Natural Speech Technology Programme Overview

Steve Renals
Centre for Speech Technology Research
University of Edinburgh

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

Natural Speech Technology Programme: Team

- **CSTR, University of Edinburgh**:

  Steve Renals, Simon King, Junichi Yamagishi

  Peter Bell, Arnab Ghoshal, Jonathan Kilgour, Heng Lu, Pawel Swietojanski, Christophe Veaux

- **Speech Research Group, University of Cambridge**:

  Phil Woodland, Mark Gales, Bill Byrne

  Pierre Lanchantin, Andrew Liu, Yanhua Long, Marcus Tomalin

- **Speech and Hearing Research Group, University of Sheffield**:

  Thomas Hain, Phil Green, Stuart Cunningham

  Heidi Christensen, Charles Fox

Overall aim

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Significantly advancing the state-of-the-art in speech technology

- making it more natural
- applied to speech recognition and speech synthesis
- approaching human levels of

  reliability

  adaptability

  fluency

State of the art: Recognition & Synthesis

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Learning from data – HMM/GMM framework

- **Context-dependent modelling**: divide and conquer using phonetic decision trees
- **Speaker adaptation**: MLLR and MAP families
- **Different training criteria**: maximum likelihood, minimum phone error, minimum generation error
- **Discriminative long-term features**: posteriograms, bottleneck features, deep networks
- **Model combination**: at the feature / distribution / state / model / utterance level

Key research themes

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Improving core speech technology
  - ***Common modelling framework*** for synthesis and recognition
  - ***Fluency***
  - Capturing ***richer context***
  - ***Personalisation***
  - ***Expression*** and prosody

Objectives

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

1. **Learning and Adaptation**
2. Natural **Transcription**
3. Natural **Synthesis**
4. Exemplar **Applications** (driven by user group):
   - ***homeService***: personalised speech technology to provide better interfaces (focus on older people & disabled people)
   - ***lifeLog***: personalised wearable devices and transcribe/index all encountered audio
   - extracting structure from ***media archives***

Learning and Adaptation

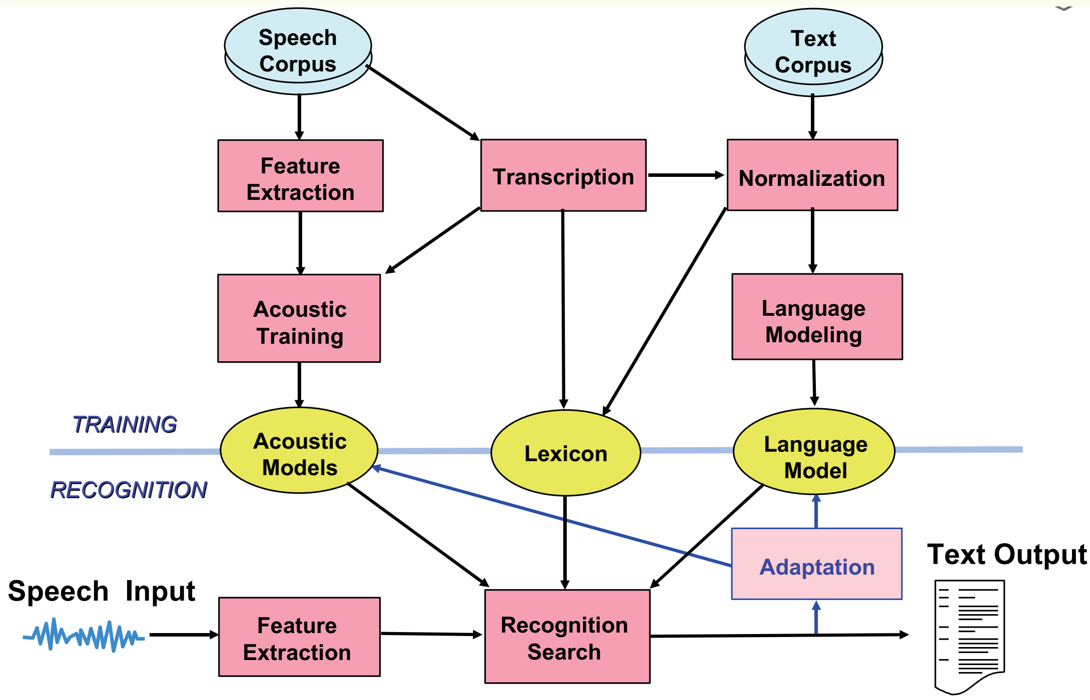
Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Speech recognition and speech synthesis based on learning statistical models from data
- Current systems can adapt to the speaker or the domain automatically
- Challenges (for both recognition and synthesis)
  - ***Factoring models*** to different causes of variability
  - Almost ***instantaneous adaptation***
  - ***Unsupervised training*** to take advantage of available data
  - ***Learning not to repeat mistakes***

Natural Transcription: Standard Systems

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Typical architecture for training / test



Natural Transcription: Standard Systems

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- ***Acoustic models*** (typically phone-context-dependent HMMs) trained from 100's to 1000's(+) hours of audio
- ***Language models*** (normally word N-grams) trained from large amounts of text data (up to billions of words) & ***lexicon*** normally fixed
- Training data is normally ***in-domain*** (including known language) with known correct transcripts (supervised).
- ***Purpose-built*** systems for different task domains e.g. meetings; broadcast news; voicemail etc
- Limited ***adaptation***

Current Transcription Systems: What's Wrong?

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Domain-specific systems (lots of point solutions)
  - All factors combined in one model & hence very data intensive (expensive, can't cover all situations)
- System performance isn't ***natural***
  - slow to adapt
  - poor generalisation (accents, acoustic environment etc)
  - can't fully exploit all known context
  - doesn't produce fluent output

Natural Transcription

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Goal: Speech recognisers that
  - output "who spoke what, when, and how"
  - give high accuracy
  - have a wide coverage of speaker, environment etc
  - are flexible and minimise in-domain training data needs
  - can be personalised
  - produce fluent output

Wide Domain Coverage

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Two approaches to transcription to cover many domains
  - Structure diverse data (from User Group + other sets) using advanced clustering techniques
  - Build canonical models (acoustic models and language models) which can be rapidly adapted to the particular sub-domain (incl speaker, environment etc)
- These overall approaches aim to allow
  - good performance on domains without any (or very limited) domain-specific training data
  - more robust speech recognition: works well for more situations

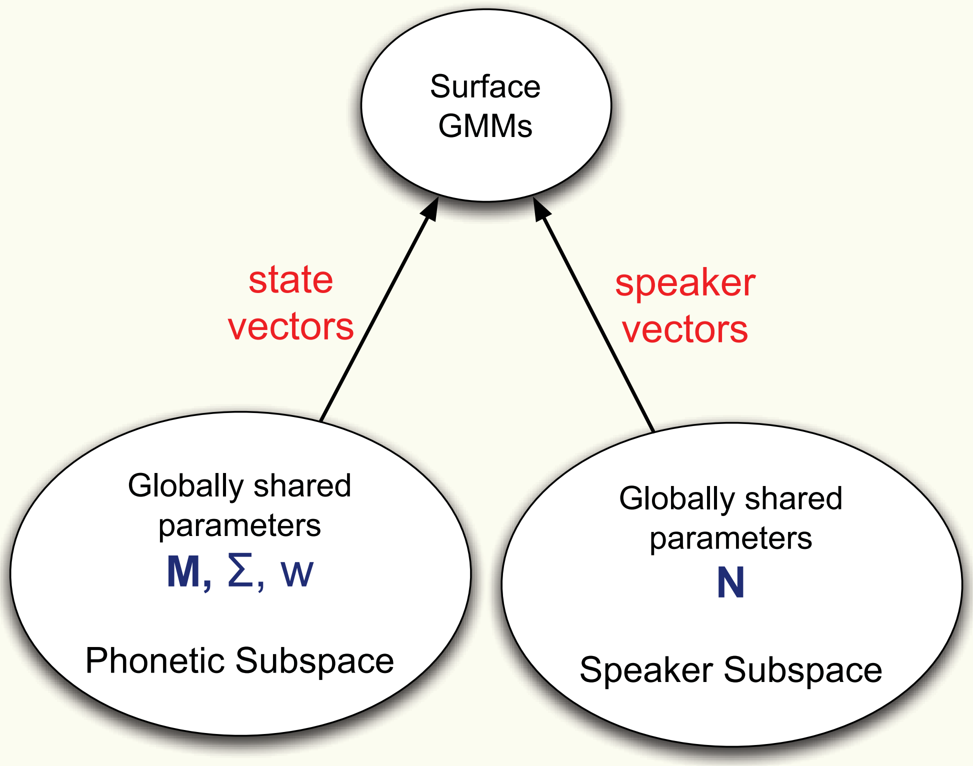
Factorisation

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Control effects of speaker, accent, style, acoustic environment, ...
- Factor different causes of variability (eg speaker, environment)
- Different approaches for both acoustic models & language models
- Allow use of speech knowledge in new situations
- Allows rapid adaptation to new situation or domain
- Can use subspace/canonical-state models for acoustic models
- Factorisation in recognition ⇔ control in synthesis

Subspace GMMs

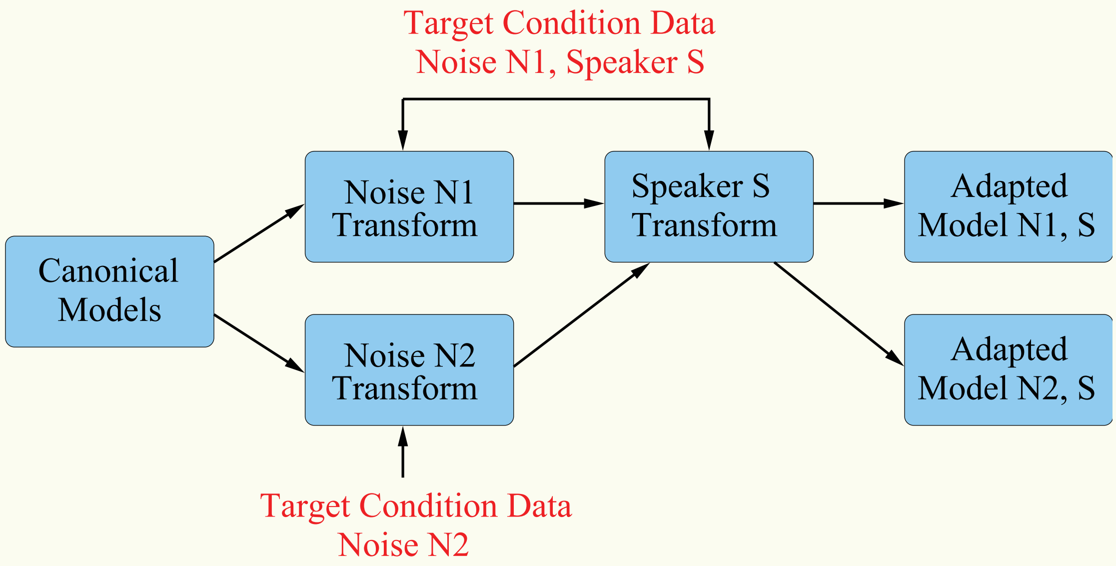
Natural Speech Technology
Edinburgh – Cambridge – Sheffield



http://kaldi.sourceforge.net/

Canonical models for acoustic factorisation

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



Use of Rich Contexts

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Extend the notion of context in speech recognition
  - beyond phone context in acoustic models
  - beyond word N-grams in language models
- More detailed context
  - Include extra information into models cf synthesis models
  - Model general situation - highly specialised domains (speaker, location etc)
  - Adapt structured acoustic and language models

Metadata Use/Generation

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Generate additional information to augment / modify output.
- Some metadata may be given and is part of rich context.
  - Use metadata to generate appropriate recognition models
  - Factorised form allows rapid model generation
- Generate information on
  - Speakers & acoustic conditions
  - Emotional state
  - Sentence boundaries (slash units), disfluencies etc
  - Genre/style (eg meeting, comedy show etc etc)

Diverse Accents/Dialects

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Structured accent models
  - Build on canonical models to capture accents & dialect variations
- Cross-lingual recognition
  - Minority language transcription with limited linguistic resources
  - Build on language independent knowledge of speakers, environments etc
  - Use models to help with areas normally requiring expert knowledge (e.g. pronunciation dictionaries)
  - Use factorised multi-level models

Environment Models

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Models that represent the physical acoustic environment. This includes representations of speaker and sound source location.

- Objective Robustness
  - Reverberation
  - Noise
  - Several moving speakers
- Recording facilities
  - Multichannel recordings (audio/video) with speaker location
  - Digital MEMS microphone array

Generating Fluent Output

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Speech recognition systems normally give literal output
- For many applications need a more ***natural, readable*** output - need to define naturalness ... (grammatical, acceptable)
- Transform/translate output as required for application
- Create suitable models of fluent data
  - Initial study based on use of N-grams and Combinatory Categorical Grammars (CCG): generation allowing substitution, permutation, insertion, deletion from ASR output
  - Promising initial results - but hard to formally evaluate!

Applications: HomeService

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Personalised, interactive speech technology which can interact with environmental control systems and home monitoring device
- Integration of synthesis and recognition, to provide better interfaces to assistive technology
- Focus on older people and disabled people (& deal with dysarthric speech)
- Closely linked to work on voice restoration and voice banking

HomeService

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- PC "box"; will run immediate audio collection and processing

- Microcone; can provide speaker diarization information.

- Infrared transmitter for controlling TV, set-top box, curtains etc.

- Android tablet; mounted on wheel chair. Interface between system and user.

homeService - use case

Natural
Speech
Technology
Edinburgh – Cambridge – Sheffield

Initial Results on Dysarthric Speech

- UASPEECH corpus
  - 18 speakers
  - 1000 different words spoken several times
  - Speakers are assessed in terms of medical speech quality level covering a range from 2% - 95%.
- Best results so far
  - General triphone HMMs - MAP adapted to domain and speakers
  - Word correctness by speaker is ranging from 95% to 5% inverse to speech quality assessment, average 55%.

Application: Transcribing Media Archives

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- BBC providing large quantities of data (aim to make all archives publicly available with searchable transcripts by BBC centenary in 2022)
  - Many genres and styles of broadcast audio data
    - radio incl. news, interviews/discussions, radio dramas
    - TV incl dramas, comedy, chat shows etc
  - Meta-data exists at various levels/types/completeness/accuracy: speakers, topics, genres & styles, subtitles, ...
  - Aim to structure models to produce targeted models given all available context
  - Pilot systems with wide range of error rates (!) from well under 20% (discussions, news) to over 60% (TV drama)

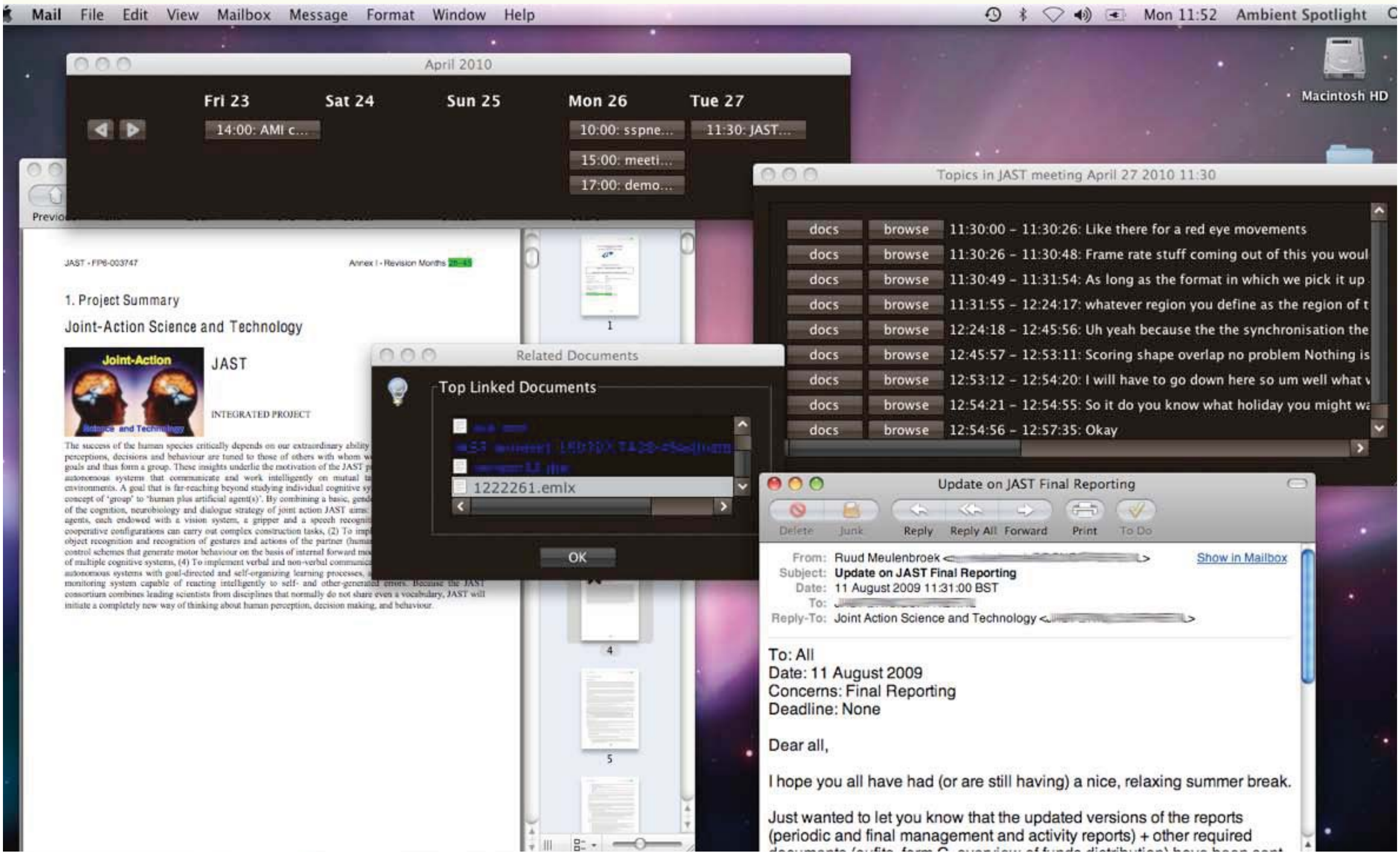
Application: Personalised Transcription

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Aim to create ***personalised*** transcription devices
  - Analyse data over extended period from particular speaker
  - High levels of adaptation/specialisation at all levels
  - Access to rich context about user and also those with which the user interacts
- Examples
  - ***lifeLog:*** Wearable, gathers information on user and the world with which the user commonly interacts
  - ***Ambient Spotlight:*** Recording, transcribing tutorials, linking to other related material

Ambient spotlight

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



Natural Synthesis: Standard Systems

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- State-of-the-art: Intelligible synthetic voices, but not perceived as natural, very limited expressivity.
- **Unit selection**
  - large, carefully segmented inventory of speech from single speaker
  - inflexible method based on cut-and-paste of phone-sized units
  - widespread commercial use
- **Statistical parametric ("HMM-based")**
  - flexible method based on a statistical model of the speech signal
  - amenable to various modifications
  - can be learned from speech of multiple speakers ('average voice model')
  - can be adapted to new speakers, etc.

Natural Synthesis

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Long term vision: Fully controllable speech synthesis, indistinguishable from a human voice, with high intelligibility in all acoustic conditions.
- Goals within NST
  - Statistical parametric synthesis,
  - controllable in terms of speech parameters,
  - adaptable without new data,
  - personalisable with minimal data,
  - high degree of expressivity if required.

Examples of synthesis work

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Creating voices from existing data, instead of high-quality studio recordings
- Speech-to-speech translation
- New models and extensions of HMM-based speech synthesis
- Clinical applications - voice output communication aids
- Voice cloning

Voice cloning from normal speech

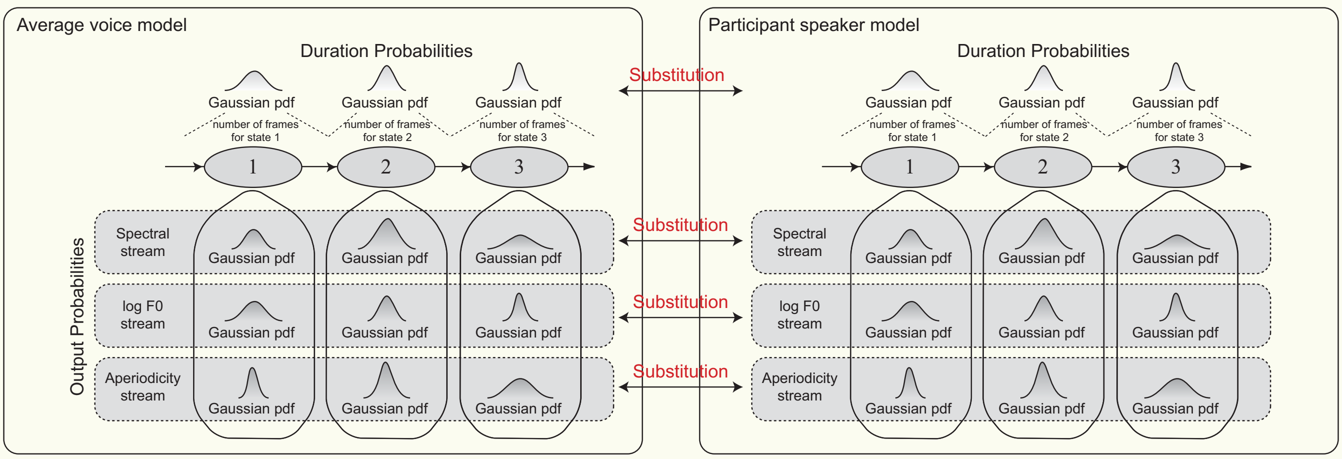
Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Example voices
  - only 100 recorded sentences
  - non-professional speakers (volunteer ‘voice donors’)
  - average voice model adapted to this data

(examples)

Reconstructing disordered speech: clinical applications

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



- “Mix-and-match” model components
  - some **copied** from an average voice model
  - some **learned** from the target speaker (patient)
  - some **adapted** from the average voice model to the target speaker

Reconstructed voices - Motor Neurone Disease

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

A personalised voice output communication aid

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Intelligibility

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- **Intelligibility** is almost a solved problem, in *good listening conditions*
- in *challenging conditions*, or for hearing-impaired listeners, much work still to be done
  - synthesisers that ***adapt*** to the environment and the listener
  - initial work has been on speech in additive noise
  - can already get intelligibility gains in some conditions, without altering signal-to-noise-ratio (SNR)
- **What we plan to do next**: adaptation to new listening environments without requiring additional data

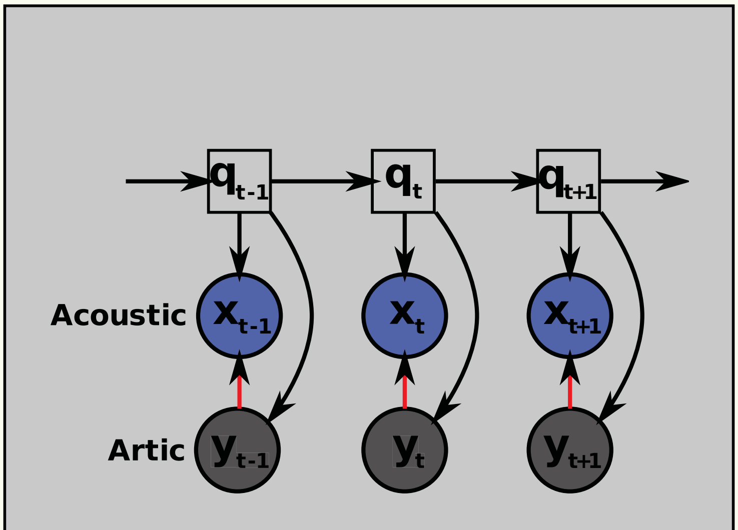
Naturalness and quality

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

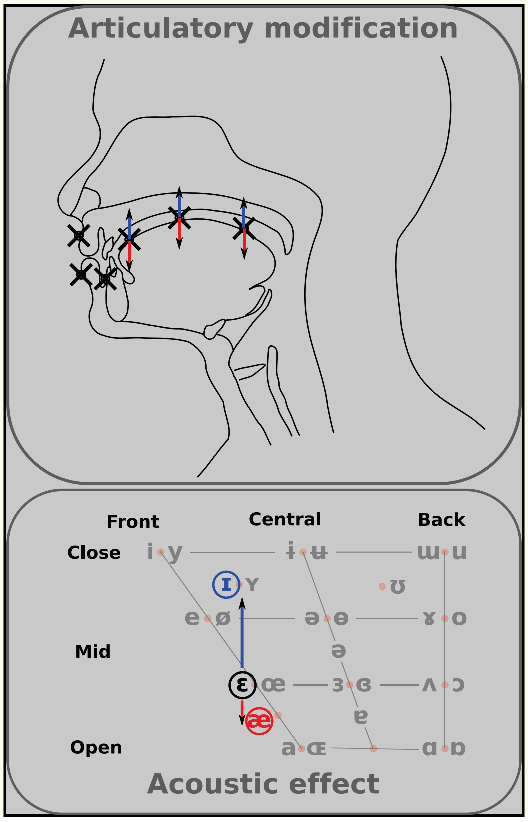
- **Naturalness** is far from being solved
  - Can achieve reasonable read-text speaking style using concatenative techniques, but these offer no flexibility or control
  - HMM-based method currently slightly less natural, but much more flexible
- **Quality** of HMM-based synthetic speech slightly behind the best unit-selection, but catching up fast
- **What we plan to do next**: factored models that explicitly represent the structure of speech processes, the input text, and the listening situation

Speech knowledge
Acoustic-Articulatory TTS

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



Feature-dependency model



Ling, Richmond, Yamagishi & Wang, 2009

Example:
Controlling tongue height

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Tongue height (cm)

+1.5	
+1.0	
+0.5	
default	peck
-0.5	
-1.0	
-1.5	

Summary

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- High profile UK funded programme on speech technology: aiming to improve naturalness for both recognition and synthesis
- Many aspects being investigated including
  - Use of rich context information to build highly adapted / focused systems from limited training data
  - Use of factorised models
  - Personalisation
  - Healthcare applications including voice restoration
- Large user group

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Thanks.

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