THE UNIVERSITY of EDINBURGH

Natural Speech Technology

Steve Renals
University of Edinburgh

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

EPSRC
Engineering and Physical Sciences Research Council

THE UNIVERSITY of EDINBURGH

Natural Speech Technology

5-year UK programme in
core speech technology research
2011–2016

Focus on
Speech Recognition
Speech Synthesis
Learning & Adaptation

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

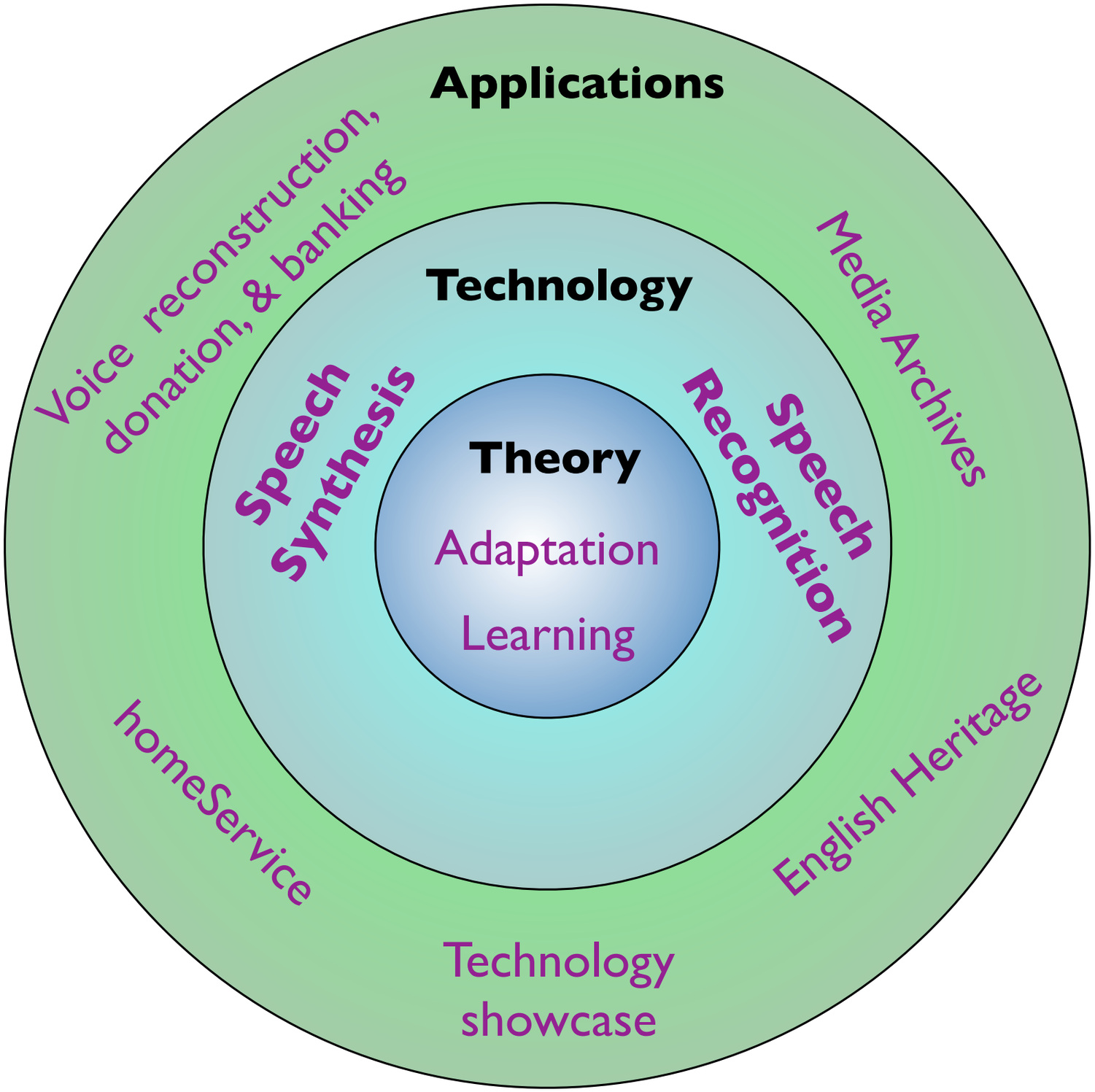
THE UNIVERSITY of EDINBURGH

Motivations

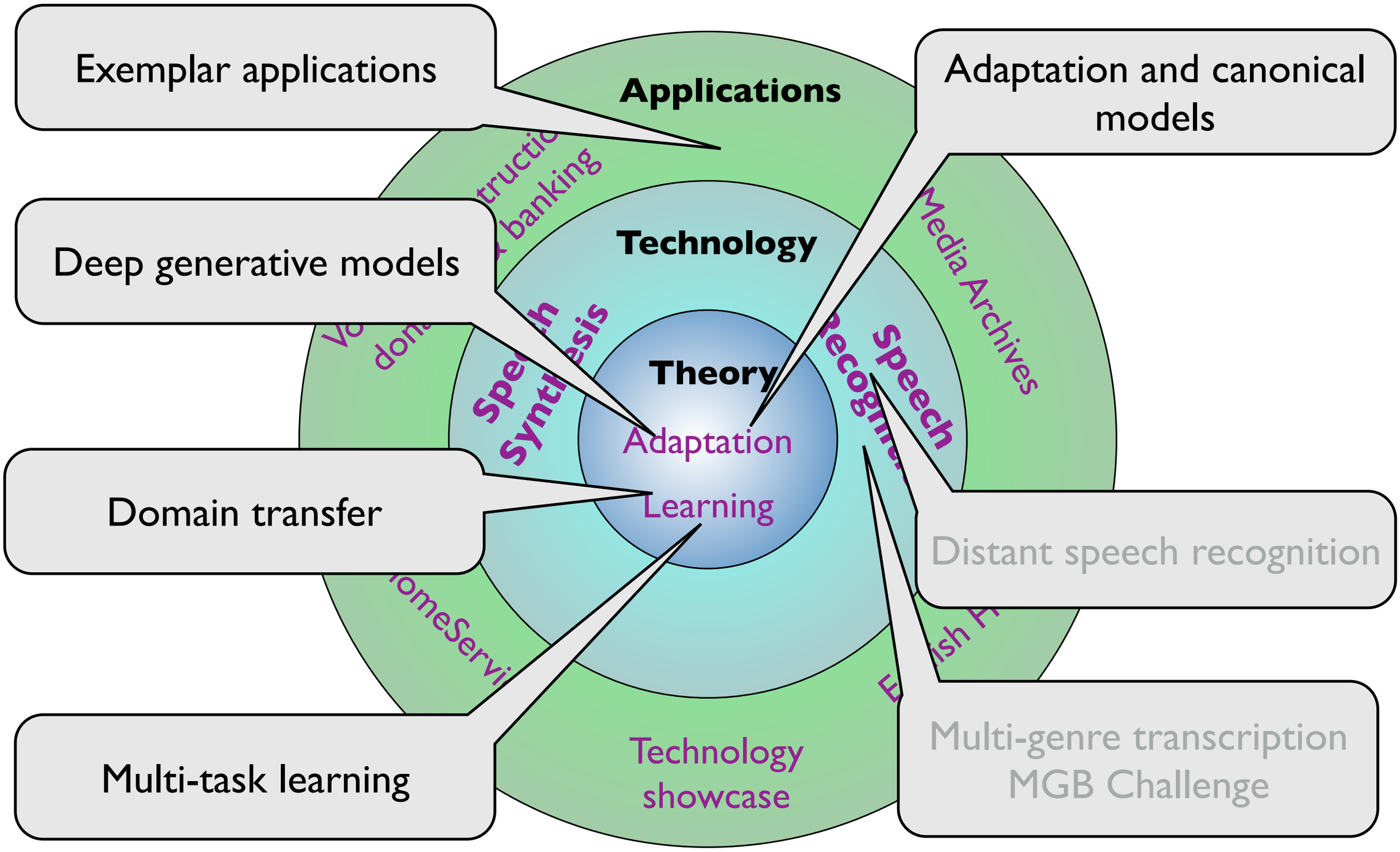
- **Weakly-factored models**
  - Factor the underlying causes of observed variability in speech
- **Domain fragility**
  - Rapid transfer to new domains, with minimal supervision
- **Synthesis and recognition developed independently**
- **Lack of reaction to the environment or context**
  - Respond and adapt to changes in the acoustic or linguistic environment
- **Relatively little speech knowledge incorporated**
- **Cannot rely on gold standard transcription**
  - Work somewhere on the supervised-unsupervised spectrum

THE UNIVERSITY of EDINBURGH

Natural Speech Technology



Natural Speech Technology



THE UNIVERSITY of EDINBURGH

Adaptation

multiple speakers
acoustic environment
different channels

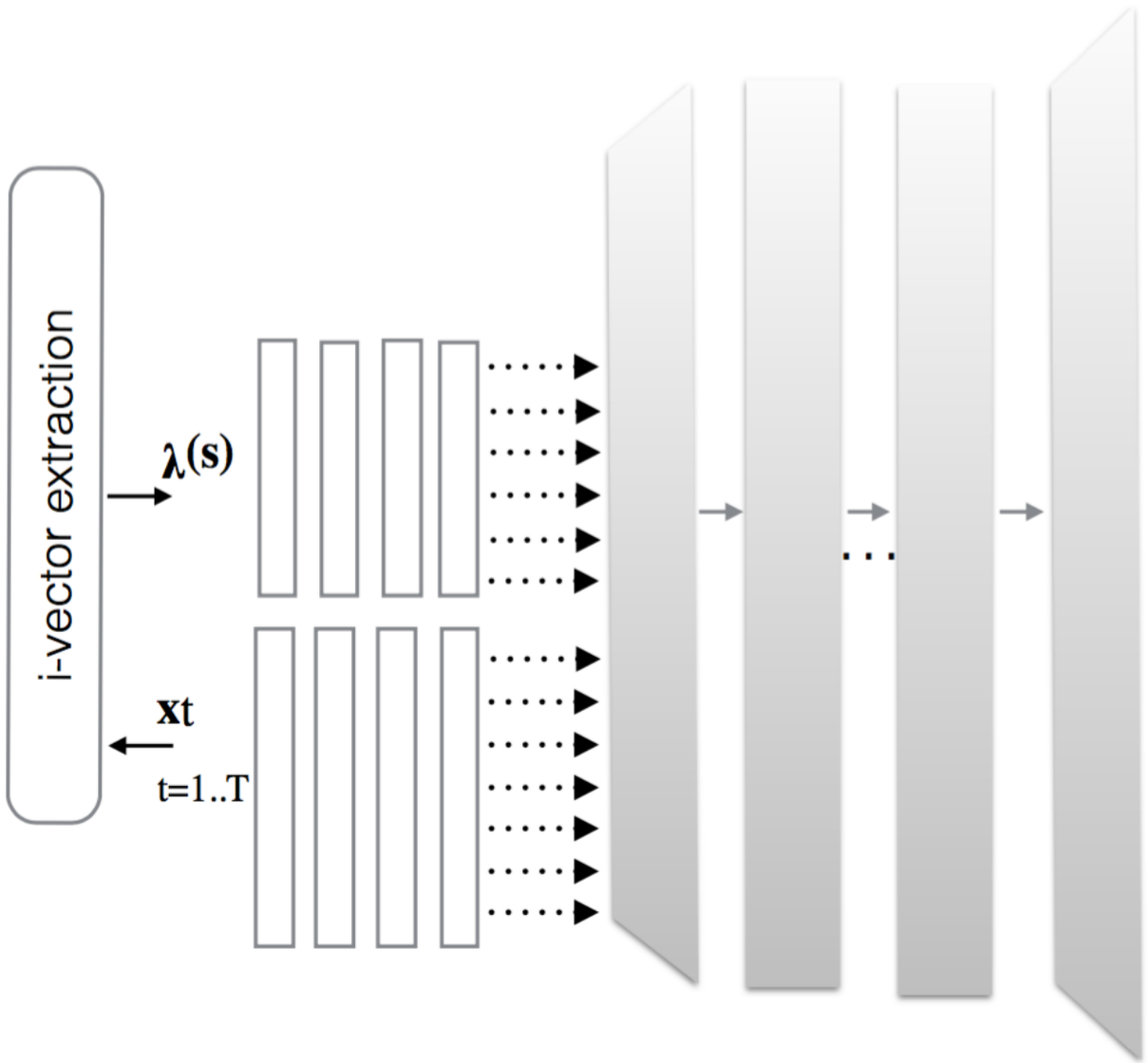
THE UNIVERSITY of EDINBURGH

Adapting NN acoustic models

- Neural network adaptation is challenging
  - models with large numbers of parameters potentially need a lot of adaptation data
  - relatively little structure in the weights
  - unsupervised adaptation is preferable to supervised
  - compact adaptation is preferable
  - joint optimisation of core acoustic model parameters and adaptation parameters
- Baseline
  - feature-space MLLR using a CD-GMM-HMM system to adapt input features for NN acoustic model

Auxiliary features

Append (and optimise) additional speaker-based features to the input



i-vectors:

- low-dimension speaker representation
- can be estimated from small amounts of data
- for ASR: Karafiat et al (ASRU-2011); Saon et al (ASRU-2013)

Factorised i-vectors

Karanasou et al, Interspeech-2014

- Extract two sets of i-vectors
  - speaker information
  - acoustic environment information
- Estimate the i-vectors as weights for a cluster adaptive training GMM system
- Orthogonal factor representations allow adaptation to account for wide range of speaker/environment conditions
- On WSJ with added noise, factorised i-vectors result in 5–10% relative reduction in WER

THE UNIVERSITY of EDINBURGH

i-Vector priors

Karanasou et al, Interspeech-2015

- With limited data (1 utterance) use a prior to improve the robustness of the i-vector estimate
- Default: Gaussian prior – sensitive to amount of data/speaker, and to mismatches between training and test duration
- Count-smoothing prior interpolates between prior and observed statistics (cf MAP)
  - speaker-independent prior statistics (estimate over all speakers)
  - gender-dependent prior statistics (two clusters)
- YouTube data, WER improves ~1–3% relative without prior, 3–5% relative with SI prior

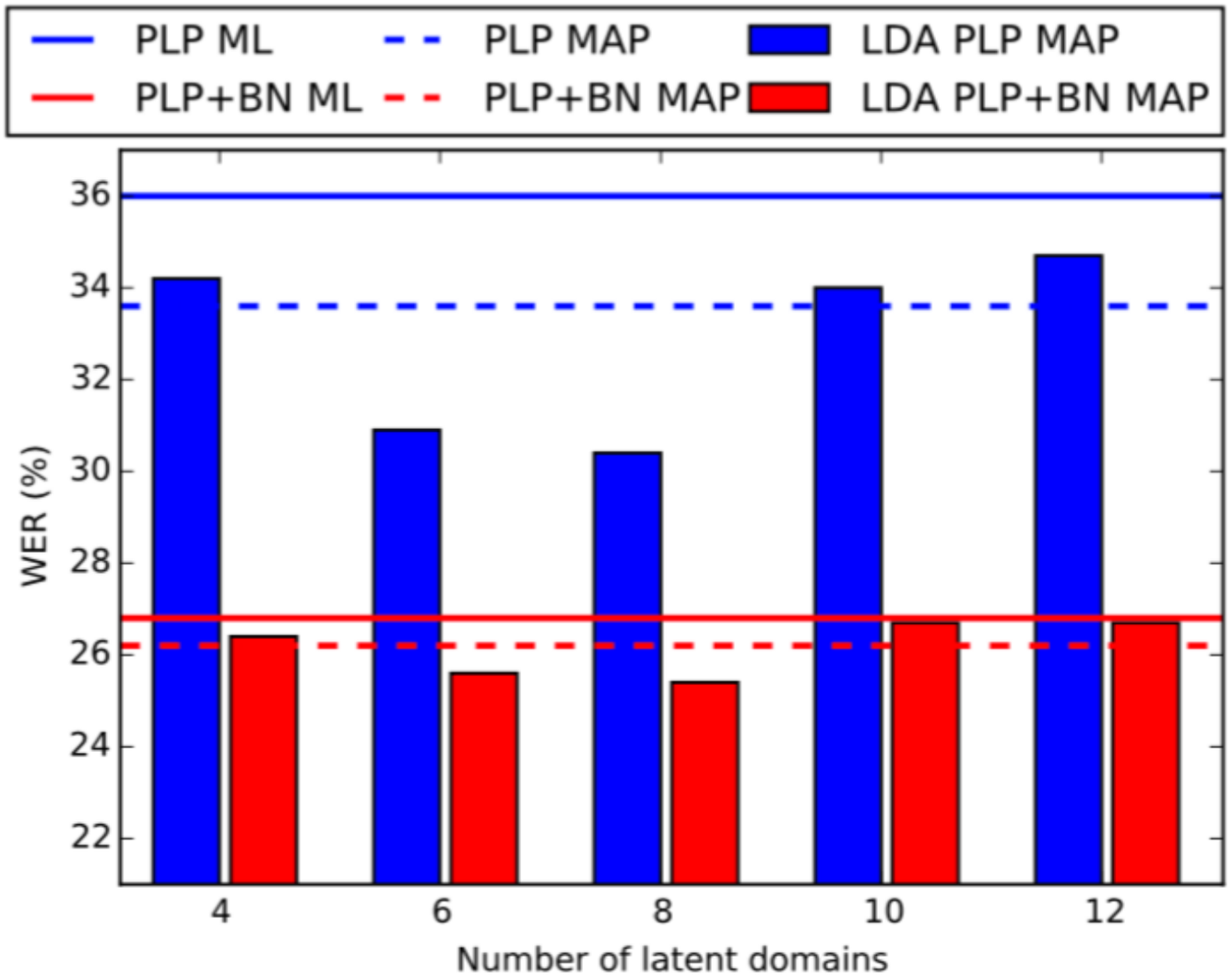
Unsupervised domain discovery

Doulaty et al, Interspeech-2015

Discovery of hidden acoustic domains using LDA

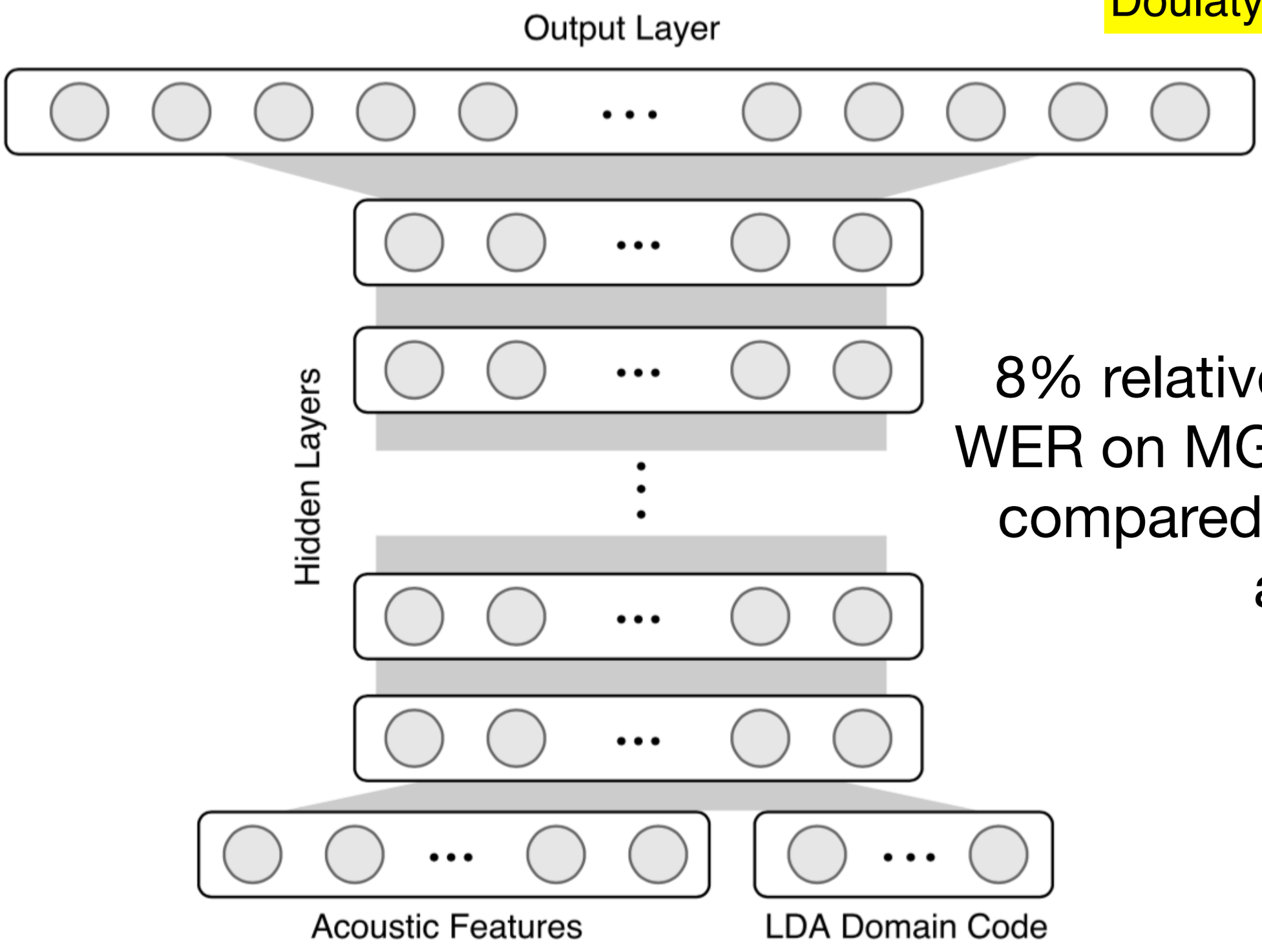
Experiments on highly diverse data

- radio
- television
- conversational telephone speech
- meetings
- read speech
- lectures



LDA-DNN

Doulaty et al, ASRU-2015



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

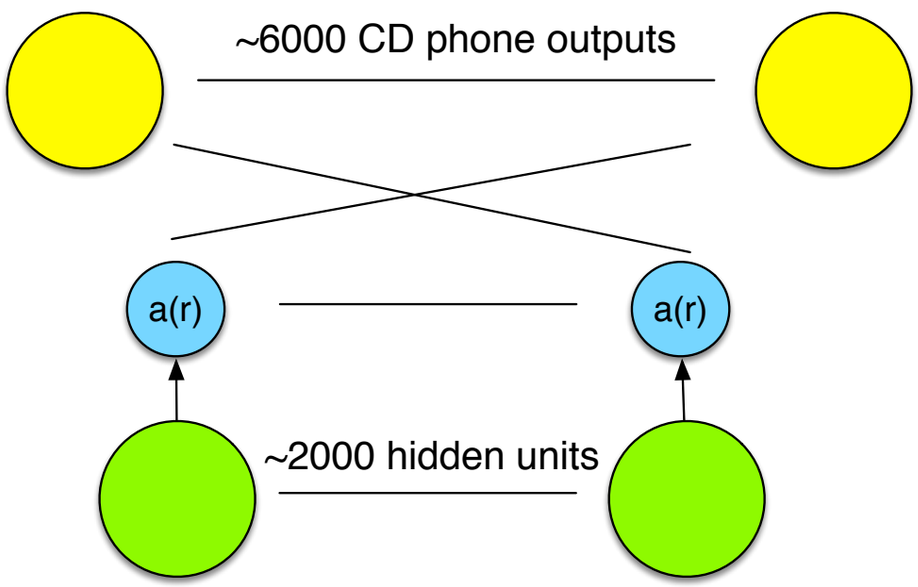
Model-based adaptation

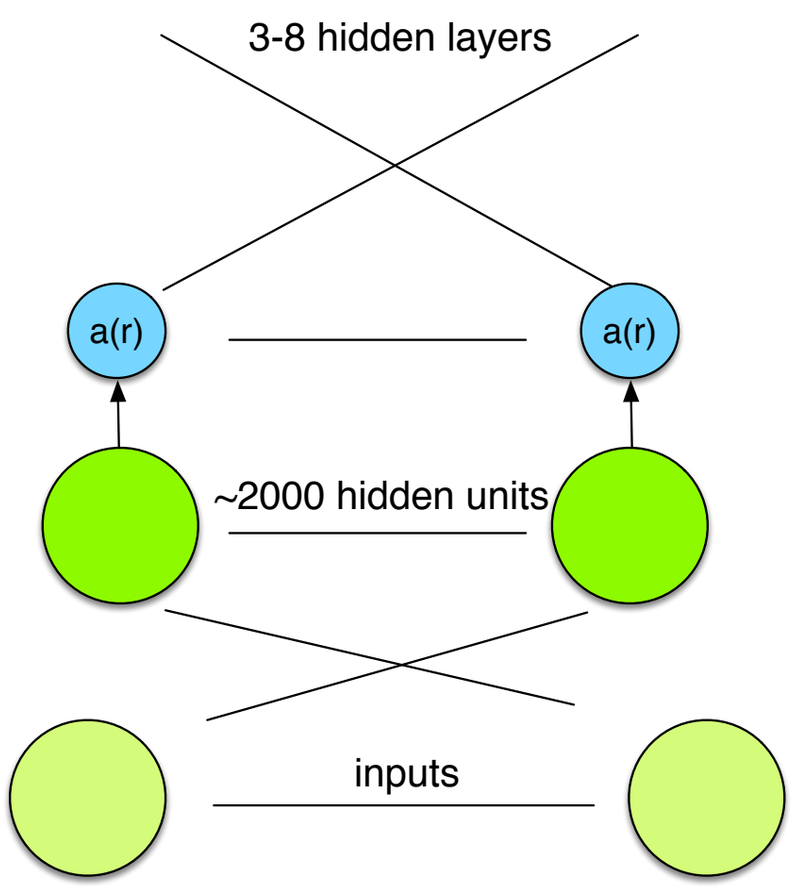
- **Speaker codes** (Bridle & Cox 1990; Abdel-Hamid & Jiang 2013) – model-based adaptation using auxiliary features
- Adaptation of different **weight subsets** (Liao, ICASSP-2013)
  - 5% relative decrease in WER when all 60M weights adapted
- Automatically adapt specific parameter subsets – **output biases** (Yao et al, SLT-2012), **slope and bias of hidden units** (Siniscalchi et al, TASLP-2013)
- Adaptation cost based on **KL divergence** between SI and SA output distributions (Yu et al, ICASSP-2013)
  - 3% relative decrease in WER on Switchboard
- Increase compactness by **SVD factorisation** of weight matrix (Xue et al, ICASSP-2014)

LHUC
Learning Hidden Unit Contributions

Swietojanski & Renals, SLT-2014

Zhang & Woodland, Interspeech-2015





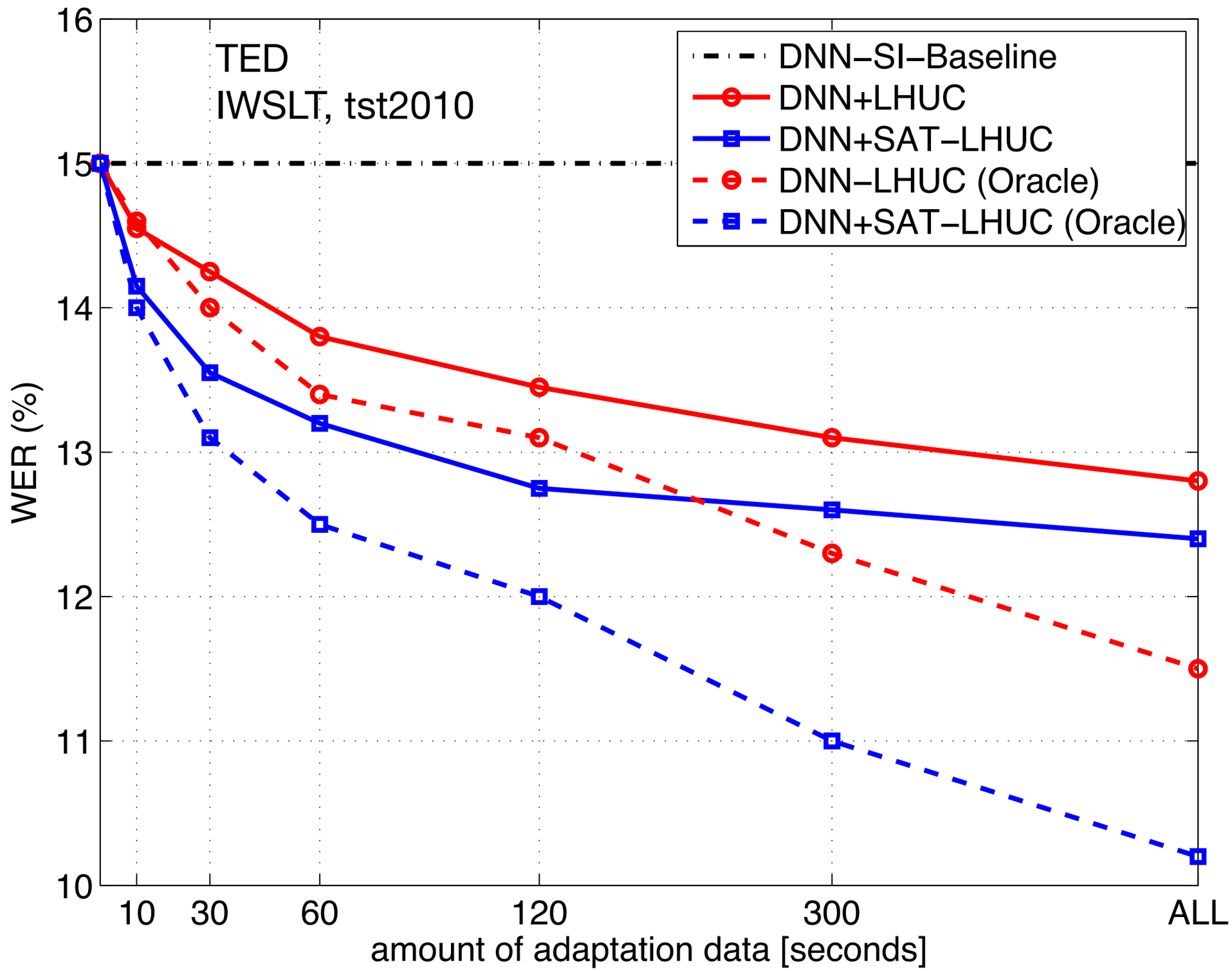
Key idea: add a learnable ***speaker-dependent amplitude*** to each hidden unit

$$\mathbf{h}_m^l = a(\mathbf{r}_m^l) \circ \sigma^l(\mathbf{W}^{l\top}\mathbf{h}_m^{l-1})$$

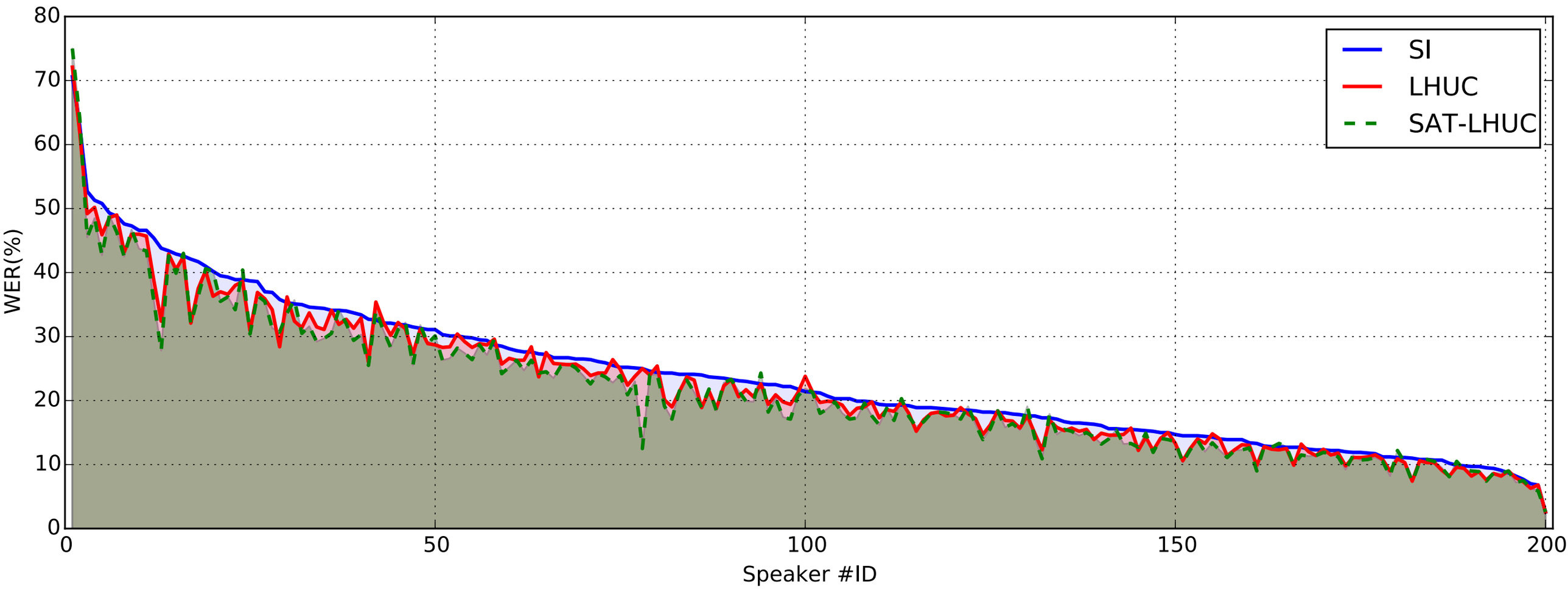
Speaker dependent parameter **r**

SI Model: set amplitudes to 1
SD Model: learn amplitudes from data, per speaker

LHUC – Adaptation data



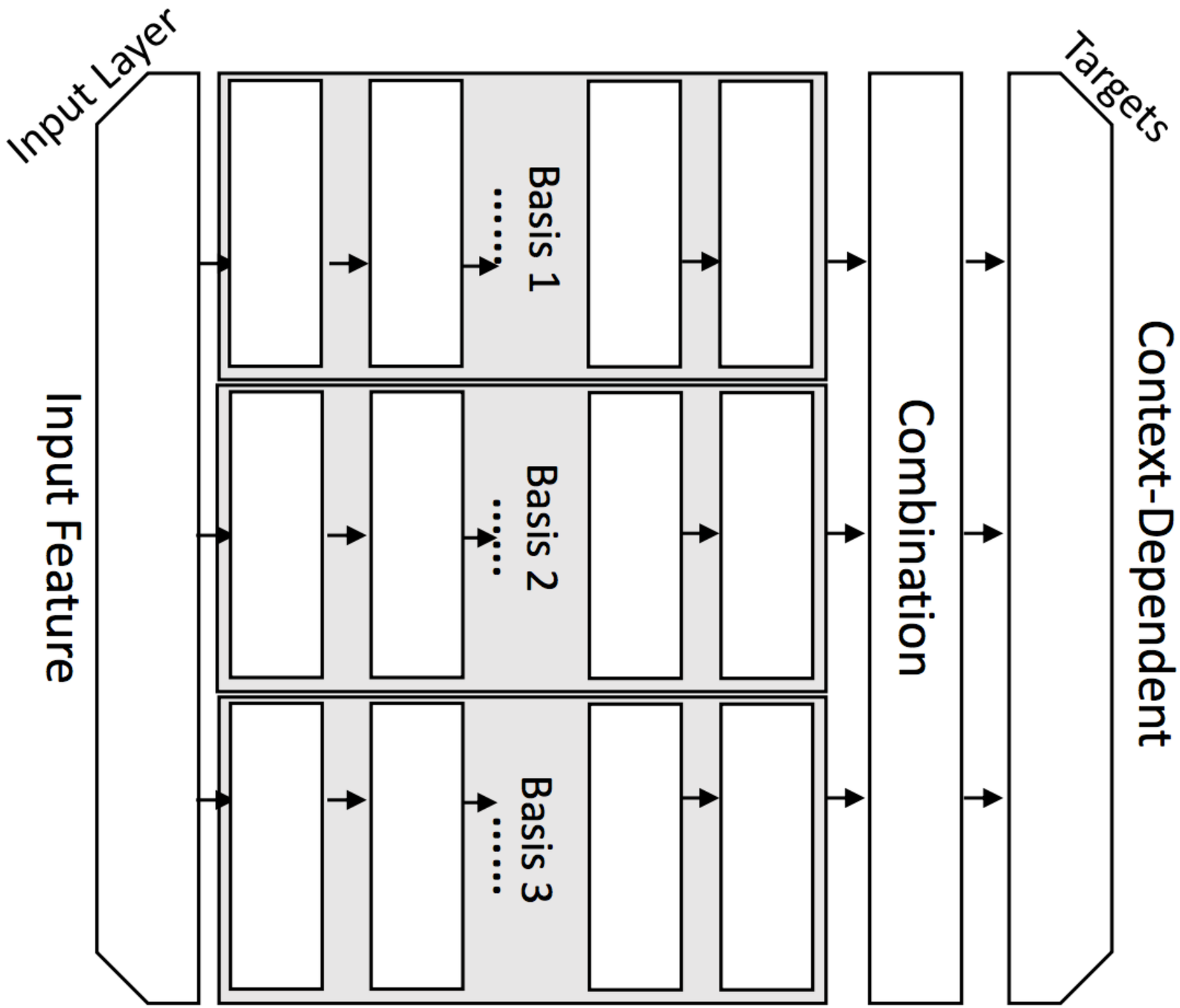
LHUC – Improvement per speaker



Combined results from TED, AMI, Switchboard

Multi-basis adaptive NN

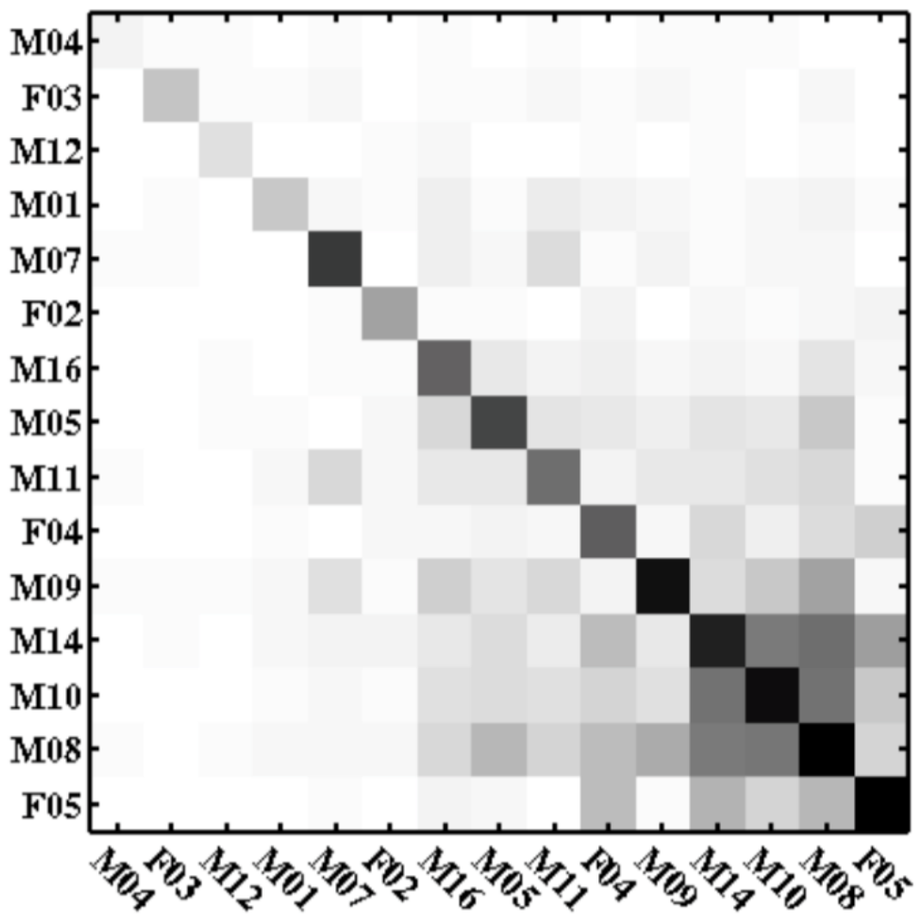
C Wu & Gales, Interspeech-2015

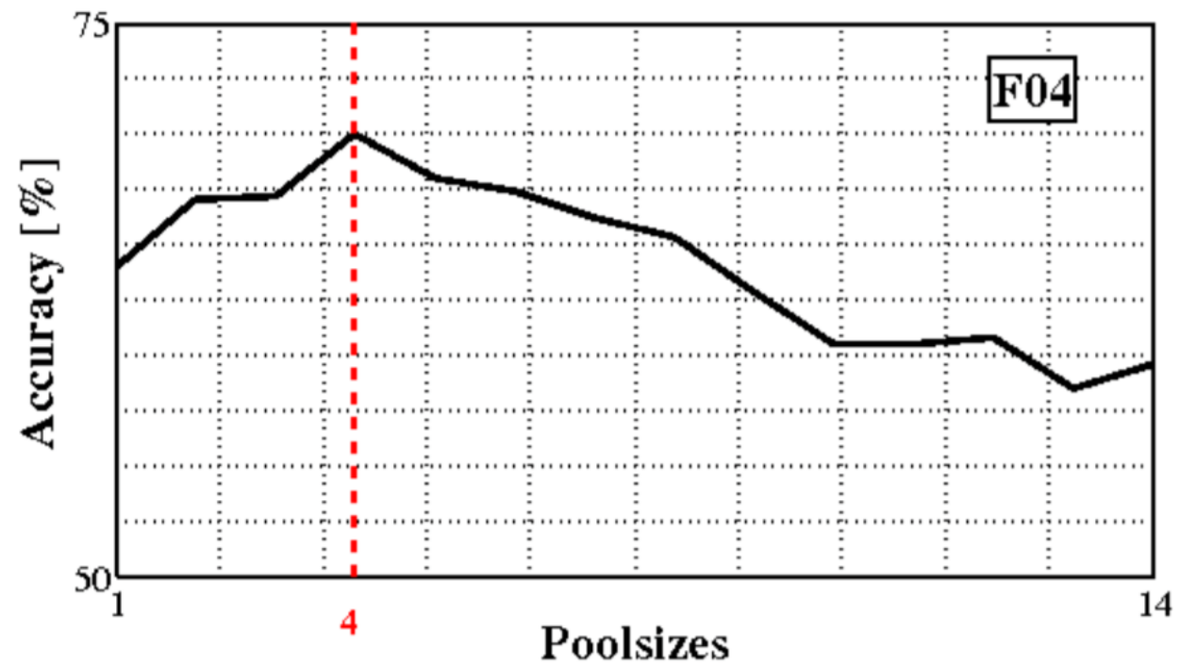


2–4% WER relative reduction (YouTube)

Adaptation by speaker selection for dysarthric speech

Christensen et al, SLT-2014



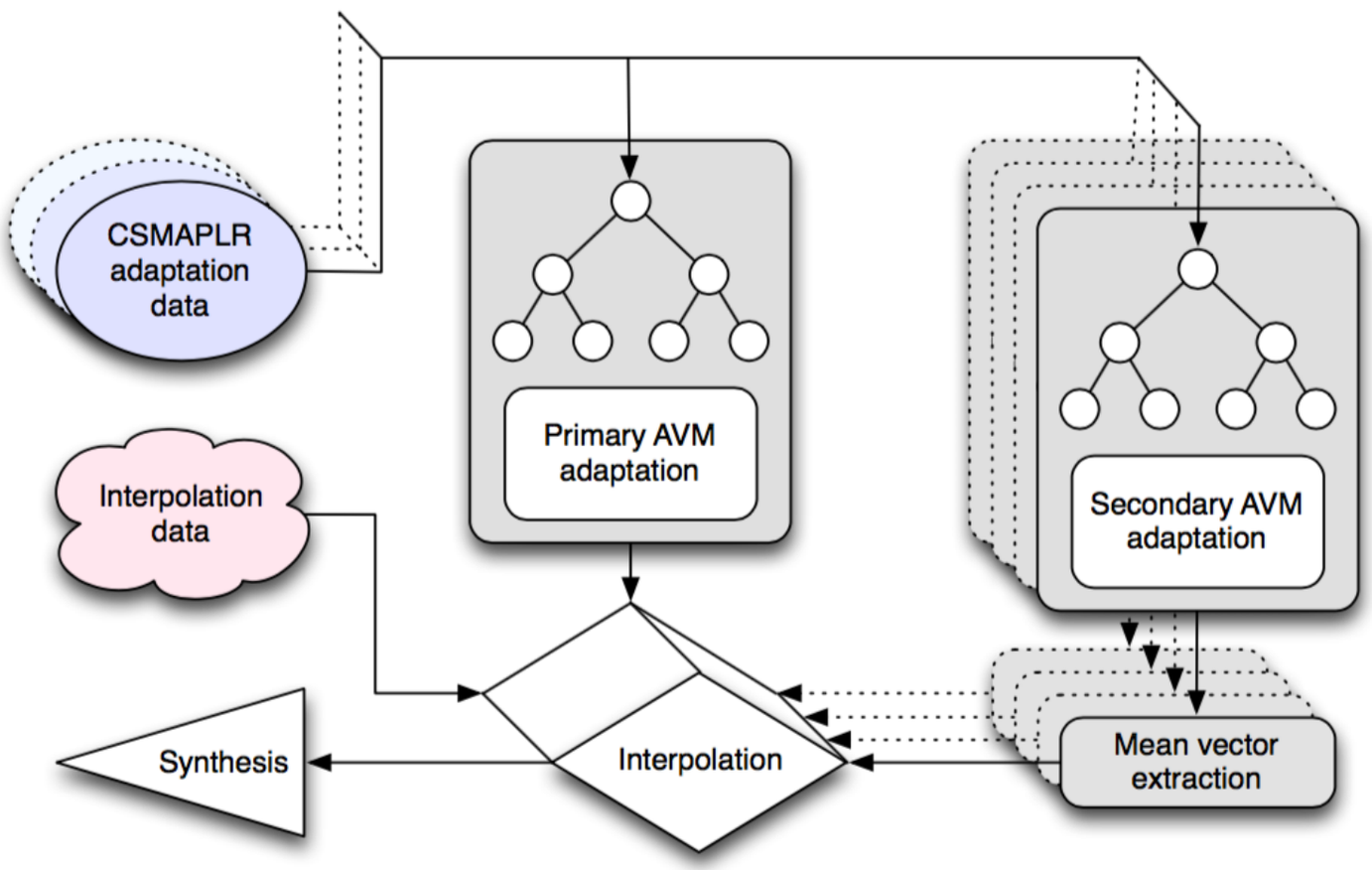


- Dysarthric speech is highly talker dependent
- UA-Speech: SD 45% WER, SI+MAP 49% WER
- Select SI speaker pool based on WER
- Pooled SI model + MAP – 40% WER

Multiple average voice model

Personalised speech synthesis for people with speech disorders

Lanchantin et al, Interspeech-2014

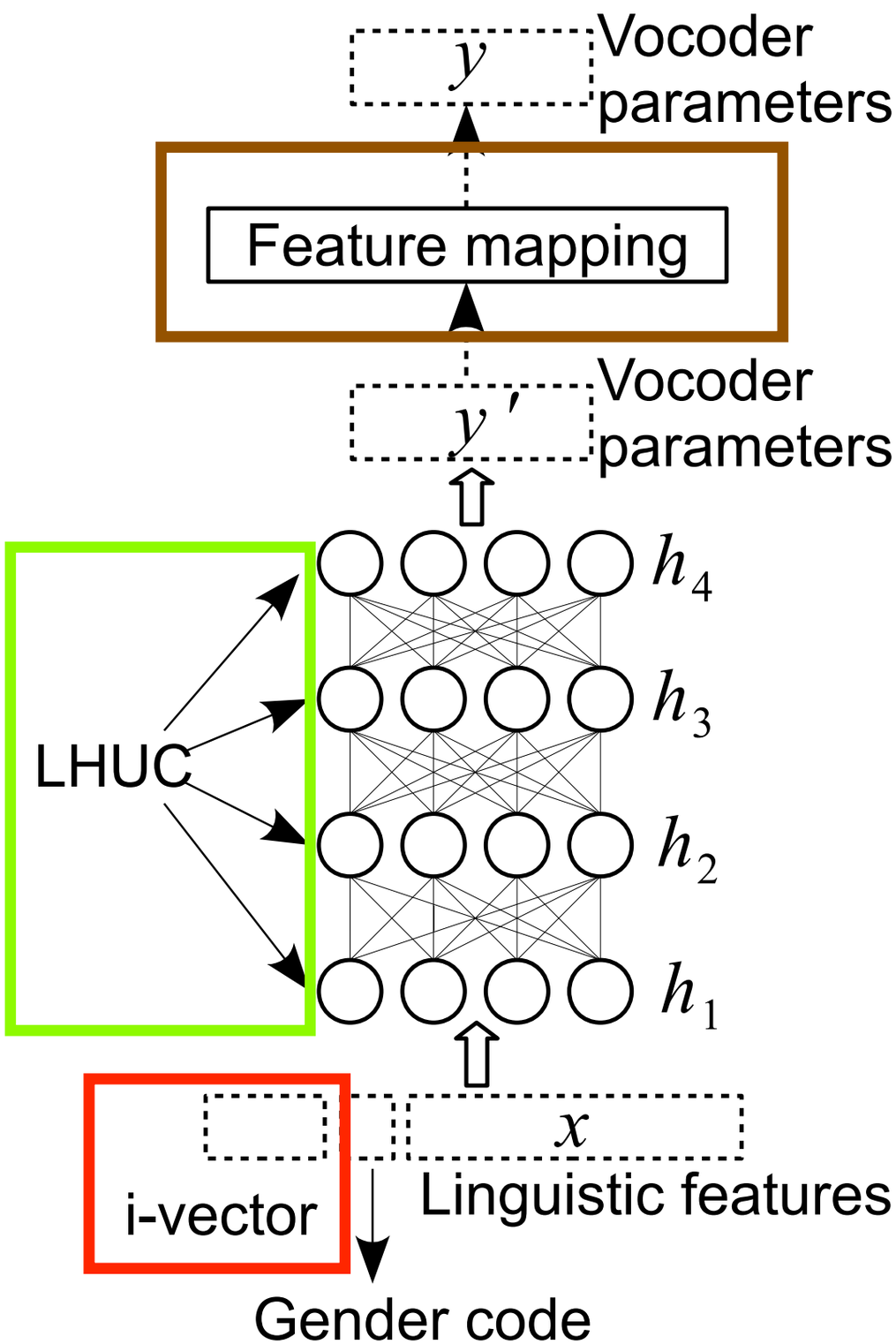


Combines cluster-adaptive training and average voice model

Adaptation by interpolating into a speaker eigenspace spanned by mean vectors of speaker-adapted AVMs

Improvements in intelligibility and naturalness over ‘tailored’ synthetic voice

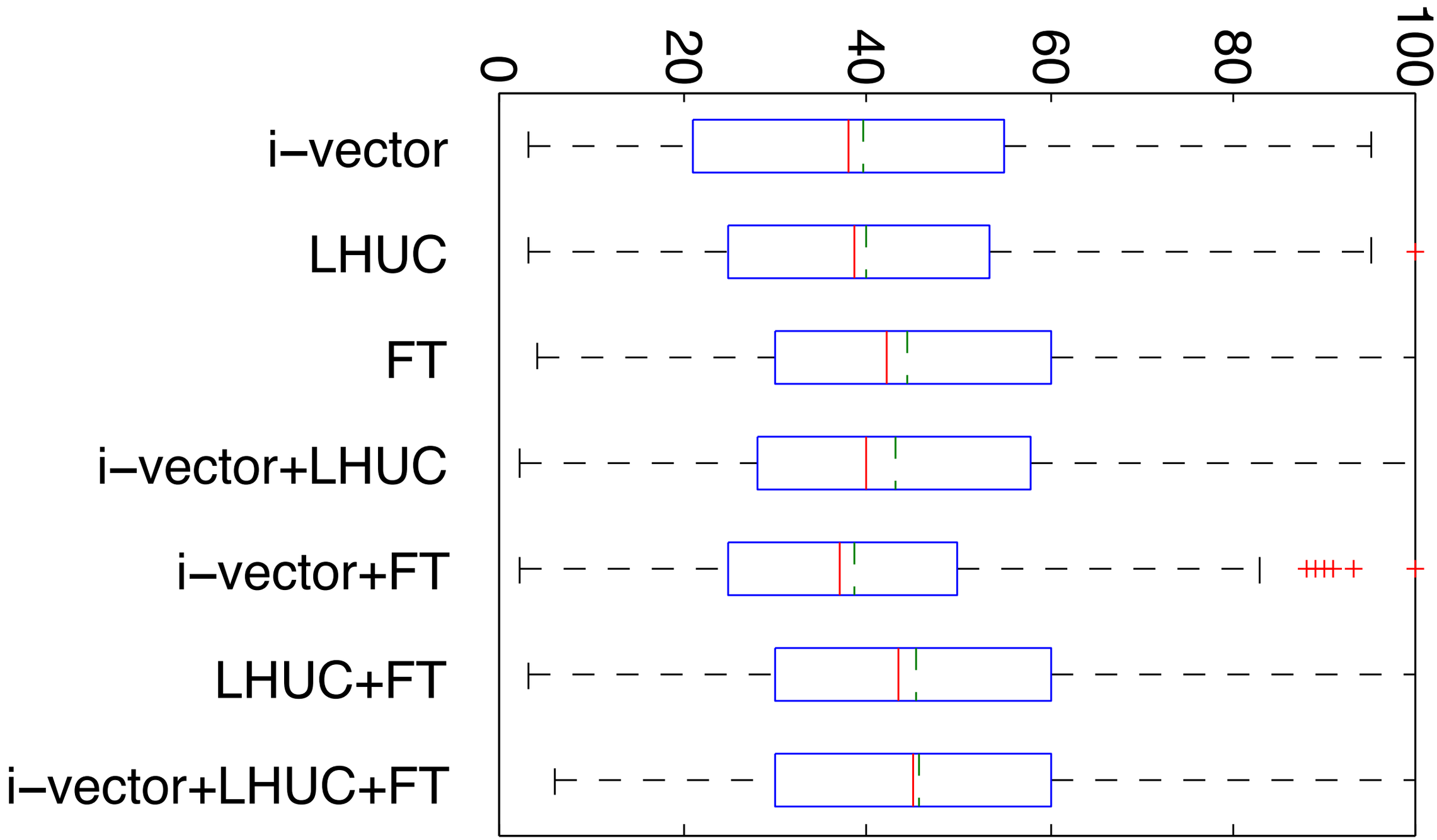
Adaptation in DNN speech synthesis



Z Wu et al, Interspeech-2015

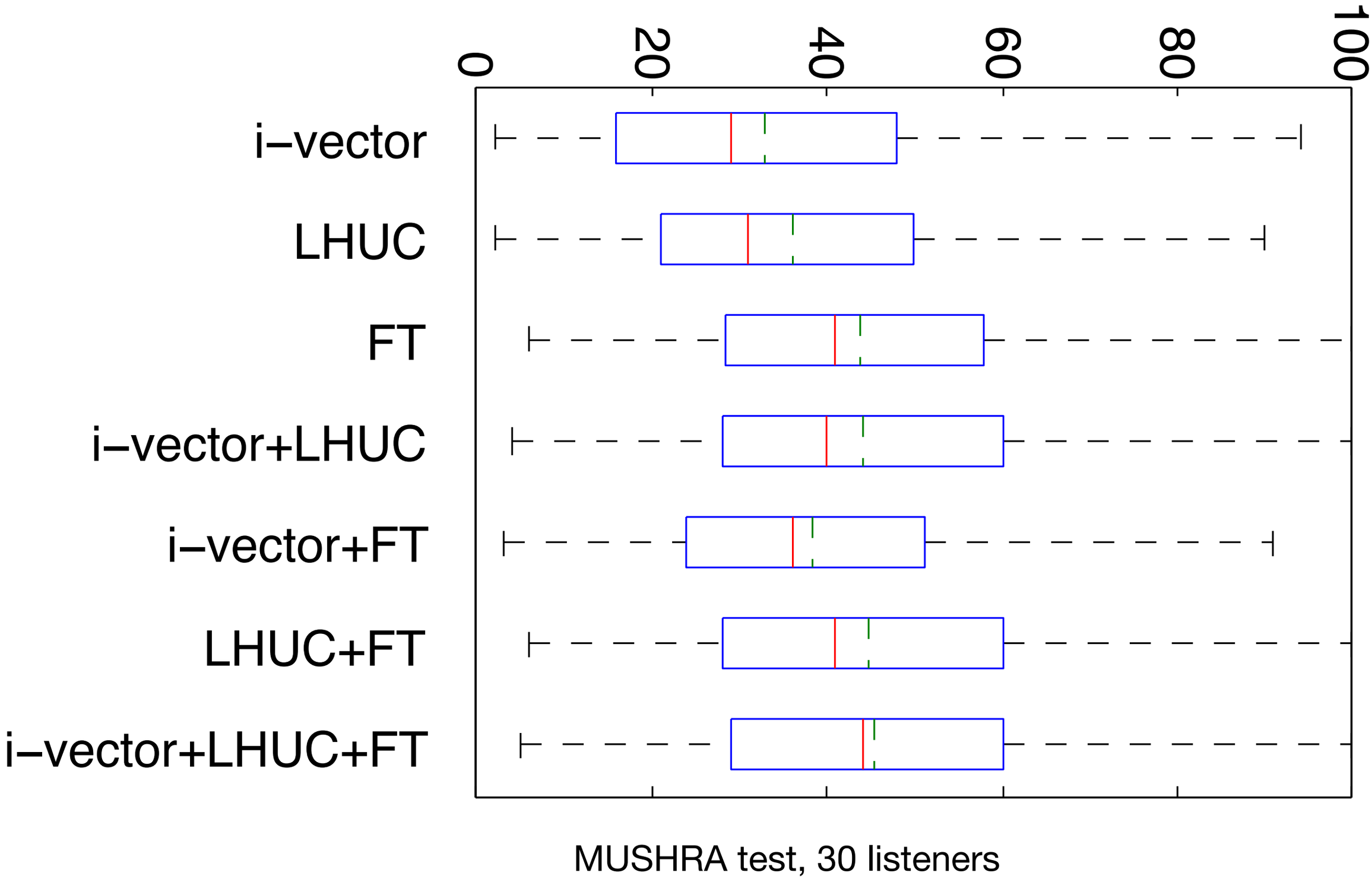
- 259D inputs
  - 60 melcep + Δ + Δ Δ
  - 25 BAP + Δ + Δ Δ
  - F0 + Δ + Δ Δ
  - Voicing
- 6 tanh hidden layers (1536 units), linear output layer
- SD normalisation of vocoder parameters

Naturalness evaluation

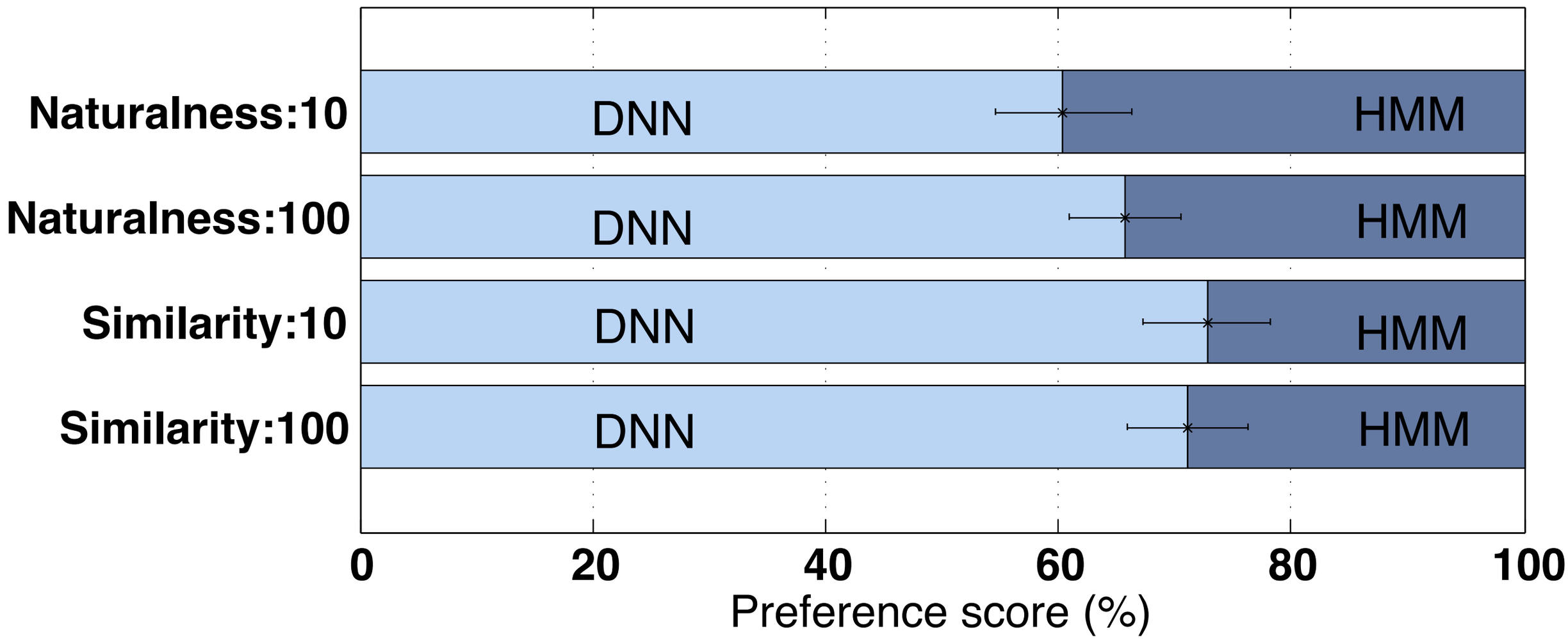


MUSHRA test, 30 listeners

Similarity evaluation



DNN vs HMM



Preference test, 30 listeners

DNN adapted using i-vector+LHUC+fMLLR
HMM adapted using CSMAPLR

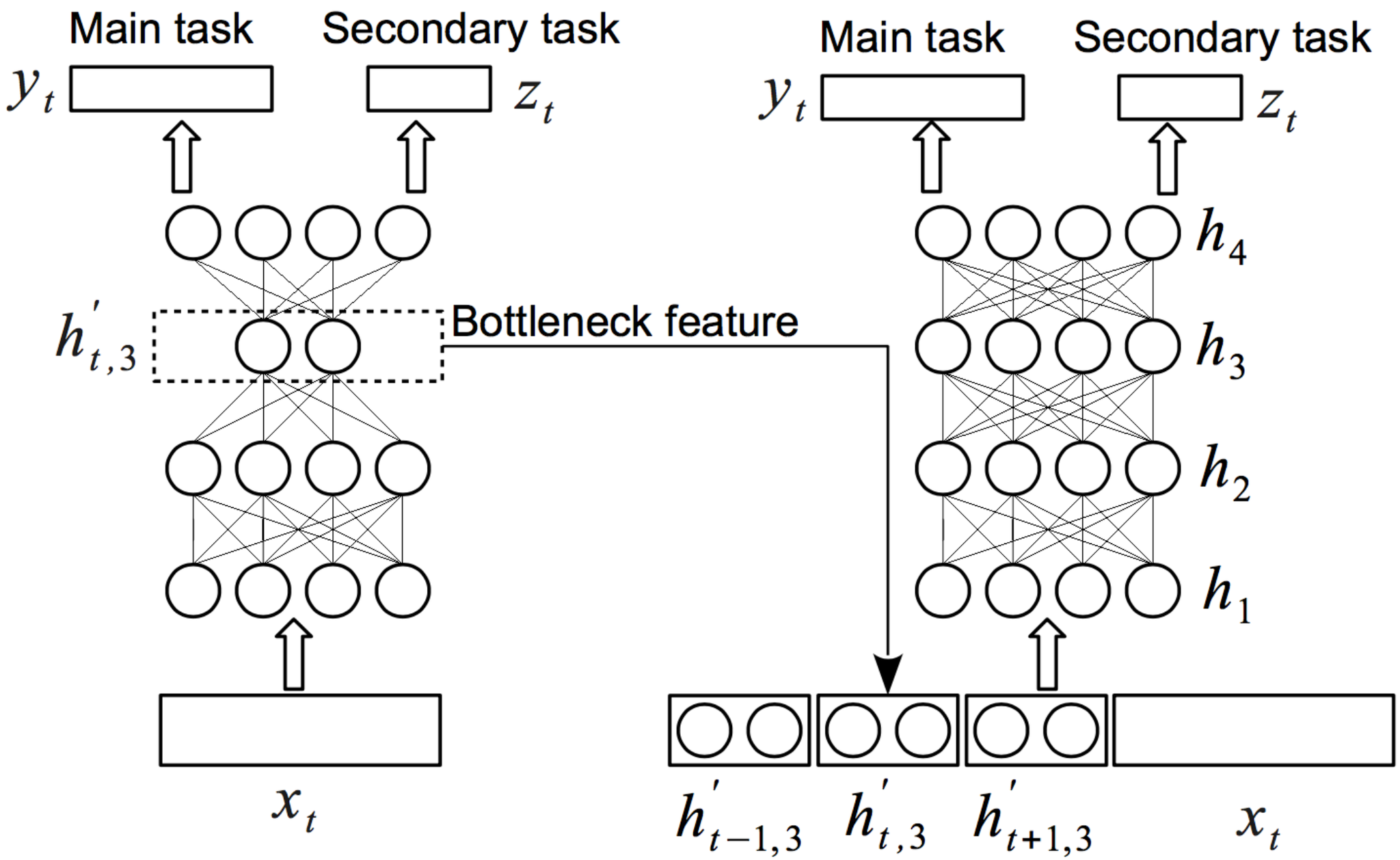
THE UNIVERSITY of EDINBURGH

Multi-task learning

Multi-task DNNs in speech synthesis

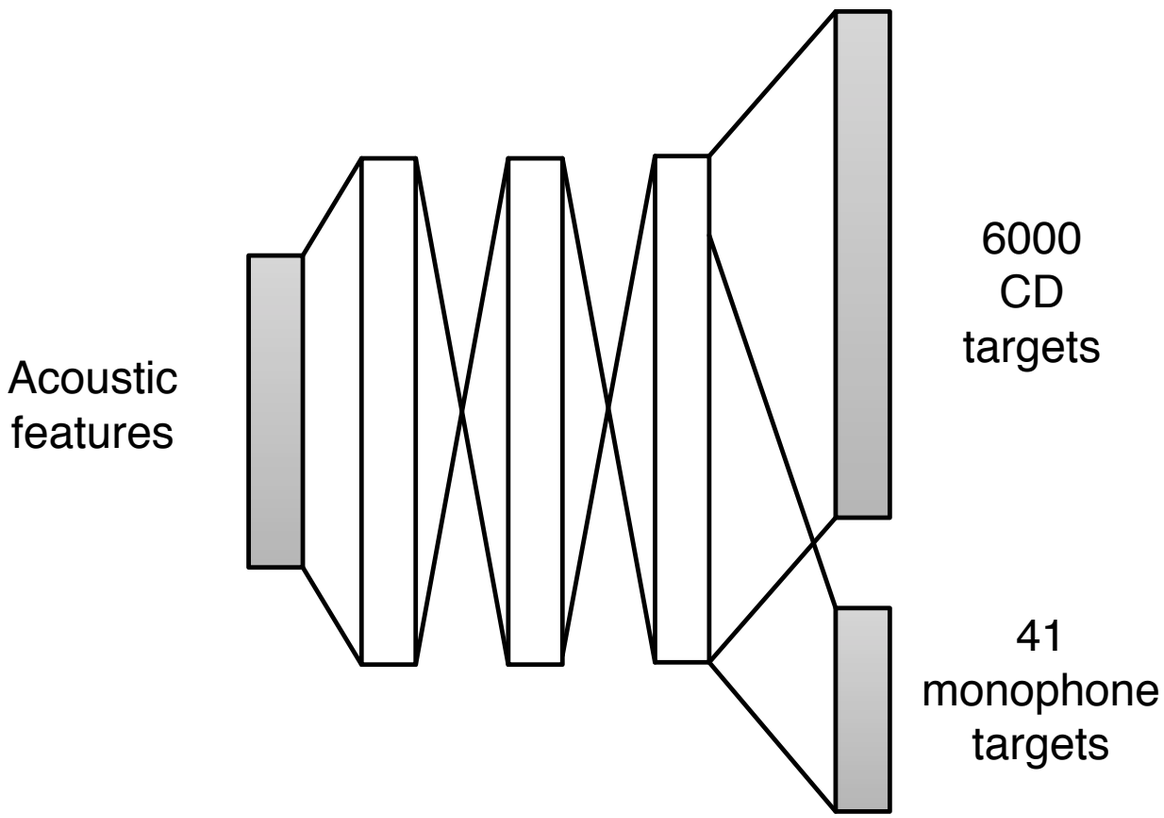
Z Wu et al, ICASSP-2015

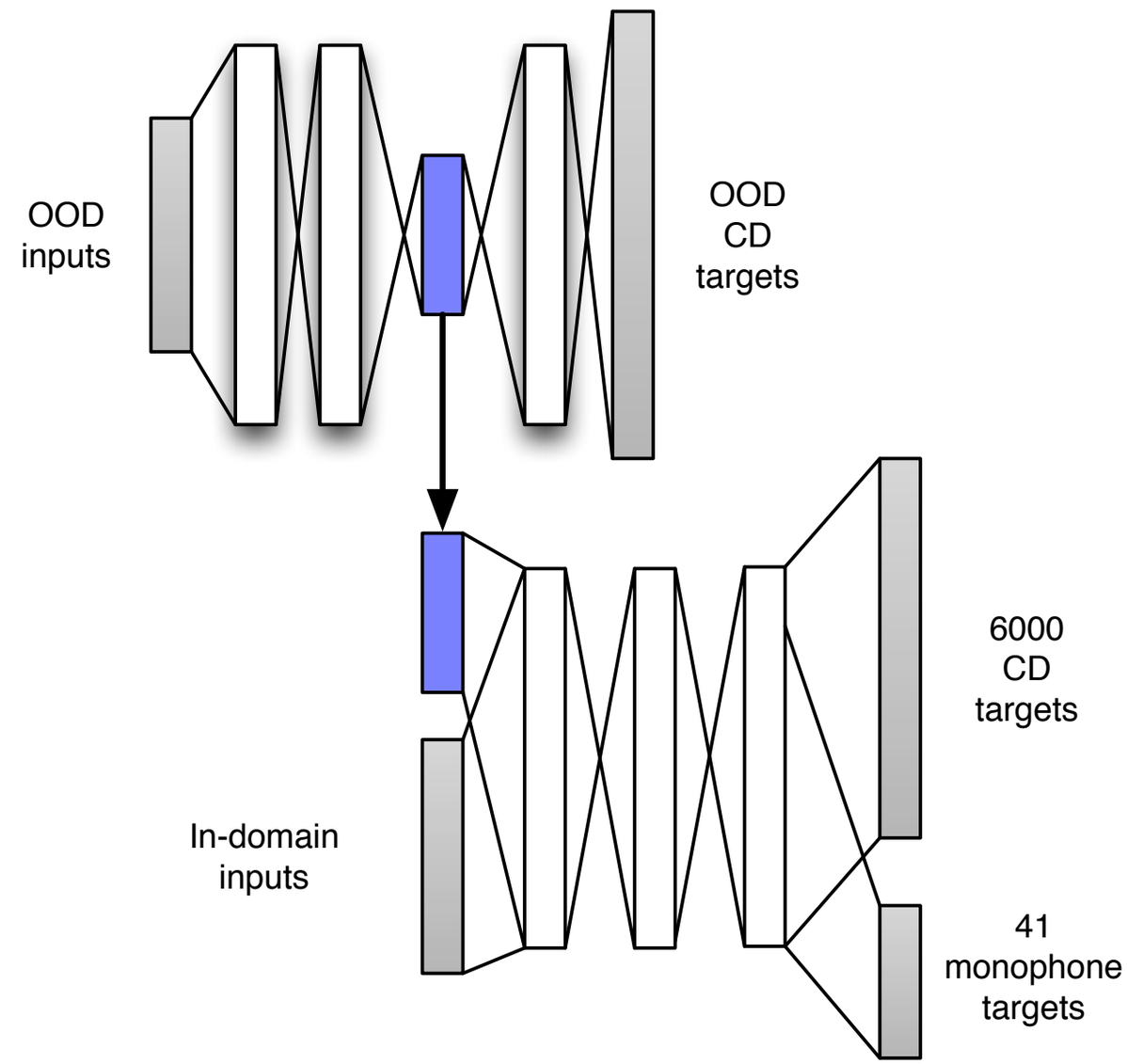
Main task: vocoder parameters
Secondary task: Glimpse-based perceptual measure (STEP)



Multi-task learning for ASR

Bell & Renals, ICASSP-2015





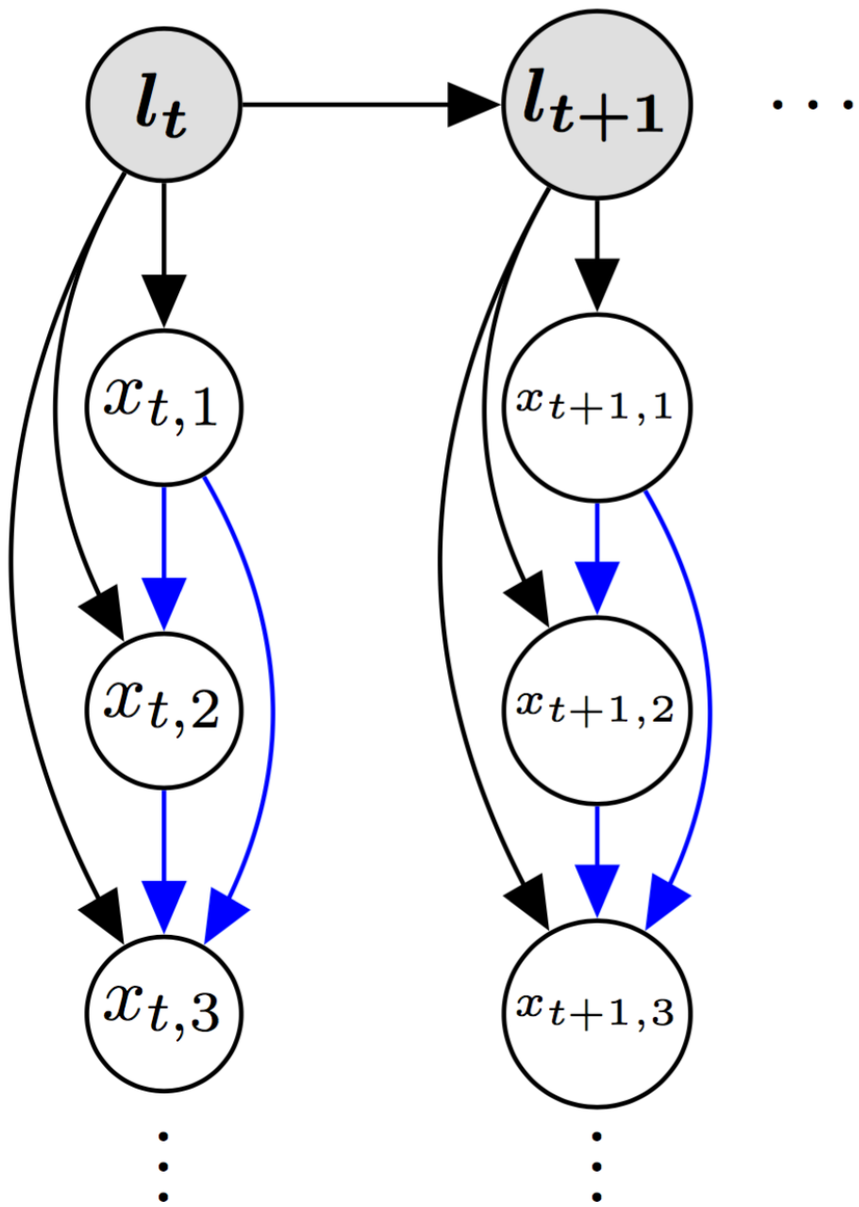
3–5% WER relative reduction (TED)

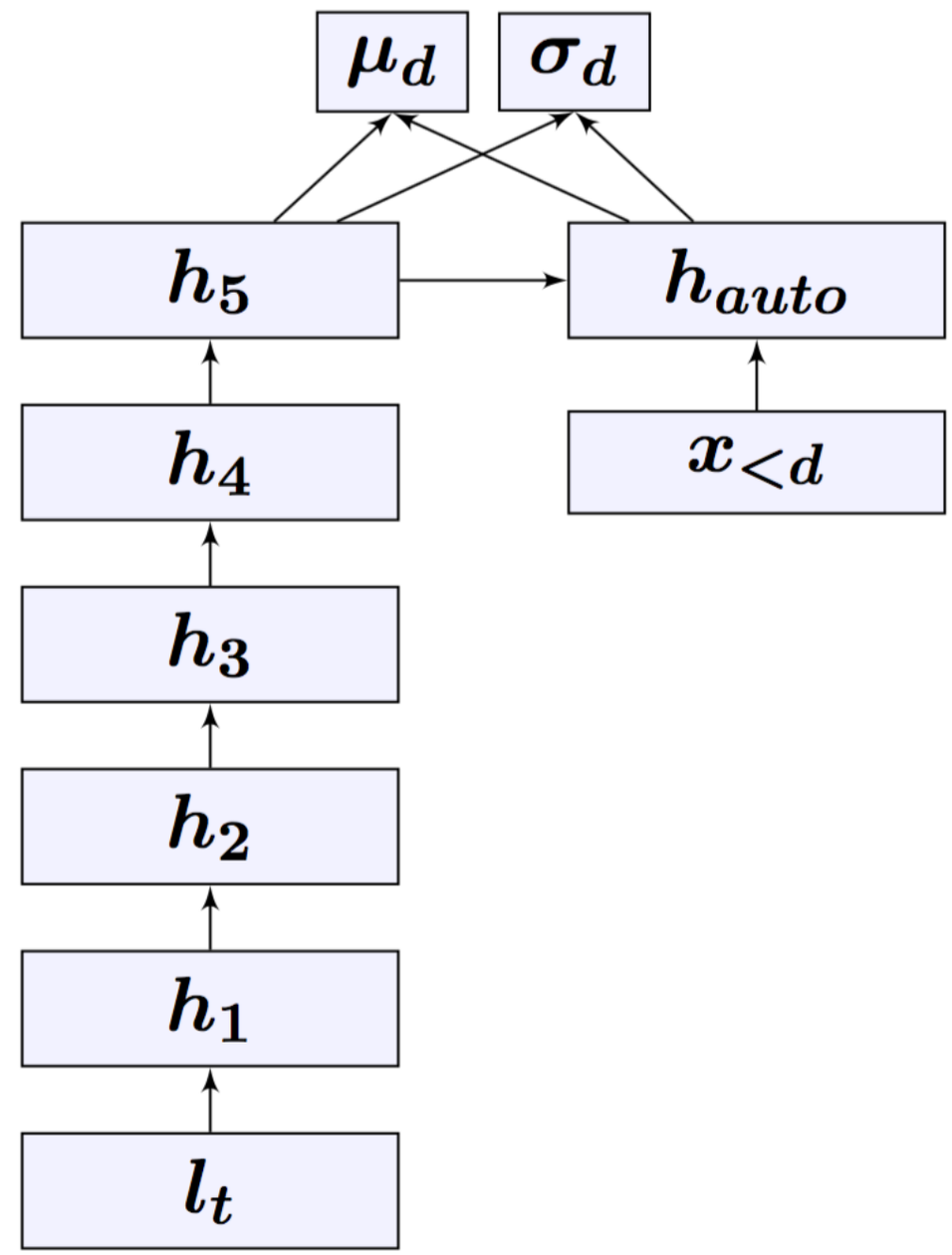
THE UNIVERSITY of EDINBURGH

Deep generative modelling

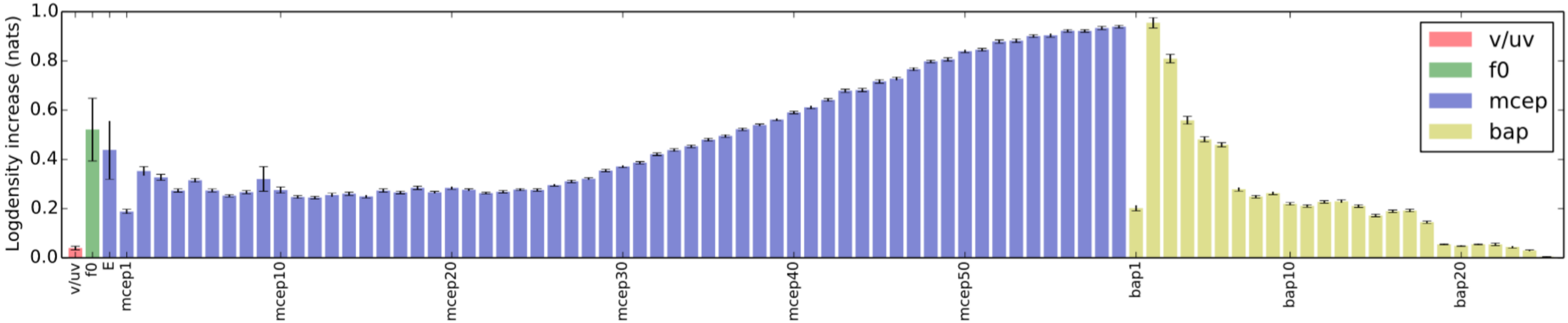
Trajectory RNADE

Uria et al, ICASSP-2015





RNADE synthesis



MDN		Traj. RNADE		n
Sample	Mean	Sample	Mean	
19.4%	-	**80.6%**	-	900
-	-	16.0%	**84.0%**	865
-	33.6%	-	**66.4%**	794

THE UNIVERSITY of EDINBURGH

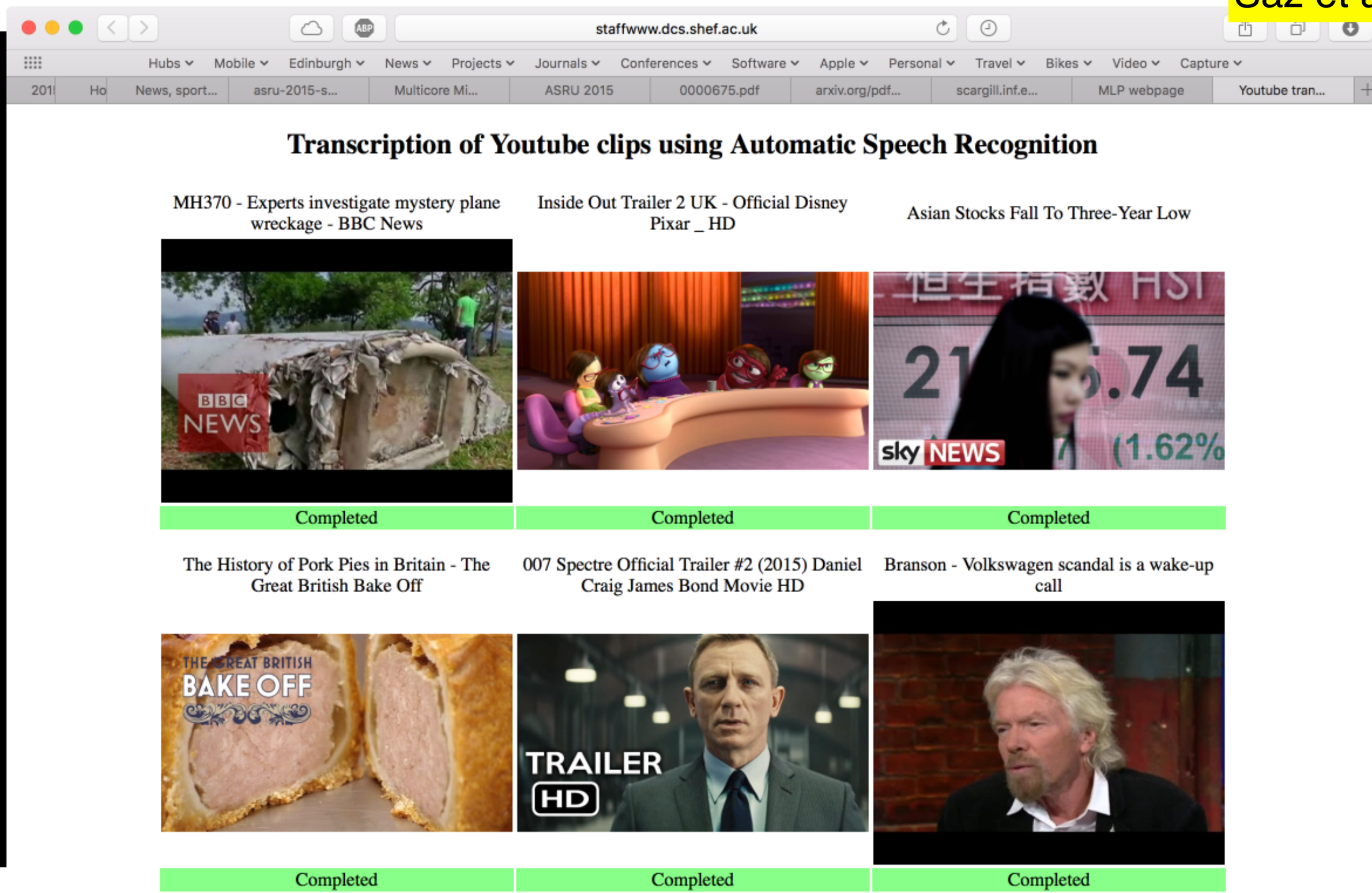
Exemplar Applications

Voice banking and personalised TTS

Veaux et al

Multi-domain ASR

Saz et al



THE UNIVERSITY of EDINBURGH

Browsing Oral Histories

Green et al

Browsing Oral History Upload

doncaster Submit

Cookie policy

user494912477
JD199078608

SOUNDCLOUD
Share

7:13 16:29

12

"and he was taking some dispatches to doncaster"

Details Summary Automatic Transcription Search

doncaster

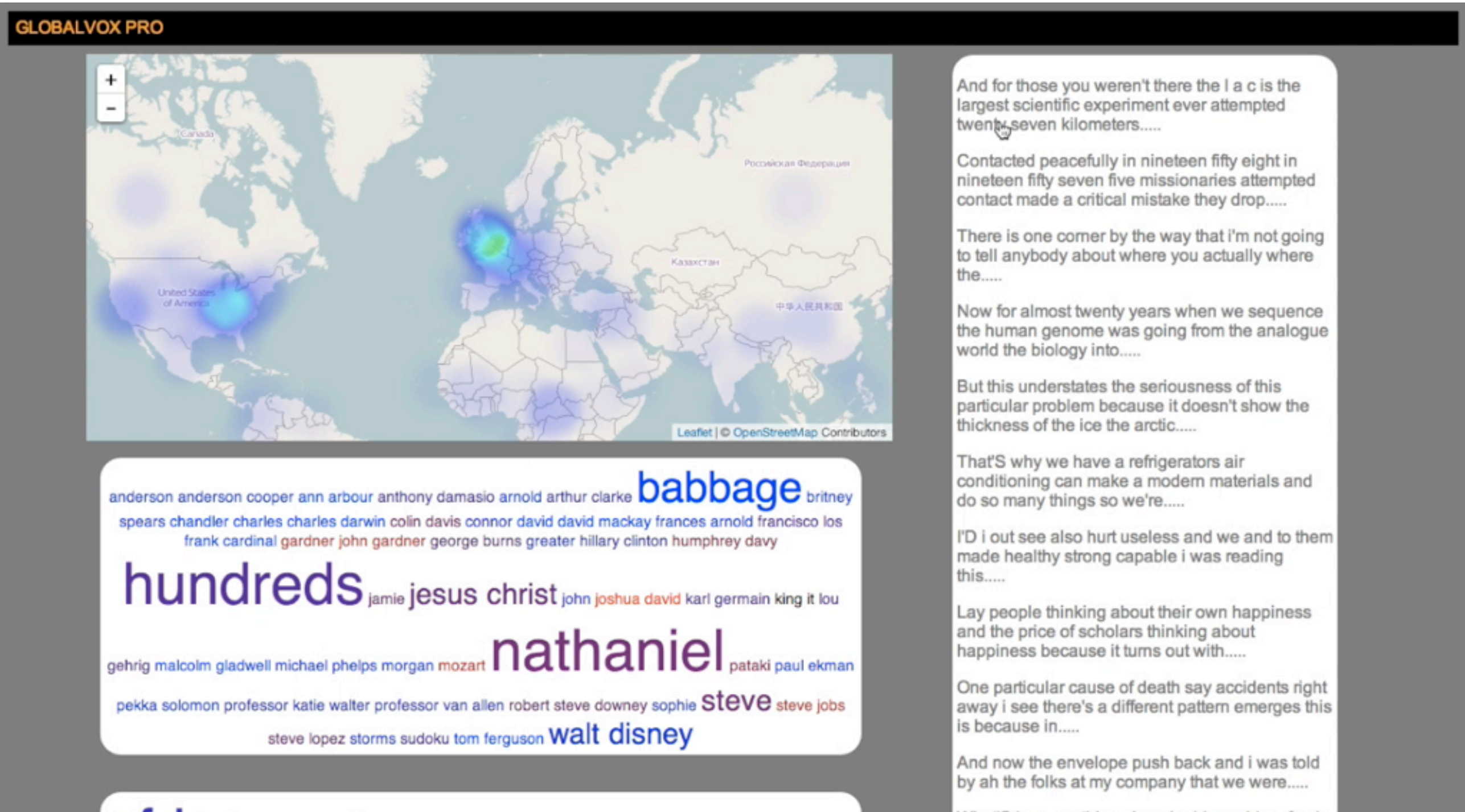
Submit

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"

THE UNIVERSITY of EDINBURGH

GlobalVox

Bell et al, Interspeech-2015



Concluding remarks

- Some recent advances from the NST project
- Other things include
  - distant speech recognition
  - disordered speech recognition
  - end-to-end RNN speech recognition
  - disfluent speech synthesis
  - speaker verification spoofing challenge
  - multilingual / cross-lingual recognition & synthesis
  - software: HTK v3.5, NN LM estimation, …

THE UNIVERSITY of EDINBURGH

NST people