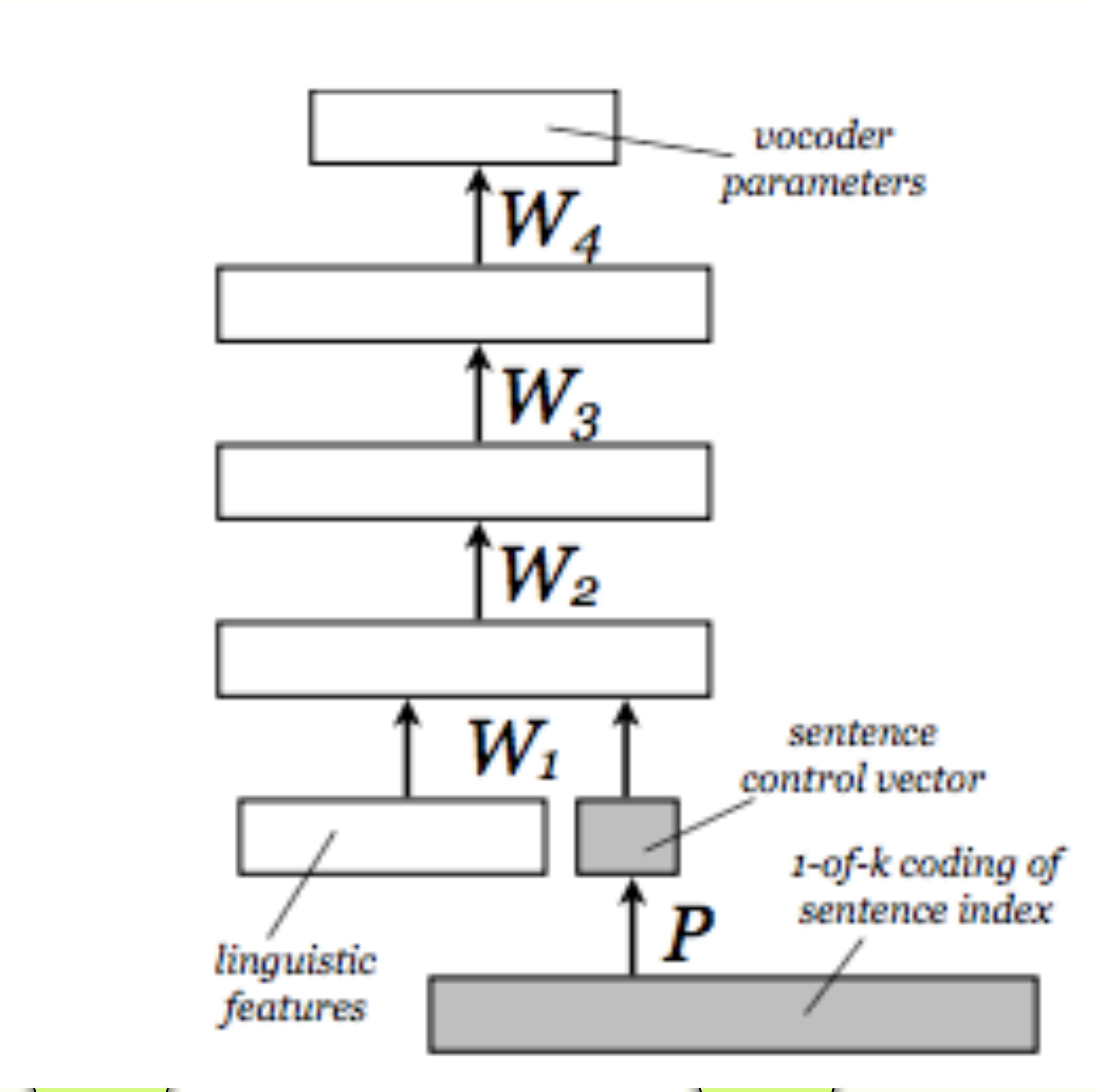
TRAINING
– optimisation, objective fn

WEIGHT SHARING
– adaptation, CNNs

ARCHITECTURES
– recurrent, convolutional, …

ACTIVATION FUNCTIONS
– pooling, RELU, gated units

(Deep) neural network acoustic models (2010s)



TRAINING

– optimisation, objective fn

ut layer

WEIGHT SHARING

– adaptation, CNNs

rned

ARCHITECTURES

on

– recurrent, convolutional, …

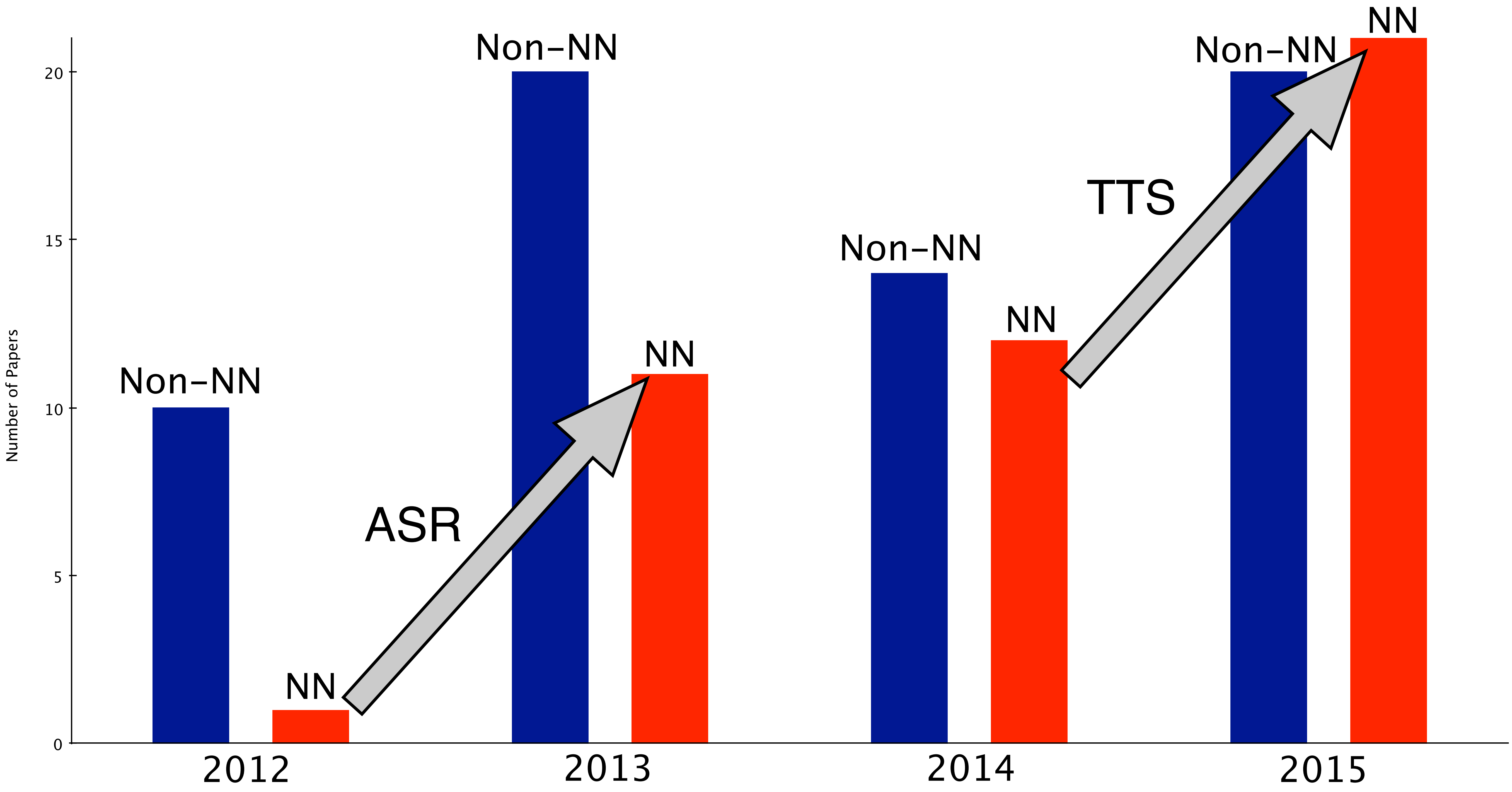
IVATION FUNCTIONS

ooling, RELU, gated units

Derived features?

Neural networks & NST

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



Today's agenda

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- 9:30 – 11:20 Intro, 4 talks, poster spotlights
- 11:20 – 13:00 Coffee + demos/posters [LR4, ground floor]
- 13:00 – 14:15 Lunch
- 14:15 – 15:15 3 talks
- 15:15 – 15:45 Coffee
- 15:45 – 16:45 Discussion: Clinical, Media, Future Challenges
- 16:45 - 17:00 Wrap-up
- *17:00 – 18:30 Advisory board meeting*
- 19:00 Dinner at Emmanuel College

EPSRC
Engineering and Physical Sciences
Research Council