EPSRC
Engineering and Physical Sciences Research Council

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Natural Speech Technology

Steve Renals
Centre for Speech Technology Research
University of Edinburgh

28 June 2016

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

Natural Speech Technology

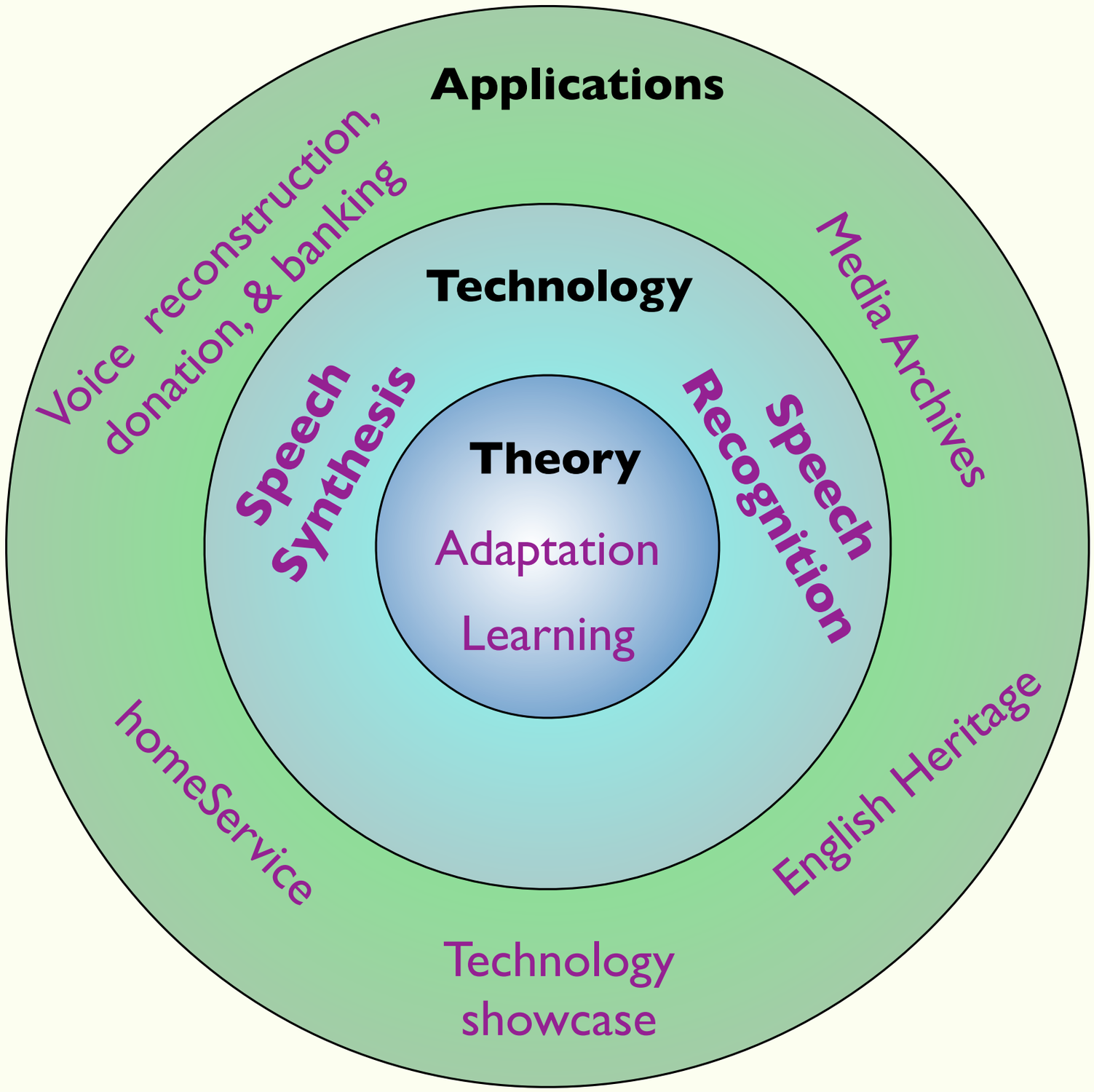
Edinburgh – Cambridge – Sheffield

- The focus
  - Learning representations for speech synthesis and recognition
  - Adapting to domains / genres / tasks / languages
  - Recognizing and generating conversational speech in "natural" acoustic conditions
  - Developing applications that drive basic technology that drives applications – user group interaction

Natural Speech Technology
Theory / Technology / Applications

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



From 2009 to 2016

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- NST Timeline…
  - 2009–2010 / construct joint vision, develop project focus
  - May 2011 / project starts
  - Nov 2011 / project running at full capacity
  - Jul 2016 / project complete
- **Things have changed a lot since 2009…**
  - People use (and like) speech-based apps – Siri, voice search, …
  - Deep neural networks have had a major impact on the field, offering significant improvements in accuracy
  - Significant increase in commercial activity

...but the big problems remain

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Challenges identified in 2009/10, addressed by NST
  - Systems are fragile when transferring across domains
    - *adaptation and domain transfer for recognition and synthesis*
  - Synthesis and recognition develop independently
    - *common approaches to ASR & TTS – representations, training*
- Reliance on manually transcribed training data
  - *MGB Challenge, speech synthesis from found data*
- Models only weakly factor the underlying sources of variability
  - *factorised approaches to modelling and adaptation*
- Speech technology models include relatively little speech knowledge
  - *new deep learning approaches – stimulated training, segmental RNNs, …*
- Systems react crudely (if at all) to the context / environment
  - *adaptively modifying intelligibility in TTS, homeService*

Broadcast media Transcription

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Transcription of broadcast TV content across all genres
- Lightly supervised training from broadcast subtitles
- Achievements
  - setting the state-of-the-art in multi-genre broadcast transcription
  - building a community around the MGB Challenges
  - working with multiple user partners including BBC and Ericsson
  - systems developed at BBC NEWSHack
  - systems deployed via WebASR
  - open source software (e.g. HTK-3.5 release)

DNN Speech Synthesis

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Many DNN architectures explored for speech synthesis
  - multi-task DNNs
  - LSTM for speech synthesis
  - deep generative models
  - DNNs to guide unit selection
  - DNN duration modelling
- **Multilingual systems – translation and generation of broadcast content**
  - automatic scaling TTS to many languages

Increasing naturalness of speech synthesis

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- **Fluency, disfluency, and naturalness**
  - exploring disfluencies in natural and synthetic
  - speech synthesis of spontaneous and disfluent speech
- Speech synthesis from diverse data
- Evaluation of speech synthesis naturalness and intelligibility
- The Spoofing Challenge
- Voice banking and reconstruction

Natural
Speech
Technology

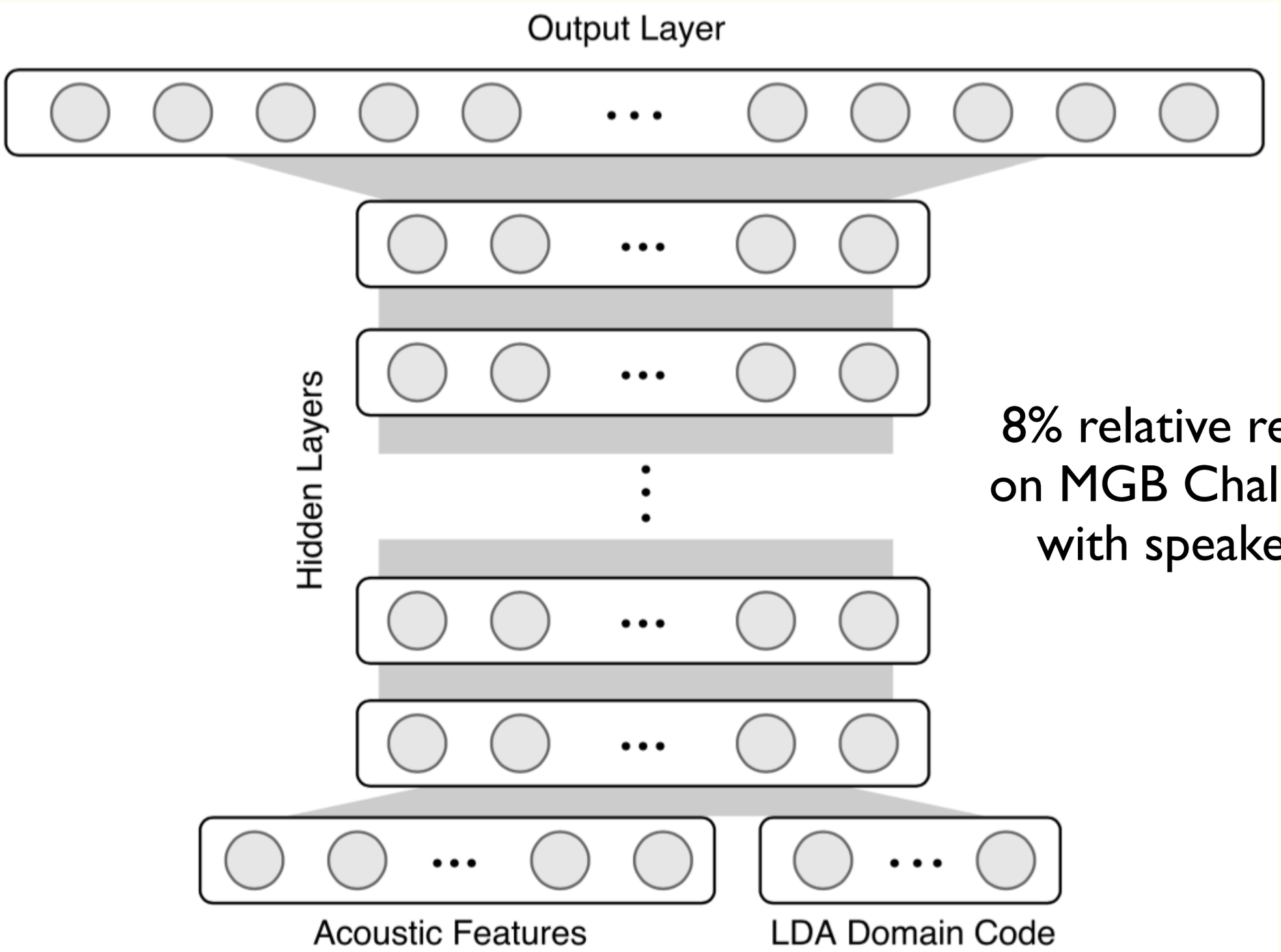
Edinburgh – Cambridge – Sheffield

Adaptation

LDA-DNN

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

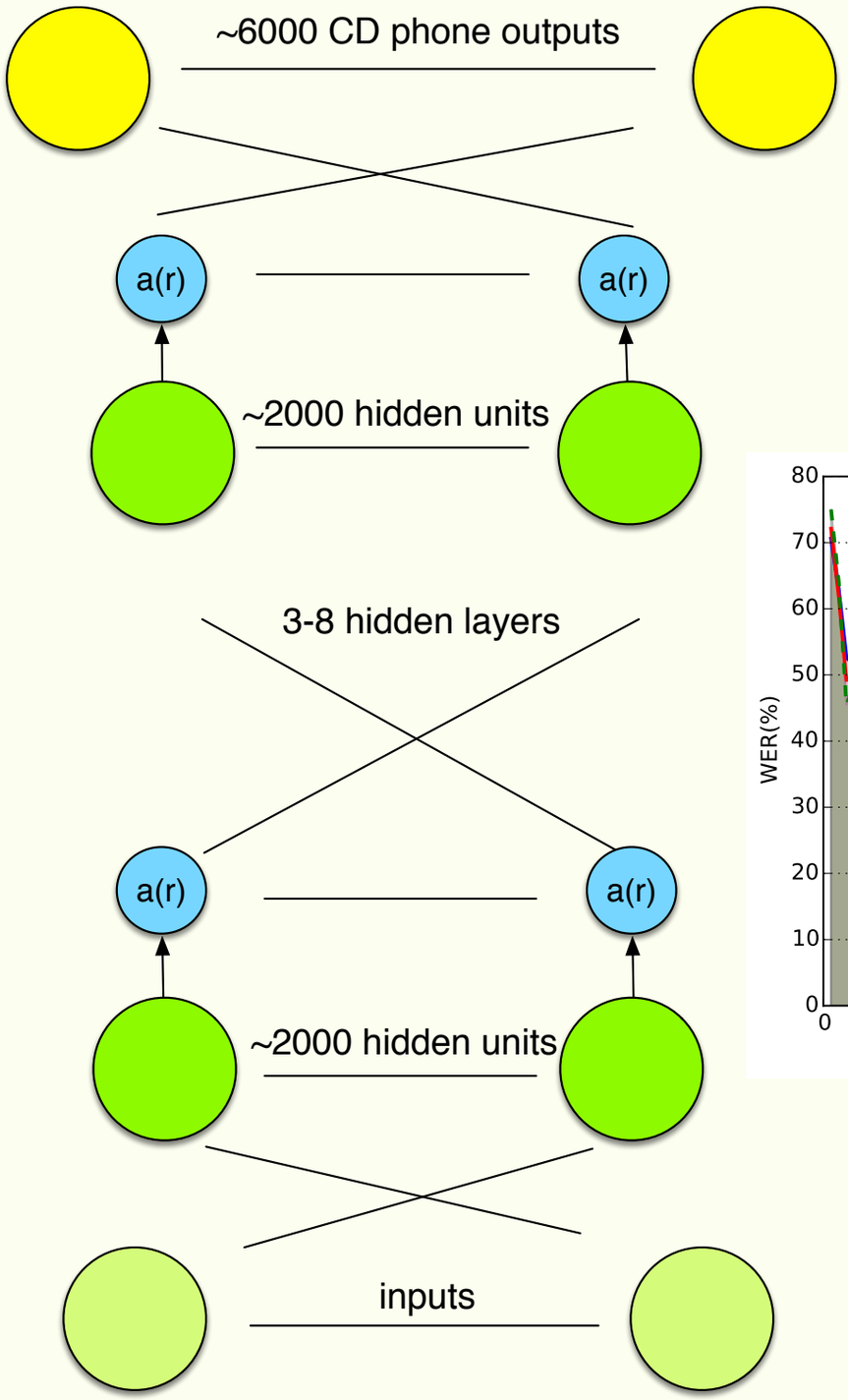


8% relative reduction in WER on MGB Challenge, compared with speaker adapted DNN

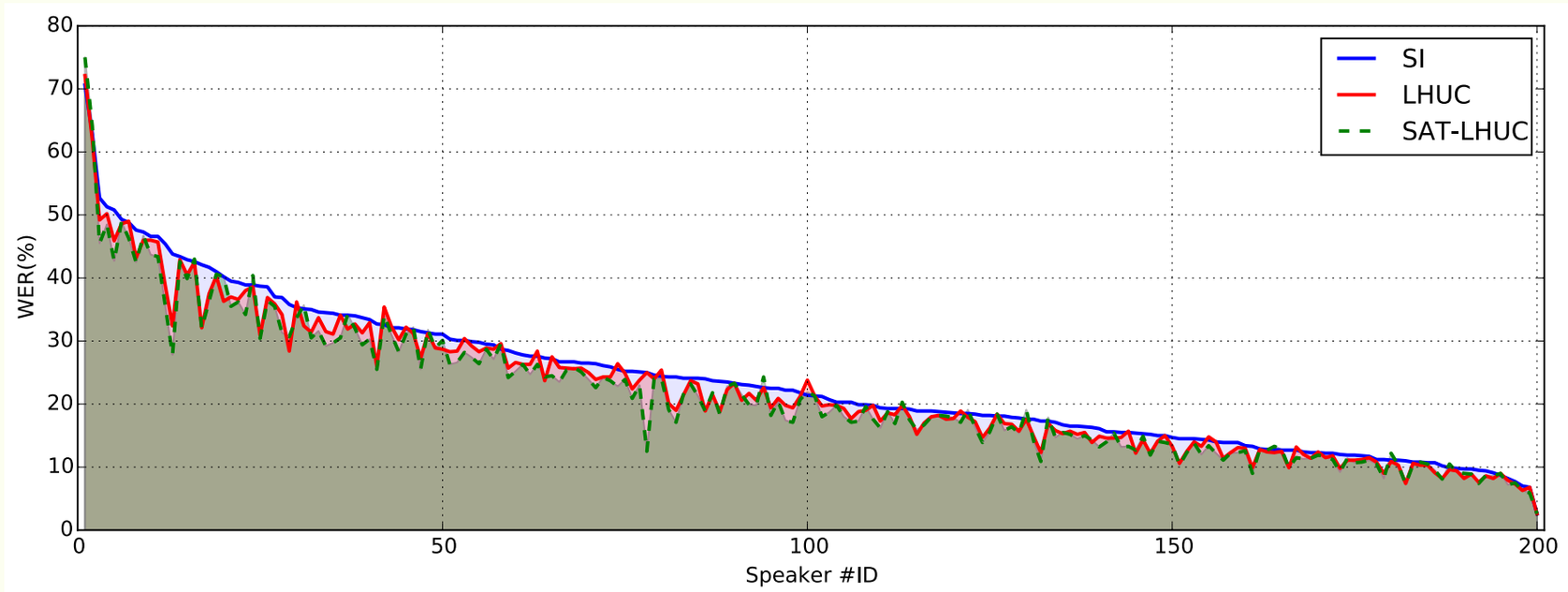
LHUC
Learning Hidden Unit Contributions

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



Key idea: add a learnable *speaker-dependent amplitude* to each hidden unit



Factorised adaptation

Natural Speech Technology

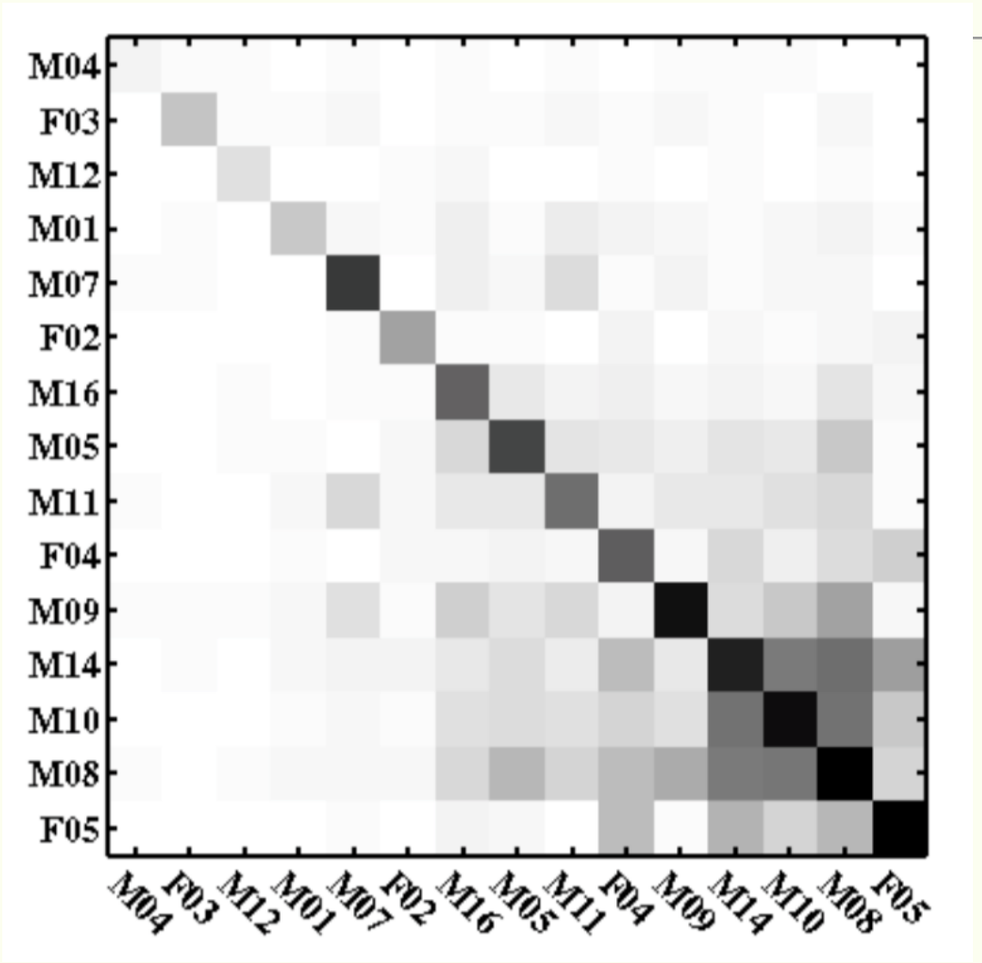
Edinburgh – Cambridge – Sheffield

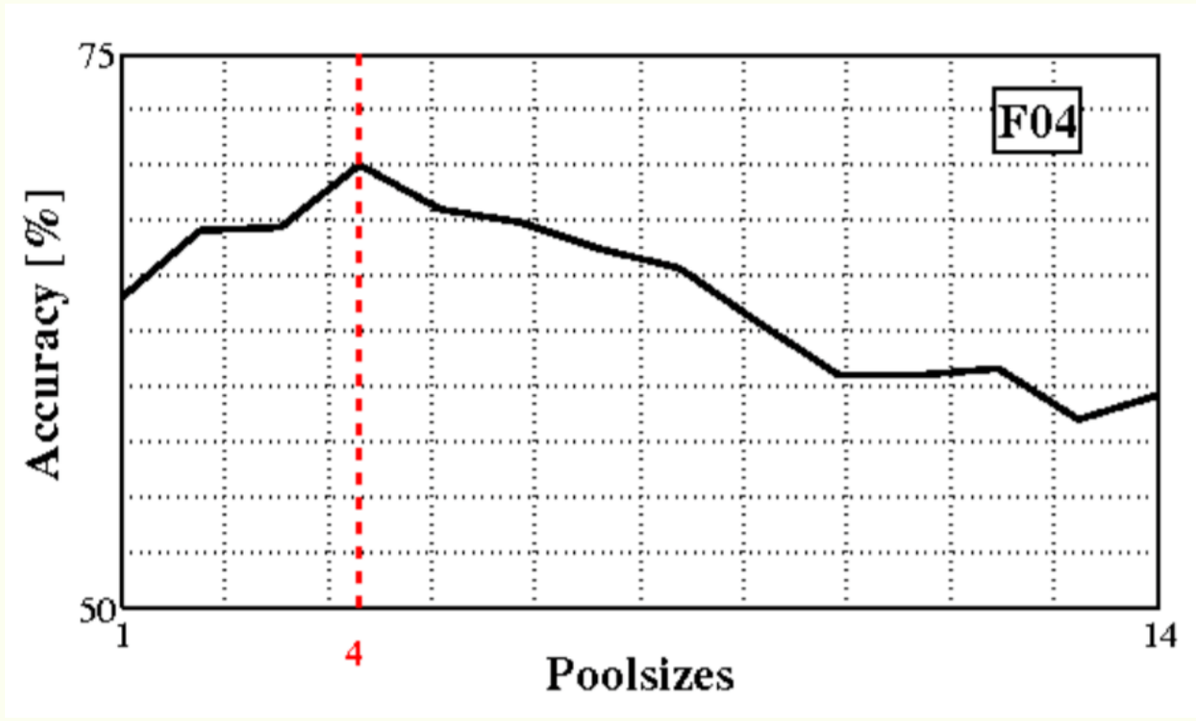
- Learn to separately adapt for environment/noise and for speaker
- Factorised i-vectors
  - Extract two sets of i-vectors – speaker information, acoustic environment information
  - Orthogonal factor representations allow adaptation to account for wide range of speaker/environment conditions
- Factorised LHUC
  - combine LHUC transforms for speaker and environment

Adaptation by speaker selection for dysarthric speech

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



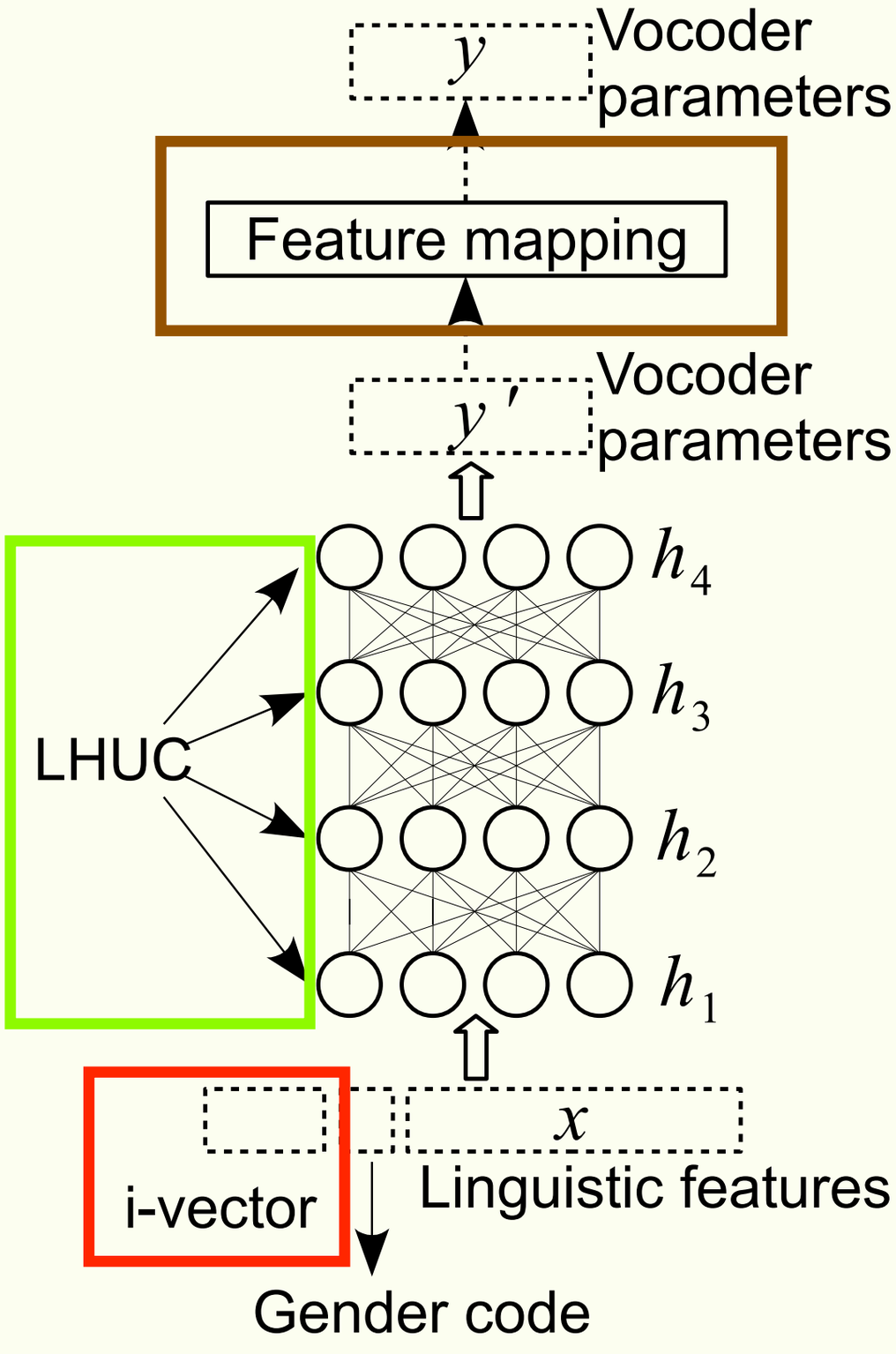


- Dysarthric speech is highly talker dependent
- UA-Speech: SD 45% WER, SI+MAP 49% WER
- Select SI speaker pool based on WER
- Pooled SI model + MAP – 40% WER

Adaptation in DNN speech synthesis

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Multi-task learning for ASR and TTS

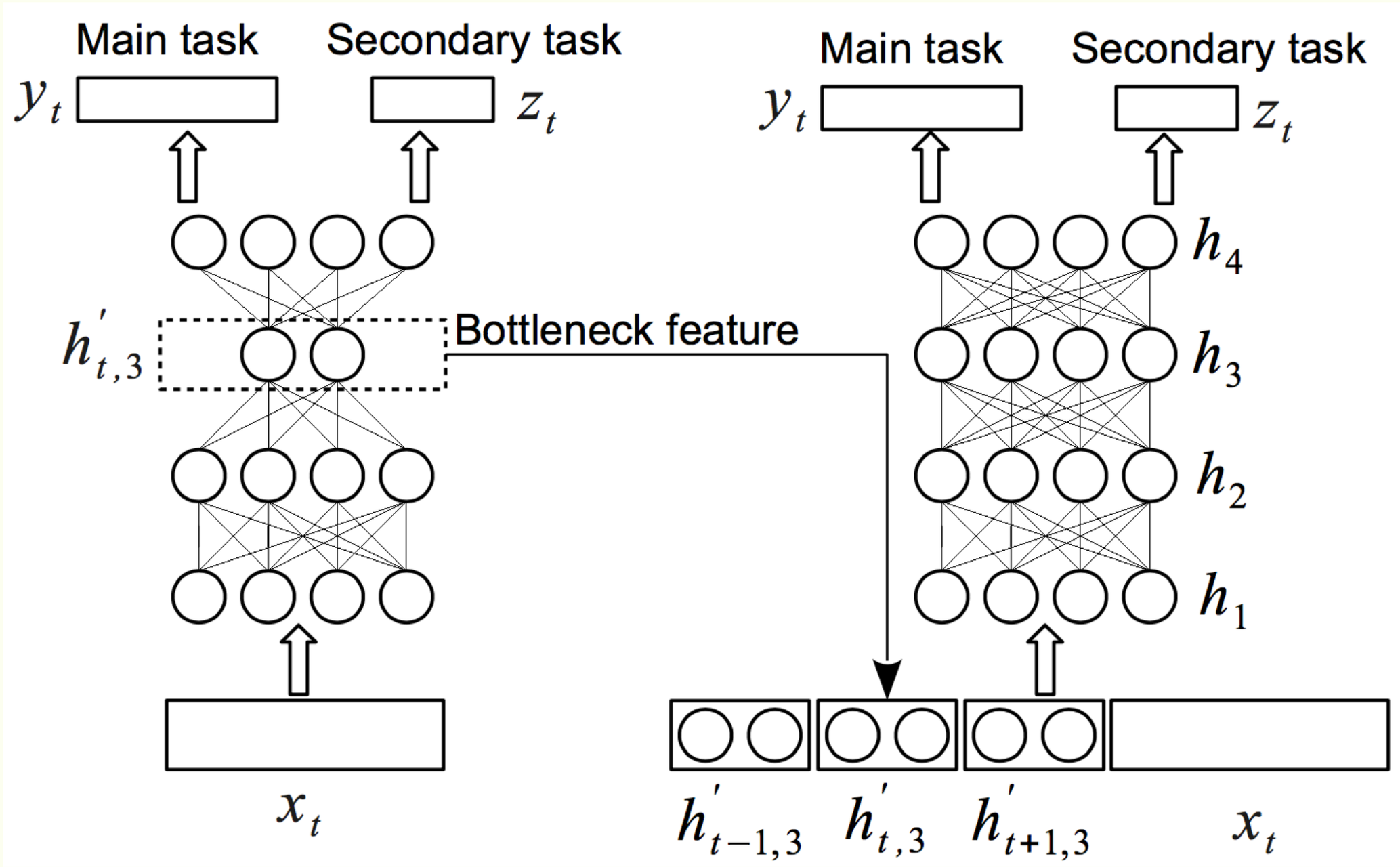
Multi-task DNNs in speech synthesis

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Main task: vocoder parameters

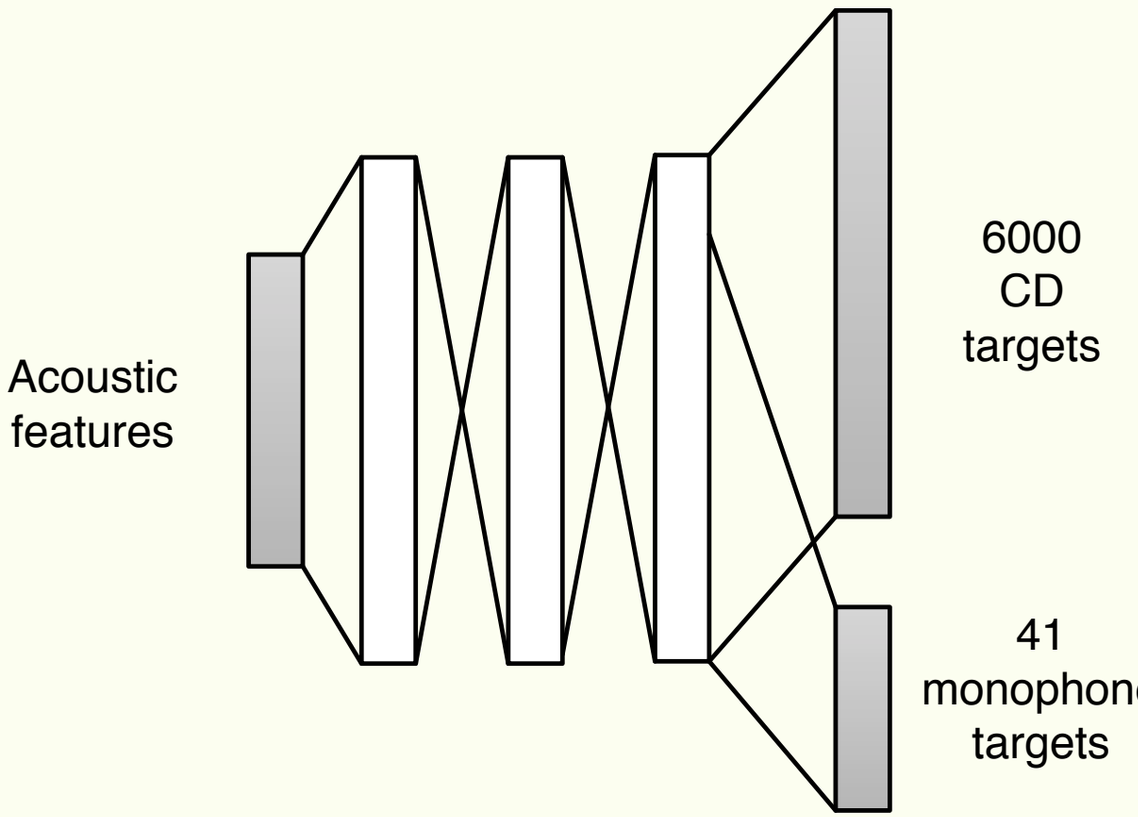
Secondary task: Glimpse-based perceptual measure (STEP)

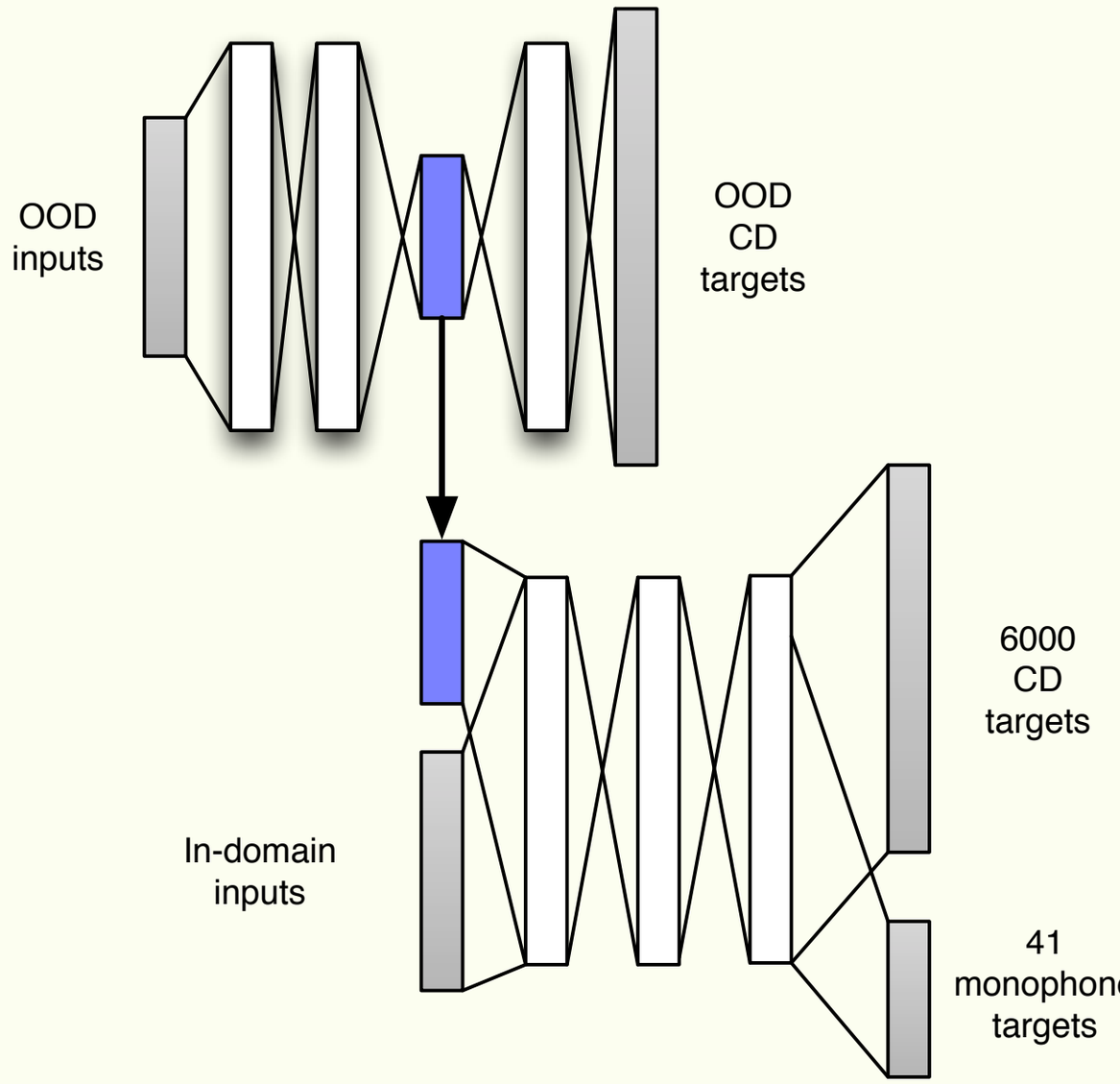


Multi-task learning for ASR

Natural Speech Technology

Edinburgh – Cambridge – Sheffield





3–5% WER relative reduction (TED)

Open source software

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

An Overview of HTK V3.5

htk3

Phil Woodland
& Cambridge HTK team
pcw@eng.cam.ac.uk

Kaldi

A state-of-the-art automatic speech recognition toolkit

http://kaldi-asr.org

HTS

CUED-RNNLM Toolkit

Merlin DNN TTS Toolkit

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Exemplar Applications

Voice banking and personalised TTS

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Multi-domain ASR

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

staffwww.dcs.shef.ac.uk

Hubs Mobile Edinburgh News Projects Journals Conferences Software Apple Personal Travel Bikes Video Capture

201 | Ho | News, sport... | asru-2015-s... | Multicore Mi... | ASRU 2015 | 0000675.pdf | arxiv.org/pdf... | scargill.inf.e... | MLP webpage | Youtube tran...

Transcription of Youtube clips using Automatic Speech Recognition

MH370 - Experts investigate mystery plane wreckage - BBC News	Inside Out Trailer 2 UK - Official Disney Pixar _ HD	Asian Stocks Fall To Three-Year Low
		
Completed	Completed	Completed
The History of Pork Pies in Britain - The Great British Bake Off	007 Spectre Official Trailer #2 (2015) Daniel Craig James Bond Movie HD	Branson - Volkswagen scandal is a wake-up call
		
Completed	Completed	Completed

Browsing Oral Histories

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Browsing Oral History Upload

doncaster Submit

user494912477
JD199078608

SOUNDCLOUD
Share

7:13
10:29
12

"and he was taking some dispatches to doncaster"

Details Summary Automatic Transcription Search

doncaster

Submit

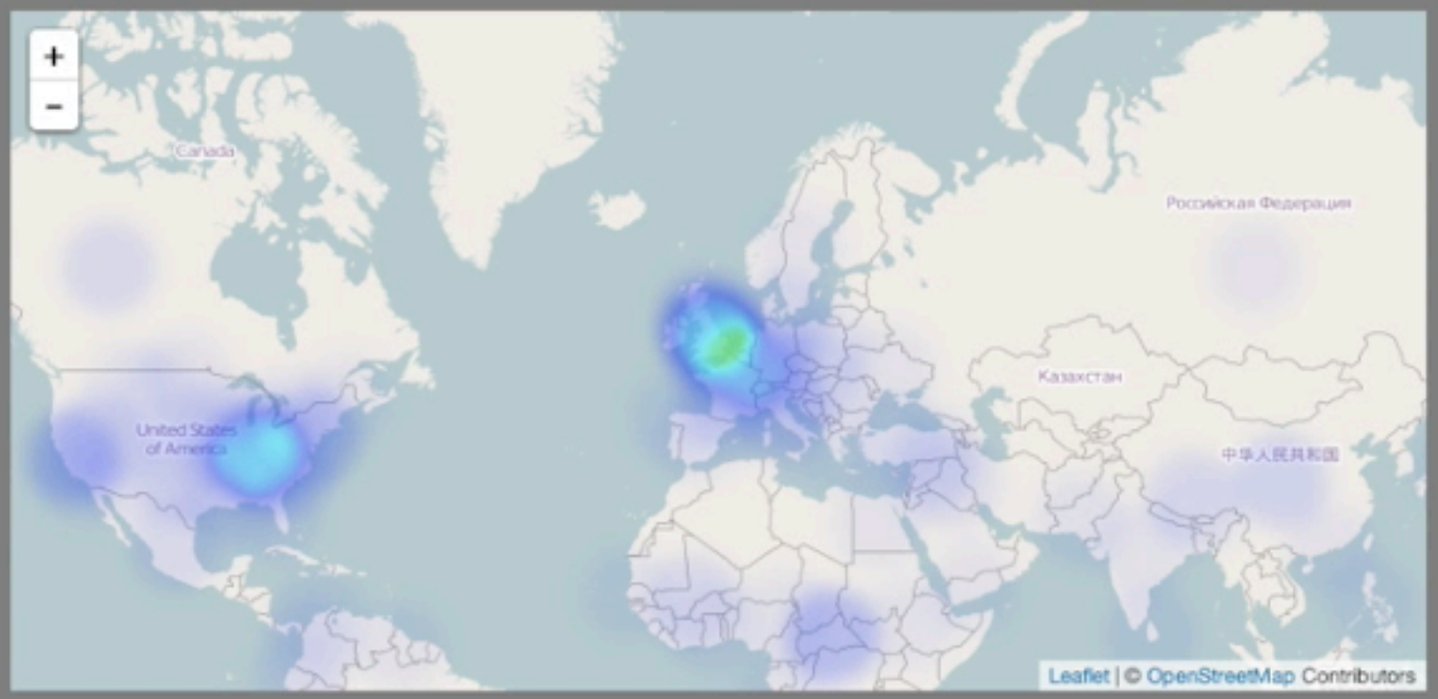
00:07:2	"travelled to doncaster"
00:07:12	"and he was taking some dispatches to doncaster"
00:01:20	"we didn't go back to school in doncaster for"
00:13:18	"and she had been a teacher at doncaster highschool actually"

GlobalVox

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

GLOBALVOX PRO



anderson anderson cooper ann arbour anthony damasio arnold arthur clarke **babbage** britney spears chandler charles charles darwin colin davis connor david david mackay frances arnold francisco los frank cardinal gardner john gardner george burns greater hillary clinton humphrey davy

hundreds jamie **jesus christ** john joshua david karl germain king it lou

gehrig malcolm gladwell michael phelps morgan mozart **nathaniel** pataki paul ekman

pekka solomon professor katie walter professor van allen robert steve downey sophie **steve** steve jobs

steve lopez storms sudoku tom ferguson **walt disney**

And for those you weren't there the l a c is the largest scientific experiment ever attempted twenty seven kilometers.....

Contacted peacefully in nineteen fifty eight in nineteen fifty seven five missionaries attempted contact made a critical mistake they drop.....

There is one corner by the way that i'm not going to tell anybody about where you actually where the.....

Now for almost twenty years when we sequence the human genome was going from the analogue world the biology into.....

But this understates the seriousness of this particular problem because it doesn't show the thickness of the ice the arctic.....

That'S why we have a refrigerators air conditioning can make a modern materials and do so many things so we're.....

I'D i out see also hurt useless and we and to them made healthy strong capable i was reading this.....

Lay people thinking about their own happiness and the price of scholars thinking about happiness because it turns out with.....

One particular cause of death say accidents right away i see there's a different pattern emerges this is because in.....

And now the envelope push back and i was told by ah the folks at my company that we were.....

Challenge Coordination

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

MGB Challenge, ASRU-2015 MGB Challenge-2, SLT-2016

Blizzard text-to-speech synthesis Challenge, 2011–2016

Spoc 015

MGB

CHALLENGE

Automatic Speaker Verification Spoofing and Countermeasures Challenge

Blizzard Challenge

Awards

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

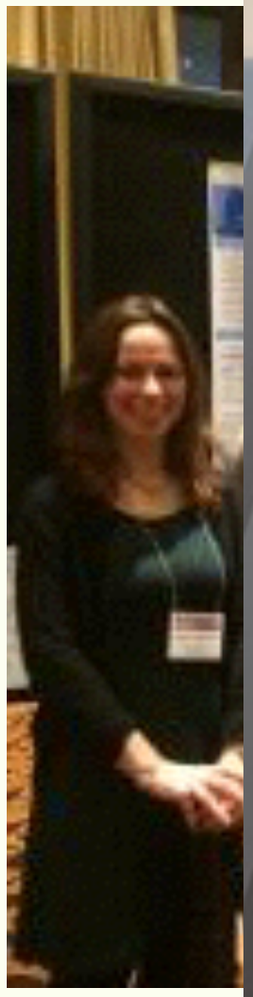
Best poster at Speech Synthesis Summer School 2015

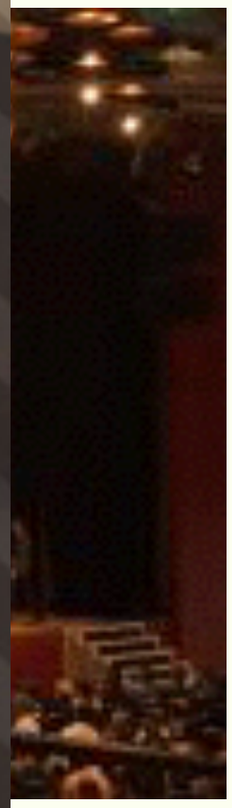
BBC NewsLabs "Surprise Us" Award 2016

BBC NewsLabs "Wimbledon Interspeech" Award 2015-2015

IBM Research Spoken Language Processing Award ICASSP-2014

Most accurate systems at Mgb Challenge ASRU 2015

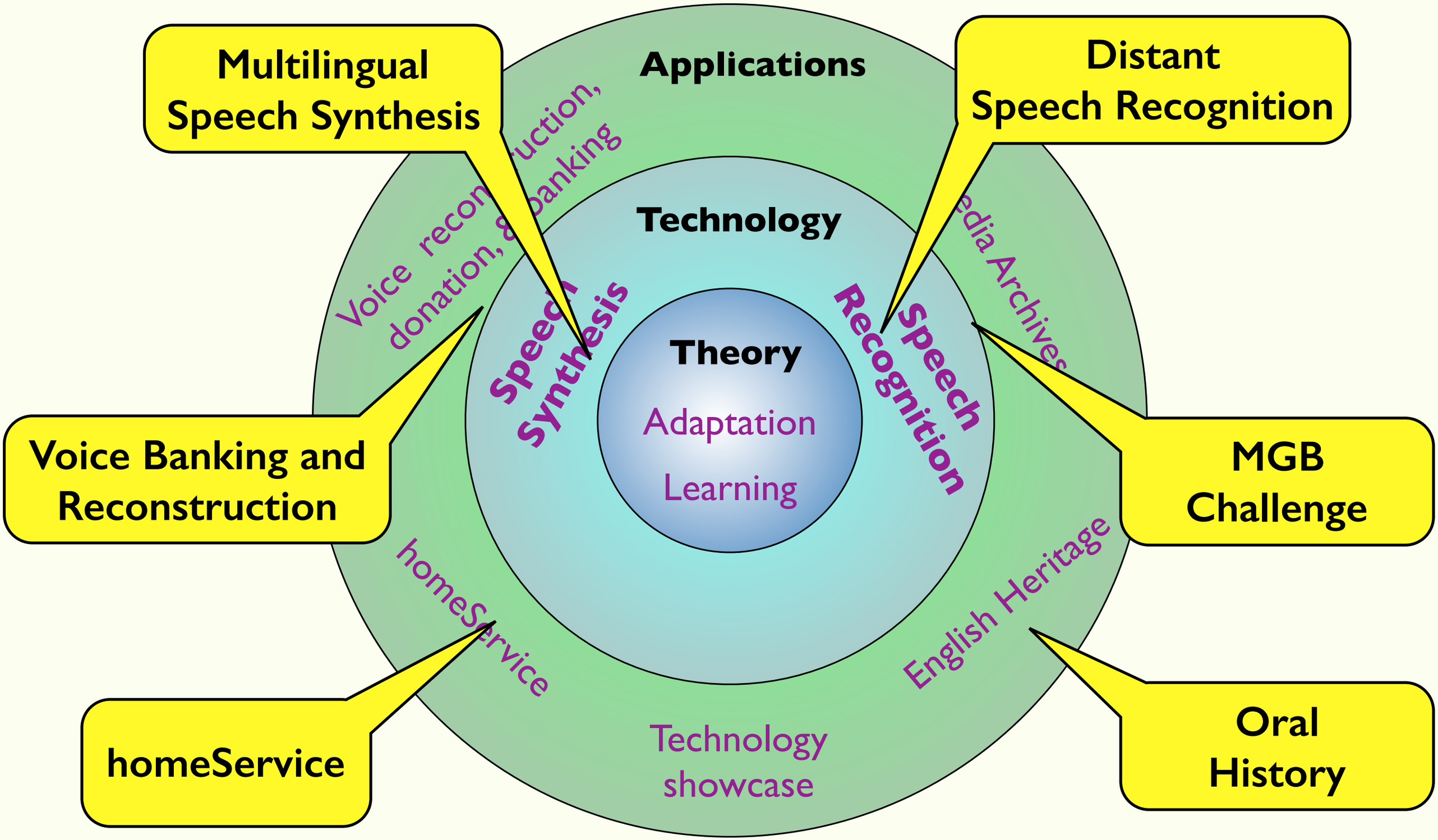




Todays talks

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



Demonstration systems

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

MGB Challenge systems
recognition, segmentation, alignment

TTS from "found data"

homeService

NewsHACK – TTS to generate multilingual broadcast content

DNN speech synthesis systems

Voice banking and reconstruction

Distant speech recognition
Sheffield Wargames Corpus

webASR

Transcribing and navigating oral history collections

NewsHACK – GlobalVox
Multilingual media monitor

NST people

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Today's agenda

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- 11:15 – 12:50 Intro + 2 talks
  - Multilingual speech synthesis (Simon King, Oliver Watts)
  - homeService (Phil Green)
- 12:50 – 13:50 Lunch
- 13:50 – 15:30 Talks
  - MGB Challenge (Phil Woodland)
  - Voice Banking (Christophe Veaux)
  - Distant Speech Recognition (Thomas Hain)
  - Oral History (Phil Green)
- 15:30 – 15:50 Break
- 15:50 – 17:30 Demos
- 17:30 – 18:00 Wrap-up and feedback
- 18:00 – 19:30 Drinks reception (Level 4 / roof)
- 20:00 – Dinner at Howies

EPSRC
Engineering and Physical Sciences Research Council

Tomorrow's agenda

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- 09:00 – 09:15 Arrival and coffee
- 09:15 – 11:00 Talks
  - Deep Learning for Speech Processing (Mark Gales)
  - System highlights from MGB (Phil Woodland)
  - Technical highlights in speech synthesis (Simon King)
- 11:00 – 11:25 Break
- 11:25 – 12:50 Posters
- 12:50 – 14:20 Lunch / *Advisory Board Meeting (rm 5.42)*
- 14:20 - 16:00 Talks
  - Stimulated training and visualising NNs (Chunyang Wu)
  - Disfluency in speech synthesis (Marcus Tomalin)
  - Structured media data using latent modeling (Mortaza Doulaty)
  - Neural segmental CRFs for sequence modeling (Liang Lu)
- 16:00 – 16:15 Wrap-up