EPSRC

Engineering and Physical Sciences Research Council

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

One minute madness

Posters and demos and coffee/tea downstairs in lecture room 4

Progress in Adaptation of DNN-based Acoustic Models

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

1. Pawel Swietojanski "*Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models*"

2. Yulan Liu/Penny Karanasou "*An investigation into speaker informed DNN front-end for LVCSR*"

3. Yulan Liu "*On the relationship between speaker informed DNN training and linear DNN input normalisation*"

4. Penny Karanasou "*I-Vector estimation using informative priors for adaptation of deep neural networks*"

5. Chunyang Wu "*Multi-basis Adaptive Neural Network for Rapid Adaptation in Speech Recognition*"

6. Peter Bell
"The UEDIN ASR Systems for the IWSLT 2014 Evaluation"

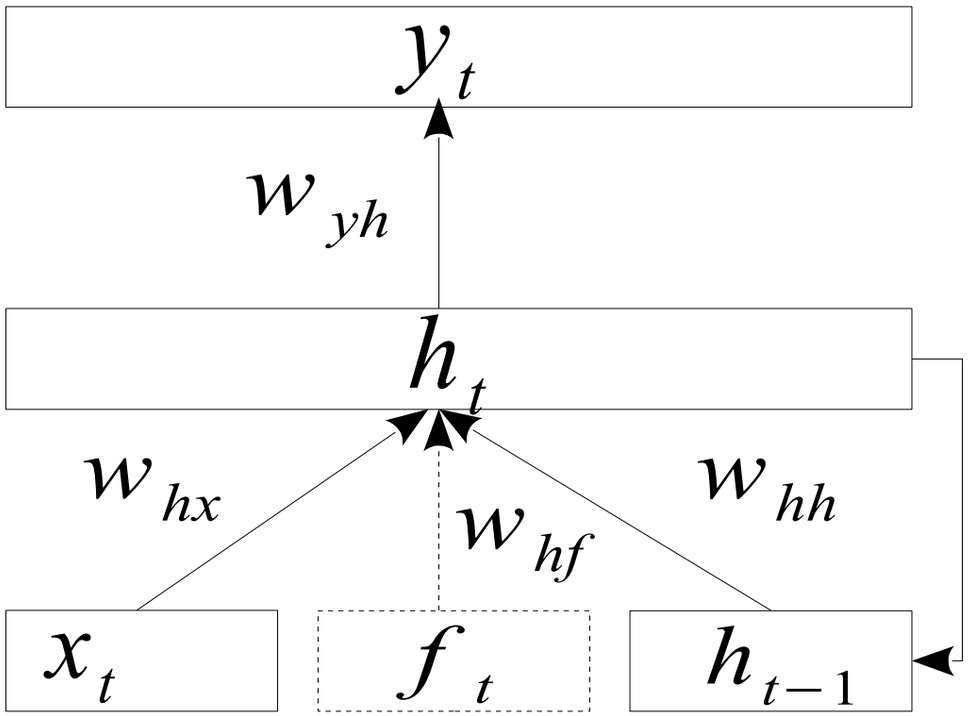
Natural Speech Technology

- This poster presents our systems for transcribing TED talks, entries to the International Workshop on Spoken Language Translation
- We entered systems for both English and German
- The highlights include:
  - hybrid DNN adaptation with LHUC
  - tandem multi-level adaptive networks
  - voice activity detection with an utterance duration prior
  - iterative dictionary refinement in German

THE UNIVERSITY of EDINBURGH

7. Prosodically-enhanced Recurrent Neural Network Language Models

Siva Reddy Gangireddy, Steve Renals, Yoshihiko Nankaku and Akinobu Lee



Recurrent neural network language model with a feature layer

- Prosody features
  - Word duration
  - Pause duration
  - Syllable duration
  - Syllable F0
- Speech Recognition
  - Switchboard
  - TED talks

8.Yanmin Qian

"*Noise-aware structured DNN for robust ASR*"

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Structured DNN
  - Each part has own function, MSE v.s. CE (MPE)
  - Different parts concatenate seamlessly
  - Decoding as normal DNN when finishing training
- Noise-aware Training
- Annealed Dropout Training

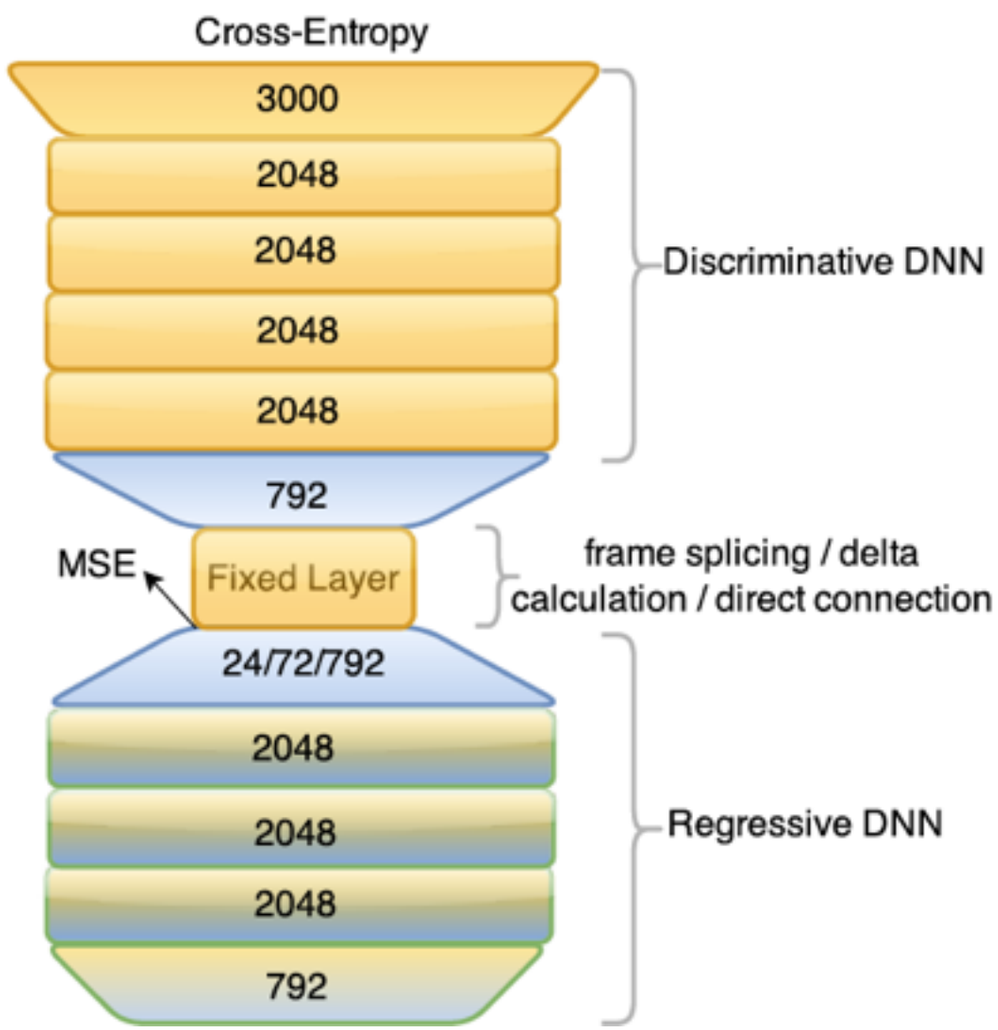


Table: WER (%) comparisons in a more realistic unseen noisy scenario.

System	SEN	2ND	AVG
DNN-HMM Baseline	15.5	24.9	20.2
Structured DNN	14.3	23.2	18.7
+ NAT + Annealed Dropout	13.6	22.8	18.2

Table: WER (%) comparisons of various systems in the literature on Aurora 4.

System	A	B	C	D	AVG
Best GMM-HMM [2]	5.6	11.0	8.8	17.8	13.4
DNN NAT DP [3]	5.4	8.3	7.6	18.5	12.4
DNN PP [1]	4.5	7.5	7.4	19.3	12.3
Spectral Mask [8]	4.5	7.9	7.5	17.7	11.4
JNAT [5]	4.5	7.4	8.1	16.5	11.1
TVWR Adap [4]	4.4	7.5	7.1	15.6	10.7
AD OSN LRF [6, 7]	4.0	7.2	6.4	14.5	10.0
Structured DNN	3.8	6.5	6.3	15.6	10.1

- The proposed Structured DNN has the good generalization, and achieves the best similar published results, ~10% WER on Aurora 4, only using the Sigmoid neurons

9. Liang Lu
"A study of RNN encoder-decoder for LVCSR"

- The model - RNN encoder-decoder
  - Mapping the variable length input sequence to the output sequence
  - The encoder maps the input sequence to a fixed length vector representation
  - The decoder computes the probability of output sequence given the vector
- The LVCSR system
  - Not hybrid - the outputs are words **not** HMM states
  - No pronunciation dictionary - since we use word outputs
  - Not explicit alignment - since we use vector representation of the whole input sequence
- The experiments - 50% WER on Switchboard without LM

10. Unsupervised Domain Discovery using Latent Dirichlet Allocation

Mortaza Doulaty, Oscar Saz and Thomas Hain

- Trying to discover domains in an unsupervised manner using Latent Dirichlet Allocation in highly diverse speech data
- Trying to find the relation of latent domains with existing manually labeled domains and meta-data
- Building / adapting latent domain models

11. Chao Zhang
"A general ANN extension for HTK"

Natural Speech Technology

- HTK has been extended to support ANNs with DCG architectures.
- HTK-ANN provides "built-in" support to Tandem and Hybrid systems.
- State-of-the-art facilities such as ANN adaptation and sequence training are included.
- HTK-ANN is compatible with most previous HTK functions.
- HTK-ANN is going to be included as part of HTK 3.5 coming in later in 2015.
- NST MGB challenge development system performance along with WSJ0 demo systems are provided.

UNIVERSITY OF CAMBRIDGE

12. Liang Lu / Pawel Swietojanski / Peter Bell
"*Kaldi extensions at Edinburgh*"

Two Kaldi recipes

- The AMI recipe in Kaldi repository now
  - Individual headset microphones
  - Multiple distant microphones
  - Single distant microphones
- The MGB challenge recipe

Interface beween Kaldi and CNTK

- Currently, CNTK is more flexible in training neural networks
- The interface support reading Kaldi features and labels
- Working on - sequence training of CNTK using Kaldi lattices

13.

Inserting Filled Pauses and Discourse Markers for Disfluent Speech Synthesis

M. Tomalin, R. Dall, M. Wester, X. Liu, B. Byrne, & S. King

Speech disfluencies (DISs) are pervasive in natural conversational speech

- *I'm getting a bit uh specific here*

DISs automatically inserted into fluent speech synthesis input text.

Four DISs modelled overtly:

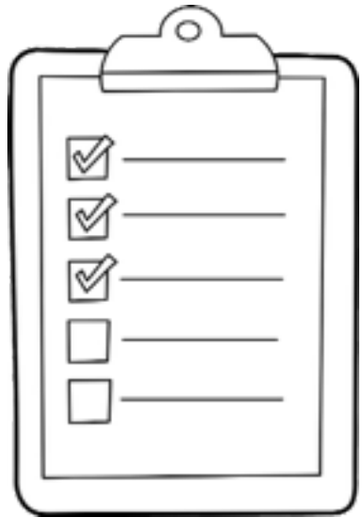
- 2 Filled Pauses: UH and UM
- 2 Discourse Markers: I MEAN and YOU KNOW

Overview of DIS-insertion system:

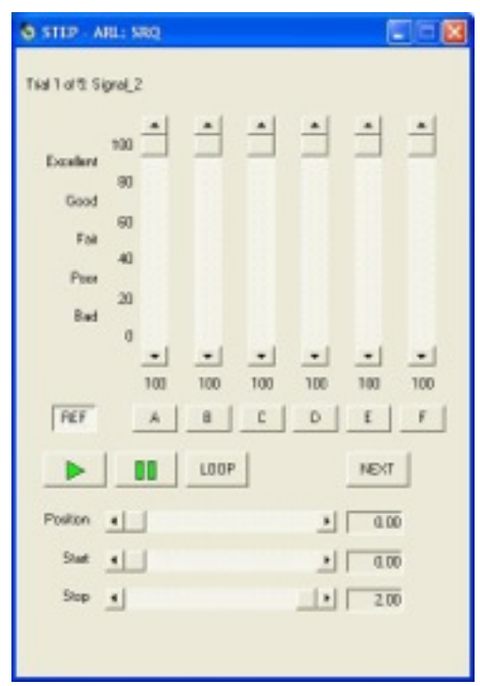
- Robust lattice-based rescoring framework
- Ngram and f-RNNLM built
- Initial lattices created with each DIS accessible from each word node
- Lattices expanded and rescored using Ngram, f-RNNLM, and Ngram + f-RNNLM
- Disfluency Parameter (DP) determines degree of disfluency in the output
- Disfluent output generated for specified DP
- Performance assessed using Precision, Recall, and F-score metrics

14. Mirjam Wester / Gustav Henter
“Subjective Evaluation of TTS”

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



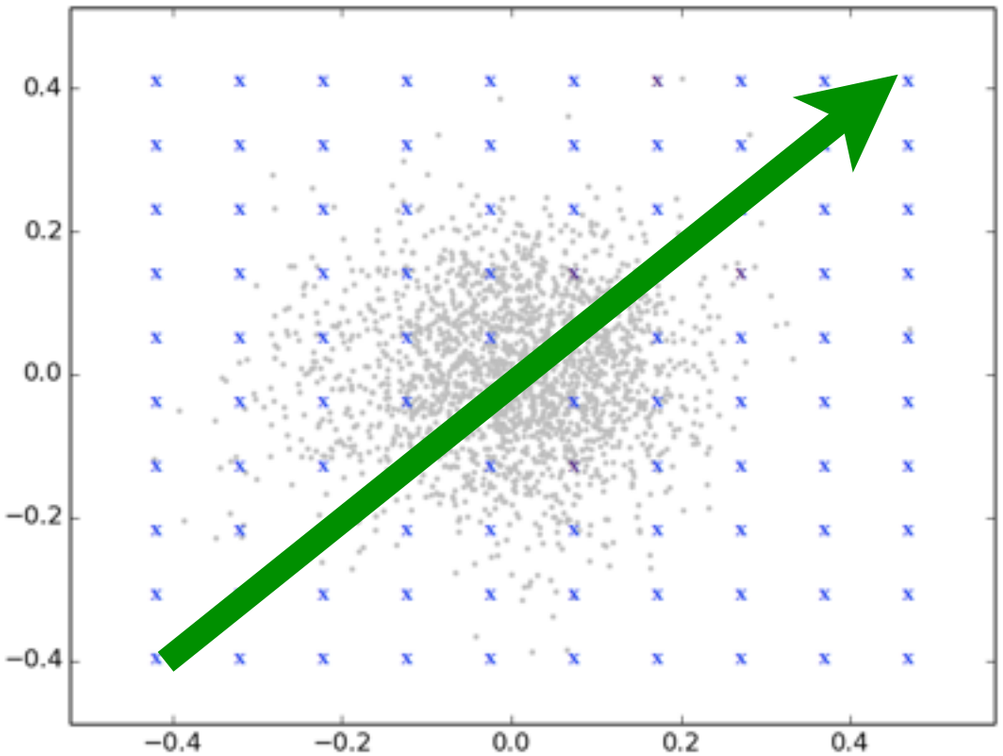
/m/	Mary came home
/p/	The puppy is playing with a rope
/b/	Bob is a baby boy
/f/	The phone fell off the shelf
/v/	Dave is driving a van
/th/ = /ð/	This hand is cleaner than the other
/n/	Neil saw a robin in a nest
/l/	A ball is like a balloon
/t/	Tim is putting on a hat
/d/	Daddy mended a door
/s/	I saw Sam sitting on a bus
/z/	The zebra was at the zoo
/sh/ = /s/	Sean is washing a dirty dish
/ch/ = /tʃ/	Charlie's watching a football match
/j/ = /dʒ/	John's got a magic badge
/y/ = /j/	The young chicks are yellow
/ng/ = /ŋ/	The bell's ringing
/k/	Karen is making a cake
/g/	Gary's got a bag of lego
/h/	Hannah hurt her hand

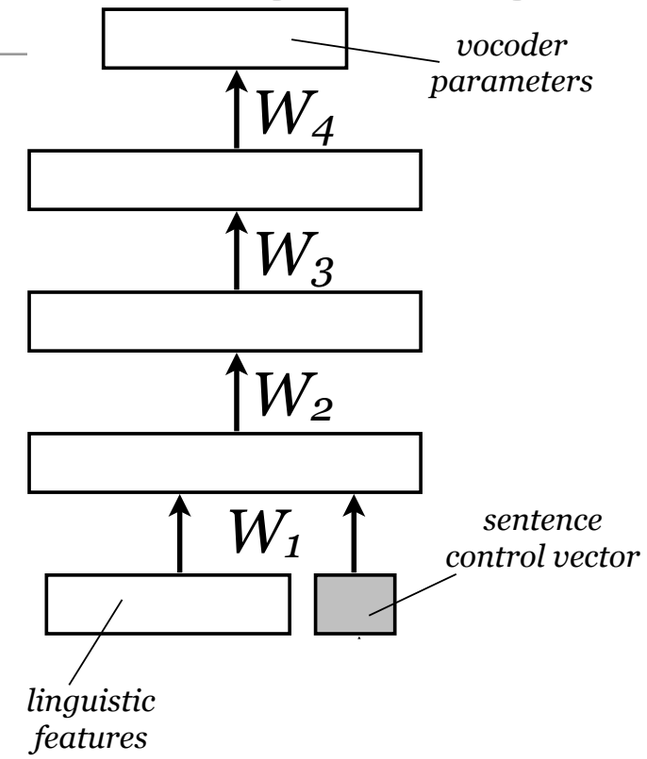


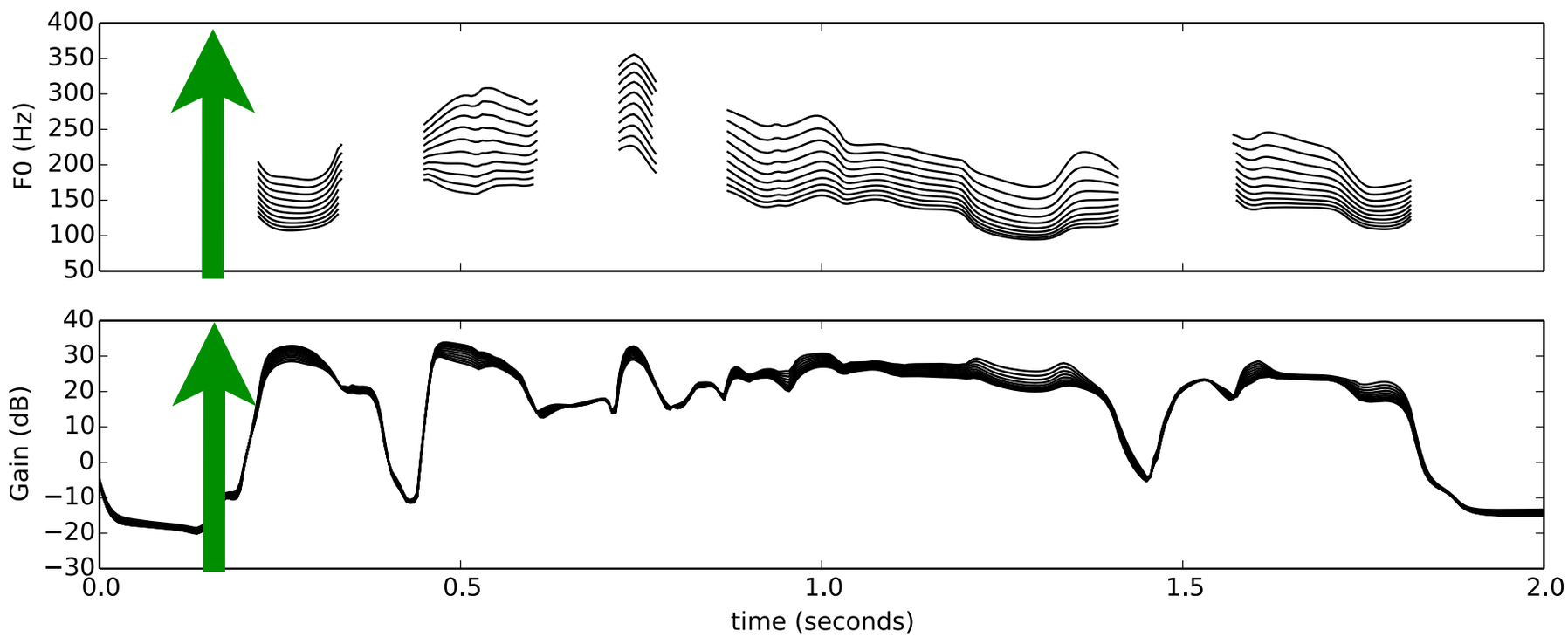
15. Oliver Watts "*Sentence-level control vectors for deep neural network speech synthesis*"

Natural Speech Technology

Edinburgh – Cambridge – Sheffield







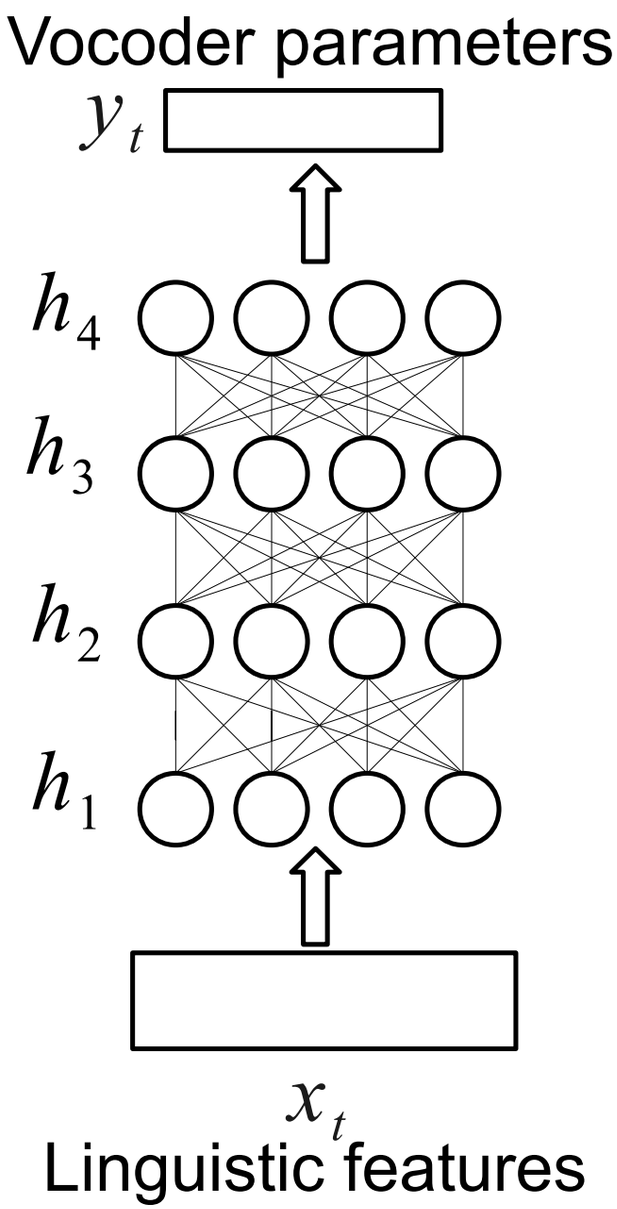
- Previous investigations highlighted:
  - Across-linguistic-context averaging harmful
  - Within-linguistic-context averaging much better
- Rich-context synthesis from literature aims to fix this issue
  - Models within-linguistic-context
  - However this uses original across-context computed leaf node as a reference for rich-context model selection
- This investigation uses DNN bottleneck features to select rich context models

17. Zhizheng Wu

"DNN Employing Multi-Task Learning and Stacked Bottleneck Features for Speech Synthesis"

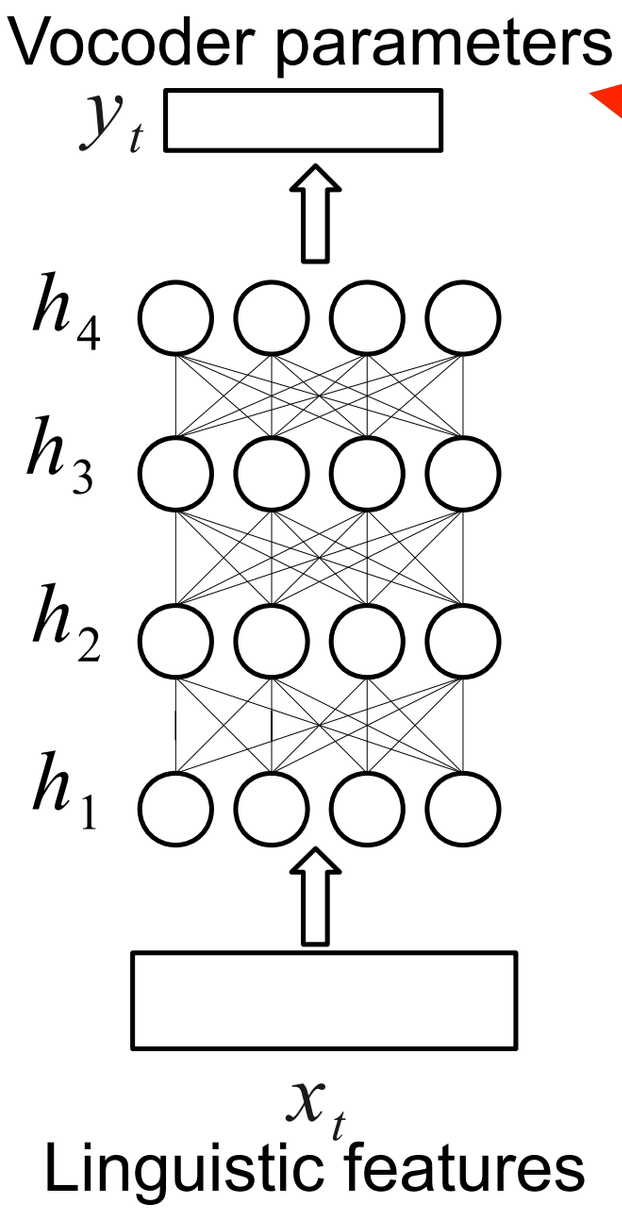
Natural Speech Technology

Edinburgh – Cambridge – Sheffield



17. Zhizheng Wu
"DNN Employing Multi-Task Learning and Stacked Bottleneck Features for Speech Synthesis"

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

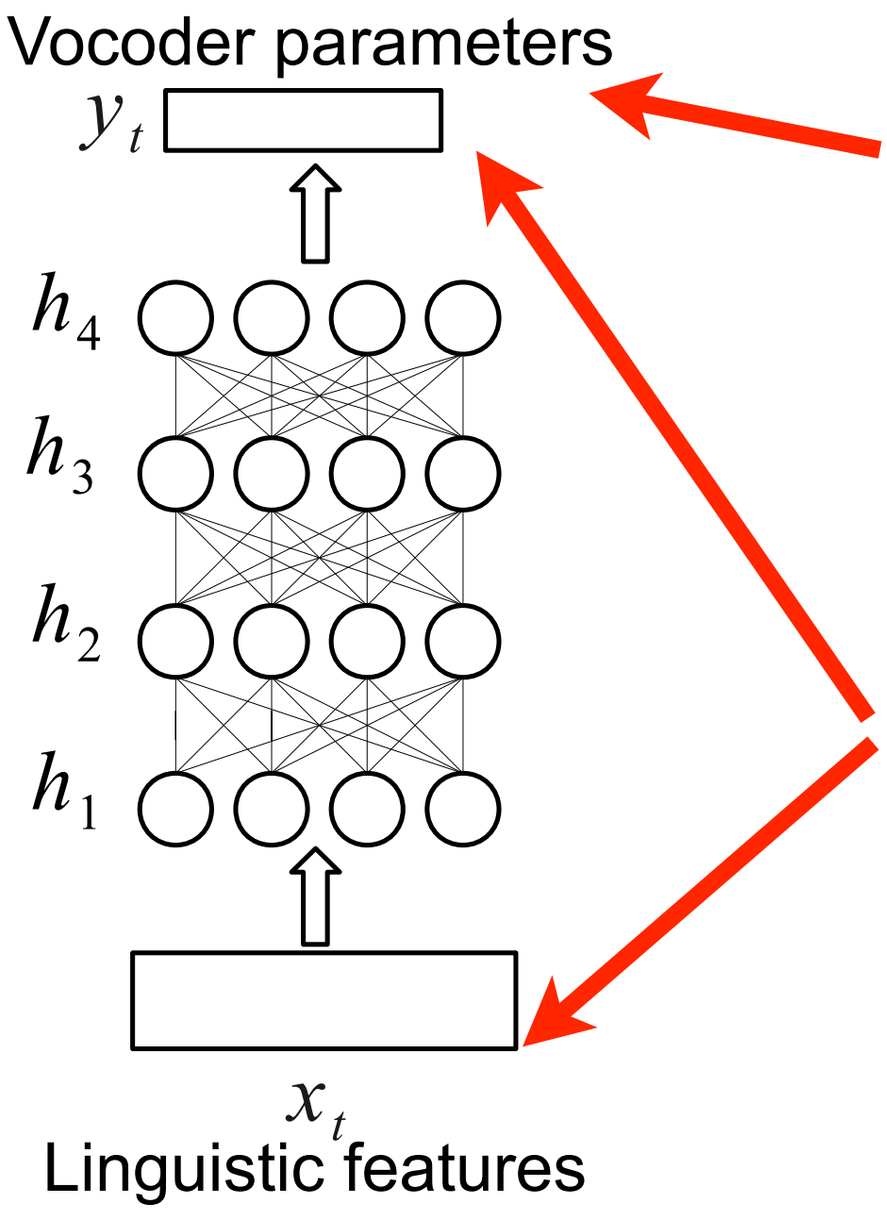


Perceptual sub-optimality

- Vocoder parameters which are invertible to speech waveform may not be correlated with human perception

17. Zhizheng Wu
"DNN Employing Multi-Task Learning and Stacked Bottleneck Features for Speech Synthesis"

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



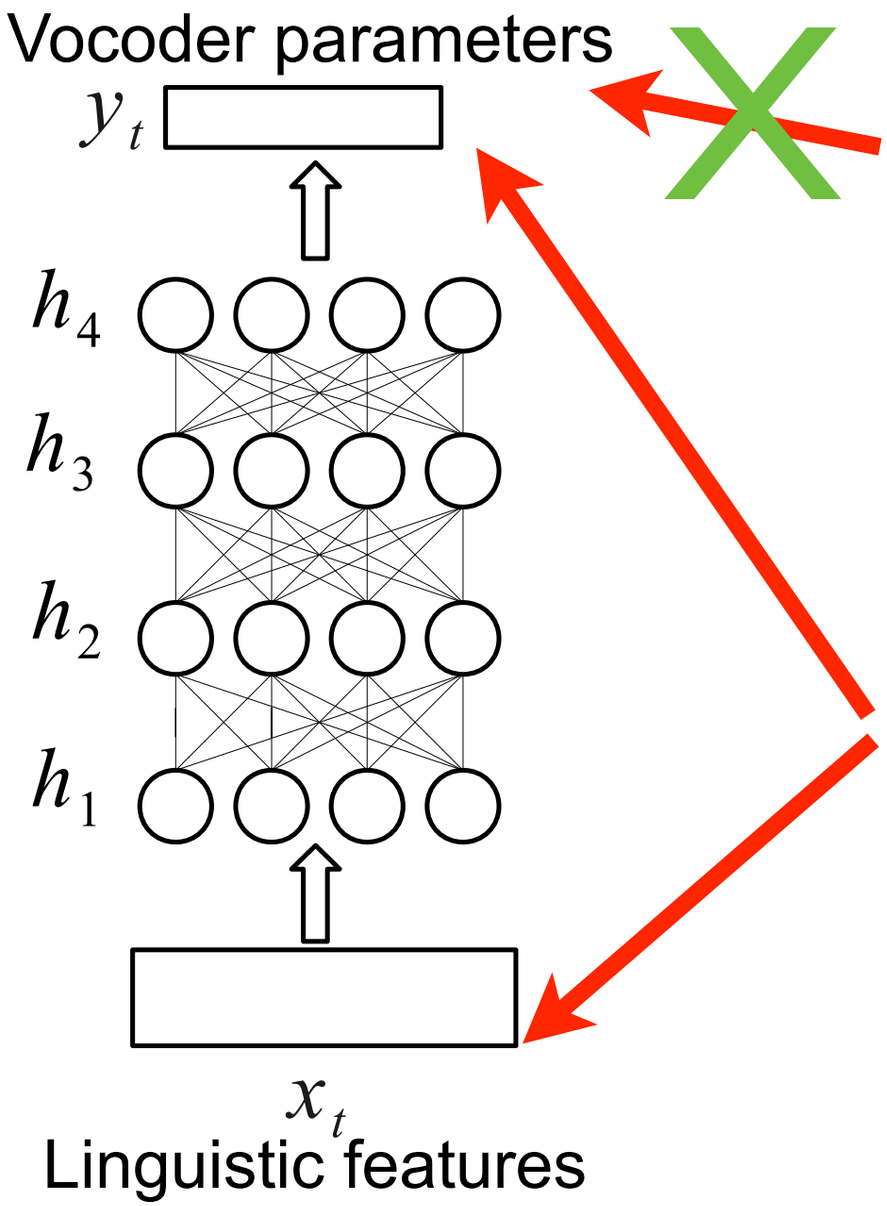
Perceptual sub-optimality

- Vocoder parameters which are invertible to speech waveform may not be correlated with human perception

Frame-by-frame independence

- Ignore contextual constraints at both input and output levels

Multi-task learning



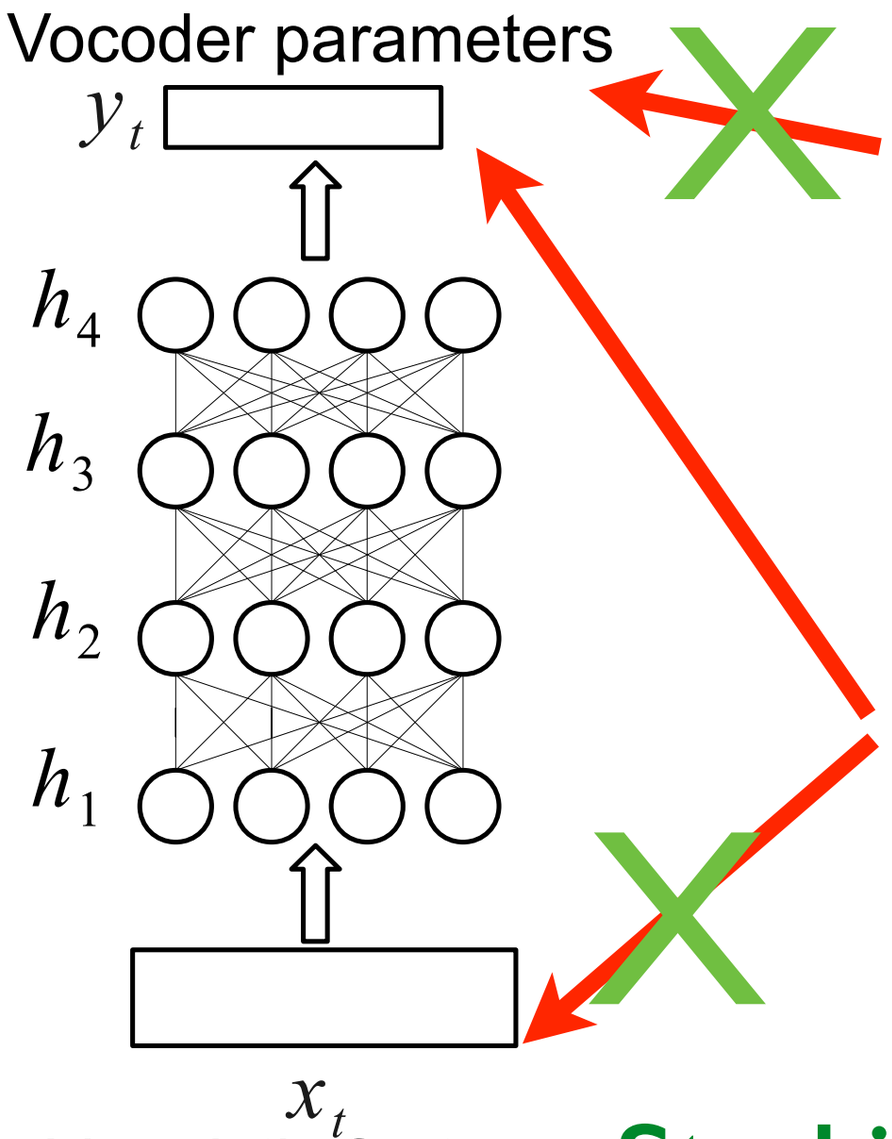
Perceptual sub-optimality

- Vocoder parameters which are invertible to speech waveform may not be correlated with human perception

Frame-by-frame independence

- Ignore contextual constraints at both input and output levels

Multi-task learning



Perceptual sub-optimality

- Vocoder parameters which are invertible to speech waveform may not be correlated with human perception

Frame-by-frame independence

- Ignore contextual constraints at both input and output levels

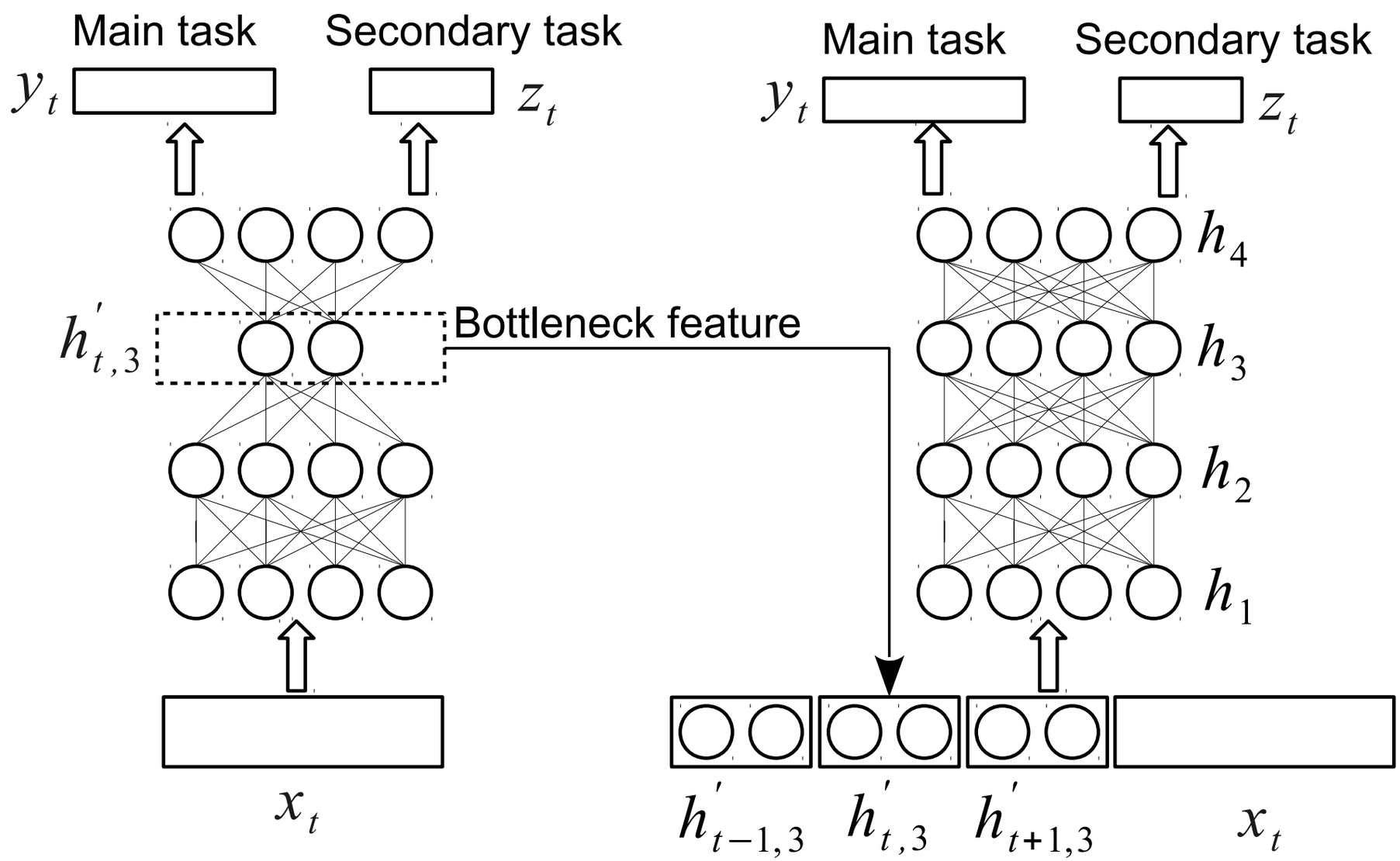
Stacking bottleneck features

17. Zhizheng Wu

"DNN Employing Multi-Task Learning and Stacked Bottleneck Features for Speech Synthesis"

Natural Speech Technology

Edinburgh – Cambridge – Sheffield



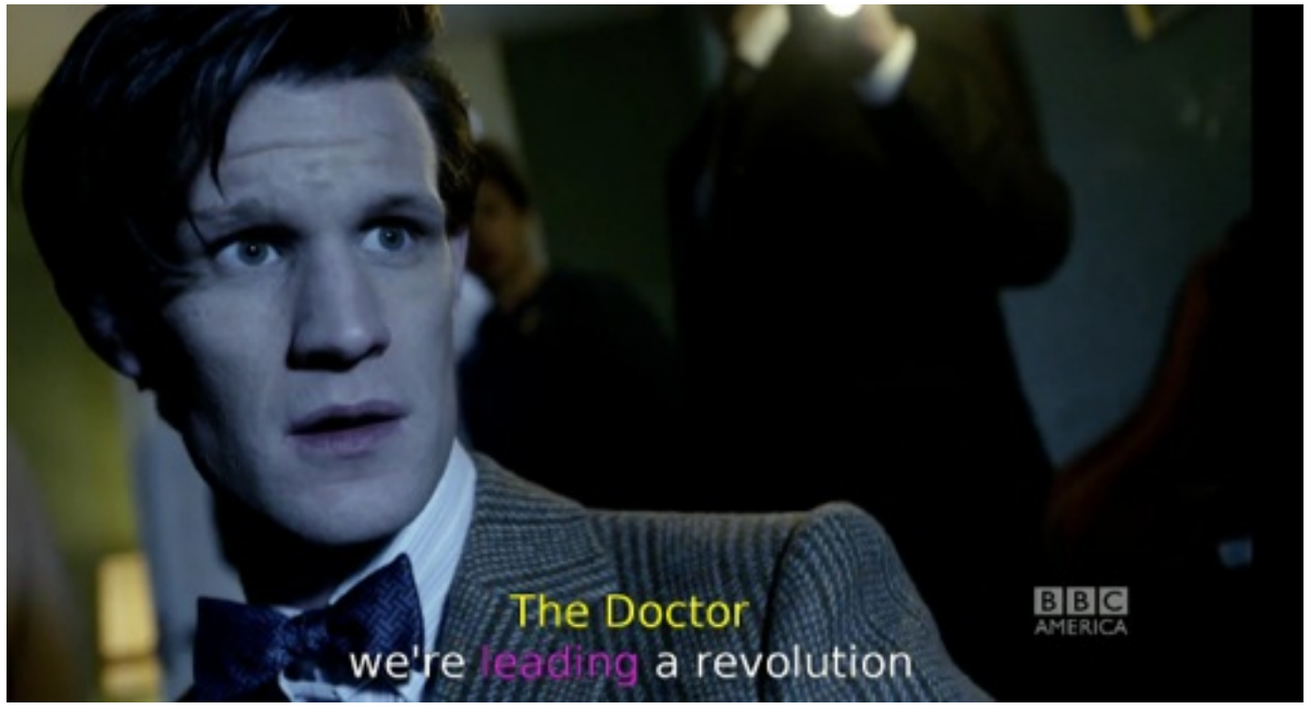
18. Pierre Lanchantin "*Details of the MGB Challenge data preparation*"

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Which data/metadata were provided to MGB challengers and how were they prepared from raw material (subtitles) ?

- Examples of **training data selection** using the provided metadata

- **Demo** of subtitles re-alignment for diverse tv-shows



19. Pierre Lanchantin "*Reconstructing voices within the Multiple-average-voice-model*"

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Personalisation of Voice output communication aids
- **Voice reconstruction**: build voices from disordered speech
- HMM-based speech synthesis approach: Adaptation+substitution of deteriorated components (risk of identity loss)
- We show that the **Multiple AVM** framework is well-suited to the Voice reconstruction task
  - complexity: requires a small quantity of data
  - flexibility: interpolation of component mean vectors can be performed in a "clean" eigenspace and interpolation weights can be fine-tuned by a practician
- We illustrate our points with **subjective assessment** of the reconstructed voice

20. Phil Green "Browsing Oral History"

- Recorded memories, reminiscences
- Long Interviews, transcriptions are rare
- Topic-focussed (but topic may be very wide)
- Many collections, no central database.. 1000s of hour
- Usage limited by retrieval problems…
- Search the ASR transcription, play back the audio.

Web site demonstrator: 'Duty Calls' project: Brodsworth Hall

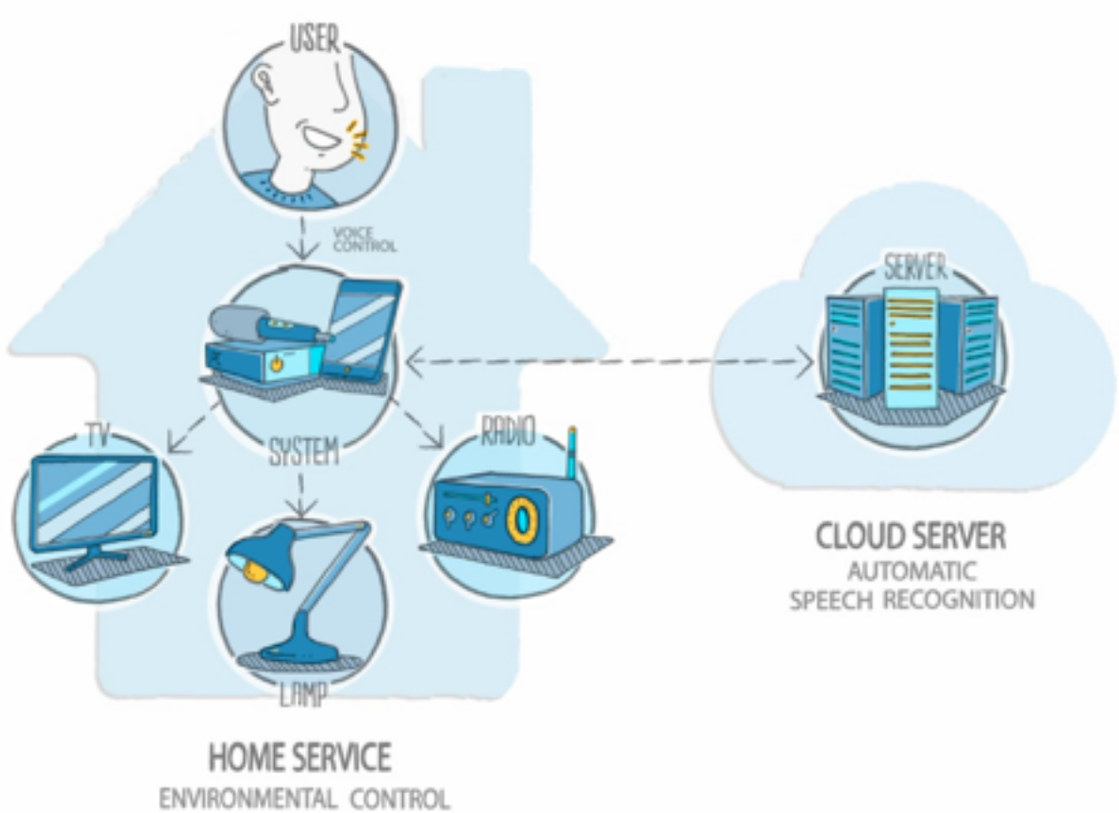
The University Of Sheffield.

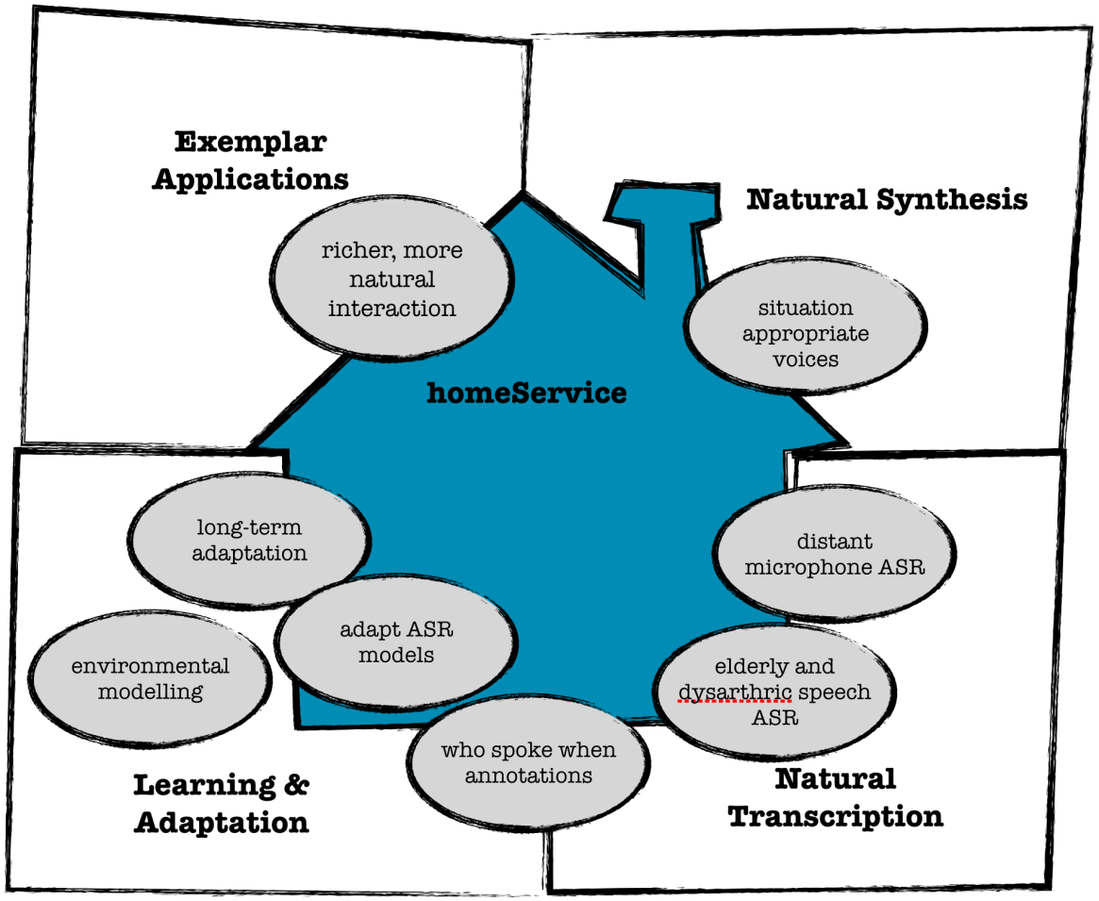
Natural Speech Technology

Edinburgh – Cambridge – Sheffield

21. Automatic speech recognition for people with disordered speech: results from online and offline experiments

Mauro Nicolao, Heidi Christensen, Salil Deena, Stuart Cunningham, Phil Green, Thomas Hain

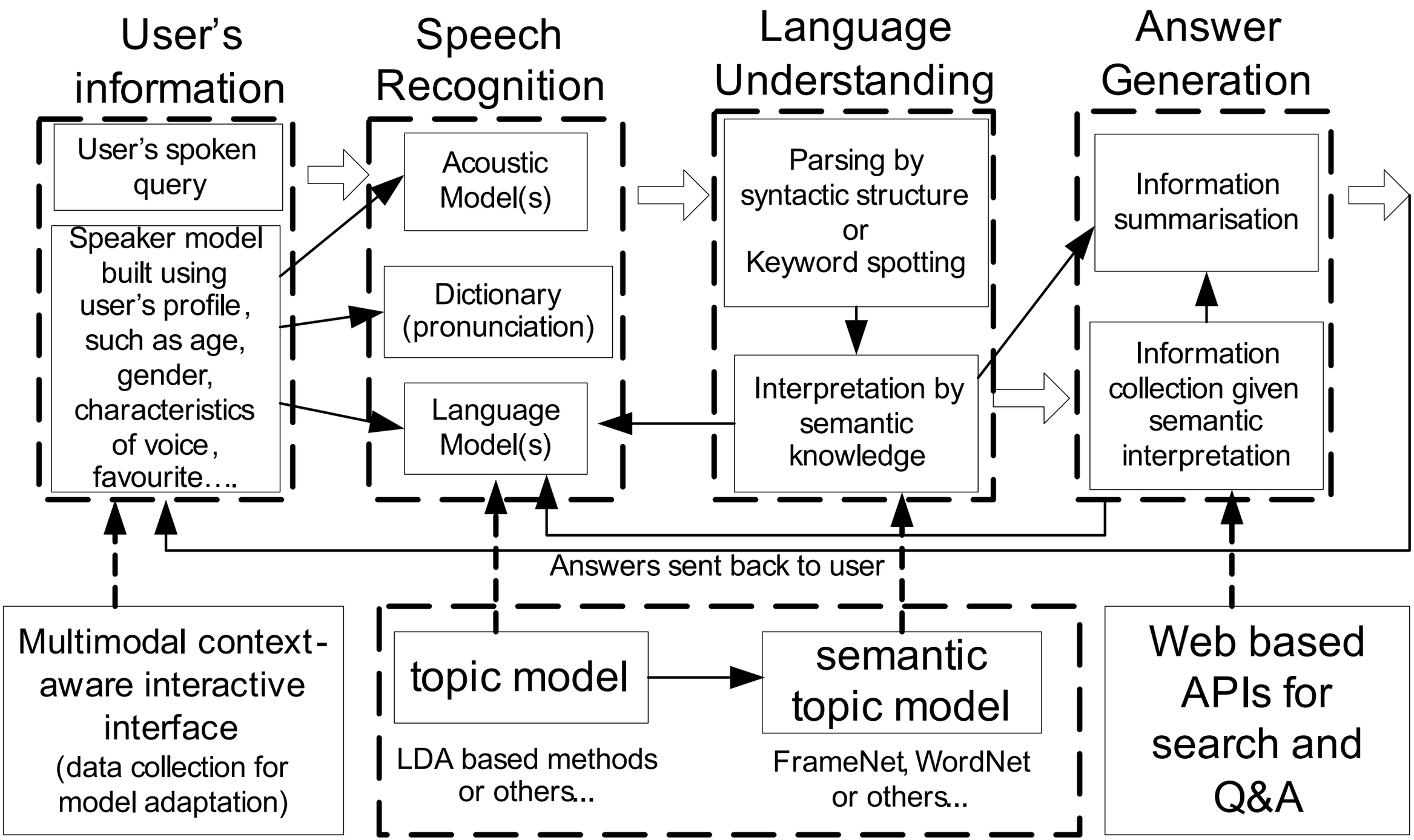




http://www.natural-speech-technology.org/homeService

22. Qiang Huang "*User-dependent interactive system*"

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



23. Peter Bell "GlobalVox Demo"

Natural Speech Technology

We demonstrate a prototype system for analysing and translating news stories, developed as part of the BBC's 2014 *newsHACK*