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

24 April 2012

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

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

NST facts and figures

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- EPSRC Programme Grant, May 2011–April 2016
- Three university partners, user group, scientific advisory board
- When at full speed
  - 8 investigators
  - ~10 postdoctoral researchers
  - 7 PhD students
- Strong links with other projects: EPSRC, EU, DARPA, NHS, JST, etc.

NST team

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- **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

EPSRC
Engineering and Physical Sciences Research Council

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

Current speech technology: 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

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

Research 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

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Goal within NST – 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
- *Talk by Phil Woodland*

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.

- *Talk by Simon King*

Application: 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
- *Talk by Heidi Christensen*

Application: Media Archives

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Long-term objective: make broadcast media archives accessible and searchable

- Many genres and styles of broadcast audio data
  - radio incl. news, interviews/discussions, radio dramas
  - TV incl dramas, comedy, chat shows etc

- Make use of existing metadata, automatically generate metadata

- Piloting systems using BBC data – very wide range of error rates (!)

- *Demo*

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
  - Rich context about user and those with whom 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
- *Demo*

User Group Members

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Barnsley Hospital, Medical Physics & Clinical Engineering
- Devices for Dignity
- Euan MacDonald Centre for Motor Neurone Disease Research
- NIHR CLAHRC for South Yorkshire
- Toby Churchill Ltd
- BBC Future Media & Technology
- Cereproc
- Cisco Systems
- EADS UK
- GCHQ
- Novauris
- Nuance Communications
- Toshiba Research Europe

User group interactions

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Support for, and co-supervision of PhD students
- Direct financial support for NST projects (and related pilot projects)
- Provision of data, domain requirements, and domain expertise
- NST involvement in specific projects at user group members
- General discussion

First year highlights

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Learning and adaptation
  - acoustic factorisation, subspace models
- Natural transcription
  - transcription of BBC broadcast data
  - cross-lingual recognition
  - fluent ASR
- Natural synthesis
  - voice reconstruction / voice banking
  - novel models for speech synthesis
- homeService application and recognition of disordered speech